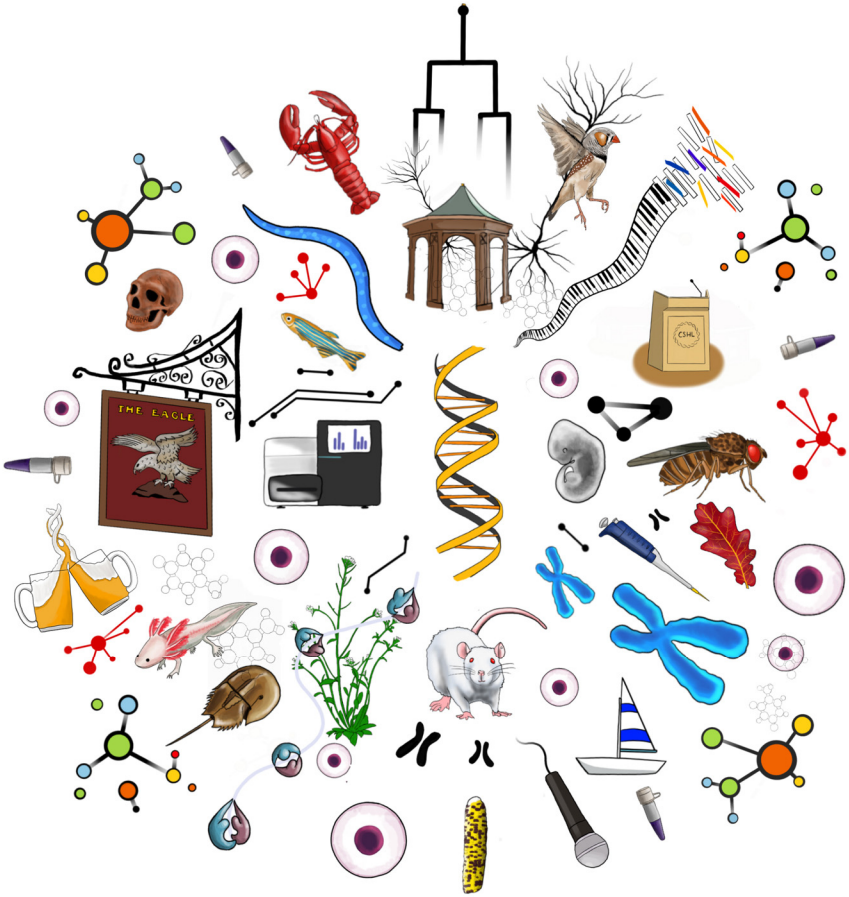
Abstracts of papers presented
at the 2016 meeting on

THE BIOLOGY OF GENOMES

May 10–May 14, 2016



CSH Cold Spring Harbor Laboratory 1890 125 2015

Abstracts of papers presented
at the 2016 meeting on

THE BIOLOGY OF GENOMES

May 10–May 14, 2016

Arranged by

Ewan Birney, *EBI/EMBL, UK*
Michel Georges, *University of Liège, Belgium*
Jonathan Pritchard, *Stanford University*
Molly Przeworski, *Columbia University*

This meeting was funded in part by the **National Human Genome Research Institute,** a branch of the **National Institutes of Health; Fluidigm; Illumina;** and **Swift Biosciences.**

Contributions from the following companies provide core support for the Cold Spring Harbor meetings program.

Corporate Sponsors

Agilent Technologies
Bristol-Myers Squibb Company
Calico Labs
Genentech, Inc.
Life Technologies (part of Thermo Fisher Scientific)
New England BioLabs

Plant Corporate Associates

Monsanto Company

Cover: Created by Alex Cagan, a PhD student at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

THE BIOLOGY OF GENOMES

Tuesday, May 10 – Saturday, May 14, 2016

Tuesday	7:30 pm	**1** Population Genomics
Wednesday	9:00 am	**2** Functional Genomics
Wednesday	2:00 pm	**3** Poster Session I
Wednesday	4:30 pm	*Wine and Cheese Party**
Wednesday	7:30 pm	**4** Computational Genomics
Thursday	9:00 am	**5** Cancer / Medical Genomics
Thursday	2:00 pm	**6** Poster Session II
Thursday	4:30 pm	**7** ELSI Panel and Discussion
Thursday	7:30 pm	**8** Evolutionary and Non-human Genomics
Friday	9:00 am	**9** Translational Genomics and Genetics
Friday	2:00 pm	**10** Poster Session III
Friday	4:30 pm	GUEST SPEAKERS
Friday	6:00 pm	Banquet
Saturday	9:00 am	**11** Genetics of Complex Traits

Workshops (immediately following morning sessions)
Illumina, Wednesday
Fluidigm, Thursday

* *Airslie Lawn*, weather permitting

Mealtimes at Blackford Hall are as follows:
Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm
Bar is open from 5:00 pm until late

Printed on 100% recycled paper.

PROGRAM

TUESDAY, May 10—7:30 PM

SESSION 1 POPULATION GENOMICS

Chairpersons: **Nels Elde,** University of Utah, Salt Lake City
Joseph Pickrell, New York Genome Center, New York

Striking differences in patterns of germline mutation in mice and humans
Sarah J. Lindsay, Raheleh Rahbari, Matt E. Hurles.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 6

Breaking the infinite sites model—Widespread mutational recurrence in exome sequence data from over 60,000 individuals
Konrad J. Karczewski, Monkol Lek, Eric V. Minikel, Kaitlin E. Samocha, Exome Aggregation Consortium, Mark J. Daly, Daniel G. MacArthur.
Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts; Broad Institute, Cambridge, Massachusetts. 7

Detecting 2,000 years of human genetic adaptation
Yair Field, Evan A. Boyle, Natalie Telis, Ziyue Gao, Kyle J. Gaulton, David Golan, Loic D. Yengo, Mark McCarthy, Jonathan K. Pritchard.
Presenter affiliation: Howard Hughes Medical Institute, Stanford, California; Stanford University, Stanford, California. 8

Happy Hour
Sponsored by **Illumina**

WEDNESDAY, May 11—9:00 AM

SESSION 2 FUNCTIONAL GENOMICS

Chairpersons: **William Greenleaf,** Stanford University, California
Deborah Winter, Weizmann Institute of Science, Rehovot, Israel

Willliam Greenleaf.
Presenter affiliation: Stanford University, California.

High-throughput, unbiased CRISPR mutagenesis of the human noncoding genome
Neville E. Sanjana, Jason Wright, Feng Zhang.
Presenter affiliation: New York Genome Center, New York, New York; New York University, New York, New York. 9

SESSION 3 POSTER SESSION I

Cancer genome assembly and structural variant detection with Bionano optical mapping and Pacific Bioscience long reads
Sara Goodwin, Maria Nattestad, Karen Ng, Timour Baslan, Tyler Garvin, James Gurtowski, Elizabeth Hutton, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John McPherson, James Hicks, Michael C. Schatz, W. Richard McCombie.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor , New York. 92

Ancient whole dog genomes show no evidence of population replacement in Neolithic Europe
Shyamalika Gopalan, Laura Botigué, Shiya Song, Amelie Scheu, Timo Seregély, Andrea Zeeb-Lanz, Rose-Marie Arbogast, Kevin Daly, Martina Unterländer, Angela Taravella, Matthew Oetjens, Amanda Pendleton, Dan Bradley, Jeffrey M. Kidd, Joachim Burger, Krishna R. Veeramah.
Presenter affiliation: Stony Brook University, Stony Brook, New York. 93

When is selection effective?
Simon Gravel.
Presenter affiliation: McGill University, Montreal, Canada. 94

Comprehensive characterization of RNA elements in the human genome
Brenton R. Graveley, Chris Burge, Xiang-Dong Fu, Eric Lecuyer, Eugene W. Yeo.
Presenter affiliation: UCONN Health, Farmington, Connecticut. 95

Sequence co-evolution predicts residue-level protein interactions
Anna G. Green, Thomas A. Hopf, Charlotta P. Schärfe, Debora S. Marks.
Presenter affiliation: Harvard Medical School, Boston, Massachusetts. 96

The human microbiome as surveyed using a rapid, culture-free whole genome assembly approach
Nicholas Putnam, Jonathan Stites, Robert Calef, Paul Havlak, Marco Blanchette, Ei Ei Min, Brendan O'Connell, Richard Green.
Presenter affiliation: Dovetail Genomics, Santa Cruz, California; University of California, Santa Cruz, Santa Cruz, California. 97

WEDNESDAY, May 11—4:30 PM

Wine and Cheese Party

WEDNESDAY, May 11—7:30 PM

SESSION 4 COMPUTATIONAL GENOMICS

Chairpersons: **Eran Segal,** Weizmann Institute of Science, Rehovot, Israel
Alexsandra Walczak, CNRS/ENS, Paris, France

THURSDAY, May 12—9:00 AM

SESSION 5 CANCER / MEDICAL GENOMICS

Chairpersons: **Serena Nik-Zainal,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
Heidi Rehm, Harvard Medical School, Boston, Massachusetts

THURSDAY, May 12—2:00 PM

SESSION 6 POSTER SESSION II

SESSION 7 ELSI PANEL AND DISCUSSION

Participant Rights to Their Sequence Data: The Pros, Cons, and Pragmatics of Returning the Incidental Genome

Moderator: **Dave Kaufman,** National Human Genome Research Institute, National Institutes of Health

Panelists

Misha Angrist, Duke University
Jason Bobe, Icahn School of Medicine at Mount Sinai
Mildred Cho, Stanford School of Medicine
Wendy Chung, Columbia University

Although the availability of direct-to-consumer genetic testing has decreased substantially over the past few years, a great deal of extant and future genomic data resides with researchers and clinicians. The tenet that study participants have a right to request and receive their genomic data when it is held by a researcher or a clinician has been discussed in depth. "Nothing about me without me", a statement made at one of the four public workshops held in 2015 to inform the design of a nationwide Precision Medicine Initiative (PMI™), encapsulates this belief. The right to at least some of one's genomic data was acknowledged and included in the official final report from the group advising the NIH on the design of the large PMI™ cohort. The report notes: "*building a true partnership with participants requires respecting participant preferences about the return of information and avoiding large disparities in data and information access between researchers and participants. Sharing information, including both data generated from biospecimens and new interpretations of data resulting from research, is critical to the success of the PMI™ …"*.

The authors of the PMI™ report were cognizant of the numerous technical and ethical challenges associated with returning genomic data. Returning clinical genomic data requires the use of CLIA certified laboratories that follow specified procedures, or FDA-cleared medical devices—elements that increase the cost of obtaining these data. Clinically actionable results may require complex interpretation by medical practitioners, as well as communication strategies and infrastructure to assure that medical follow-up is made available to all research participants. Whether and how to address inactionable results, variants of unknown significance, and results whose interpretation is likely to change can compound the questions. Research staff are often not equipped or experienced in making the myriad decisions potentially involved in returning such data.

Although the PMI™ is bringing the issue to the fore, it is not the first research initiative to consider whether an obligation exists to return genomic data, to measure how difficult this might be, or to contemplate how it might be accomplished. This session brings together a panel to discuss whether study participants have a right to obtain all of their data. The panel will also discuss the format in which data should be made available - BAM files, clinical variant reports, or something in between. It will also discuss why the return of genomic data is not currently the routine practice, and how the return of such data might be practically accomplished should it become the norm.

THURSDAY, May 12—7:30 PM

SESSION 8 EVOLUTIONARY AND NON-HUMAN GENOMICS

Chairpersons: **Felicity Jones,** Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
Ludovic Orlando, University of Copenhagen, Denmark

Felicity Jones
Presenter affiliation: Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany.

A genetic signature of flightlessness evolution in the Galapagos cormorant (*Phalacrocorax harrisi*) revealed by predictive genomics
Alejandro Burga, Weiguang Wang, Paul Wolf, Andy Ramey, Claudio Verdugo, Karen Lyons, Patricia Parker, Leonid Kruglyak.
Presenter affiliation: UCLA / HHMI, Los Angeles, California. 229

Shared genetics of obsessive compulsive disorder in dogs and humans
Elinor K. Karlsson, Hyun Ji Noh, Guoping Feng, Kerstin Lindblad-Toh.
Presenter affiliation: Umass Medical School, Worcester, Massachusetts; Broad Institute, Cambridge, Massachusetts. 230

Influence of diet, parasitism and host genetics on the biodiversity of the human gut microbiota in rural populations from Cameroon
Laure Segurel, Elise Morton, Alain Froment, Evelyne Heyer, Molly Przeworski, Ran Blekhman.
Presenter affiliation: Musée de l'Homme, Paris, France. 231

Evolutionary genomics of the horse domestication process
Ludovic Orlando.
Presenter affiliation: University of Copenhagen, Copenhagen, Denmark. 232

Species genome sequencing of the endangered Spix's macaw
Iman K. Al-Azwani, Nancy Chen, Cristina Yumi Miyaki, Yasmin A. Mohamoud, Andrew G. Clark, Cromwell Purchase, Joel A. Malek.
Presenter affiliation: Weill Cornell Medicine-Qatar, Doha, Qatar; Weill Cornell Medicine, New York, New York. 233

How social status changes the immune system—Experimental evidence from rhesus macaques
Noah Snyder-Mackler, Joaquin Sanz, Jessica Brinkworth, Jordan Kohn, Zachary Johnson, Mark Wilson, Luis Barreiro, Jenny Tung.
Presenter affiliation: Duke University, Durham, North Carolina. 234

Variation in the molecular clock of primates
Priya Moorjani, Carlos Eduardo G. Amorim, Peter F. Arndt, Molly Przeworski.
Presenter affiliation: Columbia University , New York, New York; Broad Institute, Cambridge, Massachusetts. 235

Happy Hour
Sponsored by **Swift Biosciences**

FRIDAY, May 13—9:00 AM

SESSION 9 TRANSLATIONAL GENOMICS AND GENETICS

Chairpersons: **Matthew Hurles,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
Sally John, Biogen Idec, Inc., Cambridge, Massachusetts

The prevalence and architecture of severe, dominant developmental disorders
Matthew Hurles.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 236

FRIDAY, May 13—2:00 PM

SESSION 10 POSTER SESSION III

SMASH, a fragmentation and sequencing method for genomic copy number analysis
Zihua Wang, Peter Andrews, Jude Kendall, Beicong Ma, Inessa Hakker, Linda Rodgers, Michael Ronemus, Michael Wigler, Dan Levy.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 310

Integrated genomic analysis with IOBIO
Alistair Ward, Chase Miller, Tonya Di Sera, Yi Qiao, Brent Pedersen, Aaron Quinlan, Gabor Marth.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 311

Allelic specific expression analysis of structural variation in human populations
Jia Wen, Andrew Quitadamo, Xinghua Shi.
Presenter affiliation: Jia Wen, Charlotte, North Carolina. 312

Apply empirical bayesian elastic net method to microRNA epistasis analysis in colon cancer
Jia Wen, Benika Hall, Andrew Quitadamo, Xinghua Shi.
Presenter affiliation: University of North Carolina-Charlotte, Charlotte, North Carolina. 313

Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin
Sean Whalen, Rebecca M. Truty, Katherine S. Pollard.
Presenter affiliation: Gladstone Institutes, San Francisco, California. 314

Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations
John W. Whitaker, Alice M. Walsh, Chris C. Huang, Yauheniya Cherkas, Sarah L. Lamberth, Carrie Brodmerkel, Mark E. Curran, Radu Dobrin.
Presenter affiliation: Janssen Research and Development, LLC, San Diego, California. 315

Genome-wide assessment of the contribution of short tandem repeats to de novo variation
Thomas Willems, Melissa Gymrek, David Poznik, Chris Tyler-Smith, Yaniv Erlich.
Presenter affiliation: New York Genome Center, New York, New York; Whitehead Institute for Biomedical Research, Cambridge, Massachusetts; MIT, Cambridge, Massachusetts. 316

Robust transcriptome-wide discovery of RNA binding protein binding sites with enhanced CLIP (eCLIP) and evaluation of impact of natural and disease-causing variants on RNA binding
Eric V. Nostrand, Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Fang, Balaji Sundararaman, Steven Blue, Thai Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, Eugene Yeo.
Presenter affiliation: University of California San Diego, La Jolla, California. 323

Unbiased estimation of heritability by relatedness disequilibrium regression reveals overestimation of heritability by twin studies
Alexander I. Young, Michael L. Frigge, Kari Stefansson, Augustine Kong.
Presenter affiliation: University of Oxford, Oxford, United Kingdom. 324

A role in programmed DNA deletion for the second domesticated piggyBac transposase TPB1 in *Tetrahymena thermophila*
Chao-Yin Cheng, Janet M. Young, Chih-Yi Lin, Harmit S. Malik, Meng-Chao Yao.
Presenter affiliation: Fred Hutchinson Cancer Research Center, Seattle, Washington. 325

Causal variants in metabolite quantitative trait loci
Noha A. Yousri, Khalid A. Fakhro, Amal Robay, Juan L. Rodriguez-Flores, Ronald G. Crystal, Karsten Suhre.
Presenter affiliation: Weill Cornell Medical College-Qatar, Doha, Qatar; Alexandria University, Alexandria , Egypt. 326

Insights into the performance of whole-exome sequencing technologies
Yao Yu, Hao Hu, Jerry Fowler, Yuanqing Ye, Michelle Hildebrandt, Hua Zhao, Paul Scheet, Xifeng Wu, Chad D. Huff.
Presenter affiliation: The University of Texas MD Anderson Cancer Center, Houston, Texas. 327

Real-time person identification using noisy error-prone DNA sequencing data and incomplete databases.
Sophie Zaaijer, Robert Piccone, Daniel Speyer, Yaniv Erlich.
Presenter affiliation: New York Genome Center, New York , New York; Columbia University, New York, New York. 328

Integrative analysis of multi "omics" data identifies functional Mediators as intervention points for global phenotypes
Chenchen Zhu, Christopher S. Hughes, Michelle Nguyen, Lars M. Steinmetz.
Presenter affiliation: European Molecular Biology Laboratory, Heidelberg, Germany. 336

Gene similarity network reveals sub-populations of cells in single-cell RNA-seq data
Bo Wang, Jesse Zhang, Junjie Zhu, Serafim Batzoglou.
Presenter affiliation: Stanford University, Stanford, California. 337

FRIDAY, May 13—4:30 PM

GUEST SPEAKERS

Emmanuelle Charpentier
Max Planck Institute for Infection Biology, Germany

Neil Shubin
University of Chicago

FRIDAY, May 13

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SESSION 11 GENETICS OF COMPLEX TRAITS

Chairpersons: **Carolyn McBride,** Princeton University, New Jersey
Soumya Raychaudhuri, Brigham and Women's Hospital, Boston, Massachusetts

Genomic signatures of mosquito adaptation to the amino acid content of human blood
Zhilei Zhao, Carolyn S. McBride.
Presenter affiliation: Princeton University, Princeton, New Jersey. 338

Whole genome sequence variants influence multiple amino acids levels
Bing Yu, Elena Feofanova, Donna Muzny, Alanna C. Morrison, Richard A. Gibbs, Eric Boerwinkle.
Presenter affiliation: University of Texas Health Science Center at Houston, Houston, Texas. 339

***Trans*-regulatory architecture of genetic transcriptome variation from 1,000 yeast individuals**
Frank W. Albert, Joshua S. Bloom, Jake Siegel, Laura Day, Leonid Kruglyak.
Presenter affiliation: University of California-Los Angeles, Los Angeles, California; University of Minnesota, Minneapolis, Minnesota. 340

Coupling genomic sequencing analyses with genome editing to reveal a role for alternative *GFI1B* splice variants in human hematopoiesis
Linda M. Polfus, Rajiv K. Khajuria, Ursula M. Schick, Alex P. Reiner, Santhi K. Ganesh, Vijay G. Sankaran.
Presenter affiliation: University of Texas Health Science Center , Houston, Texas. 341

Soumya Raychaudhuri.
Presenter affiliation: Brigham and Women's Hospital, Boston, Massachusetts.

THE GENEROSITY OF SELFISH GENES IN THE EVOLUTION OF IMMUNE DEFENSES

Nels C Elde, Edward B Chuong, Alesia N McKeown, Diane M Downhour, Cedric Feschotte

University of Utah, Human Genetics, Salt Lake City, UT

Transposable elements (TEs) propagate in genomes. Endogenous retroviruses (ERVs) are an abundant class of TEs containing sequences modulating transcription. The influence of ERV propagation on the evolution of gene regulation remains largely unknown. Analysis of ChIP-seq data suggests that ERVs have shaped the evolution of a transcriptional network underlying the interferon (IFN) response, a major component of innate immunity. Lineage-specific ERVs have dispersed numerous IFN-inducible enhancers independently in diverse mammalian genomes. CRISPR-Cas9 mediated deletion of a subset of these ERV elements in the human genome impaired expression of adjacent IFN-induced genes and revealed their involvement in the regulation of essential immune functions. Although these regulatory sequences likely arose in ancient viruses, they now constitute a dynamic reservoir of IFN-inducible enhancers fueling genetic innovation in mammalian immune defenses.

In addition to dispersing numerous duplicate copies, selfish genetic elements encoded by retrotransposons occasionally capture and distribute host transcripts. Most resulting copies are nonfunctional pseudogenes, but others gain novel functions as retrogenes. The birth of a new retrogene might represent a pivotal adaptation if the new encoded function promotes fitness. We discovered new classes of retrogenes providing adaptations at host-pathogen interfaces. One such retrogene excavated from the genome of squirrel monkeys encodes potent antiviral activity impeding the release of retroviruses, including HIV, from membranes of infected host cells. The retrogene is derived from a critical component of the endosomal sorting complexes required for transport (ESCRT) pathway, which is commandeered by many enveloped viruses to bud nascent particles. A retrogene copy of charged multi-vesicular body protein 3 (CHMP3) provides dominant negative inhibition of viral release. Remarkably, natural selection refined the function of retroCHMP3, such that it does not block essential ESCRT-mediated host functions while blocking viral release from host cells, highlighting the potential of retrogenes as a new class of antiviral therapeutics. Together our work suggests a new model for the origins of innate immune functions through the activity of diverse selfish genetic elements. Because selfish genes continue to propagate in populations, these elements may continuously provide new functions in the pervasive genetic conflicts defining interactions between pathogens and their hosts.

← Back

POPULATION GENOMICS OF UPPER PALEOLITHIC EUROPE

Qiaomei Fu[1,2,3], Cosimo Posth*[4,5], Mateja Hajdinjak*[3], Martin Petr[3], Swapan Mallick[2,6,7], Janet Kelso[3], Nick Patterson[6], Johannes Krause[4,5,8], David Reich[2,6,7], Svante Pääbo[3]

[1]Chinese Academy of Sciences (CAS), Key Laboratory of Vertebrate Evolution and Human Origins (IVPP), Beijing, China, [2]Harvard Medical School, Department of Genetics, Boston, MA, [3]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [4]University of Tübingen, Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, Tübingen, Germany, [5]Max Planck Institute for the Science of Human History, Jena, Germany, [6]Broad Institute of MIT and Harvard, Cambridge, MA, [7]Harvard Medical School, Howard Hughes Medical Institute, Boston, MA, [8]University of Tübingen, Senckenberg Centre for Human Evolution and Palaeoenvironment, Tübingen, Germany

Whereas genomic data from large numbers of individuals have been used to study the population history of Europe from the start of agriculture ~8,500 years ago, little is known about European population history before that time. We have analyzed genome-wide data from 51 Eurasians from ~45,000-7,000 years ago. Over this time, the proportion of Neanderthal DNA carried by European individuals decreased from 3-6% to around 2%, consistent with natural selection against Neanderthal variants in the modern human genetic background. Whereas the earliest modern humans in Europe did not contribute substantially to present-day Europeans, all individuals between ~37,000 and ~14,000 years ago descended from a single founder population and form part of the ancestry of present-day Europeans. A ~35,000-year-old individual from northwest Europe represents an early branch of this founder population which was then displaced across large parts of Europe, before reappearing in southwest Europe during the Ice Age ~19,000 years ago. About 14,000 years ago during the first major warming period after the Ice Age a new genetic component related to present-day Near Easterners appears in Europe. Thus, population turnover and migration were recurring themes in European pre-history before the advent of agriculture.

GENOMIC PATTERNS OF SELECTION THROUGH TIME IN A WILD PEDIGREED POPULATION

Nancy Chen[1,2], Elissa J Cosgrove[3], Huijie Feng[3], Ishaan A Jhaveri[3], Ashish Akshat[3], Reed Bowman[4], John W Fitzpatrick[2], Andrew G Clark[3]

[1]University of California, Davis, Center for Population Biology, Davis, CA, [2]Cornell University, Cornell Lab of Ornithology, Ithaca, NY, [3]Cornell University, Molecular Biology & Genetics, Ithaca, NY, [4]Archbold Biological Station, Avian Ecology, Venus, FL

Recent studies have demonstrated evolution on ecological timescales in a number of different organisms. Studying contemporary evolution is the only way to directly test many fundamental questions in evolutionary biology, but we usually lack the combined phenotypic and genomic data over time required for such studies in natural populations. In addition, most current approaches for inferring natural selection do not take advantage of the additional power gained from considering the full pedigree. Here, we study short-term selection using a 25-year genomic, phenotypic, and pedigree dataset in the Florida Scrub-Jay (*Aphelocoma coerulescens*), an iconic species on the U. S. Endangered Species List that has drastically decreased in number during the past half-century. A population of Florida Scrub-Jays at Archbold Biological Station has been studied since 1969, resulting in a 12-generation pedigree that is one of the most accurate and extensive for any wild vertebrate species. For all birds in the population, we have full records of individual lifespans as well as annual fecundity and lifetime fitness measures. We sequenced and assembled the Florida Scrub-Jay genome and used custom Illumina Beadchips to genotype every individual in our study population over the past two decades (3,838 individuals total) at 15,416 genome-wide SNPs. We used gene dropping to explicitly sample gametes in each generation on the known pedigree and asked whether the observed allele frequency dynamics of each SNP were consistent with a pure drift process. We identified 67 SNPs that departed significantly (FDR of 10%) from the null model and whose frequency dynamics were driven by selection. We then tested for selection acting on specific life-cycle stages by modifying existing selection component analysis frameworks to take full advantage of exhaustive population sampling. We identified a number of loci that clearly exhibited male gametic selection, sexual selection, and viability selection. By combining sensitive pedigree-based inferences of net selection with fine-scale dissection of selection components, this study provides a detailed assessment of the role of selection in perturbing allele frequency dynamics in a rapidly declining population. Results suggest a role of selection in maintaining variation even in the face of population decline and may help guide conservation efforts.

← Back

QUANTIFYING SELECTION AND DEMOGRAPHIC EFFECTS ON QUANTITATIVE GENETIC VARIATION: AN APPLICATION TO HUMAN HEIGHT

Guy Sella, Yuval B Simons

Columbia University, Biological Sciences, New York, NY

The genetic architecture of a quantitative phenotype (i.e., the number, frequency and effect size of alleles underlying variation in its value) arises from genetic and population genetic processes. Mutations affecting the trait appear at a rate that reflects the target size, and their trajectory through the population is determined by demographic processes and by the selection acting on them. Many phenotypes, including human height, appear to be under stabilizing selection, either because of selection on the trait itself or through the effects of genetic variation on other traits (i.e., via pleiotropy).

With these considerations in mind, we introduce and solve a generative model for the genetic architecture of a continuous trait under direct and pleiotropic stabilizing selection. We derive simple and robust predictions for the distribution of additive genetic variation among loci. We then relate these predictions to observations from GWAS, accounting for how the power to detect a locus depends on its contribution to additive genetic variation.

This new theory allows us to make inferences about the population genetic processes that underlie genetic variation for height in Europeans. We find an extremely good fit to GWAS findings (Wood et al. Nature Genetics 2014): by fitting a single parameter, we are able to explain the distribution of additive genetic variation over the ~700 genome-wide significant associations. Accounting for the demographic history of European populations suggests that the current GWAS is well powered to identify only loci under moderate selection. This relatively weak selection explains why the majority of loci that have been associated with height in Europeans are also segregating in African populations. We estimate the target size and distribution of selection coefficients of mutations affecting height within the range in which the current GWAS is well powered. We then employ these estimates to predict the expected increase in explained heritability with GWAS size due to variants in this range. Our results also suggest how increasing study size will enable the discovery of loci experiencing a wider range of selection effects. The framework presented here can be applied much more broadly, to investigate the genetic and selection parameters governing variation in other quantitative traits.

← Back

THE IDENTIFICATION OF GENETIC VARIANTS THAT IMPACT VIABILITY IN LARGE COHORTS

Hakhamanesh Mostafavi[1], Tomaz Berisa[2], Molly Przeworski[3], Joseph Pickrell[2,3]

[1]Columbia University, Department of Chemical Engineering, New York, NY, [2]New York Genome Center, NA, New York, NY, [3]Columbia University, Department of Biological Science, New York, NY

Large cohorts of individuals are currently being assembled with the aim of identifying genetic variants that contribute to risk of common diseases. These cohorts often include individuals from a wide range of ages. We used this property to develop a method to identify genetic variants that influence age-specific mortality. In this method, we test whether the frequency of an allele varies across individuals who have survived to a given age while accounting for variation in their ancestry. We applied our method to the Genetic Epidemiology Research on Aging (GERA) cohort. We identified one allele that significantly varies in frequency across age cohorts, which tags the APOE4 allele (rs6857, $P < 10\text{-}14$). This allele decreases monotonically in frequency after the age of 70. We further tested whether sets of genetic variants that influence a trait show consistent changes in frequency across individuals of different ages. We show tentative evidence that alleles that delay age of menarche in women are beneficial for survival to old age. This approach opens the door to direct measurement of the effects of natural selection in contemporary populations.

← Back

STRIKING DIFFERENCES IN PATTERNS OF GERMLINE MUTATION IN MICE AND HUMANS.

Sarah J Lindsay, Raheleh Rahbari, Matt E Hurles

Wellcome Trust Sanger Institute, Genomic mutation and genetic disease, Hinxton, United Kingdom

We analysed genome-wide patterns of germline mutation in two large two-generation mouse pedigrees (129S/B6, B6/129S cross) with 57 and 77 offspring respectively, and compared our findings to multi-sibling human pedigrees. We performed whole genome sequencing on 10 offspring from each pedigree, 5 from the first and last matched litters. We then validated 439 de novo mutations (DNMs) and genotyped these mutations across all offspring from the two pedigrees, as well as across additional tissues in the parents and whole genome sequenced offspring.

While the mice had a similar mutation spectrum and paternal age effect to that observed in humans, the per generational mutation rate was much lower (~60% of the human rate). We observed 70 (16%) of the validated de novo mutations in more than 1 offspring, indicating parental germline mosaicism for these sites. 17 of these 70 sites were also detected in the parental soma; there was a striking gender bias in these pre-primordial germ cell mutations, with 16/17 observed in the paternal soma, suggesting possible sexual dimorphism in the cellular genealogy relating the soma and germline.

Using information on the sharing of DNMs between offspring we reconstructed partial genealogies of the parental gametes and showed that the majority of gametes are derived from 2-4 lineages, suggesting unequal contributions of the founder primordial germ cells to mature gametes. Finally, we observed that a high proportion (~20%) of de novo mutations observed in offspring were likely the result of mutations in the first post-zygotic cell division. This is a much higher proportion than observed in humans (~4%) and suggests that the first post-zygotic cell division in mice is much more mutagenic than in humans.

← Back

BREAKING THE INFINITE SITES MODEL: WIDESPREAD MUTATIONAL RECURRENCE IN EXOME SEQUENCE DATA FROM OVER 60,000 INDIVIDUALS

Konrad J Karczewski[1,2], Monkol Lek[1,2], Eric V Minikel[1,2], Kaitlin E Samocha[1,2], Exome Aggregation Consortium[1,2], Mark J Daly[1,2], Daniel G MacArthur[1,2]

[1]Massachusetts General Hospital, ATGU, Boston, MA, [2]Broad Institute, Medical and Population Genetics, Cambridge, MA

Many population genetics models assume that if a variant is observed twice, the two observations are a result of identity by descent. However, as the number of sequenced individuals grows, the probability of observing two or more independent mutational events occurring at the same site in the genome increases. Here, we describe an analysis of widespread mutational recurrence observed in exome sequence data from 60,706 individuals from the Exome Aggregation Consortium (ExAC). This effect is most pronounced among highly mutable CpG transitions, and in this dataset, **we observe over 60% of all possible** synonymous CpG mutations and begin to reach saturation levels.

We find that approximately one-third of high-confidence validated *de novo* variants identified in external datasets of parent-offspring trios are also observed independently in the ExAC dataset, indicating that the same variant has arisen multiple times independently within the history of the sequenced populations.

This process has a marked effect on the frequency spectrum in the ExAC data, resulting in a depletion of very low-frequency variants at sites with high mutation rates, even for synonymous sites (a class of variation expected to have undergone minimal selection). Specifically, we observe a correlation between singleton rates and site mutability inferred from sequence: sites with low predicted mutability (i.e. transversions) have a singleton rate of 60%, compared to 20% for sites with the highest predicted rate (i.e. CpG transitions). Additionally, there is a strong correlation between site mutability and the probability of observing the variant in two separate populations.

We demonstrate that these patterns are only observed at a sample size greater than approximately 20,000 individuals, indicating that ExAC is the first such dataset to observe an impact on the frequency spectrum from recurrent mutation. Finally, we propose a correction factor (from models learned from synonymous variants) to properly account for the impact of mutational recurrence on the frequency spectra of various functional classes, which enables us to provide robust estimates of their deleteriousness. We note that with a moderately larger sample size, we will be able to infer selection against individual CpG variants.

DETECTING 2,000 YEARS OF HUMAN GENETIC ADAPTATION

Yair Field*[1,2], Evan A Boyle*[2], Natalie Telis*[2], Ziyue Gao[2], Kyle J Gaulton[3], David Golan[2,4], Loic D Yengo[5], Mark McCarthy[6], Jonathan K Pritchard[1,2,7]

[1]Howard Hughes Medical Institute, Stanford University, Stanford, CA, [2]Stanford University, Genetics, Stanford, CA, [3]University of California San Diego, Pediatrics, San Diego, CA, [4]Stanford University, Statistics, Stanford, CA, [5]Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 8199, Lille, France, [6]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [7]Stanford University, Biology, Stanford, CA

The extent to which adaptive evolution has shaped the genetic and phenotypic variability in modern humans is a key open question in genetics. Methods to date, such as the integrated Haplotype Score (iHS), allow us to detect in contemporary human genome sequences signals of positive selection that occurred on the timescale of the past ~25,000 years. Here we present the Singleton Density Score (SDS), the first method to detect signals of selection with specificity to an order of magnitude more recent history, with the exact timescale adaptively determined by sample size. The idea behind our new approach is to infer the recent dynamics of common variants by modeling their local association with rare variants. We applied our method to data of ~3,000 British individuals from the UK10K project and estimate that SDS reflects allele frequency changes specific to the past 2,000-2,500 years. We find evidence for several cases of significant positive selection at this short and previously uncharacterized timescale. These include large increase in frequency of few alleles known to affect Mendelian traits that are prevalent in Britain, such as lactose persistence and blond hair. However, we also observe small but highly coordinated shifts in the frequencies of many alleles that are thought to confer genetic predisposition for complex traits known to be common in Britain, with exceptional significance for tall stature. This suggests that adaptive evolution in recent human history through polygenic adaptation is likely to be more rapid and more widespread than previously observed.

HIGH-THROUGHPUT, UNBIASED CRISPR MUTAGENESIS OF THE HUMAN NONCODING GENOME

Neville E Sanjana[1,2], Jason Wright[3], Feng Zhang[3,4]

[1]New York Genome Center, New York, NY, [2]New York University, Deptment of Biology, New York, NY, [3]Broad Institute of Harvard and MIT, Cambridge, MA, [4]MIT, Dept of Brain and Cognitive Sciences, Dept of Bioengineering, Cambridge, MA

The noncoding genome plays a major role in gene regulation and disease yet we lack tools for rapid identification and manipulation of noncoding elements. Large-scale CRISPR libraries have emerged as a powerful technique for high-throughput phenotypic screens, although they have primarily targeted protein-coding regions. Here, we employ a library of ~4,000 sgRNAs targeting 200 kb of noncoding sequence in an unbiased manner surrounding the E3-ubiquitin ligase CUL3. We previously demonstrated that loss-of-function mutations of CUL3 enhance resistance to a BRAF inhibitor in BRAF mutant melanoma cells. In this work, we identify specific locations across the CUL3 locus that modulate drug resistance when mutated. Interrogation of chromatin conformation indicates that many of the sites that confer resistance physically interact with the CUL3 promoter and are evolutionarily conserved. Mutations at specific noncoding elements lead to changes in transcription factor occupancy and the local epigenetic landscape and these changes are coincident with a loss of CUL3 expression. This demonstration of an unbiased screen of the noncoding regions flanking a disease-relevant gene expands the potential of pooled CRISPR screens for fundamental genomic discovery, gene regulation, and therapeutic development to overcome drug resistance.

← Back

MUTATION AND SELECTION DURING INDUCED PLURIPOTENT STEM CELL REPROGRAMMING

Petr J Danecek, Angela Goncalves, Richard Durbin, Daniel Gaffney

Wellcome Trust Sanger Institute, Computational Genomics, Cambridge, United Kingdom

Cellular reprogramming is an inefficient process involving a series of extreme population bottlenecks, as fully differentiated adult cells are transformed to a pluripotent state. If the starting population is genetically variable, genetic drift or selection, for example on variants that confer a growth advantage, could result in a cell line that is significantly genetically diverged from the germline of the original donor. Understanding the extent of such changes is critical for the applications of IPSC technology to regenerative medicine and for modelling human disease. We genotyped 522 IPSC lines derived as part of the Human Induced Pluripotent Stem Cells Initiative (HIPSCI) and compared them to the original dermal fibroblast population from which they were derived. Our results illustrate that, although large-scale rearrangements are uncommon overall (~10% of lines) during cellular reprogramming, multiple hotspots of duplication are found throughout the genome in human IPSCs. For a subset of these, we find evidence that certain duplications enhance IPS cell growth, proliferation and survival. To better understand the source of these genetic changes, we sequenced 158 hIPSC exomes derived from 119 donors. The pattern of point mutations we observed closely resembles that produced by UV radiation. This result strongly suggests that the majority of point mutations observed in human IPSCs already exist as somatic mutations in the source population of dermal fibroblasts and increase in frequency due to random genetic drift during reprograming. Although the majority of lines exhibit a relatively low mutational burden, we also observe extreme outlier lines that harbour tens of thousands of point mutations per genome

CRISPR DELETION SCREEN REVEALS WIDESPREAD FUNCTIONAL REDUNDANCY OF MAMMALIAN *IN VIVO* ENHANCERS

Marco Osterwalder[1], Diane E Dickel[1], Iros Barozzi[1], Virginie Tissieres[2], Yoko Fukuda-Yuzawa[1], Elizabeth Lee[1], Brandon J Mannion[1], Yiwen Zhu[1], Veena Afzal[1], Ingrid Plajzer-Frick[1], Catherine Pickle[1], Momoe Kato[1], Tyler Garvin[1], Jennifer A Akiyama[1], Javier Lopez-Rios[2], Edward M Rubin[1,3], Axel Visel[1,3,4], Len A Pennacchio[1,3]

[1]Lawrence Berkeley National Laboratory, Functional Genomics, Berkeley, CA, [2]Department of Biomedicine, Development and Evolution, Basel, Switzerland, [3]US Department of Energy Joint Genome Institute, Walnut Creek, CA, [4]University of California, School of Natural Sciences, Merced, CA

Distant-acting enhancers orchestrate the expression of genes in time and space. Our lab uses epigenomic mapping techniques combined with transgenic validation in mice and has identified thousands of validated enhancers (http://enhancer.lbl.gov) and hundreds of thousands of candidate enhancers in mammalian genomes. These sequences have highly cell- and tissue-specific activity and are thought to contribute to a variety of biological processes. In particular, mammalian embryonic development is characterized by a rich and complex regulatory architecture that may involve multiple enhancers acting on the same gene. In contrast to large-scale enhancer mapping efforts, expectations regarding the functional necessity of distant-acting enhancers for developmental processes have to date relied on a limited number of anecdotal examples, with often contradictory conclusions. To systematically explore the contribution of enhancers to mammalian development, we used *in vivo* CRISPR/Cas9-mediated genome editing in mice to delete a series of enhancers active during embryogenesis. We focused on the embryonic limb, which represents a well-studied paradigm for complex mammalian morphogenetic processes and is phenotypically sensitive to perturbations of the underlying gene regulatory networks. We knocked out a series of 10 enhancers that are active in the developing limb and are associated with genes required for normal limb morphology. Surprisingly, none of these enhancer deletions affected limb morphology, and the expression levels of the associated target genes remained largely unchanged, suggesting the presence of redundant (or shadow) enhancers. Indeed, a systematic genome-wide analysis of ChIP-seq and RNA-seq data from limb tissue revealed that genes critical for limb development are typically flanked by large regulatory regions harboring multiple putative limb enhancers. To examine whether these enhancer sequences with similar activity patterns are possibly functionally redundant, we used iterative CRISPR/Cas9 engineering to generate mice in which presumptive pairs of shadow enhancers were deleted. Loss of redundant enhancers resulted in limb abnormalities that were consistent with the phenotypes resulting from loss of the neighboring limb-expressed gene. Taken together, our genome-wide analysis of the regulatory architecture suggests the pervasive presence of functional redundancy among enhancers involved in mammalian development, and our knockout studies of selected pairs of enhancers indicate that this redundancy likely provides a mechanism to confer functional robustness in complex developmental processes.

← Back

MICROGLIA DEVELOPMENT FOLLOWS A STEPWISE PROGRAM TO REGULATE BRAIN HOMEOSTASIS

Deborah R Winter[1], Orit Matcovitch-Natan[1,2], Amir Giladi[1], Stephanie Vargas Aguilar[3], Amit Spinrad[1,2], Sandrine Sarrazin[3,4,5], Eyal David[1], Meital Gury[1], Hadas Keren-Shaul[1], Christoph Thaiss[1], Keren Bahar Halpern[6], Kuti Baruch[2], Aleksandra Deczkowska[2], Shalev Itzkovitz[6], Eran Elinav[1], Michael Sieweke[3], Michal Schwartz[2], Ido Amit[1]

[1]Weizmann Institute of Science, Department of Immunology, Rehovot, Israel, [2]Weizmann Institute of Science, Department of Neurobiology, Rehovot, Israel, [3]Université Aix-Marseille, Centre d'Immunologie de Marseille-Luminy, Marseille, France, [4]Institut National de la Santé et de la Recherche Médicale, INSERM, Marseille, France, [5]Centre National de la Recherche Scientifique, CNRS, Marseille, France, [6]Weizmann Institute of Science, Department of Cell Biology, Rehovot, Israel

Microglia play important roles in life-long brain maintenance at steady state and in pathology, but are also crucial during development of the central nervous system by promoting neurogenesis and synaptogenesis. Microglia originate in the yolk sac from erythro-myeloid progenitors that migrate to the brain starting at embryonic day 8.5 and continuing until the blood brain barrier is formed; yet their regulatory dynamics during development have not been fully elucidated.

Here, we systematically study the transcriptional and epigenomic regulation of microglia throughout brain development. Genome-wide expression and chromatin profiles indicate that microglia proceed through 3 distinct developmental stages: early (embryo days 10-12), pre- (embryonic day 13-postnatal day 9), and adult microglia. Single cell transcriptome analysis revealed minor mixing of the regulatory programs across phases and the relevant markers to distinguish them. Knockout of the transcription factor MafB caused disruption of homeostasis in adulthood and activation in inflammatory pathways. Environmental perturbations, such as in germ-free mice or prenatal immune activation, also led to dysregulation of the developmental program, particularly in terms of inflammation. Together, our work identifies a stepwise developmental program of microglia integrating immune response pathways that may be associated with several neurodevelopmental disorders.

← Back

PREDICTING THE REGULATORY IMPACT OF RARE NON-CODING VARIATION

Yungil Kim[1], Xin Li[2], Farhan Damani[1], Joe Davis[2], Emily Tsang[2], Colby Chiang[3], Zachary Zappala[2], The GTEx consortium .[1], Ira Hall[3], Stephen B Montgomery[2], Alexis Battle[1]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Stanford University, Genetics and Pathology, Palo Alto, CA, [3]Washington University, McDonnell Genome Institute, St. Lous, MO

The enormous increase in availability of full human genomic sequences presents great opportunity for understanding the impact of rare genetic variants. Based on current knowledge, however, we are still limited in our ability to interpret or predict consequences of rare variants in non-coding regions of the genome. Association methods are inherently limited for variants observed in only one or a few individuals. Diverse genomic annotations such as growing resources of epigenetic data and have been shown to be informative regarding regulatory elements, but are only moderately predictive of impact for individual variants. The availability of personal RNA-seq and other cellular measurements for the same individuals with genome sequencing offers a new avenue for integrated methods for prioritizing rare functional variants. By considering informative genomic annotations along with molecular phenotyping, we are able to identify the regulatory impact of rare genetic variants largely excluded from previous analyses. We have evaluated the impact of rare regulatory variants using whole genome sequences and corresponding RNA-sequence in 54 different tissues samples from the GTEx project and report the most likely functional rare regulatory variants for each individual, demonstrating that rare variants with specific genomic annotations including enhancer and promoter elements, conserved regions, and others are associated with extreme changes in gene expression across multiple tissues. Additionally, we have developed a Bayesian machine learning approach that integrates whole genome sequencing with RNA-seq data from the same individual, leveraging gene expression levels along with diverse genomic annotations and performing joint inference to identify likely functional rare regulatory variants. We have applied this model to the GTEx data to prioritize the most likely functional rare regulatory variants for each individual. We demonstrate that integrative models perform better than predictions from DNA-sequencing or RNA-sequencing alone. Our probabilistic model of rare regulatory genetic variants offers great potential for identifying potentially deleterious non-coding genetic variants from individual genomes.

← Back

RNA SPLICING IS A PRIMARY LINK BETWEEN GENETIC VARIATION AND DISEASE.

Yang I Li[1], Bryce van de Geijn[2], Anil Raj[1], David A Knowles[3,4], Allegra A Petti[5], David Golan[2], Yoav Gilad[2], Jonathan K Pritchard[6,7]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]University of Chicago, Department of Human Genetics, Chicago, IL, [3]Stanford University, Department of Computer Science, Stanford, CA, [4]Stanford University, Department of Radiology, Stanford, CA, [5]Washington University in St. Louis, Genome Institute, St. Louis, MO, [6]Stanford University, Department of Biology, Stanford, CA, [7]Stanford University, Howard Hughes Medical Institute, Stanford, CA

Noncoding genetic variants play a central role in both the etiology of diseases and the evolution of novel traits in humans, yet we still lack understanding of the mechanisms by which most variants act. I will present work using eight molecular datasets (including data from H3K27ac ChIP-seq (n=59), DNA methylation 450K arrays (n=64), DNase-seq (n=67), 4sU-seq (n=65), RNA-seq (n=86), RNA decay (n=70), ribo-seq (n=70), and mass spectroscopy (n=62)) from a population sample of Yoruba lymphoblastoid cell lines to perform a comprehensive evaluation of inter-individual variation in gene regulation, from chromatin to proteins. Our analyses indicate that for over 73% of the cases, variation in gene transcription rates results in concordant changes in protein expression levels. We estimate that ~65% of eQTLs affect aspects of chromatin regulation, while the remaining 35% of eQTLs, not associated with chromatin-level variation, are enriched in transcribed regions and regions associated with transcriptional elongation. Using a novel method to detect variation in intronic splicing, we also identify 2,893 genetic loci that affect pre-mRNA splicing (sQTLs), at 10% false discovery rate, most of which had little or no detectable effect on overall gene expression. This indicates that pre-mRNA splicing is a primary target of common genetic variation. Finally, using statistical models to quantify the enrichment of QTLs among signals from GWAS, we found that variants affecting pre-mRNA splicing are major contributors to complex traits, roughly on a par with variants that affect gene expression levels. We conclude that pre-RNA splicing is an important regulatory mechanism by which common genetic variation modulates complex disease risk.

← Back

DIRECT IDENTIFICATION OF HUNDREDS OF EXPRESSION-MODULATING VARIANTS USING A MULTIPLEXED REPORTER ASSAY.

Ryan Tewhey[1,2], Dylan Kotliar[1,2], Daniel S Park[2], Tarjei S Mikkelsen[2], Steve F Schaffner[1,2], Pardis C Sabeti[1,2]

[1]Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA, [2]Broad Institute, Cambridge, MA

Although genetic studies have identified hundreds of loci associated with human traits and diseases, pinpointing causal alleles remains difficult, particularly for non-coding variants. To address this challenge, we adapted the massively parallel reporter assay (MPRA) to identify variants that directly modulate gene expression. We applied it to 32,373 variants from 3,642 cis-expression quantitative trait loci and control regions. Variants identified by MPRA show a strong correlation between existing measures of regulatory function, and demonstrate MPRA's capabilities for pinpointing causal alleles. In total, we identify 842 variants showing differential expression between alleles, including 53 well-annotated variants associated with diseases and traits. We investigate one in detail, a risk allele for ankylosing spondylitis, and provide direct evidence of a non-coding variant that alters expression of the prostaglandin EP_4 receptor. Thus, we have created a resource of concrete leads and illustrate the promise of this approach for comprehensively interrogating how non-coding polymorphism shapes human biology.

SOMATIC MOSAIC VARIATIONS IN HEALTHY SKIN FIBROBLASTS

Alexej Abyzov[1], Livia Tomasini[2], Bo Zhou[3], Nikolaos Vasmatzis[1], Jessica Mariani[2], Mariangela Amenduni[2], Anahita Amiri[2], Alexander E Urban[3], Flora M Vaccarino[2]

[1]Mayo Clinic, Health Sciences Research, Rochester, MN, [2]Yale University, Child Study Center, New Haven, CT, [3]Stanford University, Departments of Psychiatry and Genetics, Palo Alto, CA

Multiple studies have been performed on the analysis of somatic genomic alterations in cancer, but only a few have been conducted to understand natural somatic mosaicism, that is post-zygotic accumulation of mutations in cells of the human body. Fundamental knowledge about somatic mosaicism is not only crucial for finding determinants of cancer development and progression, but also for an understanding of various diseases and aging. We have compared genomes of 32 clonally derived human induced pluripotent stem cell (hiPSC) lines to the genomes of 11 (5 children and 6 adults) primary skin fibroblast samples, parental to the hiPSC lines. The clonal nature of hiPSC lines allows the discovery of somatic genomic variants present in the founder cell, but not in all fibroblast cells, thereby providing a mean for a high-resolution (and not compromised by amplification) analysis of single cell genomes. Adjusted for discovery sensitivity we estimated that on average, an iPSC line/a single fibroblast cell in children has 1,035 single nucleotide variants (SNVs) of which 181 could be directly confirmed as somatic by an in-depth re-analysis of fibroblast bulk genomic data with ultra-deep sequencing and digital droplet PCR, down to an allele frequency of 0.02%. Progressive increase in re-analysis' sensitivity confirmed additional SNVs as somatic, suggesting that the estimated numbers are true counts of somatic SNVs per fibroblast cell. Similar analyses in adults revealed on average ~30% increase in the count of SNVs per cell (counts ranged from 900 to 2,000) as compared to children, suggesting that a large fraction of somatic SNVs in human fibroblasts occurs during prenatal and early childhood development. Except for a few SNVs occurring at two consecutive genomic positions (likely results of exposure to UV light) SNVs were distributed randomly across the genome and their mutation spectrum was an almost perfect match to a previously uncharacterized mutation signature observed in cancers (Alexandrov et al., Nature, 2013). We, thus, propose that this cancer signature reflects normal development. Finally, in four children, allele frequency distribution for somatic SNVs had distinct narrow peaks, which, we hypothesize, are either the results of cell clonal selection or bursts of mutations during development. These new discoveries reveal a large degree of somatic mosaicism existing in healthy human tissues, link the mosaicism with development, and explain a mutational signature observed in cancers.

← Back

WHOLE GENOME SEQUENCING AND ANALYSIS OF AFLATOXIN-PRODUCING AND ATOXIGENIC *ASPERGILLUS FLAVUS* GENOTYPES.

Bishwo N Adhikari, Peter J Cotty

USDA-ARS/University of Arizona, School of Plant Sciences, Tucson, AZ

Aspergillus flavus, the fungal species most frequently implicated in episodes of crop aflatoxin contamination, varies widely in ability to produce aflatoxins. Genotypes of *A. flavus* range from production of hundreds of mg/kg to a few µg/kg. Aflatoxins are fungal metabolites that suppress the immune system, interfere with development, and cause cancer. Some *A. flavus* genotypes produce no aflatoxins, and some of these atoxigenic *A. flavus* are used as active ingredients in biopesticides used to prevent aflatoxin contamination. To better understand differences among genotypes and identify genomic bases of aflatoxin production, five *A. flavus* genotypes, with and without the ability to produce aflatoxins, were sequenced. Comparative genomic analysis among genotypes revealed genomic rearrangements and wide variation in secondary metabolite genes. Genotype-specific genes and indels may allow insights on adaptation to different environments. Subtelomeric regions exhibited high concentrations of secondary metabolite genes, densities of polymorphisms greater than the genome average, low gene density, and high frequencies of rearrangements. Features that distinguish *A. flavus* genotypes from each other will be presented and implications of genomic variability for selection of biological control agents will be discussed.

UNCOVERING THE REGULATORY LANDSCAPE OF DENDRITIC CELLS RESPONSE TO PATHOGENS

Shaked Afik[1], David S Fischer[2,3], Barbara Tabak[4], Elisa Donnard[4], Sowmya Iyer[4], Pranitha Vangala[4], Xiaopeng Zhu[4], Patrick McDonel[4], Jeremy Luban[5], Manuel Garber[4,5], Nir Yosef[2]

[1]Computational Biology Graduate Group, University of California Berkeley, Berkeley, CA, [2]Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA, [3]Department of Computer Science, ETH Zurich, Zurich, Switzerland, [4]Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, [5]Program in Molecular Medicine, University of Massachusetts Medical School , Worcester, MA

Developmental shift in cells following environmental stimuli is controlled by changes in expression of thousands of genes. Those changes are mediated by a complex regulatory network that is comprised of non-coding DNA sequences, chromatin structure and transcription factors (TFs). However, the code linking these variables in a way that temporal changes in gene expression can be predicted has yet to be deciphered. We aim to model such code based on genome-wide analysis of human dendritic cells (DCs), antigen presenting cells that help initiate the immune response, as they mature in response to lipopolysaccharide (LPS), a component of gram-negative bacteria.

We characterized temporal changes in the genomic landscape by analyzing time course ATAC-seq data. We were able to identify thousands of non-coding genomic regions that exhibit significant changes in their accessibility, revealing various temporal patterns of changes to the regulatory landscape. By using local features of those regulatory regions such as DNA composition and chromatin marks that will be collected using ChIP-seq, we will take a supervised learning approach to build a classifier in order to predict temporal binding of TFs to each region. We aim to use our prediction of TF binding along with measurements of the regulatory landscape to predict changes in gene expression. The advantage of combining prediction of TF binding with our model of gene expression changes will be the ability to distinguish functional from non-functional TF binding, as we can determine which TF binding events result in changes to gene expression levels.

The rapid response and large expression changes makes DCs activation an ideal system to understand general mechanisms of gene regulation and gain a better grasp of the human immune system. This work will define a computational platform to integrate different types of genomic data which, combined with the experimental platform, could be applied to study many other systems.

FUNCTIONAL VALIDATION OF HUMAN PROTEIN-TRUNCATING GENETIC VARIANTS

Irina M Armean[1,2], Konrad J Karczewski[1,2], Jamie L Marshall[1,2], Beryl B Cymmings[1,2], Eric Minikel[1,2], Daniel Birnbaum[1,2], Ben Weisburd[1,2], Preeti Singh[1,2], Monkol Lek[1,2], Mark Daly[1,2], Aarno Palotie[1,2], Sekar Kathiresan[2,3], Daniel G MacArthur[1,2]

[1]Massachusetts General Hospital, The Analytic and Translational Genetics Unit, Boston, MA, [2]Broad Institute of Harvard and MIT, Medical and Population Genetics, Cambridge, MA, [3]Massachusetts General Hospital, Center for Human Genetic Research and Cardiovascular Research Center, Boston, MA

Knockout model organisms have been used for decades to study the function of genes. The identification of so-called “human knockouts”, individuals with both copies of a gene inactivated by null mutations, in different populations across the globe offers a unique opportunity for investigating gene function. Homozygous protein-truncating variants (PTV) are typically rare but are more often observed in populations with historical bottlenecks or high levels of consanguineous mating, which skew the frequency spectrum and induce homozygosity, respectively. In addition to providing information about the function of human genes, human knockouts can provide valuable information about the potential efficacy and toxicity of therapeutic inhibition of specific biological targets.

Here we describe a systematic pipeline for the functional validation of predicted protein-truncating genetic variants using both public databases and newly generated data. We first identify human knockouts from samples including over 60,000 exomes assembled by the Exome Aggregation Consortium (ExAC). Secondly, we describe approaches to filtering candidate PTVs using both an in silico pipeline, LOFTEE, and publicly available RNA-seq data from over 1,000 genome-sequenced individuals. Finally, we propose an approach to perform direct functional validation of candidate gene-disrupting PTVs using participant cell lines or CRISPR-engineered equivalents differentiated into relevant tissues.

We also describe the aggregation of these variants into a database of LoF variants, dbLoF, providing a resource for pharmaceutical development, transplant biology and understanding of rare Mendelian diseases.

← Back

INFERENCE OF LOCAL ANCESTRY BASED ON ADMIXTURE GRAPHS

Georgios Athanasiadis, Mikkel H Schierup, Thomas Mailund

Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

Chromosome painting consists in assigning the most likely ancestry to each locus on a given admixed chromosome. The different ancestry labels are typically chosen from a series of extant donor populations considered to be closely related to the unobserved ancestral populations. In this method, we describe the construction of a new hidden Markov model for painting admixed chromosomes conditional on (i) a set of samples from various donor populations and (ii) the admixture graph relating donor populations with the admixed population. The method is based on a copying model and each donor chromosome is treated as a different state. Apart from a set of donor and recipient chromosomes, the method is parameterized with: (i) admixture proportions for each donor population into an ancestral admixed population, (ii) a genetic map specifying the recombination rates between adjacent loci, (iii) the mean time of coalescence within each ancestral population, and (iv) the times of admixture for each ancestral admixed population. The crux of the hidden Markov model is specifying the transition probability matrix of the model. The transition probabilities are contingent on the rate of recombination and its location on the admixture graph and encompass three cases: (i) staying on the same chromosome, (ii) changing chromosome for one from the same population, and (iii) changing population. As a first implementation, we defined different admixture graphs and used the model to simulate paths of local ancestry. We ran follow-up analyses of migrant tract distribution to evaluate the efficiency of our method.

ASSESSMENT OF FUNCTIONAL CONVERGENCE ACROSS STUDY DESIGNS IN AUTISM

Sara Ballouz, Jesse Gillis

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY

Disagreements over genetic signatures associated with disease have been particularly prominent in the field of psychiatric genetics, creating a sharp divide between disease burdens attributed to common and rare variation, with study designs independently targeting each. Meta-analysis within each of these study designs is routine, whether using raw data or summary statistics, but no method for combining the "functional" results across study designs exist.

In this work, we develop a general solution which integrates the disparate genetic contributions constrained by their observed effect sizes to determine functional convergence in the underlying architecture of complex diseases, which we illustrate on autism spectrum disorder (ASD) data. Our approach looks not only for similarities in the functional conclusions drawn from each study type individually but also those which are consistent with the known effect sizes across these studies. We name this the "functional effect size trend" and it can be understood as a generalization of a classic meta-analytic method, the funnel plot test. We took candidate disease gene data from multiple ASD studies across thousands of individuals and study designs, including whole-exome sequencing and genome-wide association studies. We split the candidate genes by variant class (common and rare) and effect size (low to high) into 14 gene sets, controlling for set size.

We detected remarkably significant trends in aggregate ($p \sim 1.92e-31$) with 20 individually significant properties (FDR<0.01), many in areas researchers have targeted based on different reasoning, such as the fragile X mental retardation protein (FMRP) interactor enrichment (FDR~0.006). We are also able to detect novel technical effects and we see that network enrichment from protein-protein interaction data is heavily confounded with study design, arising readily in control data. Our meta-analytic approach, explicitly accounting for different study designs, can be adapted to other diseases to discover novel functional associations and increase statistical power.

NEW UCSC GENOME BROWSER VIEWS: EXON-ONLY, GENE-ONLY, ALTERNATE HAPLOTYPES, AND CUSTOM REGIONS

Galt P Barber, Angie S Hinrichs, Kate R Rosenbloom, Matthew L Speir, Christopher M Lee, Ann S Zweig, Donna Karolchik, Jim Kent

University of California Santa Cruz, Genomics Institute, Santa Cruz, CA

The University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu)
is a public, freely available web-based graphical viewer for the display of genomic sequences
and their annotations with links to external public databases.

Browser users have long asked for the ability to remove intronic and intergenic regions from the display, leaving only the exons on view. These intronic and intergenic regions are often extremely large and can occupy much of the visual field in the browser display. Removing them allows the user to focus on the annotations in exonic regions in greater detail.

The Genome Browser now offers a multi-region configuration option that supports an “exon-only” display mode, and more. The exonic regions are defined by a preset gene track optimized for the selected genome assembly. Users can define the number of bases used for padding between exons or genes. The view responds dynamically to gene track settings, such as options to include non-coding or alternatively spliced transcripts. The exon-only display is useful for looking at protein-coding regions and variants. It also works well when viewing the Browser’s new data tracks from the Genotype-Tissue Expression (GTEx) project, a resource to study human tissue-specific gene expression and regulation and its relationship to genetic variation.

In addition to exon-only mode, the Genome Browser supports three other display modes that allow users to view multiple regions alongside one another in the same browser window. In the gene-only view, the intergenic spaces are removed from the display, but not the introns. An alternate-haplotype view (on newer human assemblies) shows a user-selected alternate haplotype placed in context on its reference chromosome. The custom-regions view allows the user to create a customized view of regions specified in a BED file URL. The user can then scroll around the large virtual chromosome made from the custom regions list.

The multi-region view mode and documentation can be accessed via the "multi-region" button on the Genome Browser track display.

ADAPTIVE EPISTASIS: NUCLEAR-MITOCHONDRIAL INTERACTIONS SELECT FOR DIFFERENT GENOTYPES

Tara Z Baris[1], Dominique N Wagner[1], David I Dayan[1], Xiao Du[1], Pierre U Blier[2], Nicolas Pichaud[2], Marjorie F Oleksiak[1], Douglas L Crawford[1]

[1]University of Miami/RSMAS, Marine Biology and Ecology, Miami, FL, [2]Université du Québec à Rimouski, Dept. de Biologie, Rimouski, Canada

We are investigating the impact of nucleotide divergence on oxidative phosphorylation (OxPhos) metabolism among populations of *Fundulus heteroclitus*. The OxPhos pathway is responsible for most aerobic ATP production and is the only pathway with both nuclear and mitochondrial encoded proteins. *F. heteroclitus* populations are distributed along a steep thermal cline on the east coast of the United States and have evolved by natural selection to adapt to this clinal variation in temperature. These distinct populations have sequence divergence in OxPhos genes in both mitochondrial and nuclear genomes. Two distinct mitochondrial haplotypes exist along this thermal cline, a northern and southern haplotype, with a break occurring at the Hudson River. In northern New Jersey, there is an admixture of mitochondrial haplotypes with a frequency of about 60% southern and 40% northern. In this admixed population, we examined whether mito-nuclear interactions alter allele frequencies for ~11,000 nuclear SNPs in 155 individuals. Between the two mt-haplotypes, there are significant differences in nuclear allele frequencies for 349 SNPs. These SNPs occur in genes involved in regulating metabolic processes but are not directly associated with the 79 nuclear OxPhos proteins. Therefore, we postulate that epistatic selection affects OxPhos function and is acting upstream of OxPhos. RNA sequencing will be performed on these same individuals to understand the role of gene expression on mito-nuclear genome interactions. Additionally, OxPhos function measured in cardiac tissues of these 155 individuals revealed significant differences between the two mitochondrial haplotypes. These differences are most apparent when individuals are acclimated to high temperatures with the southern mitochondrial genotype having a large acute response and the northern mitochondrial genotype having little, if any acute response. These data demonstrate a complex gene by environmental interaction affecting the OxPhos pathway.

← Back

ROAD MAP OF THE GENETIC AND EVOLUTIONARY FORCES DRIVING POPULATION DIFFERENCES IN IMMUNE RESPONSES TO INFECTION

Joaquin Sanz Remón[1,3], Yohann Nédélec[1,3], Golshid Baharian[1,3], Anne Dumaine[1], Alain Pacis[1,3], Ariane Pagé Sabourin[1], Jean-Christophe Grenier[1], Jamel Belaid Boukra[4], Vania Yotova[1], Luis B Barreiro[1,2]

[1]CHU Sainte-Justine Research Center, Department of Genetics, Montreal, Canada, [2]University of Montreal, Department of Paediatrics, Montreal, Canada, [3]University of Montreal, Department of Biochemistry, Montreal, Canada, [4]Hospital Tracadie, Department of Medicine , New Brunswick, Canada

Individuals from African and European ancestry considerably vary in their susceptibility to infectious diseases or diseases characterized by pathological inflammation such as chronic inflammatory and many autoimmune disorders. Such differences suggest inter-population variation in the immune response, possibly caused by the adaptation of Africans and Europeans to different pathogenic pressures through our their evolutionary history. However, due to the lack of comparative functional data across populations, it remains unclear the extent to which local adaptations contributed to the to the diversification of immune responses between populations, and what phenotypes have been differently selected for. Here, we infected macrophages from a panel of 68 African Americans and 103 European Americans with either *Listeria monocytogenes* (Gram-positive bacterium) or *Salmonella typhimurium* (Gram-negative bacterium). Following infection, we collected RNA-seq data from matched non-infected and infected samples, to an average of ~36 million reads sequenced per sample and a total of 546 RNA-seq profiles. By leveraging on the power of high-quality RNA-sequencing we show that 32% of genes expressed in macrophages show at least one of transcriptional difference between populations, whether in the form of differences in gene expression (26%), transcriptional response to infection (12%) or differences in isoform usage (1.3%). Our results indicate that African descent individuals elicit a stronger inflammatory response in response to infection, which is associated with an increased ability of their macrophages to control bacterial growth post-infection. Combining the transcriptional data with dense genotypic information we show that *cis* genetic variation explains at least ~25% of the transcriptional differences identified. Finally, we show that natural selection, including adaptive introgression with Neanderthal, significantly contributed to the diversification of the immune system between populations. Collectively, our data suggest that regulatory differences between African and European individuals impact on the ability of macrophages to control bacterial infections and are likely to contribute to some of the known ethnic disparities in susceptibility to inflammatory and autoimmune diseases.

← Back

FUNCTIONAL PRIORITIZATION OF STRUCTURAL VARIANTS THROUGH A COMBINATORIAL APPROACH FOR IDENTIFYING LOCI UNDER PURIFYING SELECTION

Justin R Bartanus, Fuli Yu

Baylor College of Medicine, Molecular and Human Genetics, Houston, TX

As the cost of sequencing has fallen dramatically over the past decade, it has become more feasible to sequence large cohorts of individuals from a diverse range of populations. While many studies have focused on analyzing human variation, the vast majority of the focus has been directed at single-nucleotide variants (SNVs) and short indels even though copy number variants (CNVs) account for variants covering a larger proportion of the human genome. Due to this fact, our understanding of CNVs and their relative contributions to both common and rare disease is incomplete. Though more recent tests for directional selection have begun incorporating functional properties to traditional statistical analyses of regional variations, the overall picture of the contribution of these regional variants to disease susceptibility remains unclear. It is well known that mutations in functionally important regions undergo directional selection. Through the application of population genetic principles on large cohorts such as the Exome Aggregation Consortium dataset, we can study selective signals in genomic regions with known pathogenic structural variants. We propose a composite test which combines the strengths of various orthogonal purifying signal estimators, including both gene-level and variant-level statistical tests, functional tests, and evolutionary tests, to amplify the overall selective signal of a target region. Through comparisons of target regions and their corresponding selective signals, we can functionally prioritize regions according to the relative susceptibilities to functional disruption. We can further associate regional susceptibilities to overlapping structural variations to effectively prioritize individual structural variants by their contributions to both rare and common disease.

PROTEIN RECODING BY RNA EDITING IN BACTERIA

Dan Bar-Yaacov, Ernest Mordret, Orna Dahan, Schraga Schwartz, Yitzhak Pilpel

Weizmann Institute of Science, Department of Molecular Genetics, Rehovot, Israel

Could the DNA (always) be trusted in bacteria?
Recoding of protein sequence by adenosine (A) to inosine (I) RNA editing was reported to occur in eukaryotes, but never in bacteria. Here, we report for the first time on protein recoding by RNA-editing in Escherichia coli. Analysis of multiple RNA/DNA-Seq data sets detected recoding of the hokB toxin (Y29C) that was shown to contribute to antibiotic tolerance. Remarkably, the editing recapitulates a DNA-hardcoded cysteine at the homologous position of all other members of the hok family. Predicted 3D structure suggests that this edited cysteine is likely to interact with a cysteine at position 46 through a disulfide bridge. Furthermore, the hokB edited site is embedded within a tRNA adenosine deaminase A (tadA) recognition motif, thus implicating this enzyme in the editing process. Consistent with our findings, a mutation in tadA was reported to cause resistance to the hok toxins. Our work suggests a functional link between RNA editing, formation of S-S bonds and antibiotic tolerance in bacteria. This link is currently being examined in the context of genetically engineered strains of hokB and tadA.

UNVEILING SUBPOPULATION STRUCTURES IN LARGE-SCALE SINGLE-CELL RNA-SEQ EXPERIMENTS WITH A NOVEL SIMILARITY-LEARNING FRAMEWORK

Bo Wang[1], Junjie Zhu[2], Emma Pierson[1], Grace X Zheng[3], Jessica Terry[3], Tarjei Mikkelsen[3], Serafim Batzoglou[1]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Electrical Engineering, Stanford, CA, [3]10x Genomics, Pleasanton, CA

Single-cell RNA-seq (scRNA-seq) technologies enable gene expression measurement of individual cells and allow the discovery of cell population heterogeneity. Recent advances in scRNA-seq have allowed simultaneous profiling of thousands to tens of thousands of cells and increased the sensitivity in quantifying single-cell transcriptome landscapes of complex biological systems. However, scRNA-seq data sets are noisy with high degree of dropouts, and exhibit higher levels of diversity such as cell input heterogeneity and variation in cell cycle stages,. These challenges make it difficult to define cell-to-cell similarity measures based on strict statistical assumptions that have been developed for bulk RNA-seq. Here, we propose a novel similarity-learning framework, SIMLR (single-cell interpretation via multi-kernel learning), which learns an appropriate distance metric from the data for dimension reduction, clustering and visualization.

We profiled >50,000 peripheral blood mononuclear cells (PBMCs) with the ChromiumTM system from 10x Genomics, and used SIMLR to provide an unbiased classification of all major subpopulations at expected proportions. In order to evaluate the sensitivity and accuracy of SIMLR, we further analyzed individual purified populations from PBMCs and pooled the data in silico at varying proportions. In addition, we extensively evaluated SIMLR using published scRNA-seq datasets generated by several other micromanipulation and microfluidics platforms with varying sequencing depths. For all the datasets above, we first show that simple correlation-based or distance-based similarity measures are sensitive to noise, dropouts and outlier effects among the high dimensional data. In contrast, SIMLR can uncover clear similarity block structures by automatically learning appropriate cell-to-cell similarity specific to each dataset. Furthermore, we performed dimension reduction on these high dimensional datasets using the similarity learned by SIMLR, and compared it with 8 other popular dimension reduction methods, including linear and nonlinear methods (such as PCA and tSNE), and a recently published approach (ZIFA) which is specifically designed for single-cell data sets. We evaluated the effectiveness of SIMLR's dimension reduction both quantitatively and qualitatively by considering clustering accuracy and visualization. Dimension reduction performed via SIMLR yields a substantially higher clustering accuracy. When applied to visualization, we illustrate SIMLR's advantage over other methods in projecting the high dimensional data to 2-D and 3-D where the different cell types are automatically projected in spatially distinct clusters.

A MULTI-SCALE, PROBABILITY-BASED APPROACH TO SOLVING POORLY ASSEMBLED GENOMES USING CHROMOSOME CONTACT DATA

Lyam Baudry, Martial Marbouty, Hervé Marie-Nelly, Romain Koszul

Institut Pasteur, Genomes and Genetics, Paris, France

Despite improvement in technologies and computational analysis, the costs of polishing genome assembly drafts have left most of them in an 'unfinished' state, typically displaying numerous gaps, limited-size scaffolds and the absence of well-defined chromosomes. This is especially prevalent among animals with large genomes carrying an important number of repeated sequences, making the assembling step difficult when resorting to conventional approaches.

Here we build upon published re-assembly software – dubbed GRAAL (Genome Re-Assembly Assessing Likelihood from 3D) - using tridimensional contact data from chromosome capture conformation related (3C, Hi-C, etc.) experiments. Genomic structure is thus captured by cross-linking, leading to a DNA-protein complex bound by formaldehyde. DNA is then digested by a restriction enzyme and ligated again such that restriction fragments closest to each other in space are next to each other when sequenced in paired-end. These contact counts are expected to respect a given distribution given by polymer physics at both local and global genomic scales, which is what GRAAL relies on to assess and maximize the likelihood of a genome given such contact data, and rearrange appropriate parts as needed. Using a Markov chain Monte Carlo (MCMC) method, the genome eventually converges toward one best satisfying its underlying 3D contacts.

Due to its multi-scale nature, GRAAL operates at the level of topological domains, chromosomes and species at the same time. This makes it suitable for multiple purposes at once, such as improving an existing assembly, segregating a poorly-scaffolded genome into proper chromosomes or even unveiling new genomes from a complex network of unknown species such as what may be encountered in metagenomics. Here we will present examples of large scaffolding of genomes from animals at positions of interests in the tree of life.

NATURAL SELECTION IN FUNCTIONAL PATHWAYS: AN APPROACH TO EVOLUTIONARY SYSTEMS BIOLOGY.

Jaume Bertranpetit, Begona Dobon, Mayukh Mondal, Marc Pybus, Ludovica Montanucci, Pierre Luisi, Hafid Laayouni

Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Spain

Evolutionary analysis at the molecular level provides new tools to biology by considering the action of natural selection in genes and groups of genes on their functional setting of molecular pathways of their gene products. By comparing genomic data of different species or of different populations within a single species, we can distinguish between selection at large or short scales, allowing detection (and sometimes measurement) of natural selection in the form of positive (adaptive) selection and purifying (negative) selection.

Gene products function in molecular networks, as such, the position within the network may determine the strength of selection applied to the gene. It is possible to interrogate how selection is distributed across the molecular networks. This analysis may be applied to the pathway level (with a low number of interacting units but a very detailed molecular knowledge, like the examples of glycosylation, the insulin/TOR signal transduction pathway or phototransduction), the entire metabolome, or even the whole interactome. We present here all these cases in order to relate selection to the specificity of reactions and their function.

At the level of the human interactome, genes with higher number of interactions are more likely to have been targeted by recent positive selection during recent human evolution. Our results indicate that the relationship between centrality and the impact of adaptive evolution highly depends on the evolutionary time-scale. Most likely, network adaptation occurs through intraspecific adaptive leaps affecting key network genes, followed by the fine tuning of adaptations in less important network regions. These results may remodel the shape of the traditional fitness landscapes.

Beyond topology, dynamics of a complex molecular system may have an independent influence on the impact of selection. A case study in the core metabolic network of a human erythrocytes indicated genes encoding enzymes that carry high fluxes have been more constrained in their evolution. By the other hand we have applied a comprehensive mathematical model of mammalian phototransduction to predict the degree of influence that each protein in the system exerts on the high-level dynamic behavior.

This analysis goes beyond the identification of single cases of adaptation and opens the scope of understanding how natural selection works within the biomolecular complexity of life.

THE GENOMIC AND EPIGENOMIC PROPERTIES OF SEXUAL DIMORPHISM IN HUMAN MEIOTIC RECOMBINATION

Claude Bhérer[1,2], Christopher L Campbell[1], Adam Auton[1,3]

[1]Albert Einstein College of Medicine, Department of Genetics, Bronx, NY, [2]New York Genome Center, New York, NY, [3]23andMe Inc., Mountain View, CA

Sexual dimorphism in meiotic recombination is a widespread phenomenon across many species. In humans, males tend to have considerably lower recombination rates than females over the majority of the genome, but the opposite is usually true close to the telomeres. These broad-scale differences have been known for decades, yet little is known about the fine-scale differences between the sexes. Using recombination events inferred from pedigree datasets and representing a total of over 100,000 meioses, we have constructed genome-wide sex-specific genetic maps at a previously unachievable resolution. Compared to previous maps, our refined maps display markedly improved correlation with LD-based maps and localization of events to hotspots of recombination. We show that although a substantial fraction of the human genome shows some degree of sexual dimorphism in recombination, the vast majority of hotspots are shared between the sexes, with only a small number of putative sex-specific hotspots. Nonetheless, using wavelet analysis, we show that variation in the female and male rates are increasingly correlated at broader scales, indicating that most of the sex-differences in rate can be attributed to the fine scale. Known recombination-associated genomic features, such as genes and certain DNA repeat elements, show systematic differences between the sexes. For example, male recombination appears to be more strongly associated with THE1B repeat elements, whereas females show clear, albeit small, peak of recombination in promoter regions of genes that is absent in males.

MODELING PREDICTION ERROR IMPROVES POWER OF TRANSCRIPTOME-WIDE ASSOCIATION STUDIES

Kunal Bhutani*[1,2], Abhishek Sarkar*[3], Alexander Gusev[4], Manolis Kellis[3], Nicholas J Schork[2]

[1]University of California, San Diego, Bioinformatics & Systems Biology, La Jolla, CA, [2]J. Craig Venter Institute, Human Genetics, La Jolla, CA, [3]Massachusetts Institute of Technology, Computer Science & Artificial Intelligence, Cambridge, MA, [4]Harvard School of Public Health, Department of Epidemiology, Boston, MA

Thousands of loci associated with hundreds of complex diseases have been reported in the NHGRI catalog of genome-wide association studies (GWASs), but most genome-wide significant loci are devoid of protein-coding alterations and likely instead affect transcriptional regulation. Several studies have directly investigated the role of transcriptional regulation on complex diseases by jointly considering genotype, expression, and phenotype. However, such studies require all data to be measured in all samples, which is still prohibitive at the scale of GWAS.

Recent large-scale efforts such as the Gene-Tissue Expression Project (GTEx) have produced reference profiles of transcription, enabling transcriptome-wide association studies (TWASs). TWAS tests for association between gene expression and phenotype by predicting gene expression in GWAS cohorts (where expression is not measured) using models of transcriptional regulation trained on reference transcriptomes. However, current methods for TWAS only use point estimates of imputed expression and ignore uncertainty in the prediction.

Here, we develop a method to explicitly model error in imputed expression and propagate this error through TWAS. We demonstrate that imputed expression has high uncertainty, possibly due to genetic factors not included in typical models, environmental factors which vary between reference transcriptome cohorts, and technical biases in training the predictive models. We show through simulation and application to real data that our method improves power to detect genes associated with phenotype.

* Authors contributed equally

← Back

READ CLOUDS ENABLE ACCURATE HAPLOTYPE-RESOLVED ASSEMBLY OF COMPLEX REGIONS OF THE HUMAN GENOME

Alex Bishara[1], Stephen Mussmann[1], Noah Spies[2], Arend Sidow[2], Serafim Batzoglou[1]

[1]Stanford University, Department of Computer Science, Stanford, CA,
[2]Stanford University, Department of Genetics, Stanford, CA

The widespread adoption of Next Generation Sequencing (NGS) has enabled broad variant discovery across the genomes of a large set of individuals. However, a significant amount of variation lying within high fidelity repeats remains elusive to accurate characterization. In this work, we propose a novel sequence assembly strategy leveraging the 10X Genomics platform to accurately resolve repeats.

Briefly, the 10X Genomics platform isolates long DNA fragments into hundreds of thousands of liquid partitions. Sequencing libraries are then prepared in parallel, such that all short fragments produced from the long molecules within a partition share the same barcode tag. Subsequently, once the library is sequenced, the sequenced barcodes can be used to link reads that originated from the same long molecule. Though each partition naturally contains long-range information by virtue of the input long DNA fragments, the shallow coverage per molecule represents challenges for current de-novo assembly techniques. Namely, within a sequenced molecule, significant gaps in coverage exist, limiting the opportunity to assemble longer contigs from each partition in isolation.

In an earlier work, we presented a novel alignment algorithm that used "read clouds" from a similar strategy (Moleculo, now Illumina Synthetic Long Reads) to accurately map to repeats of a reference genome and uncover variation previously dark to short reads from NGS. We have extended our computational methodology as follows.

First, we use our aligner to map read clouds using the reference genome in order to assist in determining which shallow partitions encapsulate common haploid sequences. Partitions containing a common haploid are then pooled together such that these target sequences are sufficiently covered (~25x) thereby allowing us to leverage existing short read De-Bruijn Graph Assemblers to de novo assemble each pool.

This approach yields assembled contigs from tens to hundreds of kilobases long that span repeats. These contigs allow us to accurately characterize variation in the >5% of the human genome lying within high fidelity repeats and also to discover complex structural variation that remains hidden from current state-of-the-art methods.

We apply our assembly methodology on three trios (NA12878 and her parents, and two GIAB trios) sequenced through 10X, and we show that a significant fraction of the variation obtained from our previous alignment-based strategy actually resides within repeats that are structurally different from those in the current reference genome.

ASSESSING THE CONTRIBUTION OF DNA METHYLATION TO REGULATORY EVOLUTION IN PRIMATES

Julien Roux*[1,2,3], Lauren E Blake*[3], Irene Hernando-Herraez[4], Nicholas E Banovich[3], Raquel Garcia Perez[4], Claudia Chavarria[3], Amy Mitrano[3], Jonathan K Pritchard[5,6,7], Tomas Marques-Bonet[4], Yoav Gilad[3]

[1]University of Lausanne, Department of Ecology and Evolution, Lausanne, Switzerland, [2]University of Lausanne, Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3]University of Chicago, Department of Human Genetics, Chicago, IL, [4]Universitat Pompeu Fabra, Institute of Evolutionary Biology, Barcelona, Spain, [5]Stanford University, Department of Genetics, Stanford, CA, [6]Stanford University, Department of Biology, Stanford, CA, [7]Stanford University, Howard Hughes Medical Institute, Stanford, CA

Gene regulation has long been thought to be a driving force in adaptive evolution. Despite evidence that regulatory changes contribute to many species-specific adaptations, the mechanisms of regulatory evolution remain elusive. We leverage inter-tissue and inter-species comparisons to determine the contribution of DNA methylation changes to the evolution of gene expression.

We assessed CpG methylation status across the genome by performing whole-genome bisulfite conversion followed by high-throughput sequencing across 4 tissues (heart, kidney, liver and lung) in human, chimpanzee, and macaque samples. We collected gene expression profiles from the same samples, allowing us to perform a high resolution scan for genes and pathways whose regulation evolved under selection. By integrating these methylation and expression datasets, we characterized the genomic features where methylation most contributes to expression changes. To understand how epigenetic divergence contributes to gene expression evolution, we modeled the proportion of variation in gene expression levels across tissues and species explained by changes in methylation. We discovered strong negative associations between gene expression and methylation changes across tissues but greatly reduced correlations across species. This may imply that changes in epigenetic regulation are generally not a causal mechanism of primate evolution.

(*Authors contributed equally to this work.)

← Back

RECURRING EXON DELETIONS IN THE *HP* (HAPTOGLOBIN) GENE CONTRIBUTE TO LOWER BLOOD CHOLESTEROL LEVELS

Linda M Boettger[1,2], Rany M Salem[1,2,3,4], Robert E Handsaker[1,2], Gina M Peloso[2,5], Sekar Kathiresan[2,5], Joel N Hirschhorn[1,2,3,4], Steven A McCarroll[1,2]

[1]Harvard Medical School, Department of Genetics, Boston, MA, [2]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [3]Boston Children's Hospital, Division of Endocrinology, Boston, MA, [4]Boston Children's Hospital, Center for Basic and Translational Obesity Research, Boston, MA, [5]Massachusetts General Hospital, Center for Human Genetic Research, Boston, MA

One of the first protein polymorphisms identified in humans involves the abundant blood protein, haptoglobin. Two exons of the *HP* gene exhibit copy number variation (CNV) that affects HP protein structure and multimerization. However, this variation has been largely invisible to genome-wide genetic studies, as it is not observed with high-throughput CNV detection methods and has low linkage disequilibrium to nearby SNPs. The evolutionary origins and medical significance of this polymorphism have been uncertain. We show that this variation has likely arisen from many recurring deletions – more specifically, reversions of an ancient hominin-specific duplication of these exons. An important direction in human genetics is to understand the molecular variation and genetic architectures that underlie the phenotype-genotype associations ascertained in genome-wide association studies (GWAS). Although studying this polymorphism in large cohorts has been historically challenging, we describe a way to analyze it by imputation from SNP haplotypes and find among 22,288 individuals that these *HP* exonic deletions associate with reduced LDL and total cholesterol levels. We further show that these deletions, and a SNP that affects *HP* expression, appear to drive the strong association of cholesterol levels with SNPs near *HP*. Recurring exonic deletions in *HP* likely enhance human health by lowering cholesterol levels in the blood.

← Back

IDENTIFICATION OF SEVEN NOVEL SUSCEPTIBILITY *LOCI* FOR TYPE 2 DIABETES THROUGH GENOTYPE IMPUTATION BASED META-ANALYSIS IN 70,000 EUROPEAN INDIVIDUALS

Silvia Bonàs-Guarch[1], Marta Guindo-Martínez[1], Irene Miguel-Escalada[2], Elias Rodríguez-Fos[1], Friman Sánchez[1,3], Mercè Planas-Fèlix[1], Santiago González[1], Paula Cortés-Sánchez[1], Pascal Timshel[4], Tune H Pers[4], Claire C Morgan[2], Ignasi Moran[2], Carlos Díaz[3], Rosa M Badia[3], José C Florez[5,6], Jorge Ferrer[2], Josep M Mercader*[1], David Torrents*[1,7]

[1]Barcelona Supercomputing Center(BSC-CNS), BSC-CRG-IRB Research program in Computational Biology, Barcelona, Spain, [2]Imperial College London, Department of Medicine, London, United Kingdom, [3]Barcelona Supercomputing Center (BSC-CNS), Computer Sciences Department, Barcelona, Spain, [4]Faculty of Health and Medical Sciences, University of Copenhagen, The Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Genetics, Copenhagen, Denmark, [5]Massachusetts General Hospital, Diabetes Unit and Center for Human Genetic Research, Boston, MA, [6]Broad Institute of MIT and Harvard, Medical and Population Genetics Program, Cambridge, MA, [7]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

Despite the identification of thousands of associated disease *loci* through genome-wide association studies (GWAS), only a small fraction of the estimated heritability has been explained for the vast majority of complex diseases, such as type 2 diabetes (T2D). One potential cause is the poor coverage of variants in the low and rare allele frequency range in early GWAS arrays. In this study, we used 1000 Genomes Project and UK10K reference panels to perform genotype imputation and association testing in ~70K individuals (12,931 cases and 57,196 controls) from six publicly available T2D European ancestry GWAS datasets. This approach allowed us to identify seven novel T2D-associated *loci,* including a missense low-frequency variant (minor allele frequency=0.02, OR=1.19, $p=2.72\times10^{-10}$) in the *EHMT2* gene. Our imputation procedures and dense reference panels provide a robust basis for accurate fine-mapping: for instance, our 99% credible set for the novel *EHMT2 locus* spotlighted an in-frame deletion disrupting several interactions in the *CLIC1* gene, which is a direct target of metformin. In support of the pathophysiological relevance of our findings, gene-set enrichment analysis with DEPICT showed enrichment for genes expressed in pancreas and those related to insulin response. To conclude, our findings highlight the value of data sharing initiatives by showing how the re-analysis of publicly available GWAS data using denser sequencing-based reference panels can be a powerful strategy to advance our understanding of the architecture of complex diseases.

← Back

GENETIC VARIATION REVEALS THE HISTORY OF INVASIONS IN THE INDIAN SUBCONTINENT AND ITS INFLUENCES ON ITS DEMOGRAPHICS

Aritra Bose[1], Peristera Paschou[2], Petros Drineas[3]

[1]Graduate PhD student, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, [2]Associate Professor, Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli, Greece, [3]Associate Professor, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY

Archaeological excavations have discovered artifacts used by early humans, including stone tools, which suggest an extremely early date for human habitation in the Indian subcontinent. Around 30th Century BC, the Indus Valley civilization began in the North-Western Indian subcontinent with the much matured Harappan civilization. De-urbanization of the Harappan civilization started with the Indo-Aryan migration theory and led to the initiation of the Vedic period. Following this, Indian subcontinent has come under a lot of rulers: from Greeks and Scythian to Mauryas, Guptas, Cholas, etc. through the Medieval age. The invasion of the emperor of Ghazni, laid the first seed of Muslim invasion in the subcontinent. This was followed by the Mughal Empire, which lasted over five centuries. In this phase, a range of rulers (Afghans, Turks, Mongols) ruled India.

The aforementioned summary of historical events led to widespread admixture events of the Indian population and influenced the language, culture, caste system, and other demographics in the subcontinent. We set out to explore the population structure of the Indian populations with respect to the history of the subcontinent. We studied genomic variation in 837,279 SNPs genotyped in 1275 individuals, originating from India belonging to prominent castes and language groups in India. We conducted our study by comparing the sample with the populations that invaded India. Our findings illuminate the Indo-Aryan migration theory. We also extended our study to the language groups of India, using population network analysis.

Our results demonstrate that the Afghan and Pathans are closely related to Punjabis and Gujaratis from north-western India, supporting the theory that the Pathan rulers of the 'Delhi Sultanate' expanded their kingdom across all these regions, as well as the presence of Romani tribes in the north-western provinces of India. We also find evidence supporting the presence of caste system in India. Bengalis play an important role as a bridge of gene flow from the North-West to the North-Eastern populations. This establishes the theory that Bengali language is a language of the Indo-European family and it is located in the eastern part of India, closer to Nepal and China. Our results will help identify roots of Indian languages as well as study the evolutionary relationships of South Asian and East Asian populations.

eDiVA: EXOME SEQUENCING ANALYSIS PIPELINE FOR DISEASE GENE IDENTIFICATION

Mattia Bosio[1], Oliver Drechsel[1,2], Rubayte Rahman[1,3], Stephan Ossowski[1,4]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Bioinformatics and Genomics, Dr. Aiguader 88, 08003 Barcelona, Spain, [2]Institute of Molecular Biology (IMB), Bioinformatics core facility, Mainz, Germany, [3]Netherlands Cancer Institute NKI, Research IT, Amsterdam, Netherlands, [4]Universitat Pompeu Fabra UPF, Barcelona, Spain

Novel or inherited genetic variations can lead to drastic phenotypes including rare and common diseases. Human exome analysis using next generation sequencing (Exome-seq) has recently been established as a key approach to identify genetic variations in protein coding genes. Several tools predict the impact of variants on the mutated gene products, and prioritize variants to highlight the disease causing ones. Simple and intuitive workflows leading the researcher from sequencing results towards causal variant prediction reducing the false positive rate are needed to correctly identify potentially disease-causing variants.
We developed eDiVA, available at http://www.ediva.crg.eu, an integrated pipeline that massively facilitates and accelerates the analysis of sequencing data and disease causing variant identification. eDiVA goes from exome-seq alignment and recalibration, to SNP prediction as well as insertion and deletion (InDel) detection using multiple tools to decrease false positive rate. Variants are annotated and enriched with functional information, e.g. damage predictions from SIFT and PolyPhen2, OMICs information from dbSNP, 1000 Genomes, EVS, ExAC, UCSC Genome Browser, KEGG, and OMIM. We developed a rank product based algorithm to prioritize candidate SNPs and InDels using all integrated information. eDiVA supports various disease models, (i.e. autosomal dominant, recessive, de novo, X-linked, and compound heterozygous) to identify correctly segregating mutations in small families and parent-child trios.
eDiVA proved its validity in clinical setting, finding causal variant for mendelian diseases such as familial hyperkalemia, mitral valve prolapse, congenital ataxia, myasthenia, cystic fibrosis and phenylketonuria.
We also developed a benchmark algorithm comparing eDiVA against state-of-the-art tools like Pheno-db and PhenGen measuring precision, recall, and ease to find the causal variant. It is fully reproducible, publicly available and based on real data from ClinVar. Benchmarking demonstrates that eDiVA provides superior variant prioritization compared to similar fast-setup algorithms, and achieves equally good results compared to algorithms requiring human fine-tuning and comprehensive phenotype definitions to work properly.

← Back

CENTRIFUGER: INTERACTIVE ANALYSIS OF MICROBIOMICS DATA FOR PATHOGEN IDENTIFCATION

Florian P Breitwieser, Steven L Salzberg

Johns Hopkins University, IGM / Center for Computational Biology, Baltimore, MD

We present centrifuger, a web application for metagenomics data exploration and pathogen identification. The use of metagenomics sequencing for pathogen identification is emerging as pivotal method in the treatment of infectious diseases. However, the pinpointing of pathogenic species can be difficult, as the samples are usually dominated by host sequences, contaminants and the natural microbiota.

centrifuger enables the visual dissemination of metagenomics classification results at every level of the taxonomical tree. The identifications of multiple samples can be compared and are easily queryable. This allows quicker identification of possible pathogens as well as contaminants that are present in multiple clinical samples, even for non-bioinformaticians.

centrifuger is implemented in the R language using the Shiny framework. It can be hosted on Windows, Mac OS and Linux systems, and used with any modern web browser. centrifuger is freely available under a GPL-3 license from http://github.com/fbreitwieser/centrifuger. Furthermore we provide a publicly available web interface for testing at http://ccb.jhu.edu/software/centrifuger.

GENE DISCOVERY IN CHILDHOOD-ONSET SCHIZOPHRENIA INCLUDING A NOVEL MUTATION IN ATP1A3

Catherine A Brownstein[1], Niklas Smedemark-Margulies[1], Meghan C Towne[1], Alan H Beggs[1], Pankaj B Agrawal[1], Joseph Gonzalez-Heydrich[2]

[1]Boston Children's Hospital Manton Center for Orphan Disease Research, Division of Genetics and Genomics, Boston, MA, [2]Boston Children's Hospital, Developmental Neuropsychiatry Program, Boston, MA

The study of rare Mendelian forms of Juvenile Psychosis is an effective way to discover new candidate genes for the condition. Boston Children's Hospital is developing the infrastructure needed for large-scale psychiatric research and treatment discovery with the creation of the Developmental Neuropsychiatry Program (DNP). The DNP seeks to develop therapeutics to prevent the development of schizophrenia in at-risk children by identifying causative mutations in the youngest patients presenting with psychosis, creating neuronal cell cultures and models of neural networks expressing the mutations and using these to screen for novel therapeutics.

As part of this process, our group has consented 37 probands with very early onset psychosis (defined as psychosis by age 12) and their available first-degree relatives for genetic study. We have performed genetic testing on 23 of these probands (CMA and exome) and identified 16 potentially disease causing mutations or CNVs within this sample. Thus far, 7 of these genes have been investigated and are likely to incur a high risk for psychosis.

As an example of the progress being made, we describe a patient with onset of command auditory hallucinations and behavioral regression at age 6 in the context of longer standing selective mutism, aggression, and mild motor delays. Sequencing revealed a previously unreported heterozygous de novo mutation at c.385G>A in ATP1A3. This gene codes for a neuron-specific isoform of the catalytic alpha subunit of the ATP-dependent transmembrane sodium-potassium pump. Heterozygous mutations in this gene have been reported as causing both sporadic and inherited forms of Alternating Hemiplegia of Childhood and Rapid-onset Dystonia Parkinsonism. This protein is crucial to establishing proper transmembrane ion gradients, and resting membrane repolarization, action potential conduction, secondary active transport of calcium, and neurotransmitter release and recycling. Studies have shown that heterozygous knockout in animal models (Ikeda et al., 2013) or point mutations of this protein cause altered synaptic activity and behavioral abnormalities (Clapcote et al., 2009; Kirshenbaum et al, 2011). Functional analysis is underway including characterization of the mutated protein's function in single cell models and IPSC derived neurons.

ANNOTATION OF THE CHICKEN AND OTHER AVIAN GENOMES.

David W Burt, Richard Kuo, Lel Eory

The Roslin Institute/University of Edinburgh, Genomics and Genetics, Easter Bush Campus, United Kingdom

Our knowledge of avian genomes has increased rapidly over the past few years, starting with the publication of the chicken genome in 2004, a milestone in the fields of avian genetics and comparative genomics. Advances in DNA sequencing technology now make it possible to produce draft sequences of any vertebrate genome, quickly and cheaply. We have seen the completion of draft genomes of 100's of other birds, with plans to sequence all 10,000 by the B10K Consortium. With advances in long read sequencing technologies, we are now seeing genome assemblies moving from draft quality to genome quality with N50 contigs of more than 10 Mb. However to fully exploit these resources we need comprehensive annotation of the genes and regulatory elements within these genome assemblies.

The annotation of genes in the chicken genome has been under continuous improvement, taking advantage of transcriptome data generated by short and now, long read sequencing technologies. These approaches provide experimental data, improved gene models, with complex patterns of transcription including multiple RNA isoforms. In the chicken this has worked well for both coding and non-coding RNA genes, defining more than 40K genes.

Recently, the analysis of 44 bird genomes by the *Avian Phylogenomics Consortium* has opened up new opportunities. For individual species, the sequences coupled with the initial annotations, can serve as a vehicle for basic research. On the other hand generating a multiple sequence alignment (MSA) of all these genomes can enable comparative studies, which benefits all these species. Such studies broaden our understanding of genome evolution (e.g. using the MSA we have define more than 1.5M constrained elements, with 1M unique to birds) and the evolution of traits or can help to disentangle phylogenetic relationships. Our main aim is to analyse the integrated data with a focus on creating a detailed functional map relevant to the chicken and other birds. Such a map can be used to drive the identification of novel protein-coding and non-coding genes, binding sites for transcription factors, enhancers or other functional elements.

Finally, the challenge is to move beyond genes (“What is a gene anyway?”) and start to define regulatory elements. This is now the priority of the Functional Annotation of Animal genome (FAANG) Consortium in the coming years.

Availability: http://avianbase.narf.ac.uk/index.html; http://www.ensembl.org/index.html; http://www.faang.org

Funding: This work was supported by grants from the BBSRC (UK), EC, The Wellcome Trust (UK), University of Edinburgh (UK) and Cobb-Vantress (USA).

← Back

EVIDENCE FOR ADAPTIVE GENE-FLOW IN RECENT AFRICAN HISTORY

George Busby[1], Ryan Christ[1,2], Quang Si Le[1], Gavin Band[1], Ellen Leffler[1,3], Kirk Rockett[1,3], MalariaGEN Consortium[1,3], Dominic Kwiatkowski[1,3], Chris Spencer[1]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]University of Oxford, Department of Statistics, Oxford, United Kingdom, [3]Wellcome Genome Campus, Wellcome Trust Sanger Institute, Cambridge, United Kingdom

An opportunity for adaptive gene-flow occurs when genetically differentiated groups come together. A consequence of this admixture, which is increasingly regarded as a common feature of human populations, is that most human groups harbour a mosaic of ancestries. In the human lineage, there is growing evidence that ancient hybridisation with Neanderthals and Denisovans led to the introgression of beneficial alleles. Whether more recent gene-flow has contributed to adaptation has thus far received little attention. Here we assess evidence that new haplotypes, which have been introduced into populations by gene-flow, have spread by natural selection.

We develop new statistical approaches that search the genome for regions where ancestry significantly deviates from expectations. Applying our models to the Fulani from West Africa, a population which has experienced recent Eurasian gene-flow, our methods identify genes involved with lactase persistence and the Duffy-null phenotype as the regions of the genome with the highest and lowest levels of Eurasian ancestry respectively. When we extend our analysis to a broad dataset of 48 African and 12 Eurasian populations, we identify several loci where ancestry differs from expectations, potentially as a result of natural selection. Together with a detailed description of recent admixture with the same dataset, our results suggest an important role for adaptive gene-flow in Africa within the last 4,000 years. We discuss the role of adaptive gene-flow as a key evolutionary process.

GENETIC VARIANTS CONTRIBUTING TO TAME BEHAVIOR IN DOMESTICATED ANIMALS

Alex Cagan[1], Frank Albert[2], Irina Plyusnina[3], Lyudmila Trut[3], Rimma Kozhemjakina[3], Rimma Gulevich[3], Oleg Trapezov[3], Nikolay Yudin[3], Yury Herbeck[3], Victor Wiebe[1], Gabriel Renaud[1], Frederic Romagne[1], Verena Behringer[1], Roisin Murtagh[1], Tobias Deschner[1], Torsten Schöneberg[4], Svante Pääbo[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]University of Minnesota, Department of Genetics, Cell Biology and Development, Minneapolis, MN, [3]Siberian Branch of the Russian Academy of Sciences, Institute of Cytology and Genetics, Novosibirsk, Russia, [4]University of Leipzig, Institute of Biochemistry, Leipzig, Germany

Domesticated animal species share a suite of phenotypic changes, known as the 'domestication syndrome'. The molecular basis of these changes remains largely unknown. It has been hypothesized that many of the phenotypic changes associated with animal domestication are pleiotropic effects from selection for tame behavior. To identify the genetic changes underlying animal domestication we generated whole-genome sequences from Norway rats (*Rattus norvegicus*) and American mink (*Neovison vison*) divergently selected for their behavioral response to humans. We analyzed these together with genome sequence data from populations of several pairs of domestic animals and their closest extant wild relatives. While we find no single gene with evidence of selection in all species we identify several biological pathways that appear to be consistently involved in the process of animal domestication. Among the strongest candidates are genes involved in migration of neural crest cells, supporting the hypothesis that neural crest migration is involved in aspects of domestication. One such gene is *ECE1*, which encodes endothelin converting enzyme 1. Mutations in ECE1 affect the development of the hindbrain and coat color spotting.

RATES OF EVOLUTION AMONG SPERM GENES AND IMPLICATIONS FOR SPECIATION IN A SMALL NOCTURNAL PRIMATE, GENUS *MICROCEBUS*

C. Ryan Campbell, Matthew Dubin, Anne D Yoder

Duke University, Biology, Durham, NC

The mouse lemurs, genus *Microcebus*, are 21 species of small nocturnal primates that constitute a 10 million year old evolutionary radiation on the island of Madagascar. Although these primates have been the recent focus of next generation sequencing methods, the initial applications of these data have been centered around understanding the relationships among and between the species rather than the mechanisms that underlie the diversification of the species radiation. There is general consensus that these nocturnal and matrilocal species show low paternal care and a strong propensity towards sperm competition, which is manifested both behaviorally and morphologically during breeding season. High levels of sperm competition between males could in turn reinforce nascent species boundaries. In our study, we compared the rates of positive selection within spermatogenesis-related genes to a set of random non-spermatogenesis genes to determine if the former are evolving faster than the latter. These comparisons were made between several species of *Microcebus* and show that there are significantly more nonsynonymous base substitutions in spermatogenesis-related genes relative to non-spermatogenesis genes. These results provide molecular data that match our current knowledge of the behavior and lifestyles of these primates. They also highlight what could be an underlying mechanism of speciation among these highly speciose primates of Madagascar.

GENETIC ADAPTATION TO LEVELS OF SELENIUM IN THE DIET IN HUMANS AND OTHER VERTEBRATES

Gaurab K Sarangi, Louise White, Aida M Andrés, Sergi Castellano

Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany

Selenium is an essential micronutrient in the diet of humans and other vertebrates whose deficiency causes infertility, immune dysfunction and an increased mortality risk. Diet is the most important source of selenium and its intake depends on its levels on the soil or water on which food is gathered, hunted or grown. Selenium levels, in turn, depends on the underlying bedrock from which soils are formed with the sea being an environmental sink for selenium. This has created a patchwork of selenium deficient, adequate and occasionally toxic areas which humans and other vertebrates encountered as they settled the world. Selenium is required by vertebrates due to its function in selenoproteins, which contain the amino acid selenocysteine (Sec) as one of their constituent residues. Sec is the 21st amino acid in the genetic code and is encoded by a UGA (STOP) codon. A complex machinery of dozens of proteins is necessary for the recoding and incorporation of Sec into proteins.

We show here that deficiency or abundance of selenium have distinctly shaped the evolution of vertebrate species. Humans from areas that do not provide adequate levels of selenium have polygenic signatures of adaptation in both selenoprotein genes and genes involved in the metabolism and homeostasis of this essential micronutrient, whereas fishes have additional rare selenoprotein genes and gene duplications in which one copy appears to have evolved novel functions (neo-functionalization) that rely on selenium. Hence, vertebrates appear to have adapted to the levels of selenium in their environment in different ways throughout their history.

PARTITIONING SINGLE-MOLECULE SEQUENCING FROM SEQUENCE PARALOGS DE NOVO

Mark J Chaisson, Chris Hill, David Gordon, Evan Eichler

University of Washington, Genome Sciences, Seattle, WA

The de novo assembly of genomes using single-molecule sequences (SMS) has lead to dramatic improvements in the contiguity of assemblies. An analysis of SMS-based human assemblies reveals that while effectively all unique sequences are resolved, assemblies break down over regions of segmental duplication that are over 50kbp in length and 98% identity. Thus duplication content is the Achilles heel of de novo assembly. This problem is inherent from the challenge of determining the true overlapping reads from multiple highly similar repeat paralogs, subject to low SMS read accuracy. At best, the outcome of the assembly of a segmentally duplicated sequence may be a single collapsed consensus of many different repeat paralogs. Fortunately, while duplicated paralogs may have high sequence identity, they are rarely exact. For the instance where an assembly contains a single representation of a duplicated sequence, we have developed an approach to partition de novo reads according to paralog, using an approach motivated by polyploid phasing. As an example, we are able to accurately partition sequences from a collapsed 5-copy 40kbp segmental duplication in from a de novo SMS assembly of the haploid hydatidiform mole CHM1. Our novel method allows independent assembly of each read partition giving resolution of sequence of segmental duplications, with the possibility of being subsequently scaffolded to determine segmental duplication architecture.

INTEGRATED METADATA-DRIVEN ACCESS OF ENCODE, MODENCODE, REMC, GGR, AND MODERN DATA THROUGH A COMMON PORTAL

Esther T Chan, Aditi K Narayanan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Marcus Ho, Kathrina C Onate, J Seth Strattan, Laurence D Rowe, Forrest Y Tanaka, Ulugbek K Baymuradov, Stuart R Miyasato, Matt Simison, Benjamin C Hitz, Cricket A Sloan, J Michael Cherry

Stanford University, Department of Genetics, Stanford, CA

The efforts of large-scale NIH-funded projects such as the Encyclopedia of DNA Elements (ENCODE), its model organism corollary (modENCODE), and the Roadmap Epigenomics Mapping Consortium (REMC) have resulted in the accumulation of genome-wide maps of candidate functional sequence elements and epigenetic marks in a large variety of cells in human and other model organisms. Consideration of the provenance of experimental reagents and transparency of computational analyses is crucial to the interpretation and comparison of these data. Tracking this information consistently across different biochemical assays employed in thousands of experiments across hundreds of cell and tissue types is particularly challenging at the scale of these projects.

The ENCODE Data Coordination Center (DCC) has developed a flexible and rich data model to capture information such as key experimental variables, details of experimental and analysis methods, what software and pipelines were used to produce which files for the ENCODE project, and calculated quality metrics, known collectively as metadata. This data model has since been extended to additionally integrate curated data and metadata from related projects such as REMC, modENCODE, Model organism Encyclopedia of Regulatory Networks (modERN) and the Genomics of Gene Regulation (GGR). The resulting publicly accessible data corpus encompasses nearly 10,000 experiments and is accessible through the ENCODE portal (https://www.encodeproject.org), which features a powerful faceted browsing interface, full-text search, and a REST API so users can easily search, filter, download and visualize the collection. Data integration and release is continual and ongoing. Learn more about how to get started on the ENCODE portal here: https://www.encodeproject.org/help/getting-started/.

DIFFERENTIAL MICROBIAL COMPOSITION ASSOCIATED WITH ASTHMA

Ti-Cheng Chang[1], Jason Rosch[2], Amali Samarasinghe[3]

[1]St. Jude Children's Research Hospital, Computational Biology, Memphis, TN, [2]St. Jude Children's Research Hospital, Infectious Disease, Memphis, TN, [3]University of Tennessee, Pediatrics, Memphis, TN

The airways were considered sterile under steady conditions. However, increasing number of studies has shown a commensal microbiome is present in the airway. Complex microbial communities have been observed in bronchiectasis and chronic obstructive ling disease through molecular experiments, whereas the knowledge about the constituents of airway microbiome in asthma is limited. Exposure of the diversified bacteria in the airway can elicit allergic immune response and is associated with etiology of asthma. Here, we investigated the entire respiratory microbial metagenome in airway samples from carefully phenotyped subjects across a spectrum of asthma and health mouse models.

The analysis included 39 bronchoalveolar lavage samples from murine subjects (6 asthma, 5 flu, 4 bacterial challenge, 4 flu + bacteria, 6 asthma + flu, 3 asthma + bacteria, 6 asthma + flu +bacteria). We applied 16S rRNA gene sequencing to profile the constituent of bacterial communities in each sample. Data were analyzed using the QIIME and phyloseq pipelines. The samples typically contained 20,000–36,000 reads that have taxonomic assignment. Bacterial richness was not significantly different among the groups of samples, with the exception of the asthma + bacteria (AB) samples that have a relatively low bacterial diversity. Hierarchical clustering and PCA analysis revealed that the AB samples formed a tight group with a strong support (bootstrap value 100) and subsequently clustered with the majority of the AFB samples. Large variations of the bacteria richness between AB and control samples were shown in Cyanobacteria (ML635J-21), Firmicutes (Streptococcus), and Proteobacteria (Caulobacteraceae and Burkholderiaceae). The abundance of Caulobacteraceae decreased significantly across all the groups compared to the control samples.

The results suggest the presence of a distinct airway microbiome in the asthma samples with different underlying infections. The samples with the same infections tend to comprise similar microbioal constituents. It remains to be seen whether the abundance variation of Caulobacteraceae is associated with severity of asthma.

A POLYMORPHIC ERV ELEMENT THAT IS MOBILIZED IN THE GERMLINE AT A RATE THAT VARIES BETWEEN INDIVIDUALS CAUSES CHOLESTEROL DEFICIENCY BY DISRUPTING THE BOVINE APOB GENE.

Chad Harland[1,2], Keith Durkin[1], Maria Artesi[1], Latifa Karim[1,3], Arnaud Sartelet[4], Emilie Knapp[4], Nico Tamma[1], Erik Mullaart[5], Richard Spelman[2], Wouter Coppieters[1,3], Michel Georges[1], Carole Charlier[1]

[1]Unit of Animal Genomics, GIGA-R, University of Liège, Liège, Belgium, [2]Livestock Improvement Corporation, Research & Development, Hamilton, New Zealand, [3]GIGA-Genomics Platform, University of Liège, Liège, Belgium, [4]Bovine Clinic, FARAH, University of Liège, Liège, Belgium, [5]CRV, Research & Development, Arnhem, Netherlands

A lethal hypolipidemia with autosomal recessive mode of inheritance was recently described in Holstein Friesian Cattle. The corresponding locus was assigned to a 2.6 Mb interval (BTA11:74.5-77.1Mb) by autozygosity mapping (Kipp et al., 2015). We herein show that the causative mutation corresponds to the sense insertion of a ~7kb full-length *bos Taurus* endogenous retroviral element (*Bo*ERV) in exon 5 of the *Apolipoprotein B* gene (*APOB*), resulting in complete transcriptional termination downstream to the insertion point.
We developed the '*LocaTER*' (*Localization of Transposable Endogenous Retroviral elements*) bioinformatics pipeline to identify and characterize more than 1,000 ERV insertion sites that are polymorphic in modern cattle breeds. Underrepresentation and shifts towards lower allelic frequencies of genic (vs intergenic) sense (vs antisense) insertions testifies of their capacity to cause deleterious phenotypic effects.
We took advantage of the large '*Damona*' whole genome sequence dataset of > 750 individuals - designed for the detection of *de novo* mutation - to pinpoint and validate five *de novo* germ-line ERV transposition events, including three genic sense insertions in *GARBQ*, *CHST11* and *CYTIP*. We estimated the average transposition rate at one event per ~50 gametes with indication of a ~4-fold higher transposition rate in males than in females. Intriguingly, three of the four male transposition events occurred in the germ-line of a single bull. Moreover, two of these three de novo insertions were transmitted by the same sperm cell. Such a striking - animal-specific - burst of de novo events may be related to interindividual variation in the genomic defense mechanisms that control the strength of repression of ERV in the germ-line.

← Back

SVTOOLS: SCALABLE SV DETECTION AND INTERPRETATION FOR POPULATION-SCALE WGS STUDIES

Colby Chiang[1], David E Larson[1], Abhijit Badve[1], Haley J Abel[1], Liron Ganel[1], Ryan M Layer[2], Aaron R Quinlan[2], Ira M Hall[1,3]

[1]Washington University, McDonnell Genome Institute, St. Louis, MO, [2]University of Utah School of Medicine, Department of Human Genetics, Salt Lake City, UT, [3]Washington University, Department of Medicine, St. Louis, MO

Structural variation (SV) is a broad class of genome variation that includes copy number variants, balanced rearrangements and mobile element insertions. SV is recognized to be an important source of human genetic diversity, but SVs have not been ascertained in most large-scale studies of common human disease, and their phenotypic contributions remain a matter of debate. The forthcoming wave of large-scale WGS-based studies (CCDG, TOPMed, etc.) presents an opportunity to investigate the functional consequences of SV and, in a more general sense, to improve study power by incorporating SVs into common and rare variant association analyses. However, current SV detection algorithms scale poorly due to memory usage requirements, and are unable to jointly analyze more than several hundred genomes.

Here, we present SVTools (https://github.com/hall-lab/svtools), an innovative yet practical solution to enable efficient SV analysis on tens of thousands of human genomes without sacrificing variant detection performance. Built on the backbone of the LUMPY SV detection algorithm and the SpeedSeq genome analysis pipeline, and loosely modeled after the "N+1" paradigm of the Genome Analysis Toolkit (GATK), SVTools generates lossless gVCF-like files that include the raw probability curves for each putative breakpoint. These sample-specific variant files, which can be generated in parallel for extreme computational efficiency, are then probabilistically merged into a single cohort-level VCF that simultaneously integrates SVs across samples while refining the spatial precision of each breakpoint. The resulting re-genotyped VCF is more precise, more sensitive, and reduces the technical noise from heterogeneity in data quality across the cohort. SVTools also includes several data refinement modules designed to simplify interpretation and filtering, including an SV reclassification step that revises variant types based on read-depth and mobile element annotations.

These methods will enable comprehensive SV analysis in population-scale WGS datasets, which promises to improve our understanding of both common and rare human disease. As proof of principle, we present preliminary results from our analysis of rare structural variation in ~5,000 deeply sequenced human genomes.

← Back

FULLY PHASED ASSEMBLY OF HLA GENES USING LINKED-READS

Anton Valouev, David B Jaffe, Neil I Weisenfeld, Heather Ordonez, Adrian N Fehr, Patrick Marks, Michael Schnall-Levin, Tarjei S Mikkelsen, Deanna Church

10x Genomics, R&D, Pleasanton, CA

MHC compatibility is critically important for the efficacy of allogenic transplantation therapies, including organ replacement, bone marrow transplantation during cancer treatments and other stem cell-based therapies. Successful therapy relies on accurate matching of MHC alleles between transplant donor and recipient. Variation in the MHC region is also implicated in a number of genetic diseases, though some of these associations involve non-MHC genes such as the recent association of C4 alleles with schizophrenia. Extremely high degree of allelic diversity among individuals, homology among MHC genes and the presence of numerous pseudogenes makes it challenging to accurately and unambiguously infer haplotypes by common HLA typing methods. Emerging short-read NGS-based approaches are limited in their ability to fully reconstruct the HLA locus due to lack of accurate variant phasing between MHC genes and limited sensitivity to structural rearrangements. Amplicon-based approaches (including long-read approaches) for assaying MHC are limited to genic regions and subject to challenges such as allelic dropout. We have developed the 10X GemCode Platform, which combines microfluidics and molecular barcoding with custom bioinformatics software to obtain sequence information across long (50-100+ Kb) DNA molecules using Linked-Reads. Here, we present results on the phased reconstruction of complete human MHC genes and intergenic sequences from Linked-Read data. We demonstrate this for both whole genome sequencing libraries and libraries enriched via hybrid capture of the 3.6 Mb MHC region. Our method relies on a novel Linked-Read assembler. We have evaluated this method on a number of samples with MHC regions characterized via orthogonal methodologies, including three cell lines that have been extensively characterized: NA18555, NA18532, NA12878. Our de novo MHC assemblies show high concordance with orthogonal methodologies on the key MHC genes, while also measuring extensive variation not captured by these methods. Our results demonstrate a path towards comprehensive identification and complete phasing of genic and intergenic variation across the entire MHC region using cost-efficient short read sequencers.

INHERITED DAMAGING MUTATIONS IN IMMUNE-RELATED GENES FAVOUR THE DEVELOPMENT OF GENETICALLY HETEROGENEOUS SYNCHRONOUS COLORECTAL CANCER.

Matteo Cereda[1], Gennaro Gambardella[1], Lorena Benedetti[1], Fabio Iannelli[2], Luigi Laghi[3], Jo Spencer[4], Manuel Rodriguez-Justo[5], Francesca D Ciccarelli[1]

[1]King's College London, Division of Cancer Studies, London, United Kingdom, [2]IFOM, Experimental Oncology, Milan, Italy, [3]Istituto Humanitas, Molecular Gastroenterology, Milan, Italy, [4]King's College London, Immunobiology, London, United Kingdom, [5]UCL, Research Pathology, London, United Kingdom

Despite all efforts to characterise the genomic landscape of colorectal cancer, several questions still remain unaddressed. For example, around 5% of patients present primary tumours at initial diagnosis, referred as synchronous colorectal cancer, syCRC. The causes of syCRC are however still poorly understood. Hereditary conditions such as Lynch syndrome and familial adenomatous polyposis (FAP) only account for around 10% of all syCRC, thus suggesting that other predisposing causes exist.
To understand how genetics and environment influence the development of multiple tumours, we have performed a systematic genomic profiling of syCRCs from several patients with the aim to compare their genomic landscape. We observed that syCRCs originating in the same patients have independent genetic origins, acquire discordant driver alterations, and follow different clonal developments. This inter- and intratumour heterogeneity has consequences on the clinical management of syCRC patients, in terms of both efficacy of therapy and monitoring of drug resistance. To search for evidence of genetic predispositions to the development of multiple tumours, we have analysed the inherited genotype of syCRC patients. Instead than of a single gene, we hypothesised that syCRC results from the constitutional alteration of several genes all contributing to the same biological process. To detect such altered processes, we developed the Mutation Enrichment Gene set Analysis (MEGA). MEGA systematically compares the cumulative distribution of mutations within a process between two cohorts and identifies those processes that are overall more frequently altered in only one cohort. We found that syCRC patients show a significantly higher occurrence of inherited damaging mutations in immune-related genes as compared to patients with solitary colorectal cancer and to healthy individuals from the 1000 Genomes Project. Both the normal and the tumour colonic mucosa of syCRC patients have abnormal compositions of immune cell populations and syCRCs show somatic deregulation of immune-related transcriptional processes. This suggests the presence of an environmental field effect that promotes multiple tumours in the background of inflammation.

UNLOCKING BREAD WHEAT GENOME DIVERSITY WITH NEW SEQUENCING AND ASSEMBLY APPROACHES

Matthew D Clark[1], Bernardo Clavijo[2], Luca Venturini[3], Gonzalo Garcia[2], Jon Wright[2], David Swarbrek[3], Ksenia Krasileva[4], Michael Bevan[6], Federica di Palma[5]

[1]The Genome Analysis Centre, Plant and Microbial Genomics, Norwich, United Kingdom, [2]The Genome Analysis Centre, Assembly Algorithm Development Team, Norwich, United Kingdom, [3]The Genome Analysis Centre, Regulatory Genomics, Norwich, United Kingdom, [4]The Genome Analysis Centre, Triticeae Genomics, Norwich, United Kingdom, [5]The Genome Analysis Centre, Vertebrate and Health Genomics, Norwich, United Kingdom, [6]John Innes Centre, Cell and Developmental Biology, Norwich, United Kingdom

Wheat, a plant species originating in the Levant approximately 10,000 years ago is now the most widely grown crop across the world, spanning from Canada to Argentina in the Americas and from Portugal to China in Eurasia. This ability to thrive in such a variety of environments may be enabled by its genetic complexity – the largest and most complex of all crop genomes. A complete wheat genome sequence would empower the breeding of improved wheat varieties and reveal man's selection over thousands of years to the crop we now see. Yet to date wheat's large (17Gbp) hexaploid and highly repetitive (>80%) genome has only been partially assembled and just to short sequences. The usefulness of such assemblies are limited by their partial and highly fragmented nature, with many genes missing and most genes not contained in single contigs. Here we present the most complete and contiguous wheat assembly of the reference cultivar to date, a new comprehensive gene annotation, and the first insights into wheat variation as revealed by the assembly and comparison of multiple wheat cultivars.

UNCOVERING THE DIVERSITY OF COMPLEX STRUCTURAL VARIATION IN 465 AUTISM GENOMES WITH MULTIPLE WHOLE-GENOME SEQUENCING TECHNOLOGIES

Ryan L Collins*[1], Harrison Brand*[1,2], Carrie Hanscom[1], Matthew R Stone[1], Joseph T Glessner[1,2], Claire E Redin[1,2], Caroline Antolik[1], Stephan J Sanders[3], Michael E Talkowski[1,2,4]

[1]Massachusetts General Hospital, Center for Human Genetics Research, Boston, MA, [2]Department of Neurology, Harvard Medical School, Boston, MA, [3]University of California San Francisco, Department of Psychiatry, San Francisco, CA, [4]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA

Structural variation (SV), or rearrangements of segments of hundreds to millions of nucleotides, is the predominant factor in determining the content of any individual human genome and is associated with many diseases. Despite this, our understanding of balanced and complex SV cryptic to many conventional technologies like chromosomal microarray (CMA) remains limited. Here, we applied a combination of long-insert, short-insert, and 10X Genomics linked-read whole-genome sequencing (WGS) in 465 autism spectrum disorder (ASD) subjects and 120 family members to characterize the full mutational landscape of SV in ASD. We discovered a diverse spectrum of seven core overarching SV classes, and systematically categorized 13 distinct subclasses of recurrent relatively large complex SVs, at least one of which was observed in all individuals and 85.2% of which included at least one inverted segment. Strikingly, the majority of SV involving at least one inverted sequence (60.4%) were rearrangements other than simple, canonical inversions, suggesting most large inversion variation in the human genome is non-canonical. The most frequently observed classes of complex SV were inversions flanked by CNVs at one or both breakpoints, with flanking duplications being significantly larger than flanking deletions ($p=1.0x10-5$) yet approximately three times less frequent, with implications for mechanisms of complex SV formation. Finally, by applying 10X Genomics linked-read WGS in a series of ASD families, we were able to completely resolve and phase several large multi-breakpoint complex inversions that were recalcitrant to validation by both standard siWGS and molecular methods. Upon orthogonal validation of 49.5% of all SV sites discovered and interpreting these SV with genome-wide convergent genomics assessed from exome sequencing and CMA data aggregated from 133,819 individuals, we estimated the overall contribution of de novo SVs at the resolution of liWGS (~5kb) in ASD to be at least 8.5%, or 2.1-fold greater than CMA-based CNV analyses in these same samples and on a par with the projected contribution of de novo loss-of-function exome mutations, suggesting that cryptic and complex SV represent an important and presently underappreciated component of disease etiology.

← Back

ADVANCED APPLICATIONS FOR CLINICAL RESEARCH EXOMES

Margherita Corioni, Eric Lin, Kyeong-Soo Jeong, Carlos Pabon, Arjun Vadapalli, Francisco Useche, Marc Visitacion, Gilbert Amparo, Madhuvanthi Ramaiah, Magnus Isaksson, Douglas Roberts

Agilent Technologies, Diagnostics and Genomics, Santa Clara, CA

Targeted re-sequencing enables highly sensitive and comprehensive detection of variants and provides insights into the biology behind a given phenotype. With recent major advances in this technology, combined with a better understanding of biological pathways, NGS is now extensively considered for use in clinical research. Whole exome sequencing provides deep coverage of genomic content from curated databases making it an efficient solution to characterize and catalogue variants. We describe several strategies that add additional value to whole exome analysis enabling more complete coverage and advanced applications such as detection of copy number changes and phasing. We share an augmented exome for clinical research that focuses on targets relevant to constitutional disease research with input from specific databases including Human Gene Mutation Database (HGMD®), ClinVar and Online Mendelian Inheritance in Man (OMIM™). We also describe the combination of whole exome with copy number probes to extend target enrichment applications to the detection of genome-wide copy number changes (CNCs), and copy neutral loss-of-heterozygosity (cnLOH). Finally, we demonstrate a phased exome coupled to 10x Genomics' linked read technology. The addition of "phasing" probes allows determination of long haplotype blocks. The separation of exonic reads into haplotypes further aids in structural variant and translocation determination, especially for disease causality in complex genotypes. We describe data analysis and performance attributes for each of these platforms, including SNP/INDEL, loss of heterozygosity, copy number variation, phasing, translocation, and resolution of complex compound heterozygote samples.

DECIPHERING THE REGULATORY TRANSCRIPTIONAL NETWORK CONTROLLING REGENERATION

Elena Vizcaya[1,2], Cecilia Klein[3,4], Florenci Serras[1,2], Roderic Guigo[3,4], Montserrat Corominas[1,2]

[1]Universitat de Barcelona, Genetics, Barcelona, Spain, [2]Institut de Biomedicina (IBUB), Barcelona, Spain, [3]Center for Genomic Regulation, Barcelona, Spain, [4]Universitat Pompeu Fabra, Barcelona, Spain

Most organisms possess some ability to repair and regenerate damaged tissues. Certain organisms can regenerate whole body parts or entire limbs while others can only superficially seal wounds or restore small patches of tissue. One of the main challenges in current biology is to understand the nature of these differences and thus unveil the genetic mechanisms required for regeneration. Successful regeneration processes demand a hierarchal and well- controlled balance between proliferation, differentiation and metabolic functions, which are mostly orchestrated by signaling molecules and transcriptional regulation. Although similar gene networks participate in development and regeneration, there are differences in the intensity of the signals or the levels of transcription. The ultimate goal of our research group is to understand how transcription is regulated during development and regeneration using Drosophila wing imaginal discs, epithelia that develop adult structures and are able to regenerate. Wing discs are able to regenerate upon an injury or cell death. We have used a genetic approach to study regeneration, which consists in genetic activation of apoptosis, followed by RNA-Seq and ATAC-Seq analyses at different times after induction of damage. We have found an enrichment of gene categories such as cell cycle, transcription regulation, mitochondrial and oxidative control among candidate genes that increase expression after damage. Conversely, we have found an enrichment of genes involved in categories like pattern formation, respiratory chain, cytoskeleton organization and negative regulation of cell commitment among genes whose expression is repressed. At this time we are correlating expression data with chromatin accessibility and applying motif discovery tools to search for transcription factors that could bind to the identified regulatory regions. The ultimate goal is to elucidate how transcription factors drive gene expression programs through interaction with genomic elements, such as enhancers, and contribute to the control of gene expression changes in normal development in comparison to regeneration.

INTEGRATION AND FIXATION PREFERENCES OF HUMAN AND MOUSE ENDOGENOUS RETROVIRUSES UNCOVERED WITH FUNCTIONAL DATA ANALYSIS.

Marzia A Cremona*[1], Rebeca Campos-Sanchez*[2], Pini Alessia[3], Francesca Chiaromonte[1,5], Kateryna D Makova[4,5]

[1]Penn State University, Department of Statistics, University Park, PA, [2]Universidad de Costa Rica, Centro de Investigación en Biología Celular y Molecular, San José, Costa Rica, [3]Politecnico di Milano, MOX - Department of Mathematics, Milano, Italy, [4]Penn State University, Department of Biology, University Park, PA, [5]Penn State University, Center for Medical Genomics, The Huck Institutes of the Life Sciences, University Park, PA

* These authors have contributed equally

Endogenous retroviruses (ERVs), the remnants of retroviral infections in the germ line, occupy ~8% and ~10% of the human and mouse genomes, respectively, and affect their structure, evolution, and function. Yet we still have a limited understanding of how the genomic landscape influences integration and fixation of ERVs.
Here we conducted a genome-wide study of the most recently active ERVs in the human and mouse genome. We investigated 872 fixed and 1,208 ex vivo HERV-Ks in human, and 1,624 fixed and 242 polymorphic ETns, as well as 3,964 fixed and 1,986 polymorphic IAPs, in mouse. We quantitated >40 human and mouse genomic features (e.g., non-B DNA structure, recombination rates, and histone modifications) in ±32 kb of these ERVs' integration sites and in control regions, and analyzed them using Functional Data Analysis (FDA) methodology. In one of the first applications of FDA in genomics, we identified genomic scales and locations at which these features display their influence, and how they work in concert, to provide signals essential for integration and fixation of ERVs.
The investigation of ERVs of different evolutionary ages (young ex vivo and polymorphic ERVs, older fixed ERVs) allowed us to disentangle integration vs. fixation preferences. As a result of these analyses, we built a comprehensive model explaining the uneven distribution of ERVs along the genome. We found that ERVs integrate in late-replicating AT-rich regions with abundant microsatellites, mirror repeats, and repressive histone marks. Regions favoring fixation are depleted of genes and evolutionarily conserved elements, and have low recombination rates, reflecting the effects of purifying selection and ectopic recombination removing ERVs from the genome.
In addition to providing these biological insights, our study demonstrates the power of exploiting multiple scales and localization with FDA. These powerful techniques are expected to be applicable to many other genomic investigations.

← Back

EUPATHDB: INTEGRATING EUKARYOTIC PATHOGEN GENOMICS DATA WITH ADVANCED SEARCH CAPABILITIES

Kathryn Crouch[1], Susanne Warrenfeltz[2]

[1]University of Glasgow, Wellcome Trust Centre for Molecular Parasitology, Glasgow, United Kingdom, [2]University of Georgia, Center for Tropical and Emerging Global Diseases, Athens, GA

The Eukaryotic Pathogen Database (EuPathDB.org) Bioinformatics Resource Center provides online open access to over 170 organisms within Amoebazoa, Apicomplexa, Chromerida, Diplomonadida, Trichomonadida, Kinetoplastida and numerous phyla of oomycetes and fungi. In addition to genomes (>200) and annotation, EuPathDB integrates structured sample and clinical data, and a wide range of functional data types (>500 datasets) encompassing transcript and protein expression, sequence and structural variation, epigenomics, clinical and field isolates, metabolites and metabolic pathways and host-pathogen interactions. This is supplemented by an in-house analysis pipeline that generates data such as domain predictions and orthology profiles for all genomes. Data is analyzed using a standardized workflow system ensuring that in addition to mining specific datasets, comparisons can be made across datasets.

Finding informative patterns in large volumes of diverse data is challenging. EuPathDB offers over 100 configurable searches that interrogate the underlying data, and combined with a unique graphical search strategy system and filter parameter interface, facilitates the discovery of meaningful relationships between diverse data types across organisms. Results of individual searches that return the same type of feature can be combined using set operations (union, intersect, complement) regardless of the data type queried. The nesting tool allows users to control the order in which search results are combined. Results, including those from searches that return different data types, can also be combined by genomic location (e.g., return genes from a previous search whose upstream regions contain SNPs) using the flexible co-location tool. The functionality is further enhanced by the ability to transform gene results by orthology. This feature enables users to make inferences about organisms with limited functional data or incomplete genome annotation based on existing data in closely related organisms. Strategies can be downloaded, saved, shared and re-run at any time. To complete this versatile and powerful data mining resource, EuPathDB integrates search strategies with tools for data visualization, comparative genomics, population genetics and functional enrichment analysis.

EuPathDB is updated bi-monthly with new data, tools or features. Forthcoming additions include a private user workspace for primary data analysis, and better representation of alternatively spliced genes. EuPathDB actively works with its global community of users to ensure that this NIH/NIAID and Wellcome Trust-funded initiative meets their needs. EuPathDB's active user support system includes an email helpdesk, social media presence, a YouTube channel and a worldwide program of workshops.

Authors present on behalf of the EuPathDB team.

EXPLOITING SINGLE CELL EXPRESSION HETEROGENEITY TO CHARACTERIZE CO-EXPRESSION REPLICABILITY

Megan Crow, Anirban Paul, Sara Ballouz, Josh Huang, Jesse Gillis

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY

Co-expression networks have been a useful tool for functional genomics, providing important clues about the cellular and biochemical mechanisms that are active in normal and disease processes. An outstanding question in the interpretation of these data is the relative importance of variable sample composition in real terms (e.g., the proportion of neurons vs glia) versus variability in cell state, for example the engagement of different molecular players throughout the phases of the cell cycle. The increasing availability of single cell RNA-seq data (scRNA-seq) allows us to answer these questions by comparing networks built from specified cell types, which should have only state-dependent co-variance, to networks built from ensembles of different cell types.

We have performed the first major analysis of single-cell co-expression, including a meta-analysis of single cell RNA-seq expression, sampling from 31 individual studies comprising 163 individual celltypes. This allows us to differentiate between the effects of state and compositional variation on functional connectivity, as well as compare single-cell data to that from 163 bulk RNA-seq experiments we used as an external control. From these data, we found that single cell network connectivity is less likely to overlap with known functions than co-expression derived from bulk-data, but this appeared to be due to increased technical noise in single-cell data. Most interestingly, assessing single-cell data in which cell-type was held constant in each network showed little decrease in performance, suggesting that gene sets varying from cell to cell within a cell-type are similar to those that vary from cell-type to cell-type.

To complement this analysis, we performed our own technically controlled scRNA-seq experiment using genetically targeted interneuron classes to determine features and analysis practices that contribute to functional connectivity. Chandelier cells and parvalbumin-positive fast-spiking basket cells were prepared in known batches of 16 cells to generate co-expression networks for each. This allowed us to take the same meta-analytic approach we took to cross-laboratory comparison to characterization of technical properties within our data. We found that the use of raw or batch-corrected UMI data and post-co-expression network standardization provided the highest degree of network replicability, semantic similarity and overall functional connectivity. Importantly, we found that gene expression levels are highly predictive of node degree, and this underlies expression-dependent functional connectivity in a general way.

EFFECT-SPECIFIC ANALYSIS OF PATHOGENIC SNVS IN HUMAN INTERACTOME: INSIGHTS INTO DYNAMIC ORGANIZATION OF THE MOLECULAR NETWORK UNDERLYING COMPLEX DISEASE

Hongzhu Cui[1], Dmitry Korkin[1,2]

[1]Worcester Polytechnic Institute, Bioinformatics and Computational Biology Program, Worcester, MA, [2]Worcester Polytechnic Institute, Department of Computer Science, Worcester, MA

Recent years have seen tremendous advances in Next Generation Sequencing (NGS) and high-throughput omics technology. Powered by the data generated through those technologies, computational tools and genotype–phenotype databases have been developed to bring us one step closer in understanding of the human genetic disease. Genotypic information alone rarely elucidates the mechanistic insights pertaining to disease pathogenesis. One data-driven concept that has recently been proven critical in analyzing complex genetic disorders at the molecular level is a macromolecular network. In this work, we seek to understand how the most common type of genetic variation, single nucleotide variants (SNVs), perturb and rewire a molecular network and how these changes can be linked to the phenotypes. Specifically, we integrate genetic data, protein-protein interaction (PPI) data to study the network-rewiring properties of SNVs in the interactome. We have identified the deleterious SNVs in the network using our recently developed SNP-IN tool, and provided an additional layer of information describing specific loss or gain of interactions (edges) of disease-associated mutations in the context of the PPI network. The annotation results lead us to the analysis of network robustness and efficiency. The topological network analysis suggests that, compared to truncating mutations, SNVs are more likely to be the cause of gene pleiotropy, and both, the network centrality and enrichment on protein interface of SNVs, contribute to gene pleiotropy. Functional annotation results show that a high percentage of pathogenic SNVs could cause deleterious effects in the corresponding PPIs and would play a key role in disease progress, more than previously expected. Different mutations in the same gene leading to different interaction profiles result in distinct disease phenotypes. Moreover, the network robustness analysis suggests that a recently proposed edge-based robustness measure is more suitable to characterize the PPI network rewiring behavior, and SNVs are more efficient in damaging the molecular network, compared to random attacks.

← Back

THE EFFECTS OF DEMOGRAPHIC HISTORY ON THE DETECTION OF RECOMBINATION HOTSPOTS

Amy L Dapper, Bret A Payseur

University of Wisconsin - Madison, Laboratory of Genetics, Madison, WI

In many species, meiotic recombination is concentrated in small genomic regions. These “recombination hotspots” leave signatures in fine-scale patterns of linkage disequilibrium, raising the prospect that the genomic landscape of hotspots can be characterized from sequence variation. This approach has led to inferences that recombination hotspots evolve rapidly in some species, but are conserved in others. Past demographic events, such as population bottlenecks, violate population genetic assumptions of this approach and are known to affect patterns of linkage disequilibrium across the genome. Such events are prevalent, yet demographic history is generally unaccounted for when making inferences about the evolution of recombination hotspots. To determine the effect of demography on the detection of recombination hotspots, we use the coalescent to simulate haplotypes with a known recombination landscape. We measure the ability of popular linkage disequilibrium-based programs to detect recombination hotspots under different demographic histories. We find that past demographic events, and in particular, population bottlenecks and exponential population growth, greatly reduce power to discover recombination hotspots. Additionally, past demographic events have the potential to increase the false positive rate of hotspot discovery. Ignoring demographic history likely overestimates the power to detect recombination hotspots and underestimates the degree to which recombination hotspots are shared between closely related species. We recommend that demographic inference be incorporated into population genetic inferences about recombination hotspots.

FUNCTIONAL ASSAYS FOR *IN VITRO* CHARACTERIZATION OF MULTIPLE MYELOMA CANCERS

Theodorus E de Groot[1], Jiaquan Yu[1], Caitlin A Holien[1], Jay W Warrick[1], Shigeki Miyamoto[2], David J Beebe[1]

[1]University of Wisconsin - Madison, Biomedical Engineering, Madison, WI, [2]University of Wisconsin - Madison, Oncology, Madison, WI

Multiple myeloma (MM) is the second most common bone cancer, it treatable but not curable. The impact of targeted treatment is limited by the widespread genomic heterogeneity as rare, resistant phenotypes emerge from therapy. The effectiveness of treatment is further reduced by the specific microenvironment of the bone marrow inducing a complex network of interactions, providing protection from therapy. The significance the interactions of MM with the microenvironment has been realized in recent years and is the target of several new and developing drugs. Due to the complexity of the microenvironment, current methods for predicting appropriate therapies by either genotyping or in vitro tissue culture are not well suited.

We have developed a suite of functional in vitro models to investigate patient-specific interactions between MM, its microenvironment, and treatments. These models measure several key interactions associated with cancer progression such as adhesion, drug resistance within the context of both the tumor and its microenvironment from a single patient sample. The adhesion model is a microchannel-based assay that is capable of measuring the maximum shear MM cells are capable of adhering to a surface functionalized with proteins found in the MM microenvironment. Unlike conventional adhesion assays, this technique not only measures the specific force at adhesion at a single-cell level, but maintains all cells in their exact positions at the conclusion of the assay allowing cross-referencing of adhesion with single-cell measurements from immunocytochemistry. The first drug resistance assay places MM cells and stromal cells from a patient biopsy in separate compartments of a microchannel while allowing them to maintain soluble factor signaling during treatment with therapies. The separation allows tracking MM and stromal cells throughout the assay while maintaining a similar signaling to in vivo conditions. The second drug resistance assay focuses on how individual patient-derived mesenchymal stem cells (MSC) influence the acquisition of drug resistance on MM cell lines. Since MSCs need to isolated and expanded from a patient biopsy, patient-specific microenvironmental cues are no longer present, so the context of this assay investigates how a stable, bone marrow microenvironment interacts with a "generic" tumor. We utilized this assay with reconfigurable microculture platform where MM cells and patient MSCs can be cultured in isolation, brought together then separated. By combining, treating, then separating we can observe how the microenvironment changes over the course of several treatments. Each of these assays reveals a unique perspective about an individual's cancer that genotyping alone cannot provide.

FUNCTIONAL ANNOTATION OF LONG NON-CODING RNAS IN FANTOM6

Michiel J de Hoon, Jay W Shin, Chung Chau Hon, Masayoshi Itoh, Takeya Kasukawa, Naoto Kondo, Harukazu Suzuki, Piero Carninci

RIKEN, Center for Life Science Technologies, Yokohama, Japan

FANTOM (Functional ANnoTation Of the Mammalian genome) is an international research consortium aiming at a comprehensive identification and annotation of mammalian transcripts. Whereas earlier functional annotation efforts focused on the characterization of protein-coding transcripts, recent FANTOM projects as well as ENCODE have demonstrated that most transcripts in mammalian cells are non-coding. In contrast to proteins, for which an initial functional annotation can be generated based on the amino acid sequence, the function of non-coding RNAs cannot be reliably predicted from the nucleotide sequence alone. The vast majority of non-coding transcripts therefore currently do not have any functional annotation. As the long non-coding RNAs that have been functionally characterized were oftentimes found to play important regulatory roles, we postulate that long non-coding RNAs form a class of regulators that are essential in determining cellular behavior.

Here, we will introduced the sixth edition of the FANTOM project (FANTOM6), in which we aim to systematically elucidate the function of long non-coding RNAs (lncRNAs) in the human genome using high-throughput strategies to perturb hundreds of lncRNAs in multiple cell types, followed by transcriptome profiling using CAGE to assess the molecular phenotype. We complement these perturbation experiments by genome-wide profiles to establish the basal state of the transcriptome and epigenome in each cell type. The lncRNAs we selected for perturbation include published transcripts as well as novel transcripts that are being discovered as part of FANTOM5. The FANTOM6 data generated so far show the distinct response of the human transcriptome to the lncRNA perturbations, consistent with their functional role in cellular regulation.

WHOLE GENOME SEQUENCING AND IMPUTATION FURTHER RESOLVES GENETIC RISK FOR INFLAMMATORY BOWEL DISEASE

Katrina M de Lange, Yang Luo, Loukas Moutsianas, Javier Gutierrez-Achury, Carl A Anderson, Jeffrey C Barrett, UK IBD Genetics Consortium

Wellcome Trust Sanger Institute, Human Genetics, Hinxton, United Kingdom

Over 200 risk loci have been identified for inflammatory bowel disease (IBD), nearly all of which are driven by common variants. However, the contribution of lower frequency variants (MAF<5%) has been difficult to study, as they are poorly tagged by GWAS. Whole genome sequencing can address this, but it is financially and computationally expensive, and large sample sizes will be necessary to detect associations to these variants.

Here we present an analysis of low coverage whole genome sequences from 4445 IBD cases (2-4x depth) and 3652 controls (6x). This approach allows greater sample sizes for a fixed cost, but yields less accurate individual genomes. After quality control, 22.5 million sites were available for association testing, 9 million of which were not seen in the 1000 Genomes project. However, despite the relatively large sample size, no single variant reached genome-wide significance that had not previously been implicated by GWAS. To increase power at sites of rare variation (MAF≤0.5%), we tested for a burden of rare variants in both genes and enhancers, observing enrichment of damaging missense variants in known IBD genes (p<1e-5). The most significant burden is observed in the CD risk gene NOD2 (P=4e-7), and is independent of its known common risk variants.

To better investigate low frequency and common variation, we performed a new GWAS in 18,355 individuals, and imputed variants with MAF>0.5% from our sequenced individuals. These imputed data explain up to 28.4% of variation in liability for IBD, a modest increase on the 26% recently estimated from only common variants. We meta-analyzed with previously published IBD GWAS summary statistics, leading to a total sample size of 60,087 individuals. We identify 31 new common and low frequency associations. Despite being missed by previous meta-analyses, these loci contain three integrin genes that are the targets of existing anti-inflammatory biologic therapies, highlighting that new common associations continue to identify genes and pathways relevant to therapeutic target validation.

We performed one of the largest whole genome sequence-based association studies for a complex disease to date, coupled with a large new GWAS in the same disease. While our study extends the allele frequency spectrum tested for association to IBD risk, the majority of new discoveries still come from common variants of tiny effect. Higher coverage sequencing of tens of thousands of individuals will be needed to fully elucidate the role of truly rare genetic variants in complex disease risk.

GENETIC DETERMINANTS OF GENE EXPRESSION IN A COLLECTION OF 215 HUMAN INDUCED PLURIPOTENT STEM CELLS

Christopher DeBoever, David Jakubosky, Angelo Arias, Agnieszka D'Antonio-Chronowska, He Li, Kelly A Frazer

University of California San Diego, Pediatrics, La Jolla, CA

In this study, we examined the genetic regulation of gene expression in a collection of human induced pluripotent stem cells (iPSCs) that we systematically reprogrammed from the fibroblasts of 215 individuals. We performed transcriptome sequencing for the iPSCs, high-depth (30x) germline whole genome sequencing (WGS), and Hi-C for a subset of iPSCs. We used these data to investigate the suitability of iPSCs for quantitative trait association studies, whether iPSC expression quantitative trait loci (eQTLs) recapitulate known stem cell and gene regulation biology, and the effect of CNVs and rare variants on gene expression. We identified eQTLs for 5,816 genes (eGenes) including markers of pluripotency such as *POU5F1* (*OCT4*), *LCK*, *IDO1*, and *CXCL5*. Transcription factor (TF) binding sites for NANOG, SP1, MXI1, JUND, and other TFs were highly enriched for disruption by lead eQTL variants indicating that common genetic variation affects important reprogramming factors. We evaluated the power of iPSCs for detecting eQTLs relative to GTEx tissues and lymphoblastoid cell lines (LCLs) and found that iPSCs are powered similarly to GTEx tissues and better powered compared to LCLs indicating that the reprogramming process does not diminish the genetic contribution to gene expression. Compared with other tissues in GTEx, iPSC have about twice as many unique eGenes (i.e. not found in other tissues) suggesting that the regulatory landscape and gene expression profile of iPSCs are substantially different than adult human tissues. We identified biallelic and multiallelic copy number variants (CNVs) using the WGS data and found 359 genes with significant CNV associations including 158 genes with CNV lead variants. More than half of the CNV eQTLs are intergenic and based on Hi-C data a subset appear to affect expression levels by interacting with target genes' promoters. We also used H1-hESC DNase hypersensitivity sites (DHSs) and TF ChIP-seq data to identify putative causal eQTL variants and found that those that disrupt TF binding sites are enriched for 3D interactions with gene promoters. In our analysis of rare variants we found that rare promoter SNVs and rare indels that overlap DHSs weakly disrupt gene expression. On the other hand, rare CNVs that overlap genes tend to disrupt gene expression with relatively high effect sizes. We used allele specific expression to investigate the rate of X chromosome reactivation after reprogramming and showed that it occurs at different rates across the X chromosome and is correlated with XIST and TSIX expression. This work demonstrates the utility of iPSCs for genetic association analyses, helps define the role of CNVs and rare variants in the regulation of gene expression, identifies novel TFs as potential key regulators in stem cells, and provides information on the heterogeneity of X reactivation in iPSCs.

← Back

FROM REGULATORY VARIANTS TO GENE EXPRESSION: DISENTANGLING LOCAL REGULATORY NETWORKS

Olivier Delaneau[1], Konstantin Popadin[2], Marianna Zazhytska[2], Sunil Kumar[3], Ambrosini Giovanna[3], Andreas Gschwind[2], Christelle Borel[1], Daniel Marbach[4], David Lamparter[4], Sven Bergmann[4], Philipp Bucher[3], Stylianos Antonarakis[1], Alexandre Reymond[2], Emmanouil Dermitzakis[1]

[1]University of Geneva, Dpt of Genetic Medicine and Development, Geneva, Switzerland, [2]University of Lausanne, CIG, Lausanne, Switzerland, [3]EPFL, Swiss Institute for Experimental Cancer Research, Lausanne, Switzerland, [4]University of Lausanne, Dpt of Computational Biology, Lausanne, Switzerland

Population measurements of gene expression and genetic variation enabled the discovery of thousands of expression Quantitative Trait Loci (eQTLs); a great resource to better understand the genetic component of gene expression variation and to pinpoint genes affected by non-coding GWAS hits. By completing this standard design with in-depth characterization of chromatin activity, we can now go further and dissect the molecular mechanisms underlying eQTLs. To this aim, we quantified genome-wide gene expression (mRNA) and three histone modifications tagging promoter (H3K4me3), enhancer (H3K4me1), and active regions (H3K27ac) in lymphoblastoid cell lines of 184 densely genotyped unrelated individuals. After careful data quality control, we performed a systematic correlation analysis between all the assayed genomic features and characterized the biological properties of the resulting significant associations. First, we discovered more than 20,000 chromatin QTLs (cQTLs) and show that they capture genetic variants affecting phenotypes at multiple levels: molecular (cQTLs likely fall within ENCODE transcription binding sites), cellular (eQTLs are often cQTLs), tissue (cQTLs are differentially enriched across GTEx tissues) and organismal (GWAS hits are enriched for cQTLs). Second, we find extensive correlation between all chromatin marks; a structure that results from multiple levels of modular organization of the chromatin (i.e. a given module may contain sub-modules). We propose to represent this modular organization using tree structures and show how these encapsulate chromosomal interactions (as those mapped with Hi-C) and bring multiple distinct regulatory elements together (e.g. enhancers and promoters) into larger functional units. Finally, we describe an efficient computational approach to phenotype all these functional units and to use the resulting quantifications (1) to map associations with nearby genes and genetic variants and (2) to hypothesize about the directionality of the effects (i.e. causal inference). Overall, this large-scale study that integrates gene expression, chromatin activity and genetic variation provides novel insights into the mechanisms underlying eQTLs and their effects on transcription.

← Back

TRACING THE ORIGIN OF DISSEMINATED TUMOR CELLS IN BREAST CANCER USING SINGLE-CELL SEQUENCING

Elen K Møller*[1,2], Parveen Kumar*[3,4], Jonas Demeulemeester*[5,6], Silje Nord[1,2], David C Wedge[7], April Peterson[8], Randi R Mathiesen[1,9], Renathe Fjelldal[9], Masoud Z Esteki[3], Jason Grundstad[8], Elin Borgen[9], Lars O Baumbusch[1], Anne-Lise Børresen-Dale[1,2], Kevin P White**[8], Bjørn Naume**[2,9], Vessela N Kristensen**[1,2], Peter Van Loo**[5,6], Thierry Voet**[3,4]

[1]Institute for Cancer Research, University Hospital, Oslo, Norway, [2]KG Jebsen Center for Breast Cancer Research, University of Oslo, Oslo, Norway, [3]Reproductive Genomics, University of Leuven, Leuven, Belgium, [4]Single-cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [5]Francis Crick Institute, Cancer Genomics, London, United Kingdom, [6]Human Genome Laboratory, University of Leuven, Leuven, Belgium, [7]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [8]Institute for Genomics & Systems Biology, University of Chicago, Chicago, IL, [9]Departments of Oncology and Pathology, University HospitalOslo, Norway

Background
Single-cell micro-metastases of solid tumors often occur in bone marrow. These disseminated tumor cells (DTCs) may resist therapy and lay dormant or progress to cause overt bone and visceral metastases. Unfortunately, the molecular nature of DTCs remains elusive, as well as when and from where in the tumor they originate. Here, we apply single-cell sequencing to trace the origin of DTCs in breast cancer.

Results
We sequenced the genomes of 40 single cells isolated from the bone marrow of six patients using established markers and morphologic characteristics for epithelial tumor cells. Comparison of the cells' DNA copy number aberration (CNA) landscapes with those of the primary tumors and lymph node metastasis established that a quarter of the cells are DTCs disseminating from the observed tumor. The remaining cells represented non-aberrant 'normal' cells and 'aberrant cells of unknown origin' that have CNA landscapes discordant from the tumor. Genotyping somatic mutations called on bulk tumor exomes in the single-cell sequences confirmed that these cells did not derive from the same lineages as the observed breast cancers. Evolutionary reconstruction analysis of bulk tumor and DTC genomes enabled ordering of CNA events in molecular pseudo-time and tracing the origin of the DTCs to either the main tumor clone, primary tumor subclones, or subclones in an axillary lymph node metastasis.

Conclusions
Single-cell sequencing of bone-marrow epithelial-like cells, in parallel with intra-tumor genetic heterogeneity profiling from bulk DNA, is a powerful approach to identify and study DTCs, yielding insight into metastatic processes. A heterogeneous population of CNA-positive cells of unknown origin is prominent in bone marrow.

*authors contributed equally
**joint senior authors

← Back

GENE.IOBIO - A VISUAL, WEB BASED, REAL-TIME VARIANT ANALYSIS TOOL

Tonya Di Sera[1,2], Chase A Miller[1,2], Yi Qiao[1,2], Alistair Ward[1,2], Gabor Marth[1,2]

[1]University of Utah, Department of Genetics, Salt Lake City, UT, [2]USTAR Center for Genetic Discovery, Salt Lake City, UT

With rapidly increasing access to high-throughput, low cost sequencing, analysis tools must be able to quickly identify known causative variants in patients, discover new causative variants in known disease genes, and find new disease genes. Our tool, gene.iobio, supports multiple approaches to variant analysis by 1) providing a real-time visual interface to intuitively investigate variants; 2) performing real-time variant calling to confirm or reject a variant; 3) working with variants identified from a variant prioritization tool and; 4) increasing the search space to look at genes associated with a phenotype.

Gene.iobio is built on our popular web-based iobio system, an analysis platform governed by two main philosophies: leverage bioinformatics algorithms and software by wrapping them in services that can stream small "slices" of data; and provide interactive visualizations that encourage discovery and investigation.

Gene.iobio analyzes variants at the gene level, gathering annotations (SnpEff and VEP, ExAC and 1000G allele frequencies, ClinVar, inheritance modes) and rendering visualizations in the web browser in a matter of seconds. Variants are shown in a prioritized list, rapidly directing analysts to those most likely to be causative. Variants are visualized along the genomic coordinates, colored by impact and given a shape according to variant type. An area chart shows the sequence coverage across the gene region. Hovering over a variant displays a tooltip with detailed annotations along with allele counts.

Gene.iobio has the ability to reveal variants that may have been missed by variant callers by re-calling variants in real-time with our FreeBayes software, employing parameter sets tuned for high sensitivity. The called variants appear above the loaded variants, allowing side-by-side comparison.

Currently, disease variant identification is carried out by multiple tools that take the approach of filtering the list of variants to a smaller set of likely causative variants. These variant lists can be imported into gene.iobio using the bookmarks feature, providing an intuitive means to investigate each variant in light of its annotations, the transcript set, other variants in the gene, as well as the sequencing coverage.

Analysts can further expand the search space to genes strongly associated with a set of phenotypes. We have integrated Phenolyzer (Hui Yang, USC), a phenotype-to-gene analysis tool that returns a ranked gene list. Genes selected from this list appear in the main navigation panel as gene buttons that contain badges that signify the presence/absence of high-priority variants.

Gene.iobio's real-time, iterative analysis coupled with intuitive data visualizations offer a complete workbench for variant identification and validation.

UNLEASHING THE CANCER GENOMICS CLOUD

Jack DiGiovanna, Brandi N Davis Dusenbery, Zeynep Onder, Devin Locke, Deniz Kural

Seven Bridges Genomics, CGC, Cambridge, MA

Next generation sequencing has both invigorated the genomic data space and caused an explosion of dimensionality. In parallel, electronic medical records are improving our ability to collect detailed and even longitudinal phenotype information. Genomic and phenomic samples are pooled into secure databases approaching petabytes of multidimensional information from thousands of patients. These massive and continually expanding datasets hold the promise to overcome some of the current statistical issues in Precision Medicine. However the analysis, storage, and distribution of this information becomes increasingly challenging with data size. Security is also a paramount concern to ensure protection of patient privacy. To address these concerns, governments and consortium have often ‘silo-ed' data to ensure security, facilitate data maintenance, and provide sufficient computational power. This intrinsically requires expensive local computation resources, which often sit unused, and could restrict collaboration.

Here we showcase how our Cancer Genomics Cloud Pilot, awarded by the National Cancer Institute in late 2014, overcomes these issues. The Cancer Genomics Atlas (TCGA) network has generated an exceptional database of more than 11,000 cases spanning more than 30 different cancers. We have made this 1 PB database both accessible and actionable by leveraging state of the art bioinformatics tools and optimized cloud computation. Researchers can perform fully reproducible biomedical analysis and seamlessly collaborate worldwide.

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.

ENRICHMENT OF IBD FINE MAPPING VARIANTS IN HI-C REGIONS

Julia B Dmitrieva[1], Roman Kreuzhuber[2], Biola-Maria Javierre[3], Ming Fang[1], Elisa Docampo[1], Oliver Stegle[4], Willem Ouwehand[5], Mikhail Spivakov[3], Peter Fraser[3], Michel Georges[1]

[1]University of Liege, GIGA, Liege, Belgium, [2]University of Cambridge, Department of Haematology, Cambridge, United Kingdom, [3]Babraham Institute, Nuclear Dynamics Programme, Cambridge, United Kingdom, [4]European Bioinformatics Institute , Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, [5]Cambridge Biomedical Research Centre, National Institute for Health Research, Cambridge, United Kingdom

The collaborative efforts of different groups studying genetics of inflammatory bowel disease (IBD) have identified 94 IBD regions (spanning from 120kb to 3Mb). Many of the independent variants in these regions are enriched for protein-coding changes, disruption of transcription factor binding sites and tissues specific epigenetic marks. In this study, we are interested whether the IBD fine-mapped variants are also enriched in high-resolution capture Hi-C regions (spanning from 100 to 40kb). We have used Hi-C enhancer-promoter interactome dataset generated for 17 primary human cell tissue types. We expect that, in the case of enrichment, the number of fine-mapped variants co-localized in Hi-C regions as well as the average posterior probability should be significantly higher compared to the control regions. In order to create the control regions we let the Hi-C fragments "circulate" inside the corresponding IBD regions spanning on average 100 times larger genomic distances.

← Back

BRAINCODE: HOW DOES THE HUMAN GENOME FUNCTION IN SPECIFIC BRAIN NEURONS?

Xianjun Dong[1], Zhixiang Liao[1], David Gritsch[1], Boris Guennewig[2], Yavor Hadzhiev[3], Yunfei Bai[1], Ganqiang Liu[1], Cornelis Blauwendraat[4], Charles H Adler[5], Matthew P Frosch[6], Peter T Nelson[7], Patrizia Rizzu[4], Antony A Cooper[2], Peter Heutink[4], Thomas G Beach[8], Ferenc Mueller[3], John S Mattick[2], Clemens R Scherzer[1]

[1]Harvard Medical School and Brigham & Women's Hospital, Neurology, Boston, MA, [2]Garvan Institute of Medical Research, Neuroscience, Sydney, Australia, [3]University of Birmingham, Institute of Cancer and Genomic Sciences, Birmingham, United Kingdom, [4]German Center for Neurodegenerative Diseases, German Center for Neurodegenerative Diseases, Tübingen, Germany, [5]Mayo Clinic, Neurology, Scottsdale, AZ, [6]Massachusetts General Hospital, Pathology, Boston, MA, [7]University of Kentucky, Pathology, Lexington, KY, [8]Banner Sun Health Research Institute, Banner Sun Health Research Institute, Sun City, AZ

The human brain comprises ~86 billion neurons whose function is central to human biology. How does the human genome program high performing neurons and neural networks in response to experience? What subprograms does the genome express in physiologically and morphologically distinct brain cells? The goal of the BRAIN Cell encyclOpeDia of transcribed Elements Consortium (BRAINCODE) is to provide a comprehensive map of actively transcribed elements, both protein-coding and non-coding, from specific cell types, not in culture, but directly isolated from human brains. Going beyond traditional mRNA sequencing, all polyadenylated and non-polyadenylated transcripts over 50bp were ultra deeply sequenced using ribo-depleted total RNA from 50,000 neurons laser-captured from more than 130 human post-mortem brains yielding 23 terabytes of reads. Three prototypical neuron types, dopamine neurons, pyramidal neurons, and Betz cells, were prioritized because of their key biological roles and differential vulnerability to important neurodegenerative diseases such as Parkinson's or Alzheimer's disease. Genetic variation between individuals was examined for correlation with differences in transcribed sequences to identify genomic regions that influence whether, how, and how much a transcript is expressed in specific cell types in human brains. Initial results indicate a vast universe of annotated and novel non- coding RNAs expressed in brain cells and suggest a more diverse and much more complex transcriptional architecture than previously imagined. Support: NINDS U01 NS082157; U.S. Department of Defense; Michael J. Fox Foundation; U24 NS072026; P30 AG19610; P30 AG028383; NHMRC 631668.

← Back

TOWARDS A HIGH RESOLUTION UNDERSTANDING OF THE EVOLUTIONARY FORCES SHAPING THE POPULATION STRUCTURE OF COMMON CHIMPANZEES

Janina Dordel[1], Matthew W Mitchell[1], Peter H Sudmant[2], Mary Katherine Gonder[1]

[1]Drexel University, Biology, Philadelphia, PA, [2]Massachusetts Institute of Technology, Biology, Cambridge, MA

Despite our close evolutionary relationship and our deep curiosity for the common chimpanzee we are only starting to understand the rich genetic diversity within this species.Recent advantages in utilizing genetic markers have revealed that *Pan troglodytes* can be divided into two major geographically and genetically distinct lineages consisting of two subspecies each: a western African group that includes *P. t. verus* and *P. t. ellioti* and a central/eastern African group including *P. t. troglodytes* and *P. t. schweinfurthii*.
The natural division of the two lineages is the Sanaga River in Cameroon where the ranges of *P. t. ellioti* and *P. t. troglodytes* converge. However, we have shown that neutral evolutionary processes resulting from separation across this biogeographic barrier alone cannot explain the differentiation between *P. t. ellioti* and *P. t. troglodytes*. In addition it has been shown that only the central/eastern chimpanzees carry the Simian immunodeficiency virus, the progenitor of human HIV-1, despite ongoing gene flow between *P. t. ellioti* and *P. t. troglodytes*.
In order to identify genes that show signatures of selection and might explain diversification we analyzed 32 whole genome sequences from all four subspecies. High quality single nucleotide polymorphisms (SNPs) were identified and screened to detect outlier SNPs either by analyzing western African and central/eastern African samples or evaluating *P. t. ellioti* and *P. t. troglodytes* samples.
Gene ontology (GO) enrichment analysis was carried out to determine broader characteristics of genes under selection. However, there was no clear signal for enriched gene ontologies.
Interestingly, genome wide investigation of outlier SNPs revealed a strong outlier pattern in the major histocompatibility complex (MHC) in *P. t. ellioti*. MHC is known to play a crucial role in the immune system and to infer resistance against HIV in humans. This suggests that the MHC is under positive selection in *P. t. ellioti* and may play a role in subspecies differentiation.
We present the first extensive set of candidate genes under selection that will help to understand the diversification of two chimpanzee subspecies despite ongoing gene flow. Together with a set of neutral markers this information will be used to screen ~300 genomes from wild and geo-referenced chimpanzees to further understand their diversity and the processes leading to and maintaining it.

← Back

ANALYZING THE INTERPLAY BETWEEN ENHANCERS AND CODING ELEMENTS IN THE TRANSCRIPTIONAL RESPONSE TO CELASTROL

Noah Dukler[1,3], Greg Booth[2], Ed Rice[2], Nate Tippens[2], Charles Danko[2], John Lis[2], Adam Siepel[1]

[1]CSHL, Simons Center, Laurel Hollow, NY, [2]Cornell U., BSCB, Ithaca, NY, [3]Weill Cornell Medical College, Physiology, NY, NY

Non-coding regions of the genome have recently been found to host a wide variety of regulatory elements that affect gene expression but their mechanism(s) and dynamics are still poorly understood. To study the dynamic relationship between enhancers and promoters, we generated time course PRO-seq data after treatment of human K562 cells by the small molecule celastrol. Celastrol is a steroid derivative with therapeutic potential in obesity, cancer and inflammatory disorders. It has been shown to induce a highly complex cellular response, including elements of the heat shock response (HSR) and the unfolded protein response (UPR). PRO-seq measures the 3' end of nascent transcripts, allowing us to simultaneously quantify gene expression and detect the unstable enhancer RNAs (eRNAs) that mark divergent transcription start sites (dTSS) and identify active enhancers. Since PRO-seq measures nascent transcripts, we can immediately detect concordant changes in PolII activity between disparate elements without the delay required by RNA-seq. By fine mapping the architecture of dTSS and computationally associating enhancers and promoters, we look for shared and disjoint sets of regulatory sequence elements. Using PRO-seq we observe transcriptional activity changing within a wide variety of genes as soon as 10 minutes after induction, allowing us to separate first-order and higher order responses to celastrol for the first time. We observe several co-regulated groups of genes immediately respond to celastrol treatment, including those regulating mitochondrial metabolism, the HSF1 response, and translational regulation. Roughly 60 minutes after treatment we observe the upregulation of a number of genes associated with DNA repair, consistent with the induction of a broad cellular stress response. These results allow us to explore the early regulatory underpinning of later responses and phenotypes observed in prior studies. Our results begin to illustrate the complex and multilayered signalling response to celastrol in the regulation of both coding and non-coding elements.

IDENTIFYING BIOLOGICAL CORRELATES OF THE UNDERLYING LIABILITY FOR COMMON COMPLEX DISEASES: TOWARDS NOVEL BIOMARKER SYSTEMS FOR INFLAMMATORY BOWEL DISEASE.

Mahmoud Elansary[1], Ming Fang[1], Alexander M Kurilshikov[2], Joelia Dmitrieva[1], Rob Mariman[1], Theodorus Meuwissen[3], yurii S Aulchenko[4], Michel Georges[1]

[1]Unit of Animal Genomics, GIGA-R, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium, [2]University Medical Center Groningen, Department of Genetics, Groningen, Netherlands, [3]Institute of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway, [4]Institute of Cytology and Genetics SB RAS, Novosibirsk State University, Novosibirsk, Russia

There is a pressing need for the development of predictive biomarker systems that capture the effects of both genetic and environmental risk factors for common complex diseases. We herein propose a novel approach to reach that goal. It uses the predictable inherited component of the underlying liability as a “bait” in deeply omic-phenotyped cohorts of healthy individuals. The idea is to train a liability-based polygenic model in large SNP-genotyped case-control cohorts, and to use it to predict the liability-values of healthy individuals that have been measured for a large number of omic-type phenotypes. Significant correlations are then sought in either univariate or multivariate mode.

We are applying the approach to Inflammatory Bowel Disease (IBD). The case-control cohort that is being used is the cohort of IIBDGC comprising data from 18,967 CD cases, 14,628 UC cases and 34,257 controls. The omics cohort is the CEDAR cohort comprising ~300 individuals that have been SNP genotyped, with transcriptome data on 9 disease-relevant cell types (CD4, CD8, CD14, CD15, CD19, platelets, ileum, colon, rectum), microbiome data at the three corresponding intestinal locations, and plasma metabolome data (in progress).

Latest results will be presented.

EFFECTS OF TRANS-EQTLS ACROSS MANY HUMAN TISSUES

Brian Jo[1], Yuan He[2], Amy He[2], Ian McDowell[3], Alexis J Battle[2], Barbara E Engelhardt[4]

[1]Princeton University, Quantitative and Computational Biology, Princeton, NJ, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]Duke University, Computational Biology and Bioinformatics, Durham, NC, [4]Princeton University, Computer Science, Princeton, NJ

The genetics of gene regulation is essential to understand because of the mechanistic implications for the genetic regulation of complex traits and disease risk: expression quantitative trait loci, or eQTLs, are enriched for polymorphisms that have been found to be associated with disease risk via genome-wide association studies. Local cis-eQTLs have received the bulk of scientific attention as opposed to distal, or trans-, eQTLs. In particular, cis-eQTL SNPs are in close proximity to the genes that they regulate, allowing for many orders of magnitude fewer statistical tests for association mapping; furthermore, in human studies thus far, cis-eQTLs tend to have larger effects than trans-eQTLs. However, recent work has suggested a greater role for trans-eQTLs as compared to cis-eQTLs in complex disease, necessitating a comprehensive understanding of these distal genetic effects in order to characterize the genetic mechanisms of GWAS associations. In this work, we uncover trans-eQTLs within the Genotype-Tissue Expression (GTEx) v7 study data, consisting of over 400 individuals with RNA-sequencing samples across 44 tissue types. First, we identify trans-eQTLs using statistical approaches that share strength across multiple tissues. Then, we consider the role of gene co-expression networks in the discovery and characterization of trans-eQTLs, evaluating the simplest hypothesis that the primary mechanism of trans-eQTL SNPs is that they first regulate expression of a proximal gene, which itself participates in pleiotropic regulation of one or more distal genes. We quantify this statement in the context of gene interaction networks and our collection of trans-eQTLs. Finally, we use Mendelian randomization methods to identify trans-eQTLs regulated genes that affect complex traits in a tissue-specific way. These analyses provide a comprehensive estimate of the effects of trans-eQTLs on gene expression in diverse human tissues, which contributes to an improved understanding of the tissue-dependent cellular consequences of disease-associated genetic variation.

← Back

DNA.LAND: A COMMUNITY-WIDE PLATFORM TO COLLECT GENOMES AND PHENOMES OF MILLIONS OF PEOPLE

Assaf Gordon[1], Jie Yuan[1,2], Dina Zielinski[1], Tris Hayeck[1], Joe Pickrell[1,3], Yaniv Erlich[1,2]
[1]New York Genome Center, New York, NY, [2]Columbia Univresity, Department of Computer Science, New York, NY, [3]Columbia University, Department of Biological Sciences, New York, NY

Precision medicine is a data-hungry endeavor. However, traditional cohort ascertainment strategies poorly scale and necessitate substantial investments to obtain genomics data, conduct physical exams and lab tests, and assess familial history. But are these really required in today's world? In the last decade, the human population has produced zettabytes (1021) of digital data.

Here, we will present our successes in repurposing participants' data for ultra-large scale genetic studies. In our previous studies, we built a 13-million member family tree by crowdsourcing information from the same vibrant citizen genealogy community. Building on this work, we developed a web-platform called DNA.Land (https://DNA.Land) where anyone can securely contribute her or his own DTC-generated genome data for research and connect phenotypes using questionnaires and his/her streams of social media information. A critical concept of DNA.Land is reciprocation. To serve participants' curiosity in their genomes and family histories, our platform is built to efficiently offer analyses unavailable through DTC companies, including whole-genome imputation, refined ancestry inference, and kin-matching across company cohorts. We have been working closely, trustworthily, and fruitfully with participants, to apply the platform for scientific benefit. During its four months of operations, DNA.Land collected over 12,000 genomes.

Taken together, our approach highlights the power of repurposing pre-collected digital data from research participants. We will discuss the lessoned learned from our crowd sourcing efforts, how to leverage social media to collect phenotypes and family trees, and how other efforts could benefit form our platform.Our vision is that this platform will serve the human genetics-wide community to reach the massive scale of data needed to understand complex traits.

← Back

SPURIOUS MUTATION DUE TO DNA DAMAGE IS PERVASIVE AND CONFOUNDS ACCURATE DETECTION OF LOW FREQUENCY MUTATIONS IN HUMAN GENOME.

Lixin Chen[1], Pingfang Liu[2], Thomas C Evans[1], Laurence M Ettwiller[1]

[1]New England Biolabs, Research, Ipswich, MA, [2]New England Biolabs, Application Development, Ipswich, MA

A growing body of evidence suggests that each cell in our body contains genomic sequences unique or shared with only a few other cells. Somatic mutations are pervasive, alter the genome one cell at a time, and may result in serious medical conditions. Cancer is the most studied disease caused by somatic mutations, yet recent advances in sequencing technologies, notably single cell and ultra deep sequencing, indicate many additional phenotypes and pathologies impacted by cell specific mutations.

Nevertheless, current technologies remain very limited in detecting rare and low frequency somatic mutations, a fact that considerably restrains application of NGS to a broad range of research and clinical questions. To this end, tremendous efforts are being invested in further pushing detection limits and improving sequencing fidelity. However, the effect of DNA damage on sequencing errors has been understudied except in specialized samples such as FFPE and ancient DNA.

Here, we report that damage introduced in-vitro is common in essentially all human DNA samples analyzed. Furthermore, the damage spectrum correlates with procedures used for DNA storage and handling during library preparations and confounds determination of the real mutation spectrum found in cancers. A detailed analysis of two commonly used population resources, the 1000 Genomes Project and the TCGA project, reveals that reported sequencing reads have an excess of G to T transversions, a signature of 8-oxo guanine damage, as well as further spurious errors due to damage. More importantly, we estimate that the majority of G to T transversions found in sequencing reads are due to damage for 80 % of the TCGA samples, a finding that may help to precise clinical decision making for several cancers. We experimentally confirmed the in silico predicted damage and identified key steps during DNA handling and library preparation that allow to minimize damage, leading to improved detection of rare and low frequency mutations.

Beyond improving sequencing fidelity, our study highlights the underappreciated role of DNA damage introduced during NGS sample preparation, and resulting impact on deposited genome sequences.

A NETWORK-BASED APPROACH TO EQTL INTERPRETATION AND SNP FUNCTIONAL CHARACTERIZATION

Maud Fagny[1,2], John Platig[1,2], Joseph N Paulson[1,2], John Quackenbush[1,2]

[1]Harvard TH Chan School of Public Health, Department of Biostatistics, Boston, MA, [2]Dana-Farber Cancer Institute, Department of Biostatistics and Computational Biology, Boston, MA

Expression quantitative trait loci (eQTLs) analysis is commonly used to identify genetic variants affecting gene regulation. While such studies provide insight into genetic associations, they generally consider only *cis*-acting SNPs, and fail to address tissue-specific effects that might influence eQTL associations. We applied a meta-analysis approach that jointly analyzes *cis*- and *trans*-eQTLs, representing the associated SNPs and genes as elements of a bipartite graph, and explored its structure, using data from twelve tissue types from the Genotype-Tissue Expression (GTEx) project V4.

We identified hundreds of thousands significant eQTLs in each tissue (FDR = 10%), among which 12 to 42% were *trans*-eQTLs. We represented the eQTLs as bipartite graphs and discovered that SNPs and genes organize into dense, highly modular communities. In each tissue, we found that while ~25% of communities are shared with at least half of the other tissues; ~30% of the communities are tissue-specific.

These communities tend to group genes by biological function. Tissue-specific communities are enriched for relevant processes, such as cortex development and synaptic transmission in brain cortex, and epithelium development and cell junction organization in esophagus mucosa. The genes within shared communities are enriched for general functions, such as RNA biosynthesis, and also for antigen processing and presentation via MHC class II, possibly reflecting systemic infiltration of immune cells.

Examining the communities, we identified local, community-specific network hubs. These “core SNPs” are preferentially located in functional regions of the genome, such as DNase hypersensitive regions with active transcription factor binding sites, suggesting that *cis*- and *trans*-eQTLs might influence groups of genes by disrupting transcription.

When we mapped SNPs associated with disease in the NHGRI GWAS catalog to the networks, we found many communities to be enriched in SNPs associated with particular phenotypes or diseases. This suggests disease SNPs may act together to perturb groups of functionally-associated genes, leading to alterations in the phenotype.

Overall, we find that this network-based approach to characterizing eQTLs provides new insight into the link between genotype and phenotype, identifying groups of SNPs associated with the expression of groups of functionally related genes, and leading to new hypotheses about how the large number of weak-effect SNPs identified through GWAS may work to alter function and phenotype.

← Back

THE INTERNATIONAL GENOME SAMPLE RESOURCE (IGSR): SUPPORTING AND BUILDING ON THE 1000 GENOMES PROJECT DATA

Susan Fairley, Holly Zheng-Bradley, Avik Datta, Peter Harrison, Ernesto Lowy, Ian Streeter, David Richardson, Laura Clarke, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

The 1000 Genomes Project created the largest public catalogue of human variation data. Although the project ended in 2015, these data continue to be highly used by the community. We have established the International Genome Sample Resource (IGSR) to update and expand the dataset with more populations and assays and to ensure that it remains a valuable resource for human variation research.

We realigned the 1000 Genomes Project data to GRCh38, the current reference assembly, with a new alignment pipeline that uses alt-aware BWA. Further open data sets have been aligned in this manner including the Illumina Platinum Genomes pedigree set. These data are distributed in the CRAM file format from the 1000 Genomes FTP site. We are currently identifying additional relevant data, such as RNA-seq, from the 1000 Genomes samples to create a single source for the various assays associated with the samples.

IGSR is also part of the Human Genome Structural Variation (HGSV) Consortium, which continues the work of the 1000 Genomes SV group. The HGSV Consortium is assaying three trios from the 1000 Genomes collection using a wide range of technologies both to produce a high quality structural variation map of the human genome and to advance SV detection methodology. The HGSV data are made available under similar terms to those used by the 1000 Genomes Project.

To support these new data sets, the FTP site used by the 1000 Genomes Project has been restructured. It retains key data in the same location while accommodating data originating from many projects. In addition, the 1000 Genomes website has been updated to reflect the conclusion of the project and the role of IGSR.

While existing, high quality, open data are being added to the IGSR data collection, it remains the case that many populations are not represented. Thus, these populations, and the variation they contain, are not represented in the open data sets frequently used as reference panels and for methods development. To continue to build a fully global resource for human variation research, we are seeking further collaborations in which the IGSR would support the analysis and distribution of openly consented samples.

IMPLEMENTING A POPULATION-CENTRIC REFERENCE GENOME TO FACILITATE PRECISION MEDICINE IN QATAR AND THE MIDDLE EAST

Khalid A Fakhro[1,2], Michelle Staudt[3], Amal Robay[2], Jason Mezey[3], Ronald Crystal[2,3], Juan Rodriguez-Flores[3]

[1]Sidra Medical and Research Center, Translational Medicine, Doha, Qatar, [2]Weill Cornell Medicine in Qatar, Genetic Medicine, Doha, Qatar, [3]Weill Cornell Medicine, Genetic Medicine, New York, NY

Precision medicine will depend on the quality and speed of genome interpretation, tailored to individuals in the context of their native populations. In order to facilitate precision medicine in the Middle East, we developed a population-specific reference for Qatari Arabs (QTR1) by incorporating allele frequency data from sequencing of 1,158 Qataris, (~0.5% of the population) into the standard reference (GRCh37). The new QTR1 reference was tested by alignment of an independent Qatari individual sequenced on four common platforms to both it and GRCh37, which showed a reduction in genome-wide genotype calling errors when employing QTR1. Further, aligning reads from an "n+1" individual from each of the 3 Qatari genetic subpopulations to QTR1 reduced the number of "variants" (identified as the minor allele) in the exome by 14% and genome by 23% compared with using GRCh37. This is a significant reduction in the number of sites that would usually be considered in downstream analysis and require time and resources to annotate and interpret. For example, when considering 2,330 variant alleles previously linked to disease or pharmacogenetics in public databases, 295 (12.6%) have the GRCh37 minor allele as the major alleles (>50% frequency) in Qataris and are therefore unlikely to be deleterious in this population. These would be incorporated into QTR1, and would lead to a significant reduction in number of variants per individual when a sample is aligned to this population-specific reference. This would lead to more efficient, scalable storage and analysis as thousands of Qatari samples get sequenced in the future. Conversely, we manually curated all mutations detected in the 1,158 Qataris in known OMIM genes, and found an elevated burden for several severe diseases, especially relevant in the setting of consanguinity. Specifically, 26 of these had deleterious variant allele frequencies higher in Qatar than any other global population, e.g. homocystinuria, cystic fibrosis and arterial tortuosity. Importantly for personalized medicine in this population, only a minority of all discovered mutations presented here are currently on the national premarital and newborn screening panels. The QTR1 reference sequence and analysis software developed for personalized genome interpretation in this population would serve as useful tools for precision medicine in Qatar and, by extension, genetically related Arab populations.

SCIKIT-RIBO: ACCURATE A-SITE PREDICTION AND ROBUST MODELING OF TRANSLATION CONTROL FROM RIBOSEQ AND RNASEQ DATA

Han Fang[1,2], Yifei Huang[1], Max Doerfel[1], Gholson Lyon[1], Michael Schatz[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Stony Brook University, Applied Math and Statistics, Stony Brook, NY

Ribosome profiling (Riboseq) is a powerful technique for monitoring protein translation in vivo, analogous to RNAseq for expression profiling. However, there are very few methods available to analyze Riboseq data. Here, we present scikit-ribo, a statistical learning framework for joint analysis of Riboseq and RNAseq data. We provide modules for ribosome A-site prediction, ribosome pausing site calling, joint inference of protein translation efficiency (TE) and codon elongation rate.

Based on reads occupying start codons, we used a SVM classifier to learn key features of the A-site location within Riboseq reads. We observed strong dependencies of A-site location on both read length and codon offset (p-value$<2\times10^{-16}$), mostly due to variable effects of the digestion enzyme on individual ribosomes. After improving the A-site resolution to 3bp, we built a negative binomial mixture model to identify and analyze ribosome pausing sites. From this we discovered the commonly used RPKM-based TE calculation is very sensitive to ribosome pausing events, thus negatively skewing the TE distributions in almost all previous studies and limiting their ability to differentiate translation efficiency and codon optimality. To solve this, we built a generalized linear model to simultaneously infer protein TE and codon elongation rates, while accounting for mRNA abundance and secondary structure.

To demonstrate its effectiveness, we used scikit-ribo to analyze data from wild-type and knockout yeast strains involving the N-terminal acetyltransferase Naa10, which is bound to the ribosome as part of the NatA complex. We show our prediction method has much higher accuracy for identifying A-site location than previous methods (0.86 vs. 0.64, 10-fold CV). Scikit-ribo's predicted genome-wide codon usage fraction also has a significant correlation with published estimates (ρ=0.90, p-value$<2\times10^{-16}$). We also successfully identified nearly 100 genes with over 100 ribosome pausing sites in wild-type yeast. Subsequently we discovered mRNA with stronger secondary structure tend to have pausing ribosomes (p-value$<2\times10^{-16}$). Both Riboseq and RNAseq data showed that the mutant has significant reduction in expression of conjugation/mating genes (BH adjusted p-value: 4×10^{-13}). Interestingly, signals of transcriptional control were further amplified at the translational level, showing a significant reduction of TE (BH adjusted p-value: 5×10^{-6}). Together, these results show that scikit-ribo provides robust methods for Riboseq analysis and better understanding of translational control.

← Back

RUFUS: ACCURATE AND SENSITIVE REFERENCE FREE VARIANT DETECTION

Andrew Farrell, Gabor T Marth

USTAR Center for Genetic Discovery, Salt Lake City, UT

Our K-mer based variant detection method, RUFUS, shows higher sensitivity as well as far improved specificity over mapping based approaches enabling extremely accurate de novo variant detection as well as somatic mutation detection. K-mer comparison allows rapid and unbiased variant discovery. Rapid because the analysis involves only a tiny fraction of the data harboring genetic variants, the majority of the data representing sequence shared between the genomes is disregarded. Unbiased, because analysis does not require mapping to a reference sequence but instead can be carried out in a completely reference-free fashion, avoiding the many mapping biases that prevent the detection of genetic variations in highly diverged genomic regions.

RUFUS is ideally suited to experiments where multiple, closely related, genomes can be directly compared: de novo discovery in human trios and quartets, and the detection of somatic mutations in tumor samples in comparison with normal control tissue. De novo variant detection in family trios showcases the extremely high specificity of variant detection with RUFUS. Previous research has suggested that the rate of de novo events in human populations at ~2.5 x 10-8, or roughly 75 mutations per generation. In our analysis of numerous disease family trio data sets at the University of Utah, we see between 77 and 116 genome wide de novo mutation events per child, and on average, 93% of the calls agree with mapping based call, leaving between 2 and 6 novel calls per family in RUFUS which are either false positives or novel variants that only RUFUS can detect, confirmations pending. Conversely, mapping based methods on average have 150,000 unique genome wide de novo calls per child, dominated by mapping and reference errors, which drown out any true variation, making RUFUS a far superior choice for de novo detection.

Application to the detection of somatic mutation in tumor tissue samples has supported the specificity seen in our de novo work and further showcases RUFUS' unique, unbiased, ability to detect mutations of all types and sizes. Of particularly interest are insertion deletion events between 50bp and 500bp, lying in the "blind" spot for short-variant detection methods (e.g. GATK, FreeBayes) and structural/CNV detectors (LUMPY, WHAM, etc.). In this data set RUFUS calls 98 unique calls, of these 95 appear to be true variants missed by mapping methods (freebayes and lumpy). Of the 95 RUFUS-only calls, 84 are indels between 20 and 200bp that no other method is able to detect. This makes RUFUS an ideal method for filling the current hole of medium-length de novo INDEL detection, both in family and tumor sequencing datasets.

NCBI'S VERTEBRATE REFSEQ PROJECT: ACCESSIBILITY, CURATION AND COLLABORATION

Catherine M Farrell, Terence D Murphy, Kim D Pruitt, RefSeq Curation and Development Teams

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD

The Reference Sequence (RefSeq) database at NCBI (http://www.ncbi.nlm.nih.gov/refseq/) represents a collection of transcript, protein and genomic sequences, which provide a stable reference for genome annotation, gene identification, variant calling, and other studies of biomedical relevance. RefSeqs are represented for a range of organisms, including bacteria, viruses and eukaryotes. For eukaryotes, an NCBI genome annotation pipeline uses both known and model RefSeqs to annotate reference genomes. More than 280 eukaryotic organisms have been annotated to date, including >180 vertebrates. RefSeqs and NCBI genome annotation can be accessed by FTP download from several NCBI databases, including the Genome, Assembly, Gene and RefSeq resources, and through database records and various NCBI genome browsers. RefSeq data is also provided by non-NCBI resources, though the full dataset is best accessed via NCBI resources. Each RefSeq is associated with a record in NCBI's Gene database (http://www.ncbi.nlm.nih.gov/gene/), where Gene records include various metadata such as RefSeq summaries, associated publications, INSDC accessions, and links to related public databases. RefSeqs are maintained through both automatic processing and curation, where the curation focus is mostly on higher vertebrate species for which a high quality genome assembly exists. While curation has historically been dependent on available transcript, protein, publication or homology evidence, improved large-scale datasets that have recently become available are now being incorporated in RefSeq curation. These datasets include transcriptomic data with improved exon combination evidence, such as PacBio or SLR-RNA-seq data, as well as datasets that reveal transcript end completeness, including promoter-associated epigenomic data and CAGE data supporting transcript 5' ends, and polyA-seq data supporting transcript 3' ends. Maintenance of the RefSeq database also involves extensive collaboration, where both automatic processing and curatorial methods are used to incorporate information from external sources. Collaborating groups include nomenclature authorities for various organisms, the Genome Reference Consortium, the Consensus Coding Sequence collaboration, the RefSeqGene/Locus Reference Genomic collaboration, and numerous other groups. Taken together, these efforts ensure that our data is maintained as high quality to meet scientific community standards.

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

EFFECTS OF POST-MORTEM INTERVAL ON GENE EXPRESSION ACROSS SEVERAL TISSUES

Pedro G Ferreira[1], François Aguet[2], Ayellet V Segrè[2], Reza Sodaei[3], Dmitry Pervouchine[3], Ferran Reverter[3], Roderic Guigó[3], Kristin Ardlie[2], The GTEx Consortium[4]

[1]I3S/IPATIMUP, Cancer Programme, Porto, Portugal, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Center for Genomic Regulation, Bioinformatics and Genomics, Barcelona, Spain, [4]NIH, Common Fund, Bethesda, MD

Post-mortem tissues are a valuable resource for the study of gene expression in many different diseases and disorders. Due to the unavoidable delay in the collection of tissues, post-mortem samples cannot completely escape the effects of ischemia. While DNA is known to be stable over long ischemic periods, RNA is believed to be more unstable, but this stability may vary from tissue to tissue. A systematic genome-wide tissue specific effect has not yet been fully investigated. The Genotype to Tissue Expression (GTEx) project provides a unique opportunity to study samples obtained from post-mortem donors across a wide-range of body sites.

The main goal of this study is to use the collection of GTEx samples to investigate in detail the effect of post-mortem interval (PMI) on the transcriptome state and in particular on the stability of the full range of expressed mRNA transcript levels across a large set of human tissues.

Across all samples we found a relatively weak association of PMI and RNA integrity values (RIN) ($r=-0.27$). When breaking down by tissue, esophagus or liver show the highest correlation while skin and pituitary have almost no correlation. Based on a linear model several sample and donor features are taken into account to control for the impact of PMI on gene expression. Correlation analysis shows that for the majority of the genes their expression is not associated with PMI and that only a minority exceeds an $|r| > 0.3$. Skin is one of the least affected tissues. Colon, esophagus and heart, which had a stronger association of RIN and PMI, show a small impact on the expression profiles when other co-variables are properly taken into account. Blood has the largest number of outlier genes; The majority of tissues show a symmetric pattern of correlation values with a weak sharing of highly correlated genes among the different tissues. Enrichment analysis revealed a relatively weak and sparse functional enrichment. We also compared pre and post-mortem blood samples, which revealed a significant number of differentially expressed genes. Comparison with genes correlated with PMI shows that changes induced by death are largely different from those occurring with an evolving PMI. Overall, these results seem to indicate that when correctly accounting for the effect of the different covariates, the eventual distortion on gene expression caused by PMI can be minimized across the different tissues.

← Back

THE VAAST VARIANT PRIORITIZER (VVP): RAPID, MASSIVELY SCALABLE WHOLE GENOME VARIANT PRIORITIZATION TOOL AND ITS USE TO PRIORITIZE AND ANALYZE THE ENTIRE CONTENTS OF dbSNP

Steven Flygare[1], Lon Phan[5], Man Li[1,3,4], Barry Moore[1,2], Anthony Fejes[7], Hao Hu[6], Chad Huff[6], Lynn Jorde[1,2], Martin Reese[7], Mark Yandell[1,2]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]USTAR Center for Genetic Discovery, None, Salt Lake City, UT, [3]Johns Hopkins University, Department of Epidemiology, Baltimore, MD, [4]University of Utah, Division of Nephrology, Salt Lake City, UT, [5]NCBI / NLM / NIH, None, Bethesda, MD, [6]M.D. Anderson Cancer Center, Department of Epidemiology, Houston, TX, [7]Omicia Inc, None, Oakland, CA

Prioritizing variants within tomorrow's millions of WGS sequences will be a challenging problem. In response, we have developed VVP, which scores every variant in a genome no matter where it lies — coding, noncoding, SNV, indel, etc. VVP scores integrate sequence conservation, genetic consequence, and allele frequency. VVP scores are designed to facilitate and speed both large-scale analyses and precision diagnostic efforts, as they take into account the fact that some genes exhibit more variation than others. For example, variants in TTN are observed more frequently in the population compared to BRCA2, and as a result TTN variants tend to have lower VVP scores than BRCA2 variants. VVP can also prioritize variants according to genotype. For example, the recessive cystic fibrosis variant F508del has a heterozygous score of 12 (minimal impact), whereas the homozygous VVP score is 97 (highly damaging). Finally, VVP is very fast: a 60X WGS VCF file can be processed in <5 minutes using a laptop. To demonstrate VVP's utility we have collaborated with the NCBI to annotate the contents of dbSNP, some 120 million variants. These data will be made public in the forthcoming dbSNP release. Also presented are systematic comparisons of dbSNP VVP scores to variants with clinical classifications in ClinVar and other clinical databases. Our results reveal much about the landscape of variation as well as the pitfalls of variant prioritization and cross-database comparisons. VVP is part of the VAAST toolkit supported by NIGMS R01GM104390, and is free for academic use, see www.yandell-lab.org.

← Back

STREAMLINED AND SENSITIVE GENE-EXPRESSION PROFILING OF DEGRADED SAMPLES WITH SMART-3SEQ

Joseph W Foley[1,2], Philippe Jolivet[1], Jonathan C Dudley[2], Joanna Przybyl[2,3], Shirley X Zhu[2], Sushama Varma[2], Michael J Meaney[1], Robert B West[2]

[1]McGill University, Ludmer Center for Neuroinformatics and Mental Health, Montreal, Canada, [2]Stanford University, Pathology, Stanford, CA, [3]Maria Sklodowska-Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland

RNA sequencing (RNA-seq) has emerged as the most accurate and powerful method for genome-wide expression profiling, displacing cDNA microarrays. Recent improvements have extended the sensitivity of this method to as little input material as a single cell.

We present a new method, Smart-3SEQ, that greatly reduces the cost and working time of the library preparation protocol while achieving single-cell sensitivity. This new protocol uses only standard laboratory equipment and off-the-shelf reagents, and requires only a single enzymatic reaction step to produce an amplification-ready sequencing library from total RNA or intact cells. Smart-3SEQ is a substantial improvement to the previous 3SEQ method, which measures digital gene expression by capturing only one fragment from the 3′ end of each transcript and has been adopted by numerous laboratories. This allows accurate molecule counting without technical noise due to variable transcript lengths. Smart-3SEQ also incorporates unique molecular identifiers, which further improve its quantitative accuracy by enabling the removal of technical noise from library amplification. A particular advantage of Smart-3SEQ compared with other low-input RNA-seq methods is that it can process highly degraded RNA. Therefore, one of the potential applications of Smart-3SEQ include sequencing of material extracted from laser-microdissected cells from formalin-fixed, paraffin-embedded (FFPE) tissue sections from microscope slides.

As the throughput and cost-effectiveness of massively parallel sequencing have rapidly dropped, the bottleneck for functional genomics experiments is now the preparation of sequencing libraries from challenging biological samples. By streamlining these steps, we anticipate that Smart-3SEQ will open up the possibility of large-scale studies that examine high sample sizes without insupportable expense, particularly in fields such as cancer biology where the use of poor quality, archival FFPE samples is often required. The unprecedented sensitivity of Smart-3SEQ may also enable studies of tissue organization by profiling many rare and distinct cell populations collected in situ from a single sample.

GENOME-WIDE GENERALIZED ADDITIVE MODELS

Georg Stricker[1], Alexander Engelhardt[2], Matthias Schmid[3], Achim Tresch[4], Julien Gagneur[1]

[1]Technical University Munich, Computer Science, Munich, Germany, [2]Ludwig-Maximilians Universitaet, IBE, Munich, Germany, [3]University of Bonn, IMBIE, Bonn, Germany, [4]Max Planck Institute for Plant Breeding Research , Cologne, Germany

Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) is a widely used approach to study protein-DNA interactions. To analyze ChIP-seq data, practitioners are required to combine tools based on different statistical assumptions and dedicated to specific applications such as calling protein occupancy peaks or testing for differential occupancies. Here, we present **GenoGAM (Genome-wide Generalized Additive Model)**, which brings the well-established and flexible generalized additive models framework to genomic applications using a data parallelism strategy. We model ChIP-seq read count frequencies as products of smooth functions along chromosomes. Smoothing parameters are estimated from the data **eliminating ad-hoc binning and windowing** needed by current approaches. We derived a peak caller based on GenoGAM with performance matching state-of-the-art methods. Moreover, GenoGAM provides **pointwise and region-wise significance testing for differential occupancy with controlled type I error rate and increased sensitivity** over existing methods. By analyzing a set of DNA methylation data, we further demonstrate the potential of GenoGAM as a generic analysis tool for genome-wide assays.

← Back

A GENOME-WIDE LANDSCAPE OF RETROCOPIES IN PRIMATE GENOMES

Fábio C Navarro, Pedro A Galante

Hospital Sirio-Libanes, Centro de Oncologia Molecular, Sao Paulo, Brazil

Background: Gene duplication is a key factor contributing to phenotype diversity across and within species. Although the availability of complete genomes has led to the extensive study of genomic duplications, the dynamics and variability of gene duplications mediated by retrotransposition are not well understood. **Results**: Here, we predict mRNA retrotransposition and use comparative genomics to investigate their origin and variability across primates. Analyzing seven anthropoid primate genomes, we found a similar number of mRNA retrotranspositions (~7,500 retrocopies) in Catarrhini (Old Word Monkeys, including humans), but a surprising large number of retrocopies (~10,000) in Platyrrhini (New World Monkeys), which may be a by-product of higher L1 activity in these genomes. By inferring retrocopy orthology, we dated most of the primate retrocopy origins, and estimated a decrease in the fixation rate in recent primate history, implying a smaller number of species-specific retrocopies. Moreover, using RNA-Seq data, we identified ~3,600 expressed retrocopies. As expected, most of these retrocopies are located near or within known genes, present tissue-specific and even species-specific expression patterns, and no expression correlation to their parental genes. **Conclusion**: Taken together, our results provide further evidence that mRNA retrotransposition is an active mechanism in primate evolution and suggest that retrocopies may not only introduce great genetic variability between lineages but also create a large reservoir of potentially functional new genomic loci in primate genomes.
Financial support: FAPESP

THE POPULATION GENETICS OF HUMAN DISEASE: THE CASE OF RECESSIVE LETHAL MUTATIONS

C. Eduardo G Amorim[1], Ziyue Gao[2], Zachary T Baker[1], José F Diesel[3], Joseph Pickrell*[1,4], Molly Przeworski*[1]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]Stanford University, Howard Hughes Medical Institution, Stanford, CA, [3]Universidade Federal de Santa Maria, Departamento de Biologia, Santa Maria, Brazil, [4]New York Genome Center, New York, NY

*co-supervised this project

What determines the frequencies of disease mutations? Conversely, how can we use population genetic data, such as mutation frequencies, to help identify clinically important variants? We consider these questions by focusing on one of the simplest cases: mutations that cause Mendelian, fully recessive diseases. To this end, we rely on long-standing models of mutation-selection balance in finite populations, allowing for the possibility of compound heterozygosity and complementation. We then test how well these models fit genetic variation data for 60,706 individuals made available by the Exome Aggregation Consortium (ExAC). Specifically, we focus on a hand-curated set of 409 mutations in 32 genes that were reported to cause Mendelian, lethal diseases with complete penetrance. Compared to analytic results and simulations, we find that while observed frequencies for CpG transitions are close to expectation, the frequencies observed for transversions and non-CpG transitions in ExAC are an order of magnitude higher than expected under both a constant population size model and a plausible demographic model for human populations. In principle, the discrepancy could be due to errors in reporting causal variants, compensation by other mutations, cryptic heterogeneity in the mutation rates or balancing selection. We argue instead that it likely reflects an ascertainment bias: of all the variants that cause recessive lethal diseases, those that by chance have reached higher frequency are more likely to have been identified. Beyond the specific application, this study highlights some of the challenges in relating allele frequencies to functional effects, even in relatively simple cases.

QUANTIFYING THE EPIGENETIC FLEXIBILITY OF INDIVIDUAL LOCI ACROSS DEVELOPMENTAL TIME

Minseung Choi[1,4], Diane P Genereux[12,1], Jamie J Goodson[2], Haneen Al-Azzawi[6], Shannon Q Allain[1], Stan Palasek[7], Carol B Ware[8], Chris Cavanaugh[8], Daniel G Miller[9], Winslow C Johnson[10], Kevin D Sinclair[6], Reinhard Stoger[6], Charles D Laird[1,11]

[1]University of Washington, Biology, Seattle, WA, [2]University of Washington, Pathology, Seattle, WA, [3]University of Nottingham, School of Biosciences, Nottingham, United Kingdom, [4]Princeton University, Lewis Sigler Institute for Integrative Genomics and Department of Computer Science, Princeton, NJ, [5]University of Washignton, Pathology, Seattle, WA, [6]University of Nottingham, School of Biosciences, Nottingham, United Kingdom, [7]Princeton University, Department of Mathematics, Princeton, NJ, [8]University of Washington, Institute for Stem Cell and Regenerative Medicine, Seattle, WA, [9]University of Washington, Department of Comparative Medicine, Seattle, WA, [10]University of Pittsburgh, Department of Biological Sciences,, Pittsburgh, PA, [11]University of Genome Sciences, Department of Genome Sciences, Seattle, WA, [12]University of Massachusetts Medical School, Bioinformatics and Integrative Biology, Worcester, MA

The mammalian epigenome is responsive to a diverse array of environmental factors, including cigarette smoke, famine, ethanol, and bisphenol-A. Intriguingly, the developmental time of maximal sensitivity to these exposures seems to differ among loci. We introduce a new metric, the Ratio of Concordance Preference (RCP), to quantify this epigenetic flexibility at single-locus resolution across developmental time. RCP quantifies the relative contributions from conservative processes, which yield methylation states that are stable across cell division, and random processes, which yield states that are comparatively prone to transition, in determining the methylation state of a given locus at a given developmental time point. We apply RCP to barcode-validated, double-stranded DNA methylation patterns from both single loci and a large fraction of CpGs genome-wide in human and murine cells. We find that, for the genome overall, cellular differentiation is characterized by increasing contributions from conservative, and that, while concordance preference remains substantial through embryonic totipotency and early stages of pluripotency, primordial germ cells initially have strong contributions from random processes. RCP will be useful in identifying loci that retain flexibility -- and therefore the potential for environmental sensitivity --- into adolescence and beyond, and in guiding efforts to understand the genetic basis and evolutionary origins of the disparate developmental timing of these sensitivities.

← Back

AN INTEGRATIVE FRAMEWORK FOR LARGE-SCALE ANALYSIS OF RECURRENT VARIANTS IN NONCODING ANNOTATIONS

Jing Zhang[1,2], Lucas Lochovsky[1,2], Jason Liu[3], Jayanth Krishnan[2], Donghoon Lee[1,2], Yao Fu[1,2], Ekta Khurana[4], Mark Gerstein[1,2,5]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, [3]Program in Applied Math, Yale University, New Haven, CT, [4]Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, [5]Department of Computer Science, Yale University, New Haven, CT

In cancer research, background models for mutation rates have been extensively calibrated in coding regions, leading to the identification of many driver genes, recurrently mutated more than expected. Noncoding regions are also associated with disease; however, background models for them have not been investigated in as much detail. This is partially due to limited noncoding functional annotation. Also, huge mutation heterogeneity and potential correlations between neighboring sites give rise to substantial overdispersion in mutation count, resulting in problematic background rate estimation. Here, we first address these issues with a new computational framework called LARVA. It integrates variants with a comprehensive set of noncoding functional elements, modeling the mutation counts of the elements with a beta-binomial distribution to handle overdispersion. Moreover, it is known that part of the mutation rate heterogeneity relates to confounding genomic features, such as replication timing and chromatin organization. Here, we extended our model with a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). This approach uses a Gamma-Poisson mixture model to capture the mutation rate heterogeneity across different individuals and thus models the over dispersed mutation counts by a negative binomial distribution. Furthermore, it regresses the mutation counts against 381 features extracted from REMC and ENCODE to accurately estimate the local background mutation rate. This framework can be readily extended to accommodate additional genomic features in the future.

We demonstrate our model's effectiveness on 760 whole-genome tumor sequences, showing that it identifies well-known noncoding drivers, such as mutations in the TERT promoter. In addition, our two models highlighted several novel highly mutated regulatory sites that could potentially be noncoding drivers. We further applied our models on more than 2500 whole genome sequencing data from the Pan-Cancer Analysis Working Group (PCAWG) aiming to find mutational hotspots for specific cancer types. As a result, numerous well-known and novel coding and noncoding elements have been reported as potential cancer driver events.

ADIPOSE TISSUE CELL-TYPE DECONVOLUTION TO UNCOVER BMI AND CELL-TYPE SPECIFIC REGULATORY EFFECTS

Craig A Glastonbury, Kerrin S Small

King's College London, Twin Research & Genetic Epidemiology, London, United Kingdom

Genetic regulation of gene expression is cell-type specific and variation in cell-type composition at a population level has been extensively studied in whole blood. Whole blood cell-type proportions are easily measured and are now known to vary with age, season and a range of additional exposures. However similar studies from solid tissues are lacking and large-scale separation of cells from solid tissues is difficult. Therefore we utilized a recently published *v*-SVR algorithm (CIBERSORT) to estimate the relative proportion of seven dominant cell types found in primary subcutaneous adipose tissue biopsies (SAT) (N=766, TwinsUK). We constructed a basis matrix of cell-type specific expression from RNA-seq obtained from purified cells known to be present in SAT. Bootstrapping was used to assess accuracy of cell type deconvolution in our SAT samples. A median RMSE (0.59) and Pearson correlation (0.84) across samples was observed, suggesting accurate estimation of constituent cell types. Clustering of identified signature genes recapitulated the known relationship of immune and non-immune cell fractions present in SAT. We show the dominant cell type proportions present in SAT are Adipocytes ($\mu = 0.78$, $\sigma = 0.08$), Microvascular endothelial cells ($\mu = 0.09$, $\sigma = 0.03$) and Macrophages ($\mu =0.06$, $\sigma = 0.07$). We also observe a significant correlation between BMI and Macrophages ($r = 0.30$) – consistent with published work demonstrating increased Macrophage infiltration into SAT with obesity. We validated our estimates by implementing an independent non-negative quadratic programming approach and show that cell-estimates between methods are highly concordant (Macrophages, r=0.90, Adipocytes r=0.89). Additionally, we estimated cell proportions in an independent SAT dataset (N=200) and achieve comparable accuracy. BMI is highly associated to gene expression levels in SAT. To understand if the effect of BMI on expression is driven by variability in cell-type, we carried out a transcriptome-wide association study of BMI with and without adjusting for Macrophages in TwinsUK. 14% (797) of BMI-associated genes are Macrophage dependent. Utilizing the same cell-type correction strategy for *cis*-eQTL discovery we detected 100 cell-type specific *cis*-eQTLs (FDR 5%). PCA may readily capture cell-type composition and is widely used in *cis*-eQTL analyses. Future work will focus on the effects of cell-type for *trans*-eQTL identification, in which PCs inappropriately capture and remove multi-gene *trans*-eQTL effects. We plan to extend our analysis to a diverse set of tissue samples, to understand the impact of cell-type composition on obesity throughout the body.

CANCER GENOME ASSEMBLY AND STRUCTURAL VARIANT DETECTION WITH BIONANO OPTICAL MAPPING AND PACIFIC BIOSCIENCE LONG READS

Sara Goodwin[1], Maria Nattestad[2], Karen Ng[3], Timour Baslan[2], Tyler Garvin[2], James Gurtowski[2], Elizabeth Hutton[2], Timothy Beck[3], Yogi Sundaravadanam[3], Melissa Kramer[1], Eric Antoniou[1], John McPherson[4], James Hicks[2], Michael C Schatz[2], W. Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Stanley Center for Congnative Genomics, Cold Spring Harbor , NY, [2]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor , NY, [3]Ontario Institute for Cancer Research, Cancer Genomics, Toronto, Canada, [4]University of California, Cancer Center , Davis , CA

Cancer is a disease characterized by widespread structural variations (SVs) and CNVs. While elucidating the nature of such events is not a trivial task substantial progress had been made using short read sequencing technology (SRS). Despite this, SRS methods are insufficient for detecting long range and repetitive SVs and do not provide an adequate read length to achieve a highly contiguous assembly, especially for complex genomes. Long read sequencing (LRS) approaches generate reads in excess of 10kb, providing the read lengths needed to resolved large variations and to generate contiguous assemblies.

The breast cancer line SK-BR-3 has widespread duplication and rearrangements leading to the multiplication of regions including HER3. Recent work on this line employing the Pacific Bioscience RS II platform has highlighted the utility of LRS for cancer genomes by revealing a series of nested duplications and translocations between chr17 and chr8. However, substantial effort, computational time and high costs are complications associated with LRS protocols. One solution is optical mapping methods employed by Bionano and others. In this approach, long DNA fragments are specifically labeled and visualized. The labeled fragments are aligned to create a genome map capable of resolving large SVs providing a means of identifying and/or validated SVs on the molecular level. These maps can then be used to scaffold PacBio reads generating a highly contiguous assembly with lower LRS input.

We generated 125X coverage of SK-BR-3 using 2 IrysChips. From the de novo Bionano assembly, we identified hundreds of SVs, including fusions involving genes associated cancers. Of the >10kb SVs identified in the PacBio data, more than 50% were also found in the Bionano data, including the nested translocation between chr8 and chr17. When paired with the PacBio data, the hybrid assembly increased the assembly N50 from 2.455Mb to 6.788Mb and covered >98% of the genome. Preliminary down-sampling experiments decreasing the level of PacBio coverage indicate that as much as 40% less PacBio coverage is required to generate a contiguous assembly when paired with optical maps. These results large and complex genomes can be sequenced assembled and analyzed at costs lower than existing LRS-only approaches.

ANCIENT WHOLE DOG GENOMES SHOW NO EVIDENCE OF POPULATION REPLACEMENT IN NEOLITHIC EUROPE

Shyamalika Gopalan[1], Laura Botigué[1], Shiya Song[2], Amelie Scheu[3,7], Timo Seregély[4], Andrea Zeeb-Lanz[5], Rose-Marie Arbogast[6], Kevin Daly[7], Martina Unterländer[3], Angela Taravella[2], Matthew Oetjens[2], Amanda Pendleton[2], Dan Bradley[7], Jeffrey M Kidd[2], Joachim Burger[3], Krishna R Veeramah[1]

[1]Stony Brook University, Department of Ecology and Evolution, Stony Brook, NY, [2]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [3]Johannes Gutenberg University, Palaeogenetics Group, Institute of Anthropology, Mainz, Germany, [4]University of Bamberg, Department of Prehistoric and Protohistoric Archaeology, Bamberg, Germany, [5]Generaldirektion Kulturelles Erbe Rheinland-Pfalz, Direktion Landesarchäologie, Speyer, Germany, [6]Université de Strasbourg, CNRS-UMR 7044 "ARCHIMEDE", Strasbourg, France, [7]Trinity College, Smurfit Institute of Genetics, Dublin, Ireland

Recent efforts aimed at understanding the evolutionary past of modern domesticated dogs using genetic evidence have supported conflicting hypotheses regarding the location and timing of their origins, as well as the extent of major episodes of gene-flow and migration. Analyses of modern dog DNA are limited by the fact that today's common breeds are the result of strong artificial selection and extensive crosses between relatively few lineages, particularly during the Victorian era. By using ancient DNA, it is possible to gain insights into processes that occurred before the last 200 years and improve our understanding of older events in human/dog prehistory. So far, studies of ancient dog DNA have focused on uniparental loci or specific phenotypic markers. We present the full genomes of two European Neolithic dogs from Germany sequenced to 9x coverage. Both these dogs belong to mitochondrial haplotype C, consistent with other Paleolithic and Neolithic individuals from Europe. However, we also identify significant ancestry from wolves and non-European dogs in both ancient individuals using autosomal genome-wide data. We find that the older ancient dog (~7,000 years old), shares ancestry primarily with modern European dogs and has some genetic similarity to a source population resembling modern East Asian village dogs. Interestingly, the younger ancient dog (~5,000 years old), which shows some genetic continuity with the older specimen, appears to have experienced significant admixture with a population represented by modern day Indian dogs and wolves. Our data demonstrate that the current population structure of modern village dogs was largely established prior to the Neolithic. We also find that, despite a strong component of ancestry which closely resembles present day East Asian and Indian village dogs, suggesting migration from non-European populations, this did not lead to population replacement during the Neolithic as has been previously suggested by mtDNA evidence.

← Back

WHEN IS SELECTION EFFECTIVE?

Simon Gravel

McGill University, Human Genetics, Montreal, Canada

Deleterious alleles can reach high frequency in small populations because of random fluctuations in allele frequency. This may lead, over time, to reduced average fitness. In that sense, selection is more `effective' in larger populations. Recent studies have considered whether the different demographic histories across human populations have resulted in differences in the number, distribution, and severity of deleterious variants, leading to an animated debate.

This presentation first seeks to clarify some terms of the debate by identifying differences in definitions and assumptions used in recent studies. We argue that variants of Morton, Crow and Muller's `total mutational damage' provide the soundest and most practical basis for such comparisons.

Using simulations, analytical calculations, and 1000 Genomes data, we provide an intuitive and quantitative explanation for the observed similarity in genetic load across populations.

We show that recent demography likely has substantially affected the effect of selection, and still affects it, but the net result of these accumulated differences is small.

Direct observation of differential efficacy of selection for specific allele classes is nevertheless likely possible with contemporary datasets. By contrast, identifying genome-wide differences in the efficacy of selection across populations will require many modelling assumptions, and is unlikely to provide biological insight about humans.

COMPREHENSIVE CHARACTERIZATION OF RNA ELEMENTS IN THE HUMAN GENOME

Brenton R Graveley[1], Chris Burge[2], Xiang-Dong Fu[3], Eric Lecuyer[4], Eugene W Yeo[3]

[1]UCONN Health, Genetics and Genome Sciences, Farmington, CT, [2]MIT, Biology and Biological Engineering, Cambridge, MA, [3]UCSD, Cellular and Molecule Medicine, San Diego, CA, [4]Institut de Recherches Cliniques de Montréal, Montreal, Canada

The human genome encodes hundreds of RNA binding proteins (RBPs) that recognize RNA in a sequence-specific manner, although the binding sites for only a small subset of these proteins have been identified, particularly on the genome-wide scale. We are working towards comprehensively characterizing the RNA elements recognized by all human RBPs by studying an initial set of 250 RBPs in two human cell lines (K562 and HepG2). For each RBP we are generating CLIP-Seq data to identify the binding site for each RBP, RNA-Seq after shRNA knockdown to identify the impact of each RBP on the transcriptome, and profiles of their subcellular localization by immunofluorescence. We are also characterizing subsets of these RBPs with RNA Bind-N-Seq (RBNS) to characterize the binding affinity and specificity of the RBP *in vitro* and ChIP-Seq assays to interrogate potential RBP-chromatin interactions. We have tested 695 antibodies for 538 RBPs by immunoprecipitation followed by Western blotting and validated 455 antibodies for 384 unique RBPs. To date, we have generated 735 replicated datasets including 102 CLIP-Seq, 30 ChIP-Seq, 350 RNA-Seq after shRNA knockdown experiments in K562 and HepG2 cells, subcellular localization profiles for 217 RBPs in HepG2 cells, and RNBS data for 36 RBPs with hundreds of additional datasets currently being generated. We will present an update and overview of this dataset and examples of insights that have been obtained. All data is publicly available at http://www.encodeproject.org as soon as it is generated and verified and can be used without restriction.

Additional authors who could not be fully accommodated in the antiquated online abstract submission system: Cassandra Bazile[2], Steven Blue[3], Neal Cody[4], Daniel Dominguez[2], Michael Duff[1], Keri Elkins[3], Peter Freese[2], Abigail Hochman[2], Nicole Lambert[2], Sara Olson[1], Athma Pai[2], Gabriel Pratt[3], Rebecca Stanton[3], Balaji Sundararaman[3], Taiki Tsutsui[3], Eric Van Nostrand[3], Xiaofeng Wang[4], Xintao Wei[1], Rui Xiao[3], Lijun Zhan[1], Olivia Zhang[4]

← Back

SEQUENCE CO-EVOLUTION PREDICTS RESIDUE-LEVEL PROTEIN INTERACTIONS

Anna G Green[1], Thomas A Hopf[1,2], Charlotta P Schärfe[1,3], Debora S Marks[1]

[1]Harvard Medical School, Systems Biology, Boston, MA, [2]Technische Universität München, Bioinformatics and Computational Biology, Garching, Germany, [3]University of Tübingen, Quantitative Biology Center and Department of Computer Science, Tübingen, Germany

A detailed understanding of macro-molecular interactions is critical for interpreting the effects of mutations on organismal phenotype. Such interactions remain difficult to measure in a systematic way due to biases in existing experimental methods. We recently demonstrated that computational approaches that leverage residue co-evolution are able to detect residues that are functionally coupled between proteins. These co-evolving residues are sufficiently close in space to yield three-dimensional structure of the protein complexes. Here we present that these computational approaches can be used to quantify the probability of interaction between proteins. We test the utility of this approach for all hetero- and homo-oligomeric protein complexes in *Escherichia coli* with a known structure, and demonstrate the ability to predict experimentally inaccessible interactions. We present the possibility of extending this approach to eukaryotic genomes, where the lack of sequence diversity has previously posed a challenge. This research provides tools to determine protein interactions and oligomerization from sequence alone, as well as insight into evolutionary constraint on macromolecular interactions.

THE HUMAN MICROBIOME AS SURVEYED USING A RAPID, CULTURE-FREE WHOLE GENOME ASSEMBLY APPROACH

Nicholas Putnam[1], Jonathan Stites[1], Robert Calef[1], Paul Havlak[1], Marco Blanchette[1], Ei Ei Min[1], Brendan O'Connell[1,2], Richard Green[1,2]

[1]Dovetail Genomics, Santa Cruz, CA, [2]University of California, Santa Cruz, Biomolecular Engineering, Santa Cruz, CA

Humans, like most multi-cellular organisms live in concert with a rich and diverse community of microbes that inhabit our bodies. Recent work to catalogue and classify these communities has shown the enormous diversity of microbes and how they contribute to human health and disease. A major challenge to understanding the dynamics of microbial communities is the inability to perform rapid and complete de novo genome assemblies of the constituent organisms.

We describe a simple, powerful, and fast new approach for metagenomics genome assembly. Using a culture-free in vivo proximity-ligation method, we generate libraries that reveal long-distance genome information that can be used to assemble multi-megabase scaffolds and in many cases complete genomes from metagenomic DNA preps. We apply this method to fecal and oral human microbiome samples to catalog the organism and strain diversity in these human microbiome niches. The speed of this approach, from sample to complete assemblies in less than a week, makes it amenable for use in a clinical setting.

NEW DISCOVERIES REGARDING INTROGRESSION INTO NEANDERTALS AND DENISOVANS

Ilan Gronau*[1], Melissa J Hubisz*[2], Martin Kuhlwilm*[3,4], Cesare de Filippo[3], Javier Prado-Martinez[4], Martin Kircher[3,5], Qiaomei Fu[3,6], Hernán A Burbano[3,7], Carles Lalueza-Fox[4], El Sidrón cave paleontologists[8,9], Vindija cave paleontologists[10], Tomas Marques-Bonet[4], Aida M Andrés[3], Bence Viola[3], Svante Pääbo[3], Matthias Meyer[3], Adam Siepel[11,2], Sergi Castellano[3]

[1]Herzliya Interdisciplinary Center (IDC), Herzliya, Israel, [2]Cornell University, Ithaca, NY, [3]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [4]Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain, [5]University of Washington, Seattle, WA, [6]Chinese Academy of Sciences, Beijing, China, [7]Max Planck Institute for Developmental Biology, Tübingen, Germany, [8]Universidad de Oviedo, Oviedo, Spain, [9]Museo Nacional de Ciencias Naturales, Madrid, Spain, [10]Croatian Academy of Sciences and Arts, Zagreb, Croatia, [11]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

*contributed equally

Neandertals and Denisovans are two extinct human groups that diverged from modern humans roughly 600,000 years ago. We examine here evidence of introgression into these two archaic groups from other archaic and modern human populations. It is known that the ancestors of the Denisovans admixed with an unknown group of archaic hominins that diverged from other humans more than a million years ago (Prüfer *et al.*, 2014). We have discovered an additional admixture event, in this case between the ancestors of Eastern Neandertals and an unknown population of modern humans. Evidence for this event comes from the analysis of the genealogical relationships along the genomes of several Neandertals, the Denisovan, and present-day humans. Our analysis suggests that Eastern Neandertals received gene flow roughly 100,000 years ago from a modern human population that diverged early from other modern humans in Africa. Thus, this episode of introgression happened in the other direction and many thousands of years before the interbreeding episodes that left traces of Neandertal and Denisovan ancestries in the genomes of present-day non-Africans.

We conduct a detailed investigation of the genetic traces from the two introgression events into Eastern Neandertals and Denisovans by applying methods for reconstructing the ancestral recombination graph (ARG) of archaic and modern human genomes. The ARG provides a complete description of inferred coalescence and recombination events, and thus allows us to design robust criteria for detecting introgressed haplotypes based on age, length, and local tree structure. Using these criteria, we characterize the influence of the different introgression events on patterns of divergence between archaic and present-day humans, and examine possible adaptive pressures acting on the admixed haplotypes.

NO EVIDENCE FOR TRANSGENERATIONAL GENETIC EFFECTS IN THE TRASCRIPTOME OF ISOGENIC DERIVED MOUSE OFFSPRING

Rodrigo Gularte-Merida, Carole Charlier, Michel Georges

Unit of Animal Genomics, GIGA -- Research, University of Liège, Liège, Belgium

Experiments carried out in *C. elegans* and *D. melanogaster* support the existence of paternal transgenerational genetic effects, i.e. the effect of an untransmitted paternal alleles on the offsprings' phenotype. To test whether such effects might also occur in mammals, we generated multiple cohorts of isogenic mice derived from several (C57BL/6J x A/J consomic) F1 sires — mice harbouring an A/J chromosome (MMU15, 17, 19 or X) in an otherwise C57BL/6J (B6) background— and B6 purebred females. In total we developed three distinct N2 backcrosses and one N3 backcross generating 833 male mice apparently isogenic for C57BL/6J. The analysis of global gene expression measured by RNA-seq in 50 pooled libraries from five tissues showed 53 genes with significant differential expression between any one isogenic derived cohort and B6 controls. Of these 53, only 10 were found differentially expressed in a technical replication via RT-qPCR — individual measurements of each gene in each of the 200 sequenced samples—. Further analysis to validate the latter 10 genes in an additional 800 biological replicates from all our isogenic cohorts [N2 litter-mates, N2 mice from an independent backcross, and N3 backcross mice (N2 isogenic male x B6 female)] replicated the original observation of differential expression in only three genes; Mid1, Crem, and Gm26448. The gene Mid1 was discarded as DNA-seq data showed a Strain Specific Variant (SSV) on the XY-Pseudo-Autosomal Region is responsible for the effect in that cohort. In contrast the contrast between isogenic derived and B6 controls in the genes Crem and Gm26448 have no evidence of SSV and showed a 0.63 ($p < 0.0001$) and -0.34 ($p = 0.009$) difference in expression ($2^{-\Delta\Delta Ct}$), respectively. Finally, we measured again Crem and Gm26448 in 17 additional mice from their respective strains and >17 matched controls. This final analysis revealed that only Gm26448 showed differences of -0.7 (N2 littermate) and -0.4 (N2 independent backcross) expression. However, these failed to reach significance thresholds. We have systematically explored the hypothesis paternal transgenerational genetic effects in the mouse and with our experimental design this phenomena fails to provide sufficient evidence to show its contribution to the offspring's phenotype in a consistent manner. Thus, despite that transgenerational effects have been shown to exist in *C. elegans* and *D. melanogaster*, we conclude that in the mouse, under natural allelic variation, this phenomena appears to be extremely rare, and with very small effects.

← Back

COMBINATIONS OF GENOMIC PROPERTIES THAT EXPLAIN SELECTIVE PRESSURE ALSO PREDICT FUNCTIONAL ELEMENTS.

Brad Gulko[1], Adam Siepel[2]

[1]Cornell University, Graduate Field of Computer Science, Ithaca, NY,
[2]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

We present work in progress on new methods for scaling up fitCons (2015) to large collections of functional genomic covariates.

FitCons provides a generative model that utilizes selective pressure (INSIGHT, 2013) to indicate the potential for function associated with patterns of functional annotation at each human genomic position. While fitCons demonstrated a superior ability to identify cis-regulatory elements, original studies were limited by the use of available functional genomic data (DNase-I hypersensitivity, RNA-Seq, CDS annotation, and ChromHMM classification) for three ENCODE cell lines. Increasing the diversity of functional assays and cell types was found to improve both genomic coverage and prediction accuracy for cis-regulatory elements. However, the initial fitCons method generated a separate functional class for every possible combination of assay results; yielding an exponential growth of putative functional classes with genomic properties of interest.

Our new work employs a subdivision technique that recursively picks the additional genomic property that explains the most Shannon information about selective pressure. This allows us to generate a finely grained, whole genome segmentation while simultaneously limiting the number of functional classes and expanding tissue diversity to include 14 cell types from the Roadmap Epigenomics Project. We also investigate 10 cell-type specific genomic properties including DNA methylation, splice site proximity, and transcription factor binding. The information theoretical framework allows the identification of synergistic combinations of genomic properties; automatically identifying non-convex combinations of properties that are more informative together than the simple sum of their individual information. Practically, this process can also be used to optimize experimental design by predicting which additional assay would be most informative, given an existing collection of measurements.

THE CORRELATION ACROSS POPULATIONS OF MUTATION EFFECTS ON FITNESS

Alec J Coffman[1], Aaron P Ragsdale[2], PingHsun Hsieh[3], Ryan N Gutenkunst[1,3]

[1]University of Arizona, Department of Molecular and Cellular Biology, Tucson, AZ, [2]University of Arizona, Program in Applied Mathematics, Tucson, AZ, [3]University of Arizona, Department of Ecology and Evolutionary Biology, Tucson, AZ

Divergent selection, in which the same allele has different effects on fitness in different populations, drives environmental speciation. Much is known about patterns of genetic variation near loci with given divergent selection coefficients, but little is known about overall genomic patterns of divergent selection. To fill this gap, we developed a framework for inferring the joint distribution of fitness effects (DFE) between pairs of populations, based on a diffusion approximation to the joint allele frequency spectrum. We applied this framework to African and American populations of *Drosophila melanogaster*, first estimating demographic history and then the joint DFE. As expected, we found that genome-wide mutation fitness effects were highly correlated between these two populations. Considering functional subsets of genes, however, revealed striking differences. For example, for muscle development genes the joint DFE has a correlation of roughly zero, whereas for neuronal development genes it is almost one. Divergent selection in these populations is thus operating much more strongly on the musculature than the brain, suggesting that adaptation is primarily physiological rather than behavioral. Future work will apply this general framework to other populations and species, thus quantifying how divergent selection varies with environmental and molecular context.

MASSIVELY PARALLEL SINGLE NUCLEOTIDE MUTAGENESIS USING REVERSIBLY-TERMINATED INOSINE

Gabriel Haller[1], David Alvarado[1], Kevin McCall[1], Ping Yang[1], Robi Mitra[4], Matthew Dobbs[1,5], Christina Gurnett[1,2,3]

[1]Washington University, Department of Orthopaedic Surgery, St. Louis, MO, [2]Washington University, Department of Neurology, St. Louis, MO, [3]Washington University, Department of Pediatrics, St. Louis, MO, [4]Washington University, Center for Genome Biology and Systems Biology, Department of Genetics, St. Louis, MO, [5]Shriners Hospital for Children, St. Louis, MO

Accurate, inexpensive and efficient methods of large scale mutagenesis of target DNA sequences are needed to comprehensively assess the effects of single nucleotide changes on protein function. While synthetic oligonucleotides can be used as a template for mutagenesis, these remain expensive for large projects. Here we demonstrate the construction of a systematic allelic series (SAS), a massively parallel single nucleotide mutagenesis using reversibly-terminated deoxyinosine triphosphates (rtITP). To demonstrate the utility of SAS mutagenesis, we created a mutational library containing every single nucleotide mutation in a 200bp region containing the active site of the TEM-1 β-lactamase gene, and identify critical residues that negatively impact ampicillin resistance. Because of its low cost and simple methodology, SAS-based mutational library creation has the ability to greatly expedite interpretation of human disease variants, in vitro protein evolution and the determination of critical domains for protein function.

← Back

EVOLUTION OF ABDOMINAL PIGMENTATION IN *DROSOPHILA*: A PHENOTYPE CONTROLLED BY A GENE REGULATORY NETWORK

Clair Han[1], Alisa Sedghifar[2], Mark J Rebeiz[3], Peter Andolfatto[2]

[1]Princeton University, Quantitative and Computational Biology, Princeton, NJ, [2]Princeton University, Ecology and Evolutionary Biology, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [3]University of Pittsburgh, Biological Sciences, Pittsburgh, PA

Drosophila santomea is the only species to have light abdominal tergite pigmentation in the *melanogaster* subgroup. Abdominal pigmentation is well-understood in *Drosophila*, and several of these pigmentation genes (*tan*, *yellow* and *ebony*) have been implicated in the difference seen in *D. santomea* and its sister *D. yakuba*. Multiple distinct haplotypes support a soft sweep in the enhancer region of the *tan* gene, suggesting that the loss in pigmentation has been adaptive. To see if similar patterns of selection are associated with pigmentation genes, we obtained whole-genome sequences from *D. santomea* and *D. yakuba* individuals and haplotype data for targeted regions spanning known pigmentation genes. We test for population genomic evidence of past selection acting at these loci.

RECONSTRUCTING A/B COMPARTMENTS AS REVEALED BY HI-C USING LONG-RANGE CORRELATIONS IN EPIGENETIC DATA

Jean-Philippe Fortin[1], Kasper D Hansen[1,2]

[1]Johns Hopkins University, Biostatistics, Baltimore, MD, [2]Johns Hopkins University, Institute of Genetic Medicine, Baltimore, MD

Analysis of Hi-C data has shown that the genome can be divided into two compartments called A/B compartments. These compartments are cell-type specific and are associated with open and closed chromatin. We show that A/B compartments can reliably be estimated using epigenetic data from several different platforms: the Illumina 450k DNA methylation microarray, DNase hypersensitivity sequencing, single-cell ATAC sequencing and single-cell whole-genome bisulfite sequencing. We do this by exploiting that the structure of long-range correlations differs between open and closed compartments. This work makes A/B compartment assignment readily available in a wide variety of cell types, including many human cancers.

THE ASSOCIATION BETWEEN HISTONE MODIFICATION ABUNDANCE AND GENE EXPRESSION ACROSS INDIVIDUALS

Kipper Fletez-Brant[1,2], Kasper D Hansen[1,2]

[1]Johns Hopkins University, Institute for Genetic Medicine, Baltimore, MD, [2]Johns Hopkins University, Biostatistics, Baltimore, MD

Histone modifications are known to mark genomic elements such as enhancers and promoters. But less is known about the quantitative effect of histone binding. In this work we ask, to what degree histone modification abundance is associated with gene expression levels, comparing the same genomic region across multiple individuals within a cell type. We measure histone modification abundance by ChIP-seq and gene expression by RNA-seq using publicly available data on HapMap LCL cell lines, and study the three histone modifications H3K4me3, H3K4me1 and H3K27ac over promoters for known genes and enhancers identified in the FANTOM5 project. We find that the strongest relationship between histone binding and gene expression is identified when histone binding is quantified using peak calling simultaneously on multiple biological replicates; substantially worse relationships are obtained when using peaks based on data from single samples such as the Roadmap Epigenomics peaks. To identify significant associations we define a null distribution based on input experiments on the same biological replicates; this null distribution is substantially different from a null distribution based on permuting the sample labels or by using asymptotic statistical theory, showing the usefulness of an experimentally derived null distribution. We find that 5575 genes have a significant relationship between gene expression and abundance of H3K4me3 in their promoter. Similar results hold true for the enhancer marks H3K4me1 and H3K27ac. This work increases our understanding of the degree to which quantitative variation in histone modifications are associated with phenotypic variation between individuals within a cell type.

MUTATIONAL STRAND ASYMMETRIES IN CANCER GENOMES REVEAL MECHANISMS OF DNA DAMAGE AND REPAIR

Nicholas J Haradhvala[1,2], Paz Polak[1,2,3], Michael S Lawrence[2], Gad Getz[1,2,3]

[1]Massachusetts General Hospital, Cancer Center, Boston, MA, [2]The Broad Institute, Cancer Program, Cambridge, MA, [3]Harvard Medical School, Department of Pathology, Boston, MA

Mutational processes constantly shape the somatic genome. Sometimes this leads to cancer, but it also might lead to autoimmune disease, aging, and other natural processes over time. When cancer is the outcome, we are afforded a glimpse into these processes by the clonal expansion of the malignant cell. It has been shown that heterogeneity in mutational densities along the genome are the consequence of differential damage and repair due to various properties of the genome, such as tightness of chromatin wrapping. Furthermore, asymmetries in the mutational densities between complementary transcribed and non-transcribed strands capture the strand-specific effects of transcription-coupled repair. We explore an additional layer of mutational heterogeneity by examining strand asymmetric mutation at the DNA replication fork. Analyzing whole genome sequences of 590 tumors from 14 different cancer types, we discovered widespread asymmetries with a sweeping spectrum separating mutagenic processes to two classes: “T- class”, dominated by transcriptional asymmetry; and “R-class”, dominated by replicative asymmetry. The T-class included UV- and smoking-related processes as well as a process in liver cancer by which the non-transcribed strand is damaged in addition to repair of the transcribed strand. The R-class included APOBEC-associated tumors and provided evidence of the lagging-strand template as the substrate of this form of mutagenesis. The R-class also included mismatch-repair-deficient tumors, demonstrating a careful balance in asymmetries between mismatch repair and damage introduced during DNA replication. Moving forward we leverage mutational signatures to more directly measure and classify the asymmetries of more subtle mutational and repair processes, and expand our search to the broad array of currently available cancer types.

INSERTION AND DELETION IDENTIFICATION AND CHARACTERIZATION ACROSS A SEVEN SPECIES BABOON DIVERSITY PANEL

R. Alan Harris[1], Muthuswamy Raveendran[1], Clifford J Jolly[2], Jane Philips-Conroy[3], Todd Disotell[2], Andy Burrell[2], Yue Liu[1], Donna Muzny[1], Kim C Worley[1], Richard A Gibbs[1], Jeff Rogers[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Molecular and Human Genetics, Houston, TX, [2]New York University, Anthropology, New York, NY, [3]Washington University, Anatomy and Anthropology, St. Louis, MO

We identified genome wide insertions and deletions, indels, ranging from 1-60bp across a diversity panel consisting of Papio northern clade (*P. anubis* n=4, *P. papio* n=2, *P. hamadryas* n=2) and southern clade (*P. cynocephalus* n=2, *P. kindae* n=3, *P. ursinus* n=2) species as well as a gelada baboon (*Theropithecus gelada* n=1) outgroup. Illumina reads were mapped to the papAnu2 (*P. anubis*) assembly using BWA mem and indels were called using GATK. A total of 8,498,081 indel sites, comprised of 3,724,924 insertions, 3,977,812 deletions, and 795,345 complex sequence alterations, were identified at an average sequencing depth of 21X. These indels represent 17,920,946bp of inserted sequence and 18,855,511bp of deleted sequence. Variant Effect Predictor (VEP) was used to determine the consequences of the indels based on Ensembl gene models. Putative functional consequences included 6,936 frameshift, 1,601 inframe insertion, 1,614 inframe deletion, and 5,277 splicing associated variants.

Genes demonstrating frameshift variants (3,112) were further characterized. Many of the genes are members of large families such as olfactory receptors (67), zinc finger proteins (95), and G-protein coupled receptors (19) where frameshifts in a single member of the family may be more tolerated. We examined enrichments of gene annotation terms using the Database for Annotation, Visualization and Integrated Discovery (DAVID). UP Tissue expression enrichment at an FDR < 0.001 identified Epithelium, Testis, and Brain. EnrichR analysis of GTEx Tissue Sample Gene Expression Profiles showed that frameshift containing genes were enriched for genes upregulated in skin (p=1.656e-26).

The pattern of indels across species followed expectations based on phylogenetic relationships among the species. The northern clade species, which includes the reference species, showed similar average indels per individual (*P. anubis* 1,520,212; *P. papio* 1,520,990; *P. hamadryas* 1,715,468). The southern clade showed more average indels per individual (*P. cynocephalus* 2,173,180; *P. kindae* 2,474,452; *P. ursinus* 1,958,744) and *T. gelada* showed the most (3,347,496). Indels homozygous for the alternative allele in all samples from either clade, among which are clade specific indels that could be confirmed with additional samples, also show a pattern consistent with phylogenetic relationships (northern 585, southern 2,891).

LARGE-SCALE INDEL DISCOVERY IN RHESUS MACAQUES (*MACACA MULATTA*)

M Raveendran[1], RA Harris[1], L Cox[2], G Fan[3], B Ferguson[4], J Horvath[5], S Kanthaswamy[6], HM Kubisch[7], D Liu[8], M Platt[9], D G Smith[6], B Sun[8], E J Vallender[10], R W Wiseman[11], D M Muzny[1], R A Gibbs[1], J Rogers[1]

[1]BMC, HGSC, Houston, TX, [2]SNPRC, San Antonio, TX, [3]UCLA, Los Angeles, CA, [4]ONPRC, Beaverton, OR, [5]North Carolina Museum of Natural Sciences, Raleigh, NC, [6]CNPRC, Davis, CA, [7]TNPRC, Covington, LA, [8]Anhui Univ, Anhui, China, [9]Univ. of Pennsylvania, Philidelphia, PA, [10]Univ. of Mississippi Medical Center, Jackson, MS, [11]WNPRC, Madison, WI

Among closely related species (e.g. humans and chimpanzees) small insertion-deletion differences (indels) account for more genetic divergence than single base substitutions. Indels are also known to explain a number of damaging mutations associated with human disease. Rhesus macaques (*Macaca mulatta*) are one of the most commonly used nonhuman primates in biomedical research, and genetic variation among rhesus can be useful in studying and understanding human diseases. However, little is known about the frequency or functional consequences of small indel polymorphisms among rhesus macaques. Here we describe the identification of small indels (1 to 60 bases in length) in a sample of 133 rhesus macaques (124 Indian-origin and 9 Chinese-origin). Using GATK software to call indels from whole genome sequence data (average coverage 26.7x), and excluding variant calls that do not pass quality control filters or are likely the result of errors in the reference genome, we identified 7,926,645 indels across the 133 animals. This includes 3,026,673 insertions, 3,951,905 deletions and 948,067 more complex sequence alterations with ambiguous mechanism. Out of these 7.92 million indels, 1.68 million are found only in Chinese rhesus and 2.94 million are Indian-origin specific. Across the full dataset, we observed 2.6 indels per 1kb of the rhesus genome. 48.25% of bases affected are due to insertions and 51.75% are due to deletions. Most of the indels (74.5%) are 1 to 4 basepair in length. The average number of indels per animal is 1,249,646, which is higher than observations for individual human personal genomes. To determine potential effects on gene function or protein sequence, we analyzed all rhesus indels using the Ensembl Variant Effect Predictor (VEP) tool and identified stop codons altered (365), coding sequence frameshift indels (7,664), inframe insertions (1,563), inframe deletions (2,165) and splice region indels (4,317). In order to identify possible disease models, we queried the DisGeNET database (http://www.disgenet.org) with genes containing frameshift indels. This identified several genes associated with specific human diseases, including muscular dystrophy (*DMD*), microcephaly (*MCPH1*), cleft palate X-linked (*TBX22*) and Tangier disease (*ABCA1*).

← Back

USING THE LANDSCAPE OF GENETIC VARIATION IN PROTEIN DOMAINS TO IMPROVE FUNCTIONAL CONSEQUENCE PREDICTIONS.

Jim Havrilla[1], Brent S Pedersen[1,2], Ryan M Layer[1], Aaron Quinlan[1]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, Biomedical Informatics, Salt Lake City, UT

Numerous methods exist to predict the impact of a genetic variant on protein function. For example, the RVIS (Residual Variation Tolerance Score) study from Petrovski et al., provides a gene-wide score by regressing the number of common missense variants vs. the total number of variants. In contrast, the CADD (Combined Annotation Dependent Depletion) approach from Kircher et al., utilizes annotations and the ancestral genome to determine phenotypic impact by contrasting variants that survived natural selection with simulated mutations with a Support Vector Machine. Neither of these methods, however, directly utilize protein domain information – they only look at genes in the broader sense, not the numerous small and large functional portions of a protein for which they actually code. Accordingly, by comparing the Exome Aggregation Consortium's catalog of protein-coding genetic variation from more than 60,000 human exomes with the Pfam protein domain database, we have comprehensively measured the landscape of genetic variation among all characterized protein domains. Computing a metric to estimate rare variant density as well as the distribution of the ratio of missense to silent mutations for each domain per protein has allowed us to develop a model that should more accurately predict the likelihood that a variant in a particular genomic location will actually lead to phenotypic change. The fundamental rationale of the model is that variants overlapping protein domains that are tolerant of variation are less likely to have a functional impact, with the corollary being that variants affecting less tolerant domains are more likely to perturb protein function. We present our ongoing efforts to develop and validate a predictive model that integrates this information to reduce false negative and false positive predictions of the functional impacts of genetic variation in both research and clinical settings.

IDENTIFICATION OF CpG DESERTS IN HUMAN AND MOUSE GENOMES

Ximiao He, Charles Vinson

National Cancer Institute, NIH, Laboratory of Metabolism, Bethesda, MD

Previously, we have shown CpG methylation is mutagenic to produce TFBS that are important for the emergence of new regulatory regions and ultimately, novel phenotypes to drive evolution. While the unmethylated CG dinucleotides are clustering into CpG islands (CGIs), the majority of mutagenic CpG sites are leading to the depletion of CG dinucleotides, which produce the low CG regions in mammalian genomes. To study the biological function of the low CG regions, we identified the most depleted CG regions, typically 2~3 CG dinucleotides per Kbps, termed as CpG deserts (CGDs) in both human and mouse genomes. The majority of the CGDs are located in the inter-genic regions and ~2% are in promoter regions. The GO analysis of the genes with CGDs in the promoter regions showed that olfactory receptor activity was the most enriched molecular function in both human and mouse, which imply that the transform of mutagenic CG to TG may contribute the fast evolution of olfactory receptor genes. We calculated the average PhastCons score for each CGD, and observed that most of the CGDs were not conserved. We searched all the transcription binding sites (TFBSs) in the CGDs, the results showed that TBP, NCX, POU and OCT are the most enriched motifs. Considering the TG density, TBP and NCX motifs are enriched in CGDs with low TG density, while PBX, POU and OCT motifs are enriched in CGDs with higher TG density. Our results suggested the TG-rich CGDs are involved in the evolution of gene family such as olfactory receptor genes and important housekeeping TFBSs such as PBX, POU and OCT. Taken together, our study supported the idea that the mutation of methylated CG to TG dinucleotide plays a key role in the evolution process of mammalian genomes.

NCBI STRUCTURAL VARIATION HACKATHON – DEVELOPING OPEN-SOURCE TOOLS FOR COMPARING DBVAR DATA TO OTHER DATASETS

T Hefferon, J Garner, J Lopez, J Hsu, LQ Minh Tri, M Willi, T Mansour, Y Kai, B Busby, L Phan

January 2016 NCBI dbVar Hackathon Group, National Library of Medicine, National Institutes of Health, Bethesda, MD

A growing body of evidence implicates genomic structural variation (defined as variants > 50 bp) in the etiologies of a wide variety of human diseases, from diabetes, obesity, and cancer to psychiatric disorders and beyond. dbVar is an NCBI database of large structural variation containing over 3 million submitted structural variants from 120 human studies (among other species) including copy number variants (CNVs), insertions, deletions, duplications, inversions, and translocations. These data are centralized and freely available to the public, but their usefulness in downstream analysis has been limited for a variety of reasons including ambiguity in reported variant locations. To address these issues participants in January 2016 dbVar NCBI Hackathon collaborated over the course of a few days with the goal of producing tools and datasets to: 1) improve data exchange, data mining, computation, and reporting; 2) improve searching and matching of genomic coordinates across studies; 3) facilitate the comparison of annotations such as disease and phenotype, frequency, and genomic features with regions of structural variation; and 4) simplify variant display in sequence viewers as an aggregated histogram or density track across studies. The team generated a nonredundant set of genomic regions called Structural Variant Clusters (SVC) defined by regions of concordance amongst submitted human SSVs placed on assembly GRCh38. One can easily compare these SVC data against other annotated genomic regions such as those known to contain segmental duplications or other repeat structures, genes that are clinically significant, dosage sensitive and/or essential regions, and problematic genomic regions that may yield suspected false variants. In addition we developed open-source utilities for: 1) searching and filtering SVC data in GVF format; 2) computing summary statistics and exporting data for genomic viewers; and 3) annotating SVCs using external data sources. Tools and datasets from the January Hackathon are available at https://github.com/NCBI-Hackathons/Structural_Variant_Comparison. We invite the community to build on these tools and to participate in future NCBI Hackathons to build new ones. For more information on past and future NCBI Hackathon events, including opportunities to become involved, see https://github.com/NCBI-Hackathons or subscribe to NCBI news at http://www.ncbi.nlm.nih.gov/news/.

Acknowledgments
Work at NCBI is supported by the NIH Intramural Research Program and the National Library of Medicine.

CONTROLLING FOR PHYLOGENETIC RELATEDNESS IMPROVES DISCOVERING THE GENOMIC BASIS UNDERLYING SPECIES' PHENOTYPIC DIFFERENCES

Xavier Prudent[1,2], Genis Parra[1,2], Juliana Roscito[1,2], Michael Hiller[1,2]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Computational Biology and Evolutionary Genomics, Dresden, Germany, [2] Max-Planck-Institute for the Physics of Complex Systems , Computational Biology and Evolutionary Genomics, Dresden, Germany

The growing number of sequenced genomes allows us now to address a key question in genetics and evolutionary biology: What is the genomic basis that underlies phenotypic differences between species? Previously, we developed a computational framework called Forward Genomics that associates phenotypic to genomic differences by focusing on phenotypes that are repeatedly lost in independently lineages. Here, we present two new Forward Genomics methods that (i) control for the phylogenetic relatedness between the species of interest, (ii) control for differences in evolutionary rates and (iii) compute the significance of the association between phenotypic and genomic differences. We systematically compare these methods on simulated and on real data and demonstrate that the new methods significantly improve the sensitivity to detect such associations.

We use these methods to discover genomic loci that underlie the degeneration of the visual system in blind subterranean mammals. This genome-wide screen identifies many loci that are enriched in functions related to eye development and the perception of light as well as loci associated with eye diseases in human. In addition, we find genomic loci with a function in the circadian rhythm, which might be an adaptation to the subterranean environment.

The Forward Genomics framework has broad applicability to many other phenotypic differences. The new methods presented here significantly advance our ability to discover the genomic basis underlying phenotypic differences between species, which will contribute our understanding of how nature's phenotypic diversity has evolved.

UNRAVELING PRINCIPLES OF GENE REGULATION USING THOUSANDS OF DESIGNED REGULATORY SEQUENCES

Eran Segal

Weizmann Institute of Science, Computer Science, Rehovot, Israel

Genetic variation in non-coding regulatory regions accounts for a significant fraction of changes in gene expression among individuals from the same species. However, without a 'regulatory code' that informs us how DNA sequences determine expression levels, we cannot predict which sequence changes will affect expression, by how much, and by what mechanism. To address this challenge, we developed a high-throughput method for constructing libraries of thousands of fully designed regulatory sequences and measuring their expression levels in parallel, within a single experiment, and with an accuracy similar to that obtained when each sequence is constructed and measured individually. Using this ~1000-fold increase in the scale with which we can study the effect of sequence on expression, we designed and measured the expression of libraries in which the location, number, affinity and organization of different types of regulatory elements has been systematically perturbed. Our results provide several new insights into principles of gene regulation, bringing us closer towards a mechanistic and quantitative understanding of which how expression levels are encoded in DNA sequence.

USING MULTI-OMICS DATA TO INVESTIGATE INFLAMMATORY BOWEL DISEASE IN THE INTESTINAL EPITHELIUM

Kate J Howell[1,2], Judith Kraiczy [1], Anupam Sinha [3], Komal M Nayak[1], Marco Gasparetto [1], Philip Rosentiel [3], Matthias Zilbauer [1], Oliver Stegle [2]

[1]University of Cambridge , Paediatrics, Cambridge , United Kingdom, [2]EBI, Statistical Genomics group , Hinxton , United Kingdom, [3]University of Kiel , IKMB, Kiel, Germany

Inflammatory bowel diseases (IBD) are conditions causing chronic relapsing gut inflammation of the large bowel (Ulcerative Colitis) or the entire intestinal tract (Crohn's Disease). Over the last century there has been a significant increase in the incidence of IBD, particularly in children, many presenting with a severe clinical phenotype. IBD has been the subject of some of the largest GWAS studies to date, however these SNPs are not sufficient for disease development. Environmental factors are increasingly recognised to contribute to IBD, with epigenetic mechanisms and host-microbiota interactions now providing plausible explanations for how our environment can influence disease. Given the cell type specific nature of epigenetic and gene expression profiles, purified tissue samples are desirable in order to avoid confounding effects such as inflammation. Here we have performed a multi-omics analysis of highly purified intestinal epithelial samples, which represent a key cell type in IBD disease pathogenesis. We have generated genome wide DNA methylation (DNAm) (n=78) and gene expression profiles (RNAseq, n=66) from treatment-naïve paediatric IBD patients and matched healthy controls from the colon and ileum. All patients were genotyped. Differential methylation analysis between gut segments has identified large differences in the DNAm profiles in both patients and controls. While both inflammation and disease also show large methylation differences in the two tissues; the profiles do not completely overlap, allowing us to investigate disease specific differences. The gene expression profiles of the same epithelial samples reveal a very similar gut segment and disease specific separation.
Given the disease specific DNAm and trascriptomic profiles, we went on to investigate their use as diagnostic or prognostic biomarkers. Classification models built using only the DNAm data perform best when predicting disease outcome (i.e. severity). Importantly, the performance of models were improved by combining DNAm with RNA-seq data, highlighting one of the advantages of our dataset. Together, we present data from a unique, prospectively recruited cohort of treatment-naïve patients diagnosed with IBD. The multi-omics data generated from a purified, disease relevant cell type has already revealed a number of novel aspects, which further our understanding of disease pathogenesis and allow us to investigate the potential of new clinically relevant biomarkers.

IDENTIFYING SUBSTRUCTURE IN GENETIC RISK SHARING BETWEEN DISEASES

Luke Jostins, Gilean McVean

University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

While genetic risks for complex human diseases are often shared across multiple phenotypes, such as disorders of the immune system, this risk sharing shows complicated substructure. Risk alleles can be associated with multiple diseases with similar effect sizes or with smaller or opposite effects, or uniquely with one or a few diseases. Such complexity, along with incomplete power, makes inferring the exact sharing model for any given locus, and hence structure in sharing between loci, difficult.

To address these problems, we present a statistical approach that, instead of testing for sharing at each locus, instead attempts to model the underlying biology driving genetic risk substructure across multiple loci and phenotypes. Motivated by a network model of shared and potentially pleiotropic risk components, we use a reduced rank representation of genetic covariance in which each risk allele belongs to one of a small number of pathways, where each pathway varies in complexity from simple perfectly correlated effects to more complex covariance models. The method adjusts the number and rank of pathways to infer the simplest model that explains observed association summary statistics, given uncertainty in effect sizes. Individual loci are probabilistically assigned to inferred pathways, allowing us to examine the genes and functions associated with each pathway.

Using simulations, we show the power and accuracy of the method to estimate underlying parameters and choose the correct pathway structure using the Bayesian information criterion. For example, we have high power to recover the structure of a 4-pathway model under realistic scenarios (100 loci, 5 diseases, 10000 total samples).

We apply this method to 200 loci from 5 immune-mediated diseases (Crohn's disease/CD, ulcerative colitis/UC, psoriasis/PS, ankylosing spondylitis/AS, and primary sclerosing cholangitis/PSC). Our method selected a 4-pathway model (2 complex, 2 low rank) with significant substructure ($p<1e-68$ compared to a 1-pathway model), and 126/200 loci were assigned to a pathway with posterior >80%. The pathways highlight underlying shared biology of these diseases: for example, one simple pathway consisted of strong correlated effects on CD, UC and AS, with weaker effects on PSC and PS. This pathway was enriched for genes with the lymphocyte activation GO term, and included many important inflammatory genes such as *IL10*, *JAK2* and *STAT3*.

The difficulties of cross-disease analysis scale exponentially with the number of diseases. In contrast, by providing simplified, biologically interpretable models of genetic risk sharing, our method could potentially analyse shared pathways across hundreds of diseases.

← Back

HIGH-THROUGHPUT MAPPING OF REGULATORY DNA

Nisha Rajagopal[1], Sharanya Srinivasan[1,2], Kameron Kooshesh[2], Yuchun Guo[1], Matthew D Edwards[1], Budhaditya Banerjee[2], Tahin Syed[1], Bart J Emons[2], David K Gifford[1], Richard I Sherwood[2]

[1]MIT, CSAIL, Cambridge, MA, [2]Brigham and Women's hospital and Harvard Medical School, Division of Genetics, Department of Medicine, Boston, MA

We present the multiplexed editing regulatory assay (MERA), a high-throughput CRISPR-Cas9–based approach to test regulatory sequences for function in their native context. Unlike reporter assays that test for sequence sufficiency for gene expression, MERA tests for sequence necessity. We report on the application of MERA to tile thousands of mutations across ~40 kb of cis-regulatory genomic space. MERA expresses a single guide RNA per cell from a genomic template, and uses knock-in green fluorescent protein (GFP) reporters to read out gene activity. Using this approach, we obtain quantitative information on the contribution of cis-regulatory regions to gene expression. We identify proximal and distal regulatory elements necessary for the expression of four embryonic stem cell–specific genes. We show a consistent contribution of neighboring gene promoters to gene expression and identify unmarked regulatory elements (UREs) that control gene expression but do not have typical enhancer epigenetic or chromatin features. We perform deep sequencing to discover genotypes that are required for gene activity at specific gRNA target sites, and we abstract these genotypes into base pair–resolution functional motifs of regulatory elements.

DIVERSITY OF IMMUNE RECEPTOR REPERTOIRES.

Aleksandra M Walczak

CNRS/ENS, Laboratoire de Physique Theorique, Paris, France

The recognition of pathogens relies on the diversity of immune receptor proteins. Recent experiments that sequence entire immune cell repertoires provide a new opportunity for quantitative insight into naturally occurring diversity and how it is generated. I will show how applying statistical inference to these recent experiments that sequence entire B and T-cell repertoires we can quantify the origins of diversity in these sequences and characterize selection acting on the somatic evolutionary process that leads to the observed receptor diversity.

TRANSCRIPTIONAL REGULATORS COMPETE WITH NUCLEOSOMES POST-REPLICATION

Srinivas Ramachandran[1], Steven Henikoff[1,2]

[1]Fred Hutchinson Cancer Research Center, Basic Sciences, Seattle, WA, [2]Howard Hughes Medical Institute, FHCRC, Seattle, WA

Every nucleosome across the genome must be disrupted and reformed when the replication fork passes, but how chromatin organization is re-established following replication is unknown. To address this problem, we have developed Mapping In vivo Nascent Chromatin with EdU and sequencing (MINCE-seq) to characterize the genome-wide location of nucleosomes and other chromatin proteins behind replication forks at high temporal and spatial resolution. We find that the characteristic chromatin landscape at Drosophila promoters and enhancers is lost upon replication. The most conspicuous changes are at promoters that have high levels of RNA polymerase II (RNAPII) stalling and DNA accessibility and show specific enrichment for the BRM remodeler. Enhancer chromatin is also disrupted during replication, suggesting a role for transcription factor (TF) competition in nucleosome re-establishment. Thus, the characteristic nucleosome landscape emerges from a uniformly packaged genome by the action of TFs, RNAPII and remodelers minutes after replication fork passage.

THE MOBILE ELEMENT LOCATOR TOOL (MELT): POPULATION-SCALE MOBILE ELEMENT DISCOVERY AND BIOLOGY

Eugene J Gardner[1,2], Vincent K Lam[2,3], Daniel N Harris[1,2], Nelson T Chuang[1,2,4], Ryan E Mills[5,6], 1000 Genomes Project Consortium [1], Scott E Devine[1,2,3,4]

[1]Program in Molecular Medicine, University of Maryland Baltimore, Baltimore, MD, [2]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, [3]Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD, [4]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, [5]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, [6]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI

Mobile element insertions (MEIs) represent ~25% of all structural variants in human genomes, and can have a range of effects on the host genome. As such, it is crucial to discover and characterize MEIs along with other forms of genetic variation in projects involving whole genome sequencing. Here, we describe the Mobile Element Locator Tool (MELT), which was developed as part of the 1000 Genomes Project to meet the growing demands of MEI discovery on a population scale. Using both simulations and Illumina genome sequencing data, we demonstrate that MELT outperforms existing MEI discovery tools in terms of scalability, specificity and sensitivity. In addition to using MELT to discover MEIs in over 2,500 diverse human genomes, we also used MELT to study the difference in mobilization patterns between ancient hominids and modern humans. We detected ancient MEIs in the Neanderthal and Denisovan genomes that are not found in chimpanzees or in modern humans, suggesting that *Alu*, L1, and SVA elements were likely active in these ancient hominids. Our data reveal diverse patterns of MEI population stratification in ancient hominids and modern humans that were likely shaped by several factors including source element variation across populations, diverse patterns of MEI production and inheritance, and the introgression of ancient MEIs into modern human genomes. Overall, our data demonstrate that MELT can be used in a variety of experimental settings to perform MEI discovery and to explore biological questions related to MEIs.

T32 DK067872 (NC), R01CA166661 (SED), and R01HG002898 (SED).

CHROMATIN EXTRUSION EXPLAINS KEY FEATURES OF LOOP AND DOMAIN FORMATION IN WILD-TYPE AND ENGINEERED GENOMES

Suhas S Rao[1,2], Adrian L Sanborn[1,3,4], Su-Chen Huang[1], Neva C Durand[1], Miriam H Huntley[1], Andrew I Jewett[1], Ivan D Bochkov[1], Dharmaraj Chinnappan[1], Ashok Cutkosky[1], Jian Li[1,3], Kristopher P Geeting[1], Andreas Gnirke[5], Alexandre Melnikov[5], Doug McKenna[1,6], Elena K Stamenova[1,5], Eric S Lander[5,7,8], Erez Lieberman Aiden[1,3,5]

[1]The Center For Genome Architecture, Baylor College of Medicine, Houston, TX, [2]School of Medicine, Stanford University, Stanford, CA, [3]Center for Theoretical Biological Physics, Rice University, Houston, TX, [4]Department of Computer Science, Stanford University, Stanford , CA, [5]Broad Institute of MIT and Harvard, Cambridge, MA, [6]Mathemaesthetics, Inc. , Boulder, CO, [7]Department of Biology, MIT, Cambridge, MA, [8]Department of Systems Biology, Harvard Medical School, Boston, MA

We recently used in situ Hi-C to create kilobase-resolution 3D maps of mammalian genomes. Here, we combine these maps with new Hi-C, microscopy, and genome-editing experiments to study the physical structure of chromatin fibers, domains, and loops. First, by examining the probability for short chromatin fragments to form a cycle, we show that chromatin is flexible at the kilobase scale, inconsistent with widespread 30-nm fibers in vivo. Next, we find that contact domains are inconsistent with the equilibrium state for an ordinary condensed polymer. Combining Hi-C data and novel mathematical theorems, we show that contact domains are also not consistent with a fractal globule. Instead, we use physical simulations to study two models of genome folding. In one, intermonomer attraction during polymer condensation leads to formation of an anisotropic “tension globule.” In the other, CTCF and cohesin act together to extrude unknotted loops during interphase. Both models are consistent with the observed contact domains and with the observation that contact domains tend to form inside loops. However, the extrusion model explains a far wider array of observations, such as why loops tend not to overlap, and why the CTCF-binding motifs at pairs of loop anchors lie in the convergent orientation. Furthermore, we use the extrusion model to deconvolve loops into simultaneity classes and demonstrate a novel mechanism for domain formation outside loops. Finally, we perform 13 genome-editing experiments examining the effect of altering CTCF sites on chromatin folding. The convergent rule predicts the affected loops in every case. Moreover, the extrusion model predicts in silico the 3D maps resulting from each experiment using only the location of CTCF sites in the wild type. Thus, we show that it is possible to disrupt, restore, and move loops and domains using targeted mutations as small as a single base pair.

ADVANCES IN THE UNDERSTANDING OF MUTATIONAL SIGNATURES IN HUMAN CELLS

Serena Nik-Zainal

Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Mutational signatures are the imprints of the biological processes that have gone awry in human cells. We previously outlined the methods for identifying and quantifying base substitution mutational signatures present in primary human cancers (http://cancer.sanger.ac.uk/cosmic/signatures). Here, using a highly-curated cohort of 560 whole genome sequenced breast cancers, we extend the understanding of mutational signatures to include six novel rearrangement signatures. We further demonstrate the variation in genomic distribution of mutational signatures of breast tissue, relative to replication time and strands, transcriptional strands and chromatin organisation. These mutational signatures distinguish clinical cohorts in breast cancer and have intriguing genomic properties with potential for clinical application as a biomarker.

DISSECTING THE INFLUENCE OF GENOMIC BACKGROUND IN TUMOR MUTATIONS

Ingegerd Elvers[1,2], Jason Turner-Maier[1], Ross Swofford[1], Michele Koltookian[1], Jeremy Johnson[1], Mara Rosenberg[3], Rachael Thomas[4], Gad Getz[3,5,6], Federica di Palma[7], Jaime F Modiano[8], Matthew Breen[9], Kerstin Lindblad-Toh[1,2], Jessica Alföldi[1]

[1]Broad Institute of MIT and Harvard, Vertebrate Genome Biology, Cambridge, MA, [2]Uppsala University, IMBIM, SciLifeLab, Uppsala, Sweden, [3]Broad Institute of MIT and Harvard, Cancer Genomics, Cambridge, MA, [4]North Carolina State University, College of Veterinary Medicine, Raleigh, NC, [5]Harvard Medical School, Pathology, Cambridge, MA, [6]Massachusetts General Hospital, Pathology, Boston, MA, [7]The Genome Analysis Center, Vertebrate and Health Genomics, Norwich, United Kingdom, [8]University of Minneapolis, Animal Cancer Care and Research Program; College of Veterinary Medicine; and Masonic Cancer Center, Minneapolis, MN, [9]University of North Carolina, Lineberger Comprehensive Cancer Center, Chapel Hill, NC

Tumor sequencing generally focuses on the tumor, using the germ-line variation only as a means of filtering for somatic variants. However, germ-line risk variants influence tumor mutation load and landscape. This is exemplified by pediatric tumors generally having a much lower mutation frequency compared to cancers with adult onset. To understand the influence of genomic background on tumor mutations, we sequenced tumor and matched normal tissue from dogs spontaneously developing lymphoma. Due to species and breed creation bottlenecks and artificial breeding for phenotypic factors, dog breeds display strong genetic diversity between breeds but very limited diversity within breeds. Dogs were selected from breeds with differential predisposition to B- and T-cell lymphoma. Canine lymphoma is clinically and histologically comparable to human lymphoma and dogs receive the same chemotherapy, but their cancer progress faster. We showed that tumors from two breeds predisposed to B-cell lymphoma show large overlaps in their top significantly mutated genes, and two breeds predisposed to T-cell lymphoma have very little overlap in their significantly mutated genes. We also showed that while the number of non-synonymous mutations reflected tumor type, the mutation frequency was dependent on breed. Also, when grouping one base pair substitutions into 96 substitution classes (preceding base, mutation, and following base) the relative fraction of substitutions show more similarities within breeds than between breeds with the same tumor type. Hence, dog breeds are a unique resource for understanding the influence of the genetic background on tumor mutation load as well as a tool for understanding mutational signatures. Analysis of breed-specific germ-line and somatic variation suggest differential DNA repair activity in different breeds. We are also investigating how the genetic background correlates with treatment outcome. We expect that this will allow identification of treatment response groups and better dissection of human tumors.

GENETIC CONNECTIONS BETWEEN SCHIZOPHRENIA, AUTISM AND NEURODEVELOPMENT

Tarjinder Singh, Liu He, Jeffrey C Barrett, DDD Project, UK10K Neurodevelopmental Group

Wellcome Trust Sanger Institute, Human Genetics, Hinxton, United Kingdom

Substantial recent progress has been made using exome sequencing to identify genes in which loss-of-function (LoF) rare variants or de novo mutations confer very high risk for autism and severe developmental disorders. These studies, which have until now largely been conducted separately, have revealed that many of the same genes are disrupted in patients with a wide range of diagnoses and presentations. In a recent meta-analysis of schizophrenia exomes, we identified at genome-wide significance the first such gene (SETD1A) for which LoF variants confer substantial risk for schizophrenia, an adult onset neuropsychiatric disorder where the genetic overlap with autism and developmental disorders is less clear. LoF variants in SETD1A were also found to confer risk for severe developmental disorders in children, including some affected by the same two base-pair splice disrupting variant.

Here, we describe a series of analyses based on large-scale genetic datasets that further explores the potential overlap of genetic risk across these disorders, and the phenotypic heterogeneity within them. We jointly analyzed published data from 1,077 schizophrenia exome trios, 4,264 case and 9,343 control exomes, and array-based CNV calls from 1,077 trios, 6,882 cases and 11,255 controls. Across all types of variants (SNVs and CNVs), we show a consistent LoF burden in schizophrenia cases compared to controls, as well as a significant enrichment in a wide range of gene sets implicated in autism and developmental disorders, including those subject to ExAC-based constraint ($p < 1 \times 10-8$), CHD8 binding targets ($p < 1 \times 10-5$), a curated list of dominant diagnostic neurodevelopmental disease genes ($p < 1 \times 10-5$). While previous reports have implicated pathways such as FMRP targets and the postsynaptic density, these new implicated gene sets suggest that at least part of the genetic risk for schizophrenia is shared with these other disorders. Furthermore, in a subset of 119 schizophrenia patients with comorbid learning disability we show this burden is even stronger than in the general schizophrenia population, mirroring previous results comparing low and high IQ autism probands. We will also present a parallel analysis of 4,053 published and 634 unpublished autism trios, showing a differential burden of LoF mutations depending on ascertainment strategy (population based, clinical genetics, and mixed).

In conclusion, we believe the integrated analysis across thousands of exomes from different neurodevelopmental disorders can shed new light on the genetic connections between them.

← Back

RECURRENT NONCODING REGULATORY MUTATIONS IN PANCREATIC DUCTAL ADENOCARCINOMA

Michael Feigin[1,2], Tyler Garvin[3], Peter Bailey[4], Nicola Waddell[5,6], David Chang[4,7,8,9], Shimin Shuai[10], Steven Gallinger[11,12], John D McPherson[12], Sean M Grimmond[4,6], Ekta Khurana[1], Lincoln Stein[10], Andrew Biankin[4,7,8,9], Michael C Schatz[1], David A Tuveson[1,2]

[1]Cold Spring Harbor Laboratory, CSHL, Cold Spring Harbor, NY, [2]Cold Spring Harbor Laboratory, Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, NY, [3]Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY, [4]University of Glasgow, Wolfson Wohl Cancer Research Centre, Glasglow, United Kingdom, [5]QIMR Berghofer Medical Research Institute, QIMR Berghofer Medical Research Institute, Brisbane, Australia, [6]The University of Queensland, Queensland Centre for Medical Genomics, Queensland, Australia, [7]Garvan Institute of Medical Research, The Kinghorn Cancer Centr, Sydney, Australia, [8]Department of Surgery, Bankstown Hospital, Sydney, Australia, [9]South Western Sydney Clinical School, University of NSWLiverpool, Australia, [10]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [11]Mount Sinai Hospital, Lunenfeld-Tanenbaum Research Institute, Toronto, Canada, [12]Division of General Surgery, Toronto General Hospital, Toronto, Canada

Cancer is a disease caused by genetic mutation. While the contributions of coding mutations to tumorigenesis are relatively well known, few studies have detailed alterations in noncoding DNA. Here we describe a new computational pipeline, GECCO (Genomic Enrichment Computational Clustering Operation) to analyze somatic noncoding alterations in 308 pancreatic ductal adenocarcinomas (PDA) and systematically identify commonly mutated putative regulatory regions. We find that recurrent somatic noncoding regulatory mutations are present in PDA but uncommon near canonical PDA genes. Instead, we find that the regulatory mutations are enriched in known PDA pathways, including axon guidance and cell adhesion, as well as novel processes including transcriptional regulation and homeobox genes. We identify mutations in specific protein binding sites that correlate with significant differential expression of proximal genes and reveal two genes (PTPRN2, SLC12A8) with previously unidentified clinical relevance in PDA.

To explore the effects of these candidate mutations on gene expression, we developed an expression modulation score (EMS) that quantifies the strength of gene regulation imposed by each class of regulatory elements. Surprisingly, we find that the strongest regulatory elements are also the most frequently mutated, suggesting a selective advantage for mutations in these regions. Our analysis provides one of the first detailed single-cancer views of noncoding alterations in tumorigenesis, identifies recurrent regulatory mutations as new candidates for diagnostic or prognostic markers, and suggests novel mechanisms for tumor genome evolution.

← Back

DECIPHERING THE GENOME: COMMUNITY DRIVEN APPROACHES

Heidi L Rehm[1,2,3]

[1]Partners Personalized Medicine, Laboratory for Molecular Medicine, Boston, MA, [2]Broad Institute of Harvard and MIT, Boston, MA, [3]Brigham & Women's Hospital, Harvard Medical School, Boston, MA

With the plummeting cost of sequencing, genetic data is becoming increasingly available for use in the diagnosis, treatment and prediction of disease. However, robust and accurate use of genomics in the practice of medicine will require high quality knowledgebases. To address this need, the NIH funded ClinGen to develop authoritative resources to define the clinical relevance of genes and variants for use in precision medicine and research. This includes working closely with the NCBI's ClinVar database to support the deposition and public sharing of variant-level data. As of Apr. 2016, over 500 laboratories had submitted over 125,000 unique interpreted variants. Prior analyses in May 2015 showed that ~11% of variants had interpretations submitted by more than one lab, and of those ~17% were interpreted differently. ClinGen has adopted the new ACMG/AMP-developed guidelines to assist labs in resolving differences in variant interpretation. Through pilot efforts the majority of differences in variant interpretation appear to be resolvable and efforts are underway to resolve these differences. In addition, ClinGen approves expert panels and practice guidelines (marked as 3 and 4 star submissions in ClinVar) to allow the submission of high quality, expert reviewed variant interpretations for reliable use in clinical care. Additional expert panels have been and continue to be formed by ClinGen to expand the expert review process. Although improvements in variant interpretation are critical to the accurate use of genetic information in clinical care, assessment of gene-disease relationships and supporting novel gene discovery are also critical. ClinGen has developed guidelines for assessing the strength of evidence for gene-disease relationships and is supporting numerous projects to curate genes with reported associations to specific diseases. ClinGen rules were applied to the curation of 1400 gene-disease pairs for the BabySeq project allowing improved interpretation of newborn sequencing results. In addition, in collaboration with the Global Alliance for Genomics and Health and the International Rare Disease Research Consortium, the Matchmaker Exchange has launched a federated network for sharing and matching on case-level genomic and phenotypic data to aid in building evidence for candidate gene-disease relationships.
In summary, the open sharing and curation of gene and variant information will be critical to ensure a safe and successful implementation of genomic medicine into clinical care.

Supported by NIH grants U41HG006834, U01HG006500 and U19HD077671.

← Back

DECIPHERING THE NON-CODING REGULATORY LANDSCAPE IN AUTISM SPECTRUM DISORDERS

Jingjing Li[1], Jingtian Zhou[1], Zhihai Ma[1], Minyi Shi[1], Douglas H Phanstiel[1], Guipeng Li[1], Haitao Wang[2], Deurloo Marielle[2], Qing Li[2], Bo Zhou[1,3], Yong Cheng[1], Joachim Hallmayer[3], Alexander Urban[3], Zhong-Ping Feng[2], Mathew Pletcher[4], Michael Snyder[1]

[1]Stanford University School of Medicine, Genetics, Stanford, CA, [2]University of Toronto, Physiology, Toronto, Canada, [3]Stanford University School of Medicine, Psychiatry & Behavioral Sciences, Stanford, CA, [4]Autism Speaks, Genomic Discovery, New York, NY

For decades, technical and cost hurdles have prevented the systematic investigation of non-coding sequences in complex human diseases, and thus our knowledge about autism spectrum disorders (ASD) has been primarily obtained from analysis of protein-coding sequences. We have combined the analysis of whole genome sequencing with global studies of regulatory sequences of human cortical neurons to reveal the regulatory architecture of ASD. Analysis of de novo mutations from whole genome sequencing of 261 autism families revealed the physical proximity of ASD de novo mutations specifically to the cortical expression quantitative loci (eQTLs) of synaptic genes. We performed ATAC-Seq, ChIP-Seq, RNA-Seq and Hi-C experiments on human cortical neurons, which for the first time provided a paranormal view of the regulatory landscape in these cells. We found that ASD de novo mutations preferentially affect regulatory elements, and the associated genes are shared targets of two ASD syndromic factors, CHD8 and PTEN. Analyzing 15 chromatin states across 127 human tissue/cell types revealed a significant enrichment of ASD de novo mutations in active transcription start sites and the perturbed genes implicated in neuron functions; this distribution enabled us to develop a machine-learning algorithm to assess potential ASD risk for a given individual. Taken together, our study for the first time revealed the regulatory landscape in human neurons, demonstrated the importance of the non-coding genome in ASD and provides a general framework for analyzing regulatory mutations for other complex human diseases.

GENETIC BASIS OF INNATE IMMUNITY IN HUMAN MONOCYTES

Sarah Kim-Hellmuth[1], Matthias Bechheim[2], Pejman Mohammadi[1], Veit Hornung*[2], Johannes Schumacher*[3], Tuuli Lappalainen*[1]

[1]New York Genome Center, New York, NY, [2]University of Bonn, Institute for Molecular Medicine, Bonn, Germany, [3]University of Bonn, Institute of Human Genetics, Bonn, Germany

The innate immune system recognizes microbial pathogens by employing an evolutionary conserved set of pattern recognition receptors (PRRs). Despite major advancements in our understanding of how the innate immune system senses the presence of pathogens, the genetic basis for differences in innate immune responses is only poorly defined. In this study, we characterize the dynamics of genetic effects on gene expression in human monocytes activated with diverse PPR ligands. For this purpose we isolated monocytes of 134 individuals and stimulated them with lipopolysaccharide (LPS), muramyldipeptide (MDP) and 5'-triphosphate dsRNA (ppp-dsRNA) to trigger Toll-like receptor 4 (TLR4), Nod-like receptor 2 (NOD2) and retinoic acid–inducible protein I (RIG-I), respectively. We performed transcriptome profiling at three time points (0 min/90 min/6 h) and genome-wide SNP-genotyping. Comparing expression quantitative trait loci (eQTLs) under baseline and upon immune stimulation revealed hundreds of immune response specific eQTLs (iQTLs). We showed that the iQTLs of the three different stimuli are distinct, and correspond to relevant immunological pathways. We further examined the dynamics of genetic regulation on early and late immune response, which revealed three main patterns of action for iQTLs: early-transient, late and prolonged, of which the former is less prevalent than the other two. Analysis of signs of recent positive selection and the direction of the effect of the derived allele of iQTLs on immune response suggested a trend of selection towards aggravated immune response. Finally, we show that SNPs conferring risk to primary biliary cirrhosis, inflammatory bowel disease and celiac disease are immune response eQTLs for novel candidate genes, bringing new insights into the pathophysiology of these disorders in the context of PRR-activation. This work demonstrates the importance of studying genetic variation under pathophysiologically relevant conditions to resolve functional genetic variants and the transcriptional responses associated with disease.

LOSS-OF-FUNCTION MUTATIONS IN *IFIH1* PREDISPOSE TO SEVERE VIRAL RESPIRATORY INFECTIONS IN CHILDREN

Samira Asgari[1], Luregn J Schlapbach[2], Stéphanie Anchisi[3], Christian Hammer[1], Dominique Garcin[3], Jacques Fellay[1]

[1]École Polytechnique Fédérale de Lausanne (EPFL), Global Health Institute, School of Life Sciences, Lausanne, Switzerland, [2] University of Queensland, Paediatric Critical Care Research Group (PCCRG), Mater Research, Brisbane, Australia, [3]University of Geneva, Department of Microbiology and Molecular Medicine, Faculty of Medicine, Geneva, Switzerland

Respiratory viruses are the most common pathogens leading to non-elective admission to Pediatric Intensive Care Unit (PICU) and above 50% of these infections are caused by two viruses: human respiratory syncytial virus and rhinovirus. Dramatic inter-individual differences are observed in the severity of these viral infections. This project aims at identifying and functionally characterization of rare genetic variants conferring unusual susceptibility to common viral respiratory infections in the pediatric population. 120 previously healthy children requiring intensive care support because of a severe viral respiratory infection were prospectively recruited in Swiss and Australian PICU. After exome sequencing, we used a combination of bioinformatics tools for variant calling and annotation and only included in downstream analyses single nucleotide variants (SNVs) and small insertions/deletions (indels) meeting stringent quality criteria. We identified three rare loss-of-function (LoF) variants in *IFIH1*, which encodes a RIG-I-like receptor involved in viral RNA sensing and activation of type1 interferons in the cell. Four study participants carried a rare *IFIH1* splicing variant, rs35732034: one in homozygous and three in heterozygous form. We identified two additional LoF variants in *IFIH1*, only present in heterozygous form, in a total of four study participants: a second splicing variant rs35337543 (N=3) and a stop-gained variant rs3574460 (N=1). Transcriptome analysis supported the loss of wild-type *IFIH1* expression in individuals carrying the splicing variants. Functional testing of the variants demonstrated that the mutated proteins are unable to induce interferon-β, are intrinsically less stable than wild-type IFIH1, and lack ATPase activity. We also observed a dominant negative effect in co-transfection experiments. *In vitro* assays showed significant restriction of human respiratory syncytial virus and rhinovirus in *IFIH1*-transduced Huh7.5 cells compare to wild-type Huh7.5 cells, which lack both IFIH1 and RIG-I. This study demonstrates that LoF variants in IFIH1 result in primary immunodeficiency against respiratory RNA viruses, suggesting a central role for *IFIH1* in the establishment of an efficient response against common viral respiratory infections.

EVALUATION OF MOLECULAR SUBTYPES AND CLASSIFICATIONS IN BREAST AND SKIN CANCER

Yu-Jui Ho, Molly Hammell

Cold Spring Harbor Laboratory, Watson School of Biological Science, Cold Spring Harbor, NY

Outcomes for breast cancer patients vary depending on the cancer types, disease stages, and patients' age. Adequately characterizing breast caner into distinct groups according to their biological function can have a large influence on how physicians treat patients in clinical settings and leads to a direct impact on outcome of the patient.

Here we present an approach for characterizing large cohorts of breast tissue samples collected by The Cancer Genome Atlas (TCGA) through a semi-supervised classification method. Our rationale is to find common gene co-expression patterns that can be used to classify unknown samples, compare with existing molecular subtypes, and evaluate the performance by patient survival outcome. Starting from a list of genes with the most variable expression patterns, we first assess groups that are extracted by decomposition of the expression data. We further identify genes that are essential for classification and explore their functions within each group. Orthogonally, we also apply the same analysis pipeline to another gene list utilizing the information in the co-expression network from the same samples.

Groups classified by using the two extracted gene lists are compared with the predefined molecular subtypes assigned by using PAM50 and shows a high concordance. Results show an enrichment of PAM50 genes in the two extracted gene lists, which confirm the important biological functions of these core genes that define each subtype. Moreover, Kaplan-Meier plots are used to demonstrate difference in patient survival status between groups identified using our new essential gene list. This pattern can be seen in another large cohorts of Melanoma tissue samples as well.

We seek to apply this strategy to other collection of cancer tumor samples. Moreover, we present the whole analysis pipeline as a user-friendly interface and hope this tool can help identify sub- or unknown groups from different kinds of genomics data sets.

JOINT FINE MAPPING OF GWAS AND eQTL DETECTS TARGET GENE AND RELEVANT TISSUE

Farhad Hormozdiari[1], Ayellet V Segre[2], Martijn v Bunt[3,4], Xiao Li[2], Jong Wha J Joo[1], Michael Bilow[1], Jae-Hoon Sul[5,6], Bogdan Pasaniuc[7,8], Eleazar Eskin[1,8]

[1]University of California, Los Angeles, Department of Computer Science, Los Angeles, CA, [2]The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cancer Program, Cambridge, MA, [3]University of Oxford, Oxford, Oxford Centre for Diabetes, Endocrinology & Metabolism, Oxford, United Kingdom, [4]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [5]University of California, Los Angeles, Department of Psychiatry and Biobehavioral Sciences, Los Angeles, CA, [6]University of California, Los Angeles, emel Center for Informatics and Personalized Genomics, Los Angeles, CA, [7]University of California, Los Angeles, Department of Pathology and Laboratory Medicine, Los Angeles, CA, [8]University of California, Los Angeles, Department of Human Genetics, Los Angeles, CA

The vast majority of genome-wide association studies (GWAS) risk loci fall in non-coding regions of genome. One possible hypothesis is that these GWAS risk loci alter the disease risk through their effect on gene expression in different tissues. Detecting which gene is affected in which tissue can aid in understanding an underlying mechanism of a GWAS risk locus. If the same variant responsible for a GWAS locus affects gene expression, then the relevant gene and the tissue may play a role in the disease mechanism. Identifying whether or not the same variant is causal in both GWAS and eQTL studies is challenging due to the uncertainty induced by linkage disequilibrium (LD) and the fact that some loci harbor more than one causal variant. Current methods that address this problem assume there exist at most one causal variant in a locus. In this paper, we provide a new method, eCAVIAR, to compute the quantity we refer to as the colocalization posterior probability (CLPP), which is the probability that the same variant is responsible for both the GWAS and eQTL signal, while accounting for LD. The main advantages of eCAVIAR are that it can account for more than one causal variant in any loci and eCAVIAR leverages summary statistics without accessing the individual genotype data. We use both simulated and real datasets to demonstrate the utility of our method. Utilizing data from the Genotype-Tissue Expression (GTEx) project, we are able to utilize this probability to prioritize the likely relevant tissues and the target gene for a few Glucose and Insulin-related traits loci.

← Back

DISCOVERY OF COMPLEX INVERSIONS AND MUTATIONAL PROPERTIES UNDERLYING THE ORIGIN OF SEGMENTAL DUPLICATIONS

Fereydoun Hormozdiari[1,2], Maika Malig[2], Brad Nelson[3], Mark Chaisson[3], Evan E Eichler[3,4]

[1]UC Davis, Biochemistry and Molecular Medicine, Davis, CA, [2]UC Davis, Genome Center, Davis, CA, [3]University of Washington, Department of Genome Sciences, Seattle, WA, [4]Howard Hughes Medical Institute, Seattle, WA

Recent studies from the 1000 Genomes Project suggest that the majority of predicted inversions are not "simple" inversions involving just reversal of gene order but rather show more complex patterns of deletion and duplication at the breakpoints. Single-molecule, real-time (SMRT) sequencing of human genomes confirms that a large fraction of the predicted inversions are in fact "inverted-duplications" involving a non-tandem duplication of a genomic segment mapping in inverted orientation with respect to the original locus. One haploid genome sample (i.e., CHM1) sequenced by long-read technology revealed a total of 25 inverted-duplications. We developed a novel computational algorithm to accurately predict inverted-duplications and improve our power to predict inversions using Illumina WGS. Our method was able to recover 16 (16/25=64%) of these inverted-duplications with moderately low false discovery rate (25%). Surprisingly, using this method we can also discover a subset of gene conversions that can masquerade as false inverted-duplications or inversions. We applied our method to 10 human genomes sequenced using Illumina and discovered 97 distinct inverted-duplications. Over 80% of predicted inverted-duplications passed our manual inspection as high-quality calls. We randomly selected 33 of these for further sequence resolution. We selected fosmid clones, which spanned the predicted insertion loci of the inverted-duplication, and targeted the loci for sequencing using SMRT technology (i.e., PacBio). Out of the 33 predicted inverted-duplications tested, a total of 26 (26/33=%78) were validated as true inverted-duplications. The majority of the predicted inverted-duplications are small (length ranges from 50 bp to 500 bp) and the new copy maps on average ~2000 bp away from the original copy. In some cases, inverted-duplication segments are large (12 kbp) and mapped 10 kbp from the ancestral locus. Our results provide insight into an understudied mechanism of human mutation that may be responsible for the origin of interspersed segmental duplications in the human genome.

A SCALABLE FRAMEWORK FOR INFERRING FITNESS CONSEQUENCES OF NONCODING MUTATIONS IN THE HUMAN GENOME

Yifei Huang[1], Brad Gulko[1,2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cornell University, Graduate Field of Computer Science, Ithaca, NY

Genome-wide association studies and evolutionary studies suggest that a large proportion of variants important to human diseases, phenotypes, and evolutionary adaptation are located in noncoding regions. However, methods for identifying causal noncoding variants important to human evolution and disease have only achieved limited success. Here we report a scalable framework, L-INSIGHT, for inferring fitness consequences of noncoding mutations by integrating a variety of comparative and functional genomic data. This novel framework is based on a previously developed model, INSIGHT, which was used in the fitCons framework to infer fitness consequences of noncoding mutations in human genome. As a generalized linear regression model with a special likelihood function borrowed from the INSIGHT model, L-INSIGHT significantly improved the scalability of the fitCons method and allows the integration of vastly larger numbers of genomic features. Unlike most existing methods, including GWAVA, DeepSEA, CADD, and FunSeq2, our framework is explicitly defined in evolutionary terms and can distinguish between weak and strong deleterious effects. Using noncoding disease variants curated by the ClinVar and HGMD databases, we show that our new method performs favorably compared to the state-of-the-art.

RARE VARIANT CASE-CONTROL ASSOCIATION STUDIES WITH HETEROGENEOUS SEQUENCING DATASETS

Yao Yu[1], Fulan Hu[1], Jiun-Sheng Chen[1], Shan Chen[1], Hao Hu[1], Aditya S Deshpande[1], Smruthy Sivakumar[1], Yihua Liu[1], Jerry Fowler[1], S Shankaracharya[1], Barry Moore[2], Yuanqing Ye[1], Michelle Hildebrandt[1], Hua Zhao[1], Paul Scheet[1], Xifeng Wu[1], Mark Yandell[2], Chad D Huff[1]

[1]The University of Texas MD Anderson Cancer Center, Department of Epidemiology, Houston, TX, [2]University of Utah School of Medicine, Department of Human Genetics and USTAR Center for Genetic Discovery, Salt Lake City, UT

Whole-exome sequencing data is increasingly becoming available to the research community for secondary analyses, providing new opportunities for studies that directly test rare variant, common disease hypotheses. However, the heterogeneous nature of these datasets is a major barrier to large-scale sequencing association studies that incorporate data from multiple sources. Minor differences in sample preparation and sequencing protocols often result in strong technological stratification biases that overwhelm subtle signals of disease association. These biases can be reduced through the use of joint calling/genotyping and standard quality control procedures. However, these approaches alone typically result in poor signal-to-noise ratios and unacceptably high levels of Type I error inflation in exome sequencing association studies with heterogeneous data sources. To address this problem, we developed XQC, a new toolkit to support high-throughput sequencing association studies that greatly mitigates and in some cases eliminates Type I error inflation resulting from technological stratification. XQC optimizes joint variant calling and recalibration procedures based on the target region of each platform, detects and filters variants influenced by technological stratification biases, and assesses population stratification and residual technological stratification from well-behaved markers. We applied XQC to evaluate the contribution of rare, protein-coding variation to cancer risk by conducting a series of whole-exome case-control studies using VAAST 2.1. The cases consisted of individuals of European ancestry from TCGA involving the following cancer types: breast (783 cases), colorectal (362 cases), melanoma (314 cases), ovarian (272 cases), pancreatic (156 cases), and pheochromocytoma and paraganglioma (144 cases). The controls consisted of 1726 females and 1781 males of European ancestry. Our results replicate many established susceptibility genes at $p < 0.05$ and provide odds ratio estimates for rare missense and nonsense variation in each gene. Our results also provide support for promising new candidates, including multiple genes in the Fanconi anemia pathway.

← Back

NRL MEDIATES WIDESPREAD CHANGES IN THE EPIGENOMIC LANDSCAPE OF MOUSE PHOTORECEPTORS

Andrew E Hughes, Jennifer M Enright, Connie A Myers, Joseph C Corbo

Washington University School of Medicine, Pathology & Immunology, St. Louis, MO

The mouse retina is a specialized portion of the central nervous system composed of dozens of diverse cell types. Despite considerable progress in defining the transcriptional profiles of individual retinal cell types, how cell type-specific gene expression is encoded by regulatory DNA is incompletely understood. To begin to elucidate the complex epigenomic architecture of individual retinal cell types, we isolated photoreceptor subtypes—rods, blue cones, and green cones—and performed chromatin accessibility (ATAC-seq) and transcriptome (RNA-seq) profiling. In this way, we generated genome-wide maps of cell type-specific cis-regulatory elements (CREs) in mouse photoreceptors. Globally, we find that rods have a reduction in open chromatin compared to cones and other cell types, which may reflect the unique organization of rod nuclei in nocturnal mammals. We show that these differences are regulated by the transcription factor neural retina leucine zipper (Nrl) and that knocking out Nrl in rods yields an epigenomic profile nearly identical to endogenous cones. In addition, we find that differentially expressed genes in rods and cones are flanked by differentially accessible CREs. In general, we observe strong enrichments of binding sites for known photoreceptor transcription factors among shared photoreceptor CREs. Nevertheless, differences in simple sequence features do not explain the majority of rod- and cone-specific open chromatin. Taken together, these data provide key insights into the development and transcriptional regulation of photoreceptor identity, and they provide a resource for the future study of cell type-specific cis-regulation in the mammalian nervous system.

DISSECTING THE IMPACT OF POPULATION VARIATION IN DNA METHYLATION ON TRANSCRIPTIONAL RESPONSES TO IMMUNE ACTIVATION

Lucas T Husquin[1,2], Maxime Rotival[1,2], Helene Quach[1,2], Julia L MacIsaac[3], Michael S Kobor[3], Lluis Quintana-Murci[1,2]

[1]Institut Pasteur, Human Evolutionary Genetics, Paris, France, [2]Centre National de la Recherche Scientifique, URA3012, Paris, France, [3]Centre for Molecular Medicine & Therapeutics, University of British Columbia, Vancouver, Canada

Epigenetic marks are emerging as important drivers of phenotypic diversity, particularly disease-related. In this context, although the role of DNA methylation in regulating gene expression is increasingly recognized, the mode in which it impacts variation in gene expression, particularly in response to environmental stimuli, remains largely unexplored. Here, we generated whole-genome DNA methylation profiles, genotyping data and RNA sequencing data from a unique cell type, the monocyte, in 200 individuals of African and European-descent. Specifically, we monitored how methylation patterns in resting cells condition the response of monocytes to various immune stimuli — influenza infection and activation of various TLR pathways (TLR1/2, TLR4, and TLR7/8) — and how such methylation variation correlates with genetic variants (meQTL mapping). We show that African and European populations are clearly separated based on their genome-wide methylation profiles, underlining the importance of ethnic background as a driver of methylation variation. The detection of differentially methylated genes, enriched in functions related to immunity and metabolism, provided novel insights into the inter-population differences of DNA methylation of a major immune cell type. We find that methylation variation in a number of genes highly correlates with the degree of cellular responses to environmental stimuli in a population–specific manner, highlighting the important link between epigenetic variation and host immunity to infection. Moreover, DNA methylation variation at around 10% of the total number of sites tested is associated with at least one meQTL, located in cis or trans of the methylation site. Interestingly, this proportion increases to up to 60% when focusing on DNA methylation sites that correlate with the transcriptional activity of nearby genes. Lastly, we infer the causality behind the detected associations between genetic, epigenetic and transcriptional variation to increase our understanding of the causal order of regulatory events. Collectively, our study provides new insight into the impact that both genetic and epigenetic variation have on the cellular transcriptional response to infection.

PLATYPUS HAS RECOMBINATION HOTSPOTS

Julie Hussin[1], Gang Zhang[1], Elisabeth Batty[1], Hilary Martin[2], Tasman Daish[3], Frank Grutzner[3], Simon Myers[1,4], Peter Donnelly[1,4]

[1]University of Oxford, Wellcome Trust for Human Genetics, Oxord, United Kingdom, [2]University of Cambridge, Wellcome Trust Sanger Institute, Cambridge, United Kingdom, [3]University of Adelaide, School of Molecular and Biomedical Science, Adelaide, Australia, [4]University of Oxford, Departement of Statistics, Oxford, United Kingdom

Ornithorhynchus anatinus (platypus) is a non-placental mammal that occupies a unique place in the mammalian phylogenetic tree, which is reflected in various features of its biology such as a unique reproductive system that combines egg-laying with lactation. The platypus also has a remarkable sex chromosome system with 5 pairs of X chromosomes in females and 5 X and 5 Y chromosomes in males, all 10 chromosomes forming an alternating chain in male meiosis. This raises many questions regarding meiosis, and little is known about meiotic recombination in this unusual system. Here, we characterize recombination genome-wide, by using DMC1 ChIP-Seq to map the sites of double strand breaks (DSBs) that initiate recombination. Using this fine-scale DSB map, we establish for the first time that this mammal has DSB hotspots. We located over 4000 hotspots across the platypus genome, and characterized their fine-scale properties and associated sequence features. These hotspots occur in CpG-rich and high GC content regions. Additionally, hotspots are enriched near particular types of retrotransposon elements, namely members of the most recently active subfamilies of L2 (LINE) and MIR (SINE) elements. These elements appear to be specific to the platypus lineage and have differing distributions across chromosomes. We observed further differences between autosomal and X-linked hotspots, and detected recombination hotspots at pseudoautosomal boundaries. This study provides novel insights into recombination mechanisms in an extant species from the most distinct of the contemporary mammalian lineage.

DE NOVO ASSEMBLY OF MEDAKA FISH GENOME USING SMRT SEQUENCING AND CONSTRUCTION OF CHROMOSOME MAP USING GENETIC MARKERS

Kazuki Ichikawa[1], Jun Yoshimura[1], Koichiro Doi[1], Junko Taniguchi[1], Ryohei Nakamura[2], Atsuko Shimada[2], Masahiko Kumagai[2], Hiroyuki Takeda[2], Shinichi Morishita[1]

[1]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan, [2]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

We constructed the version 1 of medaka (*Oryzias latipes*) reference genome from the Hd-rR strain in 2007, which contained low quality regions and 97,933 sequence gaps. To overcome this deficiency, we reconstructed medaka draft genomes of two inbred strains, Hd-rR and HNI, using Single Molecule Real-Time (SMRT) sequencer PacBio RS II (Pacific Biosciences) and using genetic markers from single nucleotide polymorphisms (SNPs). From the Hd-rR (HNI, respectively) strain, we collected ~109-fold (~65-fold) coverage of ~13.4 (~14.7) million subreads of average length 7338bp (4480bp). We used the FALCON assembler to generate long contigs such that the N50 contig length was ~2.5 Mbp (~1.3 Mbp).
We utilized 2,401 SNP genetic markers to anchor PacBio contigs to the chromosomes using alignment software program ispcr (in-silico PCR), and generated the version 2.1 medaka draft reference genomes (available at http://utgenome.org/medaka_v2/#!Top.md). In the version 2.1, the length of contigs anchored to the Hd-rR and HNI chromosomes were ~690 Mbp and ~657 Mbp, respectively, and the numbers of gaps were dramatically reduced to 399 and 652.
To improve the quality, we corrected erroneous indels by aligning Illumina short reads to the contigs using Pilon. We identified 14 misassembled contigs from the Hd-rR strain (13 from HNI) that contained genetic markers originating from different chromosomes. We correctly anchored misassembled contigs by dividing them according to the information of genetic markers. It was speculated that a large inversion of length >15 Mbp in chromosome 11 was present between Hd-rR and HNI; however, the contigs with the breakpoints of the inversion could not be identified in the version 1. We settled this problem in the version 2.1. Comparison between contigs of the two inbred strains that diverged ~18 million years ago revealed substantial large structural variants such as insertions, deletions, duplications, and inversions.
There still remain contigs of total size ~101 Mbp to be anchored around telomeric or subtelomeric regions. We continue our efforts of anchoring and ordering these orphan contigs using high-resolution sequence motif physical maps using Nanochannel Array (Irys system) from BioNano genomics.

← Back

LINKING GENES TO PHENOTYPES USING GTEX-TRAINED PREDIXCAN ASSOCIATIONS IN 40 HUMAN TISSUES & MILLIONS OF INDIVIDUALS

Alvaro Barbeira[1], Jason M Torres[2], Kaanan P Shah[3], Heather E Wheeler[4], Graeme I Bell[5], Dan L Nicolae[3], Nancy J Cox[5], Hae Kyung Im[3]

[1]Instituto Tecnologico de Buenos Aires, Physics, Buenos Aires, Argentina, [2]The University of Chicago, Committee on Molecular Metabolism and Nutrition, Chicago, IL, [3]The University of Chicago, Genetic Medicine, Chicago, IL, [4]Loyola University Chicago, Biology and Computer Science, Chicago, IL, [5]The University of Chicago, Medicine, Chicago, IL, [6]Vanderbilt University, Vanderbilt Genetic Institute, Nashville, TN

GWAS and sequencing studies have yielded thousands of genetic variants robustly associated genetic complex traits. However, the underlying biology of those associations needs to be further elucidated. To address this issue we have proposed a method called PrediXcan that links these genetic variants with likely causal genes using the correlations between genetically predicted expression levels and phenotypes. Our method tests whether the transcript-level consequences of genetic variation have an effect on a specific phenotype and and is similar in some respect with Mendelian randomization-based approaches. Recently, we have developed an extension called MetaXcan, which can infer PrediXcan results using only summary results from GWAS. The advantages of our approach are that it a) directly tests a biological mechanism, b) provides gene level results, c) provides direction of effects (which can be used to prioritize drug targets since positively correlated genes have the potential to reduce disease risk when knocked down), d) the multiple testing burden is reduced, and e) provides tissue specific results. To implement our approach we have developed prediction models for gene expression in 40 tissues using the GTEx Consortium and Depression Genes Network data. The results are publicly available and ready to be used by the community (http://hakyimlab.org/predictdb/). Stable versions of the software to run individual level and summary statistics based approaches are currently available on a public Github repository (https://github.com/hakyimlab/PrediXcan and https://github.com/hakyimlab/MetaXcan). We have applied our method to all publicly available meta-analysis results including GIANT, DIAGRAM, MAGIC, PGC, among others using transcriptome prediction models for 40 human tissues. We validate our approach by re-identifying established genes as well as new genes in the vicinity of known loci, which can be accessed in http://test.hakyimlab.org/metaxcan-all-preliminary. In addition, we have identified a number of novel genes not previously reported with specific phenotypes.

INTERROGATING THE GENOMIC MECHANISMS OF SCHIZOPHRENIA GENETIC RISK IN THE HUMAN BRAIN

Andrew E Jaffe[1,2], Richard E Straub[1], Jooheon Shin[1], Leonardo Collado Torres[2], Ran Tao[1], Amy Deep-Soboslay[1], Yuan Gao[1], Jeffrey T Leek[2], Thomas M Hyde[1], Joel E Kleinman[1], Daniel R Weinberger[1]

[1]Lieber Institute for Brain Development, Clinical Sciences, Baltimore, MD, [2]Johns Hopkins University, Department of Biostatistics, Baltimore, MD

Background: Recent genome-wide association studies (GWAS) have identified over 100 loci that confer risk for schizophrenia, but potential mechanisms underlying these loci are largely unknown. We therefore sought to better characterize the genomic correlates related to the causes and consequences of schizophrenia in the developing and adult human brain.
Methods: We generated genotype, polyA+ RNA sequencing (RNA-seq), and DNA methylation (DNAm) data on the dorsolateral prefrontal cortex (DLPFC) from 495 samples including 320 non-psychiatric controls across the lifespan, including 50 in the second trimester of fetal life, and 175 adult subjects diagnosed with schizophrenia. We developed a novel approach for modeling RNA degradation to improve differential expression replication comparing patients to controls.
Results: We identified 199 genes with significantly differentially expressed between schizophrenia patients compared to controls after RNA quality correction that replicated in independent samples – these genes were enriched for ion channel activity (FDR<0.001). In addition, genes within the schizophrenia risk loci showed decreased expression in patients compared to controls (p<1e-20). We next identified extensive developmental regulation of transcription, including thousands of novel alternatively-spliced transcripts also present in GTEX and GEUVADIS datasets – the expression levels of these genes diverged less from fetal life in patients with schizophrenia compared to adult controls (p<1e-20), and we found that the subset of developmentally regulated genes that switch expression patterns across the fetal-postnatal developmental transition were more prevalent within schizophrenia risk loci (p<0.01). We identified extensive genetic regulation of nearby DNAm and expression levels across convergent features that demonstrated extensive transcript specificity, and showed that the majority of schizophrenia risk variants associated with nearby expression and DNAm levels in human brain. We lastly found that the majority of developmentally-regulated genes contain a methylation quantitative trait locus (meQTL) containing schizophrenia risk SNPs (OR=4.0, p=2e-8).
Discussion: We highlight one potential mechanism of how schizophrenia genetic risk may impart function in the human brain by suggesting these risk alleles affect local epigenetic regulation in fetal life that likely remain stable through adulthood, providing molecular support underlying the neurodevelopmental hypothesis of this disorder.

← Back

DIRECT DETERMINATION OF GENOME SEQUENCES

David B Jaffe, Neil I Weisenfeld, Vijay Kumar, Kamila Belhocine, Rajiv Bharadwaj, Deanna M Church, Paul Hardenbol, Jill Herschleb, Chris Hindson, Yuan Li, Patrick Marks, Pranav Patel, Andrew Price, Michael Schnall-Levin, Ryan Wilson, Alex Wong, Indira Wu

10X Genomics, Pleasanton, CA

We introduce a new method for determining the genome sequence of an organism. Our method has the following key advantages:

1. Our starting material consists of 1 ng of high molecular weight DNA, as compared to typical requirements of 1,000-10,000 ng or more.
2. We create a single library, as compared to typical methods which require creation of multiple libraries and often multiple data types. This makes our process fundamentally more robust than typical methods.
3. Our costs are about ten times lower.
4. The entire process, including assembly, is not organism-specific. For example, there are no parameters to specify to the algorithm.
5. We produce a genome sequence that mirrors the actual chromosomes in the sample, as contrasted with prior methods for which a contig is a mélange of homologous sequences.

To accomplish this, we create a single 10X Genomics Linked-Read (Chromium Genome) library and sequence it on the HiSeq X instrument. The data type consists of barcoded pools of reads, each originating from several very long molecules, with each molecule represented by many reads, and shallowly covered. Our new turn-key software, the Supernova Assembler, exploits these pools, first to create local assemblies, that can capture difficult regions, and then to phase homologous chromosomes.

Using human genomes, we obtain contigs of size 100 kb, in scaffolds of size larger than 10 Mb, and composed of phase blocks of size 3-4 Mb. Notably, these phase block lengths greatly exceed those obtained from any other method, in spite of being constructed from dramatically less expensive data. We rigorously assess the accuracy of our assemblies, including by means of a HGP sample for which 340 Mb of finished sequence is available. We further demonstrate our method on a wide range of organisms (both animal and plant), obtained from diverse starting materials. For many purposes, our method supplants all prior methods for obtaining genome sequences by providing a direct and inexpensive path to the true sequence of the sample.

STRUCTURAL DIVERSITY, RECOMBINATION AND SELECTION IN THE 4 MB HLA REGION INFERRED FROM 100 *DE NOVO* ASSEMBLED HAPLOTYPES

Jacob M Jensen, Palle Villesen, Rune M Friborg, Mikkel H Schierup

Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

The human leukocyte antigen (HLA) is the most polymorphic region of the human genome and has been associated with a wide range of diseases, particularly autoimmune and infectious diseases. However, pinpointing the causes of these associations is hampered by the structural complexity of the HLA, which is poorly characterized with only eight annotated haplotypes, whereof six are incomplete. Although thousands of classical HLA alleles have been studied in detail, study of complete haplotypes has been impeded by the high costs of haplotype resolved genome sequencing, leaving a gap in our understanding of associations of genetic variants and disease. Here we present 100 fully assembled and annotated HLA haplotypes produced entirely from high coverage short read sequencing. As part of the Danish pan genome project 50 trios were sequenced at 78x per individual with multiple insert size libraries up to 20 kb allowing the HLA region of each individual to be de novo assembled into one or a few scaffolds. We developed a new method to infer the two full HLA haplotypes each individual carries using a combination of alignment of assembly graphs within a trio, transmission- and read-backed phasing and remapping. We present 100 complete HLA haplotypes (<1% N's) with more than 200.000 SNV and structural variants inferred and phased. We then align these 4 Mb haplotypes with the reference genome and use a novel k-mer approach to validate the variants and integrate them into the hg38 coordinates. From this set we infer a full recombination map along the HLA region and provide the most detailed description of diversity across full haplotypes to date along with a detailed catalogue of structural variation in the region providing new insights into HLA evolution.

The Danish Pangenome Consortium: Stephanie Le, Thomas Espeseth, Patrick F. Sullivan, Christian M. Hultman, Lars Bolund, Thomas D. Als, Bent Petersen, Simon Rasmussen, Kirstine Belling, Jose M. G. Izarzugaza, Maria Luisa Matey-Hernandez, Arcadio Rubio, Christian T. Have, Johan V. Beusekom, Karsten Karsten, Lasse Maretty, Anders Krogh, Jonas Sibbesen, Hans Eiberg, Jette Bork-Jensen, Mikkel H. Schireup, Jacob M. Jensen, Laurits Skov, Rune M. Friborg, Christian N. S. Pedersen, Thomas Mailund, Palle Villesen, Siyang Liu, Shujia Huang, Yuqi Chang, Weijian Ye, Junhua Rao, Ruiqi Xu, Jihua Sun, Hao Liu, Xiaosen Guo, Hongzhi Cao, Chen Ye, Ning Li, Xun Xu, Jun Wang, David Westergaard.

LINKING ROLES OF DE NOVO MUTATIONS AND COMMON VARIANTS IN SCHIZOPHRENIA

Peilin Jia, Zhongming Zhao

The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX

Schizophrenia is a chronic and socially disabling disorder whose pathophysiology remains unsolved. Emerging studies have supported the hypothesis that genetic components susceptible to schizophrenia (SCZ) involve a wide spectrum of risk factors, including common variants (CVs), rare variants, and de novo mutations (DNMs) with effect sizes ranging from small to large. However, few overlapping genes were found both impacted by CVs associated with SCZ and by extremely rare DNMs occurred in only SCZ probands. Considering that it is unlikely these two types of variants work independently in unrelated biological processes and lead to the same disease phenotype, we hypothesized that they could congregate on common biological pathways and processes to cause SCZ. To this end, we referred to these variants as network-attacking variants and set to distinguish their interaction perturbations in networks by studying their residing genes. We mapped genes with DNMs (DNMgenes) and genes with CVs (CVgenes) onto the human protein-protein interaction (PPI) network as well as its sub-networks derived by spatial and temporal expression profiles in brain. Our results showed DNMgenes had a pattern of node attacking, where DNMs lead to node removal from the network and destroyed all the interactions of the DNMgenes. In contrast, CVgenes appeared to have a pattern of edge rewire, where some, but not all, of their interactions were impacted, likely due to dysregulated expression by CVs with regulatory roles. Both patterns were replicated in spatiotemporal sub-networks of brain development, especially in the frontal cortex and sub-cortical regions in fetal developmental stage. In addition, DNMgenes and CVgenes were found to be more accessible to each other than to control genes, indicating they were likely involved in common biological processes. We then developed a network-assisted method to link DNMgenes and CVgenes in frontal cortex during the fetal brain developmental stage and built a SCZ-specific module that was enriched with both categories of genes. We found the resultant SCZ-specific module featured with major groups of genes functional in immune, chromosome modification, and neuronal pathways, which were in line with previous studies on the largest GWAS of SCZ. Strikingly, both DNMgenes and CVgenes contributed to these functional pathways, suggesting a core set of common pathways and networks underlying SCZ. In summary, we conducted a systematic investigation of common variants and de novo mutations and revealed their linking roles underlying SCZ.

SINGLE-NUCLEUS TRANSCRIPTOME SEQUENCING OF DIFFERENTIATING HUMAN MYOBLASTS REVEALS THE EXTENT OF FATE HETEROGENEITY

Weihua Zeng[1,2], Shan Jiang[1,2], Xiangduo Kong[3], Nicole El-Ali[1,2], Alexander R Ball, Jr[3], Christopher I-Hsing Ma[3], Naohiro Hashimoto[4], Kyoko Yokomori[3], Ali Mortazavi[1,2]

[1]University of California Irvine, Department of Developmental and Cell Biology, Irvine, CA, [2]University of California Irvine, Center for Complex Biological Systems, Irvine, CA, [3]University of California Irvine, Department of Biological Chemistry, School of Medicine, Irvine, CA, [4]National Center for Geriatrics and Gerontology, Department of Regenerative Medicine, Morioka, Oobu, Aichi , Japan

Myoblasts are precursor skeletal muscle cells that differentiate into fused, multinucleated myotubes . Current single-cell microfluidic methods are not optimized for capturing very large, multinucleated cells such as myotubes, thus biasing results towards smaller cells that are not as differentiated or have adopted an alternative fate. However, sequencing the transcriptome of nuclei should get around cell size bias. Using immortalized human myoblasts, we performed RNA-seq analysis of single cells (scRNA-seq) and single nuclei (snRNA-seq) and found them comparable, with a distinct enrichment for long non-coding RNAs (lncRNAs) in snRNA-seq. We then compared snRNA-seq of myoblasts and myotubes to mononucleated cells (MNCs) that remained unfused upon differentiation. We found that MNC nuclei exhibited heterogeneity compared to myoblasts and myotubes, with the majority of MNCs adopting a distinct mesenchymal state. Primary transcripts for microRNAs (miRNAs) that participate in skeletal muscle differentiation were among the most differentially expressed lncRNAs, which we validated using NanoString. Our study demonstrates that snRNA-seq provides reliable transcriptome quantification for cells that would otherwise be difficult or impossible to with current single-cell platforms. Our results further indicate that the expression changes of many miRNAs can be measured in single-cells by monitoring lncRNAs that encode their primary transcripts.

← Back

IMPROVING MAIZE GENOME RESOURCES USING LONG-READ SEQUENCING TECHNOLOGIES

Yinping Jiao[1], Bo Wang[1], Michael McMullen[2], David Rank[3], Paul Peluso[3], Jason Chin[3], Kelly Dawe[4], Alex Hastie[5], Tiffany Liang[5], Elizabeth Tseng[3], Tyson Clark[3], Andrew Olson[1], Michael Regulski[1], Michael Campbell[1], Joshua C Stein[1], Sharon Wei[1], Richard McCombie[1], Doreen Ware[1,6]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]University of Missouri, Division of Plant Sciences, Columbia, MO, [3]Pacific Biosciences, Inc., Menlo Park, CA, [4]University of Georgia, Athens, GA, [5]BioNano Genomics, San Diego, CA, [6]USDA-ARS, PSNR, Ithaca, NY

A complete and accurate reference genome is imperative for sustained progress in understanding the genetic basis of trait variation and crop improvement in maize. Although the current B73 reference sequence has seen incremental improvements in quality over the last several years, many gaps and misoriented contigs remain due to the complexity of maize genome. To remedy this, we employed Single Molecule Real-Time sequencing technology (PacBio) and NanoChannel Array from BioNano to build the next generation maize reference genome. The de novo assembly of 65X PacBio long reads reached an N50 of 1Mb with 2,908 contigs. With the BioNano genome map, the contigs were scaffolded into 625 hybrid scaffolds with an N50 of 9.6Mb. Using the B73 physical map and IBM genetic map we were able to place 96% of contigs and 99% of the sequence into chromosome-level scaffolds. The number of gaps in the final V4 assembly was reduced from 124,337 in V3 to 2,523, and size estimates for 1,111 of these gaps were obtained from the BioNano map. Total assembly size relative to V3 increased by about 50Mb with most of order and orientation errors corrected. Furthermore, the maize community is in long need of new expression evidence to improve gene annotation, especially for the reliable identification of alternative transcript isoforms. To improve annotation, we have undertaken a maize transcriptome project using PacBio Iso-seq single-molecule long-read sequencing. To gain maximal representation of transcripts we isolated full-length cDNAs from six tissues and fractionated each into six size-range bins (from <1 kb up to 10 kb), before sequencing on the PacBio RS II platform with P6-C4 chemistry. The resultant 111,151 transcripts captured ~70% of already described loci. However, many transcripts provided new evidence to enable the correction of currently misannotated genes, while 3% of transcripts corresponded to novel protein-coding genes not previously described. A large proportion of transcripts (57%) represented novel, sometimes tissue-specific, isoforms of known genes, whose structure could be validated using high-depth Illumina reads generated from matched tissues. This outcome approximately triples the number of alternative transcripts known in maize. Together with the new reference assembly, the new transcriptome evidence promises to vastly improve maize genome annotation, while uncovering the molecular components that control phenotypic variation in this important crop.

TEPEAKS: A TOOL FOR INCLUDING REPETITIVE SEQUENCES IN CHIP-SEQ ANALYSIS

Ying Jin, Yuan Hao, Molly Hammell

Cold Spring Harbor Laboratory, Genomics, Cold Spring Harbor, NY

Over half of human genome is composed of retrotransposons, which are mobile elements that can readily amplify their copy number by replicating through an RNA intermediate. Most of these elements are no longer mobile but still contain regulatory sequences that can serve as promoters, enhancers or repressors for cellular genes. Many of chromatin-associated factors such as transcription factors, histone modifiers, and other DNA binding proteins are known to bind to the repetitive regions of the genome. Chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-seq) is widely used to study the binding patterns of these factors. A key step in ChIP-seq data analysis is to count short reads mapped to a reference genome to identify peak regions. Most existing methods either discard non-uniquely mapped reads or randomly choose one from the multiple alignments. Both strategies reduce the accuracy in determining enrichment in repetitive regions such as regions with retrotransposon insertions. We have developed TEpeaks, a method for identifying ChIP-seq peaks genome-wide that includes the repetitive fraction of the genome as well as uniquely mappable sites. TEpeaks carefully distributes multiply mapped reads using the uniquely mapped reads as a guide and optimizes the assignment by the expectation maximization (EM) algorithm. Moreover, TEpeaks provides multiple normalization options and also includes a module for differential binding analysis to determine differential enrichment statistics at these candidate-binding sites when comparing between different experimental conditions or genotypes.

SCALABLE MULTI-SAMPLE VARIANT CALLER (MULTIVAC) WITH FAST AND EFFICIENT LOCAL DE NOVO ASSEMBLY

Goo Jun

University of Texas Health Science Center at Houston, Human Genetics Center/EHGES, Houston, TX

Assembly-based variant callers often have higher accuracies than conventional alignment-based callers, especially for indels and complex variations. Whole genome *de novo* assembly requires the full assembly information reside in the memory thus not yet very practical for calling more than hundreds of samples together. Local assembly is a hybrid approach that assembles mapped reads within a genomic region; thus computational burden is much alleviated. We developed a multi-sample, local assembly-based variant calling pipeline (MultiVAC) that is scalable to thousands of samples. In MultiVAC, a single colorless de Brujin graph is constructed using sequence reads from all samples for a specified local genomic region. Each node of the graph stores first and second order statistics of squared base read quality, mapping quality, read cycle, and strand information. While constructing the graph, pileup is generated for each sample, in which statistics for reference k-mers are stored according to genomic positions and statistics for non-reference k-mers are stored separately for each non-reference k-mer. In the next step, the graph is pruned by a support vector machine classifier based on quality metrics collected in the first step. Finally, we revisit each pileup file to generate genotype calls for each sample in the assembled graph. Computational scalability is obtained by employing a single colorless graph and k-mer pileups instead of colored de Brujin graph, hence the size of graph stays manageable with the sample size and sequencing depth increases. The pileups are agnostic to the graph structure, hence graphs can be easily merged and pruned without additional procedures, which means graph construction and pileup process can be easily parallelized and it provides additional scalability when samples are processed in batches or in incremental manner. We tested time and memory complexity of MultiVAC on 1Mb chromosome 20 regions from 1000 Genomes Project Phase 3 low-coverage sequences, with sample sizes 1 to 1,000. Runtime for 100 samples was less than 10 minutes with single thread and core setup, and under two hours for 1,000 samples. Peak memory usage was sublinear to the sample size, with 845MB, 7.2GB, and 55GB memory footprints for 10, 100, and 1,000 samples.

EVOLUTIONARY DYNAMICS OF ABUNDANT STOP CODON READTHROUGH IN *ANOPHELES* AND *DROSOPHILA*

Irwin Jungreis*[1,2], Clara S Chan*[1,2], Robert M Waterhouse[1,2,3,4], Gabriel Fields[1], Michael F Lin[5], Manolis Kellis[1,2]

[1]MIT, Computer Science and Artificial Intelligence Lab, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]University of Geneva Medical School, Genetic Medicine and Development, Geneva, Switzerland, [4]Swiss Institute of Bioinformatics, Geneva, Switzerland, [5]DNAnexus , Mountain View, CA

*Co-first author

Translational stop codon readthrough was virtually unknown in eukaryotic genomes until recent developments in comparative genomics and new experimental techniques revealed evidence of readthrough in hundreds of fly genes and several human, worm, and yeast genes. Here, we use the genomes of 21 *Anopheles* species and improved comparative techniques to identify evolutionary signatures of conserved, functional readthrough of 353 *A. gambiae* stop codons and 51 additional *Drosophila* melanogaster stop codons, with several cases of double and triple readthrough including readthrough of two adjacent stop codons, supporting our earlier prediction about the abundance of readthrough in pancrustacea genomes. Comparisons between *Anopheles* and *Drosophila* allow us to transcend the static picture provided within each clade to explore the evolutionary dynamics of abundant readthrough. Rather than readthrough appearing at the birth of a gene, or at the time readthrough became abundant at the root of the pancrustacea lineage, and lasting for the full life of the gene, we find that readthrough has often been gained or lost in existing genes since the *Anopheles-Drosophila* ancestor. We also determine which characteristic properties of readthrough are transient, which predate readthrough, and which are clade-specific. We estimate that there are more than 600 functional readthrough stop codons in *A. gambiae* and 900 in *D. melanogaster*. We find evidence that readthrough is used to regulate peroxisomal targeting in two genes. Finally, we use the sequenced centipede genome to refine the phylogenetic extent of abundant readthrough.

HIGHER MALE THAN FEMALE RECOMBINATION RATE LARGELY CONTROLLED BY MISSENSE VARIANTS IN *RNF212*, *MLH3*, *HFM1*, *MSH5* AND *MSH4* IN CATTLE.

Naveen K Kadri[1], Chad Harland[1,2], Pierre Faux[1], Nadine Cambisano[1,3], Latifa Karim[1,3], Wouter Coppieters[1,3], Sébastien Fritz[4,5], Erik Mullaart[6], Didier Boichard[5], Richard Spelman[2], Carole Charlier[1], Michel Georges[1], Tom Druet[1]

[1]University of Liege, Unit of Animal Genomics, GiGA-R & Faculty of Veterinary Medicine, Liege, Belgium, [2]Livestock Improvement Corporation, Corner Ruakura & Morrinsville Roads, Hamilton, New Zealand, [3]University of Liege, Genomics Platform, GIGA, Liege, Belgium, [4]Allice, 149 rue de Bercy, Paris, France, [5]Université Paris-Saclay, GABI, INRA, AgroParisTech, Jouy-en-Josas, France, [6]CRV BV, P.0.BOX 454, Arnhem, Netherlands

We herein study genetic recombination in three dairy cattle populations from France, New-Zealand and the Netherlands. We identify 2,395,177 crossover (CO) events in sperm cells transmitted by 2,940 sires to 94,516 offspring, and 579,996 CO events in oocytes transmitted by 11,461 cows to 25,332 offspring. When measured in identical family structures, the average number of CO in males (23.3) was found to be larger than in females (21.4). The heritability of global recombination rate (GRR) was estimated at 0.13 in males and 0.08 in females. The genetic correlation was equal to 0.66, indicating that shared variants are influencing GRR in both genders. Haplotype-based genome-wide association studies revealed seven genome-wide significant QTL. Variants identified by next-generating sequencing in 5 Mb windows encompassing the QTL peaks were imputed in order to perform a sequence-based association analysis. For four QTLs, we identified missense mutations in genes known to be involved in meiotic recombination among the most significantly associated variants. The *P259S* variant identified in *RNF212* had already been reported, whereas missense mutations in *MLH3* (*N408S*), *HFM1* (*S1189L*), *MSH5* (*R631Q*), *MSH4* (*C342Y*) and a second in *RNF212* (*A77T*) are new. Surprisingly, variants previously identified in *REC8* were not associated with a QTL detected on BTA10 whereas variants in *RNF212B*, a paralog of *RNF212*, showed much stronger association with the phenotype in this region. This suggests that *RNF212B* might be involved in the recombination process. Most of the identified mutations had significant effects in both genders with three of them accounting each for approximately 10% of the genetic variance in males (the allelic substitution effect being approximately equal to one additional CO per genome). Thus, a large fraction of the genetic variance is associated with missense mutations in genes known to be involved in meiotic recombination. Our results are very different from reports of recombination in other species. For instance, in human, recombination rate is higher in females, distinct variants affect recombination rate in males and females, and the genetic correlation is close to 0, whereas in cattle, we observed a higher recombination rate in males controlled by shared variants effective in both sexes.

RECONSTRUCTING THE EVOLUTIONARY HISTORY OF PRIMATE CENTROMERES USING SINGLE-MOLECULE SEQUENCING

Sivakanthan Kasinathan[1,2], Steven Henikoff[2,3]

[1]University of Washington School of Medicine, Medical Scientist Training Program, Seattle, WA, [2]Fred Hutchinson Cancer Research Center, Basic Sciences Division, Seattle, WA, [3]Howard Hughes Medical Institute, Seattle, WA

Advances in genome assembly have singularly enabled biology in the post-genome era; however, many genome assemblies remain fundamentally incomplete because of the intractability of repetitive regions such as centromeres. We have developed computational and experimental methods for characterizing repeats using single molecule, real-time sequencing, which provides a rich, assembly-independent resource for genomic analyses of repetitive sequences. Using these approaches, we have reconstructed the evolutionary history of centromeres in the primate lineage. In haplorhine primates, which include apes and monkeys, centromeres are embedded in rapidly evolving, tandemly repeated DNA comprised of ~170 bp α-satellite monomers. Tandem arrays of alphoid multimers, termed higher order repeats (HORs), are thought to be the predominant sequence motif at functional centromeres as evidenced by the placement HOR reference models at centromeres in the hg38 human genome assembly. Contrary to this prevailing view, our approach revealed a centromeric architecture wherein functional centromeric cores are characterized by abundant alphoid dimers that give rise to low-abundance peripheral HORs. We also recovered known and novel HORs consistent with evolution of centromeres by unequal cross-over and catalogued extensive centromeric structural variation across the haplorhine primates. These results provide new insights into primate evolution, offer a framework for investigating previously intractable repetitive regions of genomes, and enable scaffolding of highly repetitive regions, bringing us closer to the goal of complete, end-to-end genome assemblies.

LEVERAGING REGULATORY AND GENOTYPE-PHENOTYPE DATA TO DISCOVER AND INTERPRET THE FUNCTION OF HUMAN REGULATORY DNA IN HEALTH AND DISEASE

Aviv Madar[1], Diana Chang[1], Feng Gao[1], Aaron J Sams[1], Yedael Y Waldman[1], Deborah Cunninghame-Graham[2], Timothy Vyse[2], Andrew G Clark[1,3], Alon Keinan[1]
[1]Cornell University, Biological Statistics and Computational Biology, Ithaca, NY, [2]King's College, Genetics and Molecular Medicine and Immunology, London, United Kingdom, [3]Cornell University, Molecular Biology and Genetics, Ithaca, NY

The expanding volume and quality of human genotype-phenotype (G-P) data are amplifying our potential to understand the impacts of DNA sequence variations in normal physiology and disease. It has been argued that polymorphisms in regulatory DNA explain most phenotype variation among humans, e.g. variance in height or familial predisposition to complex diseases. Data now exists that measures the regulatory landscape of individual cell types and tissues in health and disease (e.g. ENCODE). However, it remains a challenge to connect the abundant functional genomics data with the G-P data, thereby illuminating the role of polymorphisms in regulatory DNA in health and disease. Indeed, most of the G-P associations to date are mapped to large DNA elements (commonly ~100 kb) without pinpointing causal polymorphisms. Here we integrate DNase-seq data from immune and immune-unrelated control cell types with GWAS of autoimmune diseases and immune-unrelated complex traits. We show that SNPs in immune-specific DHSs, replicate as well as genome-wide SNPs discovered at 5×10-8, while requiring lower significance, e.g. 1×10-3 for Crohn's disease (CD) and 1×10-6 for rheumatoid arthritis (RA). Consequently, we discover and replicate dozens of additional associations for these two diseases. Using this approach on six autoimmune disease GWAS (CD, RA, UC, lupus, multiple sclerosis, and type 1 diabetes) we identified 422 regulatory region associations with an FDR<0.001. Beyond improved discovery, we used cell type activity patterns (CTAPs) of disease-associated DHSs to provide insight into disease etiology. For example, we found (i) a more prominent B cell-specific regulatory activity in ulcerative colitis (UC) and lupus when compared to related diseases of CD and RA, respectively; (ii) a significant T cell-specific regulatory role for schizophrenia, but a monocyte-specific role in Alzheimer's disease. We also determined *de novo* which TF binding sites (TFBS) underlie specific CTAPs, e.g. the combination of Ets1, Runx1, STAT, and Rorc (the master regulator of Th17 T cells) is unique to human Th17-specific regulatory elements. We show that this approach can discover novel TFBSs for under-studied cell types and tissues. For example, we find two clear DNA motifs underlying fetal brain specific DHSs that do not match any known consensus TFBS. Finally, we used the identified TFBSs for each CTAP in combination with the common SNPs that intersect them, to provide testable hypotheses at base-pair resolution for hundreds of significant regulatory associations with human diseases. We conclude that incorporating functional genomics data (in this case DHSs), augmented by DHS-harbored TFBSs and SNPs as well as the cell types in which DHSs are active, can aid in the discovery and interpretability of the function of human regulatory elements in health and disease.

← Back

CHARACTERIZATION OF A LARGE VERTEBRATE GENOME AND SEX CHROMOSOMES USING SHOTGUN AND LASER-CAPTURE CHROMOSOME SEQUENCING

Melissa C Keinath[1], Stephen R Voss[1,2], Jeramiah J Smith[1]

[1]University of Kentucky, Biology, Lexington, KY, [2]University of Kentucky, Spinal Cord and Brain Injury Research Center, Lexington, KY

The salamander *Ambystoma mexicanum* (Mexican axolotl) is an important model system in regeneration, development and genome/chromosome evolution studies. All salamander genomes are greatly expanded in size relative to other extant tetrapods. Despite the size of the axolotl genome, its gene orders are highly conserved with those of reptilian and mammalian genomes and provide valuable information for reconstructing several features of ancestral tetrapod genomes. As such, the axolotl genome provides both a unique perspective on the evolution of genome biology and presents a major challenge toward the development of genomic resources. To characterize *A. mexicanum* genome structure, we sequenced whole genomic DNA (600 billion bases). Analysis of the shotgun dataset estimates the genome to be ~32Gb with repetitive sequences making up ~40%. As part of a multipronged approach to reduce assembly complexity, we developed strategies to capture, amplify and sequence individual laser-captured chromosomes (dyads). Initial analyses revealed that resulting sequence libraries provide high sensitivity and specificity to the linkage groups of the *Ambystoma* linkage map, demonstrating that chromosome-targeted sequencing presents an efficient strategy for simultaneously reducing assembly complexity and generating broad-scale scaffolding information for large genomes. In addition, these data helped refine the existing map and resolve key events in the evolution of salamander and chicken genomes.

We highlight the use of this technique in characterizing a recently evolved sex chromosome in *A. mexicanum*. Libraries developed from this chromosome have been used to create a preliminary assembly for the axolotl sex chromosome and to identify candidate sex-linked regions that vary in copy number between males and females. Conserved synteny studies with chicken corroborate previous analyses, and integration of a recently constructed linkage map for the newt (*Notophthalmus viridescens*) has allowed us to identify the corresponding chromosome in a second salamander lineage. While a ZW-type mechanism for sex determination is known for both species, it is not yet known if the homologous newt chromosome contains the newt sex-determining locus.

These studies suggest that the laser-capture sequencing approach can be readily adapted to address a broad range of biological questions, including chromosomal scaffolding of genomes for organisms that are not amenable to laboratory culture and genomic characterization of microscopically identifiable cells (e.g. cancer or germ cells).

ASSOCPLOTS: A PYTHON PACKAGE FOR STATIC AND INTERACTIVE VISUALIZATION OF MULTIPLE-GROUP GWAS RESULTS

Ekaterina A Khramtsova[1,2], Barbara E Stranger[1,2]

[1]The University of Chicago, Department of Medicine, Section of Genetic Medicine, Chicago, IL, [2]The University of Chicago, Institute for Genomics and Systems Biology, Chicago, IL

Background: Over the last decade, genome-wide association studies (GWAS) have generated vast amounts of analysis results, requiring development of novel tools for data visualization. Quantile-quantile plots and Manhattan plots are classical tools which have been utilized to present GWAS results and identify variants significantly associated with traits of interest. However, static visualizations are limiting in the information that can be shown. Recently, dynamic, interactive visualization has become more widely adopted, however it has not yet become a routine part of GWAS data analysis. Interactive data visualization not only allows clearer representation of multidimensional data, but also encourages viewer's engagement from simple data browsing to providing a platform for answering specific scientific questions, in ways that static data cannot. Here we present a package for viewing GWAS results not only using classical static Manhattan and quantile-quantile plots, but also through interactive extension which allows to visualize data interactively: zoom into dense regions, quickly obtain underlying details (e.g. SNP rs number or gene name, base pair position, p-value) by selecting a peak of interest, and visualizing the relationships between GWAS results from multiple groups. Furthermore, this platform facilitates comparison of GWAS results from multiple groups: (1) multiple phenotypes in a single group of individuals, (2) a phenotype measured among distinct cohorts, (3) expression quantitative trait loci measured across different tissues or cohorts, and (4) various experimental conditions such as before and after drug treatment, to name a few examples. Our tool makes it possible to browse multiple charts in real-time to better understand the relationships among multiple groups.
Implementation: Assocplots is implemented as a package for the Python programming language. Its basic functionality includes plotting interactive data visualization for viewing in the browser as well as static publication quality plots. The package is designed to be used both in Jupyter notebooks and in command line. Visualizing GWAS data in a web-based document (notebook), ensures data analysis reproducibility and makes it conveniently sharable with collaborators via online repositories such as GitHub. The assocplots package is open source and distributed via GitHub (https://github.com/khramts/assocplots) along with examples, documentation and installation instructions.

HISAT-GENOTYPE: A PRACTICAL APPROACH FOR ANALYZING HUMAN GENOMES ON A PERSONAL COMPUTER

Daehwan Kim[1], Steven L Salzberg[1,2]

[1]Johns Hopkins University School of Medicine, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [2]Johns Hopkins University, Biomedical Engineering, Computer Science, and Biostatistics, Baltimore, MD

Advancements in sequencing technologies and computational methods have enabled rapid and accurate identification of genetic variants in the human population. Many large-scale projects such as the 1000 Genomes Project, GTEx, and GEUVADIS have already yielded a large and growing amount of information about human genetic variation, including >110 million SNPs (in dbSNP) and >10 million structural variants (in dbVar). Although these variants represent a valuable resource for genetic analysis, computational tools do not adequately incorporate the variants into genetic analysis. For instance, >3,000 alleles of the HLA-A gene have been identified. Representing and searching through the numerous alleles of even one gene has been a challenge requiring a large amount of compute time and memory. Most methods have thus focused on genotyping one or a few genes, and analyzing whole genomes has been a formidable task.

To address these challenges, we have recently developed a novel indexing scheme that captures a wide representation of genetic variants and has low memory requirements. We have built a new alignment system, HISAT2, that enables fast search through the index. Beyond interrogating a few genes at a time, HISAT2 has the potential to genotype essentially all the genes on the human genome on a desktop within a few hours. To demonstrate the capability of our initial genotyping work, we chose two gene families: (1) Human Leukocyte Antigen genes (HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, HLA-DRB1), which are among the most diverse human genes, and (2) the breast cancer-related genes BRCA1 and BRCA2, which have a large number of known variants. The IMGT/HLA Database encompasses >12,900 alleles of the HLA gene family. The ClinVar database provides >8,400 variants (SNVs and indels) of the BRCA gene family. We incorporated these alleles and variants into our index of the human genome, requiring only a small addition in computational resources. Tests on Illumina's Platinum Genomes data showed that our method correctly identifies all 204 alleles of the six HLA genes for the 17 genomes, at a speed surpassing other currently available methods.

Because our system works well for these highly diverse genes, we anticipate it would be relatively straightforward to extend it to many, perhaps all, known variants in human genes. Based on the HLA genes, the BRCA genes, and other well-studied genes soon to be included in our index, we will provide templates that researchers who have domain knowledge and expertise can easily use to input their own custom genotyping routines into our platform. We also plan to create an online collaborative hub to encourage researchers around the globe to work together in developing the platform. Instead of genotyping one gene at a time, this platform will eventually allow for genotyping >20,000 genes within just a few hours on a personal computer.

A GENE-ENVIRONMENT INTERACTION BETWEEN COPY NUMBER BURDEN AND EXPOSURE TO TOBACCO SMOKE ASSOCIATED WITH TOTAL CHOLESTEROL

Dokyoon Kim[1,2], Anastasia Lucas[2], Molly Hall[3], Shefali S Verma[1], Yuki Bradford[2], Peggy Peissig[4], Murray Brilliant[5], Marylyn D Ritchie[1,2]

[1]Geisinger Health System, Biomedical & Translational Informatics, Danville, PA, [2]Pennsylvania State University, Department of Biochemistry & Molecular Biology, University Park, PA, [3]University of Pennsylvania, Institute for Biomedical Informatics, Philadelphia, PA, [4]Marshfield Clinic Research Foundation, Biomedical Informatics Research Center, Marshfield, WI, [5]Marshfield Clinic Research Foundation, Center for Human Genetics, Marshfield, WI

Hypercholesterolemia is associated with many other diseases such as coronary heart disease. Hypercholesterolemia or lower levels of cholesterol are associated with copy number variations (CNV). However, most previous studies were focused on a selected list of CNVs or genes within CNV regions. In addition, the relative contributions of genetic factors, environmental factors and the interactions between them and its association to total cholesterol levels are poorly understood. In this study, we have examined the relative contribution of CNV (measured as total base pairs of copy number burden), environmental exposures, and the interaction between environmental measures and copy number burden in a median total cholesterol phenotype for about 3,000 samples, extracted from the electronic health records (EHR) from the Marshfield Clinic's Personalized Medicine Research Project (PMRP). Through the CNV burden analysis, which was adjusted for age, sex, and the first three principal components, duplication and total CNV burden were significantly associated with the cholesterol phenotype, P =0.0023, P = 0.0099, respectively. Furthermore, a significant and sizable interaction was found between duplication burden and exposure to tobacco smoke (Bonferroni corrected P = 0.0075). The interaction between total CNV and exposure to tobacco smoke is also associated with total cholesterol level (Bonferroni corrected P = 0.01). The overall implication of our findings is that significant gene-environment interactions associated with total cholesterol level exist and could account for a considerable level of heritability not detected by evaluating DNA variation or environment alone. The replication using independent data set is warranted.

LANDSCAPE OF KINASE FUSION GENES BASED ON KINASE DOMAIN RETENTION ACROSS 13 MAJOR CANCER TYPES

Pora Kim, Peilin Jia, Zhongming Zhao

The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX

Kinase fusion genes (KFGs) are promising targets in the development of molecular targeted therapy in cancer. So far, a systematic view of KFGs in cancer genomes based on their kinase domain retention status - an important feature for fusion gene's function – has not been available. In this study, we explored various features of 914 kinase fusion genes covering 312 kinases across 13 major cancer types using The Cancer Genome Atlas (TCGA) data. In these analyses, we specifically focused on kinase domain retention. Based on kinase domain retention, we classified fusion genes into 229 kinase domain retained (KDR) and 269 non-retained (nonKDR) fusion genes. We found more kinase domain retention in the acceptor kinase fusion genes (i.e., 3' KFGs), than that in the donor kinase fusion genes (5' KFGs), and identified several conserved features in 3' KFGs. The results suggested the critical roles of 3' KFGs in cancer cell proliferation. Interestingly, the partner genes of 3' KDR fusion genes were enriched in 'uncontrolled phosphorylation' pathway, indicating a possible synergistic effect of 5' and 3' genes on tumorigenesis by uncontrolled signaling transduction. We further scanned the intronic sequences near the exon junction break points and found 3'KDR's motif that was highly enriched in the 'negative regulation of signal transduction' pathway and 5'KDR's motif was listed in the vial microRNA candidate hairpin structure sequence. To quantitatively measure the recurrence of fusion kinases, we introduced a Degree of Frequency (DoF) score. Based on this scoring method, we found that kinases with higher DoF scores tended to have strong gene expression change at the break point. We further interrogated all these features in a pan-cancer KDR fusion gene network. Surprisingly, we found that the top seven DoF kinases were all in thyroid carcinoma, while fusion genes in thyroid carcinoma and prostate cancer were only enriched in 'activation of MAPKK activity' pathway. In summary, our systematic analysis of kinase fusion genes provided some important insights into the mechanisms of gene fusion in cancer genomes.

THE NCBI ASSEMBLY DATABASE: A RESOURCE FOR FINDING, BROWSING AND DOWNLOADING GENOME ASSEMBLY DATA

Paul A Kitts, Avi Kimchi, Jinna Choi, Vichet Hem, Mark Johnson, Terence D Murphy, Kim D Pruitt, Robert G Smith, Françoise Thibaud-Nissen

National Center for Biotechnology Information (NCBI), NLM, NIH, Bethesda, MD

The NCBI Assembly Database (www.ncbi.nlm.nih.gov/assembly/) serves the vital function of providing stable accessioning and data tracking for genome assembly data. The Assembly database is also a convenient resource for finding, browsing and downloading genome assembly data. This presentation will highlight recent improvements that increase the utility of this resource.

We have developed a prototype web interface for the Assembly database that allows bulk downloading of genome assemblies. Investigators can use a variety of filters to select assemblies of interest and then download the file format of choice for all assemblies in the selected group. For example, all RefSeq bacterial complete genome assemblies can be selected and then their annotation in GFF format can be downloaded. This tool not only replaces the very limited number of prepackaged genome data sets that NCBI previously provided on our FTP site, but adds the flexibility for investigators to generate customized genome data sets tailored to their particular needs.

We have made it easier to get to annotation summary reports and to an interactive genome annotation viewer by adding links from the assembly details page for those genome assemblies annotated by the NCBI Eukaryotic Genome Annotation pipeline. We also provide a link to the interactive genome viewer for many other eukaryotes with chromosome-level genome assemblies in RefSeq.

We also enriched the metadata provided for each genome assembly by displaying common names for the organism or organism group in addition to the scientific name for the organism. In addition, we also show if the sequences in the genome assembly were derived from type material. Knowing which bacterial genome assemblies are from type material helps to recognize genome assemblies that have been incorrectly identified or cross-contaminated[1].

Finally, we now report the reasons a bacterial assembly was not selected for the NCBI Reference Sequence (RefSeq) project. Providing this information allows users to review the reasons and decide for themselves if they would follow RefSeq and ignore the bacterial assemblies that RefSeq skipped, or if some of the reasons would not affect their intended use of the genome assembly. A few reasons, such as chimeric, mixed culture and unverified source organism, are likely to be of concern to all users, hence, we display these reasons as "assembly anomalies".

1. Federhen, S. *et al.* Meeting report: GenBank microbial genomic taxonomy workshop (12–13 May, 2015). *Standards in Genomic Sciences* 11, 15 (2016)

HEADING FOR NEW SHORES: HIGH-RESOLUTION ANALYSIS OF DNA METHYLATION IN YET UNSEQUENCED SPECIES

Johanna Klughammer[1], Paul Datlinger[1], Dieter Printz[2], Nathan C Sheffield[1,5], Matthias Farlik[1], Johanna Hadler[1], Christoph Bock[1,3,4]

[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Medical Epigenomics Laboratory, Vienna, Austria, [2]Children's Cancer Research Institute, FACS Core Unit, Vienna, Austria, [3]Medical University of Vienna, Department of Laboratory Medicine, Vienna, Austria, [4]Max Planck Institute for Informatics, Computational Biology & Applied Algorithmics, Saarbrücken, Germany, [5]Stanford School of Medicine, Dermatology, Stanford, CA

The genome-wide study of DNA methylation at cytosines is crucial for understanding fundamental biological processes such as animal development, cellular differentiation, and maintenance of cellular identity. Assessing DNA methylation in a large variety of species and in wild populations is expected to provide insights into evolutionary processes, environmental adaptation, and mechanisms of epigenetic regulation.
To enable cost-efficient, genome-wide, high-resolution, and high-throughput analysis of DNA methylation in non-model organisms and in species that lack a high-quality reference genome, we developed an integrated approach for studying DNA methylation differences independent of a reference genome.
Experimentally, our method relies on an optimized 96-well protocol for reduced representation bisulfite sequencing (RRBS), which we have validated in nine species (human, mouse, rat, cow, dog, chicken, carp, sea bass, and zebrafish). Bioinformatically, we developed the RefFreeDMA software to deduce ad hoc genomes directly from RRBS reads and to pinpoint differentially methylated regions between samples or groups of individuals. Code and documentation along with an example data set are freely available (http://RefFreeDMA.computational-epigenetics.org).
After validating our method, we now start to apply it to a wide range of projects including epigenome-wide association studies in wild populations of passerine birds, the co-evolution of genome and epigenome, and the effect of different aquaculture rearing methods on the epigenome of food fish.

DNA EDITING OF RETROELEMENTS BY APOBECS: A SOURCE OF GENOMIC SEQUENCE DIVERSITY AND ACCELERATED EVOLUTION

Binyamin A Knisbacher, Erez Y Levanon

Bar-Ilan University, The Mina & Everard Goodman Faculty of Life Sciences, Ramat-Gan, Israel

Genome evolution is commonly viewed as a gradual process that is driven by random mutations that accumulate over time. However, APOBECs are DNA editing enzymes that can accelerate evolution by actively modifying the genomic information. The APOBECs are key players in innate immunity that can inhibit retroelements by C-to-U (cytidine to uridine) editing of retroelement DNA after reverse transcription. In some cases, a retroelement may successfully integrate into the genome despite being hypermutated. Such events introduce unique sequences into the genome and are thus a source of genomic innovation. In this study, we screened endogenous retrovirus (ERV) sequences in the genomes of 123 diverse species and identified hundreds of thousands of mutations caused by DNA editing in multiple vertebrate lineages, including humans, many additional placental mammals, marsupials and birds. Numerous edited ERVs carry high mutation loads, some with >350 edited sites, profoundly damaging their ORFs. For many of the species studied, this is also the first evidence that APOBECs are active players in their innate immune system. Unexpectedly, some birds and especially zebra finch and medium-ground finch (one of Darwin's finches), are exceptionally enriched in DNA editing. We demonstrate that edited retroelements may be preferentially retained in active genomic regions, as reflected from their enrichment in genes, exons, promoters and transcription start sites, thereby raising the probability of their exaptation for novel function. In conclusion, DNA editing of retroelements by APOBECs has a substantial role in vertebrate innate immunity and may boost genome evolution.

Initial results are in press in Molecular Biology and Evolution (Knisbacher et al., 2016).

CONTEXT-SPECIFIC eQTLS IMPLICATE DIET-INDUCED TRANSCRIPTIONAL CONTROL IN OBESITY

Arthur Ko[1,2], Elina Nikkola[1], Rita M Cantor[1], Mete Civelek[3], Aldons J Lusis[1,4], Johanna Kuusisto[5], Michael Boehnke[6], Karen L Mohlke[7], Markku Laakso[5], Paivi Pajukanta[1,2]

[1]UCLA, Dept. of Human Genetics, Los Angeles, CA, [2]UCLA, Molecular Biology Institute, Los Angeles, CA, [3]University of Virginia, Center for Public Health Genomics, Charlottesville, VA, [4]UCLA, Dept. of Medicine, Los Angeles, CA, [5]University of Eastern Finland, Dept. of Medicine, Kuopio, Finland, [6]University of Michigan, Dept. of Biostatistics and Center for Statistical Genetics, Ann Arbor, MI, [7]University of North Carolina, Dept. of Genetics, Chapel Hill, NC

Obesity is a serious risk factor for cardiometabolic disease but GWAS variants explain only 2.7% of the variance of body mass index (BMI). Environment and gene-environment interactions (GxEs) may also contribute to the world-wide epidemic of obesity. We hypothesized that the cellular obese environment modifies regulatory DNA variants, leading to context-specific effects on gene expression that can be mapped as context-specific expression quantitative trait loci (cseQTL). We investigated 3 clinically different obesity traits, BMI, BMI-adjusted fat mass percent (fat%AdjBMI), and BMI-adjusted waist-to-hip ratio (WHRAdjBMI) for transcriptome-wide trait-SNP interaction effects on gene expression. We used ~8.9M variants and adipose RNA-sequence data on 14,763 expressed genes in 789 men from the Finnish METSIM cohort to test the significance of cis (±1Mb) (FDR<5%) and trans (>1Mb) ($P<10^{-12}$) trait-SNP interaction term on expression. Of the identified 70 cis and 217 trans genes, only 6 trans and no cis genes were shared across the 3 obesity traits, reflecting their different phenotypic importance. Notably, 52.7% of the cis cseQTLs do not display a nominal eQTL effect ($P<0.05$), suggesting that they only act in an obesity-dependent manner. As 11 of the obesity cseQTL genes (BMP4, BSCL2, DHGR, EIF2A, FSTL3, GHR, LCOR, NLRP3, NOD2, PDCD4, and RASD1) have previously been implicated for high fat diet-induced obesity in mouse models, we tested whether dietary fat also affects context-specific expression of these genes in humans. Using erythrocyte membrane fatty acids as proxies for long-term dietary fatty acid intake, we identified specific fatty acid content of diet as a significant interaction factor (Bonferroni corrected $P<0.05$) for expression of two trans genes, BMP4 and GHR. BMP4 mediates changes of white adipose tissue to brown fat, improving glucose and energy homeostasis, and GHR serves as a receptor for growth hormone secreted by the pituitary gland in brain. In summary, we demonstrate cseQTLs as a form of GxEs in obesity, and the expression of two obesity cseQTL genes BMP4 and GHR significantly interacting with the fat content of diet.

← Back

ALU ELEMENTS IN BABOONS: RAPID EXPANSION AND EVOLUTIONARY INSIGHTS

Vallmer E Jordan[1], Cody J Steely[1], Thomas O Beckstrom[1], Jerilyn A Walker[1], Emily Bennett[1], Brooke Clement[1], Arinna Robichaux[1], Mark A Batzer[1], Miriam K Konkel[1], for the Baboon Genome Sequencing and Analysis Consortium[2]

[1]Louisiana State University, Department of Biological Sciences, Baton Rouge, LA, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

Baboons (genus *Papio*) are widespread across sub-Saharan Africa as well as the southern Arabian Peninsula. The history of baboon speciation appears to be complex, based on a discrepancy in mitochondrial and phenotypic data with evidence for introgressive hybridization. Baboon speciation is estimated to have begun 2-3 million years ago. A consensus of six distinct baboon species has recently emerged: olive, yellow, Guinea, hamadryas, chacma, and kinda. However, the evolutionary relationships among baboons continue to be controversial, likely in part due to the presence of stable hybrid zones where baboons meet in the wild and produce fertile offspring. The baboon-macaque split has a comparable time depth to human-chimpanzee-gorilla, allowing for direct comparisons of mobile element mobilization rates (among others) between different clades. For our comparative genomics analyses, we used the *Papio anubis* (olive baboon) draft genome assembly [papAnu2]. Our analyses support a rapid expansion of *Alu* elements in the lineage leading to *P. anubis*, with more than 45,000 *Alu* insertions specific to the baboon lineage compared to the rhesus macaque genome, suggesting a roughly 9-fold higher insertion rate compared to human-specific *Alu* insertions. *Alu* elements have been shown to be nearly homoplasy-free with a known ancestral state, thus making them ideal genetic systems for phylogenetic and population genetic analyses. To investigate the population and phylogenetic relationships within *Papio*, we analyzed a panel of 80 baboons encompassing all six baboon species. In our analyses, we included 412 polymorphic *Alu* insertions selected from the baboon reference assembly, as well as from high-throughput sequencing data of a diversity panel that contained individuals of six species within *Papio*. Our PCR analyses reveal a high degree of polymorphism across different baboon species, and suggest extensive incomplete lineage sorting, recent speciation, and/or ongoing hybridization. Our data support that chacma baboons diverged first, and that olive and yellow baboons share varying degrees of admixture. We also investigated the population structure within different baboon species. For example, our Structure analysis indicates three distinct population clusters within kinda baboons. In summary, our analyses further support the complex relationships among baboons, and provide evidence for mobile element expansion possibly linked to inter-species hybridization.

← Back

UNCOVERING HIDDEN FUNCTIONAL VARIATION IN POLYPLOID WHEAT

Ksenia V Krasileva[1,2], Hans Vasquez-Gross[3], Paul Bailey[1], Francine Paraiso[3], Leah Clissold[1], James Simmonds[4], Xiaodong Wang[3], Tyson Howell[3], Ricardo Gamirez-Gonzalez[1], Christine Fosker[1], Andy Phillips[5], Sarah Ayling[1], Cristobal Uauy[4], Jorge Dubcovksy[3,6]

[1]The Genome Analysis Centre, Genomics, Norwich, United Kingdom, [2]The Sainsbury Laboratory, TSL, Norwich, United Kingdom, [3]University of California, Plant Sciences, Davis, CA, [4]The John Innes Centre, Crop Genetics, Norwich, United Kingdom, [5]Rothamsted Research, Plant Biology and Crop Science, Harpenden, United Kingdom, [6]Howard Hughes Medical Institute, HHMI, Chevy Chase, MD

Domesticated over 10,000 years ago at the Fertile Crescent, wheat is now grown throughput the world and occupies more agricultural land than any other single crop. Wheat functional genomics have been challenged by large genome size and polyploidy of cultivated wheat. Wheat is a young polyploid species, and most of the genes are represented by more than one functional copy (homoeologs). This redundancy masks the effects of naturally occurring mutations, which are frequently missed during the selection of improved varieties. Fortunately, gene redundancy also facilitates the generation of very high densities of induced mutations allowing to sample relatively small populations to find any allele of interest.

Using Nimblegen's exome capture platform together with optimized robotic protocols, we sequenced over 3,000 mutant lines of bread wheat *Triticum aestivum* cv. Cadenza (ABD genome) and pasta wheat *Triticum turgidum* cv. Kronos (AB genome). In each line, we identified all possible mutations in the reference genes and performed computational prediction of the mutation effect on protein function. In total, we uncovered unprecedented 11,500,000 mutations in the coding regions of wheat genes (~35 mutations per kb/population).

As a reverse genetics toolbox, this dataset provides access to loss-of-function mutations in most wheat homoeologs. The combination of mutations in the different homoeologs has an enormous potential to reveal phenotypic variation that was hidden before in polyploid wheat. We are currently combining the information available in these populations with the newest wheat genome assembly resources to discover mutant wheat lines with abnormal physiological, genetic and genomic phenotypes.

SUPPORTED LIPID BILAYERS TO TURN GENOMIC SCIENCE INTO MATERIALS SCIENCE

Sam Krerowicz[1,2]

[1]Laboratory for Molecular and Computational Genomics, UW Biotechnology Center, Madison, WI, [2]UW-Madison, Chemistry, Madison, WI

The engineering mentality that guides today's thinking in the construction of DNA nano-structures and materials is heavily dependent on DNA design principles set down over the past few decades by Ned Seeman, Erik Winfree and Paul Rothemund. Such thinking has culminated in the development of "DNA origami," which makes heavy use of long, single-stranded DNA molecules. Using the bioinformatic tools and molecular modalities developed by LMCG we are developing a modern approach to the construction of very large-scale objects made of DNA that also offer novel routes to dynamic action and control through the use of supported lipid bilayers.

THE ORIGINS OF CHIMPANZEE DIVERSITY

Marc de Manuel*[1], Lukas Kuderna*[1], Peter Frandsen[2], Vitor C Sousa[4], Tariq Desai[5], Chimpanzee Genome upgrade and diversity Consortium[1], Aylwyn Scally[5], Laurent Excoffier[4], Lars Feuk[6], Andrew Sharp[7], Chris Tyler-Smith[8], Yali Xue[8], Christina Hvilsom[9], Wesley C Warren[3], Tomas Marques-Bonet[1,10,11]

[1]Institut de Biologia Evolutiva, CSIC-Universitat Pompeu Fabra, Barcelona, Spain, [2]Department of Biology, Bioinformatics, University of Copenhagen, Copenhagen, Denmark, [3]The Genome Institute, Washington University School of Medicine, St. Louis, MO, [4]Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland, [5]Department of Genetics, University of Cambridge, Cambridge, United Kingdom, [6]Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden, [7]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, [8]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, [9]Research and Conservation, Copenhagen Zoo,Copenhagen, Denmark, [10]Centro Nacional de Analisis Genomico (CNAG), PCB, Barcelona, Spain, [11]Institucio Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain

The initial sequencing of the chimpanzee genome and its comparison to our own constituted a major leap forward in the understanding of the evolutionary trajectories of our species, but the draft nature of the assembly precluded the analysis of complex genomic regions such as variation in tandem repeats and copy number. Here, we present a more complete de-novo assembly for the common chimpanzee, incorporating a wide variety of novel technologies. We increased contiguity by over 500% on contigs, and by over 300% on scaffolds. We now more clearly address the complexity of chimpanzee genome evolution by genotyping thousands of novel additional complex regions compared to the most recent assembly of the chimpanzee. To do so, we analyzed 70 complete genome sequences of chimpanzees and bonobos with known geographic origin to decipher the complex demographic history of the Pan lineage. Based on patterns of genetic diversity, linkage disequilibrium, and population modeling, we propose an origin of all chimpanzees in central Africa. We also provide genetic evidence of gene flow between chimpanzees and bonobos in the wild, and show that such events have shaped the genomic diversity of present-day Pan populations. Finally, since our sampling scheme included detailed geographic information, covering 9 countries, we explore the fine-scale population substructure within chimpanzees (sometimes at a regional level within countries) and report that genetic diversity is a remarkably good predictor of geographic origins of unknown samples, which has key implications for the conservation of this endangered species.

← Back

REGULATION OF THE *E. COLI* RNA POLYMERASE

Avantika Lal[1], Sandeep Krishna[2], Aswin Sai Narain Seshasayee[1]

[1]National Centre for Biological Sciences, National Centre for Biological Sciences, Bangalore, India, [2]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Bangalore, India

Bacteria possess a single RNA polymerase, which transcribes all genes. Any change in its level, distribution or activity could have rapid and widespread effects. In *Escherichia coli*, the sigma factor σ^{70} directs RNA polymerase to transcribe growth-related genes, while σ^{38} directs it to transcribe stress response genes during stationary phase. Their competition is regulated by Rsd, which sequesters free σ^{70}, and 6S RNA, which sequesters the RNA polymerase-σ^{70} holoenzyme. Here we use genome-wide expression studies and theoretical modeling to investigate the functions of Rsd and 6S RNA. We show that 6S RNA and Rsd act as global regulators of gene expression throughout bacterial growth, and propose a model in which 6S RNA reduces transcription by both σ^{70} and σ^{38} while Rsd specifically increases transcription by σ^{38}. Finally, we find evidence of multiple interactions between 6S RNA and Rsd, which are vital to regulating sigma factor competition.

COREGULATION OF TANDEM DUPLICATE GENES SLOWS EVOLUTION OF SUBFUNCTIONALIZATION IN MAMMALS.

Xun Lan[1,3], Jonathan K Pritchard[1,2,3]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]Stanford University, Department of Biology, Stanford, CA, [3]Stanford University, Howard Hughes Medical Institute, Stanford, CA

Gene duplication is a fundamental process in genome evolution. However, most young duplicates are degraded by loss-of-function mutations, and the factors that allow some duplicate pairs to survive long-term remain controversial. One class of models to explain duplicate retention invokes sub- or neofunctionalization, while others focus on sharing of gene dosage. RNA-seq data from 46 human and 26 mouse tissues indicate that subfunctionalization of expression evolves slowly, and is rare among duplicates that arose within the placental mammals, possibly because tandem duplicates are co-regulated by shared genomic elements. Instead, consistent with the dosage-sharing hypothesis, most young duplicates are down-regulated to match expression levels of single copy genes. Thus, dosage sharing of expression allows for the initial survival of mammalian duplicates, followed by slower functional adaptation enabling long-term preservation.

THE ROLE OF HAPLOTYPE EPISTASIS IN HUMAN GENETIC VARIATION AND DISEASE RISK

Stephane E Castel[1,2], Jimmy Z Liu[1], GTEx Consortium[1], Joseph K Pickrell[1], Tuuli Lappalainen[1,2]

[1]New York Genome Center, NY, NY, [2]Columbia University, Dept of Systems Biology, NY, NY

Most genetic studies consider genetic variants one at a time, independently of other variants. Epistasis, or non-additive genetic interaction between genetic variants has been difficult to pinpoint between distant variants. However, haplotypic epistasis between nearby variants is not only statistically more feasible to test, but also has concrete biological mechanisms of allelic gene function. In this study, we have addressed how haplotypic epistasis affects the penetrance of coding variants, studying its contribution to genetic variation in the general population and to disease risk.

First, we analyzed interaction between coding variants in the same gene in the GTEx data set, with variants phased using our novel method phASER using both WGS and RNA-seq reads. We observed an enrichment of several protein-truncating variants on the same haplotype expected as a result of relaxed selection after one inactivating mutation. Furthermore, putatively deleterious coding variants appear to enrich in exons that are excluded by splicing.

Next, we analyzed haplotypic epistasis between cis-eQTLs and coding variants in GTEx data, hypothesizing that deleterious coding variants on the more highly expressed haplotype may have higher penetrance. We found evidence of purifying selection acting to remove such haplotype combinations: in allelic expression and haplotype data rare missense coding variants are more often on the lower expressed haplotype than synonymous variants ($p < 6e-14$). Finally, we observed more negative epistatic interactions in genes that are involved in disease ($p < 1e-4$).

To identify specific instances where haplotype epistasis may contribute to disease we tested for haplotypic association to phenotypes in the GERA study. We found that for 17 of 44 significant ($p<5*10e-8$) coding SNP associations, including haplotype configuration with cis-eQTL significantly improved the power to predict phenotype from genotype (1% FDR), with one significant hit in the direction predicted by the epistasis model ($p < 0.01$; the other 16 lacked power to determine the direction of the effect). Additionally, we are currently analyzing the role of haplotypic epistasis in rare variant associations to autism.

Altogether, our results show that epistasis between variants on the same haplotype is relatively common, and has shaped the patterns of both coding and regulatory variation in humans. Our results indicating that epistasis affects disease associations highlights the importance of analyzing genetic variants in the context of the entire haplotype.

← Back

PRESERVATION OF MOLECULAR IDENTITY DURING WHOLE GENOME AMPLIFICATION TO ENABLE ACCURATE SINGLE-CELL MUTATION INFERENCE

Christopher E Laumer[1,2], Thierry Voet[2,3], John C Marioni[1,2,4]

[1]European Molecular Biology Laboratories - European Bioinformatics Institute, Single Cell Genomics Centre, Hinxton, United Kingdom, [2]Wellcome Trust Sanger Institute, Single Cell Genomics Centre, Hinxton, United Kingdom, [3]KU Leuven, Centrum Menselijke Erfelijkheid, Leuven, Belgium, [4]Cancer Research UK, Cambridge Institute, Cambridge, United Kingdom

The promise of single cell genomics lies in its ability to deliver complete measurements of the heterogeneity that differentiates a population of cells. To date, however, this field has arguably seen its greatest successes in observing functional states such as those of the transcriptome and epigenome, with single-cell level investigations of the hereditary material being more constrained in scope. Differences in target size notwithstanding, a major driver of this discrepancy has been the inefficiency of genomic DNA library preparation, which all but mandates exponential preamplification through processes such as multiple-displacement amplification (MDA). These procedures introduce erroneous bases and structural rearrangements which are propagated throughout amplification, yielding false positive variants that may outpace the magnitude of legitimate mutations; stochastic variation in amplification efficiency can also yield dropouts of entire loci and worse yet, loss-of-heterozygosity events whose existence is not readily discernible (i.e. false negatives). Here, we present efforts to overcome these challenges through novel sample preparation protocols and the concept of unique molecular identity. Cell lysis and proteolytic digestion are used to expose the naked genomic DNA, which is then enzymatically cleaved into short (<1kb) fragment sizes without appreciable denaturation or loss. These fragments are then efficiently converted into a library through cohesive-end ligation. The adapters thus introduced are then used to amplify single-stranded copies of the template, through a linear, nicking-endonuclease-driven process called Strand Displacement Amplification (SDA). DNA copies generated through SDA have favourable properties for mutation inference, such as an independent distribution of amplification errors among amplicons, and reduced coverage bias, since amplification is limited not primarily by sequence context but by the rate of enzymatic nicking. Because the inserts captured by this process are also uniquely tagged prior to amplification, it is possible to generate a consensus of the duplex sequence, thereby limiting false positive variant calls; furthermore, molecular counting renders the detection of allelic dropouts and copy number variations straightforward. We present experimental results used to estimate the library conversion efficiency under various ligation conditions, to inform the rate, linearity, and possibility of background amplification during various formulations of SDA, and finally, to address the coverage properties and variant-calling fidelity of the method using real single-cell samples.

← Back

GIGGLE: INDEXING AND SEARCH ALL GENOMIC ANNOTATION TRACKS

Ryan M Layer, Brent S Pedersen, Aaron R Quinlan

University of Utah, Human Genetics, Salt Lake City, UT

A massive effort has been put forth to characterize the functional elements of the human genome. Over 1,600 annotation tracks have been produced by the ENCODE project, 2,800 by Roadmap Epigenomics, and 8,500 by GTEx. The methods for accessing these data have remained basically the same for nearly a decade ---site-by-site visual inspection on a genome browser, like UCSC. While this visualization remains useful and important, it does not scale. Answering the common, and seemingly simple, question of what else occurs at particular locus by manually scanning every track on the genome browser (about 200) is intractable. We need some mechanism to filter the tracks down to just the subset that have some overlap with the locus, and when a larger set of intervals is considered we need a way of quickly summarizing and comparing that set's relationships among all of the tracks. Existing tools, like TABIX, could be used to perform this type of search, but they were not designed to simultaneously search thousands of tracks.

To address this we developed GIGGLE, a genomic interval search strategy that creates a unified index of the intervals and genome features from many different annotation files and formats. GIGGLE is based on a B+ tree, where the bounds of each interval serve as the key to the path and file offset of the original record. The immediate result of a GIGGLE search is a list of hits per indexed file, which on their own can be used for rapid summary and prioritization statistics. For example, given a single interval, GIGGLE reports the tracks in which at least one intersection was observed, while a multi-interval GIGGLE search can give a list of tracks ordered by an estimate of significance based on either intersection or proximity. This estimate uses a Fisher's exact test that correlates well with a permutation test but is much faster. These summaries are returned nearly instantly, enabling fast data exploration. File offsets of the original record are stored in the tree so that the full data can be accessed on demand.

GIGGLE has been designed to be useful at different scales. At a local level, GIGGLE can be used as a C API, via Python bindings, and as a command line tool to search a curated set of published and experimental data. We are also investigating the use of GIGGLE as a large-scale public search engine with an index across major projects like ENCODE, Roadmap Epigenomics, and GTex. Initial results (without optimizations) are promising with search rates topping 10,000 interval queries per second in a database with over 50 million intervals across 120 tracks. GIGGLE's speed and flexibility make it a promising framework for large-scale genomic analysis.

https://github.com/ryanlayer/giggle

A HIGH-THROUGHPUT, EXPERIMENTAL METHOD FOR QUANTIFYING THE EFFECTS OF ENHANCER METHYLATION ON GENE EXPRESSION

Amanda J Lea[1], Christopher M Vockley[2,3], Timothy E Reddy[3,4,5], Luis B Barreiro[6], Jenny Tung[1,7,8,9]
[1]Duke University, Department of Biology, Durham, NC, [2]Duke University, Department of Cell Biology, Durham, NC, [3]Duke University, Center for Genomic and Computational Biology, Durham, NC, [4]Duke University, Program in Computational Biology and Bioinformatics, Durham, NC, [5]Duke University Medical School, Department of Biostatistics and Bioinformatics, Durham, NC, [6]Sainte-Justine Hospital Research Centre, University of Montreal, Department of Pediatrics, Montreal, Canada, [7]National Museums of Kenya, Institute of Primate Research, Nairobi, Kenya, [8]Duke University, Department of Evolutionary Anthropology, Durham, NC, [9]Duke University, Duke University Population Research Institute, Durham, NC

DNA methylation is a key gene regulatory mechanism with a central role in development, disease susceptibility, and the response to environmental conditions. By altering the gene expression program of the cell, variation in DNA methylation levels can influence traits of both biomedical and evolutionary importance. However, not all changes in DNA methylation levels exert causal effects on gene expression, and only a subset of the ~28 million CpG sites in the human genome are presumed to impact gene regulation. Yet, due to methodological constraints, we are currently unable to characterize this subset in an experimental, high-throughput manner. To do so for the first time, we developed an approach, termed 'mSTARR-seq,' that combines genome-scale approaches for quantifying enhancer activity (i.e., STARR-seq) with manipulation of DNA methylation marks in vitro. The resulting assay simultaneously assess enhancer activity at thousands of loci, as well as the impact of DNA methylation on this activity, in a single experiment. To demonstrate the utility of our method, we experimentally manipulated CpG methylation levels in ~1 Mb of human BAC sequence, cloned into specially designed CpG-free vectors with self-transcribing activity. Specifically, we inserted human DNA fragments upstream of a poly-A signal and downstream of a constitutive promoter, such that regions possessing strong enhancer would transcribe themselves. We transfected these constructs into human A549 cells (n=4 methylated replicates and 4 unmethylated replicates) and sequenced the resulting mRNA to assess enhancer activity. As expected, mRNA harvested from unmethylated vectors mapped preferentially to sequences enriched for known enhancers annotated by ENCODE and the Roadmap Epigenomics Project. Further, based on comparisons between the methylated and unmethylated conditions, we found that DNA methylation tends to suppress the activity of many of these same enhancers. However, we also identified enhancer regions whose activity was not significantly affected by DNA methylation, suggesting that associations between DNA methylation levels and disease or environmental exposures may not always be functionally relevant. Taken together, our method provide a novel approach for testing where in the genome DNA methylation 'matters' for gene regulation, a question of substantial interest in functional, evolutionary, and medical genomics.

← Back

USING SYNTHETIC MOUSE SPIKE-IN TRANSCRIPTS TO EVALUATE RNA-SEQ ANALYSIS TOOLS

Dena Leshkowitz[1], Ester Feldmesser[1], Gilgi Friedlander[1], Ghil Jona[1], Elena Ainbinder[1], Yisrael Parmet[2], Shirley Horn-Saban[3]

[1]Weizmann Institute of Science, Biological Services, Rehovot, Israel, [2]Ben-Gurion University of the Negev, Industrial Engineering and Management Department, Beer Sheva, Israel, [3]Galil Genetic Analysis Ltd., Katsrin, Israel

One of the key applications of next-generation sequencing (NGS) technologies is RNA-Seq for transcriptome genome-wide analysis. Although multiple studies have evaluated and benchmarked RNA-Seq tools dedicated to gene level analysis, few studies have assessed their effectiveness on the transcript-isoform level. Alternative splicing is a naturally occurring phenomenon in eukaryotes, significantly increasing the biodiversity of proteins that can be encoded by the genome. The aim of this study was to assess and compare the ability of the bioinformatics approaches and tools to assemble, quantify and detect differentially expressed transcripts using RNA-Seq data, in a controlled experiment. To this end, in vitro synthesized mouse spike-in control transcripts were added to the total RNA of differentiating mouse embryonic bodies, and their expression patterns were measured. This novel approach was used to assess the accuracy of the tools, as established by comparing the observed results versus the results expected of the mouse controlled spiked-in transcripts. We found that detection of differential expression at the gene level is adequate, yet on the transcript-isoform level, all tools tested lacked accuracy and precision.

TUNABLE NANOCONFINEMENT FOR SINGLE-MOLECULE MANIPULATION, MODIFICATION, AND VISUALIZATION: TOWARD NEXT GENERATION GENOMIC ANALYSES

Sabrina R Leslie, Daniel Berard, Gilead Henkin, Francis Stabile

McGill University, Physics, Montreal, Canada

Long-DNA sequencing and mapping technologies often face critical bottlenecks in linearizing, chemically modifying, and visualizing long DNA molecules within controlled nanofluidic environments. Developing these technologies is important to analyzing structural variation, haplotyping, repetitive regions, and oncogenomics, in turn advancing our understanding of diseases such as cancer, and conditions such as autism. Here, we present a new high-throughput and high-fidelity loading platform for DNA into sub-50 nm cross-section nanostructures, ideal for long-DNA sequencing and mapping applications. Our platform overcomes key challenges faced by existing technologies including device clogging and molecular breakage. It works by continuously adjusting the confinement geometry, so as to dynamically untangle and squeeze long DNA polymers into embedded nanostructures. Specifically, we use the Convex Lens-induced Confinement (CLiC) single-molecule technique to gently load the DNA polymers into open-face nanogrooves from above (Berard et al, PNAS 2014), which applies much smaller (entropic) forces than other approaches (e.g. electrophoretic or pressure-based). We demonstrate reliable separation, linearization, controlled reaction and deposition of many DNA molecules in parallel, as well as high-fidelity extension for readout, using a range of substrates. This work constitutes a new pathway toward long-DNA sequencing, as well as a new platform for visualizing biomolecular complexes and dynamics.

Beyond templating molecules for visualization, we integrate controlled, in-situ chemistry procedures within the nanofluidic device and miniaturize its format. We can first template biopolymers in embedded nanostructures, and second introduce reagent molecules within the nanofluidic environment without disturbing them, while simultaneously visualizing dynamics and interactions in real time. Quantifying changes in configurations of DNA in response to binding/unbinding of reagent molecules provides new insights into a variety of processes such as DNA compaction, as well as topology-mediated processes such as strand-invasion. Overall, we demonstrate a novel, sensitive, high-throughput platform which opens new doors to a wide range of genomic and biochemical analyses.

IDENTIFYING RISK ALLELES IN *ARHGEF17* FOR INTRACRANIAL ANEURYSMS WITH MODEST SAMPLE SIZE

Jiani Li[1,2], Xinyu Yang[3], Zhen Zhang[3], Graeme Mardon[1], Yongtao Guan[1,4], Fuli Yu[1,2,3]

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [3]Tianjin Medical University General Hospital, Department of Neurosurgery, Tianjin, China, [4]Baylor College of Medicine, USDA/ARS Children's Nutrition Research Center, Houston, TX

Intracranial Aneurysms (IAs) are dilatations in the arteries that supply blood to brain. The estimated prevalence of unruptured IAs is 2%-6% in adult population. The danger posed by IAs is the possibility of rupture leading to life-threatening subarachnoid hemorrhage (SAH). The annual incidence of SAH is 4-9 per 100,000 worldwide. The IAs are complex traits, the risk of which is affected by age, smoking, hypertension, excess drinking and family history of IAs. To identify the risk alleles that predispose to IAs, we sequenced 20 IA cases (either familial or sporadic) and combined with reported unsolved IA cases that had whole exome sequencing data available (93 IA cases in total). We developed a novel statistical pipeline for IA risk variants prioritization. Through our pipeline, we identified five different deleterious variants of the *ARHGEF17* in 21 cases from nine families with different ethnicities. All those variants were rare or low frequency variants in 1000G controls or ExAC series. *ARHGEF17* highly expresses in human microvascular endothelial cells (HMEC). It localizes to actin stress fibers and are required for cell-cell adhesion, endothelial monolayer integrity and barrier function by locally activating Rho GTPas at the junctions. Four of five variants were within the portion of *ARHGEF17* encoding the catalytic cassettes. These data establish *ARHGEF17* as a IA predisposition gene.

← Back

USING BLOCBUSTER TO IDENTIFY MULTI-SNP ASSOCIATION PATTERNS IN ALZHEIMER'S DISEASE COHORTS

Zeran Li[1], Yuetiva Deming[1], Jorge Del Aguila[1], Victoria Fernandez[1], Laura Ibanez[1], Benjamin Saef[1], Bill Howells[1], ShengMei Ma[1], John Budde[1], Kathleen Black[1], David Carrell[1], Carlos Cruchaga[1], Sharlee Climer[2]

[1]Washington University in St.Louis, Psychiatry, St.Louis, MO, [2]Washington University in St.Louis, Computer Science, St.Louis, MO

BACKGROUND: Despite the substantial efforts and recourses invested into genetic risk studies for late-onset Alzheimer's disease (AD), there is still missing heritability that may be due to complex interactions among multiple genetic factors. To investigate combinatorial genetic variants associated with AD, we applied a recently developed analytical package, BlocBuster, to identify interactions appearing as multi-SNPs patterns and tested these patterns for association with disease status.
METHODS: Samples from Washington University Knight Alzheimer's Disease Research Center (KADRC) and Alzheimer's Disease Neuroimaging Initiative (ADNI) cohorts were genotyped and processed with BlocBuster package. First, a Custom Correlation Coefficient (CCC) value was calculated for each SNP pair based on a weighting score assigned to each of the four possible relationships between a pair of biallelic SNPs. Second, a network was built in which each node represented a SNP allele and each edge connected a pair of SNP alleles with a significant CCC value. Third, multi-SNP patterns associated with disease status were extracted based on their allelic frequencies in AD and control cohorts.
RESULTS: BlocBuster identified 5466 clusters with risk odds ratio (OR) > 1.3 or protective OR < 0.769 in the discovery dataset of 2092 subjects. These clusters contained 18,853 SNPs, of which 9578 (50.8%) were genotyped in the replicate dataset of 2900 subjects. There are 30 risk clusters and 73 protective clusters that were identified in the replicate dataset with similar OR values. Among those, 4 blocs with OR > 2 or < 0.5 were identified, indicating strong association with disease status. One bloc, comprised of rs12778775 and rs11191799, is located within *SH3PXD2A* at 10q24.33. An alternate SNP within *SH3PXD2A* was previously reported to be associated with AD, and it is believed this gene interacts with *ADAM12* to mediate Aβ toxicity. Another bloc, comprised of rs12778775 and rs11191799, is linked to *CDH13* at 16q23.3, which encodes a receptor for low-density lipoprotein and has shown pleiotropic effects, particularly for metabolic traits.
DISCUSSION: BlocBuster has revealed a large number of combinatorial associations with AD and may provide more specificity and power than previous studies, as demonstrated by the identification and replication of the SH3PXD2A and CDH13 haplotypes. The preliminary results suggest there may be haplotypes with strong AD associations that have been overlooked by single-marker approaches and demonstrate strong potential for using BlocBuster to identify multi-SNP association patterns in AD cohorts.

THE GENETIC ARCHITECTURE OF SHORT STATURE IN THE SOUTH AFRICAN SAN

Meng Lin[1], Julie M Granka[2], Alicia R Martin[3], Justin Myrick[4], Elizabeth G Atkinson[1], Cedric J Werely[5], Deepti Gurdasani[6,7], Cristina Pomilla[6,7], Tommy Carstensen[6,7], Brooke Scelza[4], Marlo Moller[5], Manj Sandhu[6,7], Carlos D Bustamante[3], Eileen G Hoal[5], Marcus W Feldman[2], Christopher R Gignoux[3], Brenna M Henn[1]

[1]Stony Brook University, Department of Ecology and Evolution, Stony Brook, NY, [2]Stanford University, Department of Biology, Stanford, CA, [3]Stanford University, Department of Genetics, Stanford, CA, [4]University of California Los Angeles, Department of Anthropology, Los Angeles, CA, [5]Stellenbosch University, Division of Molecular Biology and Human Genetics, Tygerberg, South Africa, [6]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, United Kingdom, [7]University of Cambridge, Department of Public Health and Primary Care, Cambridge, United Kingdom

Adult human height as a classic quantitative trait has been extensively studied in European populations. The architecture is considered additive, consisting of hundreds to thousands of small effect loci across the genome. The total number of height GWAS variants in Europeans, however, explains only a limited amount of phenotypic variance relative to the high heritability of the trait. Whether the underlying architecture of height is the same across diverse populations remains unknown. Here we characterize the genetic architecture of height in two KhoeSan communities, ≠Khomani San and Nama (n=400). In contrast to Western cohorts, these groups practice traditional hunter-gatherer or pastoralist subsistence in the rural Kalahari and Richtersveld Deserts of South Africa. They possess much shorter stature (~155cm) compared to average Europeans. Using imputed genomic data, we identified GWAS signals associated with height variation, where the top hits clustered in the intronic regions of VWA8 and KSR2, previously associated with BMI and energy regulation. The significance of these regions were confirmed though targeted resequencing at high coverage in the same cohorts, identifying nearby common variants ($p = 5.19e-8$ for top hit in VWA8, $p = 5.33e-9$ for top hit in KSR2). Gene-based and SNP-based tests demonstrate replication of these hits in the latest height association release of GIANT ($p=0.01$), which was conducted in European cohorts with ~250,000 individuals. Despite significantly smaller stature, pedigree-based heritability estimates of height in the KhoeSan are similar to or even somewhat higher than that of Europeans ($h^2=0.95$). Linear mixed models of all pairwise relationships from SNP array and exome data closely match the pedigree results, likely due to inclusion of relatives. Given greater environmental and household homogeneity in these populations, we explore the implications of our results for the role of rare variants in heritability. Taken together, these results show that endogamous cohorts can be effectively leveraged for characterization of genetic architecture and large effect variants.

FUNCTIONAL ANNOTATION GUIDED GENOTYPE-PHENOTYPE ASSOCIATION ANALYSES OF WHOLE GENOME SEQUENCE DATA

Xiaoming Liu[1], Elena V Feofanova[1], Akram Yazdani[1], Bing Yu[1], Peng Wei[1], Alanna C Morrison[1], Eric Boerwinkle[1,2]

[1]University of Texas School of Public Health, Human Genetics Center, Houston, TX, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

For many human diseases, it has been hypothesized that rare variants and variants affecting regulatory functions may play important roles in complex disease etiology. Whole genome sequence (WGS) based genotype-phenotype association studies are therefore well-suited to identify novel genes or loci associated with complex diseases. In comparison to GWAS, WGS based studies provide the unique opportunity to observe the disease causal (rare) variants directly but also suffer a much inflated noise to signal ratio due to the fact that a vast majority of the variants observed are either non-functional or functional but unrelated to the diseases in study. One approach to decrease the noise to signal ratio and to boost the power of genotype-phenotype association analysis is utilizing functional annotation to effectively reduce the variant space for testing or to mute the negative effect of non-functional variants in testing.

Here we report an ongoing functional annotation guided genotype-phenotype association analysis of disease related quantitative traits (i.e., blood cell traits, chemical measures, and cardiac- and lipid-related) with whole genomes of 1,900 African Americans. Variants were called using the goSNAP program and then annotated using the Whole Genome Sequencing Annotator (WGSA) pipeline. We employed various approaches to incorporate functional annotations into association tests. First, regulatory elements, including predicted promoters and enhancers as well as first introns, were used as the analytical units of inference in aggregated rare variant association tests. Second, a sliding window approach across the genome was applied to aggregate rare variants (minor allele frequency, MAF≤5%) within a physical window (window size of 4 kb and skip size of 2 kb). Within each window, T5 burden test and the Sequence Kernel Association Test (SKAT) were conducted adjusting for age, sex and principle components of the data. Deleteriousness prediction scores of the variants, such as Combined Annotation Dependent Depletion (CADD), were utilized as weights in T5 and SKAT tests. Third, deleteriousness prediction scores were used to set the prior for a Bayesian based genotype-phenotype association test. As we and other groups previously have shown that incorporating functional annotations can boost the power of association tests while well controlling the type I error rate, we expect that our functional annotation guided genotype-phenotype analyses will identify novel loci of regulatory function for the physiological traits studied.

SOMATIC MUTATION IN SINGLE HUMAN NEURONS TRACKS DEVELOPMENTAL AND TRANSCRIPTIONAL HISTORY

Michael A Lodato[1,2], Mollie B Woodworth[1,2], Semin Lee[3,4], Peter J Park[3,4], Christopher A Walsh[1,2]

[1]Children's Hospital Boston, Genetics and Genomics, Boston, MA, [2]Harvard Medical School, Pediatrics and Neurology, Boston, MA, [3]Harvard Medical School, Biomedical Informatics, Boston, MA, [4]Brigham & Women's Hospital, Genetics, Boston, MA

The genome is under constant pressure from mutagens, such that each individual harbors 50-100 de novo variants resulting from mutations that occurred in the germline of either parent. Interestingly, the somatic mutation rate is thought to be much higher than the germline rate. Classic experiments measuring loss-of-function of reporters such as HPRT in cultured lymphoblasts or loss-of-heterozygosity of tumor suppressor genes such as RB and APC suggest that each cell might harbor 100-1,000 somatic mutations by age 15. Thus, despite evidence of nuclear equivalence from developmental biology, from the first cleavage division of the fertilized egg until the death of an individual, somatic cells likely accumulate genetic lesions. Therefore each individual, and each tissue within that individual, exists as a mosaic of distinct genotypes.

Recent research is beginning to reveal the importance of somatic genomic mosaicism in human disease. Cancer is the classical example of a disease driven by random somatic mutations, suggesting that somatic mutations accumulate over time across a wide range of tissues in the human body. Neurological diseases such as Proteus syndrome and epilepsy can be caused by clonal expansions of somatically mutated cells during neurodevelopment. Post-mitotic neurons of the brain are long-lived and thus have extended exposure to genotoxic insults, increasing their likelihood of accruing somatic mutations, and human neurons have been confirmed to harbor somatic variants in recent studies by our lab and others. Thus, somatic mosaicism is known to be associated with human neurological disease, and since normal neurons harbor somatic variants, may play a part of normal human biology as well.

We surveyed the landscape of somatic single-nucleotide variants in the human brain using single-cell, whole genome DNA sequencing. We identified thousands of somatic SNVs by single-cell sequencing of 36 neurons from the cerebral cortex of three normal individuals. Unlike germline and cancer SNVs, which are often caused by errors in DNA replication, neuronal mutations appear to reflect damage during active transcription. Somatic mutations create nested lineage trees, allowing them to be dated relative to developmental landmarks and revealing a polyclonal architecture of the human cerebral cortex. Thus, somatic mutations in the brain represent a durable and ongoing record of neuronal life history, from development through postmitotic function.

← Back

EVALUATING THE EFFICIENCY OF PURIFYING SELECTION IN AFRICAN POPULATIONS WITH DIFFERENT MODES OF SUBSISTENCE

Marie Lopez[1], Athanasios Kousathanas[1], Hélène Quach[1], Christine Harmant[1], Alain Froment[2], Evelyne Heyer[2], Paul Verdu[2], George H Perry[3], Luis B Barreiro[4], Etienne Patin[1], Lluis Quintana-Murci[1]

[1]Institut Pasteur, CNRS URA3012, Unit of Human Evolutionary Genetics, Paris, France, [2]CNRS, MNHN, Université Paris Diderot, Paris, France, [3]Pennsylvania State University, University Park, 16802, PA, [4]Université de Montréal, CHU Sainte Justine, Montréal, Canada

Most nonsynonymous mutations are believed to be deleterious and are removed from the population by purifying selection. Recent studies have evaluated how differences in demographic history impact the number, proportion and severity of nonsynonymous variants. Here we evaluate the efficiency of purifying selection in sub-Saharan African populations representing different modes of subsistence and demographic regimes. We generated 400 whole exome sequences from populations of rainforest hunter-gatherers and neighboring farmers located in western and eastern central Africa. We identified 599,218 SNPs covering 20,726 genes and showed higher nucleotide diversity in farmers than in hunter-gatherers (increased by 9%-13%) in both geographic locations. In addition, farmers present a significantly higher number of low-frequency variants than hunter-gatherers ($P=2.2x10^{-16}$). To refine the demographic history of these populations, we used DFE-alpha and fastSIMCOAL composite likelihood methods on the synonymous site frequency spectrum. Consistent with the empirical observations, both methods estimated a 60% population contraction in hunter-gatherers and a 70% population growth in farmers. To understand how these recent changes in population sizes have affected the efficiency of purifying selection, we first inferred the distribution of fitness effects of new nonsynonymous mutations (DFE) using DFE-alpha. We found no significant differences in the DFE of hunter-gatherers and farmers, an observation that was replicated in both western and eastern settings. We then empirically quantified the deleteriousness of segregating nonsynonymous mutations within each population according to several conservation statistics (PolyPhen2, GERP++, C-scores), and found that both hunter-gatherer and farmer populations had similar proportions of segregating deleterious mutations in all allele frequency bins. Together, our results show no evidence for a difference in the efficiency of purifying selection between African populations, despite their strong differences in recent demographic history (expansion versus bottleneck). We thus provide empirical support for previous theoretical studies suggesting that recent demographic history had limited impact on the efficiency of purifying selection in several human populations.

DETECTING COPY NUMBER VARIATION LINKED TO PHENOTYPIC TRAITS AND REPEATED EVOLUTION

Craig B Lowe[1,2], Nicelio Sanchez-Luege[1], Timothy R Howes[1], Shannon D Brady[1], Rhea R Richardson[1,3], Felicity C Jones[1], Michael A Bell[4], David M Kingsley[1,2]

[1]Stanford University, Department of Developmental Biology, Stanford, CA, [2]Howard Hughes Medical Institute, Stanford University, Stanford, CA, [3]Stanford University, Department of Genetics, Stanford, CA, [4]Stony Brook University, Department of Ecology and Evolution, Stony Brook, NY

We present a method to detect copy number variants (CNVs) that are differentially present between two groups of sequenced samples. We use a multi-tape finite-state transducer where the emitted read depth is conditioned on the mappability and GC-content of all reads that could cover a given base position. In this model, the read depth within a region is a mixture of binomials, which in simulations matches the read depth more closely than the often-used negative binomial distribution. The method analyzes all samples simultaneously, preserving uncertainty as to the breakpoints and magnitude of CNVs present in an individual when it identifies CNVs differentially present between the two groups. This unified approach outperforms alternative methods that execute these tasks serially, first identifying copy number variants in individuals and then identifying which variants are consistently correlated with a trait of interest.

We apply this transducer method to identify CNVs that are recurrently associated with postglacial adaptation of marine Threespine Stickleback (Gasterosteus aculeatus) to freshwater. We identify 6664 regions of the stickleback genome, totaling 1.7Mbp, which show consistent copy number differences between multiple different marine and freshwater populations. These deletions and duplications affect both protein-coding genes and cis-regulatory elements, including a noncoding intronic telencephalon enhancer of DCHS1. The functions of the genes near or included within the 6664 CNVs are enriched for immunity and muscle development, as well as head and limb morphology. These functions match consistent phenotypic differences that have evolved repeatedly between marine and freshwater stickleback populations.

While freshwater stickleback have been iteratively derived from ancestral marine populations that are thought to have been relatively static, we show that freshwater stickleback populations can also act as reservoirs for ancient sequences that are conserved to other teleosts, but largely missing from marine stickleback due to recent selective sweeps in marine populations.

THE CRITICAL FUNCTIONS ENCODED BY SYNONYMOUS SITES

Heather E Machado[1], David S Lawrie[2], Dmitri A Petrov[1]

[1]Stanford University, Biology, Stanford, CA, [2]University of Southern California, Biological Sciences, Los Angeles, CA

Although synonymous sites have historically been considered neutral, and are still used as a neutral reference for many analyses, several processes are known to exert a selective pressure on synonymous sites. Despite evidence for selection on synonymous sites across species and for several different processes, the extent of selection on synonymous sites and the relative contributions of the selective processes are not well understood. Using genome sequence data from two *Drosophila melanogaster* populations, we perform a SFS-based maximum likelihood estimation of purifying selection on fourfold degenerate synonymous sites using short introns as a neutral control. We estimate that ~15% of fourfold sites are under strong purifying selection. We find that the selection can largely be explained by codon bias and splicing-related factors. If preferred codons and alternatively spliced genes are excluded from the analysis, we find no evidence for selection, and recover the short intron level of polymorphism. These results clearly identify the major processes contributing to purifying selection on synonymous sites, and have implications for creating a neutral reference for other species.

REGULATION OF THE TRANSCRITOME THOUGH RNA STABILITY UNDER HYPOXIA IN HUMAN COLORECTAL CANCER CELLS

Sho Maekawa[1], Sumio Sugano[1], Nobuyoshi Akimitsu[2], Yutaka Suzuki[1]

[1]The University of Tokyo, Graduate School of Frontier Sciences, Kashiwa, Chiba, Japan, [2]The University of Tokyo, Radioisotope Center, Tokyo, Japan

The eventual RNA abundance is determined by the sum of rates of transcription and RNA decay. Recent functional genomics projects have focused largely on obtaining epigenomic data to understand gene expression programmes and they have been successful to certain extent in predicting the RNA expression level; however, the understanding of mechanisms and effects of gene expression programme is far from complete. We have previously reported the extent of roles of RNA decay in regulating the eventual RNA abundance and the effects of RNA decay factors in regulating them. We found that RNA half-life measured by 5'-bromouridine immunoprecipitaiton pulse chase sequencing (BRIC-seq) was shorter for genes that have significant H3K4me3 ChIP-seq signal despite the lack of RNA expression and we found 865 genes that were predominantly controlled at the level of RNA stability of which 8% were controlled by three RNA decay factors: UPF1, EXOSC5 and STAU1. These results suggested the relevance of RNA stability in regulating RNA abundance at steady state level.
In order to understand the degree of contributions by RNA decay under an external stimulus, we have subjected a human colorectal cancer cell line to hypoxia, a condition with low oxygen potential, and profiled RNA decay using BRIC-seq. By conducting BRIC-seq results we have found 2381 and 1037 genes that had their RNA half-life elongated or shortened, in hypoxia by two-fold, respectively, suggesting that RNA decay does change upon hypoxia. We compared RNA-seq and BRIC-seq and we found of those that had their RNA half-lives elongated in BRIC-seq, we found that 169 genes were up-regulated in RNA-seq as well, which suggest cooperative regulation between transcription and RNA decay. However, 1905 genes had their RNA half-life elongated that did not result in the eventual RNA abundance from RNA-seq, which suggests potential feedback mechanisms as RNA decay changes did not result in changes of the eventual RNA abundance. Furthermore, when we compared our previous HIF1 alpha ChIP-seq in the same cell-line in hypoxia condition, only 4 out of 1905 genes that had their RNA half-lives elongated had HIF-1 alpha ChIP-seq binding. It suggests distinct control of RNA half-life from HIF-1 and that the alternative hypoxia responsive pathways are responsible in controlling RNA decay in hypoxia. Taken together, our approach allows more comprehensive understanding of mechanisms of gene expression in hypoxia.

DE NOVO SEQUENCED AND ASSEMBLED GORILLA Y CHROMOSOME SHOWS STRONG CONSERVATION WITH HUMAN BUT NOT CHIMPANZEE

Marta Tomaszkiewicz[1], Samarth Rangavittal[1], Monika Cechova[1], Rebeca Campos-Sanchez[1], Howard Fescemeyer[1], Robert Harris[1], Danling Ye[1], Rayan Chikhi[5], Oliver Ryder[2], Malcolm A Ferguson-Smith[3], Paul Medvedev[4], Kateryna D Makova[1]

[1]Penn State University, Department of Biology, University Park, PA, [2]San Diego Zoological Society, Department of Genetics, Escondido, CA, [3]University of Cambridge, Department of Veterinary Medicine, Cambridge, United Kingdom, [4]Penn State University, Department of Computer Science and Engineering, University Park, PA, [5]University of Lille, CNRS, Lille, France

The mammalian Y chromosome sequence, critical for studying male fertility and dispersal, is enriched in repeats and palindromes, and thus is the most difficult to assemble component of the genome. Previously, expensive and labor-intensive BAC-based techniques were used to sequence the Y for a handful of mammalian species. Here we present a much faster and more affordable strategy for sequencing and assembling mammalian Y chromosomes of sufficient quality for most comparative genomics analyses and for conservation genetics applications. The strategy combines flow-sorting, short-read and long-read (Pacific Biosciences) genome and transcriptome sequencing, and droplet digital PCR with novel and existing computational methods. It can be used to reconstruct sex chromosomes in a heterogametic sex of any species. We applied our strategy to produce a draft of the gorilla Y sequence. The resulting assembly allowed us to refine gene content, evaluate copy number of ampliconic gene families, locate species-specific palindromes, examine the repetitive element content, and produce sequence alignments with human and chimpanzee Y chromosomes. Our results inform the evolution of the hominine (human, chimpanzee, and gorilla) Y chromosomes. Surprisingly, the gorilla Y exhibits strong conservation with the human Y. The chimpanzee Y is very different from both the human and gorilla Y chromosomes, despite the fact that chimpanzee and human shared the most recent common ancestor. We speculate about the reasons behind species-specific patterns of Y chromosome evolution. Moreover, we have utilized the assembled gorilla Y chromosome sequence to design genetic markers for studying the male-specific dispersal of this endangered species.

← Back

EFFECT OF 184 RISK VARIANTS FOR INFLAMMATORY BOWEL DISEASE ON THE GUT MICROBIOME IN HEALTHY INDIVIDUALS

Rob Mariman, Mahmoud Elansary, Julia Dmitrieva, Elisa Docampos, Ming Fang, Emilie Theatre, Wouter Coppieters, Latifa Karim, Michel Georges

Ulg, Unit of animal genomics, Liege, Belgium

Inflammatory Bowel Disease is a chronic inflammation of the intestinal tract that is caused by the interplay between poorly defined environmental factors and a large number of genetic risk variants, of which ~200 have now been identified. The precise molecular mode of action of the vast majority of genetic risk variants remains unknown. One possible mechanism is that - conditional on the present lifestyle in industrialized societies – some risk variants perturb the composition of the intestinal microbiome, which would trigger the development of the disease.

We took advantage of our CEDAR dataset to test this hypothesis. It comprises 16S-rRNA profiles generated with three independent amplicons for the terminal ileum, transverse colon and rectum of 323 healthy European individuals. The cohort has been genotyped using the Illumina OmniExpress array followed by imputation, yielding genotype data for > 2 million sites. We used PLINK to perform "mQTL" analyses, testing the association between the genotype at 184 established IBD risk variants and the abundance of 1051 bacterial species (as deduced from corresponding 16S read depths). We devised a z-score to summarize the strength of the association across anatomical locations and amplicons for each SNP-species combination. Empirical p-values of the summary statistic were determined using a permutation test accounting for "phenotypic" correlations. None of the associations were statistically significant (experiment-wide p-value < 0.05) when properly accounting for multiple testing. However, for 14 of the top 25 associations, the z-scores were remarkably consistent across locations and amplicons and these were retained for conformation in independent replication cohorts. Latest results will be presented that focus on understanding these findings in the context of dysbiosis in IBD. This approach may reveal novel connections between the genetic risk factors for IBD and the microbiota, thereby shedding new light on the pathogenesis and management of IBD.

EFFECTS OF MUTATION INFERRED FROM GENOMIC SEQUENCES

Debora S Marks

Harvard Medical School, Systems Biology, Boston, MA

Genomic sequences contain rich evolutionary information about functional constraints on macromolecules such as proteins. This information can be efficiently mined to detect evolutionary couplings between residues in proteins and address the long-standing challenge to compute protein three-dimensional structures from amino acid sequences. Substantial progress on this problem has become possible because of the explosive growth in available sequences and the application of statistical methods that use a global probability model. In addition to three-dimensional structure of single proteins, RNA and DNA, this statistical analysis of evolutionary constraints can identify functional residues involved in ligand binding, biomolecule-interactions, alternative ensembles of conformations and even "invisible" tertiary states of disordered proteins. In addition the model can be used to predict the effect of single and higher order mutations on organism fitness, a key bottleneck in understanding the effects of genetic variation and protein design. See evfold.org

COMPARATIVE STUDY OF THE THREE-DIMENSIONAL GENOMIC STRUCTURE IN HUMANS AND PRIMATES

François Serra[1,2], Yasmina Cuartero[1,2], Marina Brasso[2], Francisca Garcia[3,4], David Izquierdo[3,6], François Le Dily[1,2], Mario Caceres[3,6], Aurora Ruiz-Herrera[3,4], Arcadi Navarro[2,5,6], Tomàs Marques-Bonet[1,2,6], Marc A Marti-Renom[1,2,6]

[1]Centre Nacional D'Anàlisi Genòmica-Centre for Genomic Regulation, (CNAG-CRG), Barcelona, Spain, [2]Universitat Pompeu Fabra, (UPF), Barcelona, Spain, [3]Institut de Biotecnologia i de Biomedicina, Bellaterra, Barcelona, Spain, [4]Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain, [5]Institute of Evolutionary Biology, (CSIC-UPF), Barcelona, Spain, [6]Institució Catalana de Recerca i Estudis Avançats, (ICREA), Barcelona, Spain

How DNA is organized in three dimensions inside the cell nucleus and how this affects the ways in which cells access, read and interpret genetic information are among the longest standing questions in cell biology. The genome is hierarchically organized into chromosome territories, sub-chromosomal compartments, Topologically Associating Domains (TADs), and globules. Based on few comparative studies, this hierarchical organization is acceptably considered as conserved in mammals. However, a systematic analysis on the evolution of the structure of genomes is still lacking.
We have generated new Hi-C interaction maps of 50 kbp resolution for a panel of humans, all the apes and mouse. All Hi-C lymphoblast maps have been replicated using two biological samples per specie. Our new Hi-C maps not only identified genome assembly errors, (i.e., by detecting artefactual rearrangements with respect to the reference genomes) but were used to estimate the rate of change in nuclear organization. Our results quantitatively relate genome divergence to the conservation of the three main levels structure organization of the genome (that is, compartments, TAD borders and intra-TAD structure). First, the organization of the genome into compartments is highly conserved between primates and mouse (86-88% conserved between human and primates and 83% conserved between human and mouse). Second, the partition of the genome into well-defined TADs correlates with the genomic divergence between species (75-82% conserved between human and primates and 60% conserved between human and mouse). And third, for the first time our data shows a correlation between sequence conservation and intra-TAD structural organization between genomes. This results will help contextualizing the impact of genome structure conservation into the genomic evolutionary rate in wide spectra of mutations during the last 20M years of human evolution.

← Back

BIRTH, EXPANSION AND DEATH OF A HUMAN Y CHROMOSOME PALINDROME.

Andrea Massaia, Sandra Louzada, Juliet Handsaker,Yali Xue, Fengtang Yang, Chris Tyler-Smith

The Wellcome Trust Sanger Institute, Human Genetics Programme, Hinxton, Cambridge, United Kingdom

The male-specific portion of the human Y chromosome (MSY) covers 95% of the Y euchromatin and comprises different classes of sequence (X-degenerate, X-transposed, and ampliconic) with different evolutionary histories. Ampliconic regions harbor eight palindromes, which collectively cover 25% of the MSY euchromatin. These are inverted repeats showing over 99.9% similarity between arms, ranging in length from 30 kb to 3 Mb, and with a central unique-sequence spacer between 2 and 170 kb in length. Palindromes contain several testis-specific genes, and are prone to frequent gene conversion events between the arms, and structural rearrangements within human populations. However, the origins and long-term evolution of palindromes are still poorly understood.

The P8 palindrome is one of the smallest: it spans 75 kb on the q arm, including a 3.4 kb spacer, and harbors the two copies of the testis-specific *VCY* gene. A recent study of the Y chromosome diversity in the 1000 Genomes Phase 3 data identified several low-frequency, partially overlapping CNVs, in a 275 kb region which contains the P8 palindrome.

We have now performed molecular combing FISH and custom PCR assays to validate these rearrangements and identify their breakpoints at high resolution. The reference structure is altered, sometimes drastically, in at least eight different ways. These include a deletion of one arm (removing one copy of *VCY*) so that the structure in no longer palindromic, duplications of flanking sequence that consequently extend the palindrome arm length, and, in the most extreme cases, duplications that create a completely new palindrome, physically detached from but near the original P8, with a different structure that results in four copies of *VCY*. Interestingly, although all of these events appear recently in the Y phylogeny, some are shared between more than one individual suggesting that they are consistent with normal fertility.

Our work underlines the importance of experimental validation in CNV studies, as the variants we observe were not fully characterized with the analysis of sequencing data alone. Most importantly, our results confirm and reinforce the view of palindromes as rearrangement hotspots, revealing drastic remodeling including *de novo* palindrome formation in this genomic region, with each new structure most likely arising as a single mutational event.

DARWIN'S DOGS: GENETIC MAPPING OF COMPLEX BEHAVIORAL TRAITS IN MIXED-BREED DOGS

Jesse McClure[1], Diane P Genereux[1], Elinor K Karlsson[1,2,3]

[1]University of Massachusetts Medical School, Bioinformatics and Integrative Biology, Worcester, MA, [2]University of Massachusetts Medical School, Program in Molecular Medicine, Worcester, MA, [3]Broad Institute of Harvard and MIT, Cambridge, MA

Dogs are emerging as an excellent model organism for the study of behavioral and psychiatric disorders as they suffer from similar conditions to humans including compulsive disorders, Alzheimer's disease, and anxiety disorders. In many cases similar genetic pathways have been implicated in the dogs and humans suffering with these conditions. Thus a better understanding of behavioral genetics in dogs will give new insight into the biology of psychiatric diseases.

We have developed a new citizen-science based approach to dog behavioral genetics to enroll any dog, regardless of breed ancestry, allowing well powered studies to be done quickly and efficiently. By taking advantage of the rapid spread of handheld devices (smart phones and tablets) and building an online community, we have engaged unprecedented numbers of dog owners in the research as citizen scientists. We are also benefiting from large catalogs of genetic variation in dogs generated by next generation sequencing technology which has facilitated the design of a much denser Affymetrix SNP array by the dog genetics research community. We are combining a social resource (a community of engaged citizen scientists / dog owners), a scientific resource (a DNA and biological samples databank), and new genomic technology into a powerful new research tool.

The positive response from the dog owners has been far greater than we initially expected: in under six months we have enrolled over 3300 dogs for whom owners have cumulatively answered nearly 300 thousand behavioral survey questions. We are currently obtaining saliva samples from participants via easy-to-use collection kits in order to add high-density genotyping array data to this database of companion dog behavioral phenotypes. In addition, we have fully sequenced 21 dogs of mixed ancestry to assess the power for genetic mapping in this diverse population. Our analysis shows that GWAS in the mixed breed population is feasible using the new Affymetrix arrays.

In the long term, by engaging directly with dog owners, Darwin’s Dogs will be a powerful new resource for investigating genome function in health and disease, providing genomic data, detailed phenotypes and facilitating recruitment for specific GWAS projects.

← Back

A POPULATION-SPECIFIC REFERENCE PANEL EMPOWERS GENETIC STUDIES OF ANABAPTISTS THROUGH IMPROVED IMPUTATION

Liping Hou[1], Rachel L Kember[2], Jared C Roach[3], Jeffrey R O'Connell[4], David W Craig[5], Maja Bucan[2], Alan R Shuldiner[4], Francis J McMahon[1]

[1]Human Genetics Branch, NIMH Intramural Research Program, Bethesda, MD, [2]Perelman School of Medicine, Univ. Pennsylvania, Phila., PA, [3]Family Genomics Group, Institute for Systems Biology, Seattle, WA, [4]Dept. Medicine, Univ. Maryland School of Medicine, Baltimore, MD, [5]Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ

Genetic isolates offer many advantages for genetic studies, including decreased genetic heterogeneity and enrichment of otherwise rare alleles. However population-specific variants and linkage disequilibrium patterns may complicate studies that rely on imputation at ungenotyped sites, and genetic drift may dramatically alter allele frequencies. Here we report the Anabaptist Genome Reference Panel (AGRP), the first panel drawn specifically from the Amish and Mennonite populations, who participate in many genetic studies.

In a preliminary assessment of the panel's value for sequencing studies, actual sequencing calls were compared to imputed genotypes. Genome-wide SNP array and high-depth whole-genome sequencing data were generated on 265 individuals, uncovering >12M variants, 13% of which were novel. Genomes were phased with SHAPEIT2, and 120 unrelated individuals were extracted. Eighty individuals were selected at random as a reference panel while the remaining 40 were set aside as a test panel, where genotypes were masked at all variable sites not represented on the Illumina Human OmniExpress array. Masked sites were then imputed by IMPUTE2 against: 1) the AGRP and 2) 100 CEU individuals from the 1000G reference panel (phase 3). Imputation accuracy was measured as the squared correlation coefficient (r^2) between imputed allele dosages and masked genotypes.

Higher accuracy was achieved with imputation against the AGRP, especially for SNVs with minor allele frequencies (MAF) <1%. The greatest advantage was observed for rarer alleles. Alleles with MAF <0.5% were imputed at mean r^2 = 0.50 with the AGRP vs. 0.27 with the 1000G panel. The AGRP also enabled a greater proportion of rare alleles to be accurately imputed. At $r^2 > 0.5$, twice as many SNVs with MAF <0.5% could be imputed with the AGRP as with the 1000G panel. Many otherwise rare alleles showed substantially higher frequencies within the AGRP. Over 43K variants that were rare or absent in 1000G reached frequencies >5% in the AGRP, consistent with genetic drift among Anabaptists.

We conclude that the AGRP improves imputation of rare variants and estimation of population-specific allele frequencies among Anabaptists, potentially enhancing power for genetic association studies in Amish and Mennonite populations.

← Back

POPULATION GENOMICS OF THE INVASIVE 'ASH DIEBACK' PATHOGEN *HYMENOSCYPHUS FRAXINEUS*

Mark McMullan, Matt Clark

The Genome Analysis Centre, Plant and Microbial Genomics, Norwich, United Kingdom

Biological invasions are a natural process relevant today because their occurrence may be accelerated by human travel and climate change. *Hymenoscyphus fraxineus* is a fungal pathogen of European ash (*Fraxinus excelsior*) that is expected to kill the majority of infected trees in Europe. The pathogen was introduced to the EU from the Far East in 1992. In Europe, this pathogen is displacing a related fungus with a wholly saprotrophic lifestyle (*Hymenoscyphus albidus*) whereas *H. fraxineus* will enter the tree and cause 'dieback' of the crown. This dieback gets worse year on year and will eventually kill the tree.

A characteristic of importance to the success of any species is its level of standing genetic variation. A large amount of standing variation facilitates rapid adaptation to new environments. We have assembled and annotated a genome for this invasive pathogen and sequenced isolates from across Europe and in Japan. Here, we focus on population genomic analyses of isolates introduced to Europe as well as those sampled from Japan. We assess the level of introduced variation and we identify a population bottleneck (or founder effect) in Europe. We observe a signal of purifying selection operating on the majority of genes in the native range of Japan but despite the bottleneck we observe relative maintenance of genic polymorphism in the introduced population. We combine population genetic measures (i.e. nucleotide diversity, F_{ST}, Tajima's D) in order to identify candidate genes important to the success of this invasion. Finally we consider how these population genetic measures can be combined in a novel setting to better understand the biology of invasions.

UNDERSTANDING CARDIAC STRUCTURE AND FUNCTION IN HUMANS USING 4D IMAGING GENETICS.

Hannah V Meyer[1], Antonio De Marvao[2], Timothy J Dawes[2], Wenzhe Shi[2], Tamara Diamond[2], Daniel Rueckert[2], Enrico Petretto[2], Leonardo Bottolo[2], Declan P O'Regan[2], Ewan Birney[1], Stuart A Cook[2]

[1]European Bioinformatics Institute, Birney Research Group, Cambridge, United Kingdom, [2]Imperical College, MRC Clinical Sciences Center, London, United Kingdom

Human health is dependent on the long lasting function of many organ systems; these in turn develop due to complex genetic programs and are maintained over a lifespan. Many human diseases are related to cardiac structure and function, from relatively common cardiac infarctions through to more rare but serious diseases such as different cardiomyopathies. Understanding the biology of the human heart is informative for both basic and translational research.

We have created the first at scale cohort of 1,500 detailed cardiac images from healthy volunteers. We used a 1.5T Philips MRI scanner to acquire detailed 4D images of the heart in a single breath hold. This provides a far more detailed and consistent cardiac measurement than the traditional combination of 2D planar cardiac images. We are able to map all these 4D images into a consistent volumetric reference, and derive over 27,000 measurements per individual representing the heart. The individuals were also genotyped on a modern SNP array and imputed using a combination of 1000 Genomes and UK10K known variants, leading to 9.4 million variants for use in association studies.

We have successfully used a dimension reduction process to reduce the large image based metrics to a more compact latent variable space (100 dimensions). Using this lower-dimensional projection, we are able to find a number of genetic loci which show strong association with the heart structure. Interestingly, some of these hits are present in enhancers of known heart development genes, and pre-existing knockout studies in mice confirm a heart phenotype. Inspired by the model organism data, we have shown that a similar phenotype, measured as the non-compacted to compacted ratio in the heart at specific points, is also present in the human population. We are replicating this finding in other human heart cohorts.

This work shows that imaging genetics provides an invaluable, unbiased discovery process for exploring the underlying biology of human organs, with an impact on our understanding of both healthy and disease physiology.

PARENTAL CHOICES AND INITIAL RESULTS FROM A COMPREHENSIVE SEARCH FOR PREDICTIVE SECONDARY GENOMIC VARIANTS IN CHILDREN UNDERGOING WHOLE GENOME SEQUENCING

M Stephen Meyn[1,2,3,4], Nasim Monfared[1], Christian R Marshall[1,5,6], Daniele Merico[1], Dmitri James Stavropoulos[6], Raveen Basran[6], Robin H Hayeems[3,7], James Anderson[10], Michael Szego[8], Marta Girdea[1,9], Gary Bader[2], Michael Brudno[1,9], Ronald D Cohn[1,2,3,4], Stephen W Scherer[1,2,3,5], Randi Zlotnik-Shaul[8,10], Cheryl Shuman[1,2], Peter N Ray[1,2,6], Sarah C Bowdin[3,4]

[1]Hospital for Sick Children, Genetics and Genome Biology, Toronto, Canada, [2]University of Toronto, Molecular Genetics, Toronto, Canada, [3]Hospital for Sick Children, Centre for Genetic Medicine, Toronto, Canada, [4]Hospital for Sick Children, Clinical and Metabolic Genetics, Toronto, Canada, [5]Hospital for Sick Children, The Centre for Applied Genomics, Toronto, Canada, [6]Hospital for Sick Children, Paediatric Laboratory Medicine, Toronto, Canada, [7]Hospital for Sick Children, Child Health Evaluative Sciences, Toronto, Canada, [8]St Joseph's Health Centre, Centre for Clinical Ethics, Toronto, Canada, [9]University of Toronto, Computer ScienceToronto, Canada, [10]Hospital for Sick Children, Bioethics, Toronto, Canada

The SickKids Genome Clinic is a multidisciplinary research project that conducts clinical whole genome sequencing (WGS) for children who are undergoing genetic evaluations. With parents' permission, we systematically search childrens' genomes for diagnostic variants that explain the patient's known phenotype and predictive secondary variants (PSVs) associated with occult or future disease. Of 321 families approached to date, 54% agreed to participate and learn their child's childhood-onset PSVs. 58% of our participants also chose to learn their child's adult-onset PSVs. Parents who declined PSVs were most concerned about perceived psychological risks and/or insurance discrimination, while many who agreed to receive PSVs saw this as self-imposed obligation to take on a potential burden.
Bioinformatics analysis of our first 100 patients identified 87 copy number variants, 571 structural variants and 2957 SNVs as potentially reportable PSVs. Two thirds of these PSVs were listed in HGMD, but manual assessment eliminated >90% of these PSVs as well as all copy number and structural variants.
Using 2015 ACMG variant classification guidelines we found pathogenic/likely pathogenic PSVs in an 2013 ACMG reportable gene in 8/100 children. Expanding our search 50 fold to include 2800+ disease genes listed in the NIH Clinical Genomic Database yielded 27 more PSVs. Including 7 additional PSVs that fell just short of "likely pathogenic", we found ~1/3 of children (34/100) had at least 1 PSV, including 4 children with2 PSVs and 2 children with 3 PSVs.
39/42 PSVs were associated with autosomal dominant phenotypes. Importantly, all reportable PSVs were inherited and two thirds of PSVs had MAFs <0.01%, indicating that the majority of PSVs were private familial mutations. Return of PSVs to parents, comparison of PSVs to a standardized three-generation pedigree, and assessment of PSV penetrance is currently underway.

← Back

ASELUX: AN ULTRA FAST AND ACCURATE ALLELIC READS ALIGNER

Zong Miao[1,2], Arthur Ko[1,3], Marcus Alvarez[1], Markku Laakso[4], Päivi Pajukanta[1,2,3]

[1]UCLA, Dept. of Human Genetics, Los Angeles, CA, [2]UCLA, Bioinformatics Interdepartmental Program, Los Angeles, CA, [3]UCLA, Molecular Biology Institute, Los Angeles, CA, [4]University of Eastern Finland, Dept. of Medicine, Kuopio, Finland

A large proportion of genes exhibit allele specific expression (ASE). Although several methods have been developed to identify ASE events from RNA-sequence (RNA-seq) data, mapping bias i.e. preferential alignment to the reference allele remains a major obstacle in ASE analysis. Since most simulation-based approaches are designed for single-end reads, they do not serve well the current paired end RNA-seq data sets. The allele-aware alignment tools such as SNP-o-matic and GSNAP are accurate but relatively slow, making them not ideal for the analysis of large RNA-seq data sets. Thus, there is an urgent need to develop computationally efficient ASE analysis tools. We developed a new software, ASElux, to align allele-specific reads fast and accurately. Since only about 10% of RNA-seq reads are allele-specific, a significant amount of time can be saved by disregarding non-allele specific reads during the alignment. Inspired by pseudoalignment (Bray et al. 2015), our new approach ASElux uses personal SNP information to generate all possible allele-specific reads before alignment, followed by direct counting of allele-specific reads instead of identifying them from alignment results. Only allele-specific reads are then globally aligned to exclude multi-alignment. Testing simulated RNA-seq data, ASElux (25000 queries/s) is 56 times faster than GSNAP (440 queries/s) while maintaining the same high mapping accuracy. We also tested the software on 20 adipose RNA-seq samples from the METSIM study. Treating the GSNAP as the standard, we compared ASElux with an RNA-seq alignment tool, STAR in ASE analysis by processing the data separately with GSNAP, STAR, and ASElux. Using the proportion of expression of reference allele compared with the total expression at the SNP site as an indicator of ASE, the average difference between ASElux and GSNAP is 0.5% which is significantly smaller ($p<10^{-5}$) than the average difference between GSNAP and STAR (2.9%). Furthermore, when we used Bonferroni corrected binomial test to identify ASE SNPs after alignment, only 55% of the ASE SNPs discovered by STAR overlapped with GSNAP whereas the overlap between ASElux and GSNAP increased to 70%. The results indicate that ASElux has the great accuracy of allele-aware alignment while it maintains the same high alignment speed as STAR. Thus, ASElux will help us better understand frequent ASE events in extensive RNA-seq studies.

IOBIO DEV KIT: RESOURCES FOR MAKING GENOMIC, REAL-TIME WEB APPLICATIONS AND SERVICES

Chase A Miller, Yi Qiao, Tony DiSera, Alistair Ward, Gabor T Marth

University of Utah, Human Genetics, Salt Lake City, UT

IOBIO (http://iobio.io) is an open-source web-based genomic platform enabling real-time analysis of large, remotely-stored, distributed datasets, with analysis algorithms operating as web servers on data-streams in standardized data formats. Several IOBIO apps have been previously released and can be found at http://iobio.io/applications.html. Here we introduce the **IOBIO Dev Kit** (https://developer.iobio.io), which provides libraries, examples, and documentation to build your own IOBIO apps and services. The dev kit consists of three main libraries: iobio.js is a client-side javascript library that streamlines creating and executing iobio commands, while hiding complicated web-socket connection code; iobio.viz is a d3-based, streaming genomic visualization library that consumes iobio data streams; and minion is a server library that wraps (with only a few lines of code) standard command line tools and converts them into iobio web services that can be mixed and matched with all other iobio web services. As an instructive example, we have rebuilt our bam.iobio.io app (http://bam.iobio.io) using dev kit components with all code found here: https://github.com/chmille4/bam.iobio.io. Using the IOBIO Dev Kit, developers will be able to create new genomic analysis apps on the web at a fraction of the time it currently takes to build desktop applications: IOBIO will be no less than an ecosystem for genomic analysis.

THE GENOMIC LANDSCAPE OF EVOLUTIONARY CONVERGENCE IN AMNIOTES

Dan Mishmar, Levin Liron

Ben-Gurion University of the Negev, Department of Life Sciences, Beer-Sheva, Israel

Many ancient genetic variants (i.e. Nodal mutations) are predicted to bare high functionality. However, it is tough to determine the reason underlying their long-term survival: either due to functional compensation, or due to their adaptive value and hence, positive selection. Here, we addressed this problem by focusing on recurrent, highly functional, ancient mutations. To this end, we screened for functional nodal mutations (fNMs) in the mitochondrial (mtDNAs) and nuclear (nDNA) genomes of 1003 and 92 species, respectively, representing the entire amniote phylogeny (i.e., Mammals, Birds and Reptiles). To identify candidate compensatory mutations for fNMs we focused on fNMs occurring in proteins with resolved 3D structures (N=~3400 PDBs), and identified amino acid changes occurring in close structural proximity (5 Å) to the fNMs, while sharing the same phylogenetic branches. We found that mtDNA fNMs, displayed higher propensity for compensation (>50%) than the nDNA fNMs (27%). Similar results were obtained by screening for compensatory mutations in mtDNA-encoded RNA genes by measuring the propensity to maintain thermostability. Thus, our analyses suggest that functional compensation play important, though not an absolute, role in the long-term survival of many mtDNA and nDNA fNMs. Nonetheless, a large proportion of fNMs have apparently not been compensated, thus highlighting them as optimal candidates for adaptiveness. Therefore, using these un-compensated mutations enabled us to reveal parsimoniously unrelated lineages enriched for recurring fNMs, thus implying convergence. Strikingly, birds and mammal shared the most fNMs. Among these recurrent fNMs we identified mutations that occurred in thermoregulation-related genes, thus being best candidates to explain the molecular basis of convergent thermoregulation in birds and mammals. Our findings suggest that functional compensations and adaptive properties equally explain the long-term survival of fNMs and that un-compensated fNMs could be used to unravel adaptive signatures in general, and evolutionary convergent events in particular.

FAMILY AND POPULATION-BASED GENOTYPE IMPUTATION IN FINLAND

A Mitchell[1], P Gormley[2,5], M Kurki[2,5], D Lal[2,5], M Hiekkala[3], P Happola[3], P Palta[3], I Surakka[3], E Hamalainen[3], M Kaunisto[3], M Wessman[3], M Kallela[4], S Ripatti[3], H Runz[1], A Palotie[2,3,5]

[1]Merck Research Labs, Boston, MA, [2]Broad Institute of Harvard and MIT, Boston, MA, [3]University of Helsinki, Helsinki, Finland, [4]Helsinki University Central Hospital, Helsinki, Finland, [5]Mass General Hospital, Harvard Medical School, Boston, MA

Imputation of whole genome sequence (WGS) data in individuals who have been sparsely genotyped can increase information obtained from a sample at a fraction of the cost. The choice of a reference population affects the quality of the imputed genotypes. Previous work has demonstrated that in a founder population, addition of a small population-specific reference set to a larger, more diverse panel can improve the accuracy and sensitivity of imputation. Here, we evaluated the impact of adding a diverse reference panel to a population-specific panel. We also evaluated family-based imputation.

We obtained DNA from 9,200 people in 2,029 kindreds from across Finland. All were genotyped on Illumina Infinium CoreExome or Psych Array, which share the HumanCore backbone and contain 480,000 variants in common. WGS was performed on 631 of the genotyped individuals from 184 of the families.

Population-based imputation was performed using ShapeIt and IMPUTE2. Two reference panels were used: 1) low coverage (4.6x) WGS Finnish panel (N=1,941), and 2) Finnish panel plus 1000 Genomes Phase 3 (N=2504) with low coverage (7.4x) WGS and high coverage (65.7x) WES. Both panels performed well for variants with MAF > 0.1%. However, with the combined panel, there was a 38% increase in high confidence calls (INFO > 0.9) relative to the Finnish panel alone. The most dramatic difference was observed for variants with MAF < 0.1% (2.4-fold increase) and MAF 0.1% to 0.5% (1.8-fold increase). Median concordance between masked and imputed genotypes was 0.994 for the Finnish panel and 0.999 for the combined panel.

GIGI was used for family-based imputation. Quality and completeness were high for family members with two sequenced parents, but both dropped sharply with increasing distance from the nearest sequenced family members. The proportion of individuals sequenced in most families was insufficient to allow high quality imputation in the extended families.

To summarize, including 1000G Phase 3 reference panel along with the Finnish panel improved performance. The larger panel provided higher sensitivity to impute a larger number of variants and the inclusion of non-Finnish alleles among the reference genomes did not negatively affect the accuracy of the genotype calls. Inclusion of inheritance information in related individuals may increase the quality of genotype calls for variants with MAF < 0.1%.

ESTIMATING TOLERATED GENETIC VARIATION IN GENE EXPRESSION FROM ALLELIC EXPRESSION DATA

Pejman Mohammadi[1,2], Stephane E Castel[1,2], Heather E Wheeler[3], Hae Kyung Im[4], GTEx Consortium[1], Tuuli Lappalainen[1,2]

[1]New York Genome Center, -, New York, NY, [2]Columbia University, Department of Systems Biology, New York, NY, [3]Loyola University Chicago, Departments of Biology and Computer Science, Chicago, IL, [4]University of Chicago, Department of Medicine, Chicago, IL

Several approaches have been presented to assess the tolerance of coding genes against functional impact of genetic variation. The lack of similar quantifications of regulatory constraint in genes is a problem given the important role of regulatory variants in disease, and the need for interpretation of putative disease-causing variants discovered in growing whole genome sequencing data sets.

Allelic expression analysis is an elegant integrative approach for capturing the effect of cis-regulatory variants, and it is minimally obscured by other confounding factors. Here we present a generative probabilistic model to describe allelic expression data as net outcome of a set of unobserved regulatory variants, and use it to estimate the heritable variation in gene expression introduced by cis-regulatory variation in population data. Our simulations demonstrate that the model is robust and accurate across a spectrum of allele frequencies, regulatory complexity, and gene expression levels.

We apply this method to allelic expression data from 499 individuals in GTEx data to estimate heritable fraction in gene expression and regulatory tolerance of over 10,000 coding genes in each of the 49 tissues. We observed that genes with lower tolerance to cis-regulatory variation tend to have lower tolerance to coding mutation as assessed by the RVIS score ($p<10^{-56}$), and are also more likely to be haploinsufficient ($p<10^{-62}$). We observe lower tolerance in genes expressed in the brain, suggesting increased evolutionary constraint. Furthermore, known transcription factor targets, as well as genes in plasma membrane and signaling pathways showed signs of higher constraint, in contrast to tolerant pathways such as translation and NMD. Gene expression heritability estimates in GTEx samples derived from our method appear better than standard expression-based estimates from the same samples, assessed by correlation to heritabilities estimated from much larger DGN cohort data ($\rho=0.40$ vs. $\rho=0.20$, respectively). We conclude that our method can accurately quantify cis-regulatory variation from allelic expression data. The obtained scores for tissue-specific intolerance for regulatory variation are valuable for interpretation and follow up analysis of disease-associated variants, including prioritization of rare variants observed in whole genome sequencing data.

← Back

EDGY: EXPORT OF DATA FROM GALAXY TO YABI, AUTOMATED WORKFLOW TRANSFER TO COMMAND LINE TOOLS

David C Molik, Ying Jin, Molly Hammell

Cold Spring Harbor Lab, Bioinformatics Shared Resource, Cold Spring Harbor, NY

The web-based bioinformatics platform, Galaxy gained popularity as a tool for allowing access to powerful compute clusters and sophisticated bioinformatics software with user-friendly point-and-click interfaces[1]. In contrast, the command line will always be more efficient and flexible for those who are comfortable working within an Unix-like environment. Likewise, Biologists who understand the computing environment are more likely to gauge what is and is not computationally feasible[2]. Therefore, an obstacle that the bioinformatics community faces is training users with little programming experience to be comfortable working at the command line. We have explored frameworks that would allow users to design their analysis workflows in a web browser environment, then transfer these workflows into pipelines suitable for running at the command line.

Yabi is an alternate web-based bioinformatics platform designed by the Center of Comparative Genomics at Murdoch University[3]. It fulfills many of the same operations as Galaxy, and is interoperable with the same tools. Yabi provides a similar user experience, providing a graphical user interface (GUI) to bioinformatics software. Moreover, because Galaxy and Yabi both use simple configuration files, many of the tool interfaces designed for Galaxy can be transferred to the Yabi format. However, Yabi additionally offers a command line tool, yabish, where users can use their web-designed workflows at the command line. Galaxy workflows can be exported to command line scripts, if the user can access the tools referenced by Galaxy. Workflow export can be an intermediate step between offering a GUI to the end user and having users do all of their analysis on the command line, while providing the benefits of logging, saved workflows, and remote data.

We implemented Yabi as an analogous software application to Galaxy and provide contrasting benefits of both platforms. Moreover, we present EDGY, which parses and reformats tool configuration files for use in either the Galaxy, Yabi, or command line format, providing tools that can help biologists combine GUI-based pipeline design with the flexibility of the command line on shared compute cluster environments.

[1]Goecks, J, A Nekrutenko, and J Taylor. 2010. “Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences.” Genome Biol.
[2]Dudley, JT, and AJ Butte. 2009. “A Quick Guide for Developing Effective Bioinformatics Programming Skills.” PLoS Comput Biol.
[3]Hunter, AA, and AB Macgregor. 2012. “Yabi: An Online Research Environment for Grid, High Performance and Cloud Computing.” Source Code for Biol.

THE GENOMIC ANALYSIS OF THE ANDAMANESE GIVES A NEW INSIGHT ON THE SPREAD OF HUMANS IN ASIA

Mayukh Mondal*[1], Ferran Casals*[2], Zheng Huang*[3], Analabha Basu[4], Giovanni M Dall'Olio[5], Marc Pybus[1], Mihai G Netea[6], David Comas[1], Hafid Laayouni[1,7], Qibin Li**[3], Partha P Majumder**[4], Jaume Bertranpetit**[1]

[1]Universitat Pompeu Fabra, Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Spain, [2]Universitat Pompeu Fabra, Servei de Genòmica, Barcelona, Spain, [3]BGI, BGI, Shenzhen, China, [4]National Institute of Bio-Medical Genomics, National Institute of Bio-Medical Genomics, Kalyani, India, [5]King's College of London, Division of Cancer Studies, London, United Kingdom, [6]Radboud University Medical Center, Department of Internal Medicine, Nijmegen, Netherlands, [7]Universitat Autonòma de Barcelona, Departament de Genètica i de Microbiologia, Barcelona, Spain

To shed light on the peopling of South Asia and the specificity of morphological adaptations, we analyzed whole genome sequences of 10 Andamanese individuals and compared them with 60 individuals drawn from populations of mainland India with different ethnic histories and other publicly available data. We show that all Asian and Pacific populations have a single origin, contradicting an earlier postulation of independent origins of East Asians and Andamanese (along with other Pacific populations). We also show that populations of South Asia, Southeast Asia and Oceania harbor a small proportion of hominin ancestry in their genomes, which is absent in Europeans or East Asians. The footprints of adaptive selection in the genomes of the Andamanese show that their characteristic distinctive phenotypes (including a very short stature) are not reflective of an ancient African origin, but the result of strong natural selection on genes related to human morphology.

* co-first authorship
** co-senior authorship

← Back

HUMAN VARIATION IN microRNA BIOGENESIS AND DISEASE

Jonathan Moody, Grzegorz Kudla, Caroline Hayward, Javier Caceres, Martin Taylor

Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom

MicroRNAs are short RNA molecules that are central to the regulation of many cellular and developmental pathways, and are often disregulated in cancer. They are processed from structured precursors in the nucleus to produce pre-miRNA and they are subsequently processed into mature microRNAs in the cytoplasm. Rare anecdotal examples of DNA sequence changes in the microRNA precursors have been found to impact processing efficiency and the abundance of the mature microRNA. We have systematically screened for both rare and common polymorphisms that overlap microRNA precursors and related them to mature microRNA levels as measured in short RNA sequencing. We find both rare and common SNPs within the precursors and mature microRNAs that are associated with the mature microRNA level, several of which have been observed as risk factors in cancer or other clinically relevant traits. We relate the sequence and structural context of the change to candidate alterations in processing and in aggregate over all microRNAs investigate how selection has shaped the distribution of variation.

SINGLE-CELL AND REAL-TIME EPITRANSCRIPTOMICS REVEALS NOVEL MECHANISMS OF CELL INDIVIDUALITY AND MEMORY

Leonid L Moroz[1], Maria Basanta Sanchez[2], Igor Lednev[2], Andrea B Kohn[3]

[1]University of Florida, Neuroscience, Gainesville, FL, [2]University of Albany, The RNA Institute, Albany, NY, [3]University of Florida, The Whitney Lab, St. Augustine, NY

Post-transcriptional changes in RNA have the potential to influence the epigenetic landscape. There are over a hundred RNA modifications. These chemical changes to nucleotides do not alter the sequence of RNA but can alter gene expression and has recently been described as part of a so-called epitranscriptome. Methylation of adenosine to N6-methyladenine (m6A) is the most prevalent internal modification on mRNA and long non-coding RNA with up to 20% of the human mRNA methylated. However, little is known about biological roles of RNA modifications and cell-specificity of the process. Here, we used Aplysia californica, a prominent model organisms to study learning and memory, and directly chemically characterized the scope of the epitranscriptome focusing at the level of individual functionally identified neurons. We determine that 4% of the total RNA was methylated in Aplysia. RNA-seq and computation analysis of the Aplysia genome shows that all the enzyme families involved in RNA methylation/demethylation are present in varying amounts from developmental states to single neurons following neuroplasticity tests. For the first time RNA modifications were identified and quantified in single neurons using ultrasensitive mass spectrometry (MS) in combination ultra-high performance liquid chromatography. Different cholinergic neurons in Aplysia showed different and distinct RNA modifications. In addition to the RNA modification, we also discover cell- and tissue specific RNA editing in Aplysia. This adds another layer to the epigenetic regulation forming a complex epitranscriptomic landscape. RNA editing is a process of targeted alterations of nucleotides in all types of RNA molecules (e.g., rRNA, tRNA, mRNA, and miRNA). As a result, the transcriptional output differs from its genomic DNA template. RNA editing can be defined both by biochemical mechanisms and by enzymes that perform these reactions. We found full a repertoire of predicted RNA-editing enzymes in the genome of Aplysia including several expansions of enzyme families compared to what has been described in mammals. In summary our data indicate that both single-cell transcriptome and epitranscriptomes are unique features of identified neurons providing novel insights into genomic bases of neuronal individuality and plasticity.

COGNITIVE ANALYSIS OF GWAS SCHIZOPHRENIA RISK GENES THAT FUNCTION AS EPIGENETIC REGULATORS OF GENE EXPRESSION

Laura Whitton[1], James Walters[2], Dan Rujescu[3], Michael Gill[4], Aiden Corvin[4], Stephen Rea[5], Gary Donohoe[1], Derek W Morris[1]

[1]Cognitive Genetics and Cognitive Therapy Group, School of Psychology and Discipline of Biochemistry, National University of Ireland Galway, Galway, Ireland, [2]MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, United Kingdom, [3]Department of Psychiatry, University of Halle, Halle, Germany, [4]Neuropsychiatric Genetics Research Group, Discipline of Psychiatry, Trinity College Dublin, Dublin, Ireland, [5]Centre for Chromosome Biology, Biochemistry, National University of Ireland Galway, Galway, Ireland

Epigenetic mechanisms govern the heritable and dynamic means of regulation of various genomic functions, including gene expression, to orchestrate brain development, adult neurogenesis and synaptic plasticity. These processes when perturbed are thought to contribute to the pathophysiology of schizophrenia. A core feature of schizophrenia is cognitive dysfunction in the form of poor memory, attention and IQ. For genetic disorders where cognitive impairment is more severe such as intellectual disability (ID), there are a disproportionally high number of genes involved in the epigenetic regulation of gene transcription. Evidence now supports some shared genetic aetiology between schizophrenia and ID. Therefore, it is pertinent to ask if an overlapping phenotypic characteristic of schizophrenia and ID (cognitive deficits) could be due to genetic variability in a shared pathobiology (epigenetic mechanisms). Our overall hypothesis is that genome wide association studies (GWAS) identified risk genes for schizophrenia that regulate epigenetic mechanisms are associated with cognitive deficits in schizophrenia. GWAS have identified 108 chromosomal regions associated with risk of schizophrenia, implicating 350 genes. The aim of this study was to identify risk genes for schizophrenia with epigenetic functions and test these genes for association with cognitive deficits in schizophrenia. We developed a list of 350 genes with epigenetic functions and cross-referenced this with the schizophrenia GWAS loci. This identified 8 candidate genes: BCL11B, CHD7, EP300, EPC2, GATAD2A, KDM3B, RERE and SATB2 . Using an Irish dataset of psychosis cases (n=905) and controls (n=330), the schizophrenia risk SNP at each gene was tested for effects on IQ, working memory, episodic memory and attention. Strongest associations were for rs6984242 and both with measures of IQ (p=0.001) and episodic memory (p=0.007). We link rs6984242 to the gene CHD7 via a long range expression quantitative trait locus. These associations were not replicated in smaller independent UK and German samples. Our study highlights that a number of newly identified risk genes for schizophrenia function as epigenetic regulators of gene expression but further studies are required to establish a role for these genes in cognition.

A MULTIKERNEL MACHINE APPROACH FOR MULTI-OMIC ANALYSIS IN CONTEXT OF ALZHEIMER'S DISEASE

Bernard Ng[1], Hans Klein[2], Ellis Patrick[2], Charles White[2], Jishu Xu[2], Lori Chibnik[3], Chris Gaiteri[4], David A Bennett[4], Philip L De Jager[3], Sara Mostafavi[1]

[1]University of British Columbia, Statistics; Medical Genetics, Vancouver, Canada, [2]Harvard University, Neurology, Boston, MA, [3]Harvard School of Public Health, Epidemology, Boston, MA, [4]Rush University, Neurological Sciences, Chicago, IL

Introduction: Joint analysis of genomics data at multiple cellular resolutions, including the genome, epigenome and transcriptome, promises a path forward for elucidating the regulatory mechanisms associated with complex traits and diseases. However, combining datasets from heterogeneous genomic assays poses significant computational and statistical challenges. The number of variables associated with each data type already imposes a difficult multiple testing problem for standard univariate analyses. Naively considering nonlinear relationships between genetic variation and downstream cellular traits would further lead to an explosion in the number of hypotheses, resulting in little statistical power.

Method: We present an approach based on multikernel machines (MKM) for associating genotype, multiple types of cellular traits, and their interactions to complex traits. Kernel machines (KM) can be shown to be equivalent to linear mixed models, which allows a set of variables, e.g. a set of genetic variants associated with a gene, to be modeled as a single random effect. With this approach, we build a kernel for each data type as well as kernels that model their interactions by employing an extension of KM known as MKM. Importantly, our approach applies MKM in a gene-by-gene manner by jointly modeling all genomic variables that are spatially proximal (cis) to each gene, enabling us to derive regulatory mechanisms that underlie trait-associated loci.

Results: Multiple types of genomics data from the ROS/MAP studies were used in this work, which comprise genotype, gene expression, DNA methylation, and histone modification data derived from the dorsolateral prefrontal cortices of 400 subjects phenotyped for traits related to clinical and pathological aspects of Alzheimer's disease (AD). In addition to significant main effects of each data type, by applying our MKM approach to these data we identified a few loci where the interactions between genotype and a cis cellular trait were found to be significantly associated with AD related traits. We also observed that gene expression only partially mediates the impact of genetic and epigenetic factors on AD. In summary, our analysis shows that both additive and non-additive relationships between regulatory factors contribute to AD and other related complex traits.

TOWARDS MEASURING NUCLEAR DOMAINS FROM HI-C DATA

Yuichi Motai, Shinichi Morishita

The University of Tokyo, Department of Computational Biology and Medical Sciences, Kashiwa, Chiba, Japan

Genome-wide chromosome conformation capture (Hi-C) has enabled to construct spatial proximity maps among genome. The maps, however, do not give direct information on three-dimensional (3D) genome organization and it is challenging to understand 3D structures corresponding to complex patterns in the maps. The reconstruction of 3D genome structure from Hi-C data is one promising way to facilitate a geometric interpretation of the Hi-C data; however, it still remains unexplored how to characterize biologically interpretable structures in the reconstructed 3D model of genome, motivating us to develop a method to determine 3D structural features from the model.

With this method, we can capture structures interpreted as nuclear domains such as loops of chromatin fibers, topologically associated domains (TAD), heterochromatin, nuclear bodies and interchromatin compartments in a 3D model of genome, and furthermore can measure some geometric feature values of each characteristic structure such as size, shape and chromatin density. Because even one kind of nuclear domains has various size and the spatial extent of a coarse-grained 3D model is unknown, the method has been designed to detect the variously sized structures from a series of 3D models with continuously changing volume in real number at a single calculation by using persistent homology. The found structures can be associated with many kinds of data mapped on reference genome (e.g., DNA modifications, histone modifications, gene expression levels, etc.) by localizing them on each structure. That association enables to provide new interpretation of Hi-C data and other genomic data, and generate new hypotheses of relation between chromatin organization and functional processes in nucleus.

We applied this method to human Hi-C data and examined the relation between the detected structural features and genome annotation data such as gene composition, gene expression and histone marks.

THE DISCOVERY OF OVER 100 NOVEL HUMAN PROTEIN-CODING GENES BASED ON CONSERVATION, NEXT GENERATION TRANSCRIPTOMICS AND MASS SPECTROMETRY

Jonathan M Mudge[1], Adam Frankish[1], Toby Hunt[1], James Wright[2], Jyoti Choudhary[2], Jennifer Harrow[1]

[1]Wellcome Trust Sanger Institute, Computational Genomics, Hinxton, United Kingdom, [2]Wellcome Trust Sanger Institute, Proteomics and Mass Spectrometry, Hinxton, United Kingdom

While the question *how many protein-coding genes are in the human genome?* may seem like a relic of the late 90s, it remains unanswered in spite of the enormous effort expended by gene annotation projects. In fact, the number of recognized protein-coding genes has steadily decreased over recent years, largely due to the manual removal of spurious *in silico* CDS predictions. As part of the GENCODE project, we have begun a major push to finalize our set of protein-coding genes. Central to this workflow is the manual analysis of thousands of strictly processed phyloCSF signals that do not correspond to GENCODE coding sequences - i.e. comparative annotation based on evolutionary conservation - coupled with the usage of next-generation RNA sequencing data, especially long-read data from PacBio and SLR-seq data from Tilgner *et al.* In parallel, we have identified or confirmed several dozen novel coding loci based on high-stringency mass spectrometry data. In particular, we have subjected the millions of raw spectra obtained by Kim *et al*. and Wilhelm *et al*. during their production of 'draft human proteome maps' to a complete reanalysis, based on the use of multiple search algorithms combined with more appropriate significance filtering. It is notable that we do not find any peptide support for novel CDS that do not also exhibit strong conservation. Currently, novel protein-coding genes, coding exons or pseudogenes are being identified with high confidence on a daily basis. Furthermore, most of our novel protein-coding genes have not been previously reported, and it is clear that many remain essentially 'invisible' to purely computational high-throughput analytical pipelines. Nonetheless, these loci were also missed by the initial GENCODE manual annotation phase, most typically because either their CDS are very short (more than half are under 200aa) or their expression is highly restricted (as demonstrated through CAGE profiling). Our efforts thus demonstrate the necessity of a fully integrated annotation workflow.

On behalf of the GENCODE consortium.

SELECTIVE SWEEPS ACROSS TWENTY MILLIONS YEARS OF HUMAN EVOLUTION

Kasper Munch[1], Kiwoong Nam[2], Mikkel H Schierup[1], Thomas Mailund[1]

[1]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark, [2]Université Montpellier 2, UMR INRA, Montpellier, France

The contribution from selective sweeps to variation in genetic diversity has proven notoriously difficult to assess, in part because polymorphism data only allow detection of sweeps in the most recent few hundred thousand years. Here we show how linked selection in ancestral species can be quantified across evolutionary timescales from patterns of incomplete lineage sorting (ILS) along the genomes of closely related species. ILS is ubiquitous on the internal branches of the great apes species tree, and the two species trios investigated here (human, chimpanzee, gorilla and human, orangutan, gibbon) both show high levels of ILS. We show that sweeps in the human-chimpanzee and human-orangutan ancestors can be identified as depletions of ILS in regions in excess of 100 kb in length. By comparing the ILS patterns along the genomes of the closely related human-chimpanzee and human-orangutan ancestors we are further able to quantify the impact of selective sweeps relative to that of background selection. We show that sweeps predicted in each ancestral species, as well as recurrent sweeps predicted in both species, often overlap sweeps predicted in humans. This suggests that many genomic regions experience recurrent selective sweeps. We conclude that sweeps are more important than background selection in reducing diversity along the genome and that sweeps have reduced diversity in the human-chimpanzee ancestor much more than in the human-orangutan ancestor.

MODELING ANCESTRY-DEPENDENT PHENOTYPIC VARIANCE REDUCES BIAS AND INCREASES POWER IN GENETIC ASSOCIATION STUDIES

Shaila Musharoff[1], Scott Huntsman[1], Celeste Eng[1], Esteban G Burchard[1,2], Noah Zaitlen[1]

[1]University of California San Francisco, Department of Medicine Lung Biology Center, San Francisco, CA, [2]University of California San Francisco, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA

Many complex human phenotypes vary dramatically in their distributions between populations. Genetic association studies typically use estimates of ancestry, such as principal components (PCs), as fixed-effect covariates to prevent confounding caused by a dependence of phenotypic mean on ancestry. However, the standard approach of including PC covariates in linear regression models assumes that different populations have the same phenotypic variance, which may not hold for recently admixed populations. In this work we consider the possibility that populations with differences in phenotypic mean also have differences in phenotypic variance. First we show this is the typical case under an additive genetic architecture. Then we develop a new likelihood-based method, based on a double generalized linear model, to account for relationships between ancestry and phenotypic variance in genetic association studies. In simulations, our test has better power than several linear regression tests that assume equal variance across groups. We observe power increases of 12 - 66% and obtain unbiased parameter estimates for data simulated with realistic effect sizes and population minor allele frequency differences of 0.45. Furthermore, we show that the current gold standard approach of linear regression with PC covariates can lead to inflation or deflation of p-values for tests of genetic association when population phenotypic variances differ. For example, simulated populations with minor allele frequencies of 0.05 and 0.5 produce test statistics with an inflation factor (lambdaGC value) of 1.56, which our method fixes.

We apply our method to the Study of African Americans, Asthma, Genes and Environments (SAGE) and find significant associations of baseline lung function (i.e. before treatment with asthma drugs) with estimated global ancestry proportion when either sex or log(BMI) is a sole covariate. By contrast, when using a standard linear regression (which assumes equal variance among groups), this association is not significant with either sex or log(BMI) as a sole covariate and requires additional covariates (age, sex, height, and weight) for significance. Our method finds significant associations with fewer covariates, possibly because phenotypic variances as well as phenotypic means differ among groups, and is promising for other studies.

COMPLEX REARRANGEMENTS AND ONCOGENE AMPLIFICATIONS REVEALED BY SINGLE MOLECULE DNA SEQUENCING OF A HIGHLY REARRANGED CANCER CELL LINE

Maria Nattestad[1], Sara Goodwin[1], Karen Ng[2], Timour Baslan[3], Fritz Sedlazeck[1,6], Tyler Garvin[1], James Gurtowski[1], Elizabeth Hutton[1], Elizabeth Tseng[4], Jason Chin[4], Timothy Beck[2], Yogi Sundaravadanam[2], Melissa Kramer[1], Eric Antoniou[1], John McPherson[4], James Hicks[1], Michael C Schatz[1,6], William R McCombie[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Ontario Institute for Cancer Research, Cancer Genomics, Toronto, Canada, [3]Memorial Sloan Kettering Cancer Center, Cancer Biology and Genetics Program, New York, NY, [4]Pacific Biosciences, Bioinformatics, Menlo Park, CA, [5]University of California Davis, Comprehensive Cancer Center, Davis, CA, [6]Johns Hopkins University, Computer Science, Baltimore, MD

Genomic instability is one of the hallmarks of cancer, leading to widespread copy number variations, chromosomal fusions, and other structural variations in many cancers. The breast cancer cell line SK-BR-3 is an important model for HER2+ breast cancers, which are among the most aggressive forms of the disease and affect one in five cases. Through short read sequencing, copy number arrays, and other technologies, the genome of SK-BR-3 is known to be highly rearranged with many copy number variations. However, these technologies cannot precisely characterize the nature and context of the identified genomic events and other important mutations may be missed altogether because of repeats, multi-mapping reads, and the failure to reliably anchor alignments to both sides of a variation.

To address these challenges, we have sequenced SK-BR-3 using PacBio long read technology. We generated more than 70X coverage of the genome with average read lengths of 9-13kb (max: 71kb). To develop a detailed map of structural variations in this highly aberrant genome, we designed two new methods of variant-calling for long reads: Sniffles using split-read alignment, and Assemblytics using an assembly-based consensus strategy. In addition, we have developed an algorithm named SplitThreader to evaluate variants and reconstruct the mutational history, which allowed us to discover a complex series of nested duplications and translocations between chr17 and chr8, two of the most frequent translocation partners in primary breast cancers. We have also carried out full-length transcriptome sequencing using PacBio's Iso-Seq technology, which has revealed a number of previously unrecognized isoforms as well as enabled tracing of gene fusions through complex nested variants using the connectivity information from the SplitThreader graph. Combining long-read genome and transcriptome sequencing technologies enables an in-depth analysis of how changes in the genome affect the transcriptome.

PERVASIVE TRANSCRIPTION DECONVOLUTION REVEALS TRANSPOSABLE ELEMENTS ACTIVITY DURING THE DEVELOPMENT OF THE HUMAN BRAIN.

Fábio Navarro[1,2], Mark Gerstein[1,2]

[1]Yale University, Computational Biology & Bioinformatics Program, New Haven, CT, [2]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT

Transposable elements (TE) constitute approximately half of the human genome. Despite the difficulties created by their highly repetitive nature, there is building evidence that transposable sequences play a significant role in the genome biology influencing gene regulation and creating variability across individuals and species. TEs show somatic activity in pathogenic and healthy cells. Among other tissues, the human brain is thought to harbor somatic transposable element somatic.
Due to their high copy number and broad distribution across the genome, assessing transposable elements expression is specially affected by RNA fragments originating from pervasive transcription. We developed a new approach that deconvolutes pervasive transcription signal from real transcription and estimates TE subfamily expression level. We used RNA-seq experiments from BrainSpan and PsychENCODE to evaluate the activity pattern of LINE1s, SVAs and HERVs in 16 regions of the human brain across 12 developmental stages and different conditions. We observe that the majority of the reads mapped to transposable elements are due to pervasive transcription and removing of this signal results in recent, and potentially active, TE subfamilies being preferentially transcribed in the human brain. Moreover, we found that early stage brain development harbor higher activity of L1HS but lower activity of HERVH and no activity of SVAs. Conversely, adult brains harbor higher activity of SVA and endogenous retrovirus. All together our methods and results demonstrate that different and evolutionary recent TEs are active during the whole development of the human brain and might impact tissue biology.

SYSTEMATIC ANALYSIS OF LARGE HUMAN RNA-SEQ DATASETS

Abhinav Nellore[1,2,3], Andrew E Jaffe[1,3,4,5], Jean-Philippe Fortin[1,3], Leonardo Collado-Torres[1,3,4], José Alquicira-Hernández[1,6], Christopher Wilks[2,3], Siruo Wang[1,7], Robert A Phillips[1,8], Nishika Karbhari[1,9], Kasper D Hansen[1,3,10], Ben Langmead[1,2,3], Jeffrey T Leek[1,3]

[1]Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [3]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [4]Johns Hopkins Medical Campus, Lieber Institute for Brain Development, Baltimore, MD, [5]Johns Hopkins University, Department of Mental Health, Baltimore, MD, [6]National Autonomous University of Mexico, Undergraduate Program on Genomic Sciences, Mexico City, Mexico, [7]Centre College, Department of Mathematics and Computer Science, Danville, KY, [8]Salisbury University, Department of Biological Sciences, Salisbury, MD, [9]University of Texas at Austin, Department of Biological Sciences, Austin, TX, [10]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

We present analyses of large public human RNA sequencing (RNA-seq) datasets and new databases that allow investigators to rapidly query expression and splicing patterns across tens of thousands of samples. These analyses were enabled by Rail-RNA, a spliced alignment program that scales to analyze thousands of RNA-seq samples in the cloud with Amazon Elastic MapReduce. In one project, we aligned thousands of RNA-seq samples released by the GTEx consortium spanning many individuals and tissues and classified expressed regions of the genome to develop a unique gene expression barcode for each tissue type, in analogy to the gene expression barcode for microarray data by Irizarry and collaborators. Expressed regions were compiled into a queryable database available at http://rail.bio. In a second project, we aligned 21,504 publicly available Illumina-sequenced human RNA-seq samples from the Sequence Read Archive (SRA) to the human genome and compared detected exon-exon junctions with junctions in several recent gene annotations. 56,865 junctions (18.6%) found in at least 1,000 samples were not annotated, and their expression associated with tissue type. Newer samples contributed few novel well-supported junctions, with 96.1% of junctions detected in at least 20 reads across samples present in samples before 2013. We compiled junction data into a resource called intropolis also available at http://rail.bio. We discuss an application of this resource to cancer involving a recently validated isoform of the ALK gene.

← Back

IDENTIFYING THE ANCESTRAL ORIGIN OF RARE ALLELES

Dominic Nelson[1], Claudia Moreau[1,2], Damian Labuda[2], Simon Gravel[1]

[1]McGill University, Human Genetics, Montreal, Canada, [2]CHU Ste-Justine, Research Centre, Montreal, Canada

Large pedigrees contain a wealth of information about the genetic history of a population. For pedigrees containing thousands to millions of individuals, such as the province-wide pedigree that exists in Quebec, Canada, the joint analysis of genetic and pedigree data is numerically challenging. We present a new method for reconstructing the complete disease history of rare diseases, including identifying their ancestral origin within the pedigree. Combined with newly available genotype data for individuals connected to the large pedigree, we discuss the possibility of reconstructing near-complete genomes of several founders of the population of Quebec, as well as improved estimates of regional disease prevalence. When such detailed genealogical information is not available, we can still infer what effects the hidden pedigree may have on individual genomes. Using a new method for efficiently simulating admixed genomes, we discuss applications for improving models of demographic history and genetic association.

GENOMIC SIGNATURES OF HYBRID SPECIATION IN INVASIVE SCULPINS (COTTUS)

Fritz J Sedlazeck[1], Jie Cheng[2], Janine Altmüller[3,4], Arnd von Haeseler[5], Arne W Nolte[6,7]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Ocean University of China, Key Laboratory of Marine Genetics and Breeding , Qingdao, China, [3]University of Cologne, Cologne Center for Genomics, Cologne, Germany, [4]University of Cologne, Institute of Human Genetics, Cologne, Germany, [5]University of Vienna, : Bioinformatics and Computational Biology, Vienna, Austria, [6]University of Oldenburg, Ecological Genomics, Oldenburg, Germany, [7]Max-Planck-Institute for Evolutionary Biology, Evolutionary Genetics, Plön, Germany

Models of homoploid hybrid speciation predict that traits from both parental species are combined in an emerging hybrid species making it fitter than either parent. Hybridization between Cottus rhenanus and Cottus perifretum has produced an invasive hybrid species. Here, we determine which alleles from one or other parent species have reached high frequency in the hybrid population by natural selection, rather than by drift or migration. We first measured allele frequencies for 52,211 SNPs from one or other parent in invasive Cottus by high-throughput sequencing a pool of genomic DNA. We used a diffusion approximation to model the expected spectrum of allele frequencies across the genome under neutral drift with gene flow. 649 high frequency alleles (linked to 212 annotated genes) from C. rhenanus did not fit our neutral expectation, suggesting that they were targets of natural selection in the hybrid lineage. In contrast, high frequency alleles from C. perifretum could be explained by our neutral model, without invoking selection, presumably because the time since admixture is too long or because the overall contribution of this species to the admixed gene pool is too large. Only 12% of selected alleles are within previously identified hybrid-incompatibility QTL, suggesting either that most intrinsic incompatibilities have not yet been discovered, or that extrinsic selection is important in shaping the hybrid genome. The evolution of invasive Cottus involves a diverse genomic basis that presumably involves genotypes that originate from C. perifretum as well, but which could not be detected using our experimental method.

NOVEL SMALL RNAs IDENTIFIED IN DEVELOPING MAIZE SEEDS

Christos Noutsos[1], Oliver H Tam[1], Petsch Katherine[1], Timmermans C Marja[2]

[1]Cold Spring Harbor Lab, Hammel, Cold Spring Harbor, NY, [2]University of Tübingen, Center for Plant Molecular Biology - ZMBP, Tübingen, Germany

Maize is an important crop with a large and complex genome whose genome annotation remains incomplete. Small RNAs annotation is particularly poor, despite their large role in regulating various biological processes. In the current project, we are using maize embryo and endosperm tissue in several inbred populations (B73, W22 and Mo17) to survey the expression of known and novel microRNAs in early development. We performed high-throughput sequencing of wild type and dicer-like1 (dcl1) mutants to identify novel microRNAs originating from both unique and repetitive regions of the genome. These candidates share many characteristics with previously annotated maize microRNAs, and show significant depletion in the dcl1 mutant. The availability of dcl1 mutant tissue allows for both a high confidence set of novel microRNA candidates, and for an assessment of the characteristics of dcl1-dependent small RNAs that are currently missed by computational predictions alone. We have complemented our studies with publicly available small RNA datasets from various B73 tissues to explore the expression of these small RNAs throughout maize development. Further investigation of these candidates to identify their regulatory targets, and to assess their conservation among other plant species is ongoing. Comparisons across inbreds and related grasses will provide valuable insights into the molecular and functional diversity of microRNAs within maize populations, as well as their evolutionary divergence across grasses.

COMPARATIVE TRANSCRIPTOMICS OF IMMUNE CELL REPROGRAMMING IN HUMAN AND MOUSE SPECIES

Ramil Nurtdinov[1,2], Alexandre Esteban[1,2], Amaya Abad[1,2], Maria Sanz[1,2], Marina Ruiz[1,2], Dmitri Pervouchine[1,2], Sebastian Ullrich[1,2], Cecilia Klein[1,2], Alessandra Breschi[1,2], Silvia Perez[1,2], Rory Johnson[1,2], Roderic Guigo[1,2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Bioinformatics and Genomics, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Bioinformatics, Barcelona, Spain

Transcriptomes reflect cellular and organismic phenotypes. Therefore transcriptome comparisons across species uncover the molecular basis of both conserved and species-specific phenotypes. Since the laboratory mouse is top choice organism for human biology, transcriptome comparisons across multiple tissues between human and mouse have been extensively carried out. However, because phenotypic differences between human and mouse tissue are difficult to quantify, it is difficult to relate transcriptional differences to phenotyipic differences. Here we have compared human and mouse transcriptional patterns in a very well controlled time course cellular differentiation process for which the time of differentiation is quantifiable phenotype that can be precisely defined in human and mouse. Specifically, we employed the BlaER cell line, a powerful *in vitro* model of haematopoietic transdifferentiation. These engineered lymphoblastic leukaemia B-cells can be reproducibly and efficiently converted to a macrophage-like identity by the activation of a single CEBPα transgene. This process is accompanied by loss of the B-cell-specific surface marker CD19, and a concomitant gain of the Mac-1 marker, allowing cellular phenotype to be monitored and quantified by flow cytometry. The process takes seven days in human, but only three in mouse. We collected and sequenced cytoplasmic RNA at several consecutive time-points (12 in human and 10 in mouse). We aligned the human and mouse time points according to transcriptional similarity, and identified both the conserved and the species-specific transcriptional programs that participated in the process. Thus we identified 2689 orthologous protein coding gene with twofold expression change and concordant profiles during the transdiferentiation process, as well as 2149 variable ones that behaved differently in human and mouse. We also identified splicing events with conserved human-mouse patterns. Based on the species specific transcriptional pattern, we build models to predict transdifferentiation time depending on gene expression, and we are currently experimentally validating these models.

DEFINING THE microRNA MUTATIONAL LANDSCAPE IN 1000 GENOMES AND PEDIATRIC ACUTE LYMPHOCYTIC LEUKEMIA DATASETS

Ninad Oak, Sharon E Plon

Baylor College of Medicine, Molecular and Human Genetics, Houston, TX

Background: MicroRNAs (miRNAs) were first linked to cancers in studies of adult leukemia and the role of altered miRNA expression as cancer drivers has since been well established. Mutations in miRNAs have the potential to alter miRNA biogenesis, miRNA expression and impact hundreds of downstream target genes. However, among the large number of cancer sequencing studies, most focus on protein-coding genes by undertaking whole exome sequencing which captures less than 50% of known miRNAs. A very limited number of whole genome sequencing (WGS) studies have reported on sequence variation in the miRNAs (primarily in adult malignancies) and of those, only a few have performed secondary validation assays. **Methods:** In order to create a comprehensive catalog of miRNA variation, we developed a miRNA variant analysis and prioritization pipeline using genome analysis toolkit (GATK) and in-house analysis tools as well as functional assays for miRNA processing defects. **Results:** We first analyzed 1000Genomes (phase3) samples for variation across 1881 miRNAs (miRBase v21). We found an average of 8 rare germline miRNA variants per individual. As expected, mature miRNAs were highly conserved and harbored only 10% of all observed miRNA variation, in contrast to other regions of the precursor and primary miRNA. A significantly large number of miRNAs (~500) did not harbor any variants, including let-7e, mir-15/16 and other miRNAs that are known to alter key developmental and cancer pathways. To understand the role of miRNA variation in pediatric cancers, we analyzed WGS dataset of 62 tumor-paired normal samples of pediatric acute lymphocytic leukemia (ALL) from PCGP (St. Jude Children's Research Hospital). We observed higher number of rare germline miRNA variants per individual, which may partly result from the significantly higher sequencing depth. We found rare and novel germline variants in miRNAs frequently deregulated in pediatric ALL. Analysis of somatic variants in PCGP dataset identified 31 somatic variants, two of which were seen in the same precursor-miRNA. We are now extending the current mutation analysis pipeline to additional pediatric sequencing datasets like TARGET (NCI). **Conclusions:** Application of a comprehensive annotation pipeline for miRNA variation in 1000Genomes data has demonstrated a wide spectrum of variability in miRNAs and has identified miRNAs that are intolerant to variation, indicating their important regulatory roles in essential cellular pathways. Analysis of small pediatric ALL WGS datasets has identified putative oncogenic driver and cancer susceptibility miRNAs highlighting the potential role of miRNA variation in pediatric cancer patients.

← Back

INTERPRETING VARIANT PATHOGENICITY: LESSONS FROM OVER 60,000 HUMAN EXOMES

Anne H O'Donnell-Luria[1,2,3,4], Eric V Minikel[2,3,4], Monkol Lek[2,3,4], Konrad J Karczewski[2,3,4], Kaitlin E Samocha[2,3,4], Mark J Daly[2,3,4], Daniel G MacArthur[2,3,4], on behalf of the Exome Aggregation Consortium[5]

[1]Boston Children's Hospital, Genetics and Genomics, Boston, MA, [2]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [3]Broad Institute of Harvard and MIT, Medical and Population Genetics, Cambridge, MA, [4]Harvard Medical School, Boston, MA, [5]Exome Aggregation Consortium, Boston, MA

Deep reference panels of human genetic variation are critically required for clinical variant interpretation and disease gene discovery. Assembling such panels requires overcoming the challenges of data heterogeneity and scale. To explore the patterns of genetic variation across human protein-coding genes we jointly analyzed exome sequencing data from a collection of 60,706 individuals collected as part of the Exome Aggregation Consortium (ExAC), leveraging new haplotype-based approaches to joint variant-calling. These approaches substantially improve the accuracy and sensitivity of variant detection, especially for small insertions and deletions, and readily scale to tens of thousands of samples. We have publicly released frequency data for the more than 10 million variants discovered in ExAC, creating the largest available variant frequency dataset for clinical interpretation.

We provide an unprecedented view of the spectrum of human functional genetic variation extending down to extremely low population frequencies. Current clinical variant databases are polluted with numerous false positive inferences, as evidenced by the average ExAC participant harboring over 40 high-quality, disease-causing variants. We now have adequate power to assess the pathogenicity of many more variants than previously possible, including in understudied populations such as South Asians and Latinos. By reassessing the literature support for 192 previously reported pathogenic variants, we reclassified over 163 such variants from pathogenic to benign. These analyses confirm the importance of allele frequency as a powerful component of clinical interpretation of genetic variants.

We further sought phenotypic data for a subset of ExAC participants homozygous for a reported severe recessive disease variant, enabling reassessment of a gene previously associated with disease.

Our results underscore the tremendous contribution of large population resources such as ExAC in deciding whether variants are pathogenic or benign. The accuracy of these determinations has profound implications for patients, as the genetic diagnosis often leads to specific diagnostic and/or treatment recommendations. The power of ExAC will continue to grow as sample sizes increase into the hundreds of thousands of individuals.

IDENTIFICATION OF SEX-BIASED EXPRESSION AND EXPRESSION QUANTITATIVE TRAIT LOCI (eQTLs) IN INNATE AND ADAPTIVE IMMUNITY

Meritxell Oliva[1,2], Charles Czysz[1,2], Barbara E Stranger[1,2]

[1]The University of Chicago, Institute for Genomics and Systems Biology, Chicago, IL, [2]The University of Chicago, Department of Medicine, Section of Genetic Medicine, Chicago, IL

Human males and females exhibit sexual dimorphism in a wide range of phenotypes. Several diseases display sex biases in prevalence, clinical features and prognosis, e.g. autoimmune diseases are more common in females and infectious diseases in males. Moreover, disease-related phenotypes such as drug responsiveness or survival rates behave in a sex-dependent manner. Herein, we characterize the influence of sex in determining inter-individual transcriptional variation in cells representing the adaptive and innate branches of the immune system in cohorts of different ancestry. We aim to: a) identify genes and regulatory networks differentially expressed (DE) between sexes, b) identify and characterize sexually dimorphic expression quantitative trait loci (eQTLs), c) pinpoint differences in sex-biased eQTL profiles between autosomal and sex chromosomes and d) determine the variation of sex biases across immune cell types and their contribution to disease.

We have analyzed genotype and gene expression data from naïve T lymphocytes and monocytes in healthy Caucasian-American, Asian-American and African-American populations (Immune Variation project, Raj et al. 2014), to discover genes and pathways regulated differently between healthy males and females. We identified 6% and 7% genes that exhibit significant (FDR<5%) sex-biased expression in naïve T-cells and monocytes, respectively. Sex-biased gene sets are enriched in X-linked genes (F-test $p<2\times10^{-16}$) and also in pathways related to apoptosis, immunity, infectious and autoimmune diseases, phenotypes for which sex bias has been previously identified but rarely characterized at a molecular level. Several immune pathways mediated by hormones (e.g. prolactin) exhibit sex bias, highlighting how sex-specific hormone profiles explain differences between male and female immunity. Interestingly, we observe distinct cell-type specific profiles: despite having nearly equal fractions of sex-biased genes, more than 80% of the sex-biased genes are specific to either monocytes or T-cells.

These results suggest strong context specificity in sex-dependent transcriptional regulation. To identify genetic variation that regulates gene expression differently between sexes, for each cell type and population we have modeled transcript abundance as a function of sex, genotype and the interaction of both, testing for significance of the latter as sex-biased eQTLs. To increase the statistical power of the association analysis, we meta-analyzed populations for each cell type. We are characterizing: a) whether genes that are DE between sexes demonstrate sex biases in the genetic basis of their regulation and b) differences in sex-biased regulation profiles between sex and autosomal chromosomes. Finally, we are exploring the contribution of sex-biased eQTLs to disease through integration with GWAS results.

DIFFERENTIALLY EXPRESSED miRNAs IN LIVER TISSUE RELATED TO FEED EFFICIENCY IN NELORE CATTLE.

Priscila S N De Oliveira[1], Polyana C Tizioto[1], Gabriela B De Oliveira[2], Aline S M César[2], Mirele D Poletti[2], Wellison J S Diniz[3], Andressa O De Lima[3], James M Reecy[4], Luis L Coutinho[2], Luciana C A Regitano[1]

[1]Embrapa Southeast-Cattle Research Center, Animal Biotechnology Laboratory, São Carlos, Brazil, [2]University of São Paulo, Animal Science, Piracicaba, Brazil, [3]Federal University of São Carlos, Department of Genetics and Evolution, São Carlos, Brazil, [4]Iowa State University, Department of Animal Science, Ames, IA

Residual feed intake (RFI) is an important and well-known measure of feed efficiency in beef cattle, with a major impact on production costs. Among the regulatory mechanisms, miRNAs have emerged as a new dimension in post-transcriptional regulation in mammals and have been associated with a number of important biological processes affecting production traits. In this study, Nelore steers genetically divergent for RFI (kg/d) were selected based on BLUP (Best Linear Unbiased Prediction) estimates and ranked in order to select the most extreme values for additive genetic merit. Sequencing of small RNA libraries from liver tissue of eight Nelore steers (n = 4 High RFI, n= 4 Low RFI) were held in MiSeq equipment using the Miseq Reagent Kit V3 150 cycles. After the sequencing quality control, the miRDeep2 software was performed to identify and quantify novel and known miRNAs using Bos taurus UMD3.1 as reference genome. Differentially expressed (DE) miRNAs were identified by DESeq2 R package and potential regulatory target transcripts were predicted by TargetScan software. Four DE miRNAs were identified; bta-miR423-5p (padj=0.0002), bta-miR30b-5p (padj=0.0143), bta-miR339 (padj=0.0143) and bta-miR378 (padj=0.0171). The genes *ATP2A2*, *SLC41A*, *FADS2*, *COL1A1* and *RNF150* were identified as target genes of these DE miRNAs, which were previously identified as DE in RNAseq study using this same set of samples. These genes are related to some important biological process for feed efficiency; like ion transport and metal ion binding, carbohydrate and/or fatty acid metabolism and glycosylation. Further studies will be carried out in order to investigate the role of miRNAs in biological mechanisms involved in feed efficiency in Nelore beef cattle.

IDENTIFYING THE TISSUE OF ACTION FOR GWAS VARIANTS AND ASSESSING TISSUE SPECIFICITY OF eQTLs IN GTEx

Halit Ongen, Andrew A Brown, Olivier Delaneau, Alexandra C Nica, GTEx Consortium, Emmanouil T Dermitzakis

University of Geneva, Department of Medical Genetics and Development, Geneva, Switzerland

Genome-wide association studies (GWAS) discover variants that correlate with a disease or a trait, the vast majority of which lie in the non-coding genome, rendering their biological interpretation difficult. Expression quantitative trait locus (eQTL) analyses find regulatory variants for genes and aid in the interpretation of GWAS variants. However, given the abundance of QTLs and GWAS variants, linkage disequilibrium, and partial knowledge of all the variants, interpreting if the same functional variant is responsible for colocalized eQTL or GWAS signals remains a challenge. We have previously described the Regulatory-Trait Concordance (RTC) score tackling this problem. We have improved this method so that it scales up to thousands of phenotypes and genotypes, and also for multiple independent signals per phenotype. We applied RTC to the GTEx dataset, which comprises RNA-seq from 44 tissues and genotypes, from 70-361 samples. We sought to answer two questions. Firstly, given the GWAS variants for certain traits and diseases, and eQTLs from 44 tissues if we can identify the relevant tissues. Secondly whether we can assess tissue specificity of eQTLs on a per variant basis. To this end we conducted *cis*-eQTL analysis identifying multiple independent eQTLs per gene and find 897-12918 eQTLs in 44 tissues. Our simulations show that the RTC score is the probability of two independently discovered variants to tag the same functional variant. By calculating the proportion of GWAS hits that have an RTC score (probability) ≥ 0.9 with an eQTL in each of the tissues, while accounting for differential power of discovery, we identify relevant tissues for certain traits. For Type II Diabetes, the main tissues are adipose tissues, artery tissues, nerve, and thyroid, whereas for a ubiquitous trait like height the tissue contributions are distributed uniformly. We tested tissue specificity of eQTLs using the RTC score, and show that RTC scores across all sites correlate well with the more commonly used π_1 analysis, highlighting the accuracy of RTC. Using RTC we assess the tissue specificity on a site by site basis. Preliminary analysis with 11 tissues identifies two sets of tissues: tissues like esophagus where majority of the eQTLs are active in all tested tissues, and tissues like blood where eQTLs have binary properties and are either very tissue specific or active in all tissues. Our results show that finding GWAS and eQTL variants tagging the same functional variant in different tissues aids in identifying the relevant tissues in disease pathophysiology, and we can ascertain tissue specificity of eQTLs in a granular way. We will further present analysis where we integrate tissue specificity with GWAS interpretation to uncover tissue genetic causality and shared mechanisms of GWAS variant action.

PROPERTIES OF FALSE-NEGATIVE VARIANT CALLS IN HUMAN EXOME SEQUENCING DATA

Jason A O'Rawe[1,2], Gholson J Lyon[1,2]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY, [2]Stony Brook University, Genetics Program, Stony Brook, NY

Accuracy is paramount if high-throughput assays like exome sequencing are to be in widespread clinical use. Substantial progress has been made toward generating highly accurate variant detection in exome sequencing, but even the most developed detection methods generate a unique subset of variants. The community has devoted substantial resources toward understanding and mitigating false positive detections, but less focus has been given to false negative detections; those variants that we miss. In the clinical realm, missing even a single variant can in some cases mean the difference between effective or ineffective care. In this presentation, we describe methods for extracting false negative calls from existing call sets using large validation data sets. We describe the important features of false negative calls, as well methods by which machine learning models were developed to differentiate true negative from false negative detections using these data sets. In general, over 200 false negative SNV calls and over 1000 false negative INDEL calls were extracted from three large MiSeq validation data sets generated from a single sample; sample K8101. Several machine learning methods were used to build predictive models of false negatives, including Support Vector Machines, Random Forests, Gradient Tree Boosting, Logistic Regression and others. Ensemble approaches were used for aggregating model predictions, to varying degrees of success. Model performance is promising, with mean AUROC measures for each model surpassing 0.9. Estimates of feature importance reveal several important variables, which give insight as to the origin of these errors.

GENETICS OF GENE EXPRESSION REGULATION IN A CASE-CONTROL STUDY FOR ACUTE MYOCARDIAL INFARCTION IN A PAKISTANI POPULATION.

Nikolaos I Panousis[1], Salih Tuna[2], Lazaros Lataniotis[2], Asif Rasheed[3], Nabi Shah[3], John Danesh[4], Emmanouil T Dermitzakis[1], Danish Saleheen[3,5], Panos Deloukas[2]

[1]University of Geneva, Dept of Genetic Medicine and Development, Geneva, Switzerland, [2]Queen Mary University, William Harvey Research Institute, London, United Kingdom, [3]Center for Non-Communicable Diseases, Karachi, Pakistan, [4]University of Cambridge, Dept of Public Health and Primary Care, Cambridge, United Kingdom, [5]University of Pennsylvania, Dept of Biostatistics and Epidemiology, Philadelphia, PA

The prevalence of coronary heart disease (CHD) has been reported to be unusually high in populations of South Asia; the risk of Myocardial infarction (MI) is 2-3 fold higher. By studying patterns of gene expression and genetic variation in monocytes, which is a relevant cell type we can discover biomarkers and develop a functional understanding of how genetic variation can alter risk of CHD. We report an RNA-seq analysis of monocytes from 71 cases of confirmed acute MI and 77 healthy individuals from the Pakistan Risk Of Myocardial infarction study (PROMIS).
We identified 5244 differentially expressed genes (5% FDR), including many that had previously been associated with CHD. PCA of expression differentiated MI from healthy individuals, demonstrating whole genome differences in expression. More specifically, genes that significantly contributing to the variance explained by the PC that differentiate the samples, were enriched in pathways of signaling of inflammatory responses suggesting a method to detect and/or validate novel biomarkers. As monocytes are central to the development of CHD, it is hoped that specific variants affecting gene expression could have downstream consequences on this and other related diseases. We explored genetic regulation of monocyte expression and we identified 4799 eQTLs (5% FDR). These variants were enriched for GWAS variants for cardiovascular disease (OR 1.8). Analyzing cases and controls separately, we identified hundreds of MI specific eQTLs, which were not associated with expression in controls. Enrichment analysis of MI-specific eQTLs and MI GWAS loci, using the Regulatory Trait Concordance (RTC) method revealed several specific loci where the eQTL and GWAS signal were tagging the same functional variant. This included eQTLs tagging GWAS hits for obesity, lipid levels and Type2 Diabetes. By combining these findings with our list of phenotypic risk factors (lipid profile, BMI), will allow to further investigate the architecture of MI genetic risk. In summary, we will show how analysis of gene expression in monocytes is informative about MI and genetic variants active in the cell can be informative on genetic causes of disease.

← Back

EFFECT OF BRAF AND RAS MUTATIONS ON ALTERNATIVE POLYADENYLATION IN PAPILLARY THYROID CARCINOMA

Ji Yeon Park[1], Jin Wook Yi[2,3], Byoung-Ae Kim[2], Brian Y Ryu[1], Bin Tian[4], Kyu Eun Lee[2,3], Ju Han Kim[1]

[1]Seoul National University Biomedical Informatics, Division of Biomedical Informatics, Seoul, South Korea, [2]Seoul National University Hospital and College of Medicine, Department of Surgery, Seoul, South Korea, [3]Seoul National University College of Medicine, Cancer Research Institute, Seoul, South Korea, [4]Rutgers New Jersey Medical School, Department of Microbiology, Biochemistry and Molecular Genetics, Newark, NJ

Alternative polyadenylation (APA) is receiving more attention in cancer progression, but its role is not clear in papillary thyroid carcinoma (PTC). Using DaPars program (Xia et al., 2014), we identified 3'UTR length changes in PTC samples harboring BRAF and RAS mutations, which derived from The Cancer Genome Atlas (TCGA). Compared to previous cancer-related APA studies, relatively small number of genes are found to be regulated in PTC samples over the matched normal ones, suggesting that the extent of APA regulation is relatively mild in differentiated cancers such as PTC. In addition, APA changes are not strongly biased toward 3'UTR shortening, but there is a considerable overlap in genes with 3'UTR shortening between PTC and other cancer types. The changes are more likely to be relevant to cancer pathogenesis. We also uncovered that BRAF-specific APA events, which can explain, at least partially, higher invasive phenotype of BRAF mutation. For example, BRAF mutation favors increased expression of shortened isoform in a gene encoding CASP8 and FADD like apoptosis regulator (CFLAR), but RAS mutation fails to do that. Its significance is supported by a previous report where the gene shows 3'UTR shortening in 5 out of 7 cancer types in TCGA. These results can be used to define the molecular characteristics in posttranscriptional regulation of PTC by mutation type.

Reference

Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J., and Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nat Commun 5, 5274.

COMPUTATIONAL DISCOVERY OF EPIGENETIC MEDIATORS IN ALZHEIMER'S DISEASE FROM IMPUTED METHYOME-WIDE ASSOCIATION STATISTICS

Yongjin Park[1,2], Abhishek Sarkar[1,2], Nick Mancuso[3], Alexander Gusev[2,4], Bogdan Pasaniuc[3], Manolis Kellis[1,2]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Broad Institute, Medical Population Genetics, Cambridge, MA, [3]University of California, Pathology and Laboratory Medicine and Human Genetics, Los Angeles, CA, [4]Harvard University, Epidemiology, Cambridge, MA

It is increasingly recognized that common genetic variants primarily affect regulatory regions, controlling gene expression patterns, rather than coding sequence. However, population-level epigenetic variation is not systematically surveyed in large cohorts, limiting the ability to dissect regulatory mechanisms in common diseases. To address this challenge, mapping of genetic variations across individuals in reference cohorts permits inference of regulatory program between non-coding variants and intermediate cellular phenotypes in cell type of action.

We leveraged a cohort of 700 individuals of whom genetic variations are measured with DNA methylations of 450k CpGs, mapped in brain tissues in the ROS/MAP. We utilized this information to learn robust mapping between genetic and methylation variation across individuals, which we then applied to the cohort of 74046 individuals for which only genetic information and Alzheimer's disease status are provided. We use this mapping to impute disease-methylation association statistics to recognize CpG dinucleotides are causally related with Alzheimer's disease.

We introduce new methodology, built on imputed transcriptome-wide association study (Gusev et al.), with a novel quantitative trait loci (QTL) model and efficient algorithm. We modeled proportion of DNA methylations in each CpG by non-linear Beta regression model, while not overfitting to reference panel by Bayesian variable selection. However the underlying probabilistic model does not permit closed form solutions as in linear Gaussian model, neither can be inferred by Monte Carlo simulation because number of cis-SNPs can be as large as 5000's in imputed genotypes. We implemented posterior inference algorithm by stochastic gradient approximation with reparameterization.

Applying our methodology we discovered 1731 CpGs are highly associated with Alzheimer's disease at FDR < 5%, 36 at FWER < 5%, of which only 1 locus has been identified by genome-wide association studies. On average methylations are genetically controlled by 3 independent SNPs. Our results suggest that amplification of polygenic effects through the lens of epigenetic regulation yields significantly better power and provide clear explanation of regulatory mechanisms, but emphasize importance of accurate QTL models with scalable algorithm.

← Back

VCFANNO: FAST, FLEXIBLE ANNOTATION OF GENOMIC VARIANTS

Brent S Pedersen[1], Ryan M Layer[1], Aaron R Quinlan[1,2]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, Biomedical Informatics, Salt Lake City, UT

It is now possible to produce high quality variant calls from high throughput sequencing data, but, without annotation, genetic variants of potential interest are indistinguishable from noise or likely targets of interest. Annotation with reference or in-house datasets that contain, for example, allele frequencies observed in a large background population or known pathogenic variants, allows interpretation, classification, and prioritization. However, comprehensive annotation with many reference datasets is cumbersome with existing software. We have developed a fast and flexible tool, **vcfanno**, that simplifies the annotation of genetic variants in VCF format with one or more data sets. The implementation includes a new parallel chromosome "sweeping" algorithm. The **vcfanno** command-line utility accepts a user-defined config file that defines the input files and operations such as mean, concat and max; it outputs a VCF with the modified header and INFO fields containing the information collected from the requested fields in the annotation files. Our implementation is able to annotate 20,000 variants per second with over 50 annotations from 17 annotation files using 12 processors. The performance of a single core is several thousand annotations per second and parallelization scales well up to about a dozen processors.

The annotation process is very customizable; for each annotation field, the name of the resulting field along with a means to summarize values when there are multiple overlaps. Common summary operations such as max, min, concat are built-in; for custom operations, a lua scripting engine is embedded so that it is possible to perform arbitrary manipulations that use annotation fields and/or fields from the query VCF. This allows the use of **vcfanno** to replace small scripts that modify parts of the INFO field with the added benefit of automatic parallelization of the scripting. Together, these features make **vcfanno** a fast, comprehensive tool for variant annotation in any species.

GENOMIC AND FUNCTIONAL BASIS OF ADAPTIVE CHANGE: THE SELECTIVE HISTORY OF CAMOUFLAGED DEER MOUSE POPULATIONS

Susanne P Pfeifer[1,2], Stefan Laurent[1,2], Ricardo Mallarino[3], Matthieu Foll[1,2], Catherine R Linnen [4], Jeffrey D Jensen[1,2], Rowan D Barrett[5], Hopi E Hoekstra[3]

[1]Ecole Polytechnique Fédérale de Lausanne (EPFL), School of Life Sciences, Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, SIB, Lausanne, Switzerland, [3]Harvard University, Department of Organismic and Evolution Biology, Cambridge, MA, [4]University of Kentucky, Department of Biology, Lexington, KY, [5]McGill University, Department of Biology, Montréal, Canada

Global warming (and climate change more generally), along with anthropogenic interference in wild habitats, forces many natural populations to adapt to novel ecological conditions. However, both the causes as well as the magnitude of the genomic changes underlying these adaptations often remain poorly understood. Despite the fact that recent progress in genomics has enhanced our understanding of adaptive trait evolution at the molecular level over the past years, relatively few studies have so far been published that link beneficial mutations (and thus the molecular mechanism) to specific selective pressures in natural populations, thereby providing a complete picture of the source of evolutionary change. In addition, the genomic basis of adaptive traits and the ecological causes of selection are often very complex, generally limiting the predictability of local adaptation.

In this study, we are investigating the genomic and functional basis of adaptive coat coloration of deer mouse populations to the ecologically distinct, recently formed Sand Hills in Nebraska. Previous research has shown that this coat color change, caused by allelic variation at the *Agouti* pigmentation locus, is an adaptation for crypsis (helping mice to avoid predation from visually hunting avian predators, which preferentially predate on poorly background-matched prey). We take advantage of this established relationship between the camouflaging coat-color and the underlying beneficial mutations, and combine ecological sampling and population sequencing with novel statistical and computational methods in order to address questions surrounding the mode and tempo of adaptive evolution as well as the predictability of phenotypic evolution.

Beginning with a transgenic study confirming that a previously identified deletion within the *Agouti* gene results in the lighter coat color of deer mice inhabiting this region, we then performed a large-scale field-based selection experiment that involved the introduction of 480 wild deer mice (with known genotypes and phenotypes) into replicated field enclosures, representing the two extremes in substrate color found in their natural habitat, enabling us to characterize selection on both the genotype and phenotype.

dbSNP IN THE ERA OF NEXT-GENERATION SEQUENCING

Lon Phan, Hua Zhang, Juliana Feltz, Wang Qiang, Eugene Shekhtman, Rama Maiti, David Shao, Ming Ward

National Center for Biotechnology Information, NLM/NIH, Bethesda, MD

dbSNP, despite its name, is a database for all short (<=50bp) genetic variations that include SNV, small indels, microsatellites, and non-polymorphic germline and somatic variants. The primary roles of the database are to process and archive submissions, assign stable accessions, aggregate information from multiple submitters, map and annotate variants on the latest genome assembly and RefSeq sequences (genomic, mRNA, and protein), and distribute them for general use. dbSNP data are used in diverse fields including personal genomics, medical genetics, and variant analysis in many different organisms. The data are also integrated with other NCBI resources including ClinVar, Gene, PubMed, Nucleotide, Protein, Structure, BioSample, and BioProject. dbSNP data are made available in many ways: Entrez searches, annotated on the genome assemblies and RefSeqs, in Sequence Viewers, and via ftp downloads as BED, VCF, and other formats. In addition many bioinformatics centers including EBI and UCSC host dbSNP data and the data is incorporated into popular open-source and commercial bioinformatics tools and pipelines.

dbSNP house over 1.7 Billion Submitted SNP (ss) records that cluster to 800 Million non-redundant Reference SNP (rs) from over 360 organisms. Human is the largest organism by data volume, with 540 Million ss and 154 Million rs including large datasets from WES and WGS projects such as 1000Genomes, GO-ESP, and ExAC. These and other large scale NGS submissions have enriched dbSNP with over 50 million rare human variants (MAF < = 0.001) as well as WES (3X over WGS) coding and splicing variants. Current NGS trends suggest thousands to millions of new samples will be sequenced in the next few years that will require new tools to annotate, prioritize, and interpret variants. To facilitate analysis and interpretation of variants from these genomes and promote development of new tools, dbSNP aims to enrich rs annotation to include information from VAAST variant priority score (VVP) (Flygare et al.), protein features such as post-translational modification sites, Conserved Domain Database (CDD), protein 2D and 3D structures, binding sites for interaction with protein, drug-targets, and small molecules, and BioSystems. This is in addition to the functional consequence, allele frequency, ClinVar clinical significance, and many other rs attributes already reported by dbSNP. This presentation will discuss the incorporation of the new annotations and their use in the analysis of dbSNP and ClinVar contents and interpreting variant biological impact.

Acknowledgments
Work at NCBI is supported by the NIH Intramural Research Program and the National Library of Medicine.

CANU: A PACBIO AND NANOPORE ASSEMBLER FOR GENOMES LARGE AND SMALL

Sergey Koren[1], Brian P Walenz[1], Konstantin Berlin[2], Adam M Phillippy[1]

[1]National Human Genome Research Institute, Computational and Statistical Genomics Branch, Bethesda, MD, [2]Invincea, Invincea Labs, Arlington, VA

Emerging single-molecule sequencing technologies have now enabled the continuous reconstruction of some human chromosome arms. However, the long read lengths and higher error rates of these technologies demand new computational methods for alignment and assembly. To address this, we previously introduced the MinHash Alignment Process (MHAP) for assembling noisy, single-molecule reads. Using this method, our *de novo* assemblies of several model organisms included completely assembled chromosomes, revealed novel heterochromatic sequences, and resolved gaps in the established GRCh38 human and *D. melanogaster* reference genomes. We have since built on this work, creating a new assembler named Canu that supports both PacBio and Oxford Nanopore technologies, lowers coverage requirements, improves runtime, and enhances the reconstruction of repetitive and diploid genomes. Using Canu we have assembled several mammalian, bird, plant, insect, and fish genomes from PacBio data. Some of these assemblies were also combined with Illumina reads, BioNano optical maps, and Hi-C or Chicago proximity ligation data to generate chromosome-scale scaffolds with high consensus and structural accuracy. For Oxford Nanopore, we have completely assembled multiple bacterial species, and co-assembled larger genomes using a combination of PacBio, Oxford Nanopore, and Illumina data. Instrument throughput is the only barrier currently preventing the assembly of large genomes using nanopore data alone. Future Canu work includes improved reconstruction of heterochromatic sequence, better separation of segmental duplications, and support for population and metagenomic assembly. Canu is available under a GPL license from https://github.com/marbl/canu

CHROMATIN STATE VARIABILITY: A GUIDE TO UNCOVER FUNCTIONAL GENOMIC REGIONS and interactions.

Luca Pinello[1], Alexander Gusev[2], Hilary Finucane[2], Jialiang Huang[1], Alkes Price[2], Guo-Cheng Yuan[1]

[1]Dana-Farber Cancer Institute, Biostatistics and Computational Biology, Boston, MA, [2]Harvard T.H. Chan School of Public Health, Department of Epidemiology , Boston, MA

Background: With the increasing amount of epigenomic data, a pressing challenge is to understand the mechanisms underlying chromatin states changes and their role in cell-type specific establishment and maintenance.

Methods: Here we propose a computational method based on information theoretic approaches to systematically quantify the variability of a chromatin structure and study how this variability is linked to gene expression variation.

Results: Using histone modification data from 9 human cell lines from the ENCODE project we identify highly plastic regions (HPRs) for chromatin state variation, and found these HPRs are enriched for many important regulatory elements such as super-enhancers, promoters and polycomb repressed regions. Moreover we find that the HPRs are highly enriched for GWAS-associated non-coding variants and, for some traits or diseases, provide significant additional power in explaining heritability than well-characterized elements such as enhancers and promoters. We also identify regions of co-variability based purely on histone modification data, and show that such regions correlate well with data from long range interactions obtained by HI-C and ChIA-PET assays. In addition, the co-variability measure shows a clear depletion in correspondence of the boundaries of topological associated domains (TAD), suggesting that this measure could be helpful to highlight or predict functional subdomains within the TAD.

Significance: Our analysis provides new insights into the organization and dynamic change of cell-type specific chromatin structure during development and a valuable tool for investigating the mechanisms of chromatin state establishment and usage. The HPRs obtained from our analysis provide an important guide to search for functionally important genetic variants.

UNDERSTANDING HOW ALTERNATIVE SPLICING RELATES TO PRIMATE GENOME EVOLUTION: A CROSS-PRIMATE ANALYSIS OF CHANGES IN ISOFORMS AND THEIR ABUNDANCE

Lenore Pipes[1,2], Philip Blood[4], Dylan R McNally[2], Adam Siepel[3], Christopher E Mason[2]

[1]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY, [2]Weill Cornell Medical College, Institute of Computational Biology, New York, NY, [3]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [4]Pittsburgh Supercomputing Center, Pittsburgh, PA

In order to comparatively identify and quantify taxonomically-restricted splicing isoforms and describe the selective pressures driving alternative splicing in primates encompassing >70 million years of evolution, we created *de novo* transcriptome assemblies using RNA-Seq data from the Non-Human Primate Reference Transcriptome Resource (NHPRTR). The NHPRTR data was sequenced with unprecedented depth and a variety of tissues therefore some assemblies were created from a remarkable >3 billion reads. Using the assemblies, we created splice junction and exon databases of one-to-one orthologous regions between humans and non-human primates. We used tissue-matched RNA-Seq data generated from the Genome-Tissue Expression (GTEx) Project for quantifying these regions in humans. Using this approach, we were able to identify and validate many species-specific isoforms to attempt to gain insight into the molecular basis for species-specific phenotypes across similar tissues. We quantified and validated these isoforms using RT-PCR in multiple individuals from the same species. We have also found evidence of negative selection in "frame switching" alternatively spliced exons as well as an excess of changes from minor-form exons to major-form exons in protein-coding genes, and decreased dN/dS rates 3' downstream of "frame switching" events. This not only represents one of the first attempts to understand splicing complexity in over a dozen primates but is also a systematic approach for gaining insights into the evolution of splicing across species and how their genotypes may relate to their phenotypes.

GENOME-WIDE HAPLOTYPING USING SINGLE-CELL SEQUENCING

David Porubsky[1], Ashley D Sanders[2], Niek v Wietmarschen[1], Ester Falconer[2], Mark Hills[2], Marianna R Bevova[1], Victor Guryev[1], Peter M Lansdorp[1,2,3]

[1]University Medical Center Groningen, European Research Institute for the Biology of Ageing, Groningen, Netherlands, [2]BC Cancer Agency, Terry Fox Laboratory, Vancouver, Canada, [3]University of British Columbia, Department of Medicine, Vancouver, Canada

Haplotype resolved genomes are important in many areas of human genetics ranging from variant-disease associations, mapping regions of loss of heterozygosity (LOH) to studying inheritance patterns in human populations. To assemble haplotypes, computational and experimental approaches have been developed. Currently the most efficient way to obtain a set of genetic variants of an individual is massively-parallel sequencing. Unfortunately, current sequencing technologies are not able to assemble haplotypes across the whole length of the chromosome. Chromosome capture techniques have been proposed to overcome this problem, however such techniques are labor intensive and have not been widely adopted in practice. Consequently, there are currently no sequencing based methods able to phase genetic variants across the whole length of the chromosome without parental information.

In this study, we introduce Strand-seq as a new experimental approach for whole genome haplotyping. Strand-seq is a single cell sequencing technique with a unique ability to retain directionality of sequencing reads, based on the DNA template strand inheritance. This allows us to map every read to a single parental chromosome in a single cell. Combining the sequencing data from several single cells allows us to reconstruct an individual's genome while maintaining phase and haplotype information.

We have used Strand-seq to build chromosome-length haplotypes for all members of a well-studied HapMap family trio. Comparison of our results with HapMap reference showed high concordance of 99.3% demonstrating the accuracy and validity of our approach. Importantly, such accuracy was achieved without parental information or statistical approaches. Furthermore, since chromosome spanning haplotypes are assembled we can harness parental haplotypes in order to map meiotic recombination evets in the child. In this way we have mapped 38 switches (including 2 on chromosome X) in the maternal and 26 on the paternal homologues of the child, which is consistent with rates of meiotic recombination observed in previous studies. Last but not least single cell resolution of this approach allowed us to observe regional LOH in small number of cells.

We propose Strand-seq as a tool able to rapidly phase individual genomes and map inheritance patterns in families along with detection of rare cell populations. Moreover chromosome spanning Strand-seq haplotypes can complement emerging long-read sequencing technologies in order to facilitate de novo assembly of haplotype aware personal genomes.

A GENETIC SIGNATURE OF FLIGHTLESSNESS EVOLUTION IN THE GALAPAGOS CORMORANT (*PHALACROCORAX HARRISI*) REVEALED BY PREDICTIVE GENOMICS.

Alejandro Burga[1,2], Weiguang Wang[3], Paul Wolf[4], Andy Ramey[5], Claudio Verdugo[6], Karen Lyons[3], Patricia Parker[7], Leonid Kruglyak[1,2]

[1]UCLA, Departments of Human Genetics and Biological Chemistry, Los Angeles , CA, [2]HHMI, Howard Hughes Medical Institute , Los Angeles , CA, [3]UCLA, Department of Molecular, Cell and Developmental Biology and the Orthopedic Hospital Department of Orthopedic Surgery, Los Angeles, CA, [4]USDA, APHIS, Saint Paul, MN, [5]USGS, Alaska Science Center, Anchorage, AK, [6]Universidad Austral, Facultad de Ciencias Veterinarias, Valdivia, Chile, [7]University of Missouri-St. Loius, Department of Biology, St. Louis, MO

The survival of species depends on their ability to explore new niches, utilize the available resources and compete for them. The invention and modification of many novel characters, especially new types of appendages, was and continues to be essential for the adaptive evolution of vertebrates and arthropods to diverse aquatic, terrestrial and aerial environments. Most of our knowledge about development has been derived from the study of loss of function mutations with strong and often severe associated phenotypes. Yet, the invention and modification of complex traits, such as limbs, is thought to occur mainly through the accumulation of small changes over thousands or millions of years and that is precisely what makes their study particularly troublesome. Here we present a novel methodology combining predictive and comparative genomics to identify the variants responsible for limb evolution in the wild. The Galapagos cormorant (*Phalacrocorax harrisi*) is unique in that is the only cormorant that lost the ability to fly among approximately 40 extant species. Their entire population is distributed along the coastline of the Isabela and Fernandina Islands in the Galapagos archipelago. P. harrisi has a pair of stubby wings, which are smaller than those of any other cormorant in spite of having the largest body mass, resulting in a significant deviation from the allometric relationship between wing length and body mass among flighted Phalacrocoracidae. In order to understand the genetic and molecular mechanism responsible for wing size reduction in P. harrisi, we sequenced and de novo assembled the genomes of *P. harrisi*, its two closest relatives (*P. auritus* and *P. brasilianus*) and the more distantly related *P. pelagicus*. We predicted the impact on protein function of each of the Galapagos Cormorant's coding variants on a genome-wide scale, identified candidates and provide experimental evidence linking the flightless phenotype to human skeletal ciliopathies through a novel transcriptional regulatory axis.

SHARED GENETICS OF OBSESSIVE COMPULSIVE DISORDER IN DOGS AND HUMANS

Elinor K Karlsson[1,2], Hyun Ji Noh[2], Guoping Feng[3,4], Kerstin Lindblad-Toh[2,5]

[1]U Mass Medical School, Bioinformatics and Integrative Biology, Worcester, MA, [2]Broad Inst, Cambridge, MA, [3]Broad Inst, Stanley Center for Psychiatric Research, Cambridge, MA, [4]McGovern Inst, Cambridge, MA, [5]Uppsala Univ, Science for Life Lab, IMBIM, Uppsala, Sweden

Canine OCD is a naturally occurring, genetically complex model for human obsessive compulsive disorder. In both dogs and humans, OCD manifests as time-consuming repetition of behaviors causing functional impairment to sufferers, with similar age of onset and response to treatment with SSRIs. In dogs, the limited diversity in breeds and selection on behavioral traits makes genetic mapping particularly powerful.

In the first GWAS of canine OCD, we compared 92 Doberman pinschers with OCD to 68 breed matched controls and identified a significant association at the neural cadherin CDH2, explaining 5-10% of the phenotype variance. We sequenced the top 14 GWAS regions, including CDH2, in 16 dogs from 4 breeds and identified four genes, all with synaptic function, strongly enriched for variants seen only in cases. Three genes (CDH2, CTNNA2 and ATXN1) act in glutamatergic signaling pathways moderately associated in a recent human OCD GWAS, while the 4th, PGCP, is involved in glutamate homeostasis. In addition, we identified two candidate causal variants in the 2 Mb gene desert between CDH2 and DSC3 that disrupt the same small, highly conserved regulatory element. We found both variants cause significant changes in gene expression in a human neuroblastoma cell line, likely due to disrupted transcription factor binding.

We have now sequenced exons and regulatory elements for 608 genes in pathways implicated in OCD in humans, dogs and mice in humans (592 cases and 560 controls) and identified evolutionarily conserved, likely functional variants. Overall, we found a burden of variants in synaptic adhesion and maintenance pathways. NRXN1, which encodes a synapse adhesion molecule, had an excess of coding variants in cases, including 7 missense variants in domains critical for adhesion in the post-synapse. Two genes involved in synapse maintenance and vesicle trafficking, CTTNBP2 and REEP3, had an excess of regulatory variants. We tested 17 of these using electrophoretic mobility shift assays and found 6 (35%) alter transcription factor-DNA binding in human neuroblastoma cells.

In conclusion, we show that genetic mapping in dogs, one of the best natural models for human psychiatric disease, can find new genes and pathways implicated in OCD. Furthermore, we illustrate how animal models, evolutionary conservation and functional annotation can be used in concert to identify potentially causative functional variants even when genomic data are limited.

INFLUENCE OF DIET, PARASITISM AND HOST GENETICS ON THE BIODIVERSITY OF THE HUMAN GUT MICROBIOTA IN RURAL POPULATIONS FROM CAMEROON

Laure Segurel[1], Elise Morton[2], Alain Froment[1], Evelyne Heyer[1], Molly Przeworski[3], Ran Blekhman[2]

[1]Musée de l'Homme, Eco-anthropology, Paris, France, [2]University of Minnesota, 1. Department of Genetics, Cell Biology, and Development, Minnesota, MN, [3]Columbia University, 3. Department of Biological Sciences, New York, NY

Humans live in close proximity with multiple microbial communities living in and on them. This intimate relationship has been challenged many times during human evolutionary history, for example at the Neolithic revolution during the transition from a nomadic hunter-gatherer to a sedentary farming mode of subsistence, and more recently with the industrialization of human societies. It has notably been shown that there is a substantial loss of biodiversity and an important shift in composition of the gut microbiome in urban industrialized populations when compared to rural populations. However, we know little about the relative contribution of factors such as climate, diet, medicine, hygiene practices, host genetics, and parasitism to these patterns. Here, we focus on fine-scale comparisons of African rural populations in order to (i) contrast the gut microbiomes of populations that inhabit similar environments but have different traditional diets (hunter-gatherers, farmers, fishers); (ii) evaluate the effect of parasitism on microbiome composition and structure; and (iii) identify host genetic variation that is associated with the microbiota in African populations. To this end, we profiled the gut microbiota, intestinal parasites, and host genetic variation in 64 individuals in Southwest Cameroon. We found that the presence of an intestinal protozoa, Entamoeba, is strongly correlated with microbial composition and diversity, such that an individual's infection status can be predicted with 79% accuracy based on his/her gut microbiome composition. We also found gut communities to vary significantly with subsistence mode, with some taxa previously shown to be enriched in other hunter-gatherers groups (notably Succinivibrio sp. and Ruminobacter sp.). Lastly, we identified host genetic variants correlated with gut microbiome composition, and showed that these variants are significantly enriched within genes involved in relevant metabolic pathways (such as vitamin B12 and vitamin D metabolism), and immune system processes. Our study thus highlights the substantial variability in gut microbiome composition among closely related populations and suggests an important role for eukaryotic gut inhabitants, one that has been largely overlooked in studies of the microbiome to date.

EVOLUTIONARY GENOMICS OF THE HORSE DOMESTICATION PROCESS

Ludovic Orlando

University of Copenhagen, Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark

The domestication of the horse in the Pontic-Caspian steppes some 6,000 years ago represents one major turning points in human history. With horses, humans could travel for the first time well above their own speed and carry their germs, culture and genes across vast geographic areas. The development of Horse-drawn chariots and cavalry also radically changed the history of warfare and was instrumental to the emergence of trans-continental empires. Beyond the battlefied, farm horses have also massively impacted agricultural productivity. The biological changes that accompanied the process of horse domestication are, however, difficult to reconstruct from current patterns of genetic diversity both due to the development of intensively selected and extremely influential breeds during the last two centuries, and the almost extinction of wild horses. Recent developments in ancient DNA research have opened for the characterization of complete genomes, epigenomes and microbiota over long time series. We have applied such approaches to a large panel of horse remains spread across Eurasia and dated to 44,000-200 years ago. This started revealing the genetic structure of horse populations prior to and during early domestication stages as well as the history of genetic changes that accompanied their further transformation in a range of cultural contexts. We found that horse domestication did not involve complete genetic isolation, but maintained instead important restocking from the wild. We found that a now-extinct horse lineage significantly contributed to the genetic makeup of the modern horse. Additionally, we identified distinct stages in the domestication process showing different dynamics in the accumulation of deleterious mutations, patterns of diversity loss, and selection targets.

SPECIES GENOME SEQUENCING OF THE ENDANGERED SPIX'S MACAW

Iman K Al-Azwani[1], Nancy Chen[2], Cristina Yumi Miyaki[3], Yasmin A Mohamoud[1], Andrew G Clark[4], Cromwell Purchase[5], Joel A Malek[1,6]

[1]Weill Cornell Medicine-Qatar, Genomics, Doha, Qatar, [2]University of California, Davis, Center for Population Biology, Davis, CA, [3]Universidade de Sao Paulo, Instituto de Biocencias, Sao Paulo, Brazil, [4]Cornell University, Dept. of Molecular Biology and Genetics, Ithaca, NY, [5]Al Wabra Wildlife Preservation, Birds/Macaw Captive Program, Al Shahaniya, Qatar, [6]Weill Cornell Medicine, Dept. of Genetic Medicine, New York, NY

The last known Spix's macaw (*Cyanospitta spixii*) in the wild disappeared in 2000. Its population dwindled to approximately 7 wild-caught birds, found in two nests, from which the modern population is derived. Since then tremendous effort has been given to successfully breed the Spix in captivity and the population now stands at over 100 individuals. However, illegal trade has made the original studbook difficult to interpret and certain pairings of birds do not produce viable offspring. With a goal of re-release into the wild in 2019 it is critical to understand the genetic challenges to this population.

We have sequenced the genomes of approximately 90% of the 2014 world population of Spix's macaw, including all of the animals available to the conservation effort. In addition we have created a high-quality reference genome sequence from single molecule long reads yielding a 1.1 Gb assembly with a contig N50 of 2.4 Mb. For annotation, 20,422 transcripts deriving from 16,589 genes were predicted with input from mRNA-sequencing of five different tissues. Analysis of the pedigree revealed that over 35% of the population show an inbreeding coefficient of greater than 0.25 highlighting the population's precarious situation. These data allow us to answer questions about the recent past history of the animals – what has been the trajectory of effective population size as the population crashed, and was the loss of diversity uniform across the genome as the collapse occurred? Even more exciting, the data provide an unprecedented opportunity to rescue the species from extinction. We have identified putative deleterious mutations in over 1200 genes within individuals of the population, and judicious breeding programs could purge these alleles from the species' genome.

To our knowledge this is the first time a species has effectively been fully sequenced. Just as the exhaustiveness of whole-genome sequencing allowed one to make claims about the absence of features in a genome, species sequencing allows us to assess fixation of deleterious alleles with complete enumeration. The general utility of this approach for species conservation depends not only on sequencing costs and practicalities, but also on the development of methods to employ these data for optimal breeding strategies.

HOW SOCIAL STATUS CHANGES THE IMMUNE SYSTEM: EXPERIMENTAL EVIDENCE FROM RHESUS MACAQUES

Noah Snyder-Mackler[1], Joaquin Sanz[2], Jessica Brinkworth[3], Jordan Kohn[4], Zachary Johnson[4], Mark Wilson[4], Luis Barreiro[2], Jenny Tung[1]

[1]Duke University, Evolutionary Anthropology, Durham, NC, [2]University of Montreal, Pediatrics, Montreal, Canada, [3]University of Illinois, Anthropology, Urbana-Champaign, IL, [4]Emory University, Developmental and Cognitive Neuroscience, Atlanta, GA

In hierarchically organized species, social status can strongly influence fertility, survival, and other fitness-related traits, and in humans, status is one of the best predictors of disease susceptibility. To understand its effects at the molecular level, we used sequential manipulations of social status in female rhesus macaques to (i) establish that social status causally changes gene regulation in immune cells; (ii) localize these effects to specific cell types and signaling pathways; and (iii) investigate how they shape the response to pathogen infection.

To do so, we constructed 9 social groups (n=5 females per group) using a well-established paradigm in which order of introduction predicts dominance rank: earlier introduced animals attain higher status. After monitoring these status hierarchies for a year (Phase 1), we reorganized group composition to maximize changes in rank, and then monitored the new hierarchies for a second year (Phase 1-Phase 2 rank $r=0.063$, $p=0.68$). Using RNA-seq, we profiled gene expression from all females in both phases, in five FACS-sorted immune cell populations. Both independent and meta-analytic approaches revealed highly heterogeneous responses to rank, concentrated in Natural Killer cells and helper T cells (1128 and 451 rank-responsive genes, respectively), but weak or absent in the other three cell types (monocytes, cytotoxic T, and B cells). Our results were highly concordant across Phases, indicating a substantial degree of plasticity upon changes in the social hierarchy.

To test how social status influences the response to pathogens, we used an ex vivo lipopolysaccharide (LPS) stimulation experiment. Gene expression data from paired control and LPS-stimulated samples clearly separated samples by condition (PC1: $r=0.86$, $p<10\text{-}15$) and, within conditions, by dominance rank (PC2: $r=0.59$, $p<10\text{-}8$). Rank x condition interactions were most robust for genes that were more strongly upregulated by LPS in low status females, including NFKB1, a master regulator of inflammation. In contrast, high status females expressed antiviral and type I interferon-related genes more highly, suggesting that social status polarizes the use of alternative arms of the LPS-responsive TLR4 signaling pathway. Together, our results provide novel insight into the molecular basis of social gradients in fitness and health, and the evolution of social hierarchies more broadly.

← Back

VARIATION IN THE MOLECULAR CLOCK OF PRIMATES

Priya Moorjani*[1,2], Carlos Eduardo G Amorim*[1], Peter F Arndt[3], Molly Przeworski[1,4]

[1]Columbia University , Department of Biological Sciences, New York, NY, [2]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA, [3]Max Planck Institute for Molecular Genetics, Molecular Genetics, Berlin, Germany, [4]Columbia University, Dept. of Systems Biology, New York, NY

Events in primate evolution are often dated by assuming a "molecular clock", i.e., a constant rate of substitution per unit time, but the validity of this assumption remains unclear. Among mammals, it is well known that there exists substantial variation in yearly substitution rates. Such variation is to be expected from differences in life-history traits, suggesting that it should also be found among primates. Motivated by these considerations, we analyze whole genomes from ten primate species, including Old World Monkeys (OWMs), New World Monkeys (NWMs) and apes, focusing on putatively neutral autosomal sites and controlling for possible effects of biased gene conversion and methylation at CpG sites. We find that substitution rates are ~65% higher in lineages leading from the hominoid-NWM ancestor to NWMs than to apes. Within apes, rates are ~2% higher in chimpanzees and ~7% higher in the gorilla than in humans. Substitution types subject to biased gene conversion show no more variation among species than those not subject to it. Not all mutation types behave similarly, however: in particular, transitions at CpG sites exhibit a more clock-like behavior than do other types, presumably due to their non-replicative origin. Thus, not only the total rate, but also the mutational spectrum varies among primates. This finding suggests that events in primate evolution are most reliably dated using CpG transitions. Taking this approach, we estimate that the average time to the most recent common ancestor of human and chimpanzee is 12.1 million years and their split time 7.9 million years.

THE PREVALENCE AND ARCHITECTURE OF SEVERE, DOMINANT DEVELOPMENTAL DISORDERS

Matthew Hurles

Wellcome Trust Sanger Institute, Genome Mutation and Genetic Disease Group, Hinxton, United Kingdom

Children with severe, undiagnosed developmental disorders (DDs), including Intellectual Disabilities as well as multi-system congenital malformations, are enriched for damaging de novo mutations (DNMs) in developmentally important genes. We exome sequenced 4,293 families with children with DDs, and meta-analysed these data with published data on 3,287 children with similar disorders. We show that the most significant factors influencing the diagnostic yield of de novo mutations are the sex of the child, the relatedness of their parents and the age of both father and mother. We identified over 90 genes enriched for damaging de novo mutation at genome-wide significance, including fifteen genes for which compelling data for causation was previously lacking. The large number of genome-wide significant findings allow us to demonstrate that, at current cost differentials, exome sequencing has much greater power than genome sequencing for novel gene discovery in genetically heterogeneous disorders. We estimate that 42.5% of our cohort likely carry pathogenic de novo single nucleotide variants (SNVs) and indels in coding sequences, with approximately half operating by a loss-of-function mechanism, and the remainder being gain-of-function. We established that most haploinsufficient developmental disorders have already been identified, but that many gain-of-function disorders remain to be discovered. Extrapolating from the DDD cohort to the general population, we estimate that de novo dominant developmental disorders have an average birth prevalence of ~1 in 300, but that this increases with increasing paternal and maternal age.

CLINICALLY ACCREDITED WGS AS A FIRST LINE DIAGNOSTIC TEST FOR PATIENTS WITH MENDELIAN DISORDERS

Mark J Cowley[1,2], Mark Pinese[1], André E Minoche[1], Tudor Groza[1], Tony Roscioli[1,2,3], Marcel E Dinger[1,2]
[1]Garvan Institute of Medical Research, Kinghorn Centre for Clinical Genomics, Sydney, Australia, [2]University of New South Wales, St Vincent's Clinical School, Sydney, Australia, [3]Sydney Children's Hospital, Department of Medical Genetics, Sydney, Australia

Targeted sequencing (TS), whole exome sequencing (WES), and more recently, whole genome sequencing (WGS) have had a remarkably rapid, and positive impact upon patient diagnosis. Diagnostic rates for patients with rare Monogenic Disorders have risen to on average 25% using WES, and 40-73% using WGS. In our laboratory, WGS has reached near cost parity with high quality WES, making WGS an attractive solution as a comprehensive genetic screen. We were among the first three groups in the world to acquire the HiSeq X, and we have subsequently established the laboratory, clinical and informatic infrastructure to be able to offer clinical WGS. This includes *Patient Archive*, for capturing machine-readable patient phenotype from unstructured clinical notes, *Sabretooth* for local- and cloud-based (ie DNANexus) genomic analysis pipelines, and *Seave*, a comprehensive variant filtration and interpretation platform. We have recently received provisional NATA (like CLIA) accreditation to offer clinical WGS for patient's with rare genetic disorders, making our group the first in Australia to achieve this. The immediate goals of our group are to offer clinical WGS, and build a scalable genome-phenome warehouse to utilise clinical genomes for research & ultimately improve health outcomes.

Through extensive validation work using Genome in a Bottle, and patient samples sequenced with WGS and WES or TS, we consistently find that WGS covers a greater proportion coding exons within disease-associated genes than WES or TS (in part due to capture bias, and missed exons), and offers higher sensitivity to detect SNVs, and especially short (1-20bp) insertions and deletions.

We have achieved 40-86% diagnostic rates from WGS, across a range of genetic disorders, including movement, kidney, heart, neurodevelopmental and mitochondrial disorders. In recent work with autosomal dominant polycystic kidney disease, we obtained an 86% diagnostic yield (24/28), far exceeding the 50% that we obtained using WES, driven by coding an essential splice variants missed by WES, and the identification of small exonic deletions. In a case study of 42 dilated cardiomyopathy (DCM) patients, we identified 100% of the pathogenic variants previously identified by TS. Furthermore, we identified an additional 9 rare, predicted damaging variants missed by TS, in known DCM genes including *TTN* and *NFKB1.pM1V*, and 10 candidate variants in new DCM genes, including a single-exon deletion in *DSC2*, all of which are being currently validated, and assessed for segregation in extended families.

WGS has the potential to become a first-line diagnostic test, capable of replacing clinical microarrays, individual gene tests, and TS panels, particularly for diseases with substantial genetic heterogeneity, such as intellectual disability where the rates of new disease gene discovery continue to climb.

IMPROVING GENETIC DIAGNOSES IN MENDELIAN DISEASE WITH WHOLE GENOME AND RNA SEQUENCING

Beryl B Cummings[1,2,3], Taru Tukiainen[2,3], Monkol Lek[2,3], Fengmei Zhao[2,3], Ben Weisburd[2,3], Leigh Waddell[4], Ana Topf[5], Sandra Donkervoort[7], Volker Straub[6], Carsten Bonnemann[7], Nigel F Clarke[4], Sandra T Cooper[4,8], Daniel G MacArthur[2,3]

[1]Harvard Medical School, Program in Biomedical and Biological Sciences, Boston, MA, [2]Broad Institute of Harvard and MIT, Medical and Population Genetics, Boston, MA, [3]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, [4]The Children's Hospital at Westmead, Sydney, Institute for Neuroscience and Muscle Research, Sydney, Australia, [5]Newcastle University, Institute for Genetic Medicine, Newcastle, United Kingdom, [6]University of Newcastle upon Tyne, Institute of Human Genetics, Newcastle upon Tyne, United Kingdom, [7]National Institute of Neurological Disorders and Stroke, NIH, Neuromuscular and Neurogenetic Disorders of Childhood Section, Bethesda, MD, [8]University of Sydney, Discipline of Pediatrics and Child Health, Sydney, Australia

Exome sequencing is a powerful and cost-effective tool that has become increasingly routine in Mendelian disease diagnosis; however, the current diagnostic rate for exome analysis across a variety of rare diseases is approximately 25-50%. Where exome data falls short, whole genome and RNA sequencing can offer additional insight. Whole genome sequencing provides more uniform coverage and improved identification of structural variants whereas RNA sequencing gives insight into the transcriptional landscape of the affected tissue allowing, for instance, the detection of aberrant splicing and allelic imbalance, all types of events rarely detectable from genotype data alone. Such analyses can empower molecular diagnosis by validating the transcriptional effects of variants proposed by exome data or by identifying novel variants. Here we describe an integrated approach of trio whole genome and patient muscle RNA sequencing in over 50 exome-unsolved cases with severe neuromuscular diseases. Our data consist of individuals for whom exome sequencing has prioritized variants predicted to alter gene expression or RNA splicing as well as those for which there are no candidates from exome sequencing. We demonstrate the power of whole genome and RNA sequencing to validate candidate splice-disrupting mutations and to identify splice-altering variants in both exonic and deep intronic regions. We also apply both technologies to a set of undiagnosed Duchenne muscular dystrophy cases and demonstrate the value of both technologies for detecting and functionally validating complex structural rearrangements (such as inversions) invisible to standard diagnostic methods. Finally, we describe an analysis framework focused on the detection of transcript level changes that are unique to the patient, relative to a database drawn from over 150 skeletal muscle RNA-seq samples from the GTEx project, as well as query heterozygous predicted deleterious variants to identify evidence of allelic imbalance. Together, our results demonstrate the value of both whole-genome and tissue transcriptome data for the detection and interpretation of variants missed by standard exome-based approaches.

← Back

COMPLEX GENETIC OVERLAP BETWEEN SCHIZOPHRENIA RISK AND ANTIPSYCHOTIC RESPONSE

Douglas Ruderfer[1], Alex Charney[1], Ben Readhead[2], Brian Kidd[2], Anna Kahler[3], Paul Kenny[4], Michael Keiser[5], Jennifer Moran[6], Christina Hultman[3], Stuart Scott[2], Patrick Sullivan[7], Shaun Purcell[1,2], Joel Dudley[2], Pamela Sklar[1,2]

[1]Mount Sinai, Psychiatry, NY, NY, [2]Mount Sinai, Genetics, NY, NY, [3]Karolinska Institutet, Epidemiology, Stockholm, Sweden, [4]Mount Sinai, Pharmacology, NY, NY, [5]UCSF, Bioengineering, San Francisco, CA, [6]Broad Institute, Stanley Center, Cambridge, MA, [7]UNC, Genetics, Chapel Hill, NC

Treatments for schizophrenia (SCZ) exist but do not alleviate symptoms for all patients and efficacy is limited by common, often severe side effects. Large-scale genetic studies of both rare and common variation have increased the number of genes and gene sets associated with SCZ risk. However, among the many important remaining questions is how these findings can inform and improve treatment. We hypothesize that genes implicated by genetic studies and those involved in therapeutic efficacy will overlap and by intersecting this information we can further our understanding of both the disorder and the manner in which to treat it. We defined SCZ risk loci as genomic regions reaching genome-wide significance in the latest schizophrenia genome-wide association study (GWAS) and loss of function variants seen only once among 5,079 individuals in an exome-sequencing study of 2,536 SCZ cases and 2,543 controls. Using two comprehensive and orthogonally created databases, we collated drug targets into 167 gene sets of pharmacologically similar drugs and examined enrichment of SCZ risk loci in these groups of drug targets. In addition, we assessed the contribution of rare variants to treatment response.

We identified significant enrichment of SCZ risk loci from both common and rare variation in both known targets of antipsychotics (70 genes, p=0.0078), and novel predicted targets (277 genes, p=0.019). Further, treatment resistant patients had a significant excess of rare disruptive variants in gene targets of antipsychotics (347 genes, p=0.0067) and in genes with evidence for a role in antipsychotic efficacy defined by PharmGKB (57 genes, p=0.0002). Our results support genetic overlap between SCZ pathogenesis and antipsychotic mechanism of action. This finding is consistent with treatment efficacy being polygenic in nature and not solely moderated by one receptor thus implying that continued reliance on single target therapeutics may be insufficient. We further provide evidence of a role for rare functional variants in antipsychotic treatment response pointing to a subset of patients where their genetic information could inform treatment. We present this approach as a framework for enhancing both the understanding of treatment mechanism of action and disease pathology of complex disorders.

← Back

LEARNING A NEW LANGUAGE: TRANSLATIONAL GENOMICS IN DRUG DISCOVERY

Sally John

Biogen, Research, Cambridge, MA

The high rate of failure for novel therapeutic mechanisms in early drug discovery and proof of concept clinical studies is a major challenge for therapeutic development. Limitations in our ability to modulate the target of interest in a way that will result in the desired efficacy and tolerable adverse events, combined with the fact that cell and animal models of disease are imperfect proxies for disease in patients, contribute to high rates of attrition. Large-scale genetic studies have led to a substantial number of DNA variants that are robustly associated with disease and physiological traits, providing clues about novel disease mechanisms in humans. These findings require systematic follow up, focused on producing a body of evidence that validates potential therapeutic mechanisms and results in actionable hypothesis tests in the clinic. Variants associated with a common disease or quantitative traits of disease risk can be used as tool in discovery of novel pathways for therapeutic intervention.
Assuming that genetic variation in humans effectively mimic the effect of drug intervention, allows us to inform likely efficacy in multiple diseases, mechanism-association adverse events and potential biomarkers. This presentation will provide examples of the application of human genetics in drug development from discovery through target validation, to development in the clinical phase.

GENOME-GUIDED DESIGN OF PERSONALIZED CANCER VACCINES

Elaine R Mardis [1], Jasreet Hundal[1], Malachi Griffith[1], Christopher Miller[1], Beatriz Carreno[2], William E Gillanders[3], Gavin Dunn[3,4], Gerald Linette [2], Matthew Gubin[4], Robert Schreiber[4]

[1]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [2]Washington University School of Medicine, Department of Medicine, Division of Oncology , St. Louis, MO, [3]Washington University School of Medicine, Department of Surgery, St. Louis, MO, [4]Washington University School of Medicine, Department of Pathology and Immunology, St. Louis, MO

Large-scale cancer genomics discovery has revealed the multitude of somatic alterations in genes that drive the development and progression of cancer, some of which also render cancer cells sensitive to targeted therapies. In addition to defining the mutational landscape of individual tumor cell genomes, these alterations also result in proteins that appear to the immune system as "non-self" targets, although the immune system seems to be suppressed to mounting a response to eradicate cells expressing the altered proteins. Our work combines genomic analysis to identify mutated peptides produced by individual cancers, then evaluates these peptides relative to the patient's HLA molecules to identify those most likely to elicit an immune response in the context of a personalized vaccine. My talk will present our computational approaches and describe early results from mouse models and first-in-human clinical trials.

PREDICTION OF COLORECTAL TUMOR MUTATIONS USING THE GUT MICROBIOME

Michael B Burns[1], Emmanuel Montassier[2], Dan Knights[2], Ran Blekhman[1]

[1]University of Minnesota, Genetics, Cell Biology, and Development, Minneapolis, MN, [2]University of Minnesota, Computer Science and Engineering, Minneapolis, MN

Recent studies have found an association between the composition of the gut microbiome and colorectal cancer, the third most diagnosed cancer in the USA. Understanding interactions between colorectal tumors and the microbiome in the tumor microenvironment is critical for elucidating the potentially causal role of bacteria in colorectal cancer, and for development of therapeutics that target the microbiota. To that end, we have investigated how different tumor characteristics, and especially tumor mutational profiles, affect the attached microbial communities. To do so, we performed whole-exome sequencing and microbiome profiling in tumors and patient-matched normal colorectal tissue samples. By jointly analyzing these data, we find a strong association between tumor mutation patterns and shifts in the microbiome. We show that somatic mutations in certain genes are associated with changes in abundance of specific bacterial taxa, including a link between loss-of-function mutations in APC and the abundance of Fusobacterium, a known cancer-associated taxon. Lastly, we built a mutation risk index from a panel of microbes associated with each of several highly prevalent tumor mutations. We show that we can use this index to accurately predict the existence of loss-of-function mutations in cancer-related genes (including in APC) and pathways (including in MAPK signaling and Wnt signaling pathways), solely based on the composition of the tumor-associated microbial communities. These results serve as a starting point for development of colon cancer prognostics and individualized microbiota-targeted therapies.

SINEUPS, A NEW CLASS OF TRANSLATION REGULATORY RNAs: FROM FUNCTION TO FUTURE GENE THERAPY

Hazuki Takahashi[1], Kazuhiro Nitta[1], Aleks Schein[1], Chung Chau Hon[1], Harshita Sharma[1], Silvia Zucchelli[2], Stefano Gustincich[2], Piero Carninci[1]

[1]RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Japan, [2]International School for Advanced Studies (SISSA), Area of Neuroscience, Trieste, Italy

High throughput transcriptome studies have identified that the majority of the genome is transcribed into long non-coding RNAs (lncRNAs). In the Functional Annotation of Mammalian Genome (FANTOM) project, we have identified more than 23,000 lncRNA both for the mouse and the human genome. A new catalogue of human lncRNAs is being prepared, which extensively expands the current GENCODE by ~ 15,000 high confidence lncRNA transcripts. Those transcripts frequently overlap GWAS SNPs, for which no functional elements were known. Assessing the function of the lncRNA is being broadly tackled by the FANTOM6 project, which is ongoing at RIKEN.

An important subset of lncRNA is constituted by antisense RNAs, which are though generally to control the expression of the sense genes. CAGE data has shown antisense transcription for the majority of the protein coding genes. We have identified a class of non-coding antisense RNAs, which have the surprising property to up-regulate protein translation of the sense RNA that they overlap. Enhancement of protein translation is mediated by SINE elements (SINEB2 in the mouse versions), which share common ancestor with tRNAs. The specificity of action is mediated by the region antisense to the 5'UTRs of the target mRNAs. We renamed these RNAs "SINEUPs", because they contain a SINE elements that can up-regulate translation.

We are extensively characterizing natural SINEUPs and the functional domain of this interesting new class of RNAs, identifying several novel mouse and human natural SINEUPs and are mapping the functional domain of these lncRNAs. Although the active SINE elements differ in sequences, they show common function. Very importantly for future gene therapeutics applications, artificial SINEUPs can also be designed to target regions around the 5'UTR of essentially any target RNAs. Although mostly used SINEUPs originate from the mouse, they are active in other mammalian and lower vertebrates, suggesting that specie-specific SINEs have function that goes across species borders, probably mediated by specific structures.

We will present the most recent data characterizing the functional elements of SINEPUPs and initial applications to correct gene expression in haploinsufficiencies and other gene unbalances in cells and in disease model using SINEUPs. We believe that this novel class of lncRNA, acting in the opposite direction of miRNAs, may find much broader applications in research and gene therapy.

IMPROVING THE REPRODUCIBILITY OF CLINICAL GENETIC TESTS: CHALLENGES AND SOLUTIONS

Stephen Lincoln[1], Leif Ellisen[2], Allison Kurian[3], David Haussler[4], Shan Yang[1], Benedict Paten[4], Robert Nussbaum[1]

[1]Invitae, San Francisco, CA, [2]Massachusetts General Hospital, Boston, MA, [3]Stanford University, Palo Alto, CA, [4]University of California, Santa Cruz, CA

Recently [1,2] we demonstrated that multigene NGS tests can produce clinically valid results, comparable to traditional genetic tests, while expanding treatment options available to physicians and patients. These findings focused on hereditary cancers, although similar results are known in cardiology, neurology, and pediatrics. Challenges however remain to making these tests ubiquitous, particularly concerns about reproducibility and accuracy in both to variant detection and interpretation.

Concerns regarding variant detection are real: About 10% of actionable variants we see are technically challenging. Many are single exon (or sub-exon) CNVs. Others are in genes with pseudogenes, low complexity sequence, or involve complex DNA changes. We find that 6 calling algorithms and manual data review can be required to address the full spectrum, even for genes well represented in GRCh37.

Validation resources have not kept up. The Genome in a Bottle (GIAB) is tremendously useful, although 15% of our clinically tested genes are in the 23% of the genome for which high quality GIAB data are not yet available. Moreover the GIAB has few complex variants in clinically tested regions. New efforts to address this gap have been launched by the GIAB consortium and the GET-RM project.

Concerns about variant interpretation are also real, although this is sometimes overstated. Some laboratories guard data as a proprietary asset, contrary to the best interests of patient care, and these labs tend to prominently highlight disagreements in public databases. Others see databases like ClinVar as a tool, allowing helpful interlaboratory peer-review and quality control.

We used ClinVar to evaluate the concordance of BRCA1/2 variant interpretations across labs, including a major lab that does not submit to ClinVar but where data are available. We find very high concordance (98.5%) when counting differences that would change the actionability of a result. Moreover because all of the discordant variants are very rare, we calculated the expected chance of a patient receiving discordant results as 0.2%, similar to the result in our prospective study [2]. New refinements to the ACMG guidelines promise to make consistent variant interpretation even more scalable across large teams of lab directors and scientists.

In summary, tools now exist to help make clinical tests increasingly robust, which will speed their adoption in routine health care.

1. Desmond, JAMAOncology 2015
2. Lincoln, JMolDiag 2015

eQTL ANALYSIS OF LUNG ADENOCARCINOMA EXPRESSION SUBTYPES

Andrew Quitadamo, Xinghua Shi

University of North Carolina at Charlotte, Bioinformatics and Genomics, Charlotte, NC

Lung cancer is the most common type of cancer world wide, and is responsible for the most cancer deaths in the United States. Approximately one quarter of US cancer deaths in 2015 were due to lung cancer, in part because the 5-year survival rate is below 20%. Investigating the biological basis and mechanisms of lung cancer could lead to a better understanding of its development and progression. Previous studies have found three distinct expression subtypes in lung adenocarcinoma, which overlap clinical features. For example patients with the Terminal Respiratory Unit, or Bronchioid, subtype have more favorable outcomes, while tumors of the Proximal-inflammatory, or Squamoid, subtype tend to be more poorly differentiated. Expression quantitative trait loci (eQTL) analysis is a useful method to investigate the impact of human genetic variation on gene expression. Using genotype and gene expression data from The Cancer Genome Atlas lung adenocarcinoma dataset we use an eQTL analysis to find commonalities and differences between the subtypes. Using the eQTL genes and the subtype specific genes as inputs, we create a protein-protein interaction network for each subtype. By comparing the results of the subtype eQTLs, the protein-protein interactions, and the clinical phenotype, we are able to elucidate some of the genetic components of these molecular subtypes.

SMRT SEQUENCING REVEALS COMPLEX STRUCTURE OF THE SEX DETERMINATION LOCUS IN ATLANTIC HERRING

Nima Rafati[1], Chungang Feng[1], Sangeet Lamichhaney[1], Alvaro Martinez Bario [2], Mats Petterson[1], Ignas Bunikis[3], Carl-Johan Rubin[1], Leif Andersson[1,4,5]

[1]Uppsala University, Medical Biochemistry and Microbiology, Uppsala, Sweden, [2]Uppsala University, Cell and Molecular Biology, Uppsala, Sweden, [3]Science for Life laboratory, National Genomics Infrastructure, Uppsala, Sweden, [4]Swedish Agricultural Sciences, Animal Breeding and Genetics, Uppsala, Sweden, [5]Texas A&M University, Veterinary Integrative Bioscience, Texas, TX

Several different sex determination systems have evolved in vertebrates, in particular in fish species. The plasticity of fish genomes has generated diverse reproductive and sex determination strategies compared to other vertebrates. Fishes (similar to some reptiles) have established two distinct sex determination systems: genetic sex determination (GSD), or environmental determination (ESD), and these two systems sometimes work in concert. GSD can be under control of sex chromosomes or master genes on autosomal chromosomes but in most fish species genes with predominant roles in sex determination have not been reported. An exception is a dmrt1 paralog (dmy) in the medaka (Oryzias latipes) Y-chromosome governing sex determination. In many teleost species, including Atlantic herring (Clupea harengus), the sex-determination system is still unknown. Atlantic herring is among few marine species reproducing in different salinities (2-35‰) and seasons. These adaptations and reproductive strategies have made herring a unique model for evolutionary studies. We generated a high-quality draft genome assembly by short read sequencing technology where we identified a large region (~100 Kb) for which males and females showed significant differentiation. We extended our analysis to unmapped reads where we found male unique sequences belonging to a member of a *sperm-associated protein* gene family. But our efforts in linking these two segments by PCR failed. To gain further insight into the herring genome, we generated a new assembly by single-molecule real-time (SMRT) sequencing technology. In this new assembly we were able to reveal the structure of the previously observed male specific sequence representing early stages of sex chromosome evolution. This is the first report on identifying a sex-determination locus and proto-Y chromosome in Atlantic herring. This study will help to add to our understanding about sex chromosome evolution in this species and other teleosts.

EMASE: ACCURATE ESTIMATION OF ALLELE-SPECIFIC EXPRESSION USING AN EM ALGORITHM

Narayanan Raghupathy, Kwangbom Choi, Steve Munger, Ron Korstanje, Gary Churchill

The Jackson Laboratory, Center for Genome Dynamics, Bar Harbor, ME,

Allele-specific expression (ASE) refers to the differential abundance of the allelic copies of a transcript in a diploid organism. An accurate estimation of ASE from RNA-seq data requires alignment strategies that accommodate polymorphisms to minimize alignment artifacts and biases. Current approaches to quantify ASE typically consider only unique mapping reads, and consider evidence only at the SNP locations, and cannot quantify allele level expression for isoforms. We present an integrated statistical approach to quantify ASE that utilizes read alignment to a diploid transcriptome and applies an Expectation-Maximization algorithm that explicitly resolves the hierarchy of reads that multiply align to genes, alleles, and isoforms respectively. Our software EMASE (https://github.com/churchill-lab/emase) can be applied to estimate ASE from RNA-seq data in any diploid organism with known genetic variations or imputed genomic sequence. We first demonstrate the utility of EMASE using simulated data by comparing ASE estimates from other methods like WASP, RSEM, and Kallisto. Then, we apply EMASE to reciprocal F1 mice data to address whether the cause of ASE is due to local genetic variants or genomic imprinting.

THE DOE SYSTEMS BIOLOGY KNOWLEDGEBASE (KBASE): FAST AND FLEXIBLE RNA-SEQ ANALYSIS OF PLANTS AND MICROBES

Srividya Ramakrishnan[1], James Gurtowski[1], Michael C Schatz[1], Sunita Kumari[1], Shinjae Yoo[2], Priya Ranjan[3], Jim Thomason[1], Vivek Kumar[1], Fei He[2], Samuel Seaver[5], David Weston[3], Doreen Ware[1,4], Nomi Harris[6], Robert W Cottingham[3], Sergei Maslov[2], Rick Stevens[5], Adam P Arkin[6]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Brookhaven National Laboratory, Upton, NY, [3]Oak Ridge National Laboratory, Oak Rigde, TN, [4]USDA ARS, ITHACA, NY, [5]Argonne National Laboratory, Argonne, IL, [6]Lawrence Berkeley National Laboratory, Berkeley, CA

The U.S Department of Energy Systems Biology Knowledgebase (KBase, http://kbase.us) provides an open web-accessible system for systems biology research focused around microbes, plants and communities. It supports biological data analysis, sharing and integration of biological data from a variety of analysis tools that include, modeling and simulation methods and visualizations.

KBase allows users to integrate new analysis tools and data, and now contains a rich set of computational methods and curated datasets for performing gene expression analysis from RNA-Seq data. This includes a selection of preprocessed high-quality reference genomes, a searchable data store of high-value RNA-Seq data sets, and a wide variety of analytical methods including algorithms for short-read mapping, identification of splice junctions, transcript and isoform detection and quantitation, differential expression and visualization. A gene expression "pipeline" example in KBase is centered around the widely used Tuxedo suite of RNA-Seq tools (Bowtie, Tophat, Cufflinks, Cuffdiff, and CummeRbund) and also includes several novel interactive visualizations and analytical components. The platform also provides other useful services that are directly integrated with gene expression profiles that predict metabolic networks and provide other gene expression related analyses. All of these components are backed by a highly-scalable, high-performance computing backend with many cores and terabytes of space available.

The RNA-Seq analysis services are available from within a highly interactive and dynamic user interface called the Narrative Interface. Within a Narrative, short reads from an RNA-Seq experiment can be uploaded into KBase to perform gene expression analysis and the results can be shared, reproduced, and the research extended by others in the KBase community. We demonstrate the utility of the Narratives by performing a point-and-click, yet detailed analysis of public RNA-Seq data from several species and tissue types, including experiments in Arabidopsis thaliana and E. coli.

IN VITRO GENE-BY-ENVIRONMENT INTERACTIONS ARE RELEVANT FOR COMPLEX TRAITS

Allison L Richards[1], Gregory A Moyerbrailean[1], Cynthia Kalita[1], Daniel Kurtz[1], Omar Davis[1], Christopher Harvey[1], Adnan Alazizi[1], Donovan Watza[1], Yoram Sorokin[3], Nancy Hauff[3,4], Xiang Zhou[2], Xiaoquan Wen[2], Roger Pique-Regi[1,3], Francesca Luca[1,3]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [3]Wayne State University, Department of Obstetrics and Gynecology, Detroit, MI, [4]Wayne State University, College of Nursing, Detroit, MI

Recent studies have highlighted the importance of gene-by-environment interactions in molecular phenotypes and complex traits. In order to study these interactions in a systematic way, we designed an experiment where we exposed 5 cell types to 35 treatments, with each treatment chosen due to humans' frequent exposure to it. These treatments included hormones (e.g., glucocorticoids, insulin), dietary components (e.g., vitamin A, selenium), pollutants (e.g., BHA, phthalates), and common over the counter drugs (e.g., aspirin). We collected and deeply sequenced the RNA (130M reads/sample on average) from each of these environments (defined by cell type and treatment) across three individuals. 32,838 genes are differentially expressed (10% FDR) in any of the 89 environments. Some of these gene responses have been previously reported, such as *FKBP5* response to dexamethasone; while many others are novel and characterize previously unknown treatment effects. We next investigated allele specific expression (ASE) in our samples. Across the 89 environments, we identified 11,772 ASE events (10% FDR) in 1,530 unique genes. While the majority of ASE occurs across multiple environments, we also identified 161 genes that demonstrate gene-by-environment interactions. Specifically, these 161 genes have conditional ASE (cASE), or ASE in a specific environment. Furthermore, we identified 63 genes where expression was very low in control samples and was increased by at least 5-fold in the treatment, resulting in ASE (inducible ASE, iASE). We found that 47% of genes reported with cASE or iASE are found among GWAS genes, highly enriched compared to the 21% of differentially expressed genes among GWAS (p-value < 10_{-15}, OR = 3.3). Many of these instances may help to shed light on the missing heritability of various complex traits. For example, we identified cASE in a gene, *LAMP3*, which has previously been associated with Parkinson's disease. This cASE occurs in response to selenium, a factor that reduces bradykinesia. By comparing the cASE SNP to that identified in the GWAS, we find that selenium increases the expression of the allele in phase with the protective allele of *LAMP3*. In this example, and with others, we are able to suggest candidate genes and environments that may play a role in various disease states through gene-by-environment interactions.

← Back

HIGH THROUGHPUT SINGLE-MOLECULE MAPPING LINKS SUBTELOMERIC VARIANTS, LONG-RANGE HAPLOTYPES, AND TELOMERE LENGTH PROFILES WITH SPECIFIC HUMAN TELOMERES

Eleanor Young[1], Steven Pastor[1], Ramakrishnan Rajagopalan[1], Jennifer McCaffrey[1], Justin Sibert[1], Angel Mak[2], Pui-Yan Kwok[2], Harold Riethman[3], Ming Xiao[1]

[1]Drexel University, Biomedical Engineering, Philadelphia, PA, [2]UCSF, Dermatology, San Francisco, CA, [3]Old Dominion University, Medical Diagnostic and Translational Sciences, Norfolk, VA

Accurate maps and DNA sequences for human subtelomere regions, along with detailed knowledge of subtelomere variation and long-range telomere-terminal haplotypes in individuals, are critical for understanding telomere function and its roles in human biology. DNA elements cis to the (TTAGGG)n tract regulate both TERRA levels and haplotype-specific (TTAGGG)n tract length and stability. Large structural variations causing altered juxtaposition of subtelomeric sequence elements and 1-copy DNA relative to the telomere may affect gene expression and the packaging of telomeric chromatin. De novo deletion of subtelomeric duplications can cause disease in some contexts, and long-range interactions of telomeres with subtelomeric genes can regulate the expression of specific subtelomeric genes in a telomere length-dependent fashion.

We have recently developed a highly automated whole genome mapping technology in nano-channel arrays. Here, we use this technique to analyze large terminal human chromosome segments extending from chromosome-specific subtelomere sequences through subtelomeric repeat regions to terminal (TTAGGG)n repeat tracts. We establish detailed maps for subtelomere gap regions in the human reference sequence, detect many new large subtelomeric variants, and demonstrate the feasibility of long-range haplotyping through segmentally duplicated subtelomere regions. Based on single molecule mapping data along with CRISPR/Cas9-directed labeling of (TTAGGG)n tracts, we were able to estimate telomere lengths linked to distinguishable telomeric haplotypes.

This method opens the door to the characterization of difficult genomic regions using long uncloned single genomic DNA molecules, making it a uniquely valuable new tool for improving the quality of genome assemblies. We demonstrate precise gap characterization as well as delineation of variant subtelomeric haplotypes relative to the current reference sequence for several individual human genomes. Finally, we demonstrate proof of principle for single-telomere genotyping, where allele-specific single telomere length profiles can be linked to subtelomeric repeat element organization and long-range haplotypes. This methodology will enable, for the first time, delineation of specific human cis elements involved in telomere length regulation.

IDENTIFYING THE SOURCE OF ROTAVIRUS VIRULENCE USING SENSITIVE SEQUENCE METHODS

Firas M Riyazuddin*[1,2], Mileidy W Gonzalez*[1], John L Spouge[1]

[1]National Library of Medicine, National Center for Biotechnology and Information, National Institutes of Health, Statistical Computational Biology Lab, Computational Biology Branch, Bethesda, MD, [2]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Adult Inter-Institute Clinical Endocrinology and Metabolism Fellowship Training Program (IETP), Bethesda, MD

* Authors contributed equally.

As one of the leading causes of childhood diarrhea across the world, Rotaviruses are an important public health problem and a prime focus for vaccine development. Antigen-antibody assays have identified different sub-groups of the virus of pathogenic importance, Rotavirus A being the predominant serogroup. Recent efforts have been focused on understanding the various pathogenic strains at the sequence level in the hope of developing more targeted diagnostics and therapeutics, as well as possible new sequence-based genomic classifications. As a segmented double stranded RNA virus, its antigenic diversity and hence pathogenic potential is critically dependent on its genomic composition. From one generation to the next, periods of "antigenic drift" are punctuated by episodes of "antigenic shift" resulting from large scale rearrangement of the segments within its genome. A recent paper by Libonati et al (2014)[1] examined serologically identical G10P11 Rotavirus A isolates from 19 symptomatic and 20 asymptomatic patients. Libonati et al found no significant sequence differences between them, leaving open the surprising question of whether phenotypic differences in virulence can be traced genotypically at the sequence level. Libonati et al failed to find differences in Rotaviruses when using sequence identities calculated by standard sequence methods like BLAST and ClustalW. Thus, we are developing an algorithm combining nonparametric statistical tests with sensitive sequence analysis to identify site-specific differences as signatures of virulence. Given the higher information content of proteins, we test for differences between the test groups at the protein level. Our study also takes advantage of the increased sensitivity afforded by profile-based methods like Hidden Markov Models and the improved alignments of T-coffee and MUSCLE. Identifying molecular signatures of virulence could be instrumental in understanding the pathogenicity of Rotaviruses and informing the development of an effective vaccine.

Disclosure: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine, Bethesda, MD, USA. The authors have no conflicts of interest to disclose.

1. Libonati MH, Dennis AF, Ramani S, McDonald SM, Akopov A, Kirkness EF, Kang G, Patton JT. Absence of genetic differences among G10P [11] rotaviruses associated with asymptomatic and symptomatic neonatal infections in Vellore, India. Journal of virology. 2014 Aug 15;88(16):9060-71.

← Back

GWAS REPLICABILITY ACROSS TIME AND SPACE

Juan A Rodriguez[1], Urko M Marigorta[4], Arcadi Navarro[1,2,3]

[1]Institute of Evolutionary Biology - Universitat Pompeu Fabra (UPF-CSIC) , Department of Experimental and Health Sciences, Barcelona, Spain, [2]Centre de Regulaciò Genòmica (CRG), Bioinformatics and Genomics, Barcelona, Spain, [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [4]Georgia Institute of Technology, School of Biology, Atlanta, GA

The key step to validating associations between genetic variants and complex diseases is the replication of findings in independent samples. This was, perhaps, the most valuable lesson learned from candidate-gene association studies that were performed prior to the era dominated by Genome-Wide Association Studies (GWAS).
In the present study we evaluated the discovery and replication of GWAS risk variants using the NHGRI GWAS Catalog. Two main aspects of replication are considered – time and geographic space.
First, we are evaluating temporal replication of 46 disease traits during the last 10 years. Three different levels (LD blocks, genes and pathways) are being considered in our evaluation of replication. Replication of associated LD blocks has been high since the beginning of the GWAS era (2005), but the discovery of novel associations has been on the decline since 2012. However, replication at the gene level has been substantially increasing, supporting the notion that several causal variants, in different LD blocks, are located within a single gene. Additionally, we find that different disease categories (immune, cancer, mental) exhibit highly heterogeneous, temporal replicability rates.
Second, we are extending on previous observations of geographic replication (1), which illustrated a remarkable replication of effect sizes among continental populations ($r \approx 0.78$). Here we are focusing on the sharing of genetic architectures among populations, and the feasibility of using polygenic risk scores to predict diseases in cohorts from another continent.
Finally, a hypothesis for the lack of replication in GWAS is being developed and tested. The hypothesis rests on the biology of gene networks and diseases being the product of disruptions among genes within a network. Thus by studying associations in the context of networks we may recover what appears to be a paucity of replication.

1.Marigorta UM, Navarro A (2013) High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. PLoS Genet 9(6):e1003566

PAPIO BABOONS: A PRESENT-DAY MODEL FOR ANCIENT HOMININ GENETIC INTROGRESSION

Jeffrey Rogers, Kim C Worley, Muthuswamy Raveendran, R. Alan Harris, Richard A Gibbs, for the Baboon Genome Sequencing and Analysis Consortium

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

An international consortium of researchers is investigating the content, evolution and diversity of genomic sequences among *Papio* baboons. This genus of Old World monkeys is widely used as a biomedical model of human disease, but is also exceptional in providing a unique model for analyses of speciation and hybridization among primates. The six species of *Papio* baboons exhibit substantial differences in coat color, anatomy and social behavior, but well-documented hybrid zones with fertile F1 and F2 hybrids exist in at least four African locations where two baboon species meet in the wild. We sequenced the genome of an olive baboon (*Papio anubis*) and produced a draft assembly (contig N50 40.3 kb; scaffold N50 529 kb). We also generated whole genome sequences for 15 additional *Papio* baboons representing all six species in the genus. Standard SNP calling methods identified 42.6 million SNVs. This variation within and between species was used to reconstruct a phylogeny and determine whether past hybridization and gene flow have been sufficient to influence genome content across species. Whole genome analyses produce a well-supported phylogeny for baboon species, but also document ancient gene flow and admixture among the diverging evolutionary lineages. Furthermore, hybrid baboons in captive colonies show elevated frequencies of craniofacial abnormalities. Given the opportunity to directly investigate active natural baboon hybrid zones where divergent species are currently interbreeding, as well as new genomic evidence for recurrent historical introgression, baboons are an outstanding model for the complex evolutionary and demographic processes that influenced the evolution of the Neanderthal, Denisovan and modern human genomes. At a deeper timescale, the last common ancestor (LCA) of baboons and macaques (genus *Macaca*) lived at about the same time as the LCA of humans, chimpanzees and gorillas. Comparisons between baboons and macaques illuminate patterns of genomic evolution, providing new perspective on human-chimpanzee-gorilla genetic divergence. Assessing genome-wide differences among species shows nucleotide substitutions occurred more slowly in humans and apes than in macaques and baboons. Analyses of retrotransposon insertions, transcriptome evolution, methylation and post-translational modifications of proteins provide additional insights and a broader context for reconstructing human genomic evolution.

DEVELOPMENT OF A HIGH-THROUGHPUT CLINICAL TUMOR SEQUENCING WORKFLOW

Jeffrey A Rosenfeld[1], Ying Chen[1], Li Liang[2], Jay Tischfield[2], David Foran[1], Amrik Sahota[2]

[1]Rutgers Cancer Institute of NJ, Pathology, New Brunswick, NJ, [2]RUCDR, Pathology, Piscataway, NJ

Next generation sequencing of cancer tissue is becoming a mainstream technique in clinical laboratories because of its potential to contribute to the design of patient-specific therapies. Here, we describe the validation of the ThunderBolts (TB) Cancer Panel from RainDance Technologies for the detection of sequence variants in DNA from FFPE tissue from patients with a variety of solid tumors. The ThunderBolts panel targets hotspot mutations within 50 genes with the highest chance of mutation. This panel is much smaller than others that are commonly used, but for the majority of cases, it is sufficient.

We have developed a complete workflow and the IT and bioinformatics services to facilitate each step starting with sample extraction and resulting in final clinical reporting. DNA is extracted from the FFPE samples, amplified on the RainDance instrument and them sequenced on a MiSeq. After the sequencing is completed, we use Base Space for initial alignment and variant calling. Next, Golden Helix's VarSeq software is used to filter the variants down to those most likely to be pathogenic. Finally, the services of N-of-One and a series of automated scripts, developed in-house are used to generate a signed clinical report. By developing this assay in-house, we can achieve much lower costs and faster turnaround time than would be required for a commercial send-out test. We have compared our results using a relatively small number of targets to those obtained from other assays, and have consistently found a high concordance of results. At the current stage of development, our team will use our assay as a first-level test and limit the send out samples to only those cases where the small panel does not identify any targetable mutations.

HTLV-1/BLV ANTISENSE RNA-DEPENDENT CIS-PERTURBATION OF CANCER DRIVERS IN LEUKEMIC AND PRE-LEUKEMIC CLONES

Nicolas Rosewick*[1,2], Keith Durkin*[1], Ambroise Marçais[3], Maria Artesi[1], Vincent Hahaut[1], Philip Griebel[4], Natasa Arsic[4], Arsène Burny[2], Carole Charlier[1], Olivier Hermine[3], Michel Georges[1], Anne Van den Broeke[1,2]

[1]Unit of Animal Genomics, GIGA-R, ULg, Liège, Belgium, [2]Experimental Hematology, Institut Jules Bordet, ULB, Bruxelles, Belgium, [3]Service d'Hématologie, Hôpital Universitaire Necker, Paris, France, [4]VIDO, University of Saskatchewan, Saskatoon, Canada

*authors contributed equally

More than 10 million humans and 50 million of dairy cattle are infected with the closely-related Human T-cell leukemia virus type-1 (HTLV-1) and Bovine Leukemia Virus (BLV) respectively. These retroviruses infect T (HTLV-1) and B (BLV) lymphocytes, provoking a polyclonal expansion that will evolve into an aggressive lethal monoclonal leukemia in ~5% of individuals following decades of latency. It is generally assumed that oncogenic changes are largely dependent on virus-encoded products and especially the trans-activating effects of the Tax and HTLV-1 bZIP HBZ oncoproteins, while progression to acute leukemia involves somatic mutations in host genes, yet is independent of proviral integration site that has been found to be very variable between tumors. We herein upset this dogma, by demonstrating that HTLV-1/BLV integrate in the vicinity of host cancer driver genes, which they affect either by provirus-dependent transcription termination (~ 25%) or as a result of viral antisense RNA-dependent cis-perturbation (~ 75%). Virus-driven interactions with the tumor genome depend on positive-strand silenced proviruses that have lost their capacity to express Tax. The same pattern was observed in early "pre-leukemic" samples obtained from asymptomatic BLV-infected sheep, suggesting that provirus-dependent host gene perturbation triggers initial amplification of the corresponding clones, requiring additional genetic and/or epigenetic changes to develop full-blown leukemia.

DEVELOPMENT AND ANALYSIS OF THE exRNA ATLAS REVEALS HIGHLY DIVERSE POPULATIONS OF SMALL-RNAs IN HUMAN BIOFLUIDS

Joel Rozowsky[1], Robert Kitchen[1], Sai Subramanian[2], William Thistlethwaite[2], Roger Alexander[3], David Galas[3], Matt Roth[2], Aleksander Milosavljevic[2], Mark Gerstein[1]

[1]Yale University, Department of Biophysics & Biochemistry, New Haven, CT, [2]Baylor College of Medicine, Bioinformatics Research Laboratory, Department of Molecular & Human Genetics, Houston, TX, [3]Pacific Northwest Diabetes Research Institute, Seattle, WA

The Extracellular RNA Communication Consortium (ERCC) has been established to explore the potential role of extracellular RNA (exRNA) for clinical diagnostics and therapy. We present the tools that have been developed for the analysis of exRNA and the construction of a comprehensive atlas of exRNAs in human body fluids. Due to the process of extracting, purifying, and sequencing of RNA from extracellular biofluids, exRNAs are more vulnerable to contamination than cellular RNA samples. Thus we have developed the extra-cellular RNA processing tool (exceRpt), optimized for the analysis of exRNA-seq datasets. This involves:
1. Pre-processing: Supports random-barcoded libraries, spike-in sequences for calibration or titration, and explicit removal of common laboratory contaminants and ribosomal RNAs.
2. Endogenous alignment: Alignment to the genome and full set of annotated, potentially spliced, endogenous RNA transcripts including all known miRNAs, tRNAs, piRNAs, snoRNAs, lincRNAs, mRNAs, retrotransposons, and circular RNAs.
3. Exogenous alignment: Alignment to all annotated exogenous miRNAs in miRBase and all exogenous rRNA sequences in the RDP. Finally, alignments to the genomes of all bacteria, viruses, plants, fungi, protists, metazoa, and select vertebrates.
The exceRpt pipeline (available at genboree.org and github.gersteinlab.org/exceRpt) generates a variety of quality control metrics, produces abundance estimates for various RNA species, and makes available alignment information for visualization and validation.
The public exRNA atlas is accessible via the consortium website at exRNA.org. The initial release of the atlas includes 519 exRNA profiles derived from over 6 billion reads processed uniformly using the exceRpt pipeline. Faceted filtering and data navigation tools are enabled by rich metadata standards developed by the consortium and metadata annotations contributed by the data producers. Users can browse, select, filter, and download processed results and raw data files of open access datasets available in the exRNA Atlas. To complement the exRNA atlas we have also created a comprehensive cellular small-RNA atlas by uniformly re-processing ~1,000 datasets in the SRA. The tools we have developed can be accessed at genboree.org and include data analysis and data submission pipelines, and exRNA metadata tracking tools.

POSITIVE SELECTION ON LOCI ASSOCIATED WITH DRUG AND ALCOHOL DEPENDENCE

Brooke Sadler[1], Gabe Haller[2], Howard Edenberg[3], Jay Tischfield[4], Andy Brooks[4], John Kramer[5], Marc Schuckit[6], John Nurnberger[7], Alison Goate[8]

[1]Washington University School of Medicine, Psychiatry, St. Louis, MO, [2]Washington University School of Medicine, Orthopedic Surgery, St. Louis, MO, [3]Indiana University, Molecular Biology, Indianapolis, IN, [4]Rutgers University, Genetics, Piscataway, NJ, [5]University of Iowa, Psychiatry, Iowa City, IA, [6]University of San Diego, Psychiatry, La Jolla, CA, [7]Indiana University, Psychiatry, Indianapolis , IN, [8]Mount Sinai Ichan School of Medicine, Neuroscience, New York City, NY

Much of the evolution of human behavior remains a mystery, including how certain disadvantageous behaviors are so prevalent. Nicotine addiction is one such phenotype. Several loci have been implicated in nicotine related phenotypes including the nicotinic receptor gene clusters (*CHRNs)* on chromosomes 8 and 15. Here we use 1000 Genomes sequence data from 3 populations (Africans, Asians and Europeans) to examine whether natural selection has occurred at these loci. We used Tajima's D and the integrated haplotype score (iHS) to test for evidence of natural selection. Our results provide evidence for strong selection in the nicotinic receptor gene cluster on chromosome 8, previously found to be significantly associated with both nicotine and cocaine dependence as well as evidence of weaker, but still detectable, selection acting on the region containing the *CHRNA5* nicotinic receptor gene on chromosome 15, that is genome wide significant for risk for nicotine dependence. To examine the possibility that this selection is related to memory and learning, we utilized genetic data from the Collaborative Studies on the Genetics of Alcoholism (COGA) to test variants within these regions with three tests of memory and learning, the Wechsler Adult Intelligence Scale (WAIS) Block Design, WAIS Digit Symbol and WAIS Information tests. Of the 17 SNPs genotyped in COGA in this region, we find one significantly associated with WAIS digit symbol test results. This test captures aspects of reaction time and memory, suggesting that a phenotype relating to memory and learning may have been the driving force behind selection at these loci. This study could begin to explain why these seemingly deleterious SNPs are present at their current frequencies.

RAPID ANONYMIZED LOOKUPS OF *DE NOVO* STRUCTURAL VARIANTS FOR WHOLE-GENOME TRIOS

William J Salerno[1], Sri Niranjan Shekar[2], Adam C English[1], Adina Mangubat[2], Jeremy Bruestle[2], Eric Boerwinkle[3], Richard A Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Spiral Genetics, Seattle, WA, [3]UT Health Science Center, Human Genetics Center, Houston, TX

Next-generation sequencing has succeeded as a tool to understand the genetic foundations of rare, Mendelian disease. The Baylor Miraca Genetics Laboratory routinely performs whole-exome sequencing for children at-risk for genetic disease and has achieved a 25% solve rate based upon the analysis of more than 10,000 probands. Increasingly, whole-genome sequencing (WGS) is being turned to for the next step in clinical diagnostics, with the ambition of solving the remaining 75% of cases. Whole-genome sequencing provides access to the full spectrum of genomic variation in both coding and non-coding regions. However, assessment of large, structural variants (SVs) in a clinical setting poses several challenges. Because the unit of observation (a read) is smaller than the desired event, methods to detect, genotype, and query SVs often require the raw data (BAMs) and are thus slow, expensive and sensitive to data heterogeneity.

The Biograph format (GBWT) provides reference-agnostic read compression that allows for rapid and scalable genotyping of SVs. Biograph compression scales less than linearly for multiple samples and implicitly anonymizes sample data by eliminating shared reads. A GBWT for 125 samples can be queried for SV genotypes in real time and is ~100 times smaller than the aggregate BAM size.

To show how Biograph could be applied to clinical whole genomes, we first assess the de novo SV rate for three data sets: a HapMap Trio, an Ashkenazi Jewish (AJ) trio, and a mother-daughter-daughter previously sequenced using three heterogeneous protocols. For each set we determine the quality of SV genotypes by determining Mendelian consistency rates and query speed, comparing standard SV detection methods to the Biograph format. We next show that querying the Biograph format can distinguish between compound events: in the AJ trio we identify a common large insertion that has differentially accrued smaller variants. Finally, we perform lookups of these SVs against batched GBWTs of three sample sets: 1000 Genomes Project light skim WGS and two disease cohorts with deep-coverage WGS.

THE PROMOTER- AND ENHANCER LANDSCAPE OF INFLAMMATORY BOWEL DISEASE

Mette Boyd[1], Jette Bornholdt[1], Morana Vitezic[1], Malte Thodberg[1], Kristoffer Vitting-Seerup[1], Anders Gorm-Pedersen[2], Kerstin Skovgaard[3], Jesper Troelsen [4], Gerhard Rogler[5], Jakob Seidelin[6], Ole Haagen Nielsen[6], Jacob Bjerrum[6], Albin Sandelin[1]

[1]University of Copenhagen, Department of Biology and BRIC, Copenhagen, Denmark, [2]Technical University of Denmark, Department of Systems Biology, Copenhagen, Denmark, [3]Technical University of Denmark, National Veterinary Institute, Copenhagen, Denmark, [4]Roskilde University, Department of Science, Systems and Models, Roskilde, Denmark, [5]UniversitätsSpital Zürich, Klinik für Gastroenterologie und Hepatologie, Zurich, Switzerland, [6]Herlev Hospital, Department of Gastroentorology, Copenhagen, Denmark

Coordinated gene regulation is essential for all aspects of cell biology, including development, differentiation and disease. Characterization of enhancers and promoters in disease has been difficult due to the lack of genome-wide methods suitable for the analysis of small tissue samples. Therefore, we know little about the regulation of genes in disease, and its variation between patients. Related to this, 85% of protein-coding genes show heritable variation in expression due to variance in gene regulation. Thus, localization of promoters and enhancers within patient material is important for disease biology and genetics.
Because promoters and enhancers are transcribed, they can be detected by RNA sequencing. Utilizing this, we have profiled promoter and enhancer usage of the descending colon in 110 patients suffering from inflammatory bowel disease (IBD). To our knowledge, this is the largest study of enhancer activity in a clinical setting ever done.
IBD is a complex group of chronic inflammatory conditions in the gut. Crohn's disease(CD) and Ulcerative Colitis (UC) are the two principal subtypes. Correct treatment depends on accurate sub-type diagnosis, which is challenging and expensive. To this end, we identified a promoter/enhancer set that with high accuracy can distinguish the shared inflammatory response, and UC-or CD-specific profiles. We identified >20.000 transcribed enhancer regions, where subsets are specifically induced in general inflammation or in UC/CD. Many of these inflammation-specific enhancers occur in clusters corresponding to so-called super-enhancers, while enhancers only used in healthy control subjects typically were singletons. Similarly responding enhancers and promoters could be linked by co-expression, and had similar DNA sequence characteristics. IBD-associated SNPs were highly enriched in these regulatory regions, enabling subsequent identification of casual regulatory mutations.

VISUALIZING STRUCTURAL VARIATION AT THE SINGLE CELL LEVEL TO EXPLORE HUMAN GENOME HETEROGENEITY

Ashley D Sanders[1], Mark Hills[1], David Porubsky[2], Victor Guryev[2], Ester Falconer[1], Peter M Lansdorp[1,2,3]

[1]BC Cancer Agency, Terry Fox Laboratory, Vancouver, Canada, [2]University Medical Centre Groningen, European Research Institute for the Biology of Ageing, Groningen, Netherlands, [3]University of British Columbia, Department of Medicine, Vancouver, Canada

Studies of genome heterogeneity and plasticity aim to resolve how genomic features underlie phenotypes and disease susceptibilities. Identifying genomic variants that differ between individuals and cells can help uncover the functional elements that drive specific biological outcomes. For this, single cell studies are paramount, as it becomes increasingly clear that the contribution of rare but functional cellular subpopulations is important for disease prognosis, management and progression. Until now, studying these associations has been challenged by our inability to map structural rearrangements accurately and comprehensively. To overcome this, we employed the template strand sequencing method, Strand-seq, to preserve the structure of individual homologues and visualize genomic rearrangements in single cells. We used this method to rapidly discover, map, and genotype human polymorphisms with unprecedented resolution. This allowed us to explore the distribution and frequency of rearrangements in a heterogeneous cell population, identify several polymorphic domains in complex regions of the genome, and locate rare alleles in the reference assembly. We then extended this analysis to comprehensively map the complete set of inversions in an individual's genome and define their unique inversion profile. We predict characterizing inversion profiles of patients will have important implications for personalized medicine. Finally, we generated a non-redundant, global reference of structural rearrangements in the human genome and better characterized their architectural features. Taken together, we describe a powerful new framework to study structural variation and genomic heterogeneity in single cell samples, whether from individuals for population studies, or tissue types for biomarker discovery.

FREQUENCY, VARIANCE AND POWER: HOW GENETIC MODEL AND DEMOGRAPHY IMPACT ASSOCIATION STUDIES.

Jaleal S Sanjak[1,2]

[1]University of California, Irvine, Department of Ecology and Evolutionary Biology, Irvine, CA, [2]University of California, Irvine, Center for Complex Biological Systems, Irvine, CA

To understand the genetic architecture of complex traits we need theoretical models that make useful predictions that are consistent with empirical observation. Simulations of complex traits are widely used for inference of population parameters and in silico testing of new experimental or analytical methods. Despite this, the approaches in the field of complex trait simulation are very heterogeneous. One common thread is that classic models consider particular mutations ("SNPS") as separate loci. However, we wish to model a genomic region as a functional unit or gene. As such, there are important implications of the structure of the relationship between mutations in the region and it's functional output in a diploid individual. Also, it is well known that demography plays an important role in shaping patterns of DNA sequence variation, but the specifics of how demography interacts with underlying genetic model are unknown. We use forward-time simulation to explore the properties of a co-dominant model and two different recessive models, in constant-sized and recently expanded populations, in the context of heritability estimation and genetic association studies. In particular, we find that the population frequency by effect-size distribution and statistical properties of association studies are both impacted by genetic model. Consequently, when explicitly modeling DNA sequence variation underlying a complex trait, it is critical to differentiate between sites within a functional unit (gene) and those in distinct functional units. Comparing the effect of population growth across multiple genetic models suggests that, perhaps, the genetic model is more important than the demographic model.

UNDERSTANDING THE SEA LAMPREY TRANSCRIPTOME DURING PROGRAMMED GENOME REARRANGEMENT

Jeramiah Smith, Cody Saraceno

University of Kentucky, Biology, Lexington, KY

The sea lamprey undergoes developmentally programmed genome rearrangements (PGR) whereby large fractions (~20%) of the genome are selectively eliminated from cells destined to form the soma during early embryonic development. While these eliminated regions are enriched for repetitive elements, they also contain a substantial number of protein-coding genes. Previous analyses suggested that deleted genes possess gene functions related to the regulation of gene expression, chromatin reorganization and the development/maintenance of stem cells. As such, PGR is conceptualized as a means of differentiating somatic and germline genomes by restricting the full developmental/regulatory potential of the genome to the germline. The genes, regulatory elements, and cellular mechanisms guiding this precise excision of DNA are largely uncharacterized and are the subject of several ongoing analyses. A fundamental resource for such studies is the development of a transcriptome assembly that represents genes that are expressed at the early developmental stages during which PGR occurs. Past efforts to generate a highly contiguous assembly from Illumina sequence data have been hampered by the exceedingly high GC content and low sequence (codon) diversity that generally characterizes lamprey protein-coding sequences. To address these issues we have used single molecule, real-time (SMRT) sequencing to develop full length reference sequences for transcripts that are expressed during PGR.

A large set of consensus reference sequences of individual genes and their isoforms (N=56155) were identified by integrating circular consensus sequencing, resampling, genome assembly and short-read sequence data. This reference transcriptome has permitted in-depth characterization of transcription during programmed genome rearrangement and the identification of full-length transcripts for a large number of genes that are eliminated by PGR. The availability of a high-quality reference transcriptome improves the sensitivity and specificity of RNA-seq experiments aimed at characterizing the dynamic expression of germline-specific genes during early embryogenesis. These full-length sequences are also expected to aid in functional assays aimed at validating the hypothesis that the misexpression of these genes in a somatic context may contribute to the development of genomic diseases (cancer).

GENOMESCOPE: FAST GENOME ANALYSIS FROM UNASSEMBLED SHORT READS

Michael C Schatz[1,2], Greg Vurture[1], Fritz J Sedlazeck[2], Maria Nattestad[1], Charles Underwood[1], Han Fang[1], James Gurtowski[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD

Current developments in de novo assembly technologies have been focused on relatively simple genomes. Even the human genome, with a heterozygosity rate of only ~0.1% and 2n diploid structure, is significantly simpler than many other species, especially plants. However, genomics is rapidly advancing towards sequencing more complex species such as pineapple, sugarcane, or wheat that have much higher rates of heterozygosity (>1% for pineapple), much higher ploidy (8n for sugarcane), and much larger genomes (16Gbp for wheat).

One of the first goals when sequencing a new species is determining the overall characteristics of the genome structure, including the genome size, abundance of repetitive elements, and the rate of heterozygosity. These features are needed to study trends in genome evolution, and can inform the parameters that should be used for the individual assembly steps. They can also serve as an independent quality control during any analysis, such as quantifying the quality of an assembly, or measuring the expected number of heterozygous bases in the genome before mapping any variants.

We have developed an analytical model and web tool GenomeScope that can infer the global properties of a genome from unassembled sequenced data. GenomeScope uses the kmer count distribution, e.g. from Jellyfish, and within seconds produces a report and several informative plots describing the genome properties. We validate the approach on simulated heterozygous genomes, as well as synthetic crosses of related strains of microbial and eukaryotic genomes with known reference genomes. GenomeScope was also applied to study the characteristics of several novel species, including pineapple, pear, the regenerative flatworm Macrostomum lignano, and the Asian sea bass. GenomeScope is available on the web at http://qb.cshl.edu/genomescope

TARGETED PROSPECTIVE SEQUENCING TO IDENTIFY AND INCORPORATE CLINICALLY ACTIONABLE PHARMACOGENOMIC VARIANTS IN ELECTRONIC HEALTH RECORDS AS A MODEL FOR PRECISION INDIVIDUALIZED HEALTH CARE.

Steven E Scherer[1], Xiang Qin[1], Donna Muzny[1], Liewei Wang[2], John L Black[2], Richard Weinshilboum[2], Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Mayo Clinic, Center for Individualized Medicine, Rochester, MN

Collaboratively, the Baylor College of Medicine's Human Genome Sequencing Center and the Mayo Clinic's Center for Individualized Medicine have undertaken a project to sequence up to 10,000 patients from the Mayo Clinic Biobank across a panel of pharmacogenomically important loci. The objective of these efforts is to prospectively incorporate the results in patient electronic health records (EHRs) to inform clinical drug prescribing practices in terms of efficacy and avoidance of adverse events. The study is using a combination of reagents including a capture-sequencing panel of seventy-six targeted genes developed originally as part of the Pharmacogenomic Research Network (PGRN). Specific targets are based on a combination of clinical drug-gene guidelines, published by the PGRN's Clinical Pharmacogenetics Implementation Consortium (CPIC) and community feedback. The targets in previous versions of the capture reagent were pared down to include gene coding regions, the entire CYP2D6 region and SNP targets aimed at characterizing both known and novel variants while keeping costs equal to or below microarray based genotyping approaches. Preliminary data was generated using 500 pilot samples characterized previously as part of the Mayo Clinic's eMERGE Network studies. We developed both data generation and analysis pipelines aimed at identification of genomic variants and star allele haplotypes influencing commonly prescribed drug efficacies and toxicities. Improvements in haplotype calling and clinical decision support are ongoing. Looking to the near future, the study will be tracking outcomes to confirm the value of this approach. This study will provide a large cohort blueprint for implementation of pharmacogenomics in precision individualized health care.

A DETAILED VIEW OF COMPLEX GENOMIC VARIATION IN HUMANS FROM HIGH-QUALITY DE NOVO GENOME ASSEMBLIES OF 50 DANISH PARENT-OFFSPRING TRIOS

Bent Pedersen[1], Jacob M Jensen[2], Siyang Liu[3], Lasse Maretty[4], Jonas A Sibbesen[4], Palle Villesen[2], Laurits Skov[2], Søren Besenbacher[8], The Danish Pan Genome Consortium[1,2,3,4], Simon Rasmussen[1], Anders Børglum[6], Thorkild I Sørensen[7], Ramneek Gupta[1], Wang Jun[5], Hans Eiberg[7], Karsten Kristiansen[8], Søren Brunak[1], Mikkel H Schierup[2]

[1]Technical University of Denmark, CBS, Lyngby, Denmark, [2]Aarhus University, BIRC, Aarhus, Denmark, [3]BGI, Europe, Copenhagen, Denmark, [4]University of Copenhagen, BINF, Copenhagen, Denmark, [5]BGI, Shenzhen, Shenzhen, China, [6]Aarhus University, Biomedicine, Aarhus, Denmark, [7]University of Copenhagen, Medical Genetics, Copenhagen, Denmark, [8]University of Copenhagen, Biology, Copenhagen, Denmark, [9]Aarhus University, MOMAAarhus, Denmark

Most known genetic variation in human genomes has been called from comparison of short reads to the reference genome. A less biased approach is to individually denovo assemble genomes and then call variants by whole genome comparisons. We performed high quality de novo assemblies of 150 individuals (50 parent-offspring trios) of Danish descent, sequenced at 78X coverage with libraries of insert sizes up to 20 kb. Genomes are assembled into non-chimeric scaffolds with N50 of >20Mb, with 90% of single genomes covered by the 100 largest scaffolds up to 120 Mb in length. We develop a new k-mer approach for genotyping a variant call set based on genome alignment and report a large set of high-quality, novel, high frequency, structural variants. We find 416.372 polymorphic insertions above 10 bp in length (75% novel) and 315.052 deletions above 10 bp in length (49% novel) alleviating a bias towards detection insertions when using a reference genome. This is in contrast to SNVs, where only 13% of the 12 million identified are novel. We show that the 4Mb HLA region is fully assembled and we can infer individual haplotypes, producing a reference set of HLA haplotypes with more than 200.000 variants for future use in association mapping. The Y chromosome is also well assembled in many regions, revealing new insight into the structural complexity of this chromosome. With our data it is possible to infer the parent of origin of >50% of the denovo mutations and indels and 80% of denovo SNVs but only 66% of denovo indels are of paternal origin. We provide the first direct evidence that older mothers contribute more new mutations to their children, directly proportional to their age. Finally, we show that with 100 independent individuals, imputation into chip-based data from the Danish population is considerably improved for both single-nucleotide and complex variation.

OFF-CHROMOSOME: UNDERSTANDING AND ACCESSING VARIATION, UPDATES AND UNCERTAINTIES IN THE HUMAN REFERENCE GENOME

Valerie A Schneider[1], Tina Graves-Lindsay[2], Kerstin Howe[3], Paul Flicek[4]

[1]NIH, NCBI, Bethesda, MD, [2]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [3]Wellcome Trust Sanger Institute, Genome Informatics, Hinxton, Cambridge, United Kingdom, [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

While a typical human genome consists only of 23 diploid chromosome pairs, the human reference assembly is a genome model capable of representing additional genetic information in the form of extra-chromosomal sequences. These fall into 3 categories: alternate loci (variation), patches (assembly updates) and unlocalized/unplaced scaffolds (sequence with ambiguous chromosomal localization). Although these scaffolds augment the value of the reference, they are historically under-utilized by researchers performing basic and clinical research due to lack of understanding and resources. Despite significant recent progress in the production of graph structures to represent a universal map of human genome variation, additional development and evaluation is still needed before their widespread adoption across the diverse community of genome consumers. Thus, these collections of reference scaffolds represent the datasets that can be used now to advance our understanding of the human genome. We will discuss characteristics and examples of the 3 types of scaffolds, including criteria and DNA sources for alternate loci representation, recent assembly updates of clinical significance and ongoing efforts and challenges in sequence localization. We will introduce approaches for working with alternate loci and patches with tools that do not have native support for these sequences, as well as those that do. We will further demonstrate how these sequences improve reference-based analyses, not only for variation calling, but for understanding of gene content, haplotype configurations and general genomic features. We will additionally present genome browser resources for visualizing and working with non-chromosomal assembly sequences, and tools for translating coordinates between alternate loci or patch scaffolds and chromosomes. Together, these data will demonstrate the value of all categories of off-chromosomal assembly sequences and the importance of resources for their use to a more complete understanding of human genomic biology.

REGULATORY VARIATION DRIVEN BY TRANSPOSABLE ELEMENTS CONTRIBUTES TO METABOLIC DISEASE

Juan Du[1,2], Amy Leung[1], Candi Trac[1], Aldons J Lusis[3], Rama Natarajan[1,2], Dustin E Schones[1,2]

[1]City of Hope, Department of Diabetes Complications and Metabolism, Duarte, CA, [2]City of Hope, Irell & Manella Graduate School, Duarte, CA, [3]UCLA, Department of Medicine, Los Angeles, CA

Gene-environment interactions are involved in the susceptibility and progression of many types of complex diseases. Despite the importance of such interactions, the most relevant work to date investigating gene-environment interaction in complex diseases has been correlative, and a functional understanding of this interaction is lacking. We are investigating the interaction of environmental and genetic factors through modifications to chromatin. We previously reported that diet-induced obesity (DIO), by means of the environmental influence of a "western" high-fat and high-sucrose (HF/HS) diet, leads to chromatin modifications across the genome in mouse liver tissue, indicating DIO is associated with pervasive changes in transcriptional regulatory programs. These environmentally induced chromatin modifications can furthermore persist even upon the reversal of HF/HS diet to a control diet, indicating these epigenetic modifications are not transient. To begin to investigate the interplay of genetic and environmental factors, we are utilizing the natural genetic variation that exists between different strains of mice from the Hybrid Mouse Diversity Panel (HMDP), which display phenotypic variability in response to HF diet. The mice in this panel have been densely genotyped and have been profiled for a variety of metabolic markers, including insulin resistance and liver metabolites. We have now profiled chromatin accessibility genome-wide in liver tissue for various strains of mice from the HMDP. According to our results, a large proportion of sites displaying chromatin variation across strains harbor transposable elements (TEs), indicating that transposable elements are major drivers of the genetic component of regulatory diversity across the strains. Given that active TEs can lead to deleterious effects, mouse and human genomes have evolved tightly controlled mechanisms to silence their activity. It has been shown that TEs can affect local gene expression by acting as regulatory elements and recent results have identified TE dysregulation as a common feature in a number of diseases, including liver disease, for which obesity is a clear risk factor. We are investigating the possible mechanisms in which TE dysregulation is involved in liver disease induced by obesity. According to our results, DIO can lead to epigenetic modifications at TEs that contribute to long-term risk for metabolic disease and associated complications.

THE IMPACT OF GENOME STRUCTURAL VARIATION ON GENE EXPRESSION IN HUMANS

Alexandra J Scott[1], Colby Chiang[1], The Genotype-Tissue Expression (GTEx) Project Consortium[2], Stephen B Montgomery[3,4], Alexis Battle[5], Don F Conrad[6], Ira M Hall[1,7]

[1]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [2]The Genotype-Tissue Expression (GTEx) Project Consortium, -, -, MA, [3]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [4]Stanford University School of Medicine, Department of Pathology, Stanford, CA, [5]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [6]Washington University School of Medicine, Department of Genetics, St. Louis, MO, [7]Washington University School of Medicine, Department of Medicine, St. Louis, MO

Structural variation (SV), including copy number variants (CNVs), balanced rearrangements, and mobile element insertions, is an important source of human genetic diversity. Using short-read DNA sequencing technologies, we can detect 5,000-10,000 SVs in the typical human genome. However, our understanding of the relationship between SV and phenotypic variation remains limited. The GTEx project presents an unprecedented opportunity to address this question due to the availability of deep (>30x) whole genome sequencing (WGS) and multi-tissue RNA-seq data for hundreds of individuals.

Here, we describe our work aimed at measuring the contribution of structural variation to human gene expression variation in an initial set of 147 GTEx individuals. We first generated a comprehensive SV map using a combination of methods including paired-end and split-read mapping using the WashU pipeline (LUMPY + SVTools), as well as read-depth based CNV detection using GenomeSTRiP. This resulted in a map of 45,968 SVs with copy number and genotype annotations, of which 24,157 are high confidence SVs detected by multiple independent alignment signals. We then used this dataset to perform SV-only eQTL mapping as well as a joint mapping of SV, SNP, and indel eQTLs in 13 tissues, resulting in 24,801 unique eQTLs affecting 10,101 distinct eGenes. Based on LD structure and heritability partitioning, we estimate that SVs are the causal variant at 2.7-5% of eQTLs, which represents a 18-33 fold enrichment relative to their abundance in the genome (0.15% of total variants). This is an order of magnitude higher than a prior estimate based on low coverage WGS and LCL gene expression. We further present several exciting results regarding (1) the magnitude, directionality and allele specificity of expression effects with respect to SV class and overlap with known regulatory elements, (2) the potential contribution of SV to previously reported GWAS hits, and (3) the notable abundance (relative to SNPs and indels) of rare high impact SVs that cause extreme gene expression outliers identified in a single GTEx individual.

← Back

A HOT L1 RETROTRANSPOSON EVADES SOMATIC REPRESSION AND INITIATES HUMAN COLORECTAL CANCER

Emma C Scott*[1,2], Eugene J Gardner*[1,2], Ashiq Masood^[2,3,4], Nelson T Chuang[1,2,5], Scott E Devine[1,2,3,4]

[1]University of Maryland Baltimore, Graduate Program in Molecular Medicine, Baltimore, MD, [2]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [3]University of Maryland School of Medicine, Greenebaum Cancer Center, Baltimore, MD, [4]University of Maryland School of Medicine, Department of Medicine, Baltimore, MD, [5]University of Maryland School of Medicine, Division of Gastroenterology, Department of Medicine, Baltimore, MD

Although human LINE-1 (L1) elements are actively mobilized in many cancers, a role for somatic L1 retrotransposition in tumor initiation has not been conclusively demonstrated. Here, we identify a novel somatic L1 insertion in the *APC* tumor suppressor gene that provided us with a unique opportunity to determine whether such insertions can actually initiate colorectal cancer (CRC), and if so, how this might occur. Our data support a model whereby a hot L1 source element on chromosome 17 of the patient's genome evaded somatic repression in normal colon tissues and thereby initiated CRC by mutating the *APC* gene. This insertion worked together with a point mutation in the second *APC* allele to initiate tumorigenesis through the classic two-hit CRC pathway. We also show that L1 source profiles vary considerably depending on the ancestry of an individual, and that population-specific hot L1 elements represent a novel form of cancer risk.

This work was funded by the following grants: T32 CA154274 (ECS), T32 DK067872 (NTC), R01CA166661 (SED), and R01HG002898 (SED).

*Authors contributed equally.
^Current address: Siteman Cancer Center, Washington University School of Medicine, St. Louis,
St. Louis, MO 63110 U.S.A.

← Back

TEASER: COMPREHENSIVE READ MAPPER BENCHMARKING IN 20 MINUTES FOR GENOMES, TRANSCRIPTOMES, METHYLOMES AND METAGENOMES

Moritz G Smolka[1], Florian Breitwieser[2], Steven L Salzberg[2,3,4], Arndt von Haeseler[1], Michael C Schatz[3], Fritz J Sedlazeck[3]

[1]Center for Integrative Bioinformatics Vienna, MFPL, Vienna, Austria, [2]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, MD, [3]Department of Computer Science, Johns Hopkins University, MD, [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

In recent years over 100 read mappers have been published to analyze high throughput sequencing data, each of which is optimized for different assays or requirements. The large number of potential mappers and the even larger number of possible parameter settings make it challenging to choose the most appropriate mapper for a given experiment. Consequently most users rely on default, unoptimized, parameters for one of a few popular methods, even when this choice performs very poorly compared to an optimized approach. This may introduce substantial biases in subsequent analyses, including reduced coverage, false determination of allele-specific expression, mis-identification of infectious agents, or other artifacts.

We previously reported Teaser, a benchmarking tool for DNA-seq mappers that has since been used by a number of large studies. Here we extend Teaser to benchmark the mapping of bisulfite, RNA-, and metagenomic sequencing data. Teaser can be applied to any number of mapping methods, and even automatically investigate their parameter settings. The benchmarks are highly customizable, so that read length, SNP rate and other key parameters can be adapted to the experiment at hand. After launching, a detailed assay-specific report is generated for each mapper configuration, often in less than 20 minutes even for mammalian-sized genomes on a desktop computer. This empowers researchers to make an informed decision on the most suitable method for their needs and allows them to fully utilize their data set.

Using Teaser, we investigated how well RNA-Seq mappers (e.g. HISAT, STAR, etc.) and quantification methods (Kallisto, Sailfish) perform on a variety of genomes and analysis tasks. Here, Teaser provides insights into the accuracy of read alignments spanning multiple exons that enables isoform-level quantification and detection of novel isoforms. Furthermore, we used Teaser to investigate the ability of metagenomics methods (e.g. Kraken, CLARK, etc.) to obtain correct predictions of different taxonomic levels (e.g. genus or species) given different read lengths and sequencing error rates, and including strains not present in the reference databases.

Teaser is available as a webserver (teaser.cibiv.univie.ac.at) or as a standalone package (github.com/Cibiv/Teaser).

ROLE OF ALTERNATIVE SPLICING IN RECOVERY FROM TRAUMATIC BRAIN INJURY

Arko Sen[1,2], Wen Qu[1], Jane Brewer[2], Douglas Ruden[2,3]

[1]Wayne State University, Department of Pharmacology, Detroit, MI, [2]Wayne State University, Institute of environmental health sciences (IEHS), Detroit, MI, [3]Wayne State University, OB/Gyn, Detroit, MI

Traumatic brain injury (TBI) can cause non-reversible pathological alternation of neuronal pathways. This may lead to cognitive dysfunctions, depression, and even increased susceptibility to life threatening diseases, such as epilepsy. To investigate the underlying genetic and molecular basis of TBI, Wasserman and colleagues developed an inexpensive and reproducible model for TBI in Drosophila melanogaster (Fruit Fly); High Impact Trauma (HIT) model. Using modified version of this model, we subjected Drosophila melanogaster to mild closed head trauma. We observed that the survivors of TBI (i.e. survive past 24 hours) showed a reduction in lifespan dependent upon the number of strikes. We collected fly heads at 2 time points; 4 hours post-TBI and 24 hours post-TBI and performed a detailed analysis of the brain transcriptome. We observed a significant up-regulation of alternative splicing (AS) events 24 hours post-TBI. Together, the data suggested that the AS changes in the brain might be critical in regulating survival post-TBI. Characterization of the AS events showed selective retention of long introns (> 81bps). We hypothesize that retention of long introns in Drosophila brain is closely regulated by the interaction of the elongating RNA polymerase with local histone modifications. To further understand this association we performed a meta-analysis of ChIP-Sequencing data (from modENCODE) for several histone modifications. We found that H3K36me3 marks alternative exons differently from constitutive exons. Furthermore KD of the corresponding histone demethylase (dKDM4a) also produced larger number of intron retention events compared to KD of other histone demethylases. This data provide preliminary evidence of association between local histone modification and alternative splicing. Further investigations studying the potential regulators of AS events especially in context of TBI, are currently underway in our lab.

FAST, SCALABLE AND ACCURATE DIFFERENTIAL EXPRESSION ANALYSIS OF SINGLE CELLS: APPLICATION TO MOUSE BRAIN AND CIRCULATING TUMOR CELLS

Debarka Sengupta [1], Say Li Kong[2], Nirmala Arul Rayan[1], An Yi Joyce Tai[2], Gek Liang Michelle Lim[1], Kok Hao Edwin Lim[2], Andrew Wu[3], Tingyuan Tu[3], Man Chun Leong[3], YiFang Lee[3], Ali Asgar Bhagat[3], Darren Wan Teck Lim[4], Daniel Shao Weng Tan[4,5], Iain Bee Huat Tan[4,5], Axel Hillmer[2], Bing Lim[6], Shyam Prabhakar[1]

[1]Genome Institute of Singapore, Computational and Systems Biology, Singapore, [2]Genome Institute of Singapore, Cancer Therapeutics & Stratified Oncology, Singapore, [3]Clearbridge Biomedics, Singapore, [4]National Cancer Centre, Singapore, Division of Medical Oncology, Singapore, [5]Genome Institute of Singapore, Singapore, [6]Genome Institute of Singapore, Cancer Stem Cell Biology, Singapore

Single-cell transcriptomes are distorted by technical biases such as RNA degradation during cell isolation, variable reagent amounts, cellular debris and PCR amplification bias. Moreover, due to limited molecular abundances (typically: 1-10 transcripts per gene), the data are inherently noisy. Scalability is also a challenge – technologies such as Drop-seq require algorithms that can quickly compare thousands of transcriptomes. Thus, the two most basic components of transcriptome analysis, normalization and detection of differentially expressed (DE) genes, are uniquely challenging on scRNA-seq data. Here we introduce pseudocounted quantile (pQ) normalization, a technique specifically tailored for single cell data that halves technical variability and substantially improves the quality of downstream analyses. We also introduce NODES, a nonparametric method for detecting differentially expressed genes that readily scales to >1,000 cells and is both more accurate and ~10 times faster than existing parametric approaches. More generally, we propose that nonparametric statistics are ideally suited for single cell analysis, since they dramatically reduce computational complexity (speed, scalability), require no prior knowledge about the properties of the data (generality), and have greater statistical power than parametric methods when the distributional assumptions of the latter are violated. We demonstrate the utility of these algorithms by applying them to the task of unsupervised, marker-free characterization of circulating tumor cells, which comprise a rare subpopulation of peripheral blood cells from colorectal cancer patients. We also use the newly developed methods to robustly identify cell types in multiple regions of embryonic and adult mouse brain.

GENETIC VARIATION IN MHC PROTEINS IS ASSOCIATED WITH T-CELL RECEPTOR EXPRESSION BIASES

Eilon Sharon[1,2], Leah V Sibener[3,4,5], Alexis Battle[6], Hunter B Fraser[2], Christopher Garcia[3,4,7], Jonathan K Pritchard[1,2,7]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]Stanford University, Department of Biology, Stanford, CA, [3]Stanford University, Department of Molecular Physiology, Stanford, CA, [4]Stanford University, Department of Structural Biology, Stanford, CA, [5]Stanford University, Department of Immunology, Stanford, CA, [6]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [7]Howard Hughes Medical Institute, Stanford University, Stanford, CA

In each individual, a highly diverse T-cell receptor (TCR) repertoire interacts with peptides presented by major histocompatibility complex (MHC) molecules. Despite extensive research, it remains controversial whether the germline-encoded TCR-MHC contacts have co-evolved to promote MHC restriction and if so, whether there are differences in the compatibility of TCR V-genes to different MHC alleles. Here we applied a genetic approach, eQTL mapping, to test for association between genetic variation and TCR V-gene usage in a large human cohort. We show that, in agreement with the co-evolution model, genetic variation in MHC residues drives usage biases in TCR V-genes. Our analysis highlights several MHC residues as determinants of TCR V-gene usage biases; remarkably, many of these residues are in direct contact or spatially proximal to either the TCR or the presented peptide. Interestingly, this is the first example of trans-eQTLs that are mediated by protein-protein interaction. Our results provide the first genetic evidence that MHC variants, several of which are linked to autoimmune diseases, can directly affect the function of TCR-MHC interaction and bias usage profiles of TCR V-genes.

TREE CONSISTENT PBWT AND THEIR APPLICATION TO RECONSTRUCTING ANCESTRAL RECOMBINATION GRAPHS AND POPULATION STRUCTURE INFERENCE

Vladimir Shchur[1], Niko Välimäki[2], Richard Durbin[1]

[1]Wellcome Trust Sanger Institute, Durbin Faculty, Cambridge, United Kingdom, [2]University of Helsinki, Department of Medical and Clinical Genetics, Helsinki, Finland

We present an efficient way to represent and process ancestral recombination graphs (ARGs). An ARG is a genealogical network which contains the complete genealogical information of a sample of individuals, providing a local coalescent tree at each position. Recombinations appear as prune-and-regraft operations that encode how neighbouring local trees are related. A proper enumeration of leaves of local trees which we call a *planar ordering*, together with a distance function between leaves, provides a data structure for efficient and scalable storage, processing and inference of possible ARGs of a sample.

The positional Burrows-Wheeler transform PBWT is a representation of a set of haplotypes that supports very efficient data compression and fast haplotype matching. We use a PBWT-inspired data structure, which we call a tree consistent PBWT, or shortly tcPBWT, which has a natural and tractable connection with local coalescent trees. tcPBWT shows some improvement in the compression rate compared to PBWT, which suggests that it has better consistency with the genealogy giving rise to the haplotypes. tcPBWT includes a great amount of planar ordering structure, so it can be used for generating possible ARG topologies. We also introduce a method of fast estimation of the coalescent times of the nodes of an inferred ARG. The shapes of inferred local trees are strongly influenced by the population structure of a sample. Our methods are scalable at least to thousands of individuals and we suggest that analysis of the inferred distribution of coalescent times in the ARG can be used to help understand the population structure of such large samples. The resulting data structures provide a very compact method for storing haplotype data as a set of highly correlated trees, which can also potentially enable other analyses.

A REFERENCE-AGNOSTIC AND RAPIDLY QUERYABLE NGS READ DATA FORMAT ALLOWS FOR FLEXIBLE ANALYSIS AT SCALE

Sri N Shekar[1], William J Salerno[2], Adam English[2], Adina Mangubat[1], Jeremy Bruestle[1], Eric Boerwinkle[2,3], Richard A Gibbs[2]

[1]Spiral Genetics, Bioinformatics, Seattle, WA, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [3]University of Texas Health Science Center, Human Genetics Center, Houston, TX

In identifying the complement of genetic variants that are associated with complex disease, larger sample sizes increase power. Studies such as the Alzheimer's Disease Sequencing Project and the CHARGE Consortium where samples are collected from a range of centers show heterogeneous data, requiring informatics that can additively scale to thousands of samples and analytics that go beyond identifying small variants in NGS data. At scale, the challenge of evaluating SNPs, indels and SVs becomes the "N+1" problem of incrementally adding samples without having to perpetually reevaluate petabytes of population read data stored in BAM files.

The Biograph Analysis Format (BAF) is a method of indexing NGS data that extends the Burrows Wheeler Transform to allow for multiple paths, effectively creating a read overlap graph of the data. A BAF of HiSeq X 30x WGS data is 8.3 Gb, 95% smaller than the original BAM. Generated from the BAM in 14 hours, the BAF can be queried up to 200,000 times a second. Multiple BAFs can be combined, resulting in a file size of approximately 3GB per individual. With multi-individual BAFs, query time grows less than linearly with the number of individuals.

For example, with 30,000 putative SV sites to be queried, SV-typing these sites across 10,000 HiSeq X WGS samples in BioGraph Analysis Format would require less than 30 TB of storage (for all the read data), 16 CPU hours, and 10 minutes (using 100 machines).

Additionally, the data are reference-agnostic, so variants can be called against any reference or against the read graph of any other set of individuals, dramatically reducing the time for data harmonization. Further, information is divided such that the "read overlap graph" created from all the individuals is separate from the information indicating that path through the graph for each individual. This allows a search for a particular variation of interest directly from the read data remotely and rapidly, without the opportunity to reveal the exact individual(s) from that the variant originates.

Because the data are essentially a read overlap graph, it is possible to accurately characterize SVs by traversing the graph from a particular location or search for a particular sequence associated with the SV. So, fast querying of small files with reasonable compute requirements provides an N+1 solution for SNPs, indels and SVs. We describe how the BAF API allows users to construct specific queries for a range of applications.

ASSESSMENT OF THE HUMAN eQTLSCAPE BY STANDARDIZED RE-ANALYSIS OF OVER 50 eQTL DATASETS

Sushila A Shenoy[1], Ronald G Crystal[1], Jason G Mezey[1,2]

[1]Weill Cornell Medicine, Department of Genetic Medicine, New York, NY, [2]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY

We report the first standardized analysis of the expression Quantitative Trait Loci (eQTL) landscape in humans (the "eQTLscape") by re-analysis of all publicly available eQTL datasets. Our analysis includes all published studies that have released both genome-wide genotype and gene expression data, providing an almost comprehensive coverage of the studies responsible for the current genome-wide picture of eQTL in humans. Our analysis places each eQTL on a continuum of presence or absence across data sets, ranging from the case of ERAP2, which has a *cis*-eQTL that replicates in every data set, to eQTL that are extremely well supported in only one of over 50 distinct eQTL data sets included in our study.

In a comparison of genes with replicated *cis*-eQTL versus genes with no *cis*-eQTL, we found that genes with *cis*-eQTL were significantly more likely to be associated with an OMIM disease or disorder but less likely to be annotated with GO terms for DNA binding and transcription regulation. We also find that for highly replicating *cis*-eQTL that have strong signals identified in studies that make use of high-coverage, whole-genome sequencing data, we can often resolve the location of the causal polymorphisms to a surprisingly small region and often to a small candidate set. This resolution often points to a single polymorphism present in a known or functionally predicted enhancer being responsible for given *cis*-eQTL, where replication of the *cis*-eQTL is likely due to whether the enhancer is in use given the sampling conditions of the eQTL study.

In contrast to the many *trans*-eQTL reported in most eQTL studies, we find that once experimental and statistical artifacts are accounted for, there are almost no strongly supported *trans*-eQTL that replicate. This suggests that *trans*-eQTL may not have been reliably detected within current eQTL studies. Our re-analysis also points to a conclusion distinct from the paradigm of discussed throughout hundreds of published studies that claim existence of tissue-specific eQTL, since we find that given the currently available eQTL data, it is impossible to make a claim that tissue-specificity is the crucial factor for why an eQTL is identified in a given study. This is particularly true given the non-standardization of experiment parameters, sample collection protocol, and data platform. Our results point to the need for more precise collection of cell types and for coordinated replication efforts, not only across independent laboratories but also across eQTL consortia, to provide a clearer picture of the conditional genome-wide landscape of human eQTL.

← Back

INTEGRATING GENETICS AND EPIGENETICS DATA TO PRIORITIZE NON-CODING RISK LOCI AND THE GENES PERTURBED IN AUTOIMMUNE DISEASES

Parisa Shooshtari[1,2], Chris Cotsapas[1,2,3]

[1]Yale University, Neurology, New Haven, CT, [2]Broad Institute of MIT-Harvard, Medical and Population Genetics, Cambridge, MA, [3]Yale University, Genetics, New Haven, CT

Genome-wide association studies have identified thousands of loci mediating risk to common, complex diseases, and we and others have shown that the majority of these effects map to regulatory DNA. These bulk enrichments support a model where disease risk is driven by alterations to gene regulation. However, there is not a systematic way yet to identify specific regulators perturbed in the complex diseases and the genes affected in this regulatory process. We have addressed this problem by designing an analytical framework that integrates genetics associations data, epigenomics data and gene expression data.

We first construct a map of regulator:gene relationships across the genome using matched expression and regulatory region assays – here, paired expression and DNase I Hypersensitivity site (DHS) data from 22 cell types profiled by the Roadmap Epigenome Project. Uniquely, we identify reliably detected regulatory regions across samples using a clustering approach. By overlaying the variants forming the credible interval (CI) in a GWAS risk locus onto this genome-wide landscape of gene regulation, we can identify specific regulators driving risk to disease. We can quantify the risk attributable to all hypersensitive sites in each disease locus, and thus evaluate the evidence that a specific risk association is mediated by noncoding effects. By incorporating the correlation between DHS and genes in the region, we can then compute the risk attributable to each gene, and thus prioritize them.

We applied this framework to association data for 9 autoimmune and inflammatory diseases and found that 301/369 (82%) of risk loci have at least one CI SNP on a local DHS. In 132/301 (44%) of these loci, at least 25% of the posterior probability of association can be attributed to SNPs on DHSs, making them reasonable candidate loci for regulatory effects. With this approach we are able to reduce the candidate DHS mediating risk from a median of 848.5 to 4.5 per locus, and identify genes controlled by those DHS in 104/132 (79%) of cases. The latter include known pathogenic genes with altered regulation: CD40 and CD58 in multiple sclerosis, CLECL1 in type 1 diabetes, and CARD6 in inflammatory bowel disease. 92/104 (88%) of these genes are not closest to the most associated variant in the region, suggesting that many risk-mediating DHS act at substantial distances. We also find strong enrichment of allelic imbalance in chromatin availability for risk-mediating DHS, suggesting that causal variants alter the activity of those regulatory regions. We are currently comparing regulatory perturbation across multiple diseases in the shared loci and will present the overall results.

DETECTING INTROGRESSED ARCHAIC HAPLOTYPES IN OCEANIC POPULATION GENOME SEQUENCES

Laurits Skov, Anders Bergstrom, Yali Xue, Chris Tyler-Smith, Richard Durbin

Wellcome Trust Sanger Institute, Computational genetics, Cambridge, United Kingdom

Introgression of archaic haplotypes into human populations is an already known phenomenon, with some haplotypes even providing a selective advantage such as adaptation to living in high altitudes or haplotypes carrying alleles of genes involved in the immune-system.
Most published methods for identification of archaic haplotypes rely on ancient DNA samples from the archaic population, to compare the modern samples with. Here we present a method for identifying regions of modern whole genome sequences that have been introgressed into a subset of modern humans from an ancestor with a long history of separation from the modern human lineage, and apply it to Oceanian genomes. The method takes advantage of the fact that introgressed regions show different patterns of LD and a high density of SNPs private to the population that received the introgression. For the Oceanian samples, we look for regions that have a clear excess of variants not seen in any 1000 Genomes Project sample, in order to identify regions introgressed from archaic populations other than Neanderthals, after the separation of Oceanians from other modern Asian ancestral populations.
We identify tens of Mb of potentially private introgressed sequence for each of the individuals in the study. The regions will be compared to regions found by other methods traditionally used for identifying introgressed regions, to known archaic sequences from ancient DNA, to each other to understand their diversity, and to other modern human sequences to estimate their original separation time.

A DEEP EVOLUTIONARY PERSPECTIVE ON VERTEBRATE GENOME BIOLOGY

Jeramiah J Smith

University of Kentucky, Biology, Lexington, KY

The lamprey genome provides unique insights into both the deep evolutionary history of vertebrate genomes and the maintenance of genome structure/integrity over development. The lamprey lineage diverged from all other vertebrates approximately 500 million years ago. As such, comparisons between lamprey and other vertebrates permit reconstruction of ancient duplication and rearrangement events that defined the fundamental architecture and gene content of all extant vertebrate genomes. Lamprey also undergoes programmatic changes in genome structure that result in the physical elimination of ~20% of its genomic DNA (~0.5Gb from a ~2 Gb genome) from all somatic cell lineages during early embryonic development. The development of a high quality genome assembly for lamprey is critical for understanding ancient and contemporary forces that shape vertebrate genomes, however, the lamprey genome presents significant challenges in this regard. For example, the lamprey genome possesses large numbers of highly identical repetitive elements (a frequency ~10x that of the human genome), is highly polymorphic (~10 polymorphisms per kilobase) and is characterized by extremely GC-rich coding regions.

Here we outline recent progress in assembly and analysis of the lamprey germline genome and progress in the development of methods for characterizing the cellular events that mediate DNA elimination. We have integrated information from several sampling approaches and sequencing technologies to achieve a highly contiguous assembly of the lamprey genome (including: Illumina fragments/mate pairs, 20X coverage in Pacific Biosciences reads, dense meiotic maps and optical mapping data). This genome assembly has dramatically improved our ability to dissect the molecular basis and genetic outcomes of programmed genome rearrangements (PGRs), and has improved our understanding of the tempo and mode of large-scale duplications and translocations within the ancestral vertebrate lineage. Analysis of the germline genome identifies several genes that are expressed in germline but physically eliminated from all somatic tissues. These eliminated genes correspond to several known oncogenes and appear to identify several other novel oncogene candidates. Complementing this assembly, the development of approaches for *in situ* analysis of 3D preserved cells has revealed that PGR unfolds through a series of dramatic cellular events that involve the programmatic alteration of several fundamental mechanisms of genome maintenance, including: alignment of chromosomes at metaphase, chromatid cohesion, separation and segregation and nuclear envelope formation.

RASCAF: GENOME ASSEMBLY SCAFFOLDING WITH RNA-SEQ DATA

Li Song[1,2], Dhruv Shankar [3], Liliana Florea[1,2]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [3]University of North Carolina, Biomedical Engineering, Chapel Hill, NC

Low sequencing costs have accelerated the pace of sequencing new genomes, making it possible for groups and even individual investigators to sequence the genome of the species they are studying. Two critical steps in every genome sequencing project are genome assembly and gene annotation. While abundant, short reads make assembly from next generation sequencing data difficult and lead to fragmented genomes, which in turn complicates gene annotation.

To enhance both assembly and annotation, we propose to use the vast resources of RNA-seq data generated by sequencing projects or already available in public databases to improve the completeness and contiguity of the assembled sequence. We developed Rascaf (RnA-seq SCAFfolder), a fast and efficient tool that leverages the long-range contiguity information from intron-spanning RNA-seq read pairs to detect new contig connections and improve the assembly, in particularly in the gene regions. Rascaf clusters overlapping RNA-seq alignments to determine gene blocks on each contig. It then builds a gene block graph by connecting blocks that are spanned by the two reads in a pair. Lastly, it chooses a path in the graph representing the gene, which is then used as guide for scaffolding.

When tested on both simulated and real data, Rascaf was both more sensitive and more precise than existing tools, namely L_RNA_scaffolder and AGOUTI. When applied to several draft Fragaria genomes, it detects 5,000-10,000 new contig connections, most of which could be verified in silico by blast searches against protein and cDNA databases. Rascaf is fast, has a small memory footprint and is easy-to-use, and can be readily incorporated into genome assembly projects to help improve an assembly and its gene annotations simultaneously. The software is available free of charge from: https://github.com/mourisl/Rascaf . Funding: NSF IOS-1339134.

READ CLOUDS REVEAL EVOLUTION OF STRUCTURAL VARIATION IN CANCER

Noah Spies[1,2,3], Ziming Weng[1,2], Justin M Zook[3], Robert B West[2], Serafim Batzoglou[4], Marc Salit[3], Arend Sidow[1,2]

[1]Stanford University, Dept of Genetics, Stanford, CA, [2]Stanford University, Dept of Pathology, Stanford, CA, [3]National Institute of Standards and Technology, Genome Scale Measurements Group, Stanford, CA, [4]Stanford University, Dept of Computer Science, Stanford, CA

Background
Structural variants, particularly distant translocations, are difficult to identify despite their fundamental importance in cancer and other diseases. Current short-read genomic approaches suffer from high rates of false positives because of the massive multiple-testing problem, and cannot detect variants in repetitive regions of the genome.
Data
The 10X Genomics platform generates barcoded short-reads from large genomic DNA fragments, which can then be clustered in silico to generate read clouds identifying the original large DNA fragments. We size-selected large (50-100kb) genomic DNA fragments from 7 spatially distinct tumor samples from a single sarcoma, as well as matched normal tissue, then applied the 10X platform to generate read clouds.
Method
We have implemented new methods to identify structural variants from these read cloud data. Our new methods dramatically improve specificity, sensitivity, and accuracy by using the patterns of dropoff in observed long fragment density at the structural variant breakpoints.
Results
We show 5-fold improvement in accuracy in structural variant detection over short-read-only approaches at similar sequence coverage. We identify over 400 large somatic structural variants, including translocations, within our sarcoma samples. About 80% of these events show no read support in a single 40x short-read library. Our improved methods are able to identify breakpoints within an average 100 bp of the correct positions, despite using fragments that average over 50kb in size. For a large fraction of these events, phasing information supports only a single haplotype, as expected for somatic events, further improving our confidence in these events.

We identify structural variants that differ between sectors of the sarcoma, although most somatic structural variants (and single-nucleotide variants) are shared within all tumor samples. This result demonstrates that even very large (in this case, >20cm in length) tumors need not show substantial subclonal diversity, and that rather a series of extreme genome rearrangements ("chromothripsis") occurred early in tumor development.

ARRAYED SYNTHESIS OF CUSTOM SINGLE GUIDE RNA LIBRARIES FOR CRISPR-CAS9 GENE EDITING

Benjamin Steyer*[1], Seyyed Alireza Aghayeemeibody*[1,2], José Rodríguez-Martínez[3], Aseem Ansari[3], Randolph Ashton[1,2,4], Krishanu Saha[1,4]

[1]University of Wisconsin-Madison, Wisconsin Institute for Discovery, Madison, WI, [2]University of Wisconsin-Madison, Department of Materials Science and Engineering, Madison, WI, [3]University of Wisconsin-Madison, Department of Biochemistry, Madison, WI, [4]University of Wisconsin-Madison, Department of Biomedical Engineering, Madison, WI

The increasing speed and decreasing cost of DNA sequencing has lead to the identification of thousands of human genetic variants associated with disease. However, there is a lag in testing the functionality of these genetic variants within human cells for a variety of applications in disease modeling, drug discovery, toxicology and regenerative medicine. Here, we describe a transformative technology that utilizes light-directed, maskless array synthesis of oligonucleotides to generate custom libraries of single guide RNAs (sgRNAs) and donor DNA templates for CRISPR-Cas9 gene editing with high complexity. Our results characterize the range of library complexity as well as the limits of sequence fidelity and biological activity of sgRNAs and donor DNA templates generated using our technology. Patient-specific stem cells and other human cells edited using this technology could be assembled into novel libraries of cells with defined genetic variants and diversity. These libraries could be used by research communities in medicine and biology to test hypotheses with new molecular precision. Further, such precision would enable more efficient screening of novel therapeutics with human model systems of disease.

* These authors contributed equally to this work

THE ENCODE ANALYSIS PIPELINES: REPEATABLE AND SHAREABLE ANALYSIS TOOLS FOR CHIP-SEQ, RNA-SEQ, DNASE-SEQ, AND WHOLE-GENOME BISULFITE EXPERIMENTS

J Seth Strattan[1], Timothy R Dreszer[1], Ben C Hitz[1], Esther T Chan[1], Jean M Davidson[1], Idan Gabdank[1], Laurence D Rowe[1], Cricket A Sloan[1], Forrest Tanaka[1], Zhiping Weng[2], Anshul Kundaje[1], J Michael Cherry[1]

[1]Stanford University School of Medicine, Genetics, Palo Alto, CA, [2]University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA

The comparability of experiments often depends on comparable experiment protocols. This applies equally to the physical manipulation of samples and to the analytical methods that transform raw data to results. Fourteen ENCODE labs have contributed over 4000 replicated ChIP-seq, RNA-seq, DNase-seq, and whole-genome bisulfite experiments on nearly 200 cell types. To ensure that the results from these experiments can be compared, the ENCODE Data Analysis Center (DAC) have specified common data processing protocols for ChIP-seq, RNA-seq, DNase-seq, and whole-genome bisulfite experiments. The Data Coordination Center (DCC) is implementing these specifications as processing pipelines, deploying these pipelines to a cloud-based platform, processing all ENCODE ChIP-seq datasets, and making the pipelines available for anyone to use. The results of these analyses, and metadata describing them, are distributed through the ENCODE Portal and illustrate general methods of accessing and interpreting ENCODE data. The cloud-based deployment of the ENCODE analysis pipelines illustrates how the DCC is creating transparent and reusable analysis tools for ENCODE data and for any primary data from experiments performed with similar protocols. The ENCODE Portal is https://www.encodeproject.org/. The DCC codebase is freely available at https://github.com/ENCODE-DCC/.

NEURODEVELOPMENTAL GENE EXPRESSION PROFILING IN HETEROZYGOUS *CHD8* MICE REVEALS PATHWAYS DRIVING MACROCEPHALY AND DEVELOPMENTAL DISORDERS.

Linda Su-Feher[1], Andrea S Gompers[1], Jacob Ellegood[2], Nycole A Copping[3], Iva Zdilar[1], Michael C Pride[3], Tyler Stradleigh[1], Deana Li[3], Christine Nordahl[3], David Amaral[3], Axel Visel[4], Len A Pennacchio[4], Diane Dickel[4], Jacqueline N Crawley[3], Jason P Lerch[2], Konstantinos Zarbalis[5], Jill L Silverman[3], Alex S Nord[1]

[1]University of California, Davis, Center for Neuroscience, Department of Neurobiology, Physiology and Behavior, College of Biological Sciences, Davis, CA, [2]The Hospital for Sick Children, Mouse Imaging Centre, Toronto, Canada, [3]University of California, Davis, MIND Institute, School of Medicine, Davis, CA, [4]Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division, Berkeley, CA, [5]Shriners Hospitals for Children, Institute for Pediatric Regenerative Medicine, Sacramento, CA

Genetic factors play a critical role in the risk and development of autism spectrum disorders (ASDs). While ASDs exhibit high genetic complexity, a number of ASD sequencing studies have identified mutations in the high-confidence risk gene *CHD8*, a chromatin remodeling factor. Genomic binding analyses in parallel with knockdown experiments in cell culture models have identified CHD8 as a potential master regulator of ASD-relevant pathways. Here, we report on a heterozygous knockout mouse model of *Chd8*, generated by CRISPR/Cas9-mediated frameshift deletion. Consistent with previous models, our $Chd8^{-/-}$ knockouts are early embryonic lethal. $Chd8^{+/-}$ mice exhibit increased brain volume as well as behavioral deficits consistent with phenotypic characteristics of human patients with *CHD8* mutations. We performed RNA-sequencing on forebrain of heterozygous and wild-type mice across development at embryonic days (e)12.5, e14.5, e17.5, and postnatal days (P)0 and P56 in order to profile the effects of *Chd8* loss on neural development. We observed large-scale gene expression changes in $Chd8^{+/-}$ mice which include changes in genes such as *Bcl11a*, *Rims2*, and *Kdm5b* that have been implicated in other ASD cases. Genes up-regulated in $Chd8^{+/-}$ mice include cell cycle, growth, and proliferation genes. Down-regulated genes are enriched across a number of pathways, the strongest of which is a large cluster of genes involved in neuron-specific mRNA processing and splicing networks. These results suggest that haploinsufficiency of *Chd8* disrupts developmental cell state, leading to reduced induction of alternative splicing pathways that regulate transition from proliferation to differentiation in neurons. By determining the transcriptional pathways affected by *Chd8* mutations, we hope to uncover the mechanisms of ASD pathogenesis and provide future targets for diagnosis and intervention in autism spectrum disorders.

← Back

A DEPENDENCE-AWARE COMPOSITE FRAMEWORK FOR IDENTIFYING AND LOCALIZING HARD SELECTIVE SWEEPS, WITH APPLICATION TO A SOUTHERN AFRICAN POPULATION

Lauren Sugden[1,2], Elizabeth Atkinson[3], Daniel Vasco[4], Ryan Hernandez[4], Brenna Henn[3], Sohini Ramachandran[1,2]

[1]Brown University, Department of Ecology and Evolutionary Biology, Providence, RI, [2]Brown University, Center for Computational Molecular Biology, Providence, RI, [3]Stony Brook University, Department of Ecology and Evolution, Stony Brook, NY, [4]University of California San Francisco, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA

To understand how human populations have evolved in response to selective forces such as novel pathogens and environments, we need statistical approaches that can effectively mine present-day genomes across the globe for signatures of adaptation. Here, we introduce a novel framework for detecting hard selective sweeps that integrates three signatures: long, shared haplotype blocks; altered site frequency spectra; and population differentiation. Many statistics have been developed to detect each of these three signatures, and recently, composite methods have shown increased power by combining multiple statistics. Our method, which uses a machine learning tool called an Averaged One-Dependence Estimator (AODE), combines statistics in a classification framework that returns probabilistically interpretable results, properly accommodates missing data, and accounts for correlations among component statistics.

Our classifier infers the probability that a locus has undergone a hard sweep based on the joint distributions of component statistics, learned from demographic simulations. In simulated data, our approach vastly outperforms state-of-the-art methods in detection and localization of sweep signals, in some cases reducing the number of false positive predictions by seven-fold. Furthermore, our method is extremely robust to misspecification of the demographic model used for training. In data from the 1000 Genomes Project, we recover known sweep targets, including the *DARC* gene in West Africans, *EDAR* in East Asians, and *SLC24A5* in Europeans, with more precise localization than we would achieve without the dependencies modeled in the AODE framework. In addition, we show that some sweep loci, including the *CD36* gene in West Africans that harbors malaria resistance alleles, can only be detected when these dependencies are modeled.

We applied our classifier to SNP array and exome data from the ≠Khomani San, a population with high genetic diversity and the largest effective population size of any present-day human population, and one that has been Southern Africa for the last ~100,000 years. When we apply our classifier to this data, using a novel scheme for modeling ascertainment bias, we find genes associated with skin pigmentation variation and height, as well as an overabundance of genes involved in neurobiological processes

META-METHYLOME ANALYSIS WITH SMRT SEQUENCING REVEALED A DIVERSITY OF DNA METHYLATION MOTIFS IN UNCULTURED HUMAN GUT MICROBIOMES

Yoshihiko Suzuki[1], Suguru Nishijima[1], Yoshikazu Furuta[2], Wataru Suda[1], Kenshiro Oshima[1], Junko Taniguchi[1], Jun Yoshimura[1], Masahira Hattori[1], Shinichi Morishita[1]

[1]The University of Tokyo, Department of Computational Biology and Medical Sciences, Kashiwa, Japan, [2]Massachusetts Institute of Technology, Institute for Medical Engineering and Science, Cambridge, MA

In prokaryotic genomes, DNA methylation has an important role mainly as a part of restriction-modification system. The diversity of DNA methylation both within and among microbiomes have been unexplored because we cannot distinguish whether motifs found in a cultured bacterial strain have existed since before the strain was cultured. To answer these questions, we sequenced 12 samples (including 6 samples from a family) of human feces with PacBio RS II sequencer, which is known to be able to characterize DNA methylation, especially 6-methyladenine and 4-methylcytosine in bacteria. We confirmed the reproducibility of our research by sequencing a biological replicate of 1 sample. DNA was extracted from the samples by using the enzymatic lysis method instead of the beads method so as not to fragment DNA into smaller pieces. Assembly of sequenced reads was basically performed with FALCON assembler and its latest "unzipping" tool. After the assembly, we used not only Quiver but also Pilon for error correction of the assembled contigs. The contigs of the PacBio assembly were much longer than those of short-read assembly, and therefore the results of gene prediction with MetaGeneAnnotator were also better in PacBio assembly as expected. Thus genes related to transposons and phages were also found more in PacBio assembly with a statistical significance. By analyzing methylome of 12 samples, we found novel DNA methylation motifs that were not registered in REBASE, which is the largest database on restriction-modification system. Moreover, we revealed a considerable diversity of motifs both within and among the uncultured samples. The contigs were clustered according to detected motifs for improving the assembly. Although there were a limited number of motifs conserved among multiple samples of the family, some such motifs suggest the inheritance of DNA methylation motifs. At present, our approach is almost the only way to comprehensively understand the real methylome of uncultured bacteria in microbiomes. Therefore, this meta-methylome analysis will be a powerful approach for elucidating the interaction between DNA methylation patterns of microbiomes and their environments.

A HUMAN DIPLOID METHYLOME USING SMRT READ KINETICS DATA

Yuta Suzuki, Shinichi Morishita

The University of Tokyo, Computational Biology, Graduate School of Frontier Sciences, Kashiwa, Chiba, Japan

Allele-specific methylation (ASM) is widespread in human genomes, and there has been a tremendous interest in how much of them are explained by parent-of-origin (i.e., imprinting), by cis-genetic regulation (mQTL), by epigenetic plasticity, or by random epigenetic drift. The best known method to investigate ASM genome-wide is statistical one (Fang et al. 2012) based on bisulfite-treated short reads and methylation patterns on them. However, the method does not assume the heterozygous variants which can distinguish two haplotypes, thereby it cannot tell, for example, which allele is actually methylated. Alternatively, haplotype-resolved variants enable that bisulfite-treated reads map onto either of the haplotypes. But the haplotype-informative sites in human genomes are too sparse to be found in majority of short-reads, limiting the number of CpGs investigated. Kuleshov et al. (2014) could recover only 6% of the bisulfite-treated reads, even with their high-quality haplotype assembly using Moleculo reads.
We developed an alternative method to observe CpG methylation using SMRT reads kinetics data, which could produce calls concordant with bisulfite sequencing data (>93% both sensitivity and precision). Our method successfully characterized methylation status of long and highly-homologous repetitive regions in human genome by utilizing long reads, which is difficult for short reads. Thus we here explore whether long reads can distinguish two haplotypes and produce methylome for each haplotype. Using high-quality haplotype-resolved variants and SMRT reads data from the Ashkenazi trio produced by Genome in a Bottle Consortium, we inferred the haplotype-of-origin for each SMRT read based on the shared heterozygous variants, 64% of the sequenced bases being attributed its haplotype-of-origin. We then called the methylation status using our method for each haplotype, yielding the "diploid methylome". We identified the genes with ASM including well-known imprinted genes such as PEG13, HYMAI, MEST. We also found the non-imprinted genes with ASM. For example, the previously documented DMR at the promoter of SDHAP3 gene showed ASM.
We will present analysis on allele-specific methylation on human genome. Advantages and disadvantages of our method will also be discussed.

NANOPORE SEQUENCING FOR GENOTYPING PATHOGENS OF TROPICAL DISEASES

Yutaka Suzuki

University of Tokyo, Department of Computational Biology, Kashiwa, Japan

Nanopore sequencer, MinION, has enabled sequencing analysis without pre-installation of expensive conventional sequencers or pre-requisite of specific skills in biological experiments. Even electric supply is not always necessary, by connecting MinION to a laptop PC. These features of MinION have opened the opportunity to enable precise genotyping of pathogens in tropical diseases in a developing country even in its filed areas. In this study, we attempted genotyping Dengue viruses regarding their serotypes (types 1-4). We directly used serum samples of Indonesian 150 Dengue patients, from which viral genomes were directly amplified by the reverse-transcription-LAMP method in an isothermal reaction condition. We directly used the amplified templates for MinION sequencing allocating one flow cell per sample. We found, although the overall sequencing quality was low (82% sequence identify to the reference genome and the quality value of QV= 10 on average), thereby obtained sequence data could discriminate different serotypes of the viruses, whose genome sequences were diverged with the sequence similarity of 70%, with the overall accuracy of >98%. To further examine whether MinION sequencing can be also applied for detecting SNVs, we conducted genotyping of presumed drug resistance-causing SNVs in malaria parasites, Plasmodium falciparum. We similarly subjected ten PCR amplicon-mixes covering these SNVs to the MinION sequencing. In spite that the sequence alignments generated by a alignment program, LAST, showed that the average sequence identity was 80%, we found that the mutations at a particular position could be called by the accuracy of 90%, when all the reads covering the corresponding positions were collectively evaluated. Taken together, we provide the first simple experimental and analytical MinION sequencing procedure, which can be easily followed in a developing country to effectively genotype pathogens of tropical diseases.

ECTOPIC EXPRESSION OF RETROTRANSPOSON-DERIVED PEG11/RTL1 CONTRIBUTES TO THE CALLIPYGE MUSCULAR HYPERTROPHY

Xuewen Xu[1,2], Fabien Ectors[3], Erica E Davis[1,4], Carole Charlier[1], Michel Georges[1], Haruko Takeda[1]

[1]University of Liège, Animal Genomics, GIGA Research Center and Faculty of Veterinary Medicine, Liège, Belgium, [2]Huazhong Agricultural University, Key Lab of Agricultural Animal Genetics, Breeding, and Reproduction of Ministry of Education & Key Lab of Swine Genetics, Wuhan, China, [3]University of Liège, Transgenic platform, FARAH and GIGA Research Center, Liège, Belgium, [4]Duke University Medical Center, Center for Human Disease Modeling and Department of Pediatrics, Durham, NC

The callipyge phenotype is an ovine muscular hypertrophy characterized by polar overdominance: only heterozygous $+^{Mat}/CLPG^{Pat}$ animals receiving the *CLPG* mutation from their father express the phenotype. $+^{Mat}/CLPG^{Pat}$ animals are characterized by postnatal, ectopic expression of Delta-like 1 homologue (DLK1) and Paternally expressed gene 11/Retrotransposon-like 1 (PEG11/RTL1) proteins in skeletal muscle. We showed previously in transgenic mice that ectopic expression of DLK1 alone induces a muscular hypertrophy, hence demonstrating a role for DLK1 in determining the callipyge hypertrophy. We here report newly generated transgenic mice that ectopically express *PEG11* in skeletal muscle, and show that they also exhibit a muscular hypertrophy phenotype. Our data suggest that both DLK1 and PEG11 act together in causing the muscular hypertrophy of callipyge sheep.

THE LANDSCAPE OF REPLICATION ASSOCIATED MUTATIONS IN THE HUMAN AND MOUSE GERMLINES

Lana Talmane[1], Martin Reijns[1], Marie Maclennan[1], Yatendra Kumar[1], Harriet Kemp[1], Sophie Marion de Proce[1], Andrew Jackson[1], Wendy Bickmore[1], Ian Adams[1], Rod Mitchell[2], Martin Taylor[1]

[1]University of Edinburgh, MRC Human Genetics Unit, Edinburgh, United Kingdom, [2]University of Edinburgh, QMRI, Edinburgh, United Kingdom

Genetic mutations provide the raw material for evolution, they are responsible for heritable disease and driving the development of cancer. We have shown that the binding of chromatin and regualtory proteins to DNA can interfere with replication and lead to region with locally elevated mutation rates. Mechanistically this process appears to involve the trapping of DNA polymerase alpha synthesised DNA in the fully replicated genome; a process we have explored with a novel method, EmRiboSeq, that tracks replicative polymerase activity in vivo. Extending this work we have measured the patterns of chromatin accessibility and protein binding specifically in the mammalian germline and related it to the distribution of polymorphism and mutation, to reveal the terrain of replication associated mutations in mice and humans. This provides a means of adjusting neutral substitution rate estimates for fine-scale mutation rate fluctuation when identifying regions of selective constraint. We also identify likely hotspots of paternal lineage mutations within functional regulatory sites.

HINTS OF RECENT POLYGENIC ADAPTATION IN NORTHERN EUROPEANS

Natalie Telis[1,2]

[1]Stanford University, Depts. of Biomedical Informatics, Genetics, and Biology, Stanford, CA, [2]Howard Hughes Medical Institute, Janelia Farms, Ashburn, VA

There are several flagship, well-annotated signals of selection in modern human history, such as lactase. In recent years, there has been increasing evidence for selection on complex traits, such as height. However, most methods are completely underpowered to evaluate all but the strongest recent selection signals. We introduce selection as measured by a novel metric (unpublished). After choosing a high-confidence curated annotated set of variants pertaining to a trait, we polarize the metric for the trait-affecting allele to evaluate mean shift in selection score for variants strongly affecting a trait. We validate that our selection metric is well-correlated with modern allele frequency differences in Northern Europe, and may serve to accurately pinpoint recent frequency shifts resulting from selection. We find evidence of positive selection on multiple variants associated with lightening pigmentation, as well as increasing height. This provides a framework for extending classic single-point selection studies to multiple variants with well-characterized effects on complex traits.

COMPARATIVE CHIP-SEQ UNCOVERS THE MOLECULAR ARCHITECTURE OF HUMAN CENTROMERES

Jitendra Thakur, Steve Henikoff

HHMI, Fred Hutchinson Cancer Research Centre, Basic Sciences, Seattle, WA

Nucleosomes containing the cenH3 (CENP-A) histone variant replace H3 nucleosomes at centromeres to provide a foundation for kinetochore assembly. CENP-A nucleosomes are part of the constitutive centromere associated network (CCAN) that forms the inner kinetochore on which outer kinetochore proteins assemble. Two components of the CCAN, CENP-C and the histone-fold protein CENP-T, provide independent connections from the ~170-bp centromeric α-satellite repeat units to the outer kinetochore. However, the spatial relationship between CENP-A nucleosomes and these two branches remains unclear. CENP-C is directly associated with CENP-A nucleosomes in most all eukaryotes, but the currently accepted model for mammalian CENP-T is that it associates with H3 nucleosomes. Testing this model has been hampered by the high repetitiveness of centromeric α-satellite DNA and the extreme insolubility of kinetochore chromatin. To overcome these technical impediments to understanding the foundation of the human kinetochore, we use a base-pair resolution genomic readout of protein-protein interactions, comparative chromatin immunoprecipitation (ChIP) with sequencing, together with sequential ChIP, to infer the in vivo molecular architecture of the human CCAN. Our results provide the first in vivo mapping of the key CCAN components to CENP-A enriched α-satellite dimers with unprecedented resolution and reveal key features of the fundamental units of human centromeres. In contrast to the currently accepted model in which CENP-T associate with H3 nucleosomes, we establish that both CENP-T and CENP-A occupy the same sequences. We find that CENP-T is centered over the CENP-B box between two well-positioned CENP-A nucleosomes on most abundant centromeric young α-satellite dimers and interacts with the CENP-B/CENP-C complex. Upon cross-linking, the entire CENP-A/CENP-B/CENP-C/CENP-T complex is nuclease-protected over an α-satellite dimer that comprises the fundamental unit of kinetochore chromatin. We conclude that CENP-A/CENP-C and CENP-T pathways for kinetochore assembly are physically integrated over young α-satellite dimers.

POLYGENIC ADAPTATION TO OPTIMUM SHIFTS

Kevin R Thornton

UC Irvine, Ecology and Evolutionary Biology, Irvine, CA

The design and interpretation of genome scans for selection have been largely influenced by models that assume either continuous directional selection on unconditionally beneficial new mutations ("hard" or "classic" sweeps) or fluctuating selection on a previously neutral or weakly-deleterious mutation ("soft" sweeps). However, many examples of adaptive phenotypes in natural population involve complex traits, and the above models may not be adequate descriptions of how such complex polygenic traits evolve. Here, I use forward-time simulation to examine the dynamics of adaptation to a sudden environmental shift, which is modeled as a shift in the optimum value of a complex trait with a broad-sense heritability less than one. The model integrates the existing concepts of "soft" sweeps from standing variation and "classic/hard" sweeps from new mutations, but the strength of selection on individual mutations changes over time as the mean trait value approaches the new optimum value. Adaptation initially proceeds via soft sweeps, fixing mutations of relatively large effect that arose prior to the optimum shift. Subsequent evolution involves hard sweeps of mutations whose effect sizes decrease as time goes on. The site frequency spectrum (SFS) and linkage disequilibrium (LD) show time-dependent deviations from equilibrium values, providing a means of inferring the magnitude of the optimum shift. Finally, the genetic variation for fitness takes a long time to return to equilibrium, and there is continued directional selection after the new optimum phenotype is reached as the population is loses mutations that increase variance in fitness.

EPIGENETIC, CYTOGENETIC AND CELLULAR ASPECTS OF PROGRAMMED DNA ELIMINATION IN THE VERTEBRATE, SEA LAMPREY (*PETROMYZON MARINUS*)

Vladimir A Timoshevskiy, Jeramiah J Smith

University of Kentucky, Bilogy, Lexington, KY

The sea lamprey represents one of the few vertebrate species known to undergo large-scale programmed genome rearrangement (PGR) over the course of its normal development. In order to shed critical light on the cellular context of PGR, we developed techniques that permit labeling of eliminated sequences, DNA modifications and proteins within optically clear and 3D-preserved embryos (which are normally opaque and highly autofluorescent). These techniques have allowed us to reconstruct several epigenetic, cytogenetic and cellular aspects of PGR. Analysis of early developmental stages has revealed that eliminated chromosomes exhibit distinct behaviors during the cell cycle, which lead to the formation of a prominent group of lagging chromosomes. These lagging chromosomes are characterized by their remarkably decelerated, poleward migration and longitudinally-stretched morphologies. After cytokinesis, lagging chromatin is packaged into micronuclei (MNi), which are characterized by several epigenetic modifications that are absent from primary nuclei: ~30% of MNi are enriched for 5-methylcytosine, ~10% are positive for repressive histone modifications H3K9me3 and H3K9me2, and >75% are positive for H3Ser10P. MNi also show clear differences in nuclear envelope formation, including a notable delay in assembly and depletion of lamin B1 and pore-O-linked glycoproteins. These features are typically considered signs of nuclear envelope collapse in cancer cells but represent a normal feature of lamprey development.

Chromosome-based comparative genomic hybridization and fluorescence *in situ* hybridization of probes derived from laser capture microdissection of lagging chromatin verify that retained and eliminated chromosomes differ substantially in sequence content. Moreover, we observe that germline-specific repetitive sequences are located at regions of contact between lagging sister chromatids. We speculate that a subset of these repeats might be functionally relevant with respect to the establishment of contact points between eliminated chromosomes or the epigenetic targeting of chromosomes for elimination.

Taken together, these studies reveal that PGR unfolds through a series of dramatic cellular events that involve the programmatic alteration of several fundamental mechanisms of genome maintenance, including: alignment of chromosomes at metaphase, chromatid cohesion, separation and segregation, and nuclear envelope formation. As such, lamprey PGR represents a novel biological model for studying several basic cellular processes in addition to the regulation of programmed genome editing in context of cellular differentiation.

UNRAVELING GENE EXPRESSION CHANGES IN *LONGISSIMUS* MUSCLE OF NELORE CATTLE DIFFERING FOR FEED EFFICIENCY

Polyana C Tizioto[1,2], Luiz L Coutinho[3], Priscila S Oliveira[1], Wellison J Diniz[4], Andressa O Lima[4], Marina I Rocha[4], Jared E Decker[2], Robert D Schnabel[2], Gerson B Mourão[3], Rymer R Tullio[1], Jeremy F Taylor[2], Luciana C Regitano[1]

[1]Embrapa Southeast Livestock, Animal Biotechnology Laboratory, São Carlos, SP, Brazil, [2]University of Missouri Columbia, Animal Sciences, Columbia, MO, [3]University of São Paulo/ESALQ, Animal Science, Piracicaba, SP, Brazil, [4]Federal University of São Carlos, Genetics and Evolution, São Carlos, SP, Brazil

Feed efficiency traits are extremely important for beef production systems due to their impact on the cost of production. These traits are driven by biological, genetic and physiological mechanisms that are not well understood in beef cattle. Genome-wide association studies (GWAS) performed in beef cattle have identified numerous quantitative trait loci likely to be caused by variation in regulatory elements. These variants affect phenotype by regulating the expression levels of key driver genes. This study identified differentially expressed (DE) genes in the *Longissimus dorsi* (LD) muscle of Nelore steers genetically divergent for Residual Feed Intake (RFI), using RNA sequencing (RNA-seq). Differential gene expression analysis between high RFI (HRFI, inefficient) and low RFI (LRFI, efficient) groups revealed 84 DE genes. From these DE genes, eleven were not annotated in the UMD3.1 bovine reference genome. Sequences for these unknown transcripts were queried by BLAST against the NCBI non-redundant nucleotide sequence database, and most were found to represent non-coding RNAs. The annotated genes are involved in the overrepresented pathways of Metabolism of xenobiotics by cytochrome P450, Butyrate and Tryptophan metabolism and Steroid hormone biosynthesis. Several of the DE genes function in the mitochondrion and in respiratory chain complex activities. Most of the animal's energy requirement is generated by the mitochondria and is produced by the process of oxidative phosphorylation. Members of cytochrome P450 gene family that are involved in these pathways catalyze reactions involved in drug metabolism and the synthesis of cholesterol, steroids, and other lipids. Our data show that expression changes in genes related to mitochondrial function and fatty acid and cholesterol metabolism influence the feed efficiency of Nelore steers. These results enhance our understanding of the metabolic mechanisms underlying feed efficiency in beef cattle.

THE PORCINE BLOOD TRANSCRIPTOMIC RESPONSE TO LIPOPOLYSACCHARIDE (LPS) IS HIGHLY SIMILAR TO THAT OF HUMAN

Christopher K Tuggle[1], Haibo Liu[1], Yet Nguyen[2], Kristina Feye[1], Anoosh Rakhshandeh[1,3], Nicholas Gabler[1], Dan Nettleton[2], Jack C M Dekkers[1]

[1]Iowa State University, Department of Animal Science, Ames, IA, [2]Iowa State University, Department of Statistics, Ames, IA, [3]Texas Tech University, Department of Animal and Food Sciences, Lubbock, TX

The genomic response to LPS, as a canonical inflammatory response, has been studied extensively in several species. Recent analyses of such transcriptomic data in humans and mice identified both similarities and differences. We are developing swine as an alternative animal model for human endotoxemia due to their high similarities in physiology and immunome. Eight pigs of ~ 63 kg body weight (BW) were I.M. challenged with E. coli O5:B55 LPS (20 µg/kg BW). Blood samples were collected immediately before LPS injection (0 hour, h), and 2, 6 and 24 h after LPS injection. Animal behaviors, febrile response and serum cytokine levels during the time course indicated pigs were responsive to the LPS treatment. Blood RNA was sequenced and the resulting RNA-seq count data were transformed into log2(counts per million total counts) and adjusted for technical variations using "voom". Using a linear mixed model spline framework implemented in the R package "lmms", we found 5,525 of 12,476 genes with detectable expression in whole blood to be differentially expressed across the time course compared to the 0 h baseline ($q \leq 0.005$). Genes involved in cell signaling, protein translation, and defense/immune response formed significant clusters. A global comparative analysis using available human leukocyte time course data in response to LPS (Calvano et al. 2005) showed that the pig and human shared more than 1,900 differentially expressed genes with similar dynamics during the time course ($q \leq 0.05$). Although we also observed significant differences between the two species in response to LPS, which could be due to LPS dosage, route of administration and E. coli serotype of the LPS used, we conclude that acute global blood transcriptomic responses to a systemic LPS challenge are very similar between humans and pigs, both at the level of individual genes, as well as in terms of gene clusters.

DE NOVO GERMLINE AND NODULAR HETEROTOPIA-ASSOCIATED POSTZYGOTIC MUTATIONS OF STXBP1 IN AN EPILEPSY PATIENT SUCCESSFULLY TREATED WITH RESECTIVE SURGERY

Mohammed Uddin[1], Cyrus Boelman[2], Ledia Brunga[1], Sylvia Lamoureux[3], Dimitri Stavropoulos[3], James Drake[4], Cecil Hahn[4], Cynthia Hawkins[3], Adam Shlien[1], Berge Minassian[4], Stephen Scherer[1,5,6]
[1]The Hospital for Sick Children, Genetics and Genome Biology, Toronto, Canada, [2]BC Children's Hospital, Division of Neurology, Vancouver, Canada, [3]The Hospital for Sick Children, Genome Diagnostics, Toronto, Canada, [4]The Hospital for Sick Children, Division of Neurosurgery, Toronto, Canada, [5]University of Toronto, McLaughlin Centre, Toronto, Canada, [6]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [7]The Hospital for Sick Children, Division of Pathology, Toronto, Canada

RATIONALE
Syntaxin-binding protein 1 (STXBP1) mutations have been reported across a spectrum of neurodevelopmental disorders, including Otahara syndrome and other epileptic encephalopathies. In order to delineate the phenotypic complexity of cases, it is necessary to analyze the STXBP1 mutation with consideration for the sequence context. Here, we report the first case of a successfully treated epilepsy surgery patient with a heterozygous STXBP1 germline mutation that was mosaic within a surgically resected area populated with dysplastic cells.
METHODS
Patients undergoing resective epilepsy surgery for refractory focal epilepsy that was associated with suspected cortical dysplasia were recruited prospectively for genetic study. For the current patient, a targeted (70 genes) sequencing approach was applied to identify rare pathogenic mutations. DNA from peripheral blood was collected from the nuclear family. DNA was also collected from resected left lateral temporal cortex tissue with dysplastic neurons. We applied droplet digital PCR (ddPCR) technology to validate the copy number variation (CNVs).
RESULTS
We have identified a de novo 4.9Kb deletion impacting exon 3-4 of the STXBP1 gene in a 6-year old proband (male) diagnosed with refractory focal-onset epilepsy with initial infantile spasms, global developmental disorder, autism spectrum disorder and MRI imaging suspicious for focal cortical dysplasia type 1 in the left anterior temporal cortex. The patient underwent left anterior temporal lesionectomy including mesial structures. Surgical pathology demonstrated nodular heterotopias in the superior temporal gyrus and deep temporal white matter with collections of cells with small round nuclei and clear cytoplasm. Hippocampus, para-hippocampal gyrus and lateral temporal cortex were normal. The ddPCR result revealed heterozygous state of the deletion in blood, whereas, evidence of the mosaic presence of heterozygous and homozygous deletions was observed within the dysplastic tissue cell population of the proband. The patient has been seizure free since surgery.
CONCLUSIONS
We report a novel de novo mutation within STXBP1 in a patient with a complex neurodevelopmental phenotype and focal epilepsy with areas of dysplastic neurons. This is the first case of STXBP1 showing formation of mosaic copy number variations within dysplastic brain cells. These findings suggest a role for STXBP1 in malformations of cortical development and a role for epilepsy surgery in the management of STXBP1-associated epilepsy.

← Back

SYSTEMATIC FUNCTIONAL DISSECTION OF COMMON GENETIC VARIATION AFFECTING TRANSCRIPTIONAL REGULATION AND HUMAN DISEASE

Jacob C Ulirsch[1,2], Satish K Nandakumar[1,2], Tarjei S Mikkelsen[2], Vijay G Sankaran[1,2]

[1]Boston Children's Hospital, Hematology, Boston, MA, [2]Broad Institute, Cambridge, MA

Genome-wide association studies (GWAS) have identified over 10,000 common single nucleotide polymorphisms (SNPs) associated with hundreds of human traits and diseases. Nevertheless, for the vast majority of GWAS loci, a causal variant remains unknown, as a result of the large number of variants in linkage disequilibrium (LD) and the challenges of assessing the function of non-coding variation. Recent advances in fine mapping are beneficial but only rarely allow for the identification of a single causal variant. Similarly, large-scale genome editing approaches have proven useful for assessing endogenous elements, but do not allow for interrogation of allele-specific variation in a scalable manner. In order to address the need for high-throughput functional screening of GWAS loci, we have developed and applied a massively parallel reporter assay (MPRA) that can identify regulatory elements that alter transcription from a minimal promoter. As a proof-of-principle, we chose to screen 2,756 variants in high LD with 75 sentinel SNPs identified from the GWAS for RBC traits, given both the predicted erythroid-intrinsic effects of these variants and our prior success in examining individual variants associated with these traits. Our MPRA identified endogenous regulatory elements based upon erythroblast DNase 1 hypersensitivity (DHS) and the activity of tested elements was predicted by the presence of key erythroid transcription factor (TF) motifs. We identified 32 functional variants, termed MFVs, that exhibited differential activity by allele at 23 loci. We derived a positive predictive value for causality from our screen of between 32-50% based upon genetics (by comparing credible and non-credible sets) and putative regulatory function from predictive algorithms. Where heterozygous samples were available, 88% of MFVs were directionally consistent with allele-specific reads in either DHS or ChIP-seq. Finally, we used genome editing to verify the endogenous effects of 3 MFVs and identified 1-3 target genes for each. In one case, we linked the target gene, RBM38, back to the original GWAS phenotype in donor-derived erythroblasts. Importantly, when we analyzed TF binding motifs, DNA shape features, and allelic skew in ChIP-seq data, we determined that common SNPs associated with RBC traits frequently affect a regulatory pathway involving the erythroid master TF GATA1. Overall, our method provides a novel, scalable, and cost-effective approach for GWAS follow-up, and its addition into the GWAS follow-up toolbox holds promise to accelerate our understanding of human disorders.

← Back

INTEGRATIVE ANALYSIS OF ESSENTIAL GENE PATTERNS CONTRIBUTING TO CANCER DRUG RESPONSE

Matthew H Ung[1], Chao Cheng[1,2]

[1]Geisel School of Medicine at Dartmouth, Gentics, Hanover, NH, [2]Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH

RNA interference (RNAi) has been used extensively to identify specific genes essential for proper functioning of a given biological process of interest. In the context of cancer, RNAi serves as a phenotypic screening procedure to identify cancer-associated genes that are necessary for neoplastic growth. Several consortia have generated these data by performing high-throughput shRNA phenotypic screens in an effort to characterize the universe of essential genes in several cancer types. These data-generating experiments have the ability to provide insight into mechanisms that underlie drug resistance and pinpoint novel genetic lesions that constitute potential therapeutic targets. Thus, in this report, we perform a systematic integrative analysis of gene essentiality data from consortia-lead projects and case-by-case studies to model drug response in diverse panels of cell lines. Our computational framework utilizes supervised and unsupervised machine learning algorithms that accurately model response to an extensive panel of experimental and FDA approved anticancer drugs using gene essentiality features. Most importantly, we show that models that incorporate gene essentiality information outperforms those that use only gene expression data when modeling drug response. Furthermore, we integrate DNA methylation, miRNA, and ENCODE ChIP-seq data into our analysis and identified regulatory modules that may contribute to gene essentiality. Overall, we maintain that incorporating gene essentiality information into translational genomic studies will yield additional insight into cancer mechanisms that go undetected when analyzing other genomic data types.

SEQUENCE MINING REVEALS INFORMATIVE AND ENRICHED ELEMENTS IN (META-)GENOMIC DATA

Niko Välimäki

University of Helsinki, Department of Medical and Clinical Genetics, Helsinki, Finland

We introduce a data mining based approach for exploring informative and enriched sequence elements in (meta-)genomic data. While traditional reference-free approaches to analyse large-scale metagenomic data have limited to search over predetermined, fixed-length k-mers, we apply so called sequence mining methodology to extract variable length k-mers (i.e. sequence elements of variable length) that discriminate two (or more) datasets. Known combinatorial solutions to the sequence mining problem range from an optimal-time algorithm to various time-memory tradeoffs (Fischer, Mäkinen, V 2008), however, due to significant memory requirements, the existing methods are practical only up to a few gigabytes of input. These scalability issues can be circumvented via a distributed algorithm that, in practice, achieves a significant reduction in both memory and time compared to state-of-the-art methods (V, Puglisi 2012).

Direct applications for the sequence mining approach were proposed in exploration and retrieval of whole-metagenome sequencing samples (Seth, V, Kaski, Honkela Bioinformatics 2014), which utilized a distributed sequence mining framework to efficiently extract all informative sequence k-mers from a pool of metagenomic samples and use them to measure the dissimilarity between two samples. The framework was evaluated on terabyte-scale metagenomics data and indicated high accuracy in, for example, discriminating between different body sites even though the method is unsupervised. Sequence mining based approaches and variable length k-mers appear to improve evaluations and give better association power compared to a predetermined, fixed k-mer length. Our on going work includes further applications in differential analysis of metagenomic sequences (Seth, V, Kaski, Honkela unpublished) and in sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes (Weinert, Chaudhuri et al. Nat Comm 2015; Lees, Vehkala, V et al. unpublished).

LEVERAGING HERITABILITY OF H3K27AC HISTONE MODIFICATIONS TO CREATE BETTER FUNCTIONAL ANNOTATIONS

Bryce van de Geijn[1], Alkes L Price[1,2]

[1]Harvard TH Chan School of Public Health, Department of Epidemiology, Boston, MA, [2]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA

Chromatin state is an important marker of gene regulation. In particular, the histone modification H3K27ac has been used to identify active enhancers. However, this process often relies on ChIP-seq peak calling and results in binary annotations that may not capture all of the underlying biology. Indeed, chromosome regions that are most active are likely to be bound by transcription factors and are therefore nucleosome-depleted. This depletion creates contours in H3K27ac signal that are further enriched for regulatory activity. Better annotations of H3K27ac could prove useful for predicting and explaining disease-causing polymorphisms.

We use an existing method, LD Score regression, to partition heritability of H3K27ac peaks in Yoruba lymphoblastoid cell lines across previously annotated functional elements. We then use heritability explained as a metric to create new H3K27ac peak annotations that are maximally enriched for regulatory function. Finally, we use our optimized algorithm to annotate H3K27ac in a wide range of cell types and show that these annotations are enriched for phenotype heritability.

SYSTEMATIC PAN-CANCER ANALYSIS OF IMMUNE INFILTRATION

Frederick S Varn[1], Chao Cheng[1,2]

[1]Geisel School of Medicine at Dartmouth, Department of Genetics, Hanover, NH, [2]Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH

The immune system is heavily involved in shaping the development and evolution of solid tumors. This interplay is highly context-dependent, with the quantities and behaviors of different immune cells changing depending on the tissue the neoplasm originated in. Computational approaches that integrate patient genomic data now allow for a systematic inference of the presence of different kinds of infiltrating immune cell types. Here, we apply our previously developed approach to patient data across twelve tumor types from The Cancer Genome Atlas (TCGA). Using this approach, we explore the relationship between immune infiltration and other genomic features, including gene expression, mutational load, and neoantigen count. Additionally, we show through hierarchical clustering that many patients share similar immune response profiles, regardless of cancer type. Finally, we examine the strength of patient immune response profiles in predicting survival and response to CTLA-4 blockade therapy. Together, these analyses provide a comprehensive examination of the tumor-immune system interaction and the genomic factors that regulate it.

FINDTRANSLOCATIONS – A STRUCTURAL VARIANT CALLING TOOLKIT

Francesco Vezzi[1], Jesper Eisfeldt[2], Daniel Nilsson[2], Anna Lindstrand[2]

[1]National Genomics Infrastructure, SciLifeLab, Stockholm University, Stockholm, Sweden, [2]Karolinska Institute, Department of Molecular Medicine and Surgery, Stockholm, Sweden

In the last decade, genomic structural variations have emerged as an important contributor to the genetic load of both rare and common disorders. A structural variant is traditionally defined as a balanced or unbalanced genetic rearrangement larger than 1kb but recent advances in genetic technology has enabled the detection of increasingly smaller rearrangements. In genetic diagnostic, currently applied techniques such as FISH and microarray studies have limited resolution. Massively parallel whole genome sequencing (WGS) is a promising technique that may be used to identify a large proportion of genomic structural variation in a single experiment. However, the detection of structural variants from WGS data is complicated by the vast amount of normal variants and reference errors and currently relies on using multiple variant callers, increasing the overall computational cost. Moreover, different variant callers are affected by different biases that might introduce a large amount of false positive calls.

FindTranslocations is a structural variant detection algorithm using discordant read pairs and coverage information to identify balanced and unbalanced structural variants while consuming less than 5 cpu hours per 30X whole genome sample. It contains a built in database function allowing the user to create local variant frequency databases, that may be used to filter out rare variants or detect variants that are common within a group. The use of a database allows not only to easily highlight rare variations, but allows also reducing the impact of false positive calls.
FindTranslocations goes beyond mere structural variant calling providing positional information for specific rearrangements to help make a correct interpretation of the clinical significance. We have used FindTranslocations in a validation set of 61 samples with previously identified clinically relevant structural variants for which we have 30x WGS data. FindTranslocations has an 89% sensitivity for unbalanced rearrangements (35/40 detected duplications, 21/23 detected deletions) sized from 5 kb to 5000 kb and 84% sensitivity for balanced rearrangements (21/25 detected).

AN OPEN SOURCE WEB APPLICATION FOR POLYGENIC TRAIT AND DISEASE RISK PREDICTION

Ümit Seren[1], Georgios Athanasiadis[2], Gaurav Bhatia[3], Jade Cheng[1], Thomas Mailund[1], Anders Borglum[4], Magnus Nordborg[1], Mikkel H Schierup[2], Bjarni J Vilhjalmsson[1,3]

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria, [2]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark, [3]TH Chan Harvard School of Public Health, Epidemiology and Biostatistics, Boston, MA, [4]Aarhus University, Department of Biomedicine, Aarhus, Denmark

In recent years, predicting polygenic traits and disease risk from the genotype has become common practice in both clinical and non-clinical settings. Interpreting such predictions for individual genomes is nontrivial and results are typically obtained with proprietary software that is not publicly available. In non-clinical settings, 23andme and other private companies have led the charge in providing direct-to-consumer genetic testing, resulting in more than one million individuals being genotyped to date. This explosion of genetic data motivated us to develop an easy-to-install open source platform for polygenic trait and disease risk prediction. The platform is developed with both clinical and non-clinical settings in mind. It implements three main computational steps: 1) imputation, 2) ancestry inference, and 3) genetic prediction. Through careful choice of algorithms and efficient implementation, the whole pipeline can be applied to a 23andme genotype in about one minute, ensuring an interactive user experience for all of the three steps. To demonstrate how the application can be used in practice, we provide an online implementation (http://thehonestgene.github.io), where users can upload their 23andme genotypes and obtain polygenic predictions for height and BMI. The polygenic risk prediction is carried out using LDpred weights, which are trained on publicly available GWAS summary statistics. To estimate the accuracy of the predictions we used self-reported height and BMI from 609 Danish high school students as validation data. The prediction R2 in the sample after adjusting for age, sex, and the first ten PCs was 25% for height and 11% for BMI.

← Back

METHYLATED CYTOSINES MUTATE TO TRANSCRIPTION FACTOR BINDING SITES THAT DRIVE TETRAPOD EVOLUTION

Ximiao He[1], Desiree Tillo[1], Jeff Vierstra[2], Khund-Sayeed Syed[1], Callie Deng[1], Jordan Ray[1], John Stamatoyannopoulos[2], Peter FitzGerald[3], Charles Vinson[1]

[1]National Cancer Institute, NIH, Laboratory of Metabolism, Bethesda, MD, [2]University of Washington , Department of Genome Sciences, Seattle, WA, [3]National Cancer Institute, NIH, Genome Analysis Unit, Genetics Branch, Bethesda, MD

In mammals, the cytosine in CG dinucleotides is typically methylated producing 5-methylcytosine (5mC), a chemically less stable form of cytosine that can spontaneously deaminate to thymidine resulting in a T•G mismatched base pair. Unlike other eukaryotes that efficiently repair this mismatched base pair back to C•G, in mammals, 5mCG deamination is mutagenic, sometimes producing TG dinucleotides, explaining the depletion of CG dinucleotides in mammalian genomes. It was suggested that new TG dinucleotides generate genetic diversity that may be critical for evolutionary change. We tested this conjecture by examining the DNA sequence properties of regulatory sequences identified by DNase I hypersensitive sites (DHSs) in human and mouse genomes. We hypothesized that the new TG dinucleotides generate transcription factor binding sites (TFBS) that become tissue-specific DHSs (TS-DHSs). We find that 8-mers containing the CG dinucleotide are enriched in DHSs in both species. However, 8-mers containing a TG and no CG dinucleotide are preferentially enriched in TS-DHSs when compared with 8-mers with neither a TG nor a CG dinucleotide. The most enriched 8-mer with a TG and no CG dinucleotide in tissue-specific regulatory regions in both genomes is the AP-1 motif (**TG**A$^{C}/_{G}$T**CA**N), and we find evidence that TG dinucleotides in the AP-1 motif arose from CG dinucleotides. Additional TS-DHS-enriched TFBS containing the TG/CA dinucleotide are the E-Box motif (G**CA**GC**TG**C), the NF-1 motif (GG**CA**—**TG**CC), and the GR (glucocorticoid receptor) motif (G-A**CA**—**TG**T-C). Our results support the suggestion that cytosine methylation is mutagenic in tetrapods producing TG dinucleotides that create TFBS that drive evolution.

RARE VARIANTS AND PARENT-OF-ORIGIN EFFECTS ON WHOLE BLOOD GENE EXPRESSION ASSESSED IN LARGE FAMILY PEDIGREES

Ana Viñuela[1], Andrew A Brown[1], Angel Martinez-Perez[2], Nikolaos I Panousis[1], Olivier Delaneau[1], Helena Brunel[2], Andrey Ziyatdinov[2], Maria Sabater-Lleal[3], Anders Hamsten[3], Juan C Souto[2], Alfonso Buil[1], Jose M Soria[2], Emmanouil T Dermitzakis[1]

[1]University of Geneva, Dpt Genetic Medicine and Development, Geneva, Switzerland, [2]Instituto de Investigaciones Biomedicas Sant Pau, Barcelona, Spain, [3]Karolinska Institutet, Dpt Medicine, Stockholm, Sweden

Studying genetic regulation of gene expression in related individuals provides insights that are inaccessible when using unrelated individuals, such as heritabilities, rare eQTL commonly observed in the pedigree, imprinting, and parent-of-origin effects. We recruited 935 individuals from 35 pedigrees, with an average of 27 individuals per pedigree and a total of 8654 related pairs. This study (GAIT2) is an effort to understand idiopathic thrombophilia, for which the cohort has been deeply phenotyped with available blood cell measurements and other clinical phenotypes. We sequenced the mRNA transcriptome from whole blood for all individuals. We found a median heritability of expression of 0.22, similar to that estimated in the TwinsUK cohort ($h2 = 0.19$, EuroBATS). To better understand the genetic regulation of expression, we identified 11,297 eQTL (corrected $p<0.01$, $\pi1 = 0.98$) using a cis association mapping (MAF>0.01), with variance components to consider the familial structure in the data. To test whether the eQTL discovered could be rare in the general population, we looked at the MAF of the variants in the CEU, GBR, IBS and TSI populations of 1000 Genomes. Compared to eQTL from the Depression Genes and Networks study with blood expression from unrelated individuals, we saw an excess of variants with MAF < 0.01 (9.6% eQTL compared to 0%, median MAF is 0.11 in GAIT2, 0.27 in DGN). Using the many trios within the pedigrees, we looked for parent-of-origin effects on expression. We performed a cis scan to find variants where the effect of the reference allele in heterozygotes depended on whether it was maternally or paternally inherited (called here parent-of-origin in expression QTL (poeQTL)). We found 4 significant poeQTL ($p<0.05$) for MEG3, CD9, NDN and SLC24A4. MEG3 and NDN are known imprinted genes with maternal and paternal expression, respectively. Additionally, pvalue enrichment analysis showed a $\pi1$ of 0.40, suggesting than parent of origin effects may be more widespread than previously anticipated in humans. In further work exploring parent of origin effects, we intend to use expression levels from trios to identify examples of imprinting, and look at mechanistic aspects of all observed parent of origin effects by integrating chromatin variation data from reference samples.

THE LANDSCAPE OF ISOFORM SWITCHES IN HUMAN CANCERS

Kristoffer Vitting-Seerup[1,2], Albin Sandelin[1,2]

[1]Section for Computational and RNA Biology (SCARB), Department of Biology, University of Copenhagen, Copenhagen, Denmark, [2]Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark

Isoform switching, which referees to the differential usage of different gene-isoforms in different conditions, are with a few exceptions largely overlooked in cancer biology. This lack of knowledge probably occurs in part because it is difficult to find and predict the functional impact of such switches and in part because the extend of isoform switching in cancers is not known.

To solve these problems we developed IsoformSwitchAnalyzeR, an easy to use R package which enables statistical identification, annotation and visualisation of isoform switches. We used IsoformSwitchAnalyzeR to identify isoform switches 12 cancer types covering almost 6000 cancer patients from The Cancer Genome Atlas (TCGA). We find that isoform switches are extremely common: across the 12 solid cancer types more than 4000 (29% of all) multi-transcript genes display differential isoform usage in at least one cancer type. In 50% of these genes (14% of all genes) the changes have easily predicted functional consequences such as domain loss, domain switch or loss of coding potential. The genes with isoform switching are not random, but are highly enriched for genes in cell signalling, adhesion and cancer signatures. Many for the found isoform switches are furthermore pan-cancer events and we both identify known isoform switches in new cancer types as well as describe novel pan-cancer isoform switches.

At The Biology of Genomes conference I will mainly focus on communicating the results from the pan-cancer analysis of isoform switching.

DETERMINING AN INFLUENZA VACCINE STRAIN USING GENOMIC SEQUENCE

Xiu-Feng (Henry) Wan[1], Tong Zhang[2], Lei Han[2], Lei Li[1], Lei Zhong[1], Feng Wen[1]

[1]Mississippi State University, Dept of Basic Sciences, College of Vet Med, Mississippi State, MS, [2]Rutgers University, Dept of Statistics, Piscataway, NJ

Influenza A virus causes both pandemic and seasonal outbreaks, leading to loss of from thousands to millions of human lives within a short time period. Vaccination is the best option to prevent and minimize the effects of influenza outbreaks. Timely identification of emerging influenza virus antigenic variants is central to the success of influenza vaccination programs. Empirical methods to determine influenza virus antigenic properties are time-consuming and mid-throughput and require live viruses. Here, we present a novel computational method for determining influenza virus antigenicity and vaccine strain selection using genomic sequence. Based on our previous sparse learning method (Sun et al. 2013. mBio, 4(4)), an integrated method using temporal multi-task learning (TMTL) and the hierarchical sparse interaction modeling (HSIM) methods are developed for phenotype-genotype association study, especially, to identify amino acid residues associated with antigenic changes. TMTL overcomes the challenges of high data sparsity and low reactors in serological data and HSIM resolves the high-order interactions among multiple residues by utilizing the heredity structure in the feature sparsity. The resulting interactive residues can be those co-evolve during viral evolution and syngenetically determine viral antigenic drift. Weights are assigned to each predicted site, and a scoring function is developed to determine if a query virus be an antigenic variant based on its genomic sequence. We applied this method in the serologic data from 1968 to 2015, and a total 58 residues are identified. Among them, 46 residues are located in antibody binding sites A-E and 41 in important high-order interactions (co-mutations). Residues associated with the 15 previously known H3N2 antigenic clusters that led to a change of the H3N2 vaccine strain were identified. This method will be useful in influenza vaccine strain selection by significantly reducing the human labor efforts for serological characterization and will increase the likelihood of correct influenza vaccine candidate selection, and it can also be useful for other phenotype-genotype studies.

NETWORK ENHANCEMENT: A GENERAL METHOD TO EXPLOIT THE TRANSITIVE EDGES IN COMPLEX NETWORKS

Bo Wang, Serafim Batzoglou

Stanford University, Computer Science, Stanford, CA

Complex networks emerge in a plethora of disciplines in natural science. They entail non-trivial topological features and patterns critical to understanding interactions within complicated biological systems. However, observed networks from data are typically noisy, and unraveling clear structures and dynamics in the networks remains a critical challenge in network science. Recent attentions have been paid to cleaning up noisy networks by inferring intrinsic links automatically [1]. In this study, we propose a novel method, Network Enhancement (NE), to improve the signal-to-noise ratio of complex networks and therefore facilitate downstream network analysis. NE leverages the transitive edges of complex networks by exploiting local structures and alleviates the corrupted links in these networks. We benchmarked NE on a popular collection of gene regulatory networks from DREAM5 Challenge [2] and show significant improvement over 10 existing methods that predict gene interactions. Additionally, we demonstrate the effectiveness of NE in community detection problems derived from several real networks. Finally, we applied NE for gene function prediction and observed higher prediction accuracy over state-of-the-art supervised learning methods. Extensive comparisons with a competing method Network Deconvolution [1] imply that proper manipulation of transitive links by NE is indeed advantageous. The results are supported by rigorous theoretical justifications. This work not only provides a method that can be easily applied to any type of noisy network, but also provides fundamental insights into how to utilize transitive edges to understand complex interactions in networks.

References:
[1] Feizi, Soheil, et al. "Network deconvolution as a general method to distinguish direct dependencies in networks." Nature biotechnology 31.8 (2013): 726-733.
[2] Marbach, Daniel, et al. "Wisdom of crowds for robust gene network inference." Nature methods 9.8 (2012): 796-804.

← Back

SMASH, A FRAGMENTATION AND SEQUENCING METHOD FOR GENOMIC COPY NUMBER ANALYSIS

Zihua Wang*, Peter Andrews*, Jude Kendall, Beicong Ma, Inessa Hakker, Linda Rodgers, Michael Ronemus, Michael Wigler, Dan Levy

Cold Spring Harbor Laboratory, Wigler Laboratory, Cold Spring Harbor, NY

Copy number variants (CNVs) underlie a significant amount of genetic diversity and disease. CNVs can be detected by a number of means, including chromosomal microarray analysis (CMA) and whole genome sequencing (WGS), but these approaches suffer from either limited resolution (CMA) or are highly expensive for routine screening (both CMA and WGS). As an alternative, we have developed a next-generation sequencing-based method for CNV analysis termed SMASH, for Short Multiply Aggregated Sequence Homologies. SMASH utilizes random fragmentation of input genomic DNAs to create chimeric sequence reads, from which multiple mappable tags can be parsed using maximal almost-unique matches (MAMs). The SMASH tags are then binned and segmented generating a profile of genomic copy number at the desired resolution. For each pair of reads (2 X 150bp), WGS averages less than one map, whereas SMASH yields more than four maps with the quality being of the same order as those seen with WGS mapping. Using correction and testing protocols optimized for WGS data, we show that on a map-for-map basis, SMASH generates read-depth copy number data that is virtually equivalent to WGS at a fraction of the cost. Because fewer reads are necessary relative to WGS to give accurate CNV data, SMASH libraries can be highly multiplexed, allowing large numbers of individuals to be analyzed at low cost. Increased genomic resolution can be achieved by sequencing to higher depth.

INTEGRATED GENOMIC ANALYSIS WITH IOBIO

Alistair Ward, Chase Miller, Tonya Di Sera, Yi Qiao, Brent Pedersen, Aaron Quinlan, Gabor Marth

University of Utah, School of Medicine, Salt Lake City, UT

With the rapid adoption of next-generation sequencing data in many research, clinical and diagnostic pipelines, it is imperative that the available tools can provide the power, functionality and ease-of-use required for all analysts. In particular, as more research MDs, clinical diagnosticians and other non-bioinformaticians are becoming actively involved in analyzing genomic data, easy to use tools with no required installation, data upload, or computational expertize, that can provide users with an interactive, yet visually intuitive environment are increasingly necessary.

We present a genomic data analysis paradigm built on the innovative IOBIO platform. This workflow starts with aligned DNA sequence and variant calls, the typical outputs of institutional level genomic pipelines, and uses IOBIO web-based applications to identify problems with underlying DNA libraries (e.g. excessive PCR duplicates), sequence alignments and variant calls. Having established data quality, gene-level variant prioritization is performed using real-time functional annotation, comparison with external variant databases, population allele frequencies, and interactive, on-demand variant calling using Freebayes to identify false-negative calls. Additionally, variant lists from external variant prioritization tools and gene-lists from phenotype-driven algorithms can be generated natively within the application or imported from external sources. All of this is achieved without the need for expensive, time-consuming data uploads or lengthy analyses.

We demonstrate this workflow using a real-world clinical example of a patient with a lactic acidosis phenotype. With access to DNA data from the proband and the parents, we confirm problems with underlying data using bam.iobio and vcf.iobio, perform variant prioritization using multiple annotations, check allele frequencies and multiple inheritance patterns using Gemini and use gene.iobio to interrogate the list of variants, generate candidate genes using Phenolyzer and ultimately identify the causative variant. We finally present some of the features and applications that will soon be augmenting this workflow making it a “one-stop-shop” genomic analysis platform for all users of sequencing data.

ALLELIC SPECIFIC EXPRESSION ANALYSIS OF STRUCTURAL VARIATION IN HUMAN POPULATIONS.

Jia Wen, Andrew Quitadamo, Xinghua Shi

Department of Bioinformatics and Genomics, Charlotte, NC

Allelic specific expression (ASE) analysis is an effective method for understanding the functional impacts of genomic variation on phenotype. Classic ASE analysis usually assesses the imbalance of allele counts that carry reference and alternative alleles over heterozygous SNP sites in an individual. Another type of genetic variation, structural variations (SV) including copy number variations, deletions, insertions and inversions, have been shown to affect gene expression. Therefore, it is important to assess how SVs affects allelic specific expression using SNPs as surrogates. In our study, we deployed a pipeline to detect candidate SVs with potential ASE effects based on nearby heterozygous SNPs. Our pipeline is based on WASP, which is a set of open-source tools for unbiased allele-specific read mapping and discovery of QTLs (Gejin et al, 2015). Different than other methods that usually choose a read with the highest mapping score, WASP randomly selects a read from the set of duplicates to correct count bias towards reference allele. We first identified SNP ASE using the binomial test, and then used these SNPs as surrogates to identify candidate SVs in the neighborhood with potential allelic expression effects. Using the RNA-seq data from GEUVADIS and genotype data from 1000 Genomes Phase 3 data set, we identified a set of SVs with potential allelic effect. We found that these candidate SVs were significantly enriched for histone modifications, DNase hypersensitive sites, chromatin open regions, and many transcription factor binding sites.

Reference:
Bryce van de Geijn, Graham McVicker, Yoav Gilad , Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods 12, 1061–1063 (2015).

APPLY EMPIRICAL BAYESIAN ELASTIC NET METHOD TO MICRORNA EPISTASIS ANALYSIS IN COLON CANCER

Jia Wen, Benika Hall, Andrew Quitadamo, Xinghua Shi

Department of Bioinformatics and Genomics, Charlotte, NC

Colon cancer is the second leading cause of cancer-related death in the United States and Europe, and it's the third most common cancer among males and females around the world. The development of colon cancer is considered to be related to a series of different genetic and epigenetic alterations. The discovery of small non-coding microRNAs (miRNAs) has opened a new scope to the regulation of gene expression for cancer research. Previous studies have shown that the aberrant expression of miRNAs in colon cancer can induce tumor development. MiRNAs regulate many genes post-translation and appear to play a key role in cancer development, progression and response to chemotherapy. Epistasis is an important genetic component in cancer and chemoresistance research. In this study, we investigate the interactions between miRNAs with epistatic effect on pathological stages of colon cancer. Specifically, we adopted a deployed pipeline based on an Empirical Bayesian Elastic Net (EBEN) method to dissect the main and pair-wise epistatic effect of miRNAs. The EBEN method usually selects one of the features which is most highly correlated with the dependent variable and applies two statistical techniques of variable selection and shrinkage operator, to infer the main effect and epistasis of high dimensional data. Using the miRNA expression profiles on colon cancer from The Cancer Genome Atlas (TCGA), we identified a set of miRNAs with a main effect and pair-wise epistatic effect on the pathological stages of colon cancer. We found that the majority pairs of epistatic miRNAs have common target genes in miRNA target databases composed of miR2Disease, TargetScan and miRDB. Our results provide a potential set of candidate miRNAs that can serve as a novel class of therapeutic targets and serve for drug resistance research in pathological stages.

ENHANCER-PROMOTER INTERACTIONS ARE ENCODED BY COMPLEX GENOMIC SIGNATURES ON LOOPING CHROMATIN

Sean Whalen[1], Rebecca M Truty[2], Katherine S Pollard[1]

[1]Gladstone Institutes, Cardiovascular Disease, San Francisco, CA, [2]Invitae Corporation, San Francisco, CA

Discriminating the gene target of a distal regulatory element from other nearby transcribed genes is a challenging problem with the potential to illuminate the causal underpinnings of complex diseases. Within each locus (or topological domain) of the human genome are many active regulatory enhancers and expressed genes. High resolution chromatin capture data reveals that a subset of enhancers consistently loop to specific promoters, but these are typically not the closest gene. To learn how enhancers find their targets, skipping over other expressed promoters in the locus, we developed TargetFinder, a computational method that reconstructs regulatory landscapes from heterogeneous features along the genome. The resulting models accurately predict individual enhancer-promoter interactions across diverse cell lines with a false discovery rate up to fifteen times smaller than using the closest gene, and a balance of precision and recall ranging from 77-90%. By evaluating the genomic features driving this accuracy, we uncover interactions between structural proteins, transcription factors, epigenetic modifications, and transcription that together distinguish interacting from non-interacting enhancer-promoter pairs. Most of this signature is not proximal to the enhancers and promoters, but instead decorates the looping DNA (i.e., the genomic interval between an enhancer and promoter). We delve into our "black box" machine learning model and present under-studied transcription factors and epigenetic modifications that offer insights into the mechanisms of DNA looping beyond well-understood marks such as CTCF and cohesin. TargetFinder also quickly screens new datasets for relevance to DNA looping, demonstrated by a case study of sumoylation ChIP-seq data, serving as a platform for vastly improved prediction and analysis of enhancer-promoter interactions towards discovery of causal non-coding variants. We conclude that complex but consistent combinations of marks on the one-dimensional genome encode the three-dimensional structure of fine-scale regulatory interactions.

INTEGRATIVE GENOMIC DECONVOLUTION OF RHEUMATOID ARTHRITIS GWAS LOCI INTO GENE AND CELL TYPE ASSOCIATIONS

John W Whitaker*[1], Alice M Walsh*[2], Chris C Huang[2], Yauheniya Cherkas[2], Sarah L Lamberth[2], Carrie Brodmerkel[2], Mark E Curran[2], Radu Dobrin[2]

[1]Janssen Research and Development, LLC, Discovery Sciences, San Diego, CA, [2]Janssen Research and Development, LLC., Immunology, Spring House, PA

Genome-wide association studies (GWAS) have identified over 100 genetic loci associated with rheumatoid arthritis (RA). The majority of RA GWAS loci do not alter protein coding sequences or obvious regulatory regions. Given that many associated GWAS loci are replicated and robust we believe the genetic findings to be true and seek to understand their function in disease. One hypothesis we are exploring is that some proportion of RA GWAS SNPs act as distal transcriptional modifiers that function in immune cell types. Thus, mapping the transcriptional regulatory roles of GWAS hits will lead to better understanding of the genetic basis for RA and the roles of different cell types.

We combined the whole-genome sequences and blood transcriptomes of 377 RA patients and identified over 6,000 unique genes with expression quantitative trait loci (eQTLs). We demonstrated the quality of the eQTL calls through comparison to RA GWAS loci and eQTL from non-RA individuals. Then we integrated the eQTLs with 20 immune cell epigenome maps, RA GWAS risk loci, and adjustment for linkage disequilibrium to link immune cell enhancers that overlap RA risk loci to their target genes. We performed a focused analysis on primary monocytes, B cells, and T cells.

We identify several gene and cell type associations with relevance to RA including the identification of FCGR2B as possessing both intragenic and B cell enhancer regulatory GWAS hits. We show that our RA patient cohort derived eQTL network is more informative for studying RA than an eQTL network derived from a healthy cohort. While not experimentally validated here, these results can prioritize future confirmation experiments with the goal of elucidating the regulatory mechanisms behind RA genetic risk associations.

GENOME-WIDE ASSESSMENT OF THE CONTRIBUTION OF SHORT TANDEM REPEATS TO DE NOVO VARIATION

Thomas Willems[1,2,3], Melissa Gymrek[1,2,4,5], David Poznik[6,7], Chris Tyler-Smith[8], Yaniv Erlich[1,2,9,10]

[1]New York Genome Center, Computational Biology, New York, NY, [2]Whitehead Institute for Biomedical Research, Computational Biology, Cambridge, MA, [3]MIT, Computational and Systems Biology, Cambridge, MA, [4]MIT, Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, [5]Broad Institute , Program in Medical and Population Genetics, Cambridge, MA, [6]Stanford University, Program in Biomedical Informatics, Stanford, CA, [7]Stanford University, Department of Genetics, Stanford, CA, [8]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [9]Columbia University, Department of Computer Science, New York, NY, [10]Columbia University, Center for Computational Biology and Bioinformatics, New York, NY

Recent studies have obtained genome-wide estimates of the number of de novo mutations for nearly every class of genetic variant, ranging from SNPs and indels to copy number variants. Despite these results, no such estimate exists for short tandem repeats (STRs), highly repetitive loci that mutate orders of magnitude more rapidly than most genetic elements. Given the role of STRs in over 40 Mendelian diseases and increasing evidence of their involvement in a wide range of complex traits, we sought to address this issue. To this end, we developed MUTEA, a new algorithm that infers STR mutation rates from population-scale high-throughput sequencing data. After extensive validation, we applied MUTEA to data from the 1000 Genomes Project and the Simons Genome Diversity Project to estimate the mutation rates of 4,500 Y-chromosome STRs. We then leveraged these estimates to construct sequence-based predictors of STR mutation rates and applied them to STRs genome-wide. The resulting prediction suggests that the mutational load of STRs exceeds that of any other known variant class. We further assessed the reliability of this prediction by genotyping STRs in a deeply sequenced trio using gold standard datasets from the Illumina Platinum Genomes and Genome In a Bottle. This analysis uncovered hundreds of putative de novo STR mutations, 85% of which replicated in an orthogonal dataset. Overall, our results underscore the putative contribution of STRs to de novo genetic variation and have broad implications for medical and population genetics.

NANOFLUIDIC APPROACHES TO CHROMOSOME SYNTHESIS

Eamon M Winden[1], David C Schwartz[1], Samuel J Krerowicz[2]

[1]University of Wisconsin, Madison, Genetics, Madison, WI, [2]University of Wisconsin, Madison, Chemistry, Madison, WI

Nanofluidic approaches have opened up new routes for comprehensive analysis of large genomes. These same approaches may also lay the basis for the fabrication of mammalian chromosomes. Accordingly, our group has been developing new systems that will leverage the unique properties and advantages of nanoconfinement to power cell-free synthesis of large genomes.

CIS AND TRANS MECHANISMS DRIVING TF BINDING, CHROMATIN, AND GENE EXPRESSION EVOLUTION

Emily S Wong*[1], Bianca Schmitt*[2], Anastasiya Kazachenka[3], David Thybert[1], Aisling Redmond[2], Frances Connor[2], Tim Rayner[2], Christine Feig[2], Anne Ferguson-Smith[3], John C Marioni[1,2], Duncan T Odom[2,4], Paul Flicek[1,4]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom, [2]University of Cambridge, Cancer Research UK, Cambridge, United Kingdom, [3]University of Cambridge, Department of Genetics, Cambridge, United Kingdom, [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

* These authors contributed equally

To elucidate the regulatory mechanisms underlying transcription factor (TF) binding variations in mammals, we isolated cis-acting genetic sequences and their associated TF binding events by examining the allele-specific TF binding of three liver-specific TFs between genetic crosses of two inbred mouse strains. We distinguished between several regulatory categories at TF binding sites by comparing the allelic differences in F1 hybrids with parental measurements. For each TF, we classified ~15,000 binding regions to one of four regulatory categories: conserved/non-differential-acting, cis, trans, both cis and trans. In striking contrast to gene expression regulation, our results highlight the dominance of cis-driven mechanisms in TF occupancy variation. Further analysis reveals that cis-acting variants leads to local coordination in TF occupancies that decays with distance, and distal coordination which may be modulated by long-range chromatin contact. Causal mechanisms underlying TF occupancy change are closely linked to their mode of inheritance and are coordinated with the regulation of gene expression levels. Our findings support an integrated model whereby transcriptional regulatory divergence can be directed by a mixture of regulatory mechanisms driving individual TF variation near expressed genes.

IMPROVED NON-HUMAN PRIMATE REFERENCE GENOME FOR THE BIOMEDICAL MODEL RHESUS MACAQUE

Shwetha C Murali[1], Adam C English[1], Yi Han[1], Vanessa Vee[1], Yue Liu[1], Daniel S T Hughes[1], Muthuswamy Raveendran[1], Min Wang[1], Evette Skinner[1], Stephen Richards[1], Donna M Muzny[1], Robert B Norgren, Jr.[2], Richard A Gibbs[1], Jeffrey Rogers[1], Kim C Worley[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center and Department of Molecular and Human Genetics, Houston, TX, [2]University of Nebraska Medical Center, Department of Genetics, Cell Biology and Anatomy, Omaha, NE

As the foundation for genomic research, the quality of a draft genome sequence has a tremendous impact on the quality of downstream analyses. The Macaca mulatta (rhesus macaque) genome was initially published in 2007(1). In the decade since it was sequenced, the sequencing technologies have evolved and assembly methods have changed. We report here an improved macaque reference genome assembled using PacBio long read data and the PBJelly(2) gap filling method, methods that we have also used to improve the genomes of three other Old World monkeys and the mouse lemur.

We generated 9x Pacific Biosciences long-read data from the original reference animal. Starting with the MacaM_assembly_v7(3) generated from the original Sanger sequences and Illumina short read data from the same animal, we used the PBJelly method(2) to fill 44,067 gaps and merge contigs. The assembled sequences were placed on chromosomes with autosomes numbered from largest to smallest as in the first reference genome(1) and the remaining unplaced contigs from MacaM_assembly_v7(3) were appended. The finished quality BAC-based Y chromosome(4), accession CM003438, is also included as part of the new assembly, Mmul_8.0 (accession GCA_000772875.3). The genome assembly is 3.2 Gb total, with only 1,536 gaps remaining on chromosomes between scaffolds. The scaffold N50 is increased by 70% to 4.19 Mb. The contig N50 is 107 kb, double(3) or quadruple(1) the previous values.

This assembly added between 57 kb and 600 kb (average 308 kb) to each chromosome and reduced the number of scaffolds, with 4,046 fewer unplaced scaffolds and 735 fewer placed scaffolds. The new assembly has increased the total ungapped genome length by 21 Mb, adding 31.5 Mb to chromosomes, including representing the 11 Mb finished Y chromosome, and merging or removing 10 Mb of unplaced scaffolds.

The improved assembly enabled the RefSeq annotation of 33,594 genes and pseudogenes including 21,574 protein coding genes, 5,731 non-coding genes, with 54,240 fully supported mRNAs and 9,401 fully supported other RNAs.

Beyond this released version of the genome assembly, further, ongoing improvements include superscaffolding using Hi-C data, followed by additional PBJelly gap filling using PacBio data and manual curation. We will provide updates.

1. Rhesus Macaque Genome et al., Evolutionary and biomedical insights from the rhesus macaque genome. Science 316, 222-234 (2007).
2. A. C. English et al., Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7, e47768 (2012).
3. A. V. Zimin et al., A new rhesus macaque assembly and annotation for next-generation sequencing analyses. Biology direct 9, 20 (2014).
4. J. F. Hughes et al., Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature 483, 82-86 (2012).

← Back

SHEEP REFERENCE GENOME SEQUENCE UPDATES: TEXEL IMPROVEMENTS AND RAMBOUILLET PROGRESS

Yue Liu[1], Shwetha C Murali[1], R Alan Harris[1], Adam C English[1], Xiang Qin[1], Evette Skinner[1], Mike Heaton[2], Timothy Smith[2], Brian Dalrymple[3], James Kijas[3], Noelle E Cockett[4], Eric Boerwinkle[1], Donna M Muzny[1], Richard A Gibbs[1], Kim C Worley[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center and Department of Molecular and Human Genetics, Houston, TX, [2]Roman L. Hruska US Meat Animal Research Center, Clay Center, NE, [3]CSIRO, Queensland, Australia, [4]Utah State University, Logan, UT

The latest methods for producing and analyzing long reads are finally improving the quality of draft genome assemblies beyond the methods employed for early Sanger sequencing. We have applied these methods to improve the existing genome of the Texel sheep and are in the process of producing a de novo assembly from a single Rambouillet ewe. The Texel ram Pacific Biosciences data was used with the PBJelly software(1) to improve Oar_v3.1. The improved assembly, Oar_v4.0 (GCA_000298735.2) that has been submitted to GenBank has improved contiguity and genomic representation. The genome is highly contiguous with a contig N50 of 150kb and scaffold N50 of 100Mb.

Further efforts to produce a high quality reference genome have shifted focus to the Rambouillet breed where we have completed sequence production using the Pacific Biosciences technology, producing 200 Gb of sequence with subread length of 12.6 kb N50 length and 8.9 kb mean length. Error correction of the reads using the Pacific Biosciences data is in progress. Initial test assemblies demonstrate that the contiguity of this assembly will exceed that of Oar_v4.0. Sample collection from the donor animal is planned for further assays including Pacific Biosciences IsoSeq for RNA sequencing and as evidence for annotation and Hi-C proximity ligation sequencing for genome scaffolding.

We will discuss ongoing research with other methods to approach finished quality genomes without using traditional expensive and manually intensive finishing efforts.

1. A. C. English et al., Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7, e47768 (2012).

NGS-SWIFT: A CLOUD-BASED VARIANT ANALYSIS FRAMEWORK USING CONTROL-ACCESSED SEQUENCING DATA FROM DBGAP/SRA

Chunlin Xiao, Eugene Yaschenko, Stephen Sherry

NCBI, NIH, Bethesda, MD

Genetic variation analysis plays an important role in elucidating the causes of various human diseases. The drastically reduced costs of genome sequencing driven by next generation sequence technologies now make it possible to analyze genetic variations with hundreds or thousands of samples simultaneously, but with the cost of ever increasing local storage requirements. The tera- and peta-byte scale footprint for sequence data imposes significant technical challenges for data management and analysis, including the tasks of collection, storage, transfer, sharing, and privacy protection. Currently, each analysis group must download all the relevant sequence data into a local file system before variation analysis is initiated. This heavy-weight transaction not only slows down the pace of the analysis, but also creates financial burdens for researchers due to the cost of hardware and time required to transfer the data over typical academic internet connections. To overcome such limitations and explore the feasibility of analyzing control-accessed sequencing data in cloud environment while maintaining data privacy and security, here we introduce a cloud-based analysis framework that facilitates variation analysis using direct access to the NCBI Sequence Read Archive through NCBI SRA Toolkit, which allows the users to programmatically access data housed within SRA with encryption and decryption capabilities and converts it from the SRA format to the desired format for data analysis. A customized machine image (ngs-swift) with preconfigured tools, including NCBI SRA Toolkit and NGS Software Development Kit, and resources essential for variant analysis has been created for instantiating an EC2 instance or instance cluster on Amazon cloud. Performance of this framework has been evaluated using dbGaP study phs000710.v1.p1 (1000Genome Dataset in dbGaP, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000710.v1.p1), and compared with that from traditional analysis pipeline, and security handling in cloud environment when dealing with control-accessed sequence data has been addressed. We demonstrate that with this framework, it is cost effective to make variant calls without first transferring the entire set of aligned sequence data into a local storage environment, thereby accelerating variant discovery using control-accessed sequencing data.

INTEGRATING LONG-RANGE INTERACTIONS IN EPIGENOMIC COMPARISONS ACROSS GROUPS OF CELL AND TISSUE SAMPLES.

Angela Yen[1,2], Manolis Kellis[1,2]

[1]MIT Computer Science and AI Laboratory, Electrical Engineering and Computer Science Department, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA

Epigenomic datasets provide critical information about the dynamic role of chromatin states in gene regulation, while long-range interactions between genes and regulatory regions can mechanistically explain the underlying reason for epigenomic changes. However, the combination of long-range interactions and chromatin state differences has not been systematically studied across groups of human tissues and cell types.

Here, we integrate long-range interactions in the context of epigenomic comparisons by extending our previous statistical framework for comparing sets of epigenomics. This framework, ChromDiff, enables the identification of epigenomic regions and signatures that differentiate groups of samples, even in the absence of gene expression changes. However, given the challenge of linking regulatory regions to their target genes, we had previously applied ChromDiff only to genic regions directly. Here, we leverage regulatory links of promoters, enhancers, and regions with accessible chromatin, to predict the regulatory regions driving cellular differences.

We find that dynamic regulatory regions reveal meaningful biological differences between cell types. For example, enhancer regions that differ between female and male samples are significantly linked to genes that escape X chromosome inactivation, suggesting the importance of epigenomic regulation in the mechanism behind the dynamic process of escape from X chromosome inactivation, which varies across tissues. These specific examples of biologically-meaningful cis- and trans-regulatory region alterations enable the discovery of relevant genes even when significant expression or genic chromatin state changes can not be detected directly. Our results emphasize the importance of long-range interactions and show that regulatory regions provide a rich lens for the identification of robust signatures distinguishing the chromatin state landscape of classes of samples. Our expanded methodology is general and provides a powerful new tool for integrating regulatory links in the study of epigenomic differences at the genome scale.

ROBUST TRANSCRIPTOME-WIDE DISCOVERY OF RNA BINDING PROTEIN BINDING SITES WITH ENHANCED CLIP (eCLIP) AND EVALUATION OF IMPACT OF NATURAL AND DISEASE-CAUSING VARIANTS ON RNA BINDING

Eric V Nostrand[1], Gabriel A Pratt[1], Alexander A Shishkin[2], Chelsea Gelboin-Burkhart[1], Mark Fang[1], Balaji Sundararaman[1], Steven Blue[1], Thai Nguyen[1], Christine Surka[2], Keri Elkins[1], Rebecca Stanton[1], Frank Rigo[3], Mitchell Guttman[2], Eugene Yeo[1]

[1]University of California San Diego, Cellular and Molecular Medicine, La Jolla, CA, [2]Caltech, Division of Biology and Biological Engineering, Pasadena, CA, [3]Ionis Pharmaceuticals, R&D, La Jolla, CA

As RNA binding proteins (RBPs) play essential roles in cellular physiology by interacting with target RNAs, binding site identification by UV-crosslinking and immunoprecipitation (CLIP) of ribonucleoprotein complexes is critical to understanding RBP function. However, current CLIP protocols are technically demanding and yield low complexity libraries with high experimental failure rates. We have developed an enhanced CLIP (eCLIP) protocol that decreases requisite amplification by ~1,000-fold, decreasing discarded PCR duplicate reads by ~60% while maintaining single-nucleotide binding resolution. By simplifying the generation of paired IgG and size-matched input controls, eCLIP improves specificity in discovery of authentic binding sites. We generated 102 eCLIP experiments for 74 diverse RBPs in HepG2 and K562 cells (available at https://www.encodeproject.org), demonstrating that eCLIP enables large-scale and robust profiling, with amplification and sample requirements similar to ChIP-seq. eCLIP enables integrative analysis of diverse RBPs to reveal factor-specific profiles, common artifacts for CLIP and RNA-centric perspectives of RBP activity. Currently, we are evaluating the impact of natural and disease-causing variants on RBP binding using our eCLIP datasets.

UNBIASED ESTIMATION OF HERITABILITY BY RELATEDNESS DISEQUILIBRIUM REGRESSION REVEALS OVERESTIMATION OF HERITABILITY BY TWIN STUDIES

Alexander I Young[1], Michael L Frigge[2], Kari Stefansson[2], Augustine Kong[2]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]deCODE Genetics, Sturlugata 8, Reykjavik, Iceland

Heritability is a fundamental quality in genetics that measures the proportion of phenotypic variance that is explained by the additive effects of inherited genetic sequence variants, independent of their correlation with the family or wider environment. It is the parameter against which the success of genome wide association studies has been measured, leading to the so-called 'problem of missing heritability'. Most estimates of heritability in humans come from twin and family studies, which have to make assumptions about the sharing of environmental effects between relatives. There has been a longstanding controversy about the validity of these assumptions and the bias of the estimates of heritability. Approaches using genomic data have promised to give independent estimates of heritability by looking at the observed sharing of sequence variants, but these approaches have also made unproven assumptions about the sharing of environmental effects. We introduce a heritability estimation method, relatedness disequilibrium regression, which makes no assumptions about the sharing of environmental effects. The method achieves this by isolating the variation in relatedness due to one generation of random Mendelian segregation, which is uncorrelated with environmental effects. It involves jointly fitting three relatedness matrices that capture the heritability, the phenotypic variance captured by parental genetic variation, and the phenotypic variance captured by genetic sharing between parents and offspring. These matrices are equal in expectation, so genetic information on both parents of each individual in the sample is necessary. In contrast to other genomic methods of estimating heritability, we do not have to restrict ourselves to distantly related individuals. We use a sample of ~55,000 Icelanders with both parents genotyped, and we estimate the three relatedness matrices using inferred sharing of segments identical-by-descent. In addition to theoretical arguments for unbiasedness, we show in simulations on real data that our method, in contrast to others, is unbiased by complex environmental confounding. We estimate heritability for 27 traits, including height (57%; 95% confidence interval: [49%,66%]) and body mass index (26%, 95% confidence interval: [15%,38%]), finding substantially lower estimates than from twin studies for most, but not all, traits. Our variance component estimates also give insight into the influence of genetic variation in parents on traits through the family environment.

A ROLE IN PROGRAMMED DNA DELETION FOR THE SECOND DOMESTICATED PIGGYBAC TRANSPOSASE TPB1 IN *TETRAHYMENA THERMOPHILA*

Chao-Yin Cheng[1], Janet M Young[3], Chih-Yi Lin[1,2], Harmit S Malik[3,4], Meng-Chao Yao[1,2]

[1]Academia Sinica, Institute of Molecular Biology, Taipei, Taiwan, [2]National Taiwan University, Genome and Systems Biology Degree Program, Taipei, Taiwan, [3]Fred Hutchinson Cancer Research Center, Division of Basic Sciences, Seattle, WA, [4]Fred Hutchinson Cancer Research Center, Howard Hughes Medical Institute, Seattle, WA

In the single-celled ciliate *Tetrahymena thermophila*, programmed DNA rearrangements eliminate thousands of deletion sequences (internal elimination sequences, or IESs) from the germline genome during conjugation and subsequent development of a new somatic nucleus. Previous studies showed that a domesticated piggyBac transposase, TPB2, likely carries out the DNA cutting step to delete the RNA-marked chromatin of most or all IESs, and is an essential gene.

We now investigate a second domesticated transposase, TPB1, which is not essential for growth. Engineered strains deficient in TPB1 produced abnormal, slow-growing progeny cells with giant vacuoles, resulting from defects suffered during conjugation. We used deep sequencing to show that even wild-type cells display a surprising level of variation in successful IES removal. We also show that TPB1-deficient cells entirely fail to eliminate a small but important set of 16 IESs during conjugation. PCR analysis of some of these IESs in additional samples supported their dependence on TPB1. Thus, TPB1 appears to act during conjugation in the deletion of a small subset of internal eliminated sequences. Unlike most IESs in the genome, which appear to have no characteristic sequence signature, the 16 TPB1-dependent IESs have a characteristic sequence motif at their edges that resembles inverted terminal repeats. The 16 TPB1-dependent IESs include some that are at or near genes that are expressed during growth. We are analyzing the functions of these genes to determine their contributions to the abnormal phenotypes of TPB1-deficient strains.

CAUSAL VARIANTS IN METABOLITE QUANTITATIVE TRAIT LOCI

Noha A Yousri*[1,2], Khalid A Fakhro*[3,4], Amal Robay[3], Juan L Rodriguez-Flores[5], Ronald G Crystal[5], Karsten Suhre[1]

[1]Weill Cornell Medical College-Qatar, Bioinformatics Core, Physiology and Biophysics, Doha, Qatar, [2]Alexandria University, Computer and Systems Engineering, Alexandria , Egypt, [3]Weill Cornell Medical College-Qatar, Genetic Medicine, Doha, Qatar, [4]Sidra Medical and Research Center, Division of Translational Medicine, Doha, Qatar, [5]Weill Cornell Medical College, Genetic Medicine, New York, NY
*Equally contributing authors

A recent study by [1] identified 145 loci associated with metabolites (metabolite ratios), using a European cohort of more than 7000 individuals, and 400 metabolites. Investigating the variants underpinning Genetically Determined Metabotypes (GDMs) requires searching the whole genome, for all potential causal variants. However, previous Metabolomics-GWASs (mGWASs) have only considered imputed array data for mGWASs, possibly missing the variants of interest. Exome sequencing provides a compromise between the expensive whole genome sequencing, and the cheaper genotype arrays for targeting those variants. In a recent study [2], the researchers use exome sequencing for fine tuning of GDMs identified using array data, yet using a small cohort of metabolites from NMR technology. Our study aims at investigating potential causal variants associated with GDMs identified by [1] using exome data. The first step is to replicate the top associations from their study and fine tune them using exonic variants. 550 serum samples were available at the time of the study, that were collected in Qatar, from healthy Qataris who are third generation. The samples were sequenced on Illumina HiSeq (100bp PE) and variants were called using a standard GATK best practices workflow. Serum samples were sent to Metabolon for measuring metabolic concentrations of 1303 known and unknown metabolites. Quality control criteria for both exome data and metabolites were used to filter them; MAF> 0.05, phwe > 10^{-6}, and genotype call rate of more than 98% were applied to exome data, and metabolites with more than 20% missing values were used, after log scaling and normalization. GenABEL package in R was used for association analysis, where all covariates and kinship were corrected for. SNPs that were in linkage disequilibrium of more than 0.5 with a target locus (locus identified in [1]) were identified based on European cohort, and then overlaid on the exome data. Exome SNPs that matched the SNPs in highest LD with a target locus were used for replication. Exact metabolite matching between metabolites used in [1] and the new Metabolon platform was done to find replicable associations. 23% of a total of 76 replicable associations (from all of the 145 loci identified in [1]), were replicated at a Bonferroni corrected p-value. For those, we identified all potential exome variants that are in LD>=0.5 with the target locus.

[1] Shin S, et. al, “An Atlas of Genetic Influences on Human Blood Metabolites”. Nature Genetics. June 2014.
[2] Dmerikan et. al, “Insight in Genome-Wide Association of Metabolite Quantitative Traits by Exome Sequence Analysis”. Plos Genetics, January 2015.

← Back

INSIGHTS INTO THE PERFORMANCE OF WHOLE-EXOME SEQUENCING TECHNOLOGIES

Yao Yu, Hao Hu, Jerry Fowler, Yuanqing Ye, Michelle Hildebrandt, Hua Zhao, Paul Scheet, Xifeng Wu, Chad D Huff

The University of Texas MD Anderson Cancer Center, Department of Epidemiology, Houston, TX

A number of new technologies have been developed to support whole-exome capture and sequencing, presenting researchers with many configuration options when designing whole-exome sequencing (WES) experiments. In addition, the availability of the new Illumina HiSeq 3000/4000 offers the potential for a sharp increase in sequencing efficiency, but the performance of WES products on this instrument remains largely unknown. To evaluate the relative performance of available technologies, we conducted a series of WES experiments varying the following design options: genome fragmentation approach (sonication vs. enzymatic), average DNA fragment length, library preparation vendor (Kapa vs. Agilent), exome capture product (Agilent Clinical Research Exome vs. Nimblegen MedExome), and whether to PCR amplify libraries prior to target capture. All sequencing was performed on an Illumina HiSeq 3000/4000 using 2x150bp paired-end (PE) reads with eight samples pooled per lane, which on average generated ~45 million read pairs per sample. Our results demonstrate that under optimal conditions, the Clinical Research Exome (CRE) and MedExome (ME) products provided comparable exome coverage, on average covering of 94.5% and 92.6% of coding bases in RefSeq coding sequencing (CDS) at ≥20X for CRE and ME, respectively. However, CRE produced more reliable coverage than ME, ranging from 92.4% to 95.4% CDS coverage at ≥20X for CRE compared to 88.1% to 96.4% coverage for ME across 16 samples. We also observed that, CRE was more appropriate for deep sequencing studies, such as those used for the detection of rare somatic events in cancer, with 31.4% of CDS bases covered at ≥100X for CRE vs. 7.6% for ME. Further, we found that sonication fragmentation using Covaris-sheared DNA generated 20% more unique fragments compared to the enzymatic fragmentation (KAPA HyperPlus). Additionally, we observed that PCR-free library preparation using KAPA Hyper kits resulted in a 3.4% increase in CDS bases covered at ≥20X. We also observed that smaller DNA fragment sizes of 200bp average length produced greater coverage than larger fragment lengths due to increased on target coverage, despite the need to trim adapters and merge overlapping reads. To support this approach, we present a new framework for quality control and reference alignment for small size libraries (~200bp) with tail overlapped read. Finally, we report the sensitivity and specificity of genotype calling results of each library preparation and target capture approach.

REAL-TIME PERSON IDENTIFICATION USING NOISY ERROR-PRONE DNA SEQUENCING DATA AND INCOMPLETE DATABASES.

Sophie Zaaijer*[1,2], Robert Piccone*[2], Daniel Speyer[2], Yaniv Erlich[1,2,3]

[1]New York Genome Center, New York , NY, [2]Columbia University, Department of Computer Science, New York, NY, [3]Columbia University, Department of Systems Biology, New York, NY

Identifying DNA samples plays a major role in a range of law enforcement and security activities. The Holy Grail being: real-time person identification. This technique can transform forensics, mass disaster control, fighting human trafficking, and provide the basis of DNA-based access control.

We will present a strategy for real-time person identification. We applied a new method using Oxford Nanopore MinION sequencer. This device has three important facets for our invention: real time data generation, long DNA reads and it is a small hand-held portable device. However, MinION poses a significant challenge since the current average error rate of MinION reads is ~10-15%, whereas the difference between two humans is 0.1%, which is about 100 times smaller than the error rate. Our new tactics involves the sample preparation for shotgun DNA sequencing using the standard MinION protocol. The comparison between the MinION reads and the incomplete database considers only a subset of genomic positions with known common polymorphisms based on open databases such as the 1000Genomes. We will present the methodology which enabled us to identify a person with 100% probability using a simulated setting that corresponds to a database of 10^7 people and only ~600 MinION reads, which we currently can obtain in 12 minutes.

ACCURATE PROMOTER AND ENHANCER IDENTIFICATION IN 127 ENCODE AND ROADMAP EPIGENOMICS CELL TYPES AND TISSUES BY GENOSTAN

Benedikt Zacher[1], Margaux Michel[3], Bjoern Schwalb[3], Patrick Cramer[3], Achim Tresch[2], Julien Gagneur[1]

[1]Ludwig-Maximilans-University, Gene Center, Munich, Germany, [2]University of Cologne, Department of Biology, Cologne, Germany, [3]Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Goettingen, Germany

Accurate maps of regulatory elements such as promoters and enhancers are of great importance to understand transcription regulation. These elements are marked by specific combinations of post-translational histone tail modifications, the 'chromatin state'. Unsupervised machine learning approaches were developed for the genome-wide annotation of chromatin states from Next-Generation-Sequening data. However current state-of-art approaches are limited because they either require extensive preprocessing of the data or make unrealistic assumptions of distribution of the data. Here we propose GenoSTAN (Genomic STate ANnotation), a hidden Markov model based method which overcomes these limitations by modeling sequencing data without the need of data transformation using count distributions (Poisson-lognormal and negative binomial). We apply GenoSTAN to provide an improved chromatin state annotation in one K562 ENCODE dataset and 127 cell types and tissues from the ENCODE and Roadmap Epigenomics projects. GenoSTAN recovers many distinctive features of promoters and enhancers and identifies them with higher accuracy than three other state-of-the-art methods. We further show that promoters and enhancers have fundamentally different transcription factor regulatory landscapes. Moreover promoters and enhancers predicted by GenoSTAN show higher enrichment of complex trait-associated SNPs than predictions from previous studies. Thus, GenoSTAN together with the chromatin state annotation inferred in this study provide a useful tool and resource for future research in (epi-)genomics.

THE GENETIC BASIS OF EVOLUTIONARY TRANSITIONS IN EARLY DEVELOPMENT

Christina Zakas, Matthew Rockman

New York University, Biology, New York, NY

Phenotypic evolution in animals is constrained by the mechanics of early development. How do major transitions in development occur? Historically, efforts to address this question have been limited to comparative methods. The polychaete annelid *Streblospio benedicti* provides a unique opportunity to use forward genetics to experimentally dissect a major transition in animal development. *S. benedicti* is ideal because it produces two distinct offspring types that differ in egg size, early development, and larval morphology. *S. benedicti* is thus a genetic model for the evolutionarily common transition between indirect and direct development. Using genetic crosses between these types, I constructed the first annelid genetic map, which reveals the distribution of genetic factors affecting a suite of genetically separable developmental phenotypes. Because early development is strongly influenced by maternal effects, my cross design disentangles maternal and zygotic genetic effects and shows that a transition from indirect to direct development requires contributions from both the zygotic and maternal genome; an increase in egg size alone is not sufficient to change development mode.

CIS-REGULATORY ANNOTATION OF GENOMES IN ENSEMBL

Daniel R Zerbino, Thomas Juettemann, Steven P Wilder, Anne Parker, Michael Nuhn, Ilias Lavidas, Avik Datta, Ernesto Lowy Gallego, Kieron Taylor, Magali Ruffier, Andrew Yates, Laura Clarke, Paul R Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom

Ensembl is one of the world's leading sources of information on the structure and function of the genome. It already provides an up-to-date, comprehensive and consistent database that brings together genome sequences, genes, non-coding RNAs, known variants, etc.

Recently we re-designed and greatly expanded our annotation of regulatory elements in the genome. The Ensembl Regulatory Build synthesises public epigenomic datasets produced by large-scale projects such as ENCODE, Roadmap Epigenomics or BLUEPRINT. We process them through a unified pipeline and make them available through a single interface. We also define functionally active regions across 38 human cell types (on both the GRCh37 and GRCh38 assemblies) and 6 mouse cell types, assigning them a function wherever possible. We are currently expanding our annotation to more cell types, and hopefully soon to more species in collaboration with the FAANG consortium.

Regulatory elements are of interest because of their action on genes; therefore we are simultaneously developing a database of cis-regulatory interactions attaching them to their target genes. Currently, two main approaches are being used to detect these interactions: genetics (e.g. eQTLs) and chromatin conformation (e.g. Hi-C). We have developed new technologies to store and display these datasets, using in particular HDF5 indexing for fast retrieval. This technology allows us to store all the GTEx summary eQTL data, as opposed to only significant correlations, thus avoiding interval censoring. Over the next year, we will be integrating more eQTL and Promoter Capture Hi-C datasets, organized by tissue.

Using our RESTful API, it will soon be possible to retrieve this data simply and efficiently for any gene, variant or region, along with all other Ensembl annotations such as LD calculations from the 1000 Genomes dataset Phase 3, conservation scores, etc. It will thus be possible to quickly develop advanced functional analysis pipelines without having to download or process massive data files.

Finally, we are also developing high performance tools for basic research in epigenomics. The GenomeStats browser allows users to remotely compute statistics on the BLUEPRINT datasets, without downloading data or software. Thanks to the underlying WiggleTools library, it is capable of processing dozens of files simultaneously and efficiently. This means that even naïve users can try out complex hypothesis testing without getting slowed down by file management issues.

UNCOVERING THE TRANSCRIPTOMIC AND EPIGENOMIC LANDSCAPE OF NICOTINIC RECEPTOR GENES IN HUMAN NON-NEURONAL TISSUES

Bo Zhang[1], Pamela Madden3[2], Ting Wang[3]

[1]Washington University School of Medicine, Department of Developmental Biology, St.Louis, MO, [2]Washington University School of Medicine, Department of Psychiatry, St.Louis, MO, [3]Washington University School of Medicine, Department of Genetics, St.Louis, MO

Nicotinic acetylcholine receptors (nAChRs) play an important role in cellular physiology and human nicotine dependence. However, the tissue specificity of nAChRs gene expression and their regulation remain unexplored. We integrated data from multiple genomics consortiums, including ENCODE, Roadmap Epigenomics, GTEx, and Fantom, to define the transcriptomic and epigenomic landscape of nAChRs across human tissues. We found that many important nAChRs exhibited strong non-neuronal tissue-specific expression patterns. CHRNA3, CHRNA5, and CHRNB4 were highly expressed in human colon and small intestine, and CHRNA4 was highly expressed in human liver. We identified a novel liver-specific alternative transcription start site (TSS) of CHRNA4, which was specifically transcribed in hepatocytes. Our findings suggest that CHRNA4 has distinct transcriptional regulatory mechanisms in human liver and brain, and such tissue-specific expression pattern is evolutionarily conserved in mouse. Finally, we found that liver-specific CHRNA4 transcription was highly correlated with the expression of CYP2A6, key enzyme of nicotine metabolism.

TISSUE-SPECIFIC ROLE OF SOMATIC MUTATIONS IN KINASE-SUBSTRATE PHOSPHORYLATION NETWORK

Junfei Zhao[1], Feixiong Cheng[2,3], Zhongming Zhao[1]

[1]UT Health Science Center at Houston, School of Biomedical Informatics, Houston, TX, [2]Harvard Medical School, Center for Cancer Systems Biology , Boston, MA, [3]Northeastern University, Center for Complex Networks Research, Boston, MA

A large number of somatic mutations have been accumulated through several large-scale cancer genome sequencing projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium projects. Phosphorylation-dependent signaling is fundamental in cellular physiology and its dysfunction plays a critical role in tumorigenesis. However, a systematic investigation of the somatic mutations in phosphorylation networks and pathways governing cell signaling has not been done yet. The rapid advancements of technologies (e.g. mass spectrometry) have made a wealth of kinase-substrate interaction data available. This provides us with a unique opportunity to interrogate somatic mutations in the context of protein functional features (i.e. phosphorylation sites) to determine their pathophysiological roles in cancer and to prioritize potentially druggable mutations that may mediate drug binding at the atom resolution. In this study, we incorporated the somatic missense mutations into kinase-substrate network to detect significantly mutated phosphorylation sites and kinase-substrate interaction modules. By analyzing 746,631 missense mutations from 4,997 tumors across 16 cancer types from TCGA, we found that more than 50% tumor samples harbored phosphorylation-associated single nucleotide variants (SNVs), and further identified 113 proteins that harbored significantly mutated phosphorylation sites (adjusted p-value < 0.05). Our network module based analysis revealed diverse recurrently mutated kinase-substrate subnetworks across 16 cancer types, implying pan-cancer heterogeneity and tissue-specific role of somatic mutations in kinase-substrate phosphorylation network.

← Back

VALOR: A HIGH-SPEED VALIDATION APPROACH FOR STRUCTURAL VARIATION USING LONG-READ SEQUENCING.

Xuefang Zhao[1], Ryan E Mills[1,2]

[1]University of Michigan Medical School, Department of Computational Medicine & Bioinformatics, Ann Arbor, MI, [2]University of Michigan Medical School, Department of Human Genetics, Ann Arbor, MI

Structural variants (SVs) are one of the major forms of genetic variation in humans and have been revealed to play important roles in various diseases including cancers and neurological disorders. Various approaches have been developed and applied to paired-end sequencing to detect SVs in whole genomes, however individual algorithms often exhibit complementary strengths. Thus, investigators typically apply and compare multiple algorithms to their samples and design their own selection strategy according to the sensitivity and specificity requirements of their research, while using orthogonal evidence from each approach as the only evidence that an actual structural rearrangement is present. The emergence of long read sequencing technology, eg. Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio), can provide direct evidence for the presence of an SV. Current strategies make use of de novo assembly to create large contigs that can be cross-referenced with a putative SV using manual inspection of the subsequent recurrence (dot) plot. Despite the high accuracy, this can be time-consuming and inefficient for the high throughput validation of large sets of SVs. Here, we present a high-speed long read based validator, VaLoR, that scores each SV prediction by autonomously analyzing the recurrence of windows within a local read against the reference genome in both their original and rearranged format according to the prediction. A positive score of each read on the altered reference, normalized against the score of the read on the original reference, supports the predicted structure. A baseline model is constructed as well by interrogating the reference sequence against itself at the query location. We show that our approach is able to quickly and accurately distinguish true from false positive predictions of both simple and complex SVs and is also able to assess the breakpoint accuracy of individual algorithms.

INVESTIGATING REGULATORY ROLES OF ASSOCIATION VARIANTS IN THREE LUNG CANCER SUBTYPES

Timothy O'Brien[1,2], Peilin Jia[2,3], Zhongming Zhao[2,3]

[1]Vanderbilt University, Vanderbilt Genetics Institute, Nashville, TN, [2]Vanderbilt University, Department of Biomedical Informatics, Nashville, TN, [3]University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX

Investigating regulatory roles of genetic variants has become an important task to decipher their functions in complex disease. In this study, we investigated the regulatory roles in three major lung cancer subtypes: small cell lung cancer (SCLC), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC). These three subtypes have distinct physiological, clinical, and genetic signatures. Many common variants (SNPs) have been found to be associated with lung cancer subtypes; however, they are typically located in the non-coding regions and their functional roles in each subtype remain largely unknown. We hypothesized that these non-coding SNPs function in common or distinct regulatory mechanisms for each subtype. To test this hypothesis, we investigated the regulatory roles of common SNPs having association with each lung cancer subtype from genome-wide association studies (GWAS) of European samples. We obtained these SNPs by applying association test $p < 10^{-5}$ in lung cancer GWAS. Then, we expanded this list to include all SNPs in linkage disequilibrium (LD) with $r^2 > 0.8$ using the 1000 Genomes Project (1KGP) data. This expansion resulted in 1427, 1788, and 1059 SNPs for LUAD, LUSC, and SCLC, respectively. Next, we searched for enrichment of the lung cancer SNPs in expression quantitative trait loci (eQTLs) for lung and an additional eight tissues using the data from the Genotype-Tissue Expression (GTEx) Project. While several SNPs were found to be eQTLs, there was no statistical enrichment compared to random SNPs selected from the GWAS chips. This lack of enrichment in eQTLs was confirmed in another lung tissue-specific eQTL dataset with a larger sample size. We further searched for regulatory enrichment using the data from five lung-related tissues and cell lines generated in the Roadmap Epigenomics Project, as well as the enhancer data from FANTOM5 and IM-PET. We found very distinct regulatory signatures that differed by lung cancer subtype and lung tissue/cell type, such as SNPs within enhancer regions in LUSC and SCLC in adult lung tissue, but not in LUAD. We also found similarities among subtypes such as the highest proportion of SNPs located in an active state identified by transcription signatures at the 3' and 5' end of genes in the A549 lung cancer cell line. Our preliminary results suggested that three major subtypes of lung cancer might share both common and distinct regulatory roles acted by the association-related common variants.

INTEGRATIVE ANALYSIS OF MULTI "OMICS" DATA IDENTIFIES FUNCTIONAL MEDIATORS AS INTERVENTION POINTS FOR GLOBAL PHENOTYPES

Chenchen Zhu[1], Christopher S Hughes*[2], Michelle Nguyen[3], Lars M Steinmetz[1,3]

[1]European Molecular Biology Laboratory,, Genome Biology, Heidelberg, Germany, [2]British Columbia Cancer Research Centre, Vancouver, Canada, [3]Stanford University, Department of Genetics, Stanford, CA

A major challenge in systems genetics concerns the identification of causal intermediates underlying the genotype to phenotype path. This is a challenge because we lack a clear understanding of which genetic variants affect complex traits and how those effects are exerted at the molecular level. To unravel the chain of underlying biological events, intermediate traits such as gene expression levels have been used to establish links between genomic variation and global phenotype. As products of cellular pathways, proteins and metabolites represent promising candidates for unraveling intermediate cellular processes and reflecting the physiological state of a cell.

In our pilot study, we have collected genome-wide gene expression profiles of a yeast cross between a laboratory strain and a clinical isolate of Saccharomyces cerevisiae in different environmental conditions. A statistical method has been developed to infer causal intermediate transcripts by exploiting environmental perturbations, thus distinguishing them from correlative effects downstream of phenotype.

Currently, we are systematically collecting dynamic profiles of proteins and metabolites for the same panel of segregants. By taking advantage of the defined genetic perturbations presented in this population, the genetic regulation of each molecular layer is dissected using quantitative trait locus (QTL) mapping. We extend the existing causal inference method to the multidimensional dataset and exploit the relationships between the different molecular layers to learn biological principles that guide the conditioning of complex phenotypes.

We present a new method to identify molecular signatures that are predictive for genes with a causal role in phenotype, integrating the proteomics and metabolomics data with existing genotypes, expression profiles, and growth rates for these strains. The integration of the proteomics and metabolomics data will permit the study of how genetic variations and transcript abundance impact cellular states. With this large compendium of datasets, we increase the sensitivity to detect causal networks at different levels. Beyond proposing a new route towards identifying specific molecular targets from high- throughput "omics" data, our results contribute to defining better models for how genotypic variation leads to phenotypic variation among closely related individuals.

← Back

GENE SIMILARITY NETWORK REVEALS SUB-POPULATIONS OF CELLS IN SINGLE-CELL RNA-SEQ DATA

Bo Wang[1], Jesse Zhang[2], Junjie Zhu[2], Serafim Batzoglou[1]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Electrical Engineering, Stanford, CA

Single-cell RNA-seq is rapidly becoming a powerful method for studying cell-to-cell variation at the transcriptional level. Due to experimental limitations, single-cell datasets suffer from technical noise in addition to variation caused by culture conditions and systematic biological effects. Traditional methods such as hierarchical clustering struggle with noisy high dimensional single-cell RNA-seq data. Some studies have tackled this problem by selecting genes based on prior knowledge and biological intuition either at the beginning of the experiment or when applying clustering methods. This approach, however, introduces bias into the computational analysis and therefore can fail to capture subtle relationships in the data that would only be identifiable using the removed genes. This work introduces Sargen (single-cell analysis using robust PCA on gene similarity network), a novel method that finds subpopulations of cells in a single-cell RNA-seq data by learning subtle gene-gene relationships in the form of a gene similarity network. Using a deep autoencoder, Sargen learns the network before using it to map the data into a reduced dimensional space. The core strategy in the method revolves around formulating a non-convex optimization problem that attempts to find a low-rank, network-aligned approximation of the input data. Sargen solves a convex relaxation of the problem and uses the solution to generate cell subpopulations. We analyzed recent single-cell datasets (Buettner et al., 2015, Kolodziejczyk et al., 2015) influenced by cell cycle effects and culture conditions, and used Sargen to successfully identifies cell subpopulations, verified by both gene ontology and prior information pertaining to the experiments. We also demonstrate from these experiments that Sargen is able to automatically select relevant genes without prior knowledge of specific gene functions in identifying cell subpopulations. Finally, we empirically show that the framework introduced by Sargen significantly improves many existing clustering (e.g., K-means and hierarchical clustering) and visualization methods (e.g., t-distributed stochastic neighbor embedding and principle components analysis), on single-cell RNA-seq data.

References:

- F. Buettner, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol., vol. 33, no. 2, pp. 155–160, 2015.
- A. A. Kolodziejczyk, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell, vol. 17, no. 4, pp. 471–485, 2015.

← Back

GENOMIC SIGNATURES OF MOSQUITO ADAPTATION TO THE AMINO ACID CONTENT OF HUMAN BLOOD

Zhilei Zhao, Carolyn S McBride

Princeton University, Princeton Neuroscience Institute and Department of Ecology & Evolutionary Biology, Princeton, NJ

Female mosquitoes use the protein they acquire in vertebrate blood to synthesize eggs. This presents a challenge, because vertebrate blood protein is eighty percent hemoglobin and hemoglobin tends to be low in isoleucine. Human hemoglobin, for example, completely lacks this essential amino acid, and several mosquito species will lay up to twice as many eggs when fed human blood supplemented with isoleucine than when fed human blood alone. How do mosquitoes cope with this challenge – especially important disease vectors that specialize in biting humans? We are addressing these questions using whole genome data from 750 individuals of the African malaria mosquitoes *Anopheles gambiae* and *coluzzii* made available through the Ag1000G project. Interestingly, mutations resulting in the loss of an isoleucine residue show elevated frequencies, while mutations resulting in gain of isoleucine show reduced frequencies. As predicted, these patterns are most striking in protein coding genes that are upregulated after a blood meal, the time in the life cycle when eggs are produced. A few other amino acids also show levels of asymmetric gain and loss not observed in animals that consume more heterogenous protein resources. We hypothesize that the amino acid composition of the proteome of this important disease vector experiences pervasive selection to increase the efficient use of human blood protein, making this an extraordinarily complex trait.

WHOLE GENOME SEQUENCE VARIANTS INFLUENCE MULTIPLE AMINO ACIDS LEVELS

Bing Yu[1], Elena Feofanova[1], Donna Muzny[2], Alanna C Morrison[1], Richard A Gibbs[2], Eric Boerwinkle[1,2]

[1]University of Texas Health Science Center at Houston, Human Genetics Center, Houston, TX, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX

Amino acids play critical roles for all of biology as precursors of other biomolecules or as intermediates in metabolism. Blood levels of amino acids in humans are important biomarkers of disease, and are affected by their biosynthesis, protein degradation, diet and interactions with the microbiome. Because of their molecular proximity to gene action, it is expected that effects of DNA sequence variation on amino acids levels may be large. We performed whole genome sequencing (WGS) and measured 70 serum amino acids in a sample of 1,458 European Americans (EAs) and 1,679 African Americans (AAs) from the Atherosclerosis Risk in Communities (ARIC) Study. An agnostic genome-wide sliding window analysis strategy (4kb window length with 2kb skip length) was applied in EAs and AAs respectively to 1) analyze common variants (MAF $\geq$ 5%) individually; and 2) aggregate low frequency variants (MAF < 5%) within a window and analyze them across the genome using a burden test (T5). Only variants or genetic regions which are significant in both ethnicities are interpreted here. A total of ~75 million variants were captured, and the number of low frequency variants per window ranged from 1 to 693 with median of 61. Common variants were identified to affect 22 amino acids levels ($p < 7.0\times10\text{-}10$, accounting for 1 million common variants and 70 traits), and among those, ten novel gene-metabolite pairs were identified in both EAs and AAs. For example, multiple common variants in DMGDH gene were associated with an average of 13.3% lower dimethylglycine levels in both EAs and AAs. For the T5 test, seven genetic regions consisting of 24 windows were identified affecting six amino acids levels at genome-wide significance ($p < 5.5\times10\text{-}10$, accounting for 1.3 million windows and 70 traits). Five out of seven identified genetic regions were non-coding regions including three ncRNA regions. Interestingly, windows upstream of the AGA gene were identified to be associated with 12-14% increase asparagine levels in EAs and AAs, and where predicted to be active promoters of transcription. Mutations in DMGDH and AGA are reported as autosomal recessive conditions for dimethylglycine and asparagine disorders (MIM: 605849, MIM: 613228), respectively. We are the first to observe that sequencing variants influence levels of dimethylglycine and asparagine in the general population. By integrating –omic technologies into deeply phenotyped populations, we showed that sequencing variants affect multiple human amino acids levels among two ethnicities. These data and results are identifying new avenues of gene function, novel molecular mechanisms and potentially diagnostic targets for multiple diseases.

← Back

TRANS-REGULATORY ARCHITECTURE OF GENETIC TRANSCRIPTOME VARIATION FROM 1,000 YEAST INDIVIDUALS

Frank W Albert[1,2], Joshua S Bloom[1], Jake Siegel[1], Laura Day[1], Leonid Kruglyak[1,3,4]

[1]University of California, Los Angeles, Department of Human Genetics, Los Angeles, CA, [2]University of Minnesota, Department of Genetics, Cell Biology, & Development, Minneapolis, MN, [3]University of California, Los Angeles, Department of Biological Chemistry, Los Angeles, CA, [4]Howard Hughes Medical Institute, UCLA, Los Angeles, CA

Genetic variation influences important traits in humans and other species. Many of these genetic effects are due to regulatory variation that influences gene expression. Regulatory variation can be identified as "expression quantitative trait loci" (eQTL). To date, all eQTL studies in any species have been hampered by low statistical power due to limited sample sizes. As a consequence, the full extent and nature of regulatory variation remains unknown.

We have addressed this limitation in the yeast *Saccharomyces cerevisiae*. We used mRNA sequencing to profile genome-wide gene expression in more than 1,000 recombinant individuals generated from a cross between two genetically different yeast strains. The statistical power of this dataset is high enough to map thousands of previously "missing" eQTL that together account for ~80% of the heritability of gene expression. Thus, our data provide a nearly exhaustive view of how genetic variation influences the transcriptome.

We identified 34,318 eQTL for 6,210 transcripts. A typical transcript is influenced by a median of 6 and by up to 20 eQTL, several fold more than previously seen. While 43% of all genes had a local eQTL that is located close to the gene, most eQTL are located elsewhere in the genome and influence gene expression in *trans*. The aggregate effect of the *trans* eQTL was larger than that of the local eQTL, illustrating the importance of *trans* acting variation.

Our near-complete view of *trans* regulatory architectures revealed several interesting features. Rather than appearing randomly across the genome, the newly discovered *trans* eQTL were highly structured, such that the vast majority fell into one of 111 hotspot regions that affect the expression of many genes. Some of these hotspots are now seen to have extraordinarily wide-reaching effects and can influence thousands of transcripts across all cellular processes, while others specifically influence certain pathways. By combining information from all genes that map to a given hotspot, we can fine-map the causal hotspot location with high precision, in 26 cases to single-gene resolution of less than 1.5 kb. This permits – for the first time – a systematic and unbiased analysis of the types of genes that act as *trans* eQTL. *Trans*-acting variation generates structure in the yeast transcriptome such that groups of genes are affected by multiple eQTL in a combinatorial fashion. Finally, despite our high statistical power, many local eQTL did not act as *trans* eQTL for other genes. This might indicate that the expression changes these eQTL cause at their local genes do not further affect cellular physiology, at least not in ways that are reflected in the transcriptome. This raises important questions about the characteristics of genes where local regulatory variation does have cellular and phenotypic consequences.

* FWA and JSB contributed equally

COUPLING GENOMIC SEQUENCING ANALYSES WITH GENOME EDITING TO REVEAL A ROLE FOR ALTERNATIVE *GFI1B* SPLICE VARIANTS IN HUMAN HEMATOPOIESIS

Linda M Polfus¶[1], Rajiv K Khajuria¶[2,3], Ursula M Schick¶[4], Alex P Reiner*[5], Santhi K Ganesh*[6], Vijay G Sankaran*[2,3]
on behalf of the CHARGE and NHLBI GO Exome Sequencing Project Hematology Working Groups
[1]University of Texas Health Science Center , Human Genetics Center, Houston, TX, [2]Boston Children's Hospital , Division of Hematology and Oncology, Boston, MA, [3]Broad Institute of Massachusetts Institute of Technology, and Harvard, Cambridge, MA, [4]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, [5]University of Washington, Department of Epidemiology, Seattle, WA, [6]University of Michigan, Division of Cardiovascular Medicine, Department of Internal Medicine, Department of Human Genetics, Ann Arbor, MI

With the increased use of exome and genome sequencing in large population-based studies, new insight into human biology is possible. Since circulating blood cell counts and indices are important indicators of hematopoietic function and clinical status, we performed whole exome sequence analyses of hematologic quantitative traits, in a large population based cohort of 15,459 individuals followed by replication in 52,024 samples. We identified a number of variants associated with blood cell traits and counts, including a low frequency synonymous variant in GFI1B associated with reduced platelet counts (rs150813342 MAF=0.009; Pdiscovery+replication=1.79x10-27), but not with other blood cell traits. GFI1B is a key transcriptional regulator of red blood cell and platelet production. Specific rare mutations of GFI1B have been identified in patients with Gray platelet syndrome, a bleeding disorder characterized by a qualitative and quantitative defect in platelets. A major question is how this low frequency synonymous variant could impact the platelet lineage selectively, while not having an impact on red blood cell production. The exact impact of the synonymous variant in GFI1B was not clear, although bioinformatic analysis suggested a potential impact on splicing involving disruption of a putative exonic splicing enhancer.
We reasoned that functional studies could enable further insight into the mechanisms underlying the GFI1B association identified. To directly assess the function of this variant, we utilized CRISPR/Cas9 genome editing to engineer the rs150813342 variant in K562 hematopoietic cell lines. Using quantitative RT-PCR, we could show that the rs150813342 variant suppressed formation of a long GFI1B isoform, which includes exon 5 that contains the rs150813342 variant. In the presence of homozygosity for rs150813342, less than 30% of the GFI1B transcripts included exon 5 relative to other exons. Isogenic cell lines harboring this variant could be robustly differentiated toward the red blood cell (erythroid) lineage, but platelet precursor cell (megakaryocyte) differentiation was markedly impaired. We further tested the lineage specific function of GFI1B isoforms in primary human hematopoietic stem/ progenitor cells using targeted knockdown with RNA interference and found the long GFI1B isoform to be necessary for megakaryocyte, but not erythroid, differentiation. Our work elucidates a novel role for distinct GFI1B isoforms in lineage bifurcation decisions that have a key role in human hematopoiesis. More generally, our findings illustrate the key insight that can be gained by coupling the results from large population-based genomic sequencing studies with functional assessment of such variation using cutting-edge genome engineering tools.

COMPREHENSIVE FINE MAPPING AND FUNCTIONAL INTERPRETATION OF HUMAN TRAITS

V Iotchkova[1,2], J Huang[1], K Walter[1], J Morris[3,4], C Barbieri[1,5], G RS Ritchie[1,2], J L Min[6], UK10K Consortium[1], I Dunham[2], N J Timpson[6], A P Reiner[7,8], P L Auer[9], E Birney[2], N Soranzo[1,10,11]

[1]Wellcome Trust Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom, [3]McGill University, Lady Davis Institute for Medical Research, Montreal, Canada, [4]McGill University, Human Genetics, Montreal, Canada, [5]San Raffaele Scientific Institute, Division of Genetics and Cell Biology, Milan, Italy, [6]University of Bristol, MRC Integrative Epidemiology Unit, Bristol, United Kingdom, [7]Fred Hutchinson Cancer Research Center, Public Health Sciences, Seattle, WA, [8]University of Washington, Epidemiology, Seattle, WA, [9]University of Wisconsin-Milwaukee, Zilber School of Public Health, Milwaukee, WI, [10]University of Cambridge, Department of Haematology, Cambridge, United Kingdom, [11]University of Cambridge, NIHR BTRU, Cambridge, United Kingdom

One of the fundamental challenges in modern human genetics is to convert our continuously improving discovery of disease risk from GWAS studies into an understanding of human biology which is useful for disease management and therapeutic development. Intermediate traits, such as metabolite levels, serum protein levels, physiological measurements and other aspects of normal physiology provide a rich phenotype of natural homeostasis. By mapping the genetic components of variation in these phenotypes we can both understand normal human physiology and explore the relationship with disease processes. This in turn provides useful biomarkers and therapeutic endpoints with clear cut causal relationships with diseases process.

To show the power of this approach we mapped 20 biomedical traits (lipids, hematological, glycemic, inflamatory, renal) in over 35,000 people from 18 studies including the UK10K cohort, replicated in over 100,000 people from 7 studies using whole genome sequencing data and imputation techniques. This has provided an unprecedented perspective on genetic components of normal variation in human physiology. We reconfirmed many previous associations, discovered 17 novel ones and refined the inter-relationship between traits and loci. Given the sample size and the whole genome sequence framework we were further able to fine map 59 loci to credible sets of under 20 variants (e.g. LIPC locus for association with HDL). We developed a robust method, GARFIELD, to associate loci to functional regions of interest (e.g. enhancers, promoters) and have found both expected and novel enrichments of functional elements and cell types with different human traits.

We have over 30 diseases (e.g. Autoimmune, Cardiovascular, Metabolic), which share at least one common locus with these physiological traits. The combination of functional enrichments and intermediate traits provide promising hypotheses of biomarkers and target gene identification.

← Back

THE HERITABILITY OF THE ORAL MICROBIOME

Brittany A Demmitt[1,2], Brooke M Huibregtse[2,3], Ivy S McDermott[1], Jaime Derringer[4], Robin P Corley[2], Matt B McQueen[2,5], John K Hewitt[2], Kenneth S Krauter[1,2]

[1]University of Colorado, Molecular, Cellular and Developmental Biology, Boulder, CO, [2]University of Colorado, Institute of Behavioral Genetics, Boulder, CO, [3]University of Colorado, Department of Psychology and Neuroscience, Boulder, CO, [4]University of Illinois at Urbana-Champaign, Department of Psychology, Champaign, IL, [5]University of Colorado, Department of Integrative Physiology, Boulder, CO

The oral microbiome is highly diverse with over 600 species identified. The human microbiome consists of complex microbial communities that reside in specific niches in and on the body. They play key roles in a growing list of human diseases including digestive syndromes, oral health, cardiac disease, and obesity among others. However, the extent to which host genetic factors influence the composition of the oral microbiome is not established. Using a cohort of 752 twin pairs we demonstrated that the overall community structure of the oral microbiome was significantly heritable measured by both beta diversity measures(p-value<0.0001) and the abundance of all taxa(p-value<0.001). In addition we determined that the genus *Streptococcus* as 31.6% heritable (p-value$=0.013$). We sought to identify the specific host genes that influence the microbiota present. We performed a genome wide association study(GWAS) on the abundance of the genus *Streptococcus*. This was completed separately in two different ancestry groups, Admixture($n=456$) and European ($n=828$), who have imputed genotypes of over 8 and 6 million single nucleotide polymorphisms (SNPs) across their genomes respectively. This analysis identified a few human genetic variants that were highly associated with the microbiome on chromosomes 7 and 16. Future research will focus on these regions of the genome to try and elucidate the mechanism by which the host genome influences the microbiome.

← Back

GENETICS OF LOCAL GENE EXPRESSION ACROSS 44 HUMAN CELL TYPES

Christopher D Brown[1], Stephen B Montgomery[2,3], GTEx Consortium[4]

[1]University of Pennsylvania, Genetics, Philadelphia, PA, [2]Stanford University, Pathology, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA, [4]GTEx Consortium, Boston, MA

The vast majority of the heritability of complex traits in humans lies in non-coding DNA with increasing evidence supporting the hypothesis that many causal variants function by modifying local gene expression. Expression QTL studies, which associate genetic variation to nearby gene expression, have improved our understanding of the mechanism of GWAS associations. However, to date they have been ascertained in a relatively limited number of cell types in diverse cohorts and individuals. Moreover, such studies have been biased towards cell types that are more readily obtained but may be less relevant for the interpretation of human disease. We present the identification of cis-eQTLs from 44 cell types examined in 70 to 361 individuals in the latest release of the GTEx project (v6). Relative to the GTEx pilot release, this analysis represents a large expansion that has now surveyed an additional 35 cell types and identified thousands of additional eSNPs and eGenes. We discover an increasing number of cell-type specific cis-eQTLs highlighting a unique identity to the impact of genetic variation in each surveyed cell type. Further, we untangle the effect sizes of shared and cell-type specific cis-eQTLs and uncouple multiple signals of association for each gene. By integrating large-scale epigenomics data and diverse genome annotation, we provide high-resolution analysis of the relationship between cross cell-type gene regulation and genetic control of gene expression. By integrating broader maps of allele-specific expression across cell types we enhance measures of sharing of cis-eQTL. The broader sampling of tissues enables us to explore the relationship between studies in large numbers of samples versus large numbers of tissues. We compare and contrast the utility of easily acquirable cell types to harder to collect cell types in the analysis and identification of novel cis-eQTLs underlying GWAS signals. Finally, we highlight the growing compendium of data available on the GTEx Portal. Together, the middle phase of the GTEx project delivers on increased understanding of the genetic control of gene expression across a broad range of cell types. These data provide exceptional utility for improving our understanding of gene regulation and human genetics.

← Back

TECHNICAL ABSTRACTS
FOR WORKSHOPS

← Back

← Back

ADVANCING GENOMICS WITH ILLUMINA'S SEQUENCING SOLUTIONS

Michael Smith Ph.D.
Sr. Sequencing Specialist
Illumina, Inc.

The growing range of Illumina Sequencing platforms has been instrumental in transforming the use of sequencing information in a wide range of genomic, genetic, and biological studies. Illumina continues to innovate, expanding into all aspects of the workflow, developing improved targeted resequencing workflows and downstream data analysis research solutions for areas such as oncology, genetic disease, and countless other areas of research with our new exome products. Illumina's expansive instrument portfolio -- from the newly introduced MiniSeq to the high throughput HiSeq X platform -- provides sequencing solutions for all of your sequencing needs from small targeted panels to large scale population sequencing studies. Join us to explore the possibilities and to learn how other customers are utilizing Illumina products.

For Research Use Only. Not for use in diagnostic procedures.

:

PRECISION MEDICINE AT COLUMBIA UNIVERSITY'S INSTITUTE FOR GENOMIC MEDICINE

Colin Malone, Ph.D.,
Director of Genomic Analysis and Technical Operations
Institute for Genomic Medicine
Columbia University Medical Center

The Institute for Genomic Medicine (IGM) was founded early in 2015 to provide precision medicine and broad genomic analysis capabilities to the Columbia University Medical Center (CUMC) community. With the capacity to process, clone, sequence and analyze over 15,000 human exomes annually, the IGM serves as the primary resource for patient and large cohort studies at CUMC. The IGM is focused on researching and understanding the genetic causes of numerous complex diseases including epilepsy, OCD, ALS, AMD, Parkinson's Disease, Schizophrenia and others. Findings from several recent studies will be reported.

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Dr. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Dr. Bishwo Adhikari
USDA-ARS
bishwoadhikari@email.arizona.edu

Mr. Shaked Afik
UC Berkeley
safik@berkeley.edu

Dr. Frank Albert
University of Minnesota
falbert@umn.edu

Dr. Carlos Eduardo Guerra Amorim
Columbia University
cg2827@columbia.edu

Ms. Andrea Anderson
GenomeWeb
anderson@genomeweb.com

Dr. Tatsiana Aneichyk
Harvard Medical School
taneichyk@mgh.harvard.edu

Dr. Misha Angrist
Duke University
misha.angrist@duke.edu

Dr. Irina Armean
Massachusetts General Hospital
iarmean@broadinstitute.org

Ms. Samira Asgari
EPFL
samira.asgari@epfl.ch

Dr. Georgios Athanasiadis
Aarhus University
athanasiadis@birc.au.dk

Dr. Elizabeth Atkinson
Stony Brook University
elizabeth.atkinson@stonybrook.edu

Dr. María Ávila-Arcos
National Autonomous University of Mexico
maricugh@gmail.com

Dr. Taejeong Bae
Mayo Clinic
bae.taejeong@mayo.edu

Dr. Orli Bahcall
Nature
o.bahcall@us.nature.com

Dr. Sara Ballouz
CSHL
sballouz@cshl.edu

Mr. Galt Barber
University of California at Santa Cruz (UCSC)
galt@soe.ucsc.edu

Ms. Tara Baris
University of Miami/RSMAS
tzbaris@gmail.com

Dr. Luis Barreiro
University of Montreal
luis.barreiro@umontreal.ca

Dr. Jeffrey Barrett
Wellcome Trust Sanger Institute
pl7@sanger.ac.uk

Mr. Justin Bartanus
Baylor College of Medicine
justin.bartanus@bcm.edu

Dr. Elizabeth Bartom
Northwestern University
ebartom@northwestern.edu

Mr. Dan Bar-Yaacov
Weizmann Institute of Science
dan.bar-yaacov@weizmann.ac.il

Dr. Alexis Battle
Johns Hopkins University
ajbattle@cs.jhu.edu

Dr. Serafim Batzoglou
Stanford University
serafim@cs.stanford.edu

Mr. Lyam Baudry
Institut Pasteur
lyambaudry@gmail.com

Mr. Christopher Bauer
Geisinger Health System
cbauer@geisinger.edu

Prof. Jaume Bertranpetit
Universitat Pompeu Fabra
jaume.bertranpetit@upf.edu

Dr. Claude Bhérer
New York Genome Center
claudebherer@gmail.com

Mr. Kunal Bhutani
University of California, San Diego
kunalbhutani@gmail.com

Dr. Minou Bina
Purdue University
bina@purdue.edu

Dr. Ewan Birney
EBI/EMBL
birney@ebi.ac.uk

Mr. Alex Bishara
Stanford University
abishara@cs.stanford.edu

Ms. Lauren Blake
University of Chicago
leblake@uchicago.edu

Dr. Ran Blekhman
University of Minnesota
blekhman@umn.edu

Dr. Jason Bobe
Icahn School of Medicine at Mount Sinai
jason.bobe@mssm.edu

Dr. Linda Boettger
The Broad Institute
boettger@broadinstitute.org

Ms. Silvia Bonas Guarch
Barcelona Supercomputing Center
silvia.bonas@bsc.es

Dr. Mark Borowsky
Novartis Institutes for BioMedical Research, INC.
mark.borowsky@novartis.com

Mr. Aritra Bose
Rensselaer Polytechnic Institute
bosea@rpi.edu

Dr. Mattia Bosio
Centre for Genomic Regulation-CRG
mattia.bosio@crg.eu

Dr. Florian Breitwieser
Johns Hopkins University
fbreitw1@jhu.edu

Dr. Christopher Brown
University of Pennsylvania
chrbro@mail.med.upenn.edu

Dr. Catherine Brownstein
Boston Children's Hospital
catherine.brownstein@childrens.harvard.edu

Dr. Alejandro Burga
UCLA
aburga@mednet.ucla.edu

Dr. David Burt
The Roslin Institute/University of Edinburgh
Dave.Burt@roslin.ed.ac.uk

Dr. George Busby
University of Oxford
george@well.ox.ac.uk

Prof. Carlos Bustamante
Stanford University
cdbadmin@stanford.edu

Mr. Alex Cagan
Max Planck Institute for Evolutionary Anthropology
alexander_cagan@eva.mpg.de

Mr. C. Ryan Campbell
Duke University
c.ryan.campbell@duke.edu

Dr. Michael Campbell
Cold Spring Harbor Laboratory
mcampbel@cshl.edu

Dr. Brandi Cantarel
UT Southwestern
brandi.cantarel@utsouthwestern.edu

Dr. Piero Carninci
RIKEN Center for Life Science Technologies
carninci@riken.jp

Dr. Karen Carniol
Cell Press
kcarniol@cell.com

Dr. Sergi Castellano
Max Planck Institute for Evolutionary Anthropology
sergi.castellano@eva.mpg.de

Dr. Mark Chaisson
University of Washington
mchaisso@uw.edu

Dr. Vered Chalifa-Caspi
Ben-Gurion University of the Negev
veredcc@bgu.ac.il

Dr. Esther Chan
Stanford University
etchan@stanford.edu

Dr. Ti-Cheng Chang
St Jude Children's research hospital
ti-cheng.chang@stjude.org

Dr. Carole Charlier
University of Liège
carole.charlier@ulg.ac.be

Dr. Emmanuelle Charpentier
Max Planck Institute for Infection Biology
charpentier@mpiib-berlin.mpg.de

Dr. Barbara Cheifet
Genome Biology
barbara.cheifet@genomebiology.com

Dr. Nancy Chen
University of California, Davis
nanchen@ucdavis.edu

Dr. Noel Chen
Novogene Corporation Inc.
noel.chen@novogene.com

Mr. Colby Chiang
Washington University
cchiang3@gmail.com

Dr. Mildred Cho
Stanford University
micho@stanford.edu

Dr. Min Cho
Nature Publishing Group
min.cho@nature.com

Dr. Wendy Chung
Columbia University
wkc15@columbia.edu

Dr. Deanna Church
10X Genomics
deanna.church@10xgenomics.com

Dr. Anne Churchland
CSHL
churchland@cshl.edu

Dr. Francesca Ciccarelli
King's College London
francesca.ciccarelli@kcl.ac.uk

Dr. Matthew Clark
The Genome Analysis Centre
matt.clark@tgac.ac.uk

Dr. Andrew Clark
Cornell University
ac347@cornell.edu

Mr. Ryan Collins
Massachusetts General Hospital
rcollins@chgr.mgh.harvard.edu

Mr. Michael Considine
Johns Hopkins University
mconsid3@jhmi.edu

Dr. Margherita Corioni
Agilent Technolgies
margherita_corioni@agilent.com

Dr. Montserrat Corominas
Universitat de Barcelona
mcorominas@ub.edu

Dr. Chris Cotsapas
Yale School of Medicine
cotsapas@broadinstitute.org

Dr. Mark Cowley
Garvan Institute of Medical Research
m.cowley@garvan.org.au

Dr. Marzia Cremona
Penn State University
mac78@psu.edu

Dr. Kathryn Crouch
University of Glasgow
kathryn.crouch@glasgow.ac.uk

Dr. Megan Crow
Cold Spring Harbor Laboratory
mcrow@cshl.edu

Mr. Hongzhu Cui
Worcester Polytechnic Institute
hcui2@wpi.edu

Ms. Beryl Cummings
Broad Institute
berylc@broadinstitute.org

Ms. Ciara Curtin
GenomeWeb
ccurtin@genomeweb.com

Dr. Jesse Dabney
University of Wisconsin - Madison
jdabney@wisc.edu

Dr. Amy Dapper
University of Wisconsin - Madison
dapper@wisc.edu

Dr. Aaron Day-Williams
Biogen
aaron.day-williams@biogen.com

Mr. Theodorus de Groot
University of Wisconsin - Madison
tedstertm@gmail.com

Dr. Michiel de Hoon
RIKEN
michiel.dehoon@riken.jp

Ms. Katrina de Lange
Wellcome Trust Sanger Institute
kdl@sanger.ac.uk

Mr. Christopher DeBoever
University of California San Diego
cdeboeve@ucsd.edu

Dr. Jacob Degner
Abbvie
jfdegner@gmail.com

Dr. Olivier Delaneau
University of Geneva
olivier.delaneau@gmail.com

Dr. Laura DeMare
Genome Research/Molecular Case Studies
ldemare@cshl.edu

Dr. Jonas Demeulemeester
The Francis Crick Institute
jonas.demeulemeester@crick.ac.uk

Ms. Brittany Demmitt
University of Colorado at Boulder
brittany.demmitt@colorado.edu

Dr. Scott Devine
Unversity of Maryland School of Medicine
sdevine@som.umaryland.edu

Dr. Federica Di Palma
The Genome Analysis Centre
federica.di-palma@tgac.ac.uk

Ms. Tonya Di Sera
University of Utah
tonyads@genetics.utah.edu

Dr. Jack DiGiovanna
Seven Bridges Genomics
jack.digiovanna@sbgenomics.com

Dr. Joelia Dmitrieva
Univerité de Liège
jbdmitrieva@ulg.ac.be

Dr. Alexander Dobin
CSHL
dobin@cshl.edu

Dr. Xianjun Dong
Brigham and Women's Hospital
xdong@rics.bwh.harvard.edu

Prof. Peter Donnelly
University of Oxford
donnelly@well.ox.ac.uk

Dr. Janina Dordel
Drexel University
jdordel@drexel.edu

Ms. Sondra Dubowsky
McLennan Community College
sdubowsky@mclennan.edu

Mr. Noah Dukler
Cold Spring Harbor Labs
ndukler@cshl.edu

Mr. Mahmoud Elansary
Unit of Animal Genomics, GIGA-R
mahmoud.elansary2012@gmail.com

Dr. Nels Elde
University of Utah
nelde@genetics.utah.edu

Dr. Ingegerd Elvers
Broad Institute & Uppsala University
ielvers@broadinstitute.org

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Dr. Jeanne Erdmann
Freelance Journalist
jeanne.erdmann@gmail.com

Dr. Michael Erdos
National Institutes of Health
mikee@mail.nih.gov

Dr. Yaniv Erlich
New York Genome Center/Columbia University
yaniv@cs.columbia.edu

Dr. Laurence Ettwiller
New England Biolabs
ettwiller@neb.com

Dr. Maud Fagny
Harvard School of Public Health
mfagny@jimmy.harvard.edu

Ms. Susan Fairley
EMBL-EBI
fairley@ebi.ac.uk

Dr. Khalid Fakhro
Sidra Medical and Research Center
kfakhro@sidra.org

Mr. Han Fang
Cold Spring Harbor Laboratory
hanfang.cshl@gmail.com

Dr. Catherine Farrell
NIH/NLM/NCBI
farrelca@ncbi.nlm.nih.gov

Dr. Andrew Farrell
University of Utah
farrelac@bc.edu

Ms. Lynn Fellman
Fellman Studios
lynn@fellmanstudio.com

Dr. Adam Felsenfeld
National Institutes of Health
adam_felsenfeld@nih.gov

Dr. Pedro Ferreira
Ipatimup
pferreira@ipatimup.pt

Dr. Yair Field
Stanford/HHMI
yairf@stanford.edu

Dr. Paul Flicek
EMBL-EBI
kwalsh@ebi.ac.uk

Dr. Liliana Florea
Johns Hopkins School of Medicine
florea@jhu.edu

Dr. Steven Flygare
University of Utah
sflygare@genetics.utah.edu

Dr. Joseph Foley
Stanford University
jwfoley@stanford.edu

Dr. Christine Fosker
The Genome Analysis Centre
business.support@tgac.ac.uk

Dr. Audrey Fu
University of Idaho
audreyf@uidaho.edu

Dr. Daniel Gaffney
Wellcome Trust Sanger Institute
aw12@sanger.ac.uk

Dr. Julien Gagneur
Technical University Munich
gagneur@in.tum.de

Dr. Pedro Galante
Hospital Sirio-Libanes
pgalante@mochsl.org.br

Dr. Xin Gao
ORISE/FDA
xin.gao1@fda.hhs.gov

Dr. Ziyue Gao
HHMI/Stanford University
ziyuegao@stanford.edu

Dr. Manuel Garber
University of Massachusetts Medical School
heidi.beberman@umassmed.edu

Dr. Eugene Gardner
University of Maryland School of Medicine
egardner@umaryland.edu

Dr. Rohit Garg
Harvard
rohitgarg@g.harvard.edu

Mr. Tyler Garvin
Cold Spring Harbor Laboratory
tgarvin@cshl.edu

Dr. Tina Gatlin
NHGRI/NIH
christine.gatlin@nih.gov

Dr. Diane Genereux
University of Massachusetts Medical School
diane.genereux@umassmed.edu

Dr. Michel Georges
ULg
michel.georges@ulg.ac.be

Dr. Mark Gerstein
Yale University
pi@gersteinlab.org

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Dr. David Gifford
Massachusetts Institute of Technology
gifford@mit.edu

Dr. Jesse Gillis
Cold Spring Harbor Laboratory
jgillis@cshl.edu

Dr. Thomas Gingeras
Cold Spring Harbor Laboratory
gingeras@cshl.edu

Mr. Craig Glastonbury
King's College London
craig.glastonbury@kcl.ac.uk

Dr. Sara Goodwin
Cold Spring Harbor Laboratory
sgoodwin@cshl.edu

Ms. Shyamalika Gopalan
Stony Brook University
shyamalika.gopalan@stonybrook.edu

Dr. Bettie Graham
National Human Genomr Research Institute/NIH
grahambj@exchange.nih.gov

Dr. Simon Gravel
McGill University
simon.gravel@mcGill.ca

Dr. Brenton Graveley
University of Connecticut Health Center
graveley@uchc.edu

Prof. Richard Green
University of California, Santa Cruz
ed@soe.ucsc.edu

Ms. Anna Green
Harvard Medical School
anna.g.green@gmail.com

Prof. William Greenleaf
Stanford University
wjg@stanford.edu

Dr. Ilan Gronau
Herzliya Interdiscipinary Center (IDC)
ilan.gronau@idc.ac.il

Dr. Roderic Guigo
Centre for Genomic Regulation (CRG)
roderic.guigo@crg.cat

Dr. Rodrigo Gularte Merida
Memorial Sloan Kattering Cancer Center
rodrigo.gularte@ulg.ac.be

Dr. Brad Gulko
Cornell University/CSHL
bgulko@cshl.edu

Dr. Zhaozhengatsun Guo
BGI
guozhaozheng@genomics.cn

Dr. Ryan Gutenkunst
University of Arizona
rgutenk@email.arizona.edu

Ms. Benika Hall
UNC Charlotte
dmrosema@uncc.edu

Dr. Ira Hall
Washington University
ihall@genome.wustl.edu

Dr. Gabriel Haller
Washington University
hallerg@wudosis.wustl.edu

Ms. Clair Han
Princeton University
clairh@princeton.edu

Mr. Bob Handsaker
Broad Institute
handsake@broadinstitute.org

Dr. Kasper Hansen
Johns Hopkins University
kasperdanielhansen@gmail.com

Mr. Nicholas Haradhvala
The Broad Institute
njharlen@gmail.com

Dr. Timothy Harkins
Swift Biosciences
zaborski@swiftbiosci.com

Dr. R. Alan Harris
Baylor College of Medicine
rharris1@bcm.edu

Dr. Tim Harris
Bajan Biotech
tjrharris@gmail.com

Dr. Christopher Hart
Ionis Pharmaceuticals
chart@ionisph.com

Dr. Shinichi Hashimoto
Kanazawa University
hashimoto@med.kanazawa-u.ac.jp

Mr. James Havrilla
University of Utah
semjaavria@gmail.com

Dr. Ximiao He
National Cancer Institute
Ximiao.He@nih.gov

Dr. Tim Hefferon
NIH/NLM/NCBI
theffero@ncbi.nlm.nih.gov

Dr. Brenna Henn
SUNY Stony Brook
brenna.henn@stonybrook.edu

Dr. Peter Heutink
DZNE
peter.heutink@dzne.de

Dr. Michael Hiller
Max Planck Institute of Cell Biology and Genetics
hiller@mpi-cbg.de

Dr. Yu-Jui Ho
Cold Spring Harbor Laboratory
yjho@cshl.edu

Ms. Rebecca Hong
NHGRI/NIH
rebecca.hong@nih.gov

Dr. Chang Pyo Hong
Theragenetex Bio Institute
changpyo.hong@theragenetex.com

Mr. Farhad Hormozdiari
University of California, Los Angeles (UCLA)
farhad.hormozdiari@gmail.com

Dr. Fereydoun Hormozdiari
UC Davis
fhormozd@ucdavis.edu

Dr. Kate Howell
University of Cambridge/EBI
kjh52@cam.ac.uk

Dr. Xiaolan Hu
Celgene
xiaolanxq@gmail.com

Prof. Lusheng Huang
Jiangxi Agricultural University
lushenghuang@hotmail.com

Dr. Yifei Huang
Cold Spring Harbor Laboratory
yihuang@cshl.edu

Ms. Melissa Hubisz
Cold Spring Harbor Laboratory
mhubisz@cshl.edu

Dr. Chad Huff
The University of Texas MD Anderson Cancer Center
chuff1@mdanderson.org

Mr. Drew Hughes
Washington University in St. Louis
hughesa@wusm.wustl.edu

Dr. Matthew Hurles
Wellcome Trust Sanger Institute
meh@sanger.ac.uk

Mr. Lucas Husquin
Institut Pasteur
lucas.husquin@pasteur.fr

Dr. Julie Hussin
University of Oxford
julieh@well.ox.ac.uk

Ms. Elizabeth Hutton
Watson School of Biological Sciences, CSHL
ehutton@cshl.edu

Dr. Oleg Iartchouk
Novartis - NIBR
oleg.iartchouk@novartis.com

Mr. Kazuki Ichikawa
The University of Tokyo
ichikawa@cb.k.u-tokyo.ac.jp

Dr. Hae Kyung Im
The University of Chicago
haky@uchicago.edu

Dr. Valentina Iotchkova
EMBL-EBI
vi@ebi.ac.uk

Dr. Andrew Jaffe
Lieber Institute for Brain Development
andrew.jaffe@libd.org

Dr. David Jaffe
10X Genomics
david.jaffe@10xgenomics.com

Dr. Jin Sung Jang
Mayo Clinic
jang.jin@mayo.edu

Mr. Jacob Jensen
Aarhus University
jmj@birc.au.dk

Dr. Peilin Jia
UTHealth
peilin.jia@vanderbilt.edu

Ms. Shan (Mandy) Jiang
University of California, Irvine
jiangs2@uci.edu

Mr. Runze Jiang
Shenzhen Luohu Hospital Group
solon.jiang@gmail.com

Dr. Yinping Jiao
Cold Spring Harbor Lab
yjiao@cshl.edu

Dr. Ying Jin
Cold Spring Harbor Laboratory
yjin@cshl.edu

Dr. Sally John
Biogen Idec, Inc.
sally.john@biogen.com

Dr. Felicity Jones
Friedrich Miescher Laboratory of the Max Planck Society
fcjones@tuebingen.mpg.de

Dr. Luke Jostins
University of Oxford
lj4@well.ox.ac.uk

Ms. Navya Josyula
Geisinger System Services
nsjosyula@geisinger.edu

Dr. Goo Jun
University of Texas Health Science Center Houston
goo.jun@uth.tmc.edu

Dr. Irwin Jungreis
MIT
iljungr@csail.mit.edu

Dr. Naveen Kadri
University of Liege
nk.kadri@ulg.ac.be

Dr. Konrad Karczewski
Massachusetts General Hospital
konradk@broadinstitute.org

Dr. Elinor Karlsson
U Mass Med School
elinor@broadinstitute.org

Mr. Sivakanthan Kasinathan
Fred Hutchinson Cancer Research Center
skasinat@fredhutch.org

Dr. Dave Kaufman
National Institutes of Health
dave.kaufman@nih.gov

Dr. Alon Keinan
Cornell University
alon.keinan@cornell.edu

Ms. Melissa Keinath
University of Kentucky
melissa.keinath@gmail.com

Dr. Manolis Kellis
MIT
manoli@mit.edu

Dr. Ekaterina Khramtsova
The University of Chicago
eakhram@gmail.com

Dr. Dokyoon Kim
Geisinger Health System
dkim@geisinger.edu

Dr. Pora Kim
The University of Texas Health Science Center
pora.kim@uth.tmc.edu

Dr. Daehwan Kim
Johns Hopkins University School of Medicine
infphilo@gmail.com

Dr. Sarah Kim-Hellmuth
New York Genome Center
skim@nygenome.org

Dr. Anthony Kirilusha
National Institutes of Health
anthony.kirilusha@nih.gov

Dr. Paul Kitts
NIH/NLM/NCBI
kitts@ncbi.nlm.nih.gov

Ms. Johanna Klughammer
CeMM Center for Molecular Medicine
jklughammer@cemm.oeaw.ac.at

Dr. Jo Knight
Lancaster University
Jo.Knight@Lancaster.ac.uk

Mr. Binyamin Knisbacher
Bar-Ilan University
binyamin.knisbacher@biu.ac.il

Mr. Arthur Ko
UCLA
a5ko@ucla.edu

Ms. Ana Leticia Kolicheski
University of Missouri
akv22@mail.missouri.edu

Dr. Miriam Konkel
Louisiana State University
konkel@lsu.edu

Dr. Ksenia Krasileva
The Genome Analysis Centre
business.support@tgac.ac.uk

Mr. Sam Krerowicz
UW-Madison
krerowicz@wisc.edu

Mr. Lukas Kuderna
IBE-Institute of Evolutionary Biology - (UPF-CSIC)
lukas.kuderna@upf.edu

Ms. Avantika Lal
National Centre for Biological Sciences
avantika@ncbs.res.in

Dr. Xun Lan
Stanford University
xlan@stanford.edu

Dr. Eric Lander
The Broad Institute of MIT & Harvard
lander@broadinstitute.org

Dr. Tuuli Lappalainen
New York Genome Center & Columbia University
tlappalainen@nygenome.org

Mr. Christopher Laumer
EMBL-EBI and the Sanger Institute
claumer@ebi.ac.uk

Dr. Ryan Layer
University of Utah
ryan.layer@utah.edu

Ms. Amanda Lea
Duke University
amanda.lea@duke.edu

Mr. Dong Jin Lee
Theragenetex Bio Institute
dongjin.lee@theragenetex.com

Dr. Charles Lee
The Jackson Laboratory for Genomic Medicine
charles.lee@jax.org

Dr. Kalle Leppälä
Aarhus University
kalle.m.leppala@gmail.com

Dr. Dena Leshkowitz
Weizmann Institute of Science
dena.leshkowitz@weizmann.ac.il

Dr. Sabrina Leslie
McGill University
sabrina.leslie@mcgill.ca

Ms. Zeran Li
Washington University in St. Louis
zeranli@wustl.edu

Dr. Jiani Li
Baylor College of Medicine
jianil@bcm.edu

Dr. Yang Li
Stanford University
yangili@stanford.edu

Ms. Meng Lin
Stony Brook University
meng.lin.1@stonybrook.edu

Dr. Stephen Lincoln
Invitae
steve.lincoln@me.com

Dr. Sarah Lindsay
Wellcome Trust Sanger Institute
sjl@sanger.ac.uk

Dr. Xiaoming Liu
University of Texas School of Public Health
xiaoming.liu@uth.tmc.edu

Dr. Xin Liu
BGI Shenzhen
liuxin@genomics.cn

Dr. Michael Lodato
Children's Hospital Boston / Harvard Medical Schoo
mlodato@gmail.com

Ms. Marie Lopez
Institut Pasteur
marie.lopez@pasteur.fr

Mr. Craig Lowe
Stanford University School of Medicine
lowec@stanford.edu

Dr. Francesca Luca
Wayne State University
fluca@wayne.edu

Dr. Daniel MacArthur
Mass General Hospital/Broad Institute
macarthur@atgu.mgh.harvard.edu

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Ms. Heather Machado
Stanford University
hmachado@stanford.edu

Mr. Sho Maekawa
The University of Tokyo
smaekawa@hgc.jp

Dr. Kateryna Makova
Penn State University
kdm16@psu.edu

Dr. Joel Malek
Weill Cornell Medical College in Qatar
jom2042@qatar-med.cornell.edu

Dr. Ankit Malhotra
Jackson Laboratory for Genomic Medicine
Ankit.Malhotra@jax.org

Dr. Colin Malone
Columbia University Medical Center
cm3556@cumc.columbia.edu

Dr. Elaine Mardis
Washington University School of Medicine
emardis@wustl.edu

Dr. Rob Mariman
GIGA R - University of Liège
rob.mariman@ulg.ac.be

Dr. Debora Marks
Harvard Medical School
annaggreen@fas.harvard.edu

Dr. Tomas Marques-Bonet
Universitat Pompeu Fabra
tomas.marques@upf.edu

Dr. Gabor Marth
University of Utah
gmarth@genetics.utah.edu

Dr. Andrea Massaia
Wellcome Trust Sanger Institute
am29@sanger.ac.uk

Dr. Carolyn McBride
Princeton University
csm7@princeton.edu

Dr. Jesse McClure
University of Massachusetts Medical School
jesse.mcclure@umassmed.edu

Dr. David McKinnon
Stony Brook University
dmckinnon@notes.cc.sunysb.edu

Dr. Francis McMahon
National Institutes of Health
mcmahonf@mail.nih.gov

Dr. Mark McMullan
The Genome Analysis Centre
mark.mcmullan@tgac.ac.uk

Dr. John McPherson
UC Davis Comprehensive Cancer Center
jdmcpherson@ucdavis.edu

Ms. Hannah Meyer
EMBL-EBI
hannah@ebi.ac.uk

Dr. Stephen Meyn
Hospital for Sick Children
stephen.meyn@sickkids.ca

Mr. Zong Miao
University of California, Los Angeles
abl0719@gmail.com

Dr. Chase Miller
University of Utah
chmille4@gmail.com

Dr. Ryan Mills
University of Michigan
remills@umich.edu

Dr. Dan Mishmar
Ben-Gurion University of the Negev
dmishmar@bgu.ac.il

Dr. Adele Mitchell
Merck Research Labs
adele.mitchell@merck.com

Dr. Pejman Mohammadi
NY Genome Center and Columbia University
pmohammadi@nygenome.org

Mr. David Molik
Cold Spring Harbor Lab
dmolik@cshl.edu

Dr. Yukihide Momozawa
RIKEN
yukihide.momozawa@riken.jp

Mr. Mayukh Mondal
Universitat Pompeu Fabra
mayukh.mondal@upf.edu

Mr. Jonathan Moody
University of Edinburgh
jonathan.moody@igmm.ed.ac.uk

Dr. Priya Moorjani
Columbia University
pm2730@columbia.edu

Ms. Lucía Morales
UNAM-LIIGH
lmorales@liigh.unam.mx

Dr. Shinichi Morishita
University of Tokyo
moris@cb.k.u-tokyo.ac.jp

Dr. Leonid Moroz
University of Florida
moroz@whitney.ufl.edu

Dr. Derek Morris
National University of Ireland Galway
derek.morris@nuigalway.ie

Dr. Sara Mostafavi
University of British Columbia
saram@stat.ubc.ca

Mr. Yuichi Motai
The University of Tokyo
motights@cb.k.u-tokyo.ac.jp

Dr. Jonathan Mudge
Wellcome Trust Sanger Institute
jm12@sanger.ac.uk

Dr. Kasper Munch
Aarhus University
kaspermunch@birc.au.dk

Ms. Shaila Musharoff
UCSF
shaila.musharoff@gmail.com

Dr. Tendai Mutangadura
University of Missouri
tendai@missouri.edu

Ms. Aditi Narayanan
Stanford School of Medicine
aditin14@stanford.edu

Dr. Maria Nattestad
Cold Spring Harbor Laboratory
mnattest@cshl.edu

Mr. Fabio Navarro
Yale University
fabio.navarro@yale.edu

Dr. Abhinav Nellore
Johns Hopkins University
anellore@gmail.com

Mr. Dominic Nelson
McGill University
nelson.dominic@gmail.com

Dr. Priscila Silva Neubern de Oliveira
Embrapa Pecuária Sudeste
priscilaneuberndeoliveira@gmail.com

Ms. Stephanie Nevins
Stanford University
snevins@stanford.edu

Dr. Khanh-Dung Nguyen
Biogen
kd.nguyen@biogen.com

Dr. Serena Nik-Zainal
Wellcome Trust Sanger Institute
snz@sanger.ac.uk

Dr. Arne Nolte
University of Oldenburg
nolte@evolbio.mpg.de

Dr. Christos Noutsos
Cold Spring Harbor Laboratory
cnoutsos@cshl.edu

Dr. Ramil Nurtdinov
Centre for Genomic Regulation (CRG)
ramil.nurtdinov@crg.eu

Mr. Ninad Oak
Baylor College of Medicine
ninad.oak@bcm.edu

Dr. Anne O'Donnell-Luria
Boston Children's Hospital/Harvard Medical School
anne.odonnell@childrens.harvard.edu

Ms. Meritxell Oliva
University Of Chicago
meritxellop@uchicago.edu

Dr. Halit Ongen
University of Geneva
halit.ongen@unige.ch

Dr. Roel Ophoff
UCLA
ophoff@ucla.edu

Dr. Jason O'Rawe
Cold Spring Harbor Laboratory
jazon33y@gmail.com

Dr. Ludovic Orlando
University of Copenhagen
Lorlando@snm.ku.dk

Dr. Stephan Ossowski
Center for Genomic Regulation
stephan.ossowski@crg.eu

Mr. Omead Ostadan
Illumina
oostadan@illumina.com

Dr. Marco Osterwalder
Lawrence Berkeley National Laboratory
mosterwalder@lbl.gov

Dr. Elaine Ostrander
National Institutes of Health
eostrand@mail.nih.gov

Dr. Toshio Ota
Kyowa Hakko Kirin Co., Ltd.
toshio.ota@kyowa-kirin.co.jp

Prof. Svante Pääbo
Max Planck Institute for Evolutionary Anthropology
mittag@eva.mpg.de

Dr. Annalise Paaby
Georgia Tech
annalise.paaby@biology.gatech.edu

Mr. Carlos Pabón-Peña
Agilent Technologies, Inc.
carlos_pabon@agilent.com

Prof. Paivi Pajukanta
University of California, Los Angeles (UCLA)
ppajukanta@mednet.ucla.edu

Mr. Nikolaos Panousis
University of Geneva
nikolaos.panousis@unige.ch

Dr. Yongjin Park
MIT
ypp@csail.mit.edu

Dr. Ji Yeon Park
Seoul National University Biomedical Informatics
parkji7@snu.ac.kr

Dr. Michael Pazin
NHGRI, NIH
pazinm@mail.nih.gov

Dr. Brent Pedersen
University of Utah
bpederse@gmail.com

Dr. Sarah Pendergrass
Geisinger Health Research
spendergrass@geisinger.edu

Dr. Joyce Peng
Novogene Corporation Inc
joyce.peng@novogene.com

Dr. Elizabeth Pennisi
Science
epennisi@aaas.org

Dr. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Alisa Petree
McLennan Community College
apetree@mclennan.edu

Dr. Susanne Pfeifer
EPFL
susanne.pfeifer@epfl.ch

Dr. Lon Phan
NIH/NLM/NCBI
lonphan@ncbi.nlm.nih.gov

Dr. Adam Phillippy
National Human Genome Research Institute
adam.phillippy@nih.gov

Dr. Joseph Pickrell
New York Genome Center
jkpickrell@nygenome.org

Dr. Luca Pinello
Dana-Farber Cancer Institute/Harvard School of Public Health
lpinello@jimmy.harvard.edu

Ms. Lenore Pipes
Cold Spring Harbor Laboratory
lpipes@cshl.edu

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Dr. Linda Polfus
University of Texas Health Science Center
Linda.M.Whitaker@uth.tmc.edu

Dr. David Porubsky
The University Medical Center Groningen
d.porubsky@umcg.nl

Dr. Jonathan Pritchard
Stanford University
ttrim@stanford.edu

Dr. Molly Przeworski
Columbia University
molly.przew@gmail.com

Prof. Francis Quetier
GIP GENOPOLE
sabine.carava@genopole.fr

Mrs. Elizabeth Quincy Rose

Dr. Aaron Quinlan
University of Utah
aaronquinlan@gmail.com

Dr. Andrew Quitadamo
University of North Carolina at Charlotte
aquitada@uncc.edu

Mr. Nima Rafati
Uppsala University
nimarafati@gmail.com

Dr. Narayanan Raghupathy
The Jackson Laboratory
narayanan.raghupathy@jax.org

Dr. Towfique Raj
Mount Sinai School of Medicine
towfique.raj@mssm.edu

Dr. Srinivas Ramachandran
Fred Hutchinson Cancer Research Center
sramacha@fredhutch.org

Dr. Srividya Ramakrishnan
Cold Spring Harbor Laboratory
sramakri@cshl.edu

Dr. Ritika Ramani
Cold Spring Harbor Laboratory
rramani@cshl.edu

Dr. Daniele Ramazzotti
Stanford University
daniele.ramazzotti@yahoo.com

Dr. Suhas Rao
Stanford University
suhas@suhasrao.com

Dr. Soumya Raychaudhuri
Brigham and Women's Hospital
soumya@broadinstitute.org

Dr. James Reecy
Iowa State University
jreecy@iastate.edu

Dr. Heidi Rehm
Harvard Medical School
hrehm@partners.org

Prof. Jun Ren
Jiangxi Agricultural University
renjunjxau@hotmail.com

Dr. Stephen Richards
Baylor College of Medicine
stephenr@bcm.edu

Dr. Allison Richards
Wayne State University
ga1765@wayne.edu

Dr. Harold Riethman
Old Dominion University
hriethma@odu.edu

Dr. Firas Riyazuddin
NIH/NLM/NCBI
firas.riyazuddin@nih.gov

Dr. Patrizia Rizzu
German Center For Neurodegenerative Diseases
patrizia.rizzu@dzne.de

Mr. Juan Rodriguez
Institute of Evolutionary Biology (UPF-CSIC)
juan.rodriguez@upf.edu

Dr. Jeffrey Rogers
Baylor College of Medicine
jr13@bcm.edu

Dr. Mostafa Ronaghi
Illumina, Inc.
mronaghi@illumina.com

Dr. Jeffrey Rosenfeld
Rutgers - New Jersey Medical School
jeffrey.rosenfeld@rutgers.edu

Dr. Nicolas Rosewick
Université de Liège
nicolas.rosewick@gmail.com

Dr. Joel Rozowsky
Yale University
joel.rozowsky@yale.edu

Dr. Yijun Ruan
The Jackson Laboratory for Genomic Medicine
maureen.sansone@jax.org

Dr. Douglas Ruderfer
Icahn School of Medicine at Mount Sinai
douglas.ruderfer@mssm.edu

Mr. Brian Ryu
Seoul National University Biomedical Informatics
brianryu87@gmail.com

Dr. Brooke Sadler
Washington University School of Medicine
sadlerb@psychiatry.wustl.edu

Ms. Tina Saey
Science News
tsaey@sciencenews.org

Dr. William Salerno
Baylor College of Medicine
ws144320@bcm.edu

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Dr. Albin Sandelin
University of Copenhagen
albin@binf.ku.dk

Dr. Ashley Sanders
BC Cancer Agency
asanders@bccrc.ca

Mr. Jaleal Sanjak
University of California, Irvine
jsanjak@uci.edu

Dr. Neville Sanjana
NY Genome Center and NYU
nsanjana@nygenome.org

Dr. Joaquin Sanz
Université de Montreal
jsanz@bifi.es

Mr. Cody Saraceno
University of Kentucky
CodySaraceno@gmail.com

Dr. Michael Schatz
Cold Spring Harbor
mschatz@cshl.edu

Dr. Paul Schaughency
Johns Hopkins School of Medicine
pschaugh@jhmi.edu

Dr. Steven Scherer
Baylor College of Medicine
sscherer@bcm.edu

Prof. Mikkel Heide Schierup
Aarhus University
mheide@birc.au.dk

Dr. Jeffery Schloss
NHGRI/NIH
schlossj@exchange.nih.gov

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Dr. Michael Schnall-Levin
10X Genomics
mike@10xgenomics.com

Dr. Korbinian Schneeberger
Max Planck Institute for Plant Breeding Research
schneeberger@mpipz.mpg.de

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@ncbi.nlm.nih.gov

Dr. Dustin Schones
Beckman Research Institute at the City of Hope
dschones@coh.org

Dr. David Schwartz
University of Wisconsin - Madison
dcschwartz@wisc.edu

Ms. Emma Scott
University of Maryland Baltimore
escott@umaryland.edu

Ms. Alexandra Scott
Washington University
ajscott@wustl.edu

Dr. Fritz Sedlazeck
Johns Hopkins University
fritz.sedlazeck@gmail.com

Dr. Eran Segal
Weizmann Institute of Science
eran@weizmann.ac.il

Dr. Laure Segurel
CNRS - Musée de l'Homme
lsegurel@mnhn.fr

Dr. Guy Sella
Columbia University
gs2747@columbia.edu

Dr. Colin Semple
University of Edinburgh
colin.semple@igmm.ed.ac.uk

Mr. Arko Sen
Wayne State University
asen@med.wayne.edu

Dr. Debarka Sengupta
Genome Institute of Singapore
debarka@gmail.com

Prof. Cathal Seoighe
National University of Ireland Galway
Cathal.Seoighe@nuigalway.ie

Dr. Eilon Sharon
Stanford University
eilon@stanford.edu

Dr. Vladimir Shchur
The Wellcome Trust Sanger Institute
vs3@sanger.ac.uk

Dr. Niranjan Shekar
Spiral Genetics
niranjan@spiralgenetics.com

Dr. Sushila Shenoy
Weill Cornell Medical College
sas2030@med.cornell.edu

Dr. Xinghua Shi
University of North Carolina at Charlotte
x.shi@uncc.edu

Dr. Parisa Shooshtari
Broad Institute of MIT-Harvard
parisa.shooshtari@yale.edu

Dr. Neil Shubin
University of Chicago
nshubin@uchicago.edu

Dr. Arend Sidow
Stanford University
arend@stanford.edu

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Mr. Yuval Simons
Columbia University
Yuval.simons@columbia.edu

Prof. Michael Simpson
King's College London
michael.simpson@kcl.ac.uk

Mr. Laurits Skov
Wellcome Genome Campus
lauritsskov2@gmail.com

Mr. Craig Smail
Stanford University
crgsml@gmail.com

Dr. Kerrin Small
Kings College London
kerrin.small@kcl.ac.uk

Dr. Michael Smith
Illumina
mismith@illumina.com

Dr. Jeramiah Smith
University of Kentucky
jjsmit3@uky.edu

Dr. Michael Snyder
Stanford University School of Medicine
scisnero@stanford.edu

Mr. Li Song
Johns Hopkins University
lsong10@jhu.edu

Dr. Noah Spies
NIST/Stanford University
nspies@stanford.edu

Dr. Arnold Stein
Purdue University
steina@purdue.edu

Mr. Benjamin Steyer
University of Wisconsin-Madison
bsteyer@wisc.edu

Dr. Barbara Stranger
University of Chicago
bstranger@medicine.bsd.uchicago.edu

Dr. J Seth Strattan
Stanford University
strattan@gmail.com

Ms. Linda Su-Feher
University of California Davis
lsu@ucdavis.edu

Dr. Lauren Sugden
Brown University
lauren_alpert@brown.edu

Dr. Zhiyi Sun
New England Biolabs, Inc.
sunz@neb.com

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Dr. Yutaka Suzuki
University of Tokyo
ysuzuki@hgc.jp

Mr. Yuta Suzuki
The University of Tokyo
ysuzuki@cb.k.u-tokyo.ac.jp

Dr. Yoshihiko Suzuki
The University of Tokyo
suzuki_yoshihiko_15@cbms.k.u-tokyo.ac.jp

Dr. Haruko Takeda
University of Liège
htakeda@ulg.ac.be

Dr. Michael Talkowski
MGH / Harvard / Broad Institute
talkowski@chgr.mgh.harvard.edu

Dr. Martin Taylor
University of Edinburgh
martin.taylor@igmm.ed.ac.uk

Dr. James Taylor
Johns Hopkins University
james@jamestaylor.org

Ms. Natalie Telis
Stanford University
ntelis@stanford.edu

Dr. Marcela Tello-Ruiz
Cold Spring Harbor Laboratory
mmonaco@cshl.edu

Dr. Ryan Tewhey
Harvard University
rtewhey@broadinstitute.org

Dr. Jitendra Thakur
Fred Hutch
jthakur@fhcrc.org

Ms. Naina Thangaraj
DNAnexus
thangarajnaina@gmail.com

Dr. Barbara Thomas
NIH
barbara.thomas@nih.gov

Dr. Jim Thomas
NIH
thomasjw4@mail.nih.gov

Dr. Kevin Thornton
University of California Irvine
krthornt@uci.edu

Dr. Vladimir Timoshevskiy
University of Kentucky
vti224@uky.edu

Dr. Polyana Tizioto
Embrapa Southeast Livestock
ptizioto@gmail.com

Dr. David Torrents Arenales
Barcelona Supercomputing Center and ICREA
david.torrents@bsc.es

Dr. Richard Trembath
King's College London
dean-folsm@kcl.ac.uk

Dr. Christopher Tuggle
Iowa State University
cktuggle@iastate.edu

Dr. Jenny Tung
Duke University
jt5@duke.edu

Dr. Andrei Turinsky
Hospital for Sick Children
turinsky@sickkids.ca

Dr. Mohammed Uddin
The Hospital for Sick Children, The Centre for App
mohammed.uddin@sickkids.ca

Mr. Jacob Ulirsch
Boston Children's Hospital
julirsch@broadinstitute.org

Mr. Matthew Ung
Geisel School of Medicine at Dartmouth
matthew.h.ung.gr@dartmouth.edu

Dr. Niko Välimäki
University of Helsinki
niko.valimaki@helsinki.fi

Dr. Anton Valouev
Grail Bio
valouev@gmail.com

Dr. Bryce van de Geijn
Harvard School of Public Health
bmvdgeijn@gmail.com

Dr. Anne Van den Broeke
GIGA-R University of Liège
anne.vandenbroeke53@gmail.com

Mr. Frederick Varn
Geisel School of Medicine at Dartmouth
Frederick.S.Varn.Jr.GR@dartmouth.edu

Prof. Byrappa Venkatesh
Institute of Molecular and Cell Biology
mcbbv@imcb.a-star.edu.sg

Dr. Francesco Vezzi
National Genomics Infrastructure (NGI)
francesco.vezzi@scilifelab.se

Dr. Bjarni Vilhjalmsson
Aarhus University
bjarni@birc.au.dk

Dr. Charles Vinson
National Cancer Institute
Vinsonc@mail.NIH.GOV

Dr. Ana Vinuela
University of Geneva
ana.vinuela@unige.ch

Mr. Kristoffer Vitting-Seerup
University of Copenhangen, Denmark
kristoffer.vittingseerup@bio.ku.dk

Dr. Natalia Volfovsky
Simons Foundation
nvolfovsky@simonsfoundation.org

Dr. Samir Wadhawan
Bristol-Myers Squib
samir.wadhawan@bms.com

Dr. Aleksandra Walczak
Laboratoire de Physique Théorique
awalczak@lpt.ens.fr

Dr. Xiu-Feng (Henry) Wan
MIssissippi State University
wan@cvm.msstate.edu

Dr. Zihua Wang
Cold Spring Harbor Laboratory
zwang@cshl.edu

Dr. JInhua Wang
NYU School of Medicine
jinhua.wang@nyumc.org

Dr. Lu Wang
NIH/NHGRI
wanglu@mail.nih.gov

Ms. Yifan Wang
University of Michigan
yifan.wang0801@gmail.com

Dr. Bo Wang
Stanford University
bowang87@stanford.edu

Dr. Alistair Ward
University of Utah
alistairnward@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory / USDA/ARS
ware@cshl.edu

Dr. Susanne Warrenfeltz
University of Georgia
swfeltz@uga.edu

Dr. Jia Wen
UNC Charlotte
jwen6@uncc.edu

Dr. Sean Whalen
Gladstone Institutes / UCSF
shwhalen@gmail.com

Dr. John Whitaker
Janssen R&D of Johnson & Johnson
jwhitak3@its.jnj.com

Mr. Thomas Willems
MIT/Whitehead Institute
twillems@mit.edu

Dr. Richard Wilson
Washington University in St. Louis
rwilson@genome.wustl.edu

Mr. Eamon Winden
University of Wisconsin, Madison
ewinden@wisc.edu

Dr. Deborah Winter
Weizmann Institute of Science
deborah.winter@weizmann.ac.il

Dr. Emily Wong
EMBL-EBI
em6@ebi.ac.uk

Dr. Kim Worley
Baylor College of Medicine
kworley@bcm.edu

Ms. Xiaoli Wu
Stony Brook University
xwu@cshl.edu

Dr. Chunlin Xiao
NIH/NLM/NCBI
xiao2@ncbi.nlm.nih.gov

Dr. Huanming Yang
Beijing Genomics Institute
yanghm@genomics.cn

Mr. Guangyu Yang
Johns Hopkins University
gyang22@jhu.edu

Dr. Mingyi Yang
Oslo University Hospital
mingyiy@ifi.uio.no

Ms. Angela Yen
MIT
angela@mit.edu

Dr. Gene Yeo
University of California, San Diego
geneyeo@ucsd.edu

Dr. Nir Yosef
UC Berkeley
niryosef@berkeley.edu

Mr. Alexander Young
University of Oxford
alextisyoung@gmail.com

Dr. Janet Young
Fred Hutchinson Cancer Research Center
jayoung@fhcrc.org

Dr. Noha Yousri
Weill Cornell Medical College-Qatar
nay2005@qatar-med.cornell.edu

Dr. Bing Yu
University of Texas Health Science at Houston
bing.yu@uth.tmc.edu

Dr. Yao Yu
The University of Texas MD Anderson Cancer Center
yyu4@mdanderson.org

Dr. Sophie Zaaijer
New York Genome Center/ Columbia University
szaaijer@nygenome.org

Mr. Benedikt Zacher
Ludwig-Maximilians-University
zacher@genzentrum.lmu.de

Dr. Laura Zahn
AAAS/Science
lzahn@aaas.org

Dr. Christina Zakas
New York University
cz12@nyu.edu

Mr. Daniel Zerbino
EMBL-EBI
zerbino@ebi.ac.uk

Dr. Bo Zhang
Washington University School of Medicine
zhangbo@wusm.wustl.edu

Dr. Zhonghua Zhang
Chinese Academy of Agricultural Sciences
zhangzhonghua@caas.cn

Dr. Shancen Zhao
BGI-Shenzhen
zhaoshancen@genomics.cn

Dr. Zhongming Zhao
University of Texas Health Science Center Houston
zhongming.zhao@uth.tmc.edu

Dr. Junfei Zhao
UT Health Science Center at Houston
Junfei.Zhao@uth.tmc.edu

Ms. Xuefang Zhao
University of Michigan
xuefzhao@umich.edu

Dr. Deyou Zheng
Albert Einstein College of Medicine
deyou.zheng@einstein.yu.edu

Mr. Chenchen Zhu
EMBL
chenchen.zhu@embl.de

Mr. Junjie Zhu
Stanford University
jjzhu@stanford.edu

Dr. Katie Zobeck
Agilent Technologies
katie.zobeck@agilent.com

Power your next big breakthrough.

Discover sequencing power
for every scale, and every lab.

Find the right sequencer
to fit your every need.
www.illumina.com/power

For Research Use Only. Not for use in diagnostic procedures.

illumina®

The New Solution for Sequencing ALL Your ChIP-Seq Samples

Accel-NGS® 2S Plus DNA Library Kit

Breaking the Limits of ChIP-Seq Sample Prep

- DNA inputs from 10 pg to 250 ng
- Accurate peak calling without the noise
- Low bias library preparation
- Compatible with ChIP-Seq samples regardless of quantity

Swift BIOSCIENCES™

www.swiftbiosci.com

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers
BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers
Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
***Helpful tips* – Use PIN# 63160** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail and printing
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar
Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late (Cash Only)
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes, ATM
Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, east wing, lower level
PIN#: **Press 63160 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($9.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
US Limousine Service	800-962-2827,ext:3 **(1047)**
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322