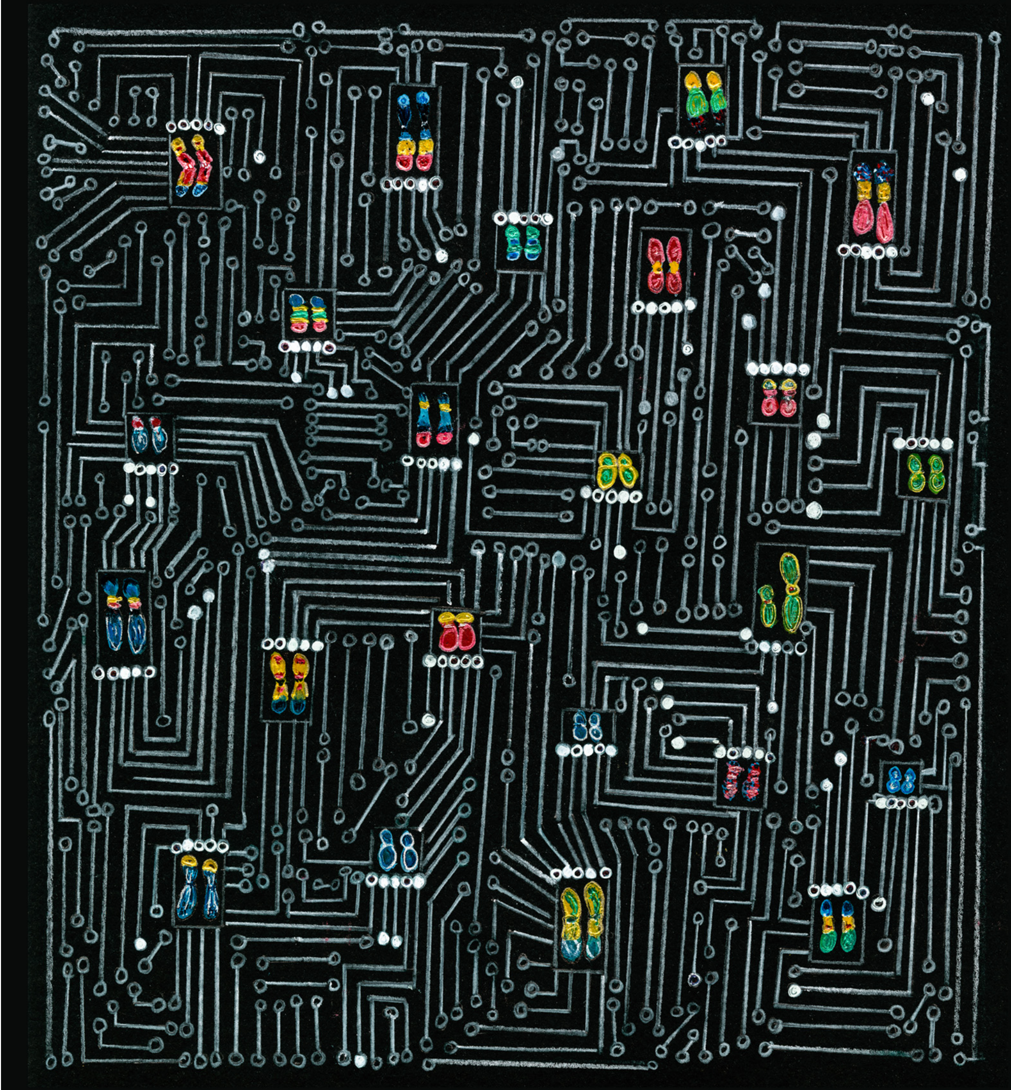
Abstracts of papers presented
at the 2017 meeting on

GENOME INFORMATICS

November 1–November 4, 2017



Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2017 meeting on

GENOME INFORMATICS

November 1–November 4, 2017

Arranged by

Janet Kelso, *Max Planck Institute for Evolutionary Anthropology, Germany*
Aaron Quinlan, *University of Utah*
Melissa Wilson Sayres, *Arizona State University*

This meeting was funded in part by the **National Human Genome Research Institute** and the **National Cancer Institute,** branches of the **National Institutes of Health;** and **Repositive.**

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

Corporate Benefactor

Regeneron

Corporate Sponsors

Agilent Technologies
Bristol-Myers Squibb Company
Calico Labs
Celgene
Genentech, Inc.
Thermo Fisher Scientific
Merck
Monsanto Company
New England BioLabs
Pfizer

Corporate Partners

Alexandria Real Estate
Gilead Sciences
Novartis Institutes for Biomedical Research
Sanofi

The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

Front Cover: Chromosomes on Chip by Manjusha Chintalapati, Max Planck Institute for Evolutionary Anthropology.

GENOME INFORMATICS

Wednesday, November 1– Saturday, November 4, 2017

Wednesday	7:30 pm	**1** Variant Discovery and Genome Assembly
Thursday	9:00 am	**2** Transcriptomics, Alternative Splicing, Gene Predictions
Thursday	1:30 pm	**3** Poster Session I
Thursday	3:30 pm	Keynote Speaker
Thursday	4:30 pm	*Wine and Cheese Party**
Thursday	7:30 pm	**4** Data Curation and Visualization
Friday	9:00 am	**5** Comparative and Metagenomics
Friday	1:30 pm	**6** Epigenomics and Non-coding Genome
Friday	4:30 pm	Keynote Speaker
Friday	5:30 pm	**7** Poster Session II
Friday	7:00 pm	Banquet
Saturday	9:00 am	**8** Personal and Medical Genomics

Workshop: Repositive, *Thursday following morning session*

.

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

WEDNESDAY, November 1—7:30 PM

SESSION 1 VARIANT DISCOVERY AND GENOME ASSEMBLY

Chairpersons: **Laura Clarke,** EMBL-EBI, Cambridge, United Kingdom
Jared Simpson, Ontario Institute for Cancer Research, Toronto, Canada

THURSDAY, November 2—9:00 AM

SESSION 2 TRANSCRIPTOMICS, ALTERNATIVE SPLICING, GENE PREDICTIONS

Chairpersons: **Mihaela Pertea,** Johns Hopkins University, Baltimore, Maryland
Oliver Stegle, EMBL-EBI, Cambridge, United Kingdom

THURSDAY, November 2—1:30 PM

SESSION 3 POSTER SESSION I

Sleuth-ALR—Improving estimation of Seq differential analysis using compositional data analysis with Sleuth
Warren A. McGee, Harold Pimentel, Lior Pachter, Jane Y. Wu.
Presenter affiliation: Northwestern University, Chicago, Illinois. 107

A variant by any other name... Ensembl's Variant Recoder
William McLaren, Sarah Hunt, Fiona Cunningham.
Presenter affiliation: European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom. 108

THURSDAY, November 2—3:30 PM

KEYNOTE SPEAKER

Maricel Kann
University of Maryland

"A protein-domain approach for the analysis of disease mutations"

THURSDAY, November 2—4:30 PM

Wine and Cheese Party

THURSDAY, November 2—7:30 PM

SESSION 4 DATA CURATION AND VISUALIZATION

Chairpersons: **Gabor Marth,** University of Utah, Salt Lake City
Ann Loraine, University of North Carolina, Charlotte

Gabor Marth.
Presenter affiliation: University of Utah, Salt Lake City, Utah.

Large-scale search of short-read sequencing experiments
Brad Solomon, Carl Kingsford.
Presenter affiliation: School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. 109

FRIDAY, November 3—9:00 AM

SESSION 5 COMPARATIVE AND METAGENOMICS

Chairpersons: **Paul Flicek,** EMBL-EBI, Hinxton, United Kingdom
Holly Bik, University of California, Riverside

FRIDAY, November 3—1:30 PM

SESSION 6 EPIGENOMICS AND NON-CODING GENOME

Chairpersons: **Elena Rivas,** Harvard University, Cambridge, Massachusetts
Adam Siepel, Cold Spring Harbor Laboratory, New York

FRIDAY, November 3—4:30 PM

KEYNOTE SPEAKER

Lior Pachter
University of California, Berkeley

“Post-procrustean bioinformatics”

FRIDAY, November 3—5:30 PM

SESSION 7 POSTER SESSION II

Jtracker—Workflow management and execution backed on Git repository with full provenance
Junjun Zhang, Linda Xiang, Brice Aminou, Chen Chen, Lincoln Stein, Vincent Ferretti.
Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada. 218

An improved assembly identifies new features of cucumber genome
Zhonghua Zhang.
Presenter affiliation: Chinese Academy of Agricultural Sciences, Beijing, China. 219

The genome organization of an autotetraploid potato（*Solanum tuberosum* L.）
Qian Zhou, Wu Huang, Chunzhi Zhang, Zhonghua Zhang, Sanwen Huang.
Presenter affiliation: Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China; Agricultural Genomics Institute At Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. 220

An integrative roadmap to PAX3 target gene networks in melanocytes and melanoma
Kirby A. Ziegler, Alan Underhill.
Presenter affiliation: University of Alberta, Edmonton, Canada. 221

Hybrid assembly of challenging genomes with MaSuRCA mega-reads
Aleksey V. Zimin, Daniela Puiu, Steven L. Salzberg.
Presenter affiliation: University of Maryland, College Park, Maryland. 222

FRIDAY, November 3

BANQUET

Cocktails 7:00 PM Dinner 7:45 PM

SATURDAY, November 4—9:00 AM

SESSION 8 PERSONAL AND MEDICAL GENOMICS

Chairpersons: **Konrad Karczewski,** Broad Institute, Cambridge, Massachusetts
Suzanne Leal, Baylor College of Medicine, Houston, Texas

AUTHOR INDEX

VARIATION AND ASSEMBLY RESOURCES AT EMBL-EBI

Laura Clarke

EMBL-EBI, Molecular Archives, Cambridge, United Kingdom

EMBL-EBI is home to many resources supporting the community in storing and discovering variation and assembly datasets.

We have hosted an archive for genome assemblies since the emergence of the first sequenced genome. This service is run as part of the European Nucleotide Archive (ENA). It currently stores more than 100,000 bacterial assemblies and around 7,000 eukaryotic assemblies. The services we provide for assemblies include a web and a programmatic submission interface; helpdesk support provides assistance in structuring and describing assembly data; exchanging assemblies with the other INSDC partners, capturing comprehensive global assembly data; and we offer data discovery services using text and sequence, supporting retrieval of assembly data programmatically and through web interfaces.

We provide open and managed access archives for variation data. We launched the European Variation Archive (EVA) three years ago to enable submission of and provide access to large-scale variation datasets. Our current database contains 564 million variants from 240 studies which cover 27 species consisting of more than 300,000 samples. In May 2017, we made a new agreement with the NCBI to share the responsibility for managing data from genetic variation experiments worldwide. The EVA will assign and manage locus accession numbers (Reference SNP rs#) to variants from all non-human species.

The European Genome-phenome Archive (EGA) provides secure long-term storage and distribution of human genetic and phenotypic data for biomedical research. As of August 2017, the EGA has data from over 1.2 million unique samples in over 1,500 studies. The EGA works closely with the Global Alliance for Genomics and Health (GA4GH) to improve data access and enhance data discovery for controlled access data.The UK-BioBank manages health information on over 500,000 individuals. The EGA has recently partnered with the UK Biobank to distribute the phenotype data for 500,000 individuals., enabling the community to use this rich data resource to better understand disease.

We also have a long track record in coordinating and building discovery resources on large- scale projects which generate variation data. We have recently joined with the Jackson Laboratory to develop the PDX Finder, an open repository for the upload and storage of Patients-derived tumour xenograft (PDX) including tumour genetic variation data. The PDX finder will facilitate PDX models discovery and distribution within the cancer community. This will be done by integrating the complex and diverse data associated with PDX mouse models. This project is also ensuring that patient confidentiality is respected and any identifiable genetic and phenotypic data will be securely stored in archives like the EGA.

This presentation will give you an overview of the assembly and variation resources we provide to the community.

← Back

TMVAR 2.0: INTEGRATING INFORMATION ON GENOMIC VARIANTS FROM BIOMEDICAL LITERATURE WITH DBSNP AND CLINVAR

Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Liu

National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD

Understanding the associations between genomic variants and disease and assessing clinical significance is critical for genomic research and precision medicine. Despite past efforts in expert curation, information about most of the 154 million dbSNP reference variants (RS) remains unknown. In contrast, there is a wealth of human knowledge about variants' biological function and impact on disease buried in unstructured literature data. Previous studies have attempted to harvest and unlock such information with text-mining techniques, but have been of limited use because their mutation extraction results were not standardized or integrated with existing curated data.

We present a novel text-mining method to extract variant mentions from the literature and normalize them to corresponding standardized dbSNP RS numbers, which are unique identifiers with aggregated genomic information such as associated gene and clinical significance. Our method demonstrates a high accuracy of ~90% in F-measure and compares favorably to the state of the art in benchmarking tests.

We applied our approach to all abstracts in PubMed, and validated the results by verifying that each text-mined SNV-gene pair matched the dbSNP annotation based on genomic position, and by analyzing variants curated in ClinVar. Our analysis revealed a plethora of novel (ie not in ClinVar) variant-disease associations, involving 41,889 RS numbers and 9,151 genes. Moreover, our results include 12,462 variants in 3,849 genes which are both predicted to be deleterious and are rare in the general population (MAF <= 0.01).

We ranked and prioritized our results, selecting a set of 10 ultra-rare variants (missense, frameshift, or nonsense) for manual curation. Several fell within genes on the ACMG's list of 58 genes in their guidelines for reporting incidental findings. Curation showed that all 10 variants, though not in ClinVar, were described in the literature as being associated with disease or cancer.

Our findings demonstrate that we can identify many high-impact variants simply by systematically surveying the literature in this way. Our results can be combined with dbSNP and ClinVar data to prioritize and rank variants according to functional consequence, allele frequency, gene annotation, and clinical significance, and therefore to gain insight into the effects variants on biological function and disease. By combining the automatically-extracted information from PubMed articles with existing variant database annotations, we can therefore significantly aid human efforts to curate and prioritize variants in genomic research.

← Back

IDENTIFICATION AND CORRECTION OF PROBLEMATIC COPY NUMBER CALLS IN TCGA

Smruthy Sivakumar[1,2], F Anthony San Lucas[1], Jerry Fowler[1], Paul Scheet[1,2]

[1]UT MD Anderson Cancer Center, Department of Epidemiology, Houston, TX, [2]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Quantitative Sciences, Houston, TX

The copy number variant (CNV) analysis pipeline of the TCGA consortium uses a circular binary segmentation method to report genomic regions and their segment mean copy number (CN) estimates from SNP genotyping arrays. Calibration of this process relies on accurate identification of non-aberrant regions of the genome to establish a baseline signal intensity representative of neutral CN. However, tumor samples exhibiting high levels of genomic instability pose a challenge for such analyses. The resulting poor calibration in the TCGA pipeline leads to erroneous CNV calls, with obvious examples of highly aberrant chromosomes (by visual inspection) being missed by the automated calling procedure and instead CNVs called across normal-appearing chromosome arms. Here we attempt to address this problem by integrating an orthogonal data source to triangulate regions more likely to be copy-neutral and then apply a correction procedure to rescue problematic cases systematically. To do so, we applied a highly sensitive allelic imbalance (AI) detector, hapLOH, to reveal allele frequency patterns consistent with acquired CNVs. In our preliminary analysis of 1079 breast (BRCA), 501 lung (LUAD) and 145 pancreatic (PAAD) cases from the Broad GDAC Firehose pipeline, we attempted corrections on 275 BRCA, 196 LUAD and 27 PAAD cases showing discordant calls between hapLOH and the deposited TCGA CNV calls. We then applied a three-step procedure to correct their CNV calls: (I) identify, using hapLOH, regions within the tumor genome that do not carry a CNV; (II) calculate a new weighted mean normal CN using the TCGA CNV estimates within these “normal” regions; (III) adjust the TCGA CNV calls using the newly estimated normal CN. A large majority of problematic cases arose from a high AI burden, where this procedure successfully corrected 67% of these cases. Interestingly, our analyses revealed that TCGA underestimated the “normal” CN in 81% of the problematic cases, which caused an overestimation of gain events and an underestimation of loss events. Our approach highlights the importance of integrating multiple data types (AI and inferred CNVs) for more robust automated inference procedures, which can run amok with single sources of data. These results should support exploration of more rigorous methods across all the cancer types in TCGA in order to improve downstream analyses and empirical discoveries, including clinical evaluations and CN-derived signatures.

SEQUENCE PRESENCE-ABSENCE DETECTION IN ASSEMBLY PAIRWISE COMPARISON WITH SCANPAV

Francesca Giordano[1], Maximilian R Stammnitz[2], Paul A Kitts[3], Elizabeth P Murchison[2], Zemin Ning[1]

[1]Wellcome Trust Sanger Institute, HPAG, Hinxton, United Kingdom, [2]University of Cambridge, Veterinary Medicine, Cambridge, United Kingdom, [3]National Center for Biotechnology information, National Library of Medicine, Bethesda, MD

The past decade has witnessed the rise of numerous new sequencing technologies as well as the development of novel de novo assembly pipelines and strategies. The exponential increase in the number of sequenced human and non-human genomes that follows demands a similar advancement in the techniques used to assess the quality and completeness of the newly generated assemblies. Here, we present scanPAV, a new pipeline that extracts Present-Absent sequences from an assembly pairwise comparison. Based on the human reference GRCh38, we demonstrate the use of scanPAV as a genome completeness assessment tool by extracting missing sequences from five publicly available human assemblies generated with data from Sanger, Illumina, PacBio and Oxford Nanopore platforms. In addition, we show how a scanPAV analysis can detect real Presence-Absence polymorphism between genomes and detect foreign dna contamination by presenting the analysis of six Tasmanian Devil's normal and tumorous samples compared with two Devil references.
The pipeline can be downloaded at
https://sourceforge.net/projects/phusion2/files/scanPAV/

← Back

MASTERING VARIANT CALLING OF SNPs AND SMALL INDELS WITH DEEP NEURAL NETWORKS

Ryan Poplin[1], Allen Day[1], Jojo Dijamco[2], Nam Nguyen[2], Dion Loy[2], Cory Y McLean[1], Mark DePristo[1]

[1]Google, Inc., Mountain View, CA, [2]Verily Life Sciences, LLC, Mountain View, CA

Next-generation sequencing (NGS) is a rapidly evolving set of technologies that can be used to determine the sequence of an individual's genome. The most common application of NGS technologies is to call genetic variants present in an individual using billions of short, errorful sequence reads. Despite more than a decade of effort and thousands of dedicated researchers, the hand-crafted and parameterized statistical models used for variant calling still produce thousands of false positive and false negative variants in each genome.

We introduce DeepVariant, a deep convolutional neural network (i.e., a deep learning machine) that can call detect variants in aligned next-generation sequencing read data by learning statistical relationships (likelihoods) between images of read pileups around putative variant sites and ground-truth genotype variant calls. This machine outperforms all existing tools, winning the "highest performance" award for SNPs in an FDA-administered variant calling challenge. The learned model generalizes across genome builds, to other ploidies, and even to other kingdoms of life, thus allowing non-human sequencing projects to benefit from the wealth of human ground-truth data.

We further show that DeepVariant can learn to call variants in a variety of sequencing technologies and experimental designs, from 10X Genomics whole genomes to Ion Ampliseq exomes, and is sensitive even in low-coverage contexts. DeepVariant represents a significant step towards the broader opportunity to shift from expert-driven statistical modeling to more automatable deep learning approaches when developing software to interpret biological instrumentation data.

← Back

CAN NANOPORE SEQUENCING FINALLY FINISH THE HUMAN GENOME?

Sergey Koren[1], Brian P Walenz[1], Arang Rhie[1], Alexander T Dilthey[1], NA12878 Consortium[2], Adam M Phillippy[1]

[1]National Institutes of Health, NHGRI, Bethesda, MD, [2]University of Nottingham, School of Life Sciences, Nottinghamshire, United Kingdom

A complete and accurate genome sequence forms the basis of all downstream genomic analyses. However, even the human reference genome remains incomplete, which affects the quality of experiments and can mask true genomic variations. For most other species, high-quality reference genomes do not exist. Long-read sequencing technologies from Pacific Biosciences and Oxford Nanopore have begun to correct this deficiency and are enabling the automated reconstruction of reference-quality genomes at relatively low cost. In a collaborative effort, we sequenced the NA12878 human genome using Oxford Nanopre MinIONs and assembled using Canu. The sequencing set includes 5-fold coverage of ‘ultra-long’ reads with an N50 of >100 kbp and max length >800 kbp. Despite the low coverage, the assembly contiguity (NG50 ~6.4 Mb) exceeds that of similar coverage assemblies using other long-read technologies. The ultra-long reads enable the complete phasing across the MHC in a single contig and close large gaps (>50 kbp) in the existing reference assembly. Additionally, we model expected assembly contiguity and predict 30-fold coverage of ultra-long sequences can exceed a 40 Mbp NG50 and match the contiguity of the current reference. Further combination of these technologies with complementary scaffolding and phasing approaches such as chromatin conformation capture (Hi-C) may soon enable the complete reconstruction of vertebrate haplotypes.

Canu source code and pre-compiled binaries are freely available under a GPLv2 license from https://github.com/marbl/canu. The complete NA12878 dataset including assembly and raw signal is available as an Amazon Web Services Open Dataset at: https://github.com/nanopore-wgs-consortium/NA12878. The ‘Cliveome’ is available from https://github.com/nanoporetech/ONT-HG1.

← Back

GRAPH-BASED DISCOVERY OF COMPLEX *DE NOVO* STRUCTURAL MUTATIONS IN *P. FALCIPARUM* EXPERIMENTAL CROSSES

Kiran V Garimella[1], Isaac Turner[1], Zamin Iqbal[1,2], Gil McVean[1]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]Wellcome Genome Campus, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom

The malaria parasite *Plasmodium falciparum* is capable of diversifying its antigenic repertoire through non-allelic homologous recombination (NAHR), presenting as successive recombinations within a several kilobase window. Such complex *de novo* structural variants (SVs) are difficult to identify in 2nd-gen sequencing data. Antigenic loci tend to be highly diverse; often the sampled haplotype bears little resemblance to the reference. Reads may not align correctly or at all, leading to mischaracterization of putative mutations. While *de novo* assembly can be a fruitful alternative, NAHR events may be predisposed to occur in highly homologous repetitive regions. Short reads often lack sufficient genomic context for an assembler to reconstruct these loci correctly. 3rd-gen sequencing can resolve larger SVs, but obtaining long-read data for hundreds of moderate-sized (~23 Mb) genomes is impractical.

We address these three concerns with novel software methods based on adding connectivity information ("links") to a de Bruijn graph (LdBG). By aligning long haplotypes to an assembly graph and noting the edges utilized at ambiguous graph junctions, we can establish greater genomic context and traverse through repeats longer than short read lengths. Using long haplotypes from parental genomes in a pedigree study provides a means of discovering large SVs and identifying parent of origin without bias towards a reference genome.

We demonstrated our method on four *P. falciparum* crosses (six parents, 150 children). We obtained short-read Illumina data (76-100 bp, 75-150x coverage, GAII to HiSeq 2000 platforms) for all samples. We constructed links from high quality PacBio RSII-based draft assemblies for all six parents. Using our LdBG approach, contig N50 lengths in the children increase by an order of magnitude, permitting vastly easier identification of NAHR events. We find ~30 such mutations across all 150 children. These long NAHR-containing contigs permit base-pair resolution of breakpoints and capture successive recombinations in a single event. Nearly all events are found in low-complexity sub-telomeric regions of the genome.

Our method permits discovery of a class of *de novo* structural variation with clinical relevance and in repetitive regions that are normally inaccessible by traditional genomic analysis. Furthermore, we are able to leverage long-read sequencing data from select samples to improve the assemblies of all members of the cohort. This can provide an enormous cost savings for genomic studies of pedigrees.

← Back

LARGE SCALE GENOMICS WITH SCALABLE REFERENCE GRAPHS

Andre Kahles, Harun Mustafa, Amir Joudaki, Gunnar Rätsch

ETH Zurich, Department of Computer Science, Zurich, Switzerland

Technological advancements in high-throughput sequencing have ushered in a new era of genomics. Even full human genomes or metagenomes of complex microbial communities can now be sequenced at the scale of several thousands within the context of a single project. Traditional methods for the storage and analysis of biological sequencing data increasingly fail to keep up with this growth and new approaches to analyze thousands of genomes in an integrative manner are needed. Thus, representing the genetic variation encoded in a large set of samples and opening it up for integrative analyses is a central research goal. Our main motivation is to capture rare or hitherto unseen genetic variation as it is commonly observed during the sequencing of cancer samples or in the analysis of microbial metagenomes. Especially the latter suffers from incomplete and often biased reference data repositories.

We present a novel, highly efficient approach to combine a large set of (meta-)genomes with raw sequencing data and variant calls into a sparse representation that can be comprehensively annotated and efficiently queried. Building on techniques from genome assembly and text compression, we encode all sequence information and associated metadata in a succinct k-mer based colored assembly graph, which not only represents single genomes but also captures inter- and intra-genome variability. The graph representation is structured as a self-index that can be used for alignment and classification of reads arising from sequencing of novel samples.

As the constructed graph structure can leverage information from both reference genomes and sequencing samples, it provides access to rare observations not yet present in reference databases. It is designed to dynamically integrate further knowledge over time (e.g., to accumulate information over many heterogeneous data sources), to provide greater sensitivity to detect unseen or rarely seen variants. Our implementation can encode a human genome together with all sequence variants collected in the gnomAD project in under 2 hours, resulting in a fully indexed reference graph of less than 3GB. We will also present applications in metagenomics, encoding over 50,000 viral sequences and show strategies to visualize the large data structures.

← Back

A NEW COMPREHENSIVE HUMAN GENE CATALOG

Mihaela Pertea[1,2], Geo M Pertea[1], Steven L Salzberg[1,2,3,4]

[1]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [4]Johns Hopkins University, Biostatistics, Baltimore, MD

Recently sequenced collections of human transcript data provide an opportunity to re-evaluate and possibly expand the catalog of human protein-coding genes and noncoding transcripts. We assembled an unprecedentedly large set of 9,795 RNA-seq samples containing almost 900 million reads with the StringTie transcript assembler (Pertea, Pertea et al. 2015), merged the results with gffcompare (Pertea, Kim et al. 2016), and applied a series of computational filters to build a new human gene catalog. The source of this data was the genotype-tissue expression (GTEx) study (GTEx Consortium 2015), which conducted RNA sequencing experiments on samples from dozens of tissues collected from hundreds of individuals. Our new human gene database, CHESS 1.0, contains 39,588 genes, of which 21,631 are protein-coding and 17,957 are noncoding, and a total of 352,014 transcripts, for an average of 9.0 transcripts per gene. Our expanded gene list includes 3,575 novel genes (1,577 coding and 1,998 noncoding) and 194,248 novel splice variants as compared to the RefSeq and GENCODE. We detected over 30 million additional transcripts at more than 650,000 sites, nearly all of which are likely to be nonfunctional, revealing a heretofore unappreciated amount of transcriptional noise in human cells. The CHESS 1.0 database of genes and transcripts, which is freely available at http://ccb.jhu.edu/chess, will be updated over time as new evidence emerges.

GTEx Consortium (2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science 348(6235): 648-660.
Pertea, M., D. Kim, G. M. Pertea, J. T. Leek and S. L. Salzberg (2016). "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown." Nat Protoc 11(9): 1650-1667.
Pertea, M., G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell and S. L. Salzberg (2015). "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." Nat Biotechnol 33(3): 290-295.

← Back

ISOCON: A NOVEL ALGORITHM COMBINED WITH TARGETED TRANSCRIPTOME SEQUENCING OF MULTICOPY GENE FAMILIES TRACES THE ORIGINS OF HIGHLY SIMILAR TRANSCRIPTS TO INDIVIDUAL GENE COPIES.

Kristoffer Sahlin*, Marta Tomaszkiewicz*, Kateryna D Makova, Paul Medvedev

Pennsylvania State University, University Park, PA

Detecting novel transcripts is particularly challenging with Illumina short-read transcriptome sequencing, failing to determine full transcript sequences in the presence of alternative splicing or duplicate genes. Long Pacific Biosciences (PacBio) reads from the Iso-Seq protocol can determine the exon order of long transcripts, however errors make distinguishing transcripts originating from duplicate gene copies difficult. A particularly challenging case is Y chromosome ampliconic gene families, each of which contains several nearly identical (up to 99.99%) gene copies. The inability of previous approaches to distinguish their respective sequences has limited our understanding of the evolution of the primate Y chromosome and the causes of male infertility disorders, for which these genes are crucial.

We present a novel approach that combines experimental and computational techniques to determine the full-length transcripts of multicopy gene families. We design RT-PCR primer pairs from the most complete conserved coding region of each gene family, and then sequence with PacBio Iso-Seq protocol. We then develop an algorithm called IsoCon to correct sequencing errors and cluster the reads into a non-redundant set of transcripts. Unlike PacBio's own Iso-Seq clustering algorithm (ICE, Gordon et al., 2015), which use a majority consensus rule, IsoCon probabilistically models the error profile and uses a statistical test to filter out redundant transcripts.

To demonstrate the power of our method, we first used simulations to show that IsoCon has both higher recall and drastically improved precision over ICE, over a range of read depths, error rates and transcript similarities. Next, we used our experimental protocol to generate data for nine ampliconic gene families from testes samples of two human individuals. Using reference transcripts present in the raw reads, IsoCon is able to capture >99% of abundant transcripts, while ICE only captures about 50%. Experimental validation using Illumina sequencing shows that >99.0% of IsoCon transcript positions are supported by Illumina alignments (with >62% of the transcripts fully supported), while only 90.2% of ICE transcript positions are supported (with only 1.1% of the transcripts having full support). Several fully supported transcripts are absent from existing databases or only predicted there in silico.The accuracy of IsoCon further allows us to separate highly similar transcripts from the same gene family into gene/pseudogene copies and their spliced variants.

← Back

THE LANDSCAPE OF ISOFORM SWITCHES IN HUMAN CANCERS

Kristoffer Vitting-Seerup[1,2], Jette Bornholdt[1,2], Albin Sandelin[1,2]

[1]University of Copenhagen, Bioinformatics Centre, Department of Biology, Copenhagen, Denmark, [2]University of Copenhagen, Biotech Research & Innovation Centre (BRIC), Copenhagen, Denmark

Differential transcript isoform usage, often called isoform switches, are known to play important roles in all aspects of development, homeostasis and diseases and are especially prevalent in cancer. It is therefore surprising that even though bioinformatics tools for isoform quantification have been around since 2010 systematic analysis of isoform switches and their potential biological consequences are to a large extend lacking.

To this end we developed IsoformSwitchAnalyzeR, and easy to use R package which facilitates identification and analysis of isoform switches with predicted consequences from RNAseq data. By applying IsoformSwitchAnalyzeR to data from >5500 cancer patients covering 12 solid cancer types we find isoform switches with predicted consequence are common affecting ~19% of multi-isoform genes. Isoform switches resulting in loss of protein domains were amongst the most common consequences highlighting their potential biological importance. By coupling isoform switches to patient survival, we also find isoform switches to be clinically relevant: 31 isoform switches can serve as biomarkers of poor prognosis – independent of cancer type origin.

In summary, our results indicate that isoform switches with predicted functional consequences are both common and important in dysfunctional cells, illustrating the potential of augmenting gene-level analysis with isoform-level analysis both in cancer and other diseases.

← Back

SQUID: TRANSCRIPTOMIC STRUCTURAL VARIATION DETECTION FROM RNA-SEQ

Cong Ma, Mingfu Shao, Carl Kingsford

Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA

Transcripts are frequently modified by structural variants, which leads to either a fused transcript of two genes (known as a fusion gene) or an insertion of a previously non-transcribing sequence into a transcript. These modifications, called transcriptomic structural variants (TSV), can lead to drastic changes in a downstream product. Several fusion-gene TSVs have been identified as cancer drivers, while much less is known about non-fusion-gene TSVs, likely partially due to the difficulty of detecting them. Detecting TSVs, especially in cancer tumor sequencing, is therefore an important and challenging computational problem. It is even more challenging since often only RNA-seq measurements are available from which to infer TSVs.

We introduce SQUID, a novel algorithm and its implementation, to accurately predict both fusion-gene and non-fusion-gene TSVs from RNA-seq alignments. SQUID takes the unique approach of attempting to reconstruct an underlying genome sequence that best explains the observed RNA-seq reads by rearranging genome segments. The inferred rearrangement is used to predict TSVs. By unifying both concordant alignments and discordant read alignments into one model for rearrangement, SQUID achieves high sensitivity with many fewer false positives than other approaches. Tested on simulation data, SQUID is 20% more precise than other tested methods or pipelines with similar sensitivity. Tested on two previously studied cell lines, high accuracy is also observed for non-fusion-gene TSV detection.

We apply SQUID to detect TSVs on TCGA tumor samples and observe that non-fusion-gene TSVs are more likely to be intra-chromosomal than fusion-gene TSVs when predicted this way in several cancer types. Prostate cancer stands out to have the largest inter-chromosomal against intra-chromosomal TSV ratio. We also quantify the propensity for breakpoint partners to be reused, and observe that more than half of recurring breakpoints are rejoined with the same partner in TSVs across the studied samples. Using SQUID, we identify several novel TSVs involving tumor suppressor genes. For example, the *ZFHX3* gene is involved in non-fusion-gene TSVs in two samples, the *ASXL1* gene is involved in a non-fusion-gene TSV in one sample and a fusion-gene TSV in another. It is reasonable to suspect that these TSVs may lead to loss-of-function in the corresponding tumor suppressor genes and play a role in tumorgenesis.

SQUID is available at https://github.com/Kingsford-Group/squid as open source and a pre-print describing the algorithm and our analysis of TCGA samples is available at http://www.biorxiv.org/content/early/2017/07/20/162776.

← Back

TRANSCRIPTOME-GUIDED GENOMIC ALIGNMENT AND ANALYSIS

Thomas D Wu

Genentech, Inc., Bioinformatics & Computational Biology, South San Francisco, CA

RNA-Seq reads can be aligned either to a genome or transcriptome, with their respective advantages and disadvantages; often pipelines perform both types of alignments separately. We propose instead an integrated strategy called transcriptome-guided genomic alignment (TGA) that generates genomic alignments, but uses a transcriptome to facilitate the process. We have implemented TGA as an option in our alignment program GSNAP, which is already one of the most accurate RNA-Seq aligners. This option allows for either a transcriptome-sufficient or transcriptome-boosted mode, with the first mode obviating a genomic search if an acceptable transcript alignment is found and the second mode performing a genomic search that can be made faster by requiring alignments to be equivalent or better than the transcriptome alignments.

Implementing TGA requires utility programs to generate suffix array and hash table indices for the genome, a suffix array for the transcriptome, and an index of transcript exon coordinates. We have addressed various technical issues, including the introduction of ambiguity when a read satisfies multiple alternate transcripts; handling of indels in transcript alignments; concordance of paired-end reads at the transcript level; and post-alignment identification of transcripts from complex genomic alignments.

TGA provides several advantages over separate transcriptome and genomic alignments. First, TGA increases the speed of genomic alignment. Second, TGA improves the accuracy of genomic alignments, especially at the ends of reads, where soft clips are often added due to uncertainty about whether mismatches derive from polymorphisms or an intron. Extending RNA-Seq alignments to the ends of reads improves the accuracy of genomic variant detection by reducing inherent bias toward reference alleles. Third, TGA facilitates analysis at the transcript level, such as quantification of isoform expression and identification of gene fusions, including fusion junctions that do not occur at annotated exon boundaries. Accordingly, we have developed analysis programs GEXPRESS and GFUSION that can analyze TGA results. GEXPRESS in particular performs a novel analysis of the transcriptome to identify expected unique alignments based on the read length.

We have performed preliminary tests on simulated paired-end reads, where reads are generated from RefSeq transcripts, but the transcriptome is drawn from Ensembl annotation. This distinction tests for cases where the transcriptome is incomplete or different from the transcripts present in the cell mRNA complement. Our results show that transcriptome-sufficient alignment increases GSNAP alignment speed by a factor of 2.8 and transcriptome-boosted alignment by a factor of 1.2 over genomic alignment, while improving the quality of genomic alignments.

HOW TO CREATE A WHOLE-GENOME HUMAN HOMOLOGY MAP IN AROUND A MINUTE

Chirag Jain[1,2], Sergey Koren[1], Alexander Dilthey[1], Srinivas Aluru[2], Adam M Phillippy[1]

[1]National Human Genome Research Institute, Genome Informatics Section, Bethesda, MD, [2]Georgia Institute of Technology, Computational Science and Engineering, Atlanta, GA

Emerging single-molecule sequencing technologies such as Oxford Nanopore have revived interest in long-read mapping algorithms. Alignment-based seed-and-extend methods demonstrate good accuracy, but face limited scalability, while faster alignment-free methods typically trade decreased precision for efficiency. Previously we have demonstrated with MashMap that the combination of a minimizer index with MinHash identity estimation can provide both scalability and precision for approximate, end-to-end read mapping.

In this new work, we have generalized our technique for local alignment problems, including split-read, assembly-to-genome, and genome-to-genome mappings. We develop a theoretical framework that defines the types of mapping targets we uncover, establish probabilistic estimates of sensitivity, and demonstrate tolerance for alignment error rates up to 20%. Additionally, we have developed an efficient plane-sweep algorithm for the prioritization and/or filtering of repetitive mappings. As a result, we were able to map a de novo human draft assembly to the human reference in just ~1 minute using 8 threads, or generate a human genome homology map in similar time. This compares favorably to the established Nucmer whole-genome alignment program, which required 23 hours to produce a similar result (no threading). These techniques will be useful for rapid mapping of long reads to large reference databases, evaluation of draft assemblies versus a reference sequence, and the scalable construction of whole-genome homology maps.

← Back

SIMULATION AND ANALYSIS TOOLS FOR SINGLE-CELL RNA SEQUENCING DATA

Luke Zappia[1,2], Belinda Phipson[1], Alicia Oshlack[1,2]

[1]Murdoch Childrens Research Institute, Bioinformatics, Melbourne, Australia, [2]The University of Melbourne, School of Biosciences, Melbourne, Australia

Single-cell RNA sequencing (scRNA-seq) is rapidly becoming a tool of choice for biologists who wish to investigate gene expression. In contrast to traditional bulk RNA-seq experiments, which measure expression averaged across millions of cells, single-cell experiments can be used to observe how genes are expressed in individual cells. Along with the dramatic increase in resolution provided by scRNA-seq comes an array of bioinformatics challenges. Single-cell data is relatively sparse (for both biological and technical reasons), quality control is difficult and it is unclear how to replicate measurements. Researchers have risen to address these challenges and there are currently more than 125 software tools available for analysing scRNA-seq data. We have catalogued these software tools in the scRNA-tools database (www.scRNA-tools.org). Analysis of this database shows that there are now methods available for a wide range of tasks, from pre-processing unique molecular identifiers to detecting allele-specific expression. However, the biggest areas of development have been in clustering cells to identify cell types and ordering of cells to understand dynamic processes. We also find that the R statistical programming language is the most popular platform for scRNA-seq analysis tools, followed by Python, and that the majority of tools have been described in peer-reviewed papers or preprints and are available under open-source software licenses.

With the ever increasing number of analysis methods available it is important to be able to assess and compare the performance, quality and limitations of an analysis tool. This is often done, at least in part, by testing methods on simulated datasets where the true answers are known. Unfortunately, current scRNA-seq simulations are frequently poorly documented, not reproducible and do not demonstrate similarity to real data or experimental designs. To address these concerns we have developed Splatter, a Bioconductor R package for reproducible simulation of scRNA-seq datasets. Splatter is a simulation framework that currently includes four previously published simulation models, allowing users to estimate parameters from real data in order to easily generate realistic synthetic scRNA-seq datasets. Here we discuss some of the challenges of simulating scRNA-seq data and present a comparison of the simulation methods available in Splatter (bioconductor.org/packages/splatter). As part of Splatter we also introduce our own simulation model, Splat, capable of reproducing scRNA-seq datasets with multiple groups of cells, differentiation paths or batch effects.

← Back

GENOMIC ANALYSIS OF GERMLINE AND SOMATIC VARIATION IN HIGH-GRADE SEROUS OVARIAN CANCER

Aaron W Adamson[1], Mihaela C Cristea[2], Lucille A Leong[2], Robert Morgan[2], Mark T Wakabayashi[3], Ernest S Han[3], Thanh H Dellinger[3], Paul S Lin[3], Amy A Hakim[3], Sharon Wilczynski[4], Linda Steele[1], Charles D Warden[5], Shu Tao[5], Yaun Chun Ding[1], Susan L Neuhausen[1]

[1]City of Hope National Medical Center, Population Sciences, Duarte, CA, [2]City of Hope National Medical Center, Medical Oncology, Duarte, CA, [3]City of Hope National Medical Center, Surgery, Duarte, CA, [4]City of Hope National Medical Center, Pathology, Duarte, CA, [5]City of Hope National Medical Center, Molecular and Cellular Biology, Duarte, CA

Background. High-grade serous ovarian cancers (HGSCs) display a high degree of complex genetic alterations involving numerous oncogenes and tumor suppressor genes. Germline mutations in *BRCA1* and *BRCA2* are associated with approximately 15% of HGSC cases and additional rarer genes of high to moderate risk (e.g., *BRIP1*, *PALB2*) have recently been identified. Many of the genes predisposing to ovarian cancer are still likely unknown, as are the combined effects of germline and somatic mutational events on disease-free recurrence. The goal of this study was to identify some of the missing genetic components associated with HGSC.
Methods. We used a targeted capture and next-generation sequencing approach to assess 74 HGSC tumors with matched blood for germline and somatic loss-of-function (LOF) mutations in 599 genes involved in the DNA damage response. Along with the exons and untranslated regions, we included 2kb upstream and 1kb downstream regions of each gene in order to detect variants that potentially disrupt transcription factor and miRNA binding sites. We also performed genome-wide analysis of somatic copy number variation (SCNV) and loss of heterozygosity (LOH) using the OncoScan assay.

Results. As observed in other studies of HGSC, we found that approximately one-third of the tumors had a LOF germline (25%) and/or somatic (12%) mutation in one or more of the 7 homologous recombination (HR) genes: *BRCA1*, *BRCA2*, *BRIP1*, *CHEK2*, *FAM175A*, *MRE11A*, and *PALB2*. Other LOF mutations included somatic mutations in *BLM*, *RB1*, *NF1*, and *PIK3CA* and the same germline mutation in *MSH6* was identified in two individuals. Of note, multiple LOF germline and somatic mutations were observed in genes not associated with ovarian cancer including *PTTG2* (4 cases), *CARD6* (3 cases), *IFNA17* (3 cases), and *CARD8*, *POLK*, and *RRM2B* (two cases each). Most of the tumors (87%) harbored somatic mutations in *TP53*. Thus far, we have performed OncoScan SCNV analysis on half of the tumors and have identified focal homozygous deletions in *BRCA1*, *BRCA2*, *PTEN*, *RB1*, *SLX4*, and *NF1*. Consistent with previous observations, we observed a high percentage of LOH on chromosome 17 (95%), copy number amplification at 8q24 encompassing *MYC* (84%), and copy number loss at 22q13 of (73%). Upon completion of the OncoScan analysis we will perform both unsupervised and supervised hierarchical clustering to determine if the tumors group together according to clinical outcomes or other features. Finally, we will perform GISTIC analysis to identify significantly altered regions unique to each cluster. Taken together, our combined sequencing and SCNV analysis approach has so far found that over 40% of the ovarian cancer patients in this study harbor deleterious mutations in HR genes and would be candidates for therapy with PARP inhibitors.

LGTSEEK - A ROBUST DISTRIBUTABLE LATERAL GENE TRANSFER PIPELINE

Ricky S Adkins, John Mattick, Robin Bromley, Karsten Sieber, David Riley, Kelly Robinson, Anup Mahurkar, Julie Dunning Hotopp

University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD

Lateral gene transfer (LGT), which is synonymous with horizontal gene transfer (HGT), is the transfer of DNA directly from one organism to another organism. In many cases, LGT is initiated by bacterial donors and is evolutionarily advantageous for the recipient organism. While there are many strategies and tools used to identify putative LGT in the consensus genome of recipient organisms, there is little in the way of tools to identify novel integration or donor sites using unassembled sequencing reads. We seek to address this by creating LGTSeek, a pipeline for the detection of putative LGT for any specified donor or recipient genome. LGTSeek can also identify novel donor integration sites if the recipient genome is unknown, or identify novel recipient integration sites if the donor genome is unknown. LGTSeek was implemented using the web-based Ergatis workflow system and is distributed as a Docker container. Using Docker gives the advantage of running LGTSeek on a variety of platforms, given the memory and space requirements are met. The LGTSeek project can be found at http://www.igs.umaryland.edu/labs/lgthgt/analysis/lgt-seek, the Docker image can be found at https://hub.docker.com/r/adkinsrs/lgtseek, and the source code for the LGTSeek pipeline can be downloaded from https://github.com/adkinsrs/ergatis-pipelines/tree/lgtseek.

← Back

SYSTEMATIC ANNOTATION OF REGULATORY ELEMENTS IN BLOOD CELL LINEAGE

Lin An[1], Cheryl A Keller[2], Elisabeth Heuston[4], Belinda Giardine[2], David Bodine[4], Yu Zhang[3], Ross Hardison[2]

[1]The Pennsylvania University, Bioinformatics & Genomics Program, State College, PA, [2]The Pennsylvania University, Department of Biochemistry and Molecular Biology, State College, PA, [3]The Pennsylvania University, Department of Statistics, State College, PA, [4]National Human Genome Research Institute, Genetics and Molecular Biology Branch, Bethesda, MD

Motivation

Epigenetic marks serve as a powerful tool to discover functional regulatory units in the mammalian genome, referred to as cis-regulatory modules (CRM), which play a vital role in cell differentiation. It is important to identify CRMs and related regulatory machinery. As an essential component of the mammalian circulation system, hematopoiesis provides an excellent system to study cis-regulatory network and chromatin state dynamics during blood cell differentiation.

Method

We previously developed the IDEAS method, an Integrative and Discriminative Epigenome Annotation System, to jointly characterize chromatin state dynamics across multiple cell types. We applied the IDEAS method to epigenetic datasets in 15 cell types in the mouse hematopoiesis lineage. To profile functional regulators, we generated 5 histone modification marks: H3K4me1, H3K4me3, H3K27ac, H3K27me3 and H3K36me3. We also included ATAC-seq data for chromatin accessibility and CTCF for transcriptional regulation. Previous hematopoietic lineage studies have shown that these epigenetic marks are closely related to hematopoietic cell differentiation.

Result

We obtained a 17 states segmentation map. The segmentation results show agreement with expression data. We found the current segmentation agrees with lineage relationship, which indicates the regulatory landscape has the ability to explain hematopoiesis development. Moreover, the identified ‘enhancer’ states from IDEAS model are overlapped with experimentally validated enhancers and match the cell-type specific sites. We also found the missing data is partially predictable by the current method as some CTCF enriched regions in the progenitor cell can be imputed based on the information from other cell types.

IN-DEPTH CHARACTERIZATION OF A HALLMARK FOR BALANCING SELECTION: HLA HETEROZYGOTE ADVANTAGE AGAINST HIV-1

Jatin Arora[1], Federica Pierini[1], Paul J McLaren[2], Mary Carrington[3], Jacques Fellay[4], Tobias Lenz[1]

[1]Max Planck Institute for Evolutionary Biology, Evolutionary Ecology, Plön, Germany, [2]JC Wilt Infectious Diseases Research Center, Public Health Agency of Canada, Winnipeg, Canada, [3]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA, [4]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Pathogen-mediated balancing selection is thought to be a key driver of host immunogenetic diversity. A hallmark for balancing selection in humans is the heterozygote advantage at genes of the Human Leukocyte Antigen (HLA), resulting in improved HIV-1 control. However, the actual mechanism through which heterozygotes obtain an advantage is still elusive. It may be conferred by the ability of HLA heterozygotes to present a broader array of viral peptides to immune cells, possibly resulting in more efficient cytotoxic T cell response. Heterozygosity may also simply increase the chance to carry the most protective HLA variants, as individual HLA alleles are known to differ substantially in their effect on HIV-1 control. Here we take advantage of available HLA genotype and set point viral load data from 6,311 HIV-1 patients of European ancestry and find a lower viral load for heterozygotes at HLA-B ($P < 0.001$) and HLA-C ($P = 0.022$). Screening the entire HIV-1 proteome, we observed that patients heterozygous at HLA-B and HLA-C are predicted to bind a broader array of HIV-1 epitopes ($P < 0.001$ for both loci). Interestingly, a patient’s viral load was negatively correlated with the breadth of the patient’s HLA-bound HIV-1 epitope repertoire for HLA-B ($tau = -0.15$, $P < 10-16$), but not for HLA-C ($tau = 0.01$, $P = 0.09$), suggesting that heterozygote advantage at HLA-B is mediated by a quantitative cytotoxic T cell response, but that a different mechanism could be involved at HLA-C. We also analyzed autologous HIV-1 sequence data and observed a significantly higher divergence of HIV-1 strains among HLA-B heterozygous patients compared to homozygotes, suggesting stronger evolutionary pressure from HLA heterozygosity.

← Back

STRUCTURAL VARIATION AND GENOME EVOLUTION IN DOMESTIC DOG

Meharji Arumilli[1,2,3], Marjo Hytönen[1,2,3], Hannes Lohi[1,2,3], Jarkko Salojärvi[4]

[1]Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland, [2]Research Programs Unit, Molecular Neurology, University of Helsinki, Helsinki, Finland, [3]The Folkhälsan Institute of Genetics, University of Helsinki, Helsinki, Finland, [4]Plant Biology, Department of Biosciences, University of Helsinki, Helsinki, Finland

The explosion of genome assemblies has enabled us to perform comparative genomics analysis to better understand the phenotypic evolution of species. Here, we infer the ancestral and recently evolved gene content in dog by performing syntenic analysis with cat (Felis catus), polar bear (Ursus maritimus) and the most recently diverged African wild dog (Lycaon pictus) genomes and study the role of structural variants (SV) in speciation process. We have detected SVs; deletions, duplications, inversions and mobile-element insertions (MEI) in 142 dogs and inferred the ancestral states with eight grey wolf genomes. We identified 1,05,912 SV loci (7,088/dog) of six major classes, including deletions (160 sites/dog), duplications (124/dog), mixed CNVs (76 sites/dog) and inversions (6 sites/dog) while SINES represent the larger fraction of SVs per genome (5,351 sites/dog) followed by LINEs (1,371 sites/dog. The size distribution of SVs varied by class with the mean size of 4.5 Kb for deletions, 8 Kb for duplications, 4 Kb for multi-allelic CNV, 4.6 Kb for inversions 181bp for SINES and 1.6 Kb for LINES together spanning 7.1 % of the dog genome. We observed that duplications exhibit fundamentally different population genetic properties and selection signatures than deletions and MEIs. We found significant over-representation of CNVs in the SINEs/LINEs, compared to segmental duplications. This study explores the structural diversity of canids and identifies the enriched functional categories among the syntenic and tandem regions in dog compared to cat, polar bear and African wild dog. Further, this study presents catalogue of novel and rare SVs identified in the dog genome at greater scale and resolution compared with the previous studies.

← Back

COPY NUMBER AND TUMOR PURITY ESTIMATION FROM TARGETED CELL FREE DNA SEQUENCING DATA

Hossein Asghari[1,2], Ibrahim Numanagić[3], Faraz Hach[2,4]

[1]Simon Fraser University, School of Computing Science, Burnaby, Canada, [2]Vancouver Prostate Centre, Vancouver, Canada, [3]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [4]University of British Columbia, Dept. of Urologic Sciences, Vancouver, Canada

Cell-free DNA (cfDNA) is the DNA obtained from both normal and tumor dying cells that have been shed into the bloodstream. Sequencing of cfDNA, which is itself easily obtainable from the blood plasma, offers inexpensive and non-invasive means for cancer data analysis, especially when compared to the alternatives such as biopsy. cfDNA sequencing can be used to estimate the copy number variation (CNV) for known cancer genes, which can further provide insights into the cancer progression. However, this is not straightforward task, mainly because of the (i) the presence of both normal and tumor DNA in the patient's blood, requiring the estimation of the tumor purity of a sample (a *priori* unavailable), and because of (ii) the low ratio of tumor DNA in the blood plasma. Furthermore, sequencing data is often affected by systematic biases, such as GC content bias and non-uniform coverage.

In order to resolve such problems, we introduce a novel combinatorial optimization framework, whose aim is to predict copy number variations from cfDNA data while estimating the tumor purity at the same time. Our method attempts to find a tumor purity and associated copy number calls by minimizing the difference between the observed coverage obtained from the sequence data and the predicted coverage (accounting for tumor purity) by modeling this as an instance of the Integer Linear Programming problem. We use copy number-neutral genes for coverage normalization, while the observed sample coverage is corrected for GC content and other biases via various smoothing methods, such as LOWES. We applied our framework on various simulated targeted sequencing samples, modeled to cover various levels of tumor content.

Our result shows that the framework is able to successfully predict the underlying copy-number values for any sample with the tumor content above 10%. Even for the lower tumor rates (from 5% to 10%), our framework was able to successfully detect general trends within genes, such as gains or losses.

We further validated our framework on a real-data prostate cancer cfDNA sample, sequenced with our in-house targeted panel. Our predictions match the validated gains and losses available for this sample. Overall, we show that our framework offers a fast approach for calling copy number variations in the cfDNA samples, providing a means for a quick and non-invasive cancer screening of the patients.

← Back

INFERRING ORIGINATING CELL TYPE OF CANCER METASTASES USING THE SPATIAL DISTRIBUTION OF MUTATIONS

Gurnit S Atwal[1], Yulia Rubanova[2], Jeff Wintersinger[2], Quaid D Morris[2,3]

[1]University of Toronto, Molecular Genetics, Toronto, Canada, [2]University of Toronto, Computer Science, Toronto, Canada, [3]University of toronto, Electrical and Computer Engineering, Toronto, Canada

Most cancer-related death follows metastasis, so understanding metastasis is critical to understanding and treating cancer. In 2% to 5% of new cancer diagnoses, only the metastatic lesion is identified resulting in a carcinoma of unknown primary origin[1]. Treating these cases is difficult because cancer therapy is tailored towards the tissue of origin of the primary lesion. Because metastases contain all of the mutations accumulated by the original primary, we propose to identify tissue-of-origin by examining the somatic mutations. Previous work[2] showed i) a strong correlation between epigenetic state of a DNA region and the somatic mutation rate, and ii) because epigenetic state varies by tissue type, that tissue-of-origin could be predicted based on correlations between mutation rate and tissue-specific epigenetic state. In this work, we investigate whether we could forego data from epigenetic marks and infer tissue-of-origin directly from the chromosomal positions of mutations. To infer metastases' originating cell types, we trained a deep neural network on whole-genome sequencing data from the Pan-Cancer Analysis of Whole Genomes (PCAWG). Each tumour was represented as copy-number corrected counts of mutations in 1 Mb bins across the genome. Counts were used as training data for a deep neural network that classifies tumour type. The originating tumour type was accurately inferred across 32 histological categories, with most misclassifications corresponding to confusion between histological categories that correspond to the same type of originating cell. Across the 16 most common tumour types in PCAWG, we achieved a classification accuracy of 90% .

This work demonstrates evidence of a relationship between the spatial distribution of mutations and tumour type.By drawing on this relationship, we developed a novel method for classifying metastases from unknown primary tumour types. Our results not only inform the biological understanding of cancer, but also suggest clinical applications.

1. Varadhachary, G. R. Carcinoma of unknown primary origin. Gastrointest. Cancer Res. 1, 229–35 (2007).
2. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518, 360–364 (2015).

← Back

MODELLING DOUBLE STRAND BREAK HOTSPOTS TO INTERROGATE STRUCTURAL VARIATION IN CANCER

Tracy J Ballinger, Colin A Semple

University of Edinburgh, MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom

Double strand breaks (DSB) occur at high levels in tumour cells due to replication stress, metabolic dysfunction and compromised repair pathways. These breaks can lead to a variety of structural variation, often with functional consequences. In particular, it is thought that structural variants (SVs) involving DSB may lead to the activation of oncogenes or de-activation of tumor suppressors, and contribute to tumour development. It is not known why particular loci are subject to recurrent DSBs, or DSB ‘hotspots’, but recent studies suggest that chromatin accessibility, replication timing, nuclear organisation and other genomic features may all play a role. Similarly, the most comprehensive studies of SVs in tumours have discovered compelling patterns, but could not discern mutational bias from selection driving SV hotspots. Currently, we lack a quantitative description of the mutational landscape of cancer cells, defining the contributions of the inherent fragility of the underlying chromatin, and of positive selection for uncontrolled cell growth. Using such a description we can estimate the functional importance of particular SVs and the genes they affect in tumour progression.

We have generated accurate quantitative models of DSB susceptibility incorporating a host of genomic and epigenomic features to predict DSB frequencies in several cell lines. We find that replication timing, chromatin accessibility and a variety of other features are important predictors for genome fragility, and that using a large combination of features, we are able to model the landscape of DSB susceptibility across the genome. Additionally, we find that most SV hotspots in cancer genomes are explicable given our models, arising due to the inherent fragility of those regions, while others may be the result of selection.

← Back

GENETIC VARIANTS OVER GENERATIONS: SPARSITY-CONSTRAINED OPTIMIZATION TOOLS FOR STRUCTURAL VARIANT DETECTION

Mario Banuelos, Lasith Adhikari, Rubi Almanza, Roummel F Marcia, Suzanne Sindi

University of California, Merced, Applied Mathematics, Merced, CA

Structural variants (SVs) – rearrangements of an individuals' genome – are an important source of genetic diversity and disease in humans and other mammalian species. SVs are typically identified by comparing fragments of DNA from a test genome to a given reference genome. Discordant arrangements of these DNA fragments signal structural differences between the test and reference genome, but errors in both the sequencing and mapping lead to high false positive rates in SV prediction. However, the presence of DNA from multiple related individuals offers the promise to substantially reduce the false positive rate of prediction.

We develop a computational method to predict germline structural variants given genomic DNA in a family lineage. Because SVs are thought to have a low rate of spontaneous appearance, any SV present in an offspring must have occurred in at least one of their parents. We seek to improve standard methods in three main ways. First, to reduce false-positive predictions, we enforce properties of inheritance to constrain the space of admissible SVs and concurrently predict SVs in family lineages. Second, we predict variant zygosity (i.e., homozygous or heterozygous). Third, we utilize a gradient-based alternating optimization approach and constrain our solution with a sparsity-promoting ι_1 penalty (since variants should be rare). We present results on both simulated genomes, parent-child trios, and larger family lineages.

← Back

EXPLORING THE SEQUENCE COMPOSITION, FUNCTIONAL CAPACITY, AND REGULATORY ROLE OF LARGE TANDEM REPEATS AND THEIR ADJACENT SEQUENCES IN REGENERATION

Sofia N Barreira, Andreas D Baxevanis

National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

The repetitive DNA content found in the majority of sequenced genomes significantly surpasses that of the coding sequence component. Repetitive DNA has been implicated in chromatin organization, gene expression regulation, genome replication, cell proliferation, and the maintenance of genome integrity. Establishing the organization and distribution of repetitive DNA within a genome is crucial to fully understanding cellular function and identifying new targets for therapeutic applications through genome editing.

We are currently sequencing and assembling the genomes of two cnidarian species, *Hydractinia echinata* and *Hydractinia symbiolongicarpus*. These colonial marine animals have already proven to be valuable model organisms for the study of regenerative medicine, early development processes, and allorecognition. Like human embryonic stem cells, *Hydractinia* stem cells (called 'i-cells') are pluripotent; homologs for many genes associated with the ability to self-renew and differentiate in bilaterians have already been identified in *Hydractinia*, reinforcing its value in the study of development and regeneration. Importantly, the regenerative process also depends on both the maintenance of telomere integrity throughout numerous cycles of cell division and the role of ribosome biogenesis in cell growth and proliferation.

To better-understand the role of repetitive DNA in regeneration and stem cell maintenance in *Hydractinia*, whose overall repeat and AT-content is quite high (47% and 65%, respectively), we are developing new strategies to identify and advance our biological understanding of important repetitive regions such as telomeres, centromeres, and rDNA in these de novo assemblies, as a first step towards determining important similarities and differences between regenerative and non-regenerative organisms. We have already identified a complete ribosomal gene consensus sequence and prospective junction sequences on either side of rDNA clusters that will help us explore ribosome regulation. Using the eukaryotic telomeric sequence, we have also identified scaffolds that contain regions that flank the telomere and contain new tandem repeats up to 380 bp.

This overall approach will not only enable us to offer a more complete assembly than those currently available for any other model organism but also provide a foundation for better understanding development, genetic variation, and key metabolic pathways – knowledge that has the strong potential to both advance basic research efforts focused on a variety of human diseases and the development of new clinical approaches that improve human health.

← Back

GENETIC ADMIXTURE AND DIFFERENTIATION STATES SHAPE THE HUMAN METHYLOME IN STEM, PROGENITOR AND SOMATIC CELLS

Boris Bartholdy[1], Julien Lajugie[1], Zi Yan[1], Shouping Zhang[1], Rituparna Mukhopadhyay[1], John M Greally[2], Masako Suzuki[2], Eric E Bouhassira[1]

[1]Albert Einstein College of Medicine, Cell Biology, Bronx, NY, [2]Albert Einstein College of Medicine, Genetics, Bronx, NY

Whole-genome bisulfite sequencing has revealed dramatic differences of methylomes of different cell types and differentiation stages. However, the global mechanisms of maintenance of DNA methylation patterns and their functional consequences are not clear. Using a combination of novel haplotype-resolved methylomes and publicly available data, we show that stem and progenitor cells, somatic cells and transformed cells exhibit distinct classes of methylomes that differ by the proportion of partially (PMDs) and highly methylated domains (HMDs). Through analysis of allele-specific data of siblings we also show that the bodies of active genes are highly methylated in all cell types and that gene transcription is likely directly responsible for this, since allele-specific expression and methylation were highly correlated, specifically in the haplo-identical regions of two sisters.

We demonstrate that PMD and HMDs are maintained mostly during S phase and that PMD formation is not specifically associated with late DNA replication, since short PMDs replicate early, while long PMDs replicate late. We found no correlation between timing of replication and DNA methylation in core asynchronously replicating domains (c-ARD), suggesting that the efficiency in the maintenance of DNA methylation does not vary significantly during S phase.

Finally, analysis of allele-specific differentially methylated regions (a-DMRs) showed that higher levels of DNA methylation were associated with a later replication time, and, importantly, revealed their enrichment within PMDs highly enriched in SNPs, such as those of Neandertal origin, suggesting that differences in primary sequence might determine DNA methylation levels in PMDs, and implying introgressed Neandertal DNA as an important modifier of the epigenome.

← Back

YEAST ARTIFICIAL CHROMOSOMES FOR BIOSYNTHETIC PATHWAY ASSEMBLY

Philipp Berninger, Markus Schwab

Evolva, Research & Development, Reinach, Switzerland

Nature contains a treasure trove of small molecule ingredients that can improve health, wellness and nutrition. However, most of these ingredients have “issues”: the organism that makes the compound of interest is too rare, too hard to grow or does not make enough of it. Hence, the ingredient is not available at the right quality, the right price nor the necessary amount. These issues need to be solved in order to allow a larger society to access these valuable ingredients in a sustainable manner at low costs.

Metabolic engineering, mainly in microbial hosts, has sought to circumvent some of these associated issues.
Previously, we have introduced a technology which allows to express a large number of biosynthetic genes in a heterologous host. The key point of this technology is the construction of Yeast Artificial Chromosomes (YAC) from heterologous genes in a random fashion. This approach allows us to identify yeasts producing the molecule of interest, improve biosynthetic pathways or identify unknown steps in pathways, however the exact composition and arrangement of YACs with second generation sequencing approaches remains challenging, due to the use of highly repetitive building blocks.

Here, we present the integration of third generation long read sequencing and bioinformatics workflows. This allows us not only fast and precise reconstruction of our YACs but also the optimization of those pathways via machine-learning techniques.

← Back

RARE VARIANT BURDEN ANALYSIS TO DECIPHER GENETIC ARCHITECTURE OF CHARCOT-MARIE-TOOTH DISEASE

Dana M Bis[1], Feifei Tao[1], Lisa Abreu[1], Patrick Sleiman[2], Hakon Hakonarson[2], Stephan Zuchner[1]

[1]University of Miami, Dr. J.T. MacDonald Department for Human Genetics, Hussman Institute for Human Genomics, Miami, FL, [2]Children's Hospital of Philadelphia, Center for Applied Genomics, Philadelphia, PA

Charcot-Marie-Tooth (CMT) is a group of rare, clinically and genetically heterogeneous diseases that lead to distal muscular atrophy and sensory loss. Mendelian high-penetrance alleles in over one hundred different genes have been shown to cause CMT; yet, more than 50% of patients with the axonal type of CMT do not receive a genetic diagnosis. A more comprehensive spectrum of genes and alleles is warranted, including causative and risk alleles, as well as oligogenic inheritance. Exome studies in the international Inherited Neuropathy Consortium are beginning to be sufficiently powered to perform rare variant burden analysis. Our approach compared the frequency of damaging alleles at the gene unit in exomes of 343 CMT cases and 935 controls. Initially, we explored rare (ExAC MAF<0.01) and damaging (non-synonymous and loss-of-function consequences) variant burden in known CMT genes. In 76 axonal CMT genes, we saw that cases carried on average 2.08 rare, damaging variants, while unrelated non-neuropathy controls harbored 1.69 variants (p-value=3.13x10-5, Mann-Whitney U-test). This result was achieved despite prior exclusion of cases carrying a mutation in a known CMT gene from exome sequencing. Thus, enrichment of damaging variants in CMT disease genes in such a cohort suggests the presence of additional 'risk' alleles in these genes, potentially supporting oligogenic inheritance models after further exploration. To expand upon this result, we performed an unbiased exome-wide rare variant burden analysis. We tested 17,637 protein coding loci for association using the C-alpha test. After filtering results by the PLINK/SEQ i-statistic and applying Bonferroni multiple-testing correction, three genes, KDM5A (p-value= 9.9x10-7, OR=3.6), EXOC4 (p-value= 6.9x10-6, OR=2.1), and CEP78 (p-value= 2.3x10-5, OR=4.4), reached experiment-wide significance (p-value=2.3x10-5, alpha=0.05). Interestingly, several known CMT genes achieved nominal p-values <0.05, serving as a 'positive control' for the ability of this approach to identify both risk and causative genes. We are currently performing molecular genetics and cell biology follow up studies and also working towards enlarging our sample size. In summary, statistical methods, traditionally reserved for more 'common' phenotypes, are becoming increasingly available for rare disease genetics such as CMT and will help to comprehensively define the genetic architecture of complex rare neurodegenerative disorders.

CHOOSING THE BEST OF ALL WORLDS: À LA CARTE ACCESS TO EXTANT AND EMERGENT BEST-PRACTICE METAGENOMIC PIPELINES

Daniel Blankenberg[1,3,4], Sarah Carnahan-Craig[2], The Intergalactic Utilities Commission[3], The Galaxy Team[4]

[1]Cleveland Clinic, Genomic Medicine Institute, Cleveland, OH, [2]Penn State, Biology, University Park, PA, [3]galaxyproject.org/iuc,[4]galaxyproject.org

The adoption of high-throughput sequencing methods has enabled individual investigators and small research groups to perform complex metagenomic analyses that were once only approachable by large centers and consortiums. Although data generation is no longer a rate-limiting factor in many metagenomic studies, generating the data does not, in itself, lead to an increase in knowledge. The scale of the data not only presents difficulties for individual researchers attempting to analyze the data, but also significant informatics issues for collaboration, adaptability, and reproducibility. This is further complicated by the availability of common clustering-first pipelines, such as QIIME or mothur, and emerging assignment-first approaches, such as those based upon Kraken or kallisto.
The utilization of well-documented and accepted best-practice pipelines is the holy grail of any informatic analysis. By relying upon previously peer-reviewed processes, a researcher can be reasonably well-assured that an analysis meets at least a modicum of vetted correctness, assuming that all steps have been followed properly (a big assumption). However, there is rarely a single must-use best option for a particular analysis as competing pipelines often have opposing advantages and disadvantages, and 'best-practice' is a constantly moving target. Although researchers are technically able to run their data through multiple analysis pipelines and to choose individual portions of each pipeline in an à la carte fashion to best accomplish their goals, from a practical view this is untenable for even the most computer-savvy among us. The magnitude of the data and the number of steps and parameters involved in these analyses exemplifies the need to adopt a standardized strategy to allow reproducible and adaptable high-throughput analysis of large-scale metagenomic data. This strategy should maintain reproducibility, accessibility, and transparency while ensuring flexibility; swapping individual or sets of steps should be as easy as drag-and-drop.

Here we describe a set of generalizable metagenomic pipelines implemented within Galaxy. We applied these pipelines to the analysis of over 400 buccal and stool samples from mother-child pairs as part of an obesity study. We also reanalyzed over 5,000 samples from the Human Microbiome Project. Beyond making many existing software packages and new algorithms available as Galaxy tools, several technological advancements were required to enable these pipelines. To allow Galaxy to scale to the analysis of many samples, Dataset Collections (functionality allowing multiple datasets to be treated as a single unit when run within Tools and Workflows) were enhanced. Furthermore, to scale Galaxy to ever-increasing numbers and sizes of analysis pipelines, we investigated the automatic generation of Galaxy Tools. A suite of ~50 production-ready Tools were programmatically generated from the Anvi'o platform. A command-line utility, R2-G2, that is able to convert any R package (including Bioconductor) into a set of Galaxy Tools was also developed.

REFEX: A REFERENCE GENE EXPRESSION DATASET AS A WEB TOOL FOR THE FUNCTIONAL ANALYSIS OF GENES

Hiromasa Ono, Hidemasa Bono

Research Organization of Information and Systems, Database Center for Life Science, Mishima, Japan

Gene expression data are exponentially accumulating; thus, the functional annotation of such sequence data from metadata is urgently required. However, life scientists have difficulty utilizing the available data due to its sheer magnitude and complicated access. We have developed a web tool for browsing reference gene expression pattern of mammalian tissues and cell lines measured using different methods, which should facilitate the reuse of the precious data archived in several public databases. The web tool is called Reference Expression dataset (RefEx), and RefEx allows users to search by the gene name, various types of IDs, chromosomal regions in genetic maps, gene family based on InterPro, gene expression patterns, or biological categories based on Gene Ontology. RefEx also provides information about genes with tissue-specific expression, and the relative gene expression values are shown as choropleth maps on 3D human body images from BodyParts3D. Combined with the newly incorporated Functional Annotation of Mammals (FANTOM) dataset, RefEx provides insight regarding the functional interpretation of unfamiliar genes. RefEx is publicly available at http://refex.dbcls.jp/.

All scripts used to produce the data and additional descriptions are available on the github site at https://github.com/dbcls/RefEx. And all data used for RefEx website are archived under the figshare Project ‘Data Archive for RefEx’ (https://doi.org/10.6084/m9.figshare.c.3812815).

← Back

METATRANSIT: A COMPREHENSIVE TOOLKIT FOR METATRANSCRIPTOMICS AND METAGENOMICS ANALYSIS

Arthur Brady

Univ. of MD School of Medicine, Inst. for Genome Sciences, Baltimore, MD

We present MetaTRANSiT (MetaTRANScriptomic and metagenomic analysis Toolkit), an open-source, fully-modularized pipeline covering all existing aspects of metatranscriptomics analysis, with several modules also applicable to metagenomics analysis without modification. The kit contains IMP-QC, a read quality control pipeline that performs base-quality control and zero-knowledge adaptor detection and removal for raw read sets; modules for identification and removal of reads matching rRNA sequences and/or specified host contaminants; IMP, a taxon-profiling pipeline applicable to both metatranscriptomics and metagenomics read sets, whose accuracy exceeds current state-of-the-art tools (e.g. MetaPhlAn2); and a counting, normalization and differential-expression analysis module which can report community-wide expression differences between sets of metatranscriptomics samples sequenced from comparable communities (like the human gut) at the level of individual genes as well as for higher-level functional objects like KEGG orthologs, modules and pathways. The DE engine can also focus on user-chosen sets of individual clades within larger communities, providing clade-specific differential expression reports. All analyses are performed at the read level; no assembly required.

← Back

SERVERLESS JBROWSE ON THE CHEAP

Scott Cain

Ontario Institute for Cancer Research, Stein Lab, Ontario, Canada

JBrowse is widely used GMOD (Generic Model Organism Database, http://gmod.org/wiki/JBrowse) software for displaying a variety of genomic feature data types. Here we present a method for implementing JBrowse in a very low maintenance fashion by loading the data and software for JBrowse into Amazon Web Service's S3 data storage service and serving the web pages and data directly out of S3, without the need for any other web server. We outline the methods for implementation as well as the issues that implementers may want to consider, including cost and security.

MACHINE LEARNING STRATEGIES TO IDENTIFY HIGH CONFIDENCE STRUCTURAL VARIANTS IN HUMAN GENOME REFERENCE MATERIALS

Lesley M Chapman[1], Justin Zook[1], Noah Spies[2], Nancy F Hansen[3], Fritz Sedlazeck[4], Peyton Greenside[5], Marc Salit[2], The Genome in a Bottle Consortium [6]

[1]National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD, [2]National Institute of Standards and Technology, Material Measurement Laboratory, Palo Alto, CA, [3]National Human Genome Research Institute, Cancer Genetics and Comparative Genomics Branch, Rockville, MD, [4]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [5]Stanford University School of Medicine, Biomedical Informatics, Palo Alto, CA, [6]The Genome in a Bottle Consortium, NIST, Gaithersburg, MD

Next generation sequencing (NGS) technologies are rapidly evolving. Yet, discordance exists amongst large indel and structural variant (SV) calls as a result of variance between NGS sequencing and analysis pipelines. Improvements in the accuracy of calling these difficult SVs is needed to enable confidence in clinical decision making. Previous work in the Genome in a Bottle Consortium (GIABC) has used visualization, heuristics, and exploratory machine learning to form preliminary benchmark large deletion calls, but these have been limited by inaccurate breakpoints and lack of genotype (GT) information. The central aim of the current study is to use machine learning to integrate data from multiple sequencing technologies and generate a high confidence list of large indels and SVs. We use extensive short, long, and linked read whole genome sequencing from an Ashkenazim Jewish mother-father-son trio (NIST RM 8392) to develop integration methods. Members of the GIABC generated over 300000 candidate large indels and SVs (>= 20bp) within this trio from 30+ informatics pipelines and 5 sequencing technologies.Three integration approaches were developed to compare large variants, many of which are located in tandem repeat regions (> 50% of all calls). This showed that 20,218 deletions in the trio where >=2 technologies had size predictions within 20%. Results from the SURVIVOR method found 27,304 deletions, 14,045 insertions, and 117 inversions where >=2 technologies had variants of the same type and breakpoints within 1000bp. To account for different representations within repetitive regions SVcomp was used. The results of this analysis showed that 30,493 (variants matched exactly), 33,533 (variants matched with <=2% difference using three distance metrics), and 45,667 (<=10% different) variants of all SV types supported by >=2 technologies. Machine learning and crowd-sourced manual curation will be used to further evaluate these potential benchmark calls. The results of a preliminary machine learning analysis where crowdsourced data was used to train a random forest (RF) model, showed that the RF model was able to predict GT labels for 2800 randomly selected deletions with a 0.93 precision score. In future studies, we will use crowdsourced labeled data for insertions, deletions, and variants ranging from 20 - 100+kb to train machine learning machine learning models.

← Back

PREVIOUSLY UNDETECTED GENOMIC SIGNATURES OF GIANT VIRUSES ARE UBIQUITOUS IN METAGENOMES

Anirvan Chatterjee, Kiran Kondabagil

Indian Institute of Technology, Biosciences and Bioengineering, Mumbai, India

Large DNA viruses, with genomes upto 2.5 mega bases, and a near complete complement of cellular machinery, have blurred the distinction between the viral and cellular world. Genomics has led the way in unravelling the evolutionary past of these viruses. Using whole genome shotgun sequencing (WGS) we discovered 6 new giant viruses from environmental samples in Mumbai. Comparative genomics demonstrated exceptional functional conservation despite extensive genomic rearrangements in NCLDVs isolated from different geographies. For instance Kurlavirus, was found to be closely related to the recently described Noumeavirus, which is classified with Lausannevirus and Port-miou virus as Marseilleviridae Lineage B. Whole genome alignment of Kurlavirus with Noumeavirus, Lausannevirus and Marseillevirus shows greater synteny within the Lineage as compared to Marseillevirus, which is reflective of the phylogeny. On comparing genome of Powai lake megavirus with 5 other megavirus genomes, we observed remarkable homology among all the isolates indicating the presence of a founding ancestor of this lineage.

To expedite discovery of more NCLDVs we increased recovery of viral DNA from unprocessed samples and developed a metagenomics pipeline to detect NCLDV signatures in environmental shot-gun sequences. We observed greater than 10% of reads in such metagenomes aligned with NCLDV genomic database. This is greater than the recent report of NCLDV in sewages samples. De novo assembly yielded several contigs > 1kb with no significant homology of NCBI NR database, indicating probable presence of novel NCLDVs yet to be discovered. The ability to study the microbial dark matter emerged as a greatest triumph of omics. However interpretation of WGS data is limited by cultivation bias in databases.

Advances in genome informatics, single cell genomics along with deep learning methods will significantly augment discovery of more NCLDV genomes, enabling the study of their role in early evolution of all life forms and microbial diversity.

← Back

INDIVIDUAL ANCESTRY ESTIMATION FROM WHOLE EXOME SEQUENCING DATA IN PATIENT-DERIVED XENOGRAFT SAMPLES

Li Chen[1], Meredith Yeager[1], Biswajit Das[1], Chris Karlovich[1], Diane Palmieri[1], Eric Karlins[1], Vivekananda Datta[1], Corinne Camalier[1], Yvonne Evrard[1], Melinda Hollingshead[2], Paul M Williams[1], James Doroshow[3]
[1]Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, [2]National Cancer Institute at Frederick, Developmental Therapeutics Program, Frederick, MD, [3]National Cancer Institute, Division of Cancer Treatment and Diagnosis, Bethesda, MD

The NCI Patient-Derived Models Repository (PDMR; pdmr.cancer.gov) provides the researchers a useful resource of patient-derived models (PDMs) from primary and metastatic tumor tissues to facilitate pre-clinical drug studies. However, the repository has limited patient demographic data for race/ethnicity. An estimation of individual ancestry using whole exome sequencing (WES) data, which is available for early-passage PDMs and some originating patient specimens, will therefore be needed to stratify the patient population when investigating associations between genetic variants and phenotypes of interest. Here, we present a workflow to infer ancestry of patient-derived xenograft (PDX) samples from WES data using SNPweights, an algorithm to infer genetic ancestry from single nucleotide polymorphism (SNP) weights precomputed from large external reference panels and further estimates the fraction (%) ancestry. In our workflow, 364,458 SNP weights precomputed from HapMap3 samples are used to infer individual ancestry in PDX samples from four populations: West African, European, East Asian and Native American. To call SNP genotype from WES data, sample pair-end reads are first aligned to a human reference database by Novoalign, a conservative aligner to remove potential mouse reads in the PDX samples. Then the sorted and de-duplicated aligned reads are further filtered based on mapping and base quality score. By using Samtools and Bcftools, variants are detected and genotype likelihoods are calculated. SNPs are further filtered for SNPweights. Finally, % ancestry from the four populations is reported for each patient with source information. Specifically, the % ancestry is estimated from a single source if that source is an originating specimen from the patient; otherwise, it is reported as an average obtained from all sequenced PDX tumor samples derived from that patient. We have tested the workflow on PDX and patient samples from 134 PDMs. Among them, 43 patients have self-reported race/ethnicity information and 29 patients have both originating specimens and PDX samples. We first compared the ancestry estimated from different sources in 29 patients and the % ancestry is consistent in both sources. Then we compared the estimated ancestry with the self-reported race/ethnicity. By using 80% as a cutoff to assign ancestry, 41 out of 43 (95%) patients have concordant race information. Of the remaining patients, one showed admixed ancestry (all <80%) and another patient showed a different ancestry from self-reported race/ethnicity. Overall, the workflow has been shown to be accurate for inferring ancestry of the PDX samples for the PDMR project. It will be an important tool for estimating ancestry in cases where race/ethnicity of the patient is unknown, and will provide high-value for researchers investigating questions related to cancer health disparities.

← Back

PROFILING OF SOMATIC ALTERATIONS IN BRCA1-LIKE BREAST TUMORS

Youdinghuan Chen[1,2,3], Yue Wang[3,4], Lucas A Salas[1], Todd W Miller[3,7], Jonathan D Marotti[5], Nicole P Jenkins[2], Arminja N Kettenbach[2,3,7], Chao Cheng[3,4,7], Brock C Christensen[1,3,7]

[1]Geisel School of Medicine at Dartmouth, Epidemiology, Lebanon, NH, [2]Geisel School of Medicine at Dartmouth, Biochemistry and Cell Biology, Lebanon, NH, [3]Geisel School of Medicine at Dartmouth, Molecular and Systems Biology, Lebanon, NH, [4]Geisel School of Medicine at Dartmouth, Genetics, Lebanon, NH, [5]Geisel School of Medicine at Dartmouth, Pathology and Laboratory Medicine, Lebanon, NH, [6]Geisel School of Medicine at Dartmouth, Biomedical Data Science, Lebanon, NH, [7]Norris Cotton Cancer Center, Lebanon, NH

Germline or somatic mutation in *BRCA1* is associated with an increased risk of breast cancer and more aggressive tumor subtypes. BRCA1-deficient tumor cells have defective homologous recombination (HR) DNA repair, exhibiting genome instability and aneuploidy. HR deficiency can also arise in tumors in the absence of *BRCA1* mutation. An HR-deficient, BRCA1-like phenotype has been referred to as “BRCAness.” BRCA1-like cancers exhibit worse prognosis but are selectively sensitive to chemotherapeutic treatments (e.g. platinum-based alkylating agents). However, the molecular landscapes of BRCA1-like breast tumors remain largely unknown in part because they are less common in the general population. By applying a copy number-based classifier, we observed that >30% of The Cancer Genome Atlas (TCGA) breast tumors are BRCA1-like even though only ~3% tumors analyzed carry a *BRCA1* mutation or promoter hypermethylation. Separately, a differential analysis controlling for hormone receptor status, subject age, tumor stage and purity revealed a significant increase in DNA methyltransferase 1 (DNMT1) protein expression in BRCA1-like tumors. In addition, differentially methylated gene sets in BRCA1-like tumors indicated a strong enrichment in developmental signaling and a moderate involvement in gene transcription. Profiling of concomitant somatic alteration landscapes in BRCA1-like breast tumors provides alternative strategies to identify this subset of tumors and insights into novel potential therapeutic approaches.

← Back

DE NOVO ASSEMBLY OF GOLDFISH USING PACBIO LONG READS

Zelin Chen[1], Hironori Wada[2], Sergey Koren[1], Kevin Bishop[1], Raman Sood[1], Koichi Kawakami[2], Adam Adam Phillippy[1], Jam C Mullikin[1], Asao Fujiyama[4], Yoshihiro Omori[3], Shawn M Burgess[1]

[1]National Institutes of Health, National Human Genome Research Institute, Bathesda, MD, [2]National Institute of Genetics and Department of Genetics, Division of Molecular and Developmental Biology, Mishima, Japan, [3]Osaka University, Institute for Protein Research, Osaka, Japan, [4]National Institute of Genetics, Comparative Genomics Laboratory, Mishima, Japan

The new next generation sequencing technologies: 10X Genomics's linked-reads, Pacbio's SMRT and Nanopore sequencing provide sequenced ~10kb to megabase long reads without costly and time-consuming BAC or Fosmid clone library construction. De novo genome assembling by these technologies can generate more continuity and complete draft genome than using previous short reads, also for non-inbreed diploid species like many fish. Here, we provide a new assembly of the goldfish genome using Pacbio SMRT technologies.

Goldfish is the most prominently domesticated and most kept aquarium fish in the world. More than one thousand years of breeding has produced many variants with different colors, shapes and fin styles, which make it quite distinct from its recent sister crucian carp (diverged only < 2.3-3.0 Mya) and common carp (diverged 8.1–11.4 Mya). Both of goldfish and common carp are supposed involving in a fourth round of whole genome duplication at ~10Mya) after diverged from zebrafish.

We sequenced the genomic DNA from a heat shock diploid common Wakin. We obtained ~16.4 M pacbio reads (~70X coverage) with peak leng of ~9Kb. Raw reads were corrected and assembled using the Canu Assembler. Accuracy of the assembly was improved via Arrow. After deleting 988 contigs that is contained in other contigs with >97% nucleic identity, most of which is of half or lower depth, the remain assembly contains 8427 contigs and 1820 Mbp. N50,N90 reaches 833.2kbp,73.7Kbp and longest contig is 12.8Mbp. 2613 BUSCO core eukaryotic genes out of 3023 total genes can be completely aligned to our assembly, compared to that only 1832 can be completely alignment to the common carp assembly. Comparative analysis of 3-way genomic and gene alignment (carp, goldfish, zebrafish) will facilitate predicting conserved functional elements and illustrate how a recent whole genome duplication could shape species.

← Back

H3K27 TRI-METHYLTRANSFERASES CLF AND SWN REDUNDANTLY BUFFER ABA-INDUCED SENESCENCE IN ARABIDOPSIS

Chunmei Liu*[1,2], Jingfei Cheng*[1,2], Yili Zhuang[1,2], Luhuan Ye[1,2], Zijuan Li[1,2], Yuejun Wang[1,2], Meifang Qi[1,2], Yijing Zhang[1,2]

[1]Chinese Academy of Sciences, Shanghai Institutes for Biological Sciences, Institute of Plant Physiology and Ecology, Shanghai, China, [2]University of the Chinese Academy of Sciences, Beijing, China

The phytohormone abscisic acid (ABA)-induced leaf senescence facilitates nutrient reuse, which is essential for enhancing plant tolerance to stresses. Senescence is a complicated process closely associated with plant lifespan as well as crop yield, while the mechanism by which it is finely tuned through different layers of control is still insufficiently understood. We found that H3K27me3 is significantly enriched for ABA-responsive elements and that the double mutant of Polycomb enzymes CURLY LEAF (CLF) and SWINGER (SWN) is hypersensitive to ABA in Arabidopsis, indicating a close relationship between the epigenetic machinery and stress hormone response in plants. To elucidate the crosstalk on genome-scale, we systematically profiled the epigenomic and transcriptomic changes triggered by ABA and revealed the redundant role of CLF and SWN in buffering ABA-induced senescence. Specifically, H3K27me3 preferentially targets ABA-induced senescence associated genes (SAGs). These SAGs were more highly expressed in the *CLF* and *SWN* double mutant and could be further induced by ABA to a higher extent when compared with the wild-type. Furthermore, ABA-triggered H3K27me3 reduction was closely associated with ABA-induced target gene expression but occurred much later in time. Thus, the presence of H3K27me3 does not block ABA-induced SAGs, but rather limits the extent of induction. Collectively, the present study revealed that PcG enzymes gated ABA-induced senescence. The findings may serve as a paradigm for a global understanding of the crosstalk between the rapid effect of the phytohormone ABA and the long-term effect of epigenetic machinery in regulating plant senescence process and environmental responses.

HARNESSING A GOLD STANDARD DATA SET FOR IMMUNO-ONCOLOGY

Stephen Chervitz, Ravi Alla, Jason Harris, Sean Boyle, Richard Chen

Personalis, Inc., Bioinformatics, Menlo Park, CA

The Personalis neoantigen prediction pipeline identifies and evaluates the immunotherapeutic potential of single-nucleotide variations (SNVs) and small insertions & deletions (indels) specific to an individual patient's tumor. To validate the sensitivity of this next-gen sequencing (NGS)-based pipeline, a well-characterized dataset of immune epitopes was identified and evaluated. These epitopes have been published in the literature as arising from SNVs or indels known to create neoantigens that induce an immune response in humans. Neoantigen-inducing variants were spiked into control NGS data (both DNA and RNA) generated by the Personalis ACE exome technology, and the spiked data sets were used as input to our neoantigen prediction pipeline. Resulting pipeline output was assessed to ascertain (1) the reported status of the variants at the DNA and RNA levels, (2) the presence of the expected neopeptides, (3) the predicted HLA types, and (4) HLA binding affinities generated by our Neoantigen pipeline. Performance of the pipeline is assessed using the set of known SNV neoantigens, spiked in at different allele frequencies.

← Back

REGULATION OF GENE EXPRESSION IN RESPONSE TO DNA BASE COMPOSITION

Kashyap Chhatbar, Timo Quante, Konstantina Skourti-Stathaki, Shaun Webb, Jim Selfridge, Adrian Bird

Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, United Kingdom

The transcriptional landscape during embryonic differentiation depends on DNA sequence-specific transcription factors that combine to control target gene activity. We explore the possibility that in addition to various transcription factors, large regions of DNA sequence that differ in average base composition also influence gene expression. By screening mouse embryonic stem cells (ESCs) for potential proteins that bind to short, frequent AT-rich sequence motifs, we found that stem cell factor Sall4 binds a specific AT-rich sequence. Genome wide analyses showed that the AT-rich sequence reflects base composition and large DNA sequence domains enriched in this sequence preferentially bind Sall4. To access the effect of AT-binding on gene expression, we mutated the DNA-binding zinc finger cluster of Sall4. Genes with high AT-rich sequence frequency across and surrounding the gene body showed elevated expression in the mutant cells. In particular, up-regulated genes tend to be tissue specific and enriched in GO-terms relating to differentiation. The results indicate that Sall4 regulates the expression of differentiation genes by targeting a simple DNA sequence motif whose frequency responds to base composition.

← Back

INTEGRATIVE GENOMIC ANALYSIS OF 176 KOREAN LIVER CANCER REVEALS DISTINCTIVE MOLECULAR PATTERN

Eunji Choi[1], Jinyoung Lee[2], Minhyuk Jung[2], Dawon Kim[1], Young-joon Kim[1,2]

[1]Yonsei University, Integrated Omics for Biomedical Science, Seoul, South Korea, [2]Yonsei University, Biochemistry, Seoul, South Korea

Liver Cancer is the second leading cause of cancer-related death worldwide as 6th most prevalent cancer in Korea. Clinical and molecular heterogeneity of liver cancer contributes the high resistance to treatments and frequent metastasis. The aim of the study is to find molecular subtypes of liver cancer and identify subtype-specific genomic features. This pilot study analyzed 176 liver cancer cases by DNA methylation, RNA expression and 30 paired samples by whole-genome sequencing. Integrative analyses of methylation, mRNA expression and lncRNA expression of 176 liver cancer patients revealed 3 distinctive subtypes with significant distinguishing features of methylation probes and genes.

← Back

GENOME ANNOTATION USING THE MAKER-P JETSTREAM CLOUD

Kapeel Chougule[1], Michael Campbell[1], George Wang[3], Joshua Stein[1], Bo Wang[1], Yinping Jiao[1], Nicholas Hazekamp[4], Upendra Kumar Devisetty[5], Nirav Merchant[5], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]United States Department of Agriculture, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, [3]Yale University, Department of Biomedical Engineering, New Haven, CT, [4]University of Notre Dame, Department of Computer Science and Engineering, Notre Dame, IN, [5]University of Arizona, BIO5, Tucson, AZ

The promise of genome research depends on our ability to accurately annotate and derive meaning from sequence data. In plants such as maize, extremes of genome size push the limits on current algorithms, expertise, and computational power needed by today's researchers. Furthermore, new sequence technologies are driving research communities to move beyond reliance on a single reference genome to represent a species. As we head into the "pan-genome age", researchers need access to reliable computational resources, standardized annotation workflows and the consistent evidence sets for annotation of multiple reference strains. Here, we describe a freely available, robust and reusable resource for automated annotation of maize genomes that utilizes the NSF JetStream cloud service. With this service, users can check out a virtual image pre-installed with the MAKER-P annotation engine and underlying software, plus configuration files and curated evidence datasets (e.g. repeat library and transcriptome data), for standardized maize-specific annotation. As proof-of-concept, we demonstrate the utility of this resource by annotating three maize lines, B73, NC350, and W22 using message parsing interface (MPI) and work queue system (WQ-MAKER) for scalability.

This work is supported in part by NSF grants DBI-1265383 (for the CyVerse project),IOS-1445025 (for the MaizeCode project) and ACI-1445604 (for the JetStream project).

← Back

UNDERSTANDING MAMMARY STEM CELL STATE REGULATION THROUGH CHROMATIN ACCESSIBILITY

Chi-Yeh Jay Chung, Christopher Dravis, Gidsela Luna, Cynthia Ramos, Geoffrey M Wahl

Salk Institute for Biological Studies, Gene Expression Laboratory, La Jolla, CA

Epigenetic mechanism plays an important role in regulating cellular state in normal mammary cells and breast cancer. As many breast tumors share key molecular pathways with their normal counterparts, studying the epigenetic regulation of normal mammary subpopulations has important implications for breast cancer treatment. However, how chromatin structure mediates mammary cell plasticity is poorly studied. Here, we aim to understand the dynamics of chromatin structure between fetal mammary stem cells (fMaSCs) and adult basal (Ba), mature luminal (ML) and luminal progenitor (LP) cells. Within these cell types, fMaSC and Ba contain multipotent stem cell activity as demonstarted by transplantation studies, whereas LP and ML do not transplant and are thus low in stem cell activity. We performed assay for transposase-accessible chromatin using sequencing (ATAC-seq) in mouse primary fMaSCs, Ba, LP and ML cells. By integrating ATAC-seq result with transcriptomic and histone modification profiling, we show that chromatin accessibility positively correlates with gene expression and active histone mark. In addition, the regulation of distal elements is more cell-type specific than gene promoters, implicating the importance of enhancers in cell state regulation. Interestingly, fMaSCs and adult basal cells, two cell types exhibiting multipotent stem cell activity, contain significantly more open chromatin regions than lineage-restricted luminal cells. Furthermore, fMaSCs and basal cells show open chromatin at markers of multiple lineages, whereas open chromatin is restricted to single lineage markers in luminal cells. This suggests that multi- and balanced-lineage open chromatin might contribute to mammary stem cell state maintenance. Finally, to identify cell-type specific epigenetic signatures, we isolated unique accessible regions (UARs) for each cell type. Most UARs are located at distal region, and correlate with cell-type specific gene expression and functionally relevant transcription factor (TF) binding. Interestingly, many Sox10 motifs are uniquely open in the least differentiated fMaSCs and closed in the most differentiated ML cells. This is consistant with our previous data showing Sox10 regulates mammary stem/progenitor activity. To investigate if Sox10 promotes stemness in breast cancer, we performed ATAC-seq on Sox10 positive and negative mouse mammary tumor cells using two distinct Sox10 reporter lines. Strikingly, we found Sox10 positive tumor cells gain chromatin features specific to normal mammary stem and progenitor cells. As SOX10 is highly expressed in basal-like human breast cancers, a highly aggressive subtype that bears normal mammary stem/progenitor expression programs, we are currently investigating the mechanism of Sox10 in regulating chromatin structure, cellular plasticity, and tumor heterogeneity in breast cancer.

THE HUMAN CELL ATLAS DATA COORDINATION PLATFORM

Laura Clarke, The Human Cell Atlas Data Coordination Team

EMBL-EBI, Molecular Archives, Cambridge, United Kingdom

The Human Cell Atlas (HCA) is taking a systematic, data-driven approach to create a reference map of all human cells. This atlas will consider cell type alongside other facets of a cell's identity such as state, transitions between cell types, lineage, cell-cell interaction and a cell's local neighbourhood. The atlas will be used as a basis for understanding human health and diagnosing, monitoring, and treating disease.

This massive undertaking requires an open, modular, and extensible approach to data coordination. The Broad Institute, the Chan Zuckerberg Initiative, EMBL-EBI, and UCSC are building a Data Coordination Platform (DCP). This will organise terabytes of data for billions of cells, across multiple modalities, generated by hundreds of labs around the world. This service will enable the community to innovate rapidly, without barriers to access, and facilitate computational researchers developing new analysis methods. The DCP has four components: the ingestion service, the data store service, the secondary analysis service, and the tertiary portals. All the software will be developed and licensed in the open.

The ingestion service will provide human support and software tools, in the form of APIs and UIs, to support the data generators in providing high quality, well structured data and descriptions into the atlas.

The data store service will provide a multi-cloud based storage service including all raw data, metadata, and certain forms of derived data generated by the project. Data will reach the data store via the ingestion service, and users will be able to access the data directly via the consumer API or via visualization and analysis tools produced as part of the tertiary portals.

The secondary analysis service provides a pipeline execution infrastructure allowing robust, community vetted pipelines to be run in a standardized way. The results will be deposited back into the data store. The HCA analysis working group will identify which pipelines to run, ensuring there is at least one pipeline for each anticipated data type (e.g sc-RNAseq and imaging). All pipelines will be built using open source code and shared via containers to ensure the entire community can take advantage of this work.

The tertiary portals will provide user facing services for the community to analyse and visualise the HCA data. We intend to provide a few simple portals as a starting point, but new portals can be developed by anyone in the scientific or computational community. The portals will address a wide diversity of use cases, including clustering, differential interference, spatial reconstruction, visualization, and graph-based analysis.
Here we present an overview of the DCP and its components. For more information, please read our website: https://www.humancellatlas.org/.

MAPPING OF R-LOOPS IN *TRYPANOSOMA BRUCEI* REVEALS CONSERVED AND NOVEL FUNCTIONS

Emma Briggs[1], Kathryn Crouch[1], Graham Hamilton[2], Richard McCulloch[1]

[1]University of Glasgow, Wellcome Centre for Molecular Parasitology, Glasgow, United Kingdom, [2]University of Glasgow, Glasgow Polyomics, Glasgow, United Kingdom

R-loops are stable RNA-DNA structures that form within the DNA helix. As well as arising during transcription, active roles have recently been described in a wide variety of genomic processes, including activation and termination of transcription and replication. R-loops are also associated with genome rearrangement events such as during immunoglobulin class switching. Two ribonuclease H (RNaseH) enzymes have been described in eukaryotes with the ability to degrade the RNA in hybrid structures. Here, we describe R-loop distribution in the novel genome of the eukaryotic parasite *Trypanosoma brucei*, revealing conserved and diverged R-loop functions. The ~8000 genes in the core of the genome are arranged into ~200 polycistronic RNA Polymerase (Pol) II transcription units from which mRNAs are generated by coupled trans-splicing and polyadenylation. mRNA levels are regulated post-transcriptionally. Initiation and termination of transcription are poorly understood. *T. brucei* also employs a complex immune evasion mechanism. The variable surface glycoprotein (VSG) coat is expressed from one of ~15 telomeric, Pol I transcribed, multigenic expression sites (ES). Switching of the VSG coat can occur through transcriptional events within the ES, or recombination events between the ES and silent subtelomeric VSG arrays that account for around 20% of the genome. Mapping of DNA-RNA hybrids in *T. brucei* by DRIP-seq was carried out in wild type cells, RNaseH1 null mutants and after RNAi silencing of RNaseH2 (which is lethal). ~30,000 sites of R-loop enrichment are found, greatly exceeding in silico predictions. 85% of the R-loops map within the Pol II transcription units, predominantly co-localising with polyadenylation sites. Thus, in these locations R-loops relate to progression rather than termination of transcription. Loss of either RNaseH results in R-loop enrichment that is particularly marked in genes with low levels of mRNA, indicating a role for R-loops in RNA processing and turnover. R-loops are also observed within the so-called strand switch regions that separate multigene transcription units. Here, enrichment is most marked at sites of transcription initiation, with levels increased in RNaseH1 mutants and decreased after RNaseH2 loss. Processing of RNA in R-loops therefore has a complex relationship with initiation of multigenic transcription in *T. brucei*. Lastly, in RNAseH1 mutants, we observed increased VSG coat switching, which is associated with accumulation of R-loops throughout the ES and subtelomeric arrays. A notable site of accumulation is within repeat sequences only found immediately upstream of VSG genes. We therefore propose that R-loops have been harnessed to induce switching of the VSG coat.

← Back

GDCWEBAPP: FILTERING, EXTRACTING, AND CONVERTING GENOMIC AND CLINICAL DATA FROM THE GENOMIC DATA COMMONS PORTAL

Fabio Cumbo[1,2,3], Anton Nekrutenko[2], Giovanni Felici[3]

[1]Roma Tre University, Department of Engineering, Rome, Italy, [2]The Pennsylvania State University, Department of Biochemistry and Molecular Biology, State College, PA, [3]National Research Council of Italy, Institute for Systems Analysis and Computer Science "Antonio Ruberti", Rome, Italy

In the Big Data era, one of the most critical issues is how to define data. The problem of standardizing data is currently a challenge that has to be won in order to really get advantage of data and retrieve significant insights from them. In the field of Life Sciences a big effort to fight this problem has already been done with some projects like ENCODE and the Genomic Data Commons (GDC), that collect a large amount of biomedical data concerning different experiments and clinical information about thousand of ill and healthy individuals. Unfortunately they do not provide this data in a unique standard, leaving the problem of data standardization still unresolved.

Here we make a step towards the standardization of biomedical data presenting GDCWebApp: a novel web service for querying, filtering, extracting, and converting all public data from GDC. The service has a minimalistic and extremely simple interface that allows to select multiple data sets at once. A data set is defined by three attributes such as a program (i.e. The Cancer Genome Atlas (TCGA), and Therapeutically Applicable Research to Generate Effective Treatments (TARGET)), a tumor name, and a data type (i.e. Clinical and Biospecimen Supplements, Copy Number and Masked Copy Number Segment, Gene, Isoform, and miRNA Expression Quantification, Masked Somatic Mutation, and Methylation Beta Value). Furthermore, it is possible to filter out a priori only the experiments conducted on some specific patients by selecting one or more clinical and biospecimen attributes. Additionally, one of the most important feature of GDCWebApp consists in the conversion of a data set to BED, GTF, CSV, or JSON format. It is also possible to select what kind of information will be printed in the converted data (just the chromosome, start position, end position, and strand will be added by default).

The aim of GDCWebApp is to standardize different kind of data using the same format. This will allow the researcher to easily investigate these data integrating them and extracting significant insights. Other software tools are able to extract and convert genomic data to different formats (e.g. IRIS-TCGA and TCGA2BED), but the service discussed here presents some relevant differences such as the reproducibility of a request by submitting an XML file containing a list of data sets with their descriptions, the possibility to customize the content of the data, and the complete integration in Galaxy. This last feature is extremely significant because it breaks the problem of storing the requested data sets by sending them directly to the user workspace in the Galaxy platform, and automatically structure them into a list of data collections for an optimal organization of the data.

Availability: GDCWebApp is available at http://bioinf.iasi.cnr.it/gdcwebapp/ and on the official Galaxy Tool Shed under the name "gdcwebapp".

← Back

MOVING TOWARDS COMPARATIVE ANALYSIS OF HUNDREDS OF VERTEBRATE GENOMES

Carla Cummins, Mateus Patricio, Wasiu Akanni, Matthieu Muffato, Bronwen Aken, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

With whole genome sequencing becoming more attainable year on year, the volume and diversity of available genomes continues to grow at a rapid rate. Various large sequencing projects such as Genome10K [1] or B10K will generate a large amount of high-quality genomes in the coming years and many services must adapt to deal with this rate of output. At its inception, the Ensembl project contained only a single genome; in the most recent update, it contains more than 100. A major goal of the entire Ensembl project is to expand our capabilities to allow us to introduce hundreds new genomes with each new update, in order to create the most effective resources for using genomic data in biology.

Comparative analysis, by its nature, grows quadratically with each new genome analysed, giving it the potential to become a bottleneck when dealing with datasets in the order of thousands. Several important upgrades to Ensembl's comparative pipelines have already been made to cope with these demands - often with the aim to reduce search space. Gene annotations are projected amongst closely-related species, while methods such as Cd-hit [2] allow for a greatly reduced BLAST database. Hidden Markov Model (HMM) profiles allow for more rapid and sensitive categorisation of sequences for tree building and we are aiming to replace our Pecan- and Ortheus-based whole-genome alignments with progressiveCactus [3], which has been developed to handle hundreds of genomes. These algorithmic changes, in conjunction with several useful efficiency upgrades to the eHive workflow-management system, currently allow us to compute gene trees and homologies for 85 genomes in 36 hours. Here we will present our ongoing algorithmic changes with an overview of the work done to automate the data compute in a controlled, reliable and reproducible manner.

[1] Genome 10K Community of Scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. Journal of Heredity. 2009;100(6):659-674
[2] Weizhong Li, Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–1659
[3] Paten et al. Cactus: Algorithms for genome multiple sequence alignment. Genome Research. 2011;21(9):1512-1528

STREAMLINING THE INSTALLATION OF THOUSANDS OF BIOINFORMATICS SOFTWARE PACKAGES WITH BIOCONDA

Ryan Dale[1], Johannes Köster[2], Björn Grüning[3]

[1]National Institutes of Health, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, [2]University of Duisburg-Essen, University Hospital Essen, Genome Informatics, Essen, Germany, [3]University of Freiburg, Department of Computer Science, Freiburg, Germany

Installing bioinformatics software is more time-consuming and frustrating than it should be. Package managers such as apt, yum, CPAN, CRAN, BiocLite, pip, and Homebrew help greatly in their respective domains, but setting up reproducible analyses often requires coordination across these tools. Difficulties arise if a required package or its dependencies are incompatible with system libraries, and lack of root privileges (common in HPC environments) further complicates the task.

One solution is the Conda package manager. Conda started as a tool used by the Anaconda Python distribution, but has since matured into a full-fledged, language-agnostic, and platform-agnostic general package manager. It does not require root privileges and can work alongside or even replace the aforementioned package managers. Another solution for reproducible environments is to use containers (Docker, rkt, Singularity) with pre-installed tools and dependencies.

Bioconda (*https://bioconda.github.io*) is a flexible, scalable, and sustainable system that uses both Conda and containers to simplify the installation of bioinformatics software. Bioconda is a GitHub repository of bioinformatics-related conda recipes coupled to an automated build system. Contributing a new recipe or updating an existing one is easy, with extensive documentation and a welcoming and supportive community. Once the build system successfully builds and tests a recipe, a package is uploaded to the Bioconda channel to be installed via Conda. That same package is also installed into a minimal Docker container, which is uploaded to the BioContainers repository. Future reproducibility is ensured by mirroring source code used in building Bioconda recipes to Cargo Port, the distribution center of the Galaxy project, for long-term archival storage.

An example usage is the command *conda create -n rnaseq-env fastqc multiqc bedtools samtools star subread bioconductor-deseq2*, which will install a complete environment for RNA-seq analysis that is isolated from the main system, using tools written in Java, Python, C++, C, and R -- including Java, Python and R themselves plus all dependencies -- in under 5 minutes.

Bioconda seems to have struck a nerve in the community: over the past two years, 200 contributors worldwide have helped build more than 2300 packages which have collectively been downloaded over 4.6 million times.

← Back

LEVERAGING LINKED READS FOR SINGLE-SAMPLE SOMATIC VARIANT CALLING

Charlotte A Darby[1], Ben Langmead[1,2,3], Michael Schatz[1,4,5]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Johns Hopkins Bloomberg School of Public Health, Biostatistics, Baltimore, MD, [3]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [4]Johns Hopkins University, Biology, Baltimore, MD, [5]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

In contrast to germline (inherited) variants, mutations occurring early in development are only present in some cells of the developed individual. A healthy human is thought to harbor many benign "somatic mutations" throughout their body, but a single additional one can be disease-causing. Somatic mutations have been implicated in autism, rare diseases, including those where the skin has a visible "mosaic" pattern, and many forms of cancer.

Short-read sequencing of paired samples or a pedigree is currently used to identify somatic variants based on statistical analysis of allele frequency. We suggest new strategies using "linked reads" (10X Genomics) in hopes of distinguishing candidate somatic mutation from copy number variation and sequencing error in a single sample, a case where short reads are often inadequate.

A linked read is a group of short reads sequenced from the same original long DNA molecule. Linked reads enable high-quality, contiguous phasing of heterozygous SNPs, a task impossible with short reads alone. We use this phasing to assign specific reads to haplotypes, thus supplementing the allele fraction at a site with haplotype-specific allele counts. Cells bearing a somatic mutation effectively form a third haplotype, as they were present at some unknown frequency alongside normal cells in the sequenced sample. We identify informative features of the reads and linked reads mapped to a certain genomic position, including several new metrics of linked read quality. At the site, we evaluate how these allele, haplotype, short-read and linked-read features differ from the distribution for homozygous, heterozygous, or CNV sites (the false positives) to call candidate somatic mutations.

The method requires aligned linked reads and phased variant calls. We evaluate the distribution of our selected features on simulated data, and test strategies such as filters, thresholds, and feature classification. Finally, we apply the method to 10X Genomics data from healthy and diseased samples.

← Back

COMPREHENSIVE GENOMIC LANDSCAPE OF MOST COMMONLY USED BREAST CANCER CELL LINES AND PATIENT DERIVED XENOGRAFT MODELS

Niantao Deng[1,2], Andre Minoche[3], Kate Harvey[1], Alex Swarbrick[1,2]

[1]Garvan Institute of Medical Research, Cancer Division, Sydney, Australia, [2]UNSW Sydney, St Vincent's Clinical School, Faculty of Medicine, Sydney, Australia, [3]Garvan Institute of Medical Research, Genomics and Epigenetics Division, Sydney, Australia

Breast cancer cell lines (BCCLs) are the most frequently used models in cancer research. According to PubMed search, the top six studies BCCLs are MCF7, MDAMB231, T47D, SKBR3, MCF10A and MDAMB468, which coverage more than 90% of all BCCL associated studies. Breast cancer derived xenograft models have also been widely studied and served as important preclinical tools. Despite the widespread usage of these models in breast cancer research, no comprehensive whole genome-wide sequencing has been performed on these models, with previous studies only focus on sequencing of a targeted gene panel or whole exome. Here we describe for the first time whole genome sequencing (WGS) of these models by Illumina X10 platform with average more than 60x coverage. We showed that WGS can identify additional critical genomic alterations, including point mutations and genomic rearrangements, compared to previous exome data. Through integrative analysis with public available RNA-Seq and ChIP-Seq data, our WGS data provides as a useful resource of comprehensive characterisation of these models for subsequent studies.

← Back

NOVOLoci: TARGETED ASSEMBLY AND VARIANCE DETECTION FROM WHOLE GENOME DATA.

Nicolas Dierckxsens[1], Patrick Mardulyn[3], Guillaume Smits[1,2]

[1]ULB-VUB, Interuniversity Institute of Bioinformatics Brussels (IB2), Brussels, Belgium, [2]Hôpital Universitaire des Enfants Reine Fabiola (HUDERF), Genetics, Brussels, Belgium, [3]Université Libre de Bruxelles (ULB), Evolutionary Biology and Ecology Unit, Brussels, Belgium

The continuous evolution in massive parallel sequencing (MPS) made it relatively easy and affordable to produce whole genome data from a DNA sample. When there is a close reference genome available, it is possible to map the reads for variance detection or to assemble the complete genome against the reference. Unfortunately only a small fraction of the species have an assembled genome, which makes it generally a long and complicated process to deduct reliable results from a MPS dataset. In most cases, a graph based approach is used to assemble a set of contigs from the whole genome dataset. Such assembly requires time and provides approximations to be efficient on a genome-wide basis. Therefore we developed a fast and accurate seed-extend *de novo* local assembler that will initiate assemblies for your regions of interest. NOVOLoci makes it possible to assemble regions around set of genes or markers the same day you obtained your MPS dataset. Illumina short reads are required to start the assembly, but can be supported by long reads from Nanopore or PacBio. If a reference is available, NOVOLoci can also create a vcf file with all the detected variances. Our benchmark study shows that it produces longer and more accurate contigs than the most used graph based assemblers and is able to accurately detect variances from SNV to larger structural variations.

EXPANSIONHUNTER: A SOFTWARE TOOL TO DETECT LONG REPEAT EXPANSIONS FROM PCR-FREE WHOLE-GENOME SEQUENCE DATA

Egor Dolzhenko[1], Joke J van Vugt[2], Kristina Ibáñez[3], Giuseppe Narzisi[4], Marka van Blitterswijk[5], Rosa Rademakers[5], Russell McLaughlin[6,7], Orla Hardiman[6,7], Ammar Al-Chalabi[8], Chris Shaw[8], Nancy S Wexler[9], David E Housman[10], Mark Caulfield[3], Ryan J Taft[1], Leonard H van den Berg[2], David R Bentley[11], Jan H Veldink[2], Michael A Eberle[1]
[1]Illumina Inc, Clinical Genomics Research, San Diego, CA, [2]University Medical Center Utrecht, Neurology, Utrecht, Netherlands, [3]Queen Mary University London, Genomics England, London, United Kingdom, [4]New York Genome Center, New York, NY, [5]Mayo Clinic, Neuroscience, Jacksonville, FL, [6]Trinity Biomedical Sciences Institute, Neurology, Dublin, Ireland, [7]Beaumont Hospital, Neurology, Dublin, Ireland, [8]King's College London, Basic and Clinical Neuroscience, London, United Kingdom, [9]Columbia University, The US-Venezuela Collaborative Research Group, New York, NY, [10]Massachusetts Institute of Technology, Koch Institute for Integrative Cancer Research, Cambridge, MA, [11]Illumina Cambridge Ltd, Clinical Genomics Research, Little Chesterford, United Kingdom

Identifying large repeat expansions such as those that cause amyotrophic lateral sclerosis (ALS) and fragile X syndrome (FXS) is challenging for short-read (100-150 bp) whole genome sequencing (WGS) data. A solution to this problem is an important step towards integrating WGS into precision medicine. To this end, we developed a software tool called ExpansionHunter that, combined with PCR-free WGS short-read data, can genotype repeats at a locus of interest even if the expanded repeat is larger than the read length. To test our method we applied ExpansionHunter to a set of 152 samples harboring repeat expansions associated with Huntington's disease, fragile X syndrome, Friedreich's ataxia and five other genetic disorders. All but one of the repeat expansions in these samples were assessed correctly even though many of the repeats were significantly longer than the read lengths. Motivated by this success, we used ExpansionHunter to analyze repeat sizes in two large cohorts of samples.

First, we applied our algorithm to WGS data from 3,001 ALS patients who have been tested for the presence of the C9orf72 repeat expansion with repeat-primed PCR (RP-PCR). ExpansionHunter correctly classified all (212/212) of the expanded samples as either expansions (208) or potential expansions (4). Additionally, 99.9% (2,786/2,789) of the wild type samples were correctly classified as wild type by this method with the remaining three identified as possible expansions.

We next applied ExpansionHunter to WGS data from 4,298 individuals with unidentified genetic diseases that were sequenced as part of the 100,000 Genomes Project Rare Disease Programme. Our findings included pathogenic expansions in C9orf72, HTT, and FMR1 genes that were subsequently validated. In one exemplary family, the mother had the FMR1 premutation (spanning 102 repeat units) and the child had the full expansion (223 repeat units). We will present the validation results of the ongoing ExpansionHunter discovery efforts in the 100,000 Genomes Project Rare Disease Programme project.

← Back

GMOVE : A TOOL FOR EUKARYOTIC GENE PREDICTION USING VARIOUS EVIDENCE

Marion Dubarry, Benjamin Noel, Tsinda Rukwavu, Sarah Farhat, Corinne Da Silva, Jean-Marc Aury

Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Evry, France

More and more laboratories can afford to sequence organisms thanks to the technological improvements that made it faster and available at a lower cost. Thus, the number of sequenced and assembled genomes has increased but they may not be annotated. Structural genome annotation is an important step of genomics analysis but it is still an issue in bioinformatics. It helps to highlight some parts of the genome that are coding for a protein to understand the functional roles of DNA. Genomes of interest are more and more complex and it could be difficult to train an annotation tool on an unknown organism.

Here, we present Gmove (Gene Modelling using Various Evidence), a eukaryotic genome annotation tool with no need of calibration step needed. It can take diverse data as input such as RNAseq alignments, cDNA alignments, proteic alignments and ab initio prediction and outputs potential coding genes. By taking advantage of all these types of sources, Gmove combines data and finds a consensus. Predicting multiple spliced transcripts is a rather difficult task due to short reads assembly but it is now possible with long RNAseq reads. The exon succession information contained in these long reads makes it possible for us to predict alternative gene forms. Indeed, Gmove is able to improve an existing annotation by combining the former annotation with additional data. As Gmove uses alignment data, it does not rely on canonical intronic bases.

We already ran this tool on a variety of organism such as a plant, a fungus, an insect and a dinoflagellate (genomes not published yet). In this study, we compare Gmove with Augustus, an *ab initio* tool and Maker, an annotation pipeline. We reannotated partially three reference genomes: *Arabidopsis thaliana, Danio rerio* and *Caenorhabditis elegans*. We also reannotate the *Mus musculus* genome with Nanopore data to demonstrate that Gmove can perform well with long reads information.

← Back

NGSEP3: ACCURATE, EFFICIENT AND USER FRIENDLY PRODUCTION AND ANALYSIS OF GENOMIC VARIATION DATASETS THROUGH STR-AWARE INTEGRATED REALIGNMENT

Jorge Duitama[1,2], Daniel Tello[1], Juan F De La Hoz[2], Cristian Loaiza[3], John J Riascos[3]

[1]Universidad de los Andes, Systems and computing engineering, Bogotá, Colombia, [2]International Center for Tropical Agriculture (CIAT), Agrobiodiversity research Area, Cali, Colombia, [3]Centro de investigación de la caña de azúcar de Colombia (Cenicaña), Biotechnology, Cali, Colombia

The widespread use of high throughput sequencing (HTS) technologies allowed the production of large genomic variation datasets for different populations of several species. These datasets are a great information resource in different research fields, enabling for example the development of genetic tools for molecular breeding of different crops. Although small labs are now able to efficiently produce large amounts of HTS data, comprehensive analysis of these data integrating different solutions remains a challenging task. We initially developed NGSEP as an open-source package that tightly integrates novel java implementations of algorithms for discovery of single nucleotide variants (SNVs), indels, and copy number variants, called either from a command line or from a rich graphical interface implemented in an Eclipse Plugin. NGSEP includes a one step wizard for parallel automated assembly of variation datasets from HTS data of entire populations, as well as filtering, statistics and format conversion of these datasets for integration with tools for assessment of population structure, GWAS, genomic prediction, among others. NGSEP can also be integrated with the Galaxy environment and is available within the CyVerse platform.

Understanding that incorrect alignments around small indels and short tandem repeats (STRs) are one of the main sources of false positives in variants discovery, we improved the variant detection procedure to identify variable mononucleotide runs from aligned reads and perform consistent realignments. STRs predicted by other tools can be provided, allowing NGSEP not only to perform variant-aware realignments but to directly genotype each STR as a single variation locus. This procedure is nearly transparent to the user, and more efficient in memory and disk usage compared to similar solutions. Up-to-date benchmarks using both simulated datasets and real data from cassava and rice populations show that NGSEP provides similar or better accuracy for variants detection compared to GATK, Samtools and Freebayes. We also developed new modules for downstream analysis including genotype imputation and construction of distance matrices and neighbor joining dendograms directly from VCF files. We expect that all these development efforts facilitate the use of NGSEP for a growing number of researchers in different fields of basic and applied genomics.

← Back

UNIFYING DATA SUBMISSION AT EMBL-EBI: A USER-FOCUSED APPROACH

Károly Erdős, Neil Goodgame, Flávia Penim, Rolando Fernandez, Galabina Yordanova, David Richardson, Amélie Cornélis, Upendra Kumbham, Laura Clarke

European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Molecular Archival Resources, Hinxton, United Kingdom

EMBL-EBI hosts archives for many different classes of biological data. Each provide their own system to accept data submissions. These are used by over 1000 submitters per month, who often submit data to multiple archives. This is time consuming and opens up the possibility of inconsistencies in the data entered across the different archives.

The Unified Submission Interface (USI) project aims to simplify the process by providing a single interface. Our user experience (UX) team carried out extensive research into the current users' practices and needs. As a result of our streamlining efforts, USI will allow users to deposit data with BioSamples, BioStudies, ENA, EGA, EVA, ArrayExpress, Metabolights and Pride through a single system. The new design will support both web-based and programmatic submissions.

The new web-based portal will offer a substantially enhanced user experience for submitters, based on information learnt during the UX discovery process. One of the early benefits is a single sign-on system, where users will no longer have to remember various passwords and wait for authentication for each logon. This will help to increase productivity by saving time, and by only having to learn one system in the future. In the new system, all archives will share a common workflow, but each archive will capture and validate only the specific metadata fields required for their own data types. It will also support community-defined rule sets to allow the different groups submitting to EMBL-EBI archives to define their own standards and improve their own archive.

In order to make the new system scaleable and sustainable in the future, it has been designed using an event-driven microservices architecture with agile development methodology, which ensures that we can extend and improve the platform with minimal disruption to the ongoing submission services.

We continue to engage with our submitters, and recently held a hackathon which provided valuable feedback. Our prospective users' response to our prototypes has been very positive. We are planning to go to production in the 3rd quarter of 2017 initially just with sample submissions to the BioSamples archive; then enabling submissions of sequencing data to the ENA archive in the 1st quarter 2018; and finally roll out support for all other archives later in the year.

We are confident that our new service will significantly improve user experience for scientists globally.

← Back

DXM: AN ALGORITHM TO DECONVOLVE GENOMIC DNA METHYLATION DATA TO UNDERSTAND EPIGENETIC CLONALITY.

Jerry Fong[1], Jacob R Gardner[2], Yu Sun[2], Kilian Q Weinberger[2], John R Edwards[1]

[1]Washington University in St. Louis School of Medicine, Center for Pharmacogenomics, Department of Medicine, St. Louis, MO, [2]Cornell University, Department of Computer Science, Ithaca, NY

DNA methylation refers to the presence of 5-methylcytosine in a CG dinucleotide context. In addition to a functional role in X-inactivation, genomic imprinting, and mammalian retrotransposon silencing, DNA methylation shares a complex relationship with gene expression. There is broad interest to understand how changes in DNA methylation contribute to human disease. For instance, increased DNA methylation heterogeneity in patients with diffuse large B-cell lymphoma (DLBCL) has been correlated with worse prognosis. However, it is difficult to study how DNA methylation changes functionally impact disease progression because many datasets are generated from sequencing of heterogeneous clinical samples. Several algorithms have been developed to deconvolve this data by using reference profiles of the expected cell types in the sample, which limits their applicability to broader research questions.

Thus, we developed a novel algorithm, DXM, which can deconvolve bisulfite sequencing data from a heterogeneous sample into its major subpopulations, their percent composition, and their respective methylation patterns without explicit use of reference cell-type methylation profiles. Briefly, for each gene, DXM solves the number of subpopulations present iteratively. The percent composition and methylation profiles are solved with an alternating optimization scheme including a modified Hidden Markov Model. Afterwards, the major methylation profiles for all genes are assigned to subpopulations by solving a mixed-integer quadratic program.

Tuning for DXM was conducted with bisulfite sequencing data from the Roadmap Epigenomics Project. To test the performance of DXM, reads from bisulfite sequencing experiments of various lymphocytes (Blueprint Epigenome Project) were subsampled and combined to form synthetic “mixtures,” where the read counts from each underlying cell type reflected its expected percent composition in the mixture. DXM offered reasonable estimates of the number of major subpopulations, their percent composition, and their respective methylation patterns across a range of mixture conditions. To further demonstrate the robust nature of DXM, we evaluated its performance in synthetic mixtures that included non-lymphocytic cell types as well as lymphoma cells. We are currently planning analyses to better understand how DNA methylation changes impact patients with DLBCL. In the future, we hope that DXM can be broadly applied to analysis of DNA methylation data from heterogeneous samples.

IMPROVING CREST WITH THE SENTIEON PYTHON API

Donald Freed[1], Scott Newman[2], Renke Pan[1], Michael Rusch[2], Stephen Rice[2], Luoqi Chen[1], Jinghui Zhang[2]

[1]Sentieon Inc, Mountain View, CA, [2]St Jude Children's Research Hospital, Department of Computational Biology, Memphis, TN

Structural variants in cancer genomes give rise to oncogenic gene fusions and other important gene disruptions. Detecting these events accurately and quickly from >30X tumor/normal paired whole genome sequencing represents a huge bioinformatics challenge. Many structural variant prediction algorithms exist but CREST (**C**lipping **RE**veals **ST**ructure) [PMID: 21666668], although originally published in 2011, is still one of the industry's leading tools. CREST has been extensively validated on real patient data and used to make numerous novel discoveries as part of the Pediatric Cancer Genome Project [PMID: 22641210]. Although the CREST output has been extensively validated on over 700 published samples, running the algorithm itself takes between 5 and 24 hours on a 20-core node depending on the structural complexity of the sample. This computational bottleneck represents a serious problem if CREST is to be used in a time-sensitive clinical environment or at scale in the cloud.

To address our need for more efficient data processing, we have implemented the CREST algorithm as a Python recipe that calls the Sentieon data processing engine. Communication between the recipe and the engine through the Sentieon Python API allows for a clean separation of low-level functionality for data processing and higher-level structural variant detection logic improving readability and maintainability. On test data, the reimplementation consistently achieves ~10x performance improvement over original CREST with no increase in RAM usage while maintaining fidelity to the original algorithm. For a 30X paired tumor-normal sample on a 20-core node, total wall time is reduced to less than one hour for most samples, with the most complex samples taking up to three hours. The Sentieon API provides a generic solution for efficient manipulation of genomics data enabling variant calling algorithms to be implemented in Python and other scripting languages.

← Back

ROBUST ANALYSIS OF SINGLE CELL TRANSCRIPTOMES USING SAKE IDENTIFIES MARKERS OF TARGETED INHIBITOR RESISTANCE IN MELANOMA

Ray Ho, David Molik, Molly Gale Hammell

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

The unprecedented resolution of Single-cell RNA-seq (scRNA-seq) data at a transcriptome wide scale enables us to address questions that are inaccessible using methods that profile gene expression by averaging over bulk cell extracts. Single cell transcriptomes have enabled the identification of new cell types in previously characterized tissues as well as new markers that specify subtypes within cellular populations. However, the high transcript drop-out rates and stochastic transcription events present in scRNA-seq data require robust classification methods.

Here we present SAKE, a comprehensive analysis package that includes quality control steps, cluster identification, quantitative cluster assignment for each cell, and gene expression analysis. Quantitative cluster assignments are identified using non-negative matrix factorization (NMF), a method that has been successfully applied to identify validated molecular subtypes in bulk transcriptome profiling studies. SAKE also provides a module for t-distributed stochastic neighbor embedding (t-SNE), a popular visualization method for high-dimensional datasets that has been widely used to identify clusters in single-cell studies. NMF not only identifies clusters that largely recapitulate sample distributions on the t-SNE plot, but also provides quantitative confidence estimates for the number of clusters present and for the assignment of each sample to a particular cluster. Finally, SAKE also provides downstream analysis modules for differential expression statistics between clusters and gene set enrichment analysis within each cluster. Importantly, SAKE is robust and computationally efficient, such that it can also be used for large-scale library pools, such as those produced from Drop-seq methods.

We have applied SAKE to scRNA-seq transcriptomes of melanoma cells responding to small molecule targeted inhibitors of the BRAF oncogene (vemurafenib). SAKE identified two major gene expression clusters that represent cells either sensitive or resistant to the BRAF inhibitors, as well as several subtypes within each population. Surprisingly, markers of inhibitor resistant cells are already present in a subset of the naïve cells that have never been exposed to BRAF inhibitors, suggesting that intrinsic population heterogeneity may give rise to resistant populations.

PLOT ANY DATA ON ANY GENOME WITH KARYOPLOTER

Bernat Gel, Eduard Serra

Germans Trias i Pujol Research Institute (IGTP), Hereditary Cancer, Badalona, Spain

Motivation: Data visualization is a crucial tool for data exploration, analysis and interpretation. It efficiently summarizes complex data, facilitates careful examination and can reveal non-obvious patterns in the data. The natural representation for genomic data is positioned along the genome next to the ideograms of the different chromosomes. Various genomic visualization tools are available, but they are either limited to a few species, offer limited customization options or create only circular plots. There is a lack of a tool to create customizable non-circular plots of whole genomes from any species.
Results: We have developed karyoploteR, an R/Bioconductor package to create linear chromosomal representations of any genome with genomic annotations and experimental data plotted along them. The philosophy behind karyoploteR was inspired by R base graphics, with a single function responsible for the creation and initialization of the plot and a rich set of graphical functions available to add data elements (points, lines, rectangles…) and non-data elements (labels, axis, titles…). In addition to the low-level graphical primitives, a set of higher level functions exist to add more complex graphical representations of the data such as the density of elements, a rainfall plot or individual markers for genes and other genomic features. All standard R graphical parameters are available to control colors, sizes, glyphs, etc, and there are additional arguments to specify the positioning of the data with total freedom. This allows the creation of highly customizable plots from arbitrary data with complete freedom on data positioning and representation. Users can also create their own custom graphical functions and integrate them with karyoploteR to extend it's functionality. Finally, karyoploteR can be used with any genome, either the ones included in the package or a custom one provided by the user.
Conclusions: We have developed an R/Bioconductor package, karyoploteR, to plot arbitrary genomes with arbitrary data positioned on them. It offers a flexible API inspired by R base graphics. The plots are highly customizable in data positioning and appearance and it is possible to extend the package functionality using user created custom functions.
Availability: karyoploteR is released under Artistic-2.0 License. Source code and documentation are freely available through Bioconductor (http://www.bioconductor.org/packages/karyoploteR) and at the examples and tutorial page at https://bernatgel.github.io/karyoploter_tutorial.

← Back

DYNAMIC CHANGE OF TRANSCRIPTION PAUSING THROUGH MODULATING NELF PROTEIN STABILITY REGULATES GRANULOCYTIC DIFFERENTIATION

Xiuli Liu[1,2,3], Aishwarya A Gogate[2,3], Melodi Tastemel[1,2,3,4], Venkat S Malladi[2,3], Huiyu Yao[5], Kim Nguyen[1,2,3], Lily Jun-Shen Huang[5], Xiaoying Bai[1,2,3,4]

[1]Laboratory of Molecular Genetics of Blood Development, Obstetrics and Gynecology, Dallas, TX, [2]Cecil H. and Ida Green Center for Reproductive Biology Sciences, Obstetrics and Gynecology, Dallas, TX, [3]Division of Basic Reproductive Biology Research, Obstetrics and Gynecology, Dallas, TX, [4]Genetics, Development and Disease Graduate Program, Obstetrics and Gynecology, Dallas, TX, [5]Department of Cell Biology, University of Texas Southwestern Medical Center, Cell Biology, Dallas, TX

The negative elongation factor (NELF) is known to be essential for establishing transcription pausing. The cellular regulation of the NELF protein and its role in specific lineage differentiation remains largely unknown. Adopting mammalian hematopoietic differentiation as a model system, we identified a dynamic change of NELF-mediated transcription pausing as a novel mechanism regulating hematopoietic differentiation. We observed that granulocytic differentiation triggered a decrease of NELF protein abundance with a subsequent genome-wide reduction of transcription pausing along with activation of granulocyte-related genes and a reduced expression of progenitor markers. To test this genome-wide reduction, we performed custom analysis of GRO-seq data from CD34+ HSPCs and early differentiated granulocytes. Consistent with previous studies showing Pol II pausing at majority of metazoan genes, metagene analysis of GRO-seq data from progenitor cells revealed a typical genome-wide distribution of transcriptionally engaged Pol II with high signal around TSS and low signal along the gene body representing elongating Pol II. In contrast, metagene analysis of early differentiated granulocytes revealed a drastic reduction of Pol II occupancy around TSS showing an overall reduction of pausing upon granulocytic differentiation. These analyses were further verified by a pausing index analysis. The early differentiated granulocytes showed a drastic reduction of Pol II occupancy around the TSS. Functional studies also revealed that sustained expression of NELF inhibits granulocytic differentiation whereas NELF depletion leads to premature differentiation towards granulocytic lineage. Taken together, our results uncovered a previously unrecognized mechanism of regulation of transcription pausing by modulating NELF protein abundance.

← Back

RANDOM ACCESS TO SEQUENCE GRAPHS STORED IN LARGE GFA FILES

Giorgio Gonnella, Stefan Kurtz

University of Hamburg, Center for Bioinformatics (ZBH), Hamburg, Germany

The GFA (Graphical fragment assembly) formats GFA1 and GFA2 are emerging formats for the representation of sequence graphs, including assembly and variation graphs.

The graph structure of a GFA file is recorded in nodes (segments, representing sequences) connected by different kind of arcs (links, containments, edges and gaps). In GFA files no particular order of the lines is required and lines defining arcs contain two references to segments. For these reason, traversing the graph usually requires to store the contents of the entire GFA file in memory. For large GFA files this becomes infeasible due to limitations of the memory.

We present a method for traversing a GFA graph, without reading the entire GFA file in memory. In particular, we implemented a software tool, which, in a preliminary phase employs an external sorting method to handle GFA files of unlimited size. Thereby, arcs referring to the segment as their first segment reference are sorted immediately after the segment line. The tool also outputs an index containing, for each segment name, the position of the segment line in the sorted GFA file, and a list of positions of the arcs referring to the segment as their second segment position.

As an example application, we implemented a tool, which extracts a subgraph from a GFA file, from a specified segment and traverses the arcs in breadth-first fashion, until a specified depth is reached. Thereby, only the index file must be kept in memory. The advantage of using the index in terms of memory requirement is particularly significant if the GFA file contains long segment names, or data other than the topology information, such as metadata, sequences, alignments, or assignments of reads to contigs.

ACCELERATING CONGENITAL HEART DEFECT VARIANT ANALYSIS THROUGH BIG DATA

David Gordon[1], Harkness Kuck[1], Ben Kelly[1], Grant Lammi[1], James Fitch[1], Stephanie LaHaye[1], Sara Fitzgerald-Butt[1,2], Vidu Garg[2,3], Kim McBride[1,2,3], Peter White[1,3]

[1]Nationwide Children's Hospital, Institute for Genomic Medicine, Columbus, OH, [2]Nationwide Children's Hospital, Center for Cardiovascular Research and Heart Center, Columbus, OH, [3]The Ohio State University, Department of Pediatrics, Columbus, OH

Congenital Heart Defects (CHD) are present in nearly 1% of newborns in the United States, and are responsible for more neonatal deaths than any other type of birth defect. The number of genes known to be associated with the molecular pathogenesis of CHD has grown with the advent of next generation sequencing; however, a significant number of cases remain unsolved by traditional methods of exome sequencing and single nucleotide variant analysis. By using techniques such as whole genome sequencing, computational structural variant detection, and linked read sequencing, we can greatly broaden the scope of our studies.

However, this increase in scope brings with it challenges in both data size and integration. For example, pipelines created for SNV analysis often do not handle structural variants, and may not scale well to whole genome data. Out of the box linked read sequencing tools do not perform multisample structural variant calling, making segregation analysis challenging.

To this end, we have developed Varhouse, a variant warehouse and associated suite of tools, using the Apache Spark and Apache HBase ecosystems, to accelerate the analysis of variants in our congenital heart disease study, which includes both single family and large cohort cases. Varhouse incorporates genotype, variant annotation, and gene annotation data in a single warehouse, enabling efficient retrieval and analysis of annotated variants by researchers. Varhouse also allows researchers to more easily make queries spanning sequencing technologies, aggregate results from multiple sequencing runs, and perform cohort level variant analysis.

Through the use of Varhouse, we have successfully identified segregating structural variants in multiplex CHD families from linked read sequencing data, analyzed enriched gene sets within large cohort CHD studies, and identified candidates for oligogenic disease causation in CHD families. Our future plans include additional feature development specific to CHD analysis and generalization of our algorithms for use in other single family and large cohort genomic studies.

← Back

ABAENRICHMENT AND *GOFUNCR*: TWO R-PACKAGES FOR ONTOLOGY ENRICHMENT ANALYSES

Steffi Grote, Kay Prüfer, Janet Kelso, Michael Dannemann

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

Enrichment analyses using ontologies that describe various aspects of biology have been widely used to evaluate sets of candidate genes identified in biomedical studies. Perhaps the most widely used of these is the Gene Ontology (GO), which provides a graph representation of concepts of molecular biology to which genes are annotated based on scientific literature. The Allen Human Brain Atlas projects have developed an ontology that describes the brain anatomy and collected extensive information about the location and timing of gene expression in the human brain. The age of donors from whom expression data were obtained ranges from prenatal to adult, and for the adult brain over 400 different brain regions were measured. This therefore provides a valuable resource that can be used to pinpoint sets of genes that are characteristic of particular brain regions and/or developmental stages.
However, none of the tools developed allows for the statistical evaluation of gene candidates in terms of expression enrichment in specific regions or stages using the brain ontology.

We present here *ABAEnrichment* a freely available R package that combines fine-scale spatial and ontogenetic gene expression data from the Allen Brain Atlas with a brain anatomy ontology, and annotates genes with expression exceeding a user-defined threshold to nodes of the ontology.
It is then possible to test whether the candidate genes are over-represented (i.e. over-expressed) in particular brain regions compared to a set of background genes. We also present *GOfuncR* which performs a classic GO enrichment analysis based on the previously published FUNC but now with extended options and implemented in R.
For easy usage both packages have integrated annotations and ontologies. The packages provide the standard candidate vs. background enrichment analysis using the hypergeometric test, as well as three additional tests: (i) the Wilcoxon rank-sum test that is used when genes are ranked, (ii) a binomial test that can be used when genes are associated with two counts, e.g. amino acid changes since a common ancestor in two different species, (iii) a Chi-square or Fisher's exact test that is used in cases when genes are associated with four counts., e.g. synonymous and non-synonymous variants that are fixed between or variable within species. To correct for multiple testing and interdependency of the tests, family-wise error rates are computed based on random permutations of the gene-associated variables. Both packages also provide tools for exploring the ontology graph and the annotations, and options to take gene-length or spatial clustering of genes into account. Both packages can be integrated into workflows for analyzing gene sets and their associations.

ABAEnrichment and *GOfuncR* are available from https://bioinf.eva.mpg.de/ and Bioconductor.

← Back

GENOMECHRONICLER: THE PGP-UK GENOMIC REPORT GENERATOR

Jose Afonso Guerra-Assuncao, Lucia Conde, Javier Herrero, on behalf of the PGP-UK Consortium
University College London, Cancer Institute, London, United Kingdom

The Personal Genome Project UK (PGP-UK) is part of the PGP Global Network. We provide genomic, transcriptomic, epigenomic and self-reported phenotypic data under an open-access model with full ethical approval. This is achieved by using a strict enrolment process, which ensures the participant understands the potential consequences of making these data freely available.

To date, whole genome sequencing and respective reports have been generated for 100 participants. Our ethical approval also allows us to accept genome donations, where participants can submit external genomic information instead. Genome donors are subject to the same enrolment rules as the remaining participants.

After sample collection, lab work and data analysis, participants receive a detailed non-medical grade genome report describing some known phenotypes associated with the genotypes they present. They then have time to use their data in confidence before both data and report become public under open-access. At any point, they can withdraw consent and request deletion of their data. When approved, whole genome sequencing data are submitted to the European Nucleotide Archive, inferred genomic variants to the European Variation Archive and methylation and transcriptome data to ArrayExpress.

Here we present the computational pipeline that links the variants, found within each participant, with existing genotype to phenotype resources and displays the results in a report targeted at lay persons.

The report includes summary statistics, highlighting the vast numbers of variants that characterise each individual human genome compared with the reference. The provided trait data is based primarily upon information available in SNPedia, and it provides the participants with a list of phenotypes likely associated with their genotypes. It also integrates information from ClinVar to provide additional details on clinical implications of the variants, as well as ExAC to show population frequency of certain protein coding variants.

Besides exploring phenotypes previously associated with certain genotypes, the ancestry of each participant is inferred after merging a subset of unlinked single nucleotide polymorphisms from the participant with those of different populations from the 1000 Genomes Project.

This report generation pipeline is currently undergoing final optimisation before general release. We aim to release a standalone version to allow individuals with their genome sequence to produce a personalised genome report. While this has been developed specifically for the PGP-UK project, the code will be freely available and can be relevant to other projects with similar aims.

← Back

SPLICE-QTLs IN THE CONTEXT OF PREDISPOSITION TO COLORECTAL CANCER

Toby Gurran, Victoria Svinti MacLeod, Maria Timofeeva, Malcolm Dunlop, Li Yin Ooi, Peter Vaughan-Shaw, Alison Meynert, Susan Farrington, Colin Semple

University of Edinburgh, Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom

The heritability of colorectal cancer (CRC) has been estimated at approximately 12% based on twin studies and family lineage analyses. Certain rare, high penetrance variants are well characterized; though these are estimated to account for only ~5% of all CRC cases. The majority of GWAS-identified risk SNPs for CRC fall within non-coding regions, and the mechanisms by which the majority of these mutations contribute to disease predisposition are yet to be elucidated. However, recent data has suggested that variants altering splicing patterns may play roles in cancer predisposition and progression.

This study has analyzed 231 samples of colonic mucosa (the target tissue in CRC) from a cohort of Scottish CRC patients and controls. Two separate methods of splice-QTL identification were used and the union of the two sets analyzed. Over 600 non-coding SNPs were identified as splice-QTLs associated with changes in the expression ratios of transcripts produced by a gene. As expected the identified SNPs fall predominantly within the introns of genes, and are predicted to disrupt intron-exon boundaries and spliceosome binding sites. Genes previously implicated in the development and progression of CRC were found to have associated splice-QTLs. The splice-QTLs were found to be significantly enriched within risk SNPs identified by a large GWAS meta-analysis for CRC predisposition using 20,000 cases and 37,000 matched controls. These findings indicate that alternative splicing may provide an explanation of the underlying functional mechanisms by which some SNPs predispose to colorectal cancer.

← Back

IMPROVING COMMUNITY SEARCH ON AN IDENTITY-BY-DESCENT GRAPH WITH MILLIONS OF INDIVIDUALS

Harendra Guturu[1], and the AncestryDNA science team[1,2]

[1]Ancestry.com DNA, LLC, San Francisco, CA, [2]Ancestry.com DNA, LLC, Lehi, UT

AncestryDNA is a direct-to-consumer DNA testing company that specializes in offering customers insights into their genetic genealogy. We recently launched Genetic Communities^{TM}, a new product that allows individuals to learn about their more recent ancestry at a finer resolution compared to traditional ethnicity estimation products. The Genetic Communities^{TM} product leverages the identity-by-descent (IBD) genetic network representing the shared DNA between related individuals. Using individuals consented to research, we previously demonstrated that community detection on the massive IBD graph allowed us to resolve population sub-structure resulting from recent human migration and mating patterns. Furthermore, we showed that machine learning classification techniques allow us to assign (i.e., search) communities for new individuals without the computationally expensive community detection process. Using classification, we are able to go beyond the single community assignment nature of typical community detection algorithms and assign individuals to multiple communities (more representative of natural populations). Expanding on this work, we revisited the community search process to improve assignment accuracy for new individuals. We explore two orthogonal methods to improve multi-community assignment accuracy and the stability of assignments among related individuals. First, we characterize further the classifiers and the features being used to assign communities. Second, we investigate more graph theoretic community search measures, such as minimum degree and k-truss. Our investigations reveal insight into improving the community search process for new individuals to offer detailed resolution of their genetic ancestry, while continuing to satisfy the need to scale to millions of individuals.

← Back

ESTIMATING RNA EXPRESSION USING PERSONAL GENOMES

Gisli H Halldorsson[1], Snaedis Kristmundsdottir[1], Birte Kehr[2], Reynir L Guðmundsson[1], Bjarni Gunnarsson[1], David Sverrisson[3], Eirikur Hjartarson[3], Gisli Masson[3], Daniel Guðbjartsson[1], Pall Melsted[1], Kari Stefansson[4], Bjarni V Halldorsson[1]

[1]deCODE genetics/Amgen, Statistics, Reykjavik, Iceland, [2]Berlin Institute of Health, Genome Informatics, Berlin, Germany, [3]deCODE genetics/Amgen, Bioinformatics, Reykjavik, Iceland, [4]deCODE genetics/Amgen, Population genomics, Reykjavik, Iceland

RNA expression is generally quantified by mapping RNA sequence reads to reference transcript databases and genomes. This introduces a mapping bias as reads containing non-reference alleles are less likely to align than reads matching the reference. We use the long range phasing of sequence variants to reconstruct its maternally and paternally inherited genomes. We map reads using the STAR aligner separately to the maternal and paternal genome references. Subsequently we annotate the primary alignment and represent it in GRCh38 genome coordinates. Incorporating personal genomes into our alignment pipeline improves estimates of alternative splicing and allele specific expression by minimizing allelic mapping bias.

← Back

THE SWEETPOTATO GENOMICS RESOURCE

John P Hamilton, Kin Lau, Krystle Wiegert-Rininger, C. Robin Buell

Michigan State University, Department of Plant Biology, East Lansing, MI

Sweetpotato (*Ipomoea batatas*) is an important food crop throughout the world, especially in developing countries due to its caloric and nutritional benefits. The sweetpotato genome is large and complex due to its ploidy (hexaploid) and heterozygosity. To facilitate genome-enabled breeding in hexaploid sweetpotato, participants in the Bill & Melinda Gates Foundation-funded Genomic Tools for Sweetpotato Improvement project have sequenced and assembled pseudomolecules (version 3) for two diploid, inbred *Ipomoea* species: *Ipomoea trifida* (NSP306) and *Ipomoea triloba* (NSP323) to serve as reference genome sequences for the hexaploid sweetpotato genome. We have annotated the assemblies of both species by creating an annotation pipeline that incorporates: creating a curated custom repeat library, repeat masking, training gene finders using RNA-Seq evidence, and annotating protein-coding genes using a set of transcript and protein evidence. The resulting gene models were then improved with PASA2. The *I. trifida* NSP306 pseudomolecules are 462 Mb with 32,301 annotated high confidence gene models, whereas the *I. triloba* NSP323 pseudomolecules are 457 Mb with 31,426 annotated high confidence gene models. The Sweetpotato Genomics Resource (http://sweetpotato.plantbiology.msu.edu) has been updated for the release of the version 3 pseudomolecules and contains a set of search and query tools for the annotation, including a BLAST server, Jbrowse genome browsers for the reference genome annotation, and comprehensive gene report pages for all annotated genes in the two species. The gene report page and genome browsers also include hexaploid sweetpotato variant data from whole genome sequencing and RNA-Seq to assist breeders, including variants from the sequencing of 16 cultivated sweetpotato varieties and gene expression data for 15 biotic and abiotic stress experiments.

CREATING PIPELINES THAT ARE REPEATABLE, TRACEABLE, AND SHAREABLE FOR CLINICAL GENOMICS AND RESEARCH

Nathan A Hammond, Sowmithri Utiramerur

Stanford Healthcare, Clinical Genomics Proram, Palo Alto, CA

There are significant challenges in sharing genomics data and analysis pipelines. Large, distributed workflows are often designed to use a specific HPC scheduler and depend on locally installed executables or other features of a particular environment, which limits portability and fails to guarantee reproducibility if the original compute environment is modified or unavailable.

Recently there have been a number of efforts to address the challenge of sharing data and reproducible pipelines in genomics. The Cancer Genomics Pilot from SBGenomics provides a cloud environment that collocates compute resources with TCGA data and lets users run existing or custom workflows. The Broad Institute’s FireCloud is a centrally managed portal that lets users run pipelines with their own Google billing account. Commercial offerings such as DNAnexus, SBGenomics, and Arvados offer hosted environments to run workflows and manage data. CWL has defined a standard language for pipeline definitions, and has accumulated a small ecosystem of software tools that supported it.

Nevertheless a typical research lab or medical center, if they are unwilling to accept technical lock-in with one of the commercial products both for themselves and their sharing partners, will find it difficult to assemble a production-ready software stack that ensures reproducibility and shareability of distributed workflows.

In order to address these challenges at the Stanford Clinical Genomics Program (a joint initiative by Stanford Health Care and Lucile Packard Children’s Hospital) we have developed Loom, a workflow engine and database that is free and open source under the Gnu Affero General Public License.

Key features of Loom include:

- Platform-agnostic workflow definitions
- Portability and reproducibility of runtime environments through use of Docker containers
- Support for parallel workflows, including multi-level nested iteration
- Record-keeping for reproducibility and traceability. Workflows, analysis runs, and results are saved as database records and JSON flat files
- Support for Google Cloud Platform and stand-alone Linux-based servers, with plans to support other platforms in the future
- Command-line client
- Browser interface

We present Loom both as a free tool available for use, and as a case study for those designing their own pipeline software or workflow specification language for clinical or research use.

PRECISE DETECTION AND SPECIFICATION OF STRUCTURAL VARIATION IN GENOMES

Nancy F Hansen, James C Mullikin

Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, NHGRI, NIH, Bethesda, MD

Structural variants (SVs) are genomic insertions, deletions, inversions, and translocations at least 50 base pairs in length, and have been shown to account for genomic disorders with grave clinical implications. While great progress has been made in developing algorithms to detect these variants from sequence data, discovery of SVs remains challenging.

One important factor exacerbating this difficulty is the association of SVs with repetitive DNA. Sequence similarity near the breakpoints of SVs leads to ambiguity both in the evidence they exhibit in sequencing read data and in their exact specification. Now, as new technologies enable us to determine long-range, accurate consensus for sequenced samples, it is appropriate and of great importance that we examine SVs carefully to describe them precisely and proceed to determine their exact effect on gene function.

Here we present a software toolkit for the precise detection and specification of structural variants, called SVanalyzer. The tools in SVanalyzer allow users to characterize the ambiguity of SVs with respect to nearby sequence similarity (SVwiden), detect equivalent SV predictions by comparing altered sequences (SVcomp), genotype known SVs in new datasets (SVbackgenotype), and refine SV predictions using long-read assemblies (SVrefine). SVanalyzer is currently being used by the National Institutes of Standards and Technology's "Genome in a Bottle" project to integrate SV calls from multiple sequencing platforms.

In addition, we present here an example of how imprecise specification of structural variation can lead to inaccurate variant annotation, potentially leading to misguided functional studies. Examples like this one demonstrate how critical it is to consider not only the genomic location and size of a structural variant, but also the surrounding landscape of sequence similarity.

ROSLIN: A PORTABLE AND REPRODUCIBLE WORKFLOW INFRASTRUCTURE FOR CANCER GENOMIC SEQUENCING ANALYSIS

Chris Harris, Jaeyoung Chun, Ronak Shah, Nikhil Kumaf, Ewa Reza, Aaron Gabow, Oliver A Hampton, Nicholas D Socci

Memorial Sloan Kettering Cancer Center, Center for Molecular Oncology, New York, NY

Reproducibility and portability are persistent problems in the analysis of genomic sequence data. Usually, sequence analysis pipelines need to be run in their local environments or the specific cloud computing resource for which they were designed. Even when constructed for portability, correctly maintaining the versions for all required software (e.g., alignment, variant calling), reference builds, and annotations becomes complicated to the point of impracticality in the least and impossibility at the most. As a solution to these technical challenges, we present *Roslin*–a portable pipeline system developed at Memorial Sloan Kettering Cancer Center for the calling of variants in sequencing data.

Roslin is written in the Common Workflow Language (CWL), a specification standard requiring tasks to be modularized and inputs and outputs be explicitly defined. The requirements of explicitness and modularization enable CWL workflows to be flexible, portable, scalable, and amenable to container technologies such as Docker and Singularity. *Roslin* utilizes the Toil workflow manager from the University of California at Santa Cruz, a portable, open-source workflow engine that supports CWL and is designed to securely and reproducibly run scientific workflows efficiently at scale. *Roslin* leverages Singularity, the Lawrence Berkley National Laboratory container system that is notionally similar to Docker. The Singularity container system is designed for mobility of compute and reproducibility of scientific analysis. Singularity containers are used to package complete scientific workflows, software and libraries, and data. Combining these makes *Roslin* well suited to run versioned bioinformatics workflows on cluster, cloud, and high performance computing environments at scale.

Roslin has been deployed and tested on multiple high performance computational clusters and cloud computing resources. *Roslin* supports complete versioning of its workflows, the underlying software and libraries, and associated resource files. It offers end users GUI driven workflow logging, run reporting and real time tracking. The *Roslin* CWL workflows are also suitable for deployment and execution on platforms without Toil, as Docker versions of every Singularity container are also provided.

← Back

A MAP OF HIGHLY CONSTRAINED CODING REGIONS IN THE HUMAN GENOME.

James M Havrilla[1,2], Brent S Pedersen[1,2], Ryan M Layer[1,2], Aaron R Quinlan[1,2,3]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]University of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]University of Utah, Department of Biomedical Informatics, Salt Lake City, UT

Interspecies sequence conservation summarizes the degree of genetic constraint over vast evolutionary periods. In contrast, catalogs of genetic variation from thousands of exomes and genomes enable the inference of more recent constraint from extreme paucities of genetic variation. While existing techniques such as pLI and RVIS summarize constraint for entire genes, it is clear that single metrics do not capture the variability in constraint that exists within each protein coding gene. To address this limitation, we have charted a detailed map of constrained coding regions (CCRs) within the human genome by leveraging coding variation observed among 123,136 humans from version two of the Exome Aggregation Consortium (ExAC). Constrained coding regions arise when the observed distance between missense variants in ExAC v2 — a proxy for constraint — is much greater than expected by chance.

We demonstrate that, as expected, our most constrained coding regions are significantly enriched (154-fold) for ClinVar pathogenic variants over benigns in CCRs greater than the 95th percentile. Moreover, pathogenic de novo variants, identified in individuals with developmental delay, severe intellectual disability, and epileptic encephalopathy, are substantially (6.35-fold) enriched in CCRs in the 95th percentile when compared to benign de novo variants from unaffected siblings of autism patients. The most constrained CCRs are found in many genes known to be associated with severe autosomal dominant phenotypes and important cellular function. CCRs also reveal protein domain families under extreme constraint, suggest unannotated functional domains from the degree of constraint, and perform equal to or above existing tools for the prioritization of previously unseen variation in studies of disease. Finally, CCRs under the highest constraint hold the promise of revealing coding regions under extreme purifying selection and genes with yet unobserved human phenotypes owing to embryonic lethality.

← Back

ALLELE SPECIFIC HLA LOSS IS A PERVASIVE MECHANISM OF IMMUNE EVASION AND IS PERMISSIVE FOR NON-SMALL CELL LUNG CANCER EVOLUTION

Nicholas McGranahan[1], Rachel Rosenthal[1,5], Andrew J Rowan[2], Thomas B Watkins[2], Gareth A Wilson[1,2], Nicolai J Birkbak[1,2], Selvaraju Veeriah[1], Peter Van Loo[3,4], Javier Herrero[5], Charles Swanton[1,2], on behalf of the TRACERx consortium[1]

[1]UCL Cancer Institute, Cancer Research UK Lung Cancer Centre of Excellence, London, United Kingdom, [2]The Francis Crick Institute, Translational Cancer Therapeutics Laboratory, London, United Kingdom, [3]The Francis Crick Institute, Cancer Genomics Laboratory, London, United Kingdom, [4]University of Leuven, Department of Human Genetics, Leuven, Belgium, [5]UCL Cancer Institute, Bill Lyons Informatics Centre, London, United Kingdom

Cancer cells adopt a variety of mechanisms to evade T-cell recognition. HLA down-regulation has been proposed as an immune escape strategy in many cancers, including lung cancer. However, mutations in HLA class I genes are infrequent (<5%) in early stage non-small cell lung cancers (NSCLC), suggesting alternative mechanisms of HLA disruption may be common. To systematically explore the prevalence and importance of genomic HLA disruption, we present LOHHLA (Loss Of Heterozygosity in Human Leukocyte Antigen), a tool to accurately determine HLA haplotype specific copy number in tumors.

In essence, LOHHLA realigns the reads from the tumour and the germline on the HLA alleles inferred by Polysolver (Shukla *et al*, 2015) or Optitype (Szolet *et al*, 2014) to obtain relative tumour coverage (logR) and B-allele frequencies (BAF) for each HLA locus making use of all polymorphic sites. Allele-specific HLA copy number is determined, taking into account tumour purity and ploidy estimates obtained from ASCAT (Van Loo *et al*, 2010) or FACETs (Shen *et al*, 2016). Furthermore, in TRACERx (Jamal-Hanjani *et al*, 2017) for which we have inferred clonal tumors' evolutionary trees using CITUP (Malikic *et al*, 2015), we have developed a method to assign the HLA allele losses to particular branches.

We find HLA LOH, where either the maternal or paternal HLA allele is deleted, occurs in 40% of early-stage NSCLCs from TRACERx. The focal nature of these alterations, their subclonal frequencies, and their occurrence as parallel events on distinct branches of tumors' evolutionary trees strongly suggests that HLA LOH is positively selected as a later event in NSCLC tumor evolution. Futhermore, NSCLCs exhibiting HLA LOH were characterized by a high subclonal nonsynonymous burden, enriched for neoantigens binding to the lost HLA alleles, APOBEC mediated mutagenesis, and up-regulation of signatures of cytolytic activity. Characterizing HLA haplotype specific copy number with LOHHLA refines neoantigen prediction and may have implications for vaccine or TIL based immunotherapeutic approaches.

← Back

UCSC VARIANT ANNOTATION INTEGRATOR COMMAND-LINE WRAPPER AND HGVS VARIANT NOMENCLATURE SUPPORT

Angie S Hinrichs, Christopher M Lee, Christopher Villarreal, Robert M Kuhn, Brian T Lee, Cath Tyner, Luvina Guruvadoo, Maximilian Haeussler, Ann S Zweig, W J Kent

UC Santa Cruz, Genomics Institute, Santa Cruz, CA

The UCSC Genome Browser's Variant Annotation Integrator (VAI) web tool (https://genome.ucsc.edu/cgi-bin/hgVai, Hinrichs et al. 2016) offers an easy way to input variants and obtain functional effect predictions from any of UCSC's many gene tracks as well as scores from several protein damage estimation algorithms, estimates of conservation, and more. However, as a web tool it is limited in the number of variants that it can process in a single request, and requires variants to be transmitted to UCSC over the Internet which may not be permitted for human subject data. To support off-line processing of much larger numbers of private variants, we have added a command-line wrapper script (vai.pl) so that VAI can be run on a local host, using locally downloaded data and/or remotely querying the UCSC database depending on configuration.

VAI now supports the Human Genome Variation Society (HGVS) variant nomenclature, as variant input in the web tool and/or annotation output in both web and offline modes. Variant input HGVS terms must be at the nucleotide level (g., n., c.) and well-formed with RefSeq or Locus Reference Genomic (LRG) transcript accessions. When RefSeq Genes are selected as the transcript set for functional effect prediction, VAI can optionally add HGVS genomic (g.), coding (c.), noncoding (n.) and protein (p.) nomenclature to the output for any type of variant inputs. HGVS 3'-shifting rules are applied so sometimes the HGVS nomenclature terms are several bases away from the directly mapped genomic variant location. The conversion of genomic variants to HGVS terms has been rigorously tested by comparing results to Mutalyzer (https://mutalyzer.nl/, Vis et al. 2015), VariantValidator (https://variantvalidator.org/) and a curated set of test variants (https://github.com/personalis/hgvslib, Lee et al. 2017).

HGVS variant nomenclature and similar notations are now accepted by the UCSC Genome Browser's position/search function as well.

References

Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, Kuhn RM, Rosenbloom KR, Zweig AS, Haussler D, Kent WJ. UCSC Data Integrator and Variant Annotation Integrator. 2016. Bioinformatics. 32(9):1430-2.

Vis JK, Vermaat M, Taschner PE, Kok JN, Laros JF. 2015. An efficient algorithm for the extraction of HGVS variant descriptions from sequences. Bioinformatics. 31(23):3751-7.

Yen JL, Garcia S, Montana A, Harris J, Chervitz S, Morra M, West J, Chen R, Church DM. 2017. A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med. 9(1):7.

← Back

HIGHLY PARALLEL AND MEMORY EFFICIENT COMPACTED DE BRUIJN GRAPH CONSTRUCTION

Guillaume Holley, Páll Melsted, Trausti Saemundsson

University of Iceland, Faculty of Computer Science, Reykjavík, Iceland

De Bruijn graphs are the core data structure for a wide number of whole genome and transcriptome assemblers processing High Throughput Sequencing datasets. However, memory consumption of such assemblers is often prohibitive, due to the large number of vertices and edges in the graph, to the point of hindering the use of assemblers on large and complex genomes. Most short-read assemblers based on the de Bruijn graph paradigm reduce the assembly complexity and memory usage by compacting first all maximal non-branching paths of the graph into single vertices. Yet, such a compaction is challenging as it requires the uncompacted de Bruijn graph to be available in memory. We present a new parallel and memory efficient algorithm enabling the direct construction of the compacted de Bruijn graph without producing the intermediate uncompacted de Bruijn graph. Our method relies on a space and time efficient data structure, the Bloom filter, enhanced with minimizer hashing to increase cache performance. Despite making exstensive use of a probabilistic data structure, our algorithm guarantees that the produced compacted de Bruijn graph is deterministic. Furthermore, the algorithm features de Bruijn graph simplification steps used by assemblers such as tip clipping and isolated unitig removal. In addition, as disk-based software performance is significantly affected by the discrepancy of speed among disk storage technologies, our method uses only main memory storage. Experimental results show that our algorithm is competitive with state-of-the-art de Bruijn graph compaction methods.

A MANUAL ANNOTATION WORKFLOW INTEGRATING SIGNALS OF PROTEIN-CODING CONSERVATION TOGETHER WITH NEXT-GENERATION SEQUENCING TO IDENTIFY NOVEL PROTEIN-CODING AND PSEUDOGENE LOCI

Toby Hunt[1], Jonathan M Mudge[1], Irwin Jungreis[2], Jose Gonzalez[1], Manollis Kellis[2], Adam Frankish[1], Paul Flicek[1]

[1]EMBL-EBI, Vertebrate Genomics, Cambridge, United Kingdom,
[2]Massachusetts Institute of Technology, CSAIL, Cambridge, MA

Despite the completed first draft of the Human genome becoming available as far back as 2001, a plethora of genome annotation projects have subsequently not been able to compile a definitive Human geneset. Not only is the complete set of protein-coding genes still not accurately established but for many loci the total number of transcripts, and their putative functionality, are still not definitively known or agreed upon.

As part of the GENCODE project we produce the gold standard, manually curated human geneset, available in all major genome browsers. Here we take a highly filtered subset of the output of phyloCSF, a comparative genomics method which identifies evolutionary conserved regions of coding potential, to guide the identification and annotation of previously unknown novel protein-coding and pseudogene loci.

Utilising the standard manual annotation workflow of the GENCODE project we systematically manually analysed over 1000 high-scoring human phyloCSF regions and leveraged recently released, publicly available RNAseq models, together with long-read PacBio/SLR-seq data, to annotate 147 novel protein-coding genes and 170 pseudogenes not previously found in the GENCODE geneset, as well as adding additional exonic sequences within 277 existing protein-coding genes or pseudogenes.

The majority of these novel loci have not previously been reported, highlighting gaps in the annotation strategies used even on the most commonly studied of organisms. These new genes represent an exciting set of loci for further experimental characterization as well as clearly demonstrating the advantages of combining distinct evidence sets within discovery-based workflows.

On behalf of the GENCODE consortium.

← Back

THE COMPLEX SEQUENCE LANDSCAPE OF MAIZE REVEALED BY SINGLE-MOLECULE TECHNOLOGIES

Yinping Jiao[1], Paul Peluso[2], Jinghua Shi[3], Michelle Stitzer[4], Bo Wang[1], Michael Campbell[1], Joshua Stein[1], Xuehong Wei[1], Nathan Springer[5], Richard McCombie[1], Gernot Presting[6], Michael McMullen[7], Jeffrey Ross-Ibarra[8], R. Kelly Dawe[9], Alex Hastie[3], David Rank[2], Doreen Ware[1,10]

[1]Cold Spring Harbor Laboratory, Plant Biology, Cold Spring Harbor, NY, [2]Pacific Biosciences, Menlo Park, CA, [3]BioNano Genomics, San Diego, CA, [4]University of California, Davis, Department of Plant Sciences and Center for Population Biology, Davis, CA, [5]University of Minnesota, Department of Plant Biology, St. Paul, MN, [6]University of Hawaii, Department of Molecular Biosciences and Bioengineering, Honolulu, HI, [7]USDA-ARS, Plant Genetics Research Unit, Columbia, MO, [8]University of California, Davis, Department of Plant Sciences, Center for Population Biology, and Genome Center, Davis, CA, [9]University of Georgia, Athens, GA, [10]USDA-ARS, NEA Robert W. Holley Center for Agriculture and Health, Ithaca, NY

Complete and accurate reference genomes and annotations provide fundamental tools for characterization of genetic and functional variation. These resources facilitate elucidation of biological processes and support translation of research findings into improved and sustainable agricultural technologies. The current assembly of the maize genome, based on Sanger sequencing, was first published in 2009. Although this initial reference enabled rapid progress in maize genomics, the original assembly is composed of more than 100,000 small contigs, many of which are arbitrarily ordered and oriented, significantly complicating detailed analysis of individual loci and impeding investigation of intergenic regions crucial to our understanding of phenotypic variation and genome evolution. Here we report a next-generation assembly and gene annotation of maize, generated using sequence data obtained using single-molecule technologies, and characterization of structural variations using a high-resolution whole-genome optical map. The pseudomolecules of maize B73 RefGen_v4 are assembled nearly end-to-end(contig N50 = 1.2 Mb; scaffold N50 = 9.6 Mb), representing a 52-fold improvement of contiguity relative to the previous reference genome. Gene annotations were updated using 111,000 transcripts obtained by single-molecule sequencing, thereby improving our knowledge of gene structure and transcript variation. Using the updated reference, we demonstrated that the rate of loss of conserved genes is higher in maize than in other grass species. Characterization of the repetitive portion of the genome revealed over 130,000 intact transposable elements (TEs), allowing us to identify TE lineage expansions unique to maize. In addition, comparative optical mapping of two other inbred lines revealed a prevalence of deletions in regions of low gene density and genes specific to the maize lineage.

The new maize reference genome project is supported by NSF Gramene grant IOS-1127112, NSF Cereal Gene Discovery grant 1032105, USDA-ARS CRIS 1907-21000-030-00D and NSF Plant Genome award 1238014

← Back

CHARACTERIZING EPIGENETIC INTRATUMORAL HETEROGENEITY IN GLIOMA USING SINGLE-CELL REDUCED REPRESENTATION BISULFITE SEQUENCING

Kevin C Johnson [1], Marcos R Estecio[2], Roel G Verhaak[1]

[1]The Jackson Laboratory for Genomic Medicine, Computational Biology, Farmington, CT, [2]The University of Texas MD Anderson Cancer Center, Epigenetics and Molecular Carcinogenesis, Houston, TX

Determining the cellular mechanisms that govern glioma heterogeneity can impact the development of novel therapies, protect patients from side effects of unnecessary treatment, and prevent glioma recurrence. Emerging evidence suggests that molecular subtypes in glioma, based on genotype (e.g., IDH1 mutations) and DNA methylation profiles (glioma CpG Island Methylator Phenotype, G-CIMP), can provide clinically relevant tumor classifications. However, traditional bulk sampling of gliomas to profile molecular features fails to adequately capture the full complement of epigenomic heterogeneity in tumor cells, and may mask deadly features present in less abundant glioma cells. Therefore, single-cell epigenomic resolution is needed to characterize the epigenetic intratumoral heterogeneity in glioma, detect rare cell populations, and to identify therapeutic vulnerabilities to prevent recurrence. To optimize our investigation of the glioma epigenome we initially performed single-cell Reduced Representation Bisulfite Sequening (scRRBS) on 134 patient-derived glioma sphere-forming cells. Our results highlight that the scRRBS assay is able to cover an average of 150,000 DNA methylation sites and is highly reproducible across biological replicates. Analyses of single-cell DNA methylation profiles from primary IDH-mutant tumors are now underway. Together, our study aims to generate a cellular hierarchy of primary IDH-mutant gliomas shaped by epigenetic programs that drive tumor growth.

← Back

CO-LOCALIZATION ANALYSES OF GENOMIC ELEMENTS: ESSENTIAL FACETS AND POTENTIAL PITFALLS

Chakravarthi Kanduri[1,2], Christoph Bock[3,4], Eivind Hovig[1,5,6], Geir Kjetil Sandve[1,2]

[1]University of Oslo, Department of Informatics, Oslo, Norway, [2]J CoDiRC, K. G. Jebsen Coeliac Disease Research Centre, Oslo, Norway, [3]CeMM, Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria, [4]University of Vienna, 4Department of Laboratory Medicine, Vienna, Austria, [5]Institute for Cancer Research, Department of Tumor Biology, Oslo, Norway, [6]The Norwegian Radium Hospital, Institute for Cancer Genetics and Informatics, Oslo, Norway

Assessment of the co-localization of genomic elements on a linear reference genome has become a routine methodological approach to understand the interaction between various functional genomic elements in biological processes. Although several statistical models and tools exist for this purpose, the breadth of choices that are to be made during the formulation, analyses, and interpretation of such investigations highly influence the conclusions that one could draw. Here, we survey the essential aspects of co-localization analyses and discuss the potential pitfalls involved in each step.

THE FORGE BIOINFORMATICS PIPELINE

Sam Khalouei[1,3], Karin Ng[1,3], Yue Jiang[1,3], Sergey Naumenko[1,3], Justin Foong[1,3], Mathieu Bourgey[2,3], Guillaume Bourque[2,3], Arun Ramani[1,3], Michael Brudno[1,3]

[1]Hospital for Sick Children, The Centre for Computational Medicine, Toronto, Canada, [2]McGill University & Genome Quebec Innovation Center (MUGQIC), McGill Genome Centre, Montreal, Canada, [3]Canadian Centre For Computational Genomics (C3G), Toronto, Montreal, Canada

There are a large variety of bioinformatics pipelines that process next generation sequencing data and users are usually faced with choosing many different parameters and filters that can significantly affect the outcome. Also most of these pipelines are custom-made scripts, which make the results reproducibility between institutions difficult. There have been some attempts in recent years to provide freely-accessible standardized sequencing pipelines such as bcbio-nextgen and GenAP pipelines. These platforms not only provide the means for the scientific community to access a common established platform but also allow them to report bugs and provide feedback for improvement. Furthermore, users can modify the pipeline towards their particular needs. Here we demonstrate the modification of a GenAP sequencing pipeline termed the Forge pipeline. The Forge pipeline was designed based on the merging and optimization of two existing pipelines and addition of custom-made scripts. Some of the main highlights of the Forge pipeline are performing combined genotyping while making use of individual sample parameters for filtering variants and also making use of an ensemble of five different variant callers. The Forge pipeline improves the elimination of false positive variants and also incorporates a Gemini annotation step that allows for filtering the variants based on different criteria such as quality, allele frequency and mode of inheritance.

← Back

A GENERAL WEB PLATFORM FOR INTEGRATING EXPERIMENTAL PROTEOMICS DATA AND RESULTS FROM GWAS AND EXOME SEQUENCING PROJECTS

April Kim[1,2], Edyta Malolepsza[1,2], Justin Lim[1,3], Kasper Lage[1,2]

[1]Broad Institute, Stanley Center, Cambridge, MA, [2]Massachusetts General Hospital, Department of Surgery, Boston, MA, [3]Massachusetts Institute of Technology, Cambridge, MA

The ongoing genomic revolution has identified thousands of common variant loci significantly associated with traits through genome-wide association studies (GWAS). In parallel, specific genes or protein-coding mutations have been linked to particular diseases through exome sequencing technologies. Recent advances in stem cell technologies and quantitative interaction proteomics methods has made the experimental interrogation of physical interactions of proteins in tissue- or cell-type-specific manner possible at scale. Integrating genetic datasets and experimental proteomics results can lead to new insight into the cellular processes implicated in disease. However, this integration is technically challenging. To address this issue, we have developed Genoppi, a general web platform that quality controls experimental proteomics data and provides robust statistical integration of proteomics results with the output from exome sequencing projects or GWAS. Genoppi is a user-friendly analytical framework for computational experts and non-experts alike. It allows the elucidation of unexpected biological connections between risk genes or loci in a particular trait based on custom proteomics datasets from a tissue- or cell-type of relevance to that trait. With more and more genetic and proteomics datasets becoming available to the community, we expect Genoppi to become an increasingly valuable tool for geneticists and other life scientists in the future.

← Back

MANUALLY CURATED 16S rRNA DATABASE AND ASSOCIATED SEAMLESS UPDATING PLATFORM

Seok-Won Kim, Todd D Taylor

RIKEN IMS, Laboratory for Integrated Bioinformatics, Yokohama, Japan

Data growth in DDBJ/EMBL-EBI/NCBI is rising exponentially due to the increase of novel bacteria isolation and metagenomic studies. To manipulate these data, the primary data, including its associated metadata and sequences, needs to be checked in the next update stage against its own secondary database. This may result in a bottleneck when updating such massive datasets. Here, we present a massive sequence tracking and management platform for solving this issue. We constructed a manually edited 16S ribosomal RNA (rRNA) gene database called GRD. In GRD, both the 5' regions and 3' regions, including the anti-SD sites, have been carefully checked and contaminating sequences have been removed. Because of this careful manual checking of the 16S rRNA sequences, our database can be considered the most reliable reference source for downstream analyses. In addition, we are including PCR-based sequences which are published in public primary databases. We developed this platform for continuous updating and maintenance of the sequences and taxonomy information in this database. In particular, recently changed taxonomic names are updated according to the NCBI Taxonomy database. As with the genomic-based sequences, we confirm all amplified sequences which have 5’ and 3’ regions by manual curation. Our platform can be applied not only for rRNA genes, but also for other marker genes.

← Back

LONG-NONCODING RNA BASED PROGNOSTIC SIGNATURE FOR GLIOMAS

Manjari Kiran[1], Daniel M Keenan[2], Anindya Dutta[1]

[1]University of Virginia School of Medicine, Biochemistry and Molecular Genetics, Charlottesville, VA, [2]University of Virginia School of Medicine, Statistics, Charlottesville, VA

Low-grade glioma (LGG) develops in the supporting glial cells of brain. A more aggressive form of brain cancer is known as glioblastoma multiforme (GBM). With 5-year relative survival rate of 43% for LGG and 6% for GBM, it is estimated that more than 16,000 adults in the USA will die from primary brain cancer this year. The poor survival rate of LGGs indicates that there is a need for a better prognostic marker that will enable the physician to give more aggressive therapy at the outset, even for LGGs. Long non-coding RNAs are gaining widespread attention as a potential biomarker for cancer diagnosis and prognosis. In our previous study, we determined the expression of many known and novel lncRNAs in over 750 brain cancer and normal brain tissue RNA-seq dataset from the Cancer Genome Atlas (TCGA) and other publicly available dataset. We reported over a thousand lncRNAs induced or repressed in glial tumors relative to normal brains (Reon et al, PLoS Medicine 2016).

We have now developed a computational model for prognosis of low-grade gliomas by combining lncRNA expression, Cox regression and L1-LASSO penalized method.We trained the model on a set of 267 patients with gene expression data and clinical information and identified 16 lncRNAs that could act as a prognostic indicator independent of grade, patient age or IDH1/2 mutational status to predict the overall survival of patients. Patients in the training set, as well as a separate validation set, could be dichotomized according to their risk score calculated by adding the lncRNA expression and the coefficient obtained by multivariate cox regression done for the 16 lncRNA (Hazard Ratio (HR)=6.83, p=7e-10 for training set (n=267), HR=4.66, p=7.8e-05 for validation set (n=179)). To test whether this model is useful on a completely different cohort of patients, gene expression data from the Chinese Glioma Genome Atlas (CGGA) for 274 Chinese LGG and GBM patients was used. We can also cluster this new set of gliomas patients into two clusters with the same method established from the TCGA training set. The patients in low risk score group have significantly higher survival than the patients in high-risk group (HR=1.58, p=0.0073). Finally, we applied the method to GBMs in TCGA and could distinguish two groups with different prognosis (HR=1.54, p=0.03, n=143). Thus, there is a clear prognostic signature based on expression level of 16 lncRNA that can be applied to diverse populations of glioma and GBM patients. Besides their use as a biomarker, these lncRNAs need to be studied in detail to determine how they are affecting patient outcome.

← Back

DIPLOID GENOME ASSEMBLIES AT NCBI

Paul A Kitts, Karen Clark, Avi Kimchi, Terence D Murphy, Françoise Thibaud-Nissen, Valerie A Schneider

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

Older sequencing and assembly technologies would collapse the two haplotypes from a diploid organism into an assembly that was a composite haploid representation of the genome. Sequencing and assembly technologies have now advanced to the point that it is possible to produce genome assemblies that have multi-megabase long stretches of haplotype-resolved (phased) sequence, in which both haplotypes of a diploid genome are represented.

The two leading methods for generating long stretches of phased sequence are Pacific Biosciences' Single Molecule, Real-Time (SMRT) long sequence reads assembled using Pacific Biosciences' FALCON-Unzip software[1] and 10X Genomics' Linked-Reads assembled using 10X Genomics' Supernova software[2].

Several groups have submitted diploid genome assemblies generated using these technologies to GenBank within the last year. We expect to receive increasing numbers of diploid genome assemblies as these new sequencing and assembly technologies become more widely adopted.

NCBI manages diploid genome assemblies using an assembly model that places the two haplotypes into different assembly-units. This is an extension of the *haploid-with-alt-loci* assembly model that was originally developed to accommodate the alternate loci in the human reference assembly produced by the Genome Reference Consortium (GRC).

Because diploid assemblies contain two representations for much of the genome, they pose a number of challenges as to how best to display the assemblies, annotate the assemblies and make the genome data available to users. NCBI's approach to these challenges will be presented.

1. Chin CS *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods.* **13**, 1050-1054
2. Weisenfeld NI *et al.* (2017) Direct determination of diploid genome sequences. *Genome Res.* **27**, 757-767

THE ROLE OF REARRANGEMENTS IN THE DIVERGENT EVOLUTION OF EUROPEAN GREEN LIZARDS

Sree Rohit Raj Kolora[1,4,2], Anne Weigert[7], Henrike Indrischek[1], Amin Saffari[1], Steffanie Kehr[1], Rui Faria[6], Klaus Henle[3], Katja Nowick[5], Peter Stadler[1], Martin Schlegel[2]

[1]University of Leipzig, Bioinformatics, Leipzig, Germany, [2]University of Leipzig, Biology, Leipzig, Germany, [3]Helmholtz Centre for Environmental Research, Conservation Biology, Leipzig, Germany, [4]German center for Integrative Biodiversity Research, Flexpool, Berlin, Germany, [5]Frier University, Biology, Chemistry, and Pharmacy, Berlin, Germany, [6]University of Sheffield, Animal and Plant Sciences, Sheffield, United Kingdom, [7]Max Planck Institute for Evolutionary Anthropology, Genetics, Leipzig, Germany

A comparative genomic study of closely related lineages enables us to understand lineage-specific evolution at high resolution. *Lacerta viridis* (Eastern green lizard) and *Lacerta bilineata* (Western green lizards) are two parapatric lineages of the *Lacerta viridis* complex which diverged in the Pleistocene. These lineages have a well-studied phylogeographic history complemented by breeding experiments involving inter-specific crosses, thus providing a good model system to study evolutionary divergence. We generated high quality genome assemblies for each lineage, with hybrid approaches using Illumina and PacBio sequencing data. In addition to this, we assembled transcripts from different tissues for annotation and selection analysis. These high quality genomes provided the resolution to study the heterogametic W-chromosome of the female lizards that could contribute to hybrid sterility according to Haldane's rule. Through the comparison of the orthologous coding sequences from transcripts, we estimated the divergence between the lacertids. We observed a higher mutation rate in sex-chromosomes compared to autosomes and high levels of CpG densities on the W-chromosome indicating accelerated evolution that could have lead to changes in chromosomal stability. The genomic rearrangements between the lacertids were detected through whole-genome alignments and read-based methods and these regions were tested for signals of positive selection. The prediction of conserved-sites which included multi-genome comparisons with *Anolis carolinensis* and other well-studied vertebrate genomes revealed patterns of selection specific to the lacertid lineage. Repeat elements such as satellite DNA were associated with genes involved in nucleic acid binding and mRNA splicing. These genes were also observed to have differences in selection patterns between lacertids, which could affect patterns of transcriptional regulation. This study helps us understand the role of rearrangements in different genomic features towards the divergence of lacertids.

← Back

ANALYSIS OF MENDELIAN INHERITANCE ERRORS IN DEEP SEQUENCED WHOLE GENOMES FROM 1314 TRIOS IDENTIFIES POPULATION-SPECIFIC STRUCTURAL VARIANTS

Prachi Kothiyal, Wendy Wong, Dale Bodian, John Deeken, John Niederhuber

Inova Health System, Inova Translational Medicine Institute, Fairfax, VA

Trio-based whole genome sequencing (WGS) data can contribute significantly towards the development of quality control methods that can be applied to non-family WGS. Mendelian inheritance errors (MIEs) in parent-offspring trios are commonly attributed to erroneous sequencing calls, as the rate of true de novo mutations is extremely low compared to the incidence of MIEs. Here, we analyzed WGS data from 1314 trios across diverse human populations with the goal of studying the characteristics and distribution of MIEs. Our results indicate that MIE density increases with repeat density where Short Interspersed Nuclear Elements (SINEs) show the strongest correlation. We applied filters based on genotype call quality and observed ~86% reduction in total number of MIEs in the cohort. Filtering has a greater impact on frequent errors where >100 of the 1314 trios have MIE at any given variant site when compared to unique MIEs with a Mendelian violation in a single trio at a site. We further analyzed the distribution of unique MIEs and observed that they could be enriched in regions overlapping putative chromosomal aberrations such as copy number variations (CNVs) and structural variants (SVs). Therefore, in certain cases the errors can be useful for detecting underlying genomic anomalies. We created population-specific MIE profiles and discovered regions that present different error distributions across populations. Finally, we have created population-specific and overall MIE tracks that can be loaded in UCSC Genome Browser. These profiles can be used for variant filtering, annotating candidate de novo mutations, discovering putative SVs, and for distinguishing between regions that have errors due to sequence quality vs. chromosomal anomalies.

← Back

UNCALLED: AN ALIGNER FOR QUICKLY MAPPING RAW NANOPORE SIGNALS TO LARGE REFERENCES

Sam Kovaka[1], Taher Mun[1], Yunfan Fan[2], Michael C Schatz[1,3]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [3]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Nanopore sequencing works by measuring ionic current levels as a DNA strand passes through a pore. These signals are then segmented into “events”, each of which is meant to represent a single k-mer of the full read. After segmentation, the sequences of events are typically basecalled using either a neural network or a hidden Markov model. This is computationally expensive, and most methods require the full read to be sequenced before any bases can be called. This is especially problematic for Oxford Nanopore’s “read until” method, where the sequencing of a read can be stopped in real-time if it is decided the read is not of interest. One way a read can be deemed “of interest” is by seeing if it aligns to a certain reference genome, a process which until now required a fully basecalled read.

Here we present UNCALLED (Utility for Nanopore Current ALignment to Large Expanses of DNA), an aligner which can map nanopore reads to a reference genome without basecalling first. We do this by first translating every 6-mer in the reference genome into the expected nanopore signal, based on the 6-mer models released by Oxford Nanopore. We then create an FM index of this translated reference, where the alphabet for the index consists of the 4096 possible 6-mers. We align short segments of a read (seeds) using a modified version of the standard FM index mapping algorithm, where Nanopore events can match multiple possible 6-mers and therefore must branch to find every possible alignment. This algorithm can also handle consecutive events that represent the same nucleotides on the read, which occur due to errors in segmentation. Event alignments from one seed are recorded so that the next seed can save time by not re-aligning those events to the same locations. After all seeds are aligned, the reference locations with support from the most seeds are found and chosen as potential locations for the full read. Depending on the application, these potential locations could be further refined through dynamic time warping alignment, or used to decide to stop sequencing in a read-until sequencing run.

← Back

SEMANTIC ANNOTATION AND KNOWLEDGE EXTRACTION USING iCLiKVAL

Naveen Kumar, Todd D Taylor

RIKEN Center for Integrative Medical Sciences, Laboratory for Integrated Bioinformatics, Yokohama, Japan

There is a myriad of information hidden in scientific media in the forms of texts, images, audios, videos, datasets and various kinds of software but it is often not discoverable due to absence of structured and semantic annotations.

iCLiKVAL is a web-based application (http://iclikval.riken.jp) that collects annotations for scientific media found online by a growing community using crowdsourcing. The application is built upon a freely accessible and secure API to save semantic as well as free-form but structured annotations as “key-value” pairs with optional “relationships” between them. The philosophy behind iCLiKVAL is to identify the online media by a unique URI (Uniform Resource Identifier) and attach highly structured semantic annotations to various concepts related to the media to identify and mark occurrences of ontological entities and relationships. These annotations are stored in human-readable as well as machine-readable formats which helps computers to easily index and interpret information and allows for much more sophisticated data searches and knowledge extraction and discovery by linking with other heterogeneous data sources.

We hope this application simplifies the process for users to securely and conveniently submit their valuable annotations to iCLiKVAL and to facilitate knowledge extraction for the scientific community.

← Back

INTEGRATING DATA, TOOLS AND KNOWLEDGE TO ACCELERATE SCIENTIFIC DISCOVERY USING OPEN-SOURCE, BIOLOGICAL, DATA-SCIENCE PLATFORM OF KBASE

Vivek Kumar[1], Sunita Kumari[1], Jim Thomason[1], Doreen Ware[1,2], Priya Ranjan[3], Sean McCorkle[4], Shinjae Yoo[4], Nomi Harris[5], Bob Cottingham[3], Christopher Henry[6], Adam Arkin[5]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]USDA ARS NEA, Ithaca, NY, [3]Oak Ridge National Laboratory, Oak Ridge, TN, [4]Brookhaven National Laboratory, Upton, NY, [5]Lawrence Berkeley National Laboratory, Berkeley, CA, [6]Argonne National Laboratory, Argonne, IL

The U.S. Department of Energy Systems Biology Knowledgebase (KBase, http://kbase.us) is an open-source software and data platform designed to meet the grand challenge of systems biology: predicting and ultimately designing biological function. KBase supports the sharing and integration of reference and experimental data with analysis tools that enable researchers to design computational experiments, test hypotheses, and share findings that can be reproduced and extended by other researchers.

KBase has a socially-oriented user interface that supports a persistent and provenanced online work environment. Currently KBase provides support for genome assembly and annotation, transcriptomics, metabolic modeling, and comparative genomics. KBase enables users to upload their own data or access public data. KBase services are available from within an interactive, Jupyter-based user interface that supports the creation of dynamic workflow documents called Narratives that enable experimental and computational biologists to work together to share and publish data, approaches, workflows, and conclusions, leading to transparent and reproducible computational experiments. The Narratives can be kept private, shared with collaborators, or made public for the benefit of the wider research community. This presentation will include some Narratives that users can start with.

Visit http://kbase.us/ to learn how KBase might be useful in your research.

HYBRID ASSEMBLY OF SMALL GENOMES IN GALAXY

Delphine Lariviere, The Galaxy Team

Penn State University, http://galaxyproject.org, University Park, PA

Continuously dropping cost of massively parallel sequencing (NGS) combined with the evolution of long read technologies makes genome sequencing possible for small labs and individual research groups. Yet, as is the case with other applications of NGS, generating data is relatively straightforward while the analysis remains challenging. This is especially true of genome assembly.

Two currently available technologies are the most suitable for small genome sequencing. Illumina's technology offers high coverage and accuracy at relatively low cost but can at most generate reads 300 bp in length. On the other hand, Oxford Nanopore's molecular ratcheting through nanopore technology (ONT) generates multi-kilobase reads.

Combining both technologies amplifies their relative strengths and enables producing complete, high quality assemblies. This combination analysis, called hybrid assembly, has been made available in Galaxy. It includes the following components. The quality control of reads is performed using FastQC/multiQC that provides a set of metrics on reads quality. The genome assembly is performed with Unicycler, a pipeline based on Spades and Pilons tools and dedicated to small genome assembly. The quality of the assembly is evaluated with Quast, and the annotation is performed through Prokka. Result can be visualized on a local instance of IGV or Galaxy's own Trackster browser.

To show the utility of this pipeline we have performed sequencing of E. coli C - a previously unsequenced strain frequently used in experimental evolution experiments - using Illumina and ONT approaches. We show the results of comparative analysis of this isolate against the K12 genome. Our pipeline makes it possible to perform complete assembly and annotation analysis on the web without the needs for any additional software of hardware resources.

← Back

LEARNING VARIABLE GAPPED SEQUENCE-STRUCTURE MOTIFS FOR RNA-BINDING PROTEINS

Kaitlin U Laverty[1], Shankar Vembu[2], Arttu Jolma[2,3], Jussi Taipale[3,4], Timothy R Hughes[1,2], Quaid D Morris[1,2,5,6]

[1]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [2]University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada, [3]Karolinska Institutet, Department of Medical Biochemistry and Biophysics, Stockholm, Sweden, [4]University of Helsinki, Genome-Scale Biology Program, Helsinki, Finland, [5]University of Toronto, Department of Computer Science, Toronto, Canada, [6]University of Toronto, Department of Electrical and Computer Engineering, Toronto, Canada

RNA-binding proteins (RBPs) display binding specificity based on both the primary nucleotide sequence and the secondary structure of an RNA transcript. The recently published method "RNAcompete-S" combines an *in vitro* competitive binding reaction with a computational pipeline to simultaneously determine the sequence and structural preference of an RBP. The method is capable of detecting long (>12 base) Sequence Structure Models (SSMs). Each SSM is composed of two position weight matrices, representing sequence and structure, respectively, that can be used to scan the transcriptome to discover potential binding sites. However, this computational method is not amenable to alternative sources of experimental data and is limited to discovering ungapped SSMs.

To overcome these constraints, a new version of the method was developed with the addition of a Hidden Markov Model (HMM) to learn the spatial relationships between short (<8 base) SSMs. These associations can be used to assemble multiple gapped (or ungapped) motifs for a single RBP. We tested our method on the RBP La-related protein 6 (LARP6). LARP6 is thought to recognize primary sequence motifs on both sides of an internal loop structure of variable size and nucleotide composition in the 5’UTR of collagen transcripts. Using high-throughput RNA-SELEX data for LARP6 as input, our HMM-based pipeline was able to closely recapitulate sequence and structural binding preferences that were characterized by a gel mobility shift assay. The improved RNAcompete-S method can now be applied to open source experimental data to discover RBP motifs *de novo*.

← Back

DISCOVERY OF MEDIUM AND LONG-SIZED INSERTION VARIANTS WITH ACCURATE BREAK POINTS AND FLANKING SEQUENCES

Young Gun[1,2], Jin-young Lee[3], Young-joon Kim[2,3]

[1]Yonsei University, Severance Hospital, Seoul, South Korea, [2]Yonsei University, Integrated Omics For Biomedical Science, Seoul, South Korea, [3]Yonsei University, Biochemistry, Seoul, South Korea

Structural variations(SVs) are defined as genomic rearrangements greater than 50bp in length, distinguished from single nucleotide variants(SNVs) and shorter insertion/deletion variants(indel). The comprehensive discovery of SVs could expand our understanding of the human genome, since the SVs have a substantial impact on the human diseases and the human genetic and phenotypic diversity.

Previous SV detection algorithms based on short read whole-genome sequencing(WGS) data have been developed based on read-depth, paired-read, split-read, and assembly approaches and shows reasonable performances on small indels and large deletions. Nonetheless, the discovery of medium and long-sized insertions has been elusive except for mobile element insertions(MEIs), due to the limited nature of short reads and the incomplete haplotype representation of the reference genome.

To overcome these limitations, we focused on the determination of accurate break points of SVs and the retrieval of the longest possible sequences around the called break points, instead of the full-length insertion events. We extracted paired reads with one end properly mapped and with the other end unmapped, soft-clipped outwards, mapped at distant loci or mapped in improper directions, which we defined as one-end anchored read pairs. Then we clustered and locally assembled those reads and remapped the contigs to nearby reference sequences to accurately map break points.

We compared the performance of our approach with other insertion discovery tools (ANISE, MindTheGap, and PopIns) and showed that out tool has better specificity (>95%) and comparable sensitivity (~85%) in the discovery of novel and non-unique sequence insertions, using simulation datasets. Furthermore, out tool showed more accuracy in the discovery of insertion break points.

← Back

LANDSCAPE OF SOMATIC MUTATIONS IN INFLAMMATORY BREAST CANCER WHOLE-GENOME SEQUENCES

Xiaotong Li[1,2], Savitri Krishnamurthy[3], Sushant Kumar[2], Arif Harmanci[2], Shantao Li[2], Robert Kitchen[2], Vikram B Wali[1], Yan Zhang[2], Sangeetha Reddy[3], Wendy Woodward[3], James Reuben[3], Jeffery Chuang[4], Christos Hatzis[1], Naoto T Ueno[3], Mark Gerstein[2], Lajos Pusztai[1]

[1]Yale Cancer Center, Yale School of Medicine, New Haven, CT, [2]Yale University, Computational Biology and Bioinformatics, New Hvaen, CT, [3]The University of Texas MD Anderson Cancer Center, Morgan Welch Inflammatory Breast Cancer Research Program, Houston, TX, [4]The Jackson Laboratory, Genomic Medicine, Farmington, CT

Goal: Inflammatory breast cancer (IBC) is a rare, aggressive form of breast cancer that is characterized by a highly metastatic phenotype. Numerous previous attempts failed to identify, recurrent, IBC-specific gene expression or DNA copy number alterations. We performed whole genome sequencing (WGS) of IBC biopsies obtained before any therapy to define a comprehensive genomic landscape of this disease.

Methods: Illumina paired-end whole genome sequencing (WGS) of 20 IBC (n=9 ER+, n=11 ER-) and matched normal samples were performed with median coverage of 60X and 40X for cancer and normal, the percentages of mapped reads were 99.3% and 99.2%, respectively. We identified germ-line and somatic variants, indels as well as large scale structural variants, using GATK Haplotype Caller, MuTect and Meerkat, respectively. We performed the same analysis on WGS data from 23, age, race and ER and HER2 matched, non-IBC (n=12 ER+, n=11 ER-) from the TCGA for comparison. Variants in both coding and noncoding sequences were categorized by FunSeq to identify potential drivers. DeconstructSigs were used to decompose the mutational spectrum of each cancer into 30 validated, mutational signatures provided by COSMIC.

Results: We identified 118,818 somatic variants in the IBC samples (median: 3,856; minimum: 1,109; maximum: 24,815) including 1,060 variants (~0.9%) in coding regions. 5,287 somatic indels and 5,959 large scale structural variants were detected including 1,028 insertions and 1,857 deletions. Recurrent, non-synonymous mutations were detected in the coding region of *GRIN2A* gene in 3/20 IBC samples (15%), (previously reported as a potential driver mutation in 1.7% of breast cancers). Other significant mutations in coding regions included *GRHL1*, *PIK3R2*, *ESR1*, *FLG2* and etc. In non-coding regions, recurrent and deleterious mutations were identified in *MAST2* gene in 4/20 IBC samples (20%) vs. 0/23 non-IBC samples. Contributions of mutational signature 9, that is associated with polymerase η, were significantly higher in IBC cohort than non-IBC cohort (p-value=0.056).

Conclusion: This is the first whole genome sequencing analysis of IBC and comparison with the results from non-IBC. We identified promising candidate drivers in the coding sequence and in non-coding regulatory modules of expressed genes. We also identified mutational signature 9, and mutations in several DHS as significantly more frequent alterations in IBC compared to non-IBC.

← Back

A HUMAN-SPECIFIC SWITCH OF ALTERNATIVELY SPLICED *AFMID* ISOFORMS CONTRIBUTES TO *TP53* MUTATIONS AND TUMOR RECURRENCE IN HEPATOCELLULAR CARCINOMA

Kuan-Ting Lin[1], Wai-Kit Ma[1], Juergen Scharner[1], Yun-Ru Liu[2], Adrian R Krainer[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Taipei Medical University, Joint Biobank, Office of Human Research, Taipei, Taiwan

Pre-mRNA splicing can contribute to the switch of cell identity that occurs in carcinogenesis. Here we analyze a large collection of RNA-Seq datasets and report that splicing changes in hepatocyte-specific enzymes, such as AFMID and KHK, are associated with HCC patients' survival and relapse. The switch of *AFMID* isoforms is an early event in HCC development, and is associated with driver mutations in *TP53* and *ARID1A*. Finally, we show that the switch of *AFMID* isoforms is human-specific and not detectable in other species, including primates. The integrative analysis uncovers a mechanistic link between splicing switches, *de novo* NAD$^+$ biosynthesis, driver mutations, and HCC recurrence.

← Back

EVIDENCES FOR THE ROLE OF ZBTB33 (KAISO) IN HETEROCHROMATIN PRIMING

Xiaoxuan Lin, Khadija Rebbani, Sudhakar Jha, Touati Benoukraf

National University of Singapore, Cancer Science Institute of Singapore, Singapore

ZBTB33, also known as Kaiso, is a member of zinc finger and BTB/POZ family. In contrast to many transcription factors, ZBTB33 has the ability to bind both methylated and unmethylated DNA. In this study, we aim to investigate the role of ZBTB33 as a methylated DNA binding factor. We took advantage of the latest releases of the ENCODE sequencing datasets, including ZBTB33 ChIPseq, Whole Genone Bisulfte Sequencing (WGBS), histone marks ChIPseq and sequencing assays determining the chromatin states, to characterize the chromatin landscapes surrounding ZBTB33 methylated binding sites. Interestingly, our integrative analysis brought to light that majority of ZBTB33 methylated binding sites are located in heterochromatin carrying a specific histone post-translational modification feature, with significant enrichment of mono-methylation at lysine 4 of histone 3 (H3K4me1) and total absence of repressive histone marks. These observations suggest that ZBTB33 has the unique ability to bind closed and methylated DNA loci, suggesting a role in priming heterochromatin.

← Back

NOVEL EXON DISCOVERY IN CELLULAR DIFFERENTIATION AND HUMAN DISEASE BY UTILIZING THE SNAPTRON FRAMEWORK

Jonathan Ling[1], Christopher Wilks[2,3], Abhinav Nellore[4,5], Ben Langmead[2,3]

[1]Johns Hopkins University, Neuroscience, Baltimore, MD, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [4]Oregon Health & Science University, Biomedical Engineering, Portland, OR, [5]Oregon Health & Science University, Surgery, Portland, OR

De novo identification of novel transcripts is an exceptionally challenging task and researchers commonly rely on annotated transcript databases to quantify expression or alternative splicing. However, unannotated splicing events can be crucial to understanding disease and discovering new therapies. As an example, we recently developed a method for identifying novel and unannotated cryptic exons that are linked to neurodegeneration (*1*-*3*) and neuronal differentiation (*4*). However, this method requires extensive manual annotation and is difficult to scale across many samples.

Motivated by the vast amount of splicing data available in public, archived RNA sequencing datasets, we have extended the Snaptron software and web service (*5*) to enable rapid, large-scale screens for tissue and cell type-specific splicing patterns. Snaptron is the query-answering portion of a larger search engine (*5*-*8*) for splice junctions observed in tens of thousands of RNA-seq samples from the Sequence Read Archive and other large projects such as GTEx and TCGA. Using this framework, we have identified hundreds of highly incorporated, previously unannotated, cell type-specific exons and the splicing factors that regulate these exons. Snaptron has also allowed us to screen cryptic exons found in human disease (*1*-*3*) across all published datasets to identify surprising insights into etiology.

Finally, we demonstrate an intuitive web interface for visualizing a query exon's “percent spliced in” frequency across various datasets of choice (cell types, tissues, cancer subgroups, gene knockdowns, etc.). Snaptron provides a framework that allows for extremely versatile queries and enables researchers to leverage vast datasets that would otherwise be too difficult to obtain or too computationally unwieldy to analyze from scratch. We hope that this ability to cross reference all published datasets will accelerate interdisciplinary approaches in ways that have yet to be conceived.

1. Ling JP et al, Science (2015) PMID 26250685
2. Jeong YH et al, Mol Neuro (2017) PMID 28153034
3. Sun M et al, Acta Neuropath (2017) PMID 28332094
4. Ling JP et al, Cell Rep (2016) PMID 27681424
5. Wilks C et al, bioRxiv (2017) doi: 10.1101/097881
6. Nellore A et al, Bioinformatics (2016) PMID 27592709
7. Nellore A et al, Bioinformatics (2016) PMID 27153614
8. Collado-Torres L et al, Nat Biotech (2017) PMID 28398307

← Back

INVESTIGATING THE ASSOCIATION BETWEEN POLYGENIC RISK SCORE AND CORONARY ARTERY CALCIFICATION

Zhi Liu[1], Leslie G Biesecker[2], Nancy F Hansen[1], James C Mullikin[1]

[1]Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [2]Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

Coronary artery disease (CAD) is the leading cause of mortality world-wide. Advanced CAD usually includes plaque calcification, which can be quantified by coronary artery calcification (CAC) scoring. High CAC scores (>300) are strong predictors of future cardiac events. Previous genome-wide association studies (GWAS) have identified over 60 single nucleotide polymorphisms (SNPs) that are associated with CAD. However, many of these SNPs have relatively small effects by themselves. To evaluate cumulative effects of these SNPs and their contributions to CAC, a polygenic risk score based on these SNPs is being constructed to investigate its association with CAC. The ClinSeq® cohort, which includes 1,014 participants of predominantly European ancestry, is being utilized. Exome sequencing was performed on gDNA samples from all participants. The CAC score for each participant was calculated using the Agatston method. Due to a considerable proportion of participants with a CAC score of zero, the value of one was added to all CAC scores and log transformed, before adjusting for appropriate covariates, such as age, sex, BMI, and cholesterol medication usage. The polygenic risk score will be calculated using R package GenABEL by summing weighted risk alleles from previously published SNPs. In addition, we will further evaluate the constructed risk score and its ability to predict more severe cardiac phenotypes, such as coronary artery stent placement. By constructing a polygenic risk score and testing its association with CAC, we are able to evaluate contributions of multiple genetic factors to CAD related phenotypes.

← Back

EFFICIENT DETECTION OF HIGHLY MUTATED REGIONS WITH MUTATIONS OVERBURDENING ANNOTATIONS TOOL (MOAT)

Lucas Lochovsky[1,2], Jing Zhang[1,2], Mark Gerstein[1,2,3]

[1]Yale University, Program in Computational Biology and Bioinformatics, New Haven, CT, [2]Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT, [3]Yale University, Department of Computer Science, New Haven, CT

Identifying genomic regions with a higher-than-expected mutation burden has a number of useful applications. Regions overburdened with cancer somatic variants may be associated with cancer progression. Accumulations of germline rare variants could be indicators of positive selection, or precursors for genetic disease. We introduce a new software tool, called MOAT (Mutations Overburdening Annotations Tool), to perform mutation burden analysis with great speed. MOAT makes no assumptions about the mutation process, except that the background mutation rate (BMR) changes smoothly with other genomic features. This nonparametric scheme randomly permutes the variants (or target regions) on a relatively large scale where the BMR is assumed to be constant to provide robust burden analysis in cancer driver detection. Furthermore, the MOAT software suite incorporates a somatic variant simulator called MOATsim, which randomly permutes the input variants with effective covariate control. MOAT also offers the option to evaluate the functional impact within annotations for burden analysis. In conclusion, MOAT is a useful for a broad range of analyses that would benefit from variant permutation. MOAT is available at moat.gersteinlab.org

← Back

SCIAPPS: A CLOUD-BASED PLATFORM FOR REPRODUCIBLE BIOINFORMATICS WORKFLOWS

Zhenyuan Lu[1], Liya Wang[1], Peter Van Buren[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY

There are increasing needs to store and analyze data on distributed storage and computing systems imposed by the rapid growth of both sequence and phenotype data generated by high-throughput methods. A workflow management system is needed to ensure efficient data management across heterogeneous systems, simplify the task of analysis through automation, and make large scale bioinformatics analysis accessible and reproducible. To address these needs, we have developed SciApps (https://www.sciapps.org), a cloud-based platform for reproducible bioinformatics workflows. The platform is powered by a federated CyVerse system located at Cold Spring Harbor Laboratory (CSHL), XSEDE clusters, and Amazon EC2 services. The system is fully integrated with CyVerse Cyber infrastructure (CI) through the Agave platform for job management and iRODS-based CyVerse Data Store for data management. To create a workflow, each analysis job is submitted, recorded, and accessed through the web portal. Part or all of a series of recorded jobs can be saved as reproducible, sharable workflows for future execution using the original or modified inputs and parameters. The platform is designed to automate the execution of modular Agave apps and make it easy to bring reproducible workflows to both local and cloud-based computing systems. Two workflows, association and annotation, are provided as exemplar scientific use cases.

This work is supported in part by NSF grants DBI-1265383 (for the CyVerse project) and IOS-1445025 (for the MaizeCode project).

← Back

HUMAN MICROBIOME PROJECT (HMP) DATA RESOURCE: A WEB PORTAL FOR EXPLORING AND ACCESSING HMP DATA, ANALYSIS PRODUCTS, AND TOOLS

Anup Mahurkar[1], Heather Huot Creasy[1], Victor Felix[1], James Matsumura[1], Jonathan Crabtree[1], Arjun Kumar[1], Mike Schor[1], Lance Nickel[1], Justin Wagner[2,3,4], Michelle Giglio[1], Owen White[1]

[1]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [2]University of Maryland, Department of Computer Science, College Park, MD, [3]University of Maryland, Center for Bioinformatics and Computational Biology, College Park, MD, [4]University of Maryland Institute for Advanced Computer Studies, College Park, MD

We present here the Human Microbiome Project Data Resource, a central repository for querying and accessing multi-omic data and analysis products generated by the NIH Common Fund supported Human Microbiome Project (HMP). As the Data Analysis and Coordination Center for the Human Microbiome Project (HMP), and the Data Coordination Center (DCC) for the Integrative Human Microbiome Project (iHMP), we have collected and analyzed over 29,000 16S metagenomes, 4,000 whole metagenomes, 3,000 reference genomes, 900 metatranscriptomes, 400 host transcriptomes, 70 host genomes, 700 metabolomes, and over 500 other ‘omic datasets. In many cases we have downstream analysis products derived from the raw files including metagenome assemblies, annotations, abundance profiles, and functional profiles. Overall, we host over 15 TB of data comprised of 80,000 files generated from 30,000 samples. Much of this data is decorated with rich metadata that was collected during the course of this program which includes the Healthy Human Subjects cohort, as well as disease-focused cohorts including Type 2 Diabetes, Inflammatory Bowel Disease, and Pregnancy & Preterm Birth.

To host and maintain this data resource we have worked closely with data managers from each of the four iHMP institutions to develop a comprehensive data storage schema based upon the Open Science Data Framework (OSDF). OSDF serves as the backbone for our HMP Data Portal, a data exploration and analysis tool derived from the Genomic Data Commons data portal. This portal allows intuitive and interactive search functionality, as well as the ability to build custom cohorts/data sets for download or direct incorporation into analysis tools such as Metaviz. We have also built an API for users to upload and download the data from the repository. All schemas, associated API and submission tools are publicly available at https://github.com/ihmpdcc. The HMP Data Portal, as well as project details, SOPs and tools, are available at the comprehensive HMP/iHMP DCC website, www.hmpdacc.org.

← Back

MUMMER4: A FAST AND VERSATILE GENOME ALIGNMENT SYSTEM

Guillaume Marçais[1,3], Arthur L Deltcher[2], Adam M Phillipy[5], Rachel Coston[2], Steven L Salzberg[2,4], Aleksey Zimin[1,2]

[1]University of Maryland, Institute for Physical Science and Technology, College Park, MD, [2]Johns Hopkins School of Medicine, Center for Computational Biology, Baltimore, MD, [3]Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA, [4]Johns Hopkins University, Departments of Biomedical Engineering, Computer Science, and Biostatistics, Baltimore, MD, [5]National Institutes of Health, National Human Genome Research Institute, Bethesda, MD

We introduce the 4th version of the MUMmer sequence alignment program. The MUMmer3 sequence alignment software package was released in 2004. Beside its age and its limitations, the nucmer program of MUMmer remains one of the most used alignment programs, in particular to align genome to genome. MUMmer4 is a significant upgrade to the MUMmer3 package. While it is fully backward compatible, it lifts the limit on the input sequences length, and it effectively uses the multi-core architecture of modern machines. MUMmer4 is a very versatile aligner. It is able to align genome against genome or sequencing reads (short and long) against genome. It is competitive in accuracy and speed with aligners for next-generation sequencing data.

In addition, new programming interfaces in C++ and scripting languages (Perl, Python, Ruby) make MUMmer4 easier to incorporate in other bioinformatics tools.

MUMmer4 is available as an open source software package at https://mummer4.github.io/.

RESOLVING THE FULL SPECTRUM OF HUMAN GENETIC VARIATION USING LINKED-READS

Álvaro Martínez Barrio, Stephen R Williams, Andrew Wei Xu, Sarah Garcia, Claudia Catalanotti, Nikka Keivanfar, Jill Herschleb, Michael Schnall-Levin, Patrick Marks, Deanna M Church

10x Genomics, Inc., R&D, Pleasanton, CA

Highly accurate and affordable, standard short-read based methods fail to give a complete picture of a genome and are limited by the lack of long-range information. They cannot reliably detect many clinically important variant types in repetitive regions of the genome, copy neutral structural changes, and single exon deletions/duplications. To address variant detection shortcomings, we developed a technology that retains long-range information while maintaining the power, accuracy, and scalability of short read sequencing.
The 10x Genomics Chromium^{TM} Genome solution utilizes haplotype-level dilution of high molecular weight DNA molecules into >1 million barcoded partitions to create a novel data type referred to as 'Linked-Reads' (LR). This approach enables high-resolution genome analysis with minimal DNA input (~1 ng).
Our reference-based pipeline, Long Ranger^{TM}, leverages the unique properties of LR to both improve alignment quality of reads across more of the genome and to call a broader range of variant types. Linking heterozygous variants to distinct barcodes enables haplotype reconstruction, providing increased power for variant calling across all variant types. Lastly, algorithms assessing both barcode coverage as well as unexpected barcode overlap allow for robust identification of large, complex structural variants.
To test these approaches, we first compared NA12878 genome (lrWGS) and exome (lrWES) LR libraries to NA12878 genome (WGS) and exome (WES) standard short read libraries to assess small variant calling. We estimate that we can recover 20-40 Mb of previously inaccessible sequence, including sequence in clinically relevant genes such as CYP2D6, PMS2, SMN1, and STRC with lrWGS. An additional 74,667 novel SNPs are called in degenerate regions using lrWGS versus only 32,295 for WGS.
Single exon deletions and duplications are challenging to call using standard WES, particularly in the heterozygous state. The barcode information allows us to detect duplications robustly when phasing is present over the affected region. For NA12878 lrWES, ~50 exon-level deletions are called with a heterozygous sensitivity of 72.7% (PPV= 72%) and a homozygous sensitivity of 83.3% (PPV= 100%) without the requirement of large control populations.
We next performed 30x lrWGS sequencing on a set of 24 samples with known balanced or unbalanced SVs from either the GetRm CNVPanel (unbalanced) or the Coriell Cell Repository (balanced) with multiple assays confirming the presence of the variant. 22/24 variants were identified, and another known event was called as a candidate. With current algorithms, CNVs can be called with as little as 1-2x sequencing depth while balanced events require on the order of 10x coverage. We show that there is specific signal in the data for balanced events down to ~2x coverage.

← Back

EXPLORING THE CHEMISTRY AND BIOLOGY OF NUCLEOTIDE MODIFICATIONS IN MAMMALS, PARASITES AND DISEASE

Sergio Martínez Cuesta[1,2], Dario Beraldi[3], Fumiko Kawasaki[2], Robyn Hardisty[2], Zhe Li[2], Guillem Portella[2], Pierre Murat[2], Eun-Ang Raiber[2], Shankar Balasubramanian[1,2]

[1]University of Cambridge, Cancer Research UK - Cambridge Institute, Cambridge, United Kingdom, [2]University of Cambridge, Department of Chemistry, Cambridge, United Kingdom, [3]University of Glasgow, Cancer Research Centre, Institute of Cancer Sciences, Glasgow, United Kingdom

Our development of chemical and antibody methods to map modifications in DNA at different levels of resolution [1,2] is beginning to reveal the natural diversity of chemical changes in the four canonical nucleotide bases adenine (A), thymine (T), cytosine (C) and guanine (G). The combination of these maps with data obtained from genome-wide approaches measuring transcript levels (RNA-seq), distribution of nucleosomes (MNase-seq) and protein levels (also datasets available in GEO, ENCODE and literature) is starting to unveil potential biological functions for C and T modifications in nature and disease.

This presentation will provide an overview of our strategies to develop computational approaches to detect nucleotide modifications at single-base-resolution and at the genomic region levels. I will also introduce our approaches to integrate different types of datasets generated using our in vitro and in vivo experimental techniques. Finally, I will present our efforts to create an open environment where computational and experimental researchers share computer code, data and figures robustly and reproducibly within our chemical biology laboratory and beyond.

In addition to the extensive knowledge about the biology of 5-methylcytosine (5mC), we are trying to understand the functional roles of two other C modifications. Our recent whole genome single-base-resolution maps in tumour/margin samples of a glioblastoma patient revealed a depletion of 5-hydroxymethylcytosine (5hmC) during brain tumour development and also gave insights into how genetic and epigenetic changes are interrelated [3]. Our related research has also found 5-formylcytosine (5fC) to be a stable modification that can alter the structure of DNA and spans specific genomic profiles in samples from mouse tissues [4-6]. We are also mapping T modifications, namely 5-hydroxymethyluracil (5hmU), 5-(β-glucopyranosyl)hydroxymethyluracil (base J) and 5-formyluracil (5fU) in parasites (Trypanosoma and Leishmania) and human cell lines [7], and investigating their role in transcriptional regulation.

[1] Booth M. J. et *al.*, *Chem. Rev.* **2015** *115*:2240–2254
[2] Hardisty R. E. et *al.*, *J. Am. Chem. Soc.* **2015** *137*:9270–9272
[3] Raiber E-A. et *al.*, *npj Genomic Medicine* **2017** *2*:6
[4] Iurlaro M. et *al.*, *Genome Biology* **2016** *17*:141
[5] Bachman M. et *al.*, *Nature Chemical Biology* **2015** *11*:555–557
[6] Raiber E-A. et *al.*, *Nature Structural & Molecular Biology* **2015** *22*:44–49
[7] Kawasaki F. et *al.*, *Genome Biology* **2017** *18*:23

← Back

AN EFFICIENT ALGORITHM FOR LEARNING A GENE NETWORK UNDERLYING CLINICAL PHENOTYPES UNDER SNP PERTURBATIONS

Calvin McCarter[1], Seyoung Kim[2]

[1]Carnegie Mellon University, Machine Learning Department, Pittsburgh, PA, [2]Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA

Recent technologies are generating an abundance of 'omics' data, demanding new approaches for integrating data across the genome, transcriptome, and phenome to understand the genetic architecture and gene networks underlying diseases. Such methods for integrative analysis must also confront high-dimensional genomics data, comprising hundreds of thousands of SNPs and tens of thousands of gene expressions. Previous integrative methods have generally performed two-way 'omic' data analysis, examining genome-transcriptome interactions as in eQTL mapping, or genome-phenome interactions as in genome-wide association studies. Many such methods fall back on simple univariate analyses of the influence of a single SNP on a single gene expression or phenotype for statistical and computational simplicity, ignoring multiple correlated gene expressions due to gene-gene interactions in gene regulatory network. Network-based methods have been proposed in recent years, but their viability is limited by computational efficiency for high-dimensional datasets. In this work, we propose a statistical approach for multi-omic data analysis that tackles all these challenges by efficiently learning a cascade of sparse networks under genetic influence. Our statistical approach is based on Gaussian chain graph models that represent how genomes control transcriptome regulation in gene regulatory network, which in turn influence phenomes in a network of clinical traits. We propose an optimization algorithm for learning our model that is extremely efficient in both computation time and memory usage. We apply our method to analyze asthma data from patients from the Childhood Asthma Management Program (CAMP) study. With roughly 500,000 SNPs and over 10,000 genes our method learns the network cascade model for gene networks and phenotype networks under the influence of SNPs in less than two hours.

← Back

SCALING UP REFERENCE QUALITY ASSEMBLY OF VERTEBRATE GENOMES

Shane A McCarthy[1], Iliana Bista[1], Dirk-Dominik Dolle[1], Francesca Giordano[1], Hannes Svardal[1], Milan Malinsky[2], William Chow[1], Jingtao Lilue[1], Michelle Smith[1], Karen Oliver[1], Michael Quail[1], Thomas Keane[3], Kerstin Howe[1], Zemin Ning[1], Richard Durbin[1,4]

[1]Wellcome Trust Sanger Institute, Hinxton, United Kingdom, [2]University of Basel, Zoological Institute, Basel, Switzerland, [3]European Bioinformatics Institute, Hinxton, United Kingdom, [4]University of Cambridge, Department of Genetics, Cambridge, United Kingdom

Increased read length and throughput of long read sequencing technologies such as PacBio and Oxford Nanopore are now enabling high contiguity and accuracy de novo reference assemblies for vertebrate scale genomes. In association with the Vertebrate Genomes Project/Genome 10k Consortium, we are working on sequencing and assembly of genomes from a wide diversity of vertebrate species aiming to produce reference quality genomes from ~50 species of vertebrates, initially targeting fish, caecilian amphibians and some species of rodents. We define the reference quality assembly target here as greater than 1Mb contig N50, 10Mb scaffold N50, 90% assignment to chromosomes and base quality Q40, designated "3.4.2q40". As of August 2017 we have reached this target for three of our first five genomes based on a combination of ~50x PacBio long reads and 10X Genomics linked-reads, and expect to be able to report on ~25 more that are currently in various stages of data collection and processing. Our primary strategy consists of assembling the long read sequence, followed by scaffolding with linked reads. This is followed by gap-filling and polishing before assigning to synteny groups. Genomes with higher heterozygosity, higher repeat content or higher divergence from existing good quality assemblies may benefit from orthogonal technologies such as BioNano optical mapping or HiC. Upfront estimates of genome size, heterozygosity and repetitiveness using k-mer counting methods on the 10X Genomics Illumina data appear to be informative about the ease of assembly.

← Back

PIPELINE FOR SNP DISCOVERY IN RNA SEQUENCES

Carrie L McCracken[1], Amol C Shetty[1], Ricky S Adkins[1], Heather Huot Creasy[1], Theresa Hodges[1], Susan Dorsey[2], Michelle Giglio[1], Anup Mahurkar[1], Owen White[1]

[1]University of Maryland, Baltimore, Institute for Genome Sciences, Baltimore, MD, [2]University of Maryland, Baltimore, School of Nursing, Baltimore, MD

Many studies take advantage of the strengths of RNA sequencing (RNA-seq) such as measuring gene expression levels or characterizing splice sites. It can be advantageous to characterize the single nucleotide polymorphisms (SNPs) in the RNA-seq results to avoid the additional cost of DNA sequencing. At the Institute for Genome Sciences Informatics core, we include a SNP pipeline among our fee-for-services available. We have built and tested a pipeline for SNP discovery and characterization. We used the workflow for GATK (Genome Analysis ToolKit, https://software.broadinstitute.org/gatk/) and Annovar (http://annovar.openbioinformatics.org/). We tested the pipeline with human samples. We are able to find SNPs and to characterize whether a SNP is exonic, intronic, or intergenic. We can characterize the synonymous and nonsynonymous changes. We have demonstrated that our pipeline can find potentially biologically interesting SNPs.

← Back

SLEUTH-ALR: IMPROVING ESTIMATION OF *SEQ DIFFERENTIAL ANALYSIS USING COMPOSITIONAL DATA ANALYSIS WITH SLEUTH

Warren A McGee[1], Harold Pimentel[2], Lior Pachter[3], Jane Y Wu[1]

[1]Northwestern University, Department of Neurology, Chicago, IL,
[2]Stanford University, Department of Genetics and Biology, Stanford, CA,
[3]Caltech, Division of Biology and Biological Engineering, Pasadena, CA

Molecular probing *Seq techniques (e.g. RNA-Seq) generate "compositional" datasets, where the number of fragments sequenced is not directly proportional to the total RNA in the cell population, but rather directly proportional to other experimental factors. Thus, the dataset carries only relative information constrained by the arbitrary dataset size, even though absolute copy numbers are often of interest. Unless external information is added (e.g. spike-ins), or unless there is no change to the overall composition (i.e. only a few features change), then conclusions drawn from the relative information can be misleading. Critically, many datasets do not have external information available and have contexts where many features are expected to change (e.g. cancer cells versus control cells; knockout of a transcription factor).

We propose a new approach that combines an "additive logratio transformation" from Compositional Data Analysis with the strengths of the current pipeline using the R package *sleuth*, which we call **sleuth-ALR**. In addition, we have performed for the first time a simulation that directly addresses the conversion of absolute copy numbers into relative abundances, which allows testing the robustness of assuming no overall composition change. **Sleuth-ALR** has approximately equivalent performance with current pipelines when there is no overall change or a small change in composition, but has much improved performance when the compositional changes dramatically (similar to real datasets with verified violations of the assumption).

← Back

A VARIANT BY ANY OTHER NAME... ENSEMBL'S VARIANT RECODER

William McLaren, Sarah Hunt, Fiona Cunningham

European Molecular Biology Laboratory, European Bioinformatics Institute, Genome Analysis, Cambridge, United Kingdom

A single genomic variant can be referred to in many ways, including identifiers from databases such as dbSNP, commonly exchanged file formats such as Variant Call Format (VCF), and standardised notations such as Human Genome Variation Society (HGVS) nomenclature. Within each of these representations there is room for ambiguity, inconsistency and difficulty of interpretation that can lead to the same variant being referred to by scores of different encodings.

Standards are improving considerably and being increasingly adopted by software and services that report variant data. There remains, however, a need to translate between different standards, especially where those standards are open to variance and ambiguity. HGVS is a standard that is often misused, particularly when reporting variants in the literature: reference accessions are frequently invalid (gene symbols, chromosome names e.g. "chr1"), incomplete (missing version numbers) or entirely absent; standards for describing changes are not adhered to (using "ins" instead of "dup", differing encodings of STOP codons); changes are reported only at the protein level that do not resolve uniquely to a single genomic change.

We have created Variant Recoder, a tool that translates between a broad range of variant encodings including database identifiers, VCF and HGVS. Input standards are relaxed to allow interpretation of accession synonyms, complex multi-alleleic VCF entries via normalisation, and non-canonical, ambiguous or partial HGVS notations. Output standards are strictly enforced, with HGVS being reported on correctly accessioned genome, transcript and protein reference sequences from Ensembl, RefSeq and Locus Reference Genomic (LRG). Variant Recoder is available via a language-independent RESTful API [1], Ensembl's Perl API, or as a local executable as part of the Ensembl VEP package [2]. We anticipate a diverse range of use cases, including simple ID lookups, mapping protein variants to the genome, and translating between Ensembl and RefSeq gene sets.

[1]: http://rest.ensembl.org/documentation/info/variant_recoder
[2]: https://github.com/Ensembl/ensembl-vep#recoder

LARGE-SCALE SEARCH OF SHORT-READ SEQUENCING EXPERIMENTS

Brad Solomon, Carl Kingsford

School of Computer Science, Carnegie Mellon University, Computational Biology, Pittsburgh, PA

Public databases such at the NIH Sequencing Read Archive (SRA) now contain hundreds of thousands of short-read sequencing experiments. A major challenge now is making that raw data accessible and useful for biological analysis --- researchers must be able to find the relevant and related experiments on which to perform their analyses. A fundamental computational problem towards that effort is the problem of searching for short-read experiments by sequence. Specifically, given a query string Q and a very large collection of short-read sequencing experiments we want to quickly find the experiments that contain reads that make it likely that Q was among the sequences present, and we want to do this without appealing to a reference sequence or reference annotation (in order to support searching metagenomic and cancer experiments, for example).

We present an approach, called Sequence Bloom Trees, for solving this sequence search problem. Sequence Bloom Trees create an index of a hierarchy of Bloom filters summarizing short-read sequencing experiments. Sequence Bloom Trees appeared in Nature Biotechnology, 34:300–302 (2016). A refinement of these called Split Sequence Bloom Trees (which appeared in RECOMB 2017) further improves the search speed and reduces the size of the index. These approaches allow for the search of terabytes of raw short-read sequencing experiments in minutes using a single thread on a desktop-class computer. For example, we search over 2,500 RNA-seq experiments for every known human transcript sequence in about 3 days using a single thread using Sequence Bloom Trees (without the "Split" improvement).

Reference implementations of Sequence Bloom Trees and Split Sequence Bloom Trees are available as open source https://github.com/Kingsford-Group/bloomtree and https://github.com/Kingsford-Group/splitsbt.

STIX: A SCALABLE INDEX FOR MINING LARGE WHOLE-GENOME SEQUENCING COHORTS FOR RELIABLE STRUCTURAL VARIANT POPULATION ALLELE FREQUENCY ESTIMATES

Ryan M Layer[1,2], Brent S Pedersen[1,2], Aaron R Quinaln[1,2,3]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, USTAR Center for Genetic Discovery, Salt Lake City, UT, [3]University of Utah, Department of Biomedical Informatics, Salt Lake City, UT

When triaging variants observed in a patient with a genetic disease, the frequency of an SNV in a healthy cohort like gnomAD is vital to our assessment of pathogenicity. Unfortunately, there is no equivalent resource for structural variants (SVs). This gap is due to the complexity inherent to identifying SVs. Unlike SNV detection, which considers every sequence for every individual at all 3 billion positions, the number of possible SV configurations (3 billion squared) makes it intractable to interrogate all possible SVs. SV detection instead clusters evidence into a large set of possible variants. These sets are then filtered to maximize true positives and minimize false positives. This filtering makes it difficult to draw conclusions about the prevalence of SVs observed in new samples. It is impossible to discern whether the variant is absent from the population or if it was filtered out. We need a new method that can provide a full accounting of SVs among the deep SV data generated from a project such as TOPMed and the Centers for Common Disease Genetics.

Here we propose STIX, a structural variant index that enables rapid searches of efficient, lossless profiles of SV evidence across thousands of samples. For a given SV, STIX reports a per-sample count of all concurring evidence. From these counts we can, for example, conclude that an SV with high-level evidence in many samples is common and an SV with no evidence is rare. By representing the raw signal, we avoid the previously described false negative issue. We indexed the 2,504 genomes from the 1000 Genomes Project (1KG) and quantified the frequency of 14,146 cancer-related deletions from the COSMIC SV database. Each search took 0.1 seconds, and we found that 27% of SVs has some evidence in 1KG and 3% had evidence in at least 10% of the samples. We also use STIX to interpret SVs from families afflicted by rare disorders. In one case, a STIX query of an SV thought to be private to individuals affected by Treacher Collins identified the SV in one 1KG sample.

STIX can be useful for large-scale SV genotyping. With a compact representation of the reference evidence, STIX can quickly genotype new SVs across all indexed samples. This is vital to large sequencing projects that sequence cohorts in batches and must re-genotype new SVs across existing samples and existing SVs across new samples. STIX can also empower true population scale SV detection by jointly considering all samples.

← Back

GVCFLIB - AN EXTENSIBLE LIBRARY TO ANALYZE AND ACCURATELY ANNOTATE CLINICALLY SIGNIFICANT WILD TYPE AND MUTANT ALLELES FROM GVCF FORMATTED CALLSETS

Scott Mottarella, Sowmithri Utiramerur

Stanford Health Care, Clinical Genomics Program, Palo Alto, CA

The versatility of the Genomic VCF (GVCF) provides a solution to clinical applications in genomics where the confidence of all calls, variant and reference, must be maintained. To that aim, there is a need to provide bioinformatics solutions to traverse, parse, and manipulate GVCF files. gvcflib, which expands upon the existing library for VCF files, vcflib (written by Erik Garrison; erik.garrison@bc.edu), has been created to perform these functions. gvcflib improves upon the original library for faster processing while adding the required features for interpreting the additional constraints of the GVCF. While the core library allows for quick and simple development of project specific algorithms involving genomic variant data, several included functions have been created to provide easy access of GVCF parsing to all users. For example, a new variant annotator has been created using gvcflib that includes several features to improve annotation accuracy, a critical step for clinical applications where one missed annotation can prevent diagnosis or cause misdiagnosis. This new annotator correctly parses multiallelic variants that can be written either as multiple alternate alleles on a single variant line or as a new variant line for each possible alternate allele at a given position. While many annotators claim to handle multiallelic variants from one of these two formats, none that were tested were capable of parsing both formats simultaneously. In addition, this new annotator performs a sequence normalization step that allows for comparison between variants regardless of how they overlap. This ensures that no annotation is missed because of nuances in the format that can conceal the sequence equivalence of the variant and the annotation source. This algorithm can annotate an entire exome from multiple annotation sources in under twenty minutes from a laptop.

GENOME BROWSING ON SOMEONE ELSE'S COMPUTER

Ann Loraine[1], Nowlan Freese[1], David Norris[2], John Eckstein[2]

[1]University of North Carolina at Charlotte, Department of Bioinformatics and Genomics, Charlotte, NC, [2]StackLeader Inc., Concord, NC

Genome browsers help users understand genomic data. In principle, interactive browsers with sophisticated visual analytics ought to be better at generating insight than less interactive displays or static images. However, introducing new modes of interaction and new types of displays requires finesse. Developers face a challenge: How do we implement novel, never-seen-before methods of interaction in ways that are inherently intuitive to users? Fortunately, this process of introducing new ideas is easier thanks to genome browser ubiquity in biology. Users are already familiar with standard genome browser features such as zooming, adding and removing tracks, and fine-tuning track appearance. The genome browser user community seems ready and eager to adopt new approaches that will help them solve problems. One new direction involves using cloud computing resources to power visual analytics functions that would otherwise be impossible to achieve on a user's desktop computer. The Integrated Genome Browser project is now exploring this idea. Previously, we experimented with connecting the Galaxy workflow system to IGB, using Galaxy's external viewer API. We also linked IGB to the popular InterProScan protein analysis tool hosted at the European Bioinformatics Institute using its REST API. The usefulness of these features prompted us to propose linking IGB to other cloud and cloud-like resources via APIs. We are exploring linking IGB with CyVerse computational resources, using Agave REST APIs. I will summarize project goals and attempt to illustrate ideas for mitigating technical problems such as network latency via strategic interface design.

← Back

BIOARCH: A RECONFIGURABLE HARDWARE ACCELERATOR DESIGNED FOR BIOINFORMATICS WORKLOADS

S. Karen Khatamifard[1], Meisam Razaviyayn[2], Ulya R. Karpuzcu[1]

[1]University of Minnesota, Electrical Engineering, Minneapolis, MN,
[2]University of Southern California, Industrial and Systems Engineering, Los Angeles, CA

Recent advances in sequencing technologies have revolutionized medicine and biology. Modern sequencing platforms can sequence tens of billions of bases per each run. Processing these massive datasets can take up to hours or days even in the presence of significant amount of computational resources. These computationally expensive tasks motivate the use of hardware-level acceleration with optimized computing architectures. In this talk, we discuss how widely-used computational tasks in bioinformatics can significantly benefit from the use of optimal hardware architectures and algorithms. In particular, we introduce BioArch, a reconfigurable hardware accelerator designed for bioinformatics workloads. BioArch aims to accelerate major reference-guided and de novo tasks in computational biology.

BioArch, capable of analyzing both short and long reads, has two central hardware components: one for pre-aligning reads to reference sequence(s), Filter Unit (FilterU); and the other for finding exact pairwise similarity score between two given short/long reads, Match Unit (MatchU). FilterU can efficiently prune the candidate hits with the possibility of using parallel FilterUs to prune even more aggressively. MatchUs, on the other hand, evaluates the similarity of two given sequences in constant time independent of the sequencing error rate or read lengths. MatchU's design is based on the clever use of processing in-memory (PIM) technologies, capable of handling simple (mostly integer) computations inside the memory where the data reside. PIMs are recently introduced as an energy efficient novel form of computing. PIMs can effectively remove all data transfers to and from memory, which is the main performance and energy bottleneck of today's data-intensive applications.

In addition to standard k-mer hashing strategies, BioArch benefits from a novel hash-based similarity evaluation which has been recently introduced in the image-processing community. This hashing function is developed with the training of deep convolutional neural networks on sequencing data. Our hash-based neural network leads to a linear time alignment of similar PacBio sequences with more than 98% accuracy.

In the last part of the talk, we share our numerical experiments on BioArch for two important case studies: 1) short-read alignment of Illumina reads and 2) De novo transcriptome sequencing with PabBio long reads. Our simulations show orders of magnitude higher throughput (7.5x) and energy efficiency (109.0x), when compared to representative, optimized state-of-the-art software-based algorithms.

← Back

ENHANCING PRE-DEFINED WORKFLOWS WITH AD HOC ANALYTICS USING GALAXY, DOCKER AND JUPYTER

Anton Nekrutenko

Penn State, GalaxyProject, University Park, PA

Trees, rivers, and the analysis of next generation sequencing (NGS) data are examples of branching systems so ubiquitous in nature. Indeed, numerous types of NGS applications share the same initial processing steps (quality control, read manipulation and filtering, mapping, post-mapping thresholding, etc.), making up the trunk and main branches of this tree. Each of these main branches subsequently gives off smaller offshoots (variant calling, RNA-seq, ChIP-seq, and other "seqs") that, in turn, split further as analyses become focused towards the specific goals of an experiment. As we traverse the tree, the set of established analysis tools becomes increasingly sparse, and it is up to an individual researcher to come up with statistical and visualization approaches necessary to reach the leaves (or fruits) that represent conclusive, publishable results. Consider transcriptome analysis as an example. Initial steps of RNA-seq analysis (in our tree analogy, these are trunk and main branches), such as quality control, read mapping, and transcript assembly and quantification are reasonably well established. Yet completion of these steps does not produce a publishable result. Instead, there is still the need for additional analyses (progressively smaller branches of our tree), ranging from simple format conversion to statistical tests and visualizations. Thus, every NGS analysis can, in principle, be divided into two stages. The first stage involves processing of raw data using a small set of common, generic tools. This stage can be scripted and automated and also lends itself to building graphical user interfaces (GUIs) such as Galaxy. The second stage involves a much greater variety of tools that need to be customized for every given experiment (in many cases, there are no tools at all, and custom scripts need to be developed). As a result, it is not readily coerced into a handful of automated routines or generic GUIs.

Historically Galaxy system has been ideally suited for the first analysis stage described above. Galaxy users can utilize a large number of tools and workflows. Yet what they could not previously do is run ad hoc scripts and arbitrary tools within their Galaxy instance to perform the second, exploratory stage of analyses. This was very limiting, as initial analyses of data often involve interactive exploration with tools like Jupyter or RStudio—powerful platforms that are becoming increasingly popular in life sciences. Here, we showcase Galaxy Interactive Environment framework, designed to combine Galaxy's tools and workflows with environments such as Jupyter.

The main motivation for this work was the development of a system wherein biomedical researchers can perform both stages of data analysis: initial steps using established tools and exploratory and data interpretation steps with ad hoc approaches. Merging both steps into a unifying platform will lower entry barriers for individuals interested in data analysis, significantly improve reproducibility of published results, ease collaborations, and enable straightforward dissemination of best analysis practices.

← Back

METAVIZ AND THE HUMAN MICROBIOME PROJECT DATA PORTAL: INTERACTIVE STATISTICAL AND VISUAL ANALYSIS OF METAGENOMIC DATA FROM THE HMP

Justin Wagner*[1,2,3], Jayaram Kancherla*[1,2,3], Jonathan Crabtree[4], Anup Mahurkar[4], Hector Corrada Bravo[1,2,3]

[1]University of Maryland College Park, Computer Science, College Park, MD, [2]University of Maryland College Park, Center for Bioinformatics and Computational Biology, College Park, MD, [3]University of Maryland College Park, Institute for Advanced Computer Studies, College Park, MD, [4]University of Maryland Baltimore, Institute for Genome Sciences, Baltimore, MD

*These authors contributed equally

In this work, we present the integration of Metaviz, a web browser-based tool for interactive exploratory microbiome sequencing data analysis, with the Human Microbiome Project Data Portal. Metaviz relies on a navigation mechanism that is designed to effectively explore the hierarchically organized features of microbial community abundance profiles. We visualize abundance counts with interactive heatmaps, stacked bar plots, and boxplots that are dynamically updated as a user selects taxonomic features for inspection. Metaviz supports common data exploration techniques, including PCA scatter plots to interpret variability in the dataset and alpha diversity boxplots for examining ecological community composition. We implemented Metaviz to tightly integrate with R/Bioconductor packages for statistical analysis of metagenomic data so users can iteratively perform statistical analysis to drive visualization of abundance data in an interactive R session. Metaviz can also visualize abundance matrices that are stored in a persistent graph database. Metaviz is free and open source, documentation and links to github repositories are available at http://www.metaviz.org. The UMD Metagenome Browser is an instance of Metaviz hosted at http://metaviz.cbcb.umd.edu.

We have recently worked on connecting Metaviz to the Human Microbiome Project Data Portal, a service that provides an interactive catalog for users to explore the multiple 'omics' data and metadata generated by the NIH Common Fund's Human Microbiome Project. Microbial community abundance profiles can be viewed in the UMD Metagenome Browser that are directly linked to the HMP Data Portal which is available at https://portal.hmpdacc.org. A user can also select multiple samples using the Data Portal, save a manifest file of those samples, and use the manifest for creating a Metaviz workspace to inspect them. The integration of the extensive HMP data with an interactive visualization tool improves access of this public resource and enables hypothesis generation using the HMP data.

← Back

A *DE NOVO* ASSEMBLY OF THE NEANDERTAL GENOME PROVIDES INSIGHTS INTO HUMAN STRUCTURAL VARIATION.

Manjusha Chintalapati[1], Rohit Kolora[2], Kay Prufer[1]

[1]Max Planck Institute for Evolutionary Anthropology, Genetics, Leipzig, Germany, [2]German center for integrative biodiversity research, Bioinformatics, Leipzig, Germany

The analysis of nucleotide differences between present-day human genomes and the Neandertal genome has shown that the ancestors of present-day non-Africans admixed with Neandertals and yielded a catalogue of nucleotide changes that are unique to present-day humans. While some analysis of the Neandertal genome, based on read-coverage and paired-end sequences, yielded candidates of larger structural differences to present-day human genomes, the information on these type of changes is much more limited. The main reason for this limitation is the short fragment length and the substantial fraction of extraneous DNA.

Here we used short-read de novo assembly software based on de Bruijn graphs to assemble the ~50x coverage Altai Neandertal genome. We show that an initial step of read-correction based on k-mer frequencies is efficient in removing a large part of the substitutions caused by ancient DNA damage. Assembling the corrected sequences with the SOAPdenovo and minia assemblers resulted in N50 contig lengths of up to 2kb. While the longest assembled contigs show similarity to bacterial genomes, we found that a large fraction (50.8%) of the contigs match the human genome. Most human-aligned contigs yielded one uninterrupted alignment and most of the mappable human genome is covered by these alignments. However, a small fraction of contigs showed split alignments that could be caused by structural differences between the human reference and the assembled Neandertal genome. By applying stringent filters, that are trained on the ratio of inter-chromosomal to intra-chromosomal rearrangements and include sequence coverage as an additional source of information, we arrive at a set of candidate rearrangements, including some that overlap exons. Our analysis indicates that further information can be gained from the assembly of ancient DNA and is a first step towards a more complete Neandertal genome sequence.

← Back

THE EVOLUTION OF LIFESPAN AND THE EPIGENOME ASSESSED BY CPG FREQUENCY IN CONSERVED PRIMATE AND VERTEBRATE PROMOTERS

Adam T McLain[1], Christopher Faulk[2]

[1]College of Arts and Sciences, SUNY Polytechnic Institute, Department of Biology & Chemistry, Utica, NY, [2]College of Food, Agricultural, and Natural Resource Sciences, University of Minnesota, Department of Animal Sciences, Saint Paul, MN

CpG dinucleotides are estimated to mutate 10 to 50 times faster than non-CpG sites within vertebrate genomes, resulting in evolutionary and fitness consequences. Here we computationally analyzed the genomes of 131 mammal species (including 28 primates) within highly conserved promoter regions for the presence of CpG density correlated with a quantifiable trait. An initial database of 25,503 promoter regions from the human genome was obtained from the Eukaryotic Promoter Database and used as the basis for constructing a database of conserved promoters across mammalian genomes. Lifespan data obtained from the AnAge database was used as the quantifiable trait. Primates and all other mammals were analyzed as two separate datasets, and the results were compared. Briefly, we found that 987 (3.8% of) conserved promoter regions in primates showed a significant correlation between CpG density and lifespan and 94% of these displayed a positive correlation. Across the entire mammalian dataset 1079 conserved promoter regions were identified and again 94% displayed a positive correlation between CpG density and lifespan. Our results suggest that the most rapidly mutating sites within the genome, CpG sites outside of coding regions, are strongly and positively associated with life history traits. The list of genes identified through our method should be priority targets for epigenetic assessment or modification as it affects the correlated trait.

← Back

GLOBAL ANALYSIS OF HUMAN MRNA FOLDING DISRUPTIONS IN SYNONYMOUS VARIANTS DEMONSTRATES SIGNIFICANT POPULATION CONSTRAINT

Grant Lammi[1], James Li[1], Jeffrey Gaither[1], David Gordon[1], Harkness Kuck[1], Ben Kelly[1], James Fitch[1], Peter White[1,2]

[1]Nationwide Children's Hospital, The Institute for Genomic Medicine, Columbus, OH, [2]The Ohio State University, Department of Pediatrics, Columbus, OH

Guidelines for the interpretation of sequence variants currently state that novel synonymous variants are likely benign, unless a role in RNA-splicing can be demonstrated. *In silico* RNA folding studies suggest mRNA secondary structure may influence selection in humans and mammals, yet the potential for pathogenic synonymous variants that impact RNA folding in human genetic disease is largely unknown. As such, we set out to test the hypothesis that synonymous variants predicted to have disruptive impacts on RNA stability would show significant constraint in the human population.

To test this hypothesis, we performed a systematic whole transcriptome study of SNPs impacting RNA stability. First, we developed a big data pipeline to generate RNA folding statistics for every possible polymorphism in the known transcriptome (~0.5 billion variants). Second, we utilized population allele frequencies (AF) from 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals found in the Genome Aggregation Database (gnomAD), to determine if highly disruptive SNP mRNA folding values were constrained, thereby supporting our hypothesis that RNA stability plays a role in human health and disease.

All known RefSeq mRNA transcript sequences (42,726 transcripts) were retrieved and processed to generate 101 nucleotide flanking sequences for each position in the given transcript. Three alternate allele sequences were generated for each position and 10 RNA folding disruption metrics were calculated with the ViennaRNA package. The entire process was parallelized using Apache Spark to generate metrics for 445,740,246 total SNPs. These SNPs were then mapped back to the GRCh37 human reference genome, using a Spark wrapper for Picard tools Liftover, and joined with gnomAD AFs. Finally, a Spark implementation for SnpEff was developed that enabled functional annotation of all variants.

We observed that allele frequencies were highly constrained in SNPs that were predicted to disrupt the reference allele mRNA structure. To rigorously test this apparent relationship between allele frequency and our RNA disruption metrics, we utilized chi-squared tests. In each case we were able to strongly reject the hypothesis that AF and the disruption metric in question were independent quantities. These tests thereby clearly demonstrated that a relationship exists between RNA disruption and allele conservation, supporting our hypothesis that RNA stability plays a role in human health and disease. Given that upto 75% of rare disease whole exome sequencing studies have no clinically relevant finding, this dataset has the potential to enable discovery of new pathogenic variants that impact RNA stability, supporting a novel mechanism by which synonymous variants may contribute to human genetic disease.

← Back

LINKING MOLECULES WITH MORPHOLOGY IN THE -OMICS AGE: COMPUTATIONAL TAXONOMY PIPELINES FOR MICROBIAL METAZOA

Holly M Bik

University of California, Riverside, Department of Nematology, Riverside, CA

Microbial metazoa (organisms <1mm, including nematodes, tardigrades, platyhelminthes, other “minor” metazoan phyla, and eggs/larval stages of larger species) are abundant and ubiquitous in most soil/sediment habitats, performing key functions such as nutrient cycling and sediment stability in marine and terrestrial ecosystems. Yet, their unexplored diversity represents one of the major challenges in biology and currently limits our capacity to understand, mitigate and remediate the consequences of environmental change. Microbial metazoa have a strong history of morphological taxonomy (formal species descriptions, specimen drawings, monographs, etc.), but most of this information is offline and thus effectively inaccessible to -Omics studies. In addition, rRNA databases and genome collections lag far behind other groups of microbial organisms such as bacteria, archaea, fungi, and single-celled protists. This sparsity of computational resources severely limits the ecological and evolutionary insights that can be gained from high-throughput sequencing approaches focused on microbial eukaryotes. Here, I will discuss recent efforts to improve molecular databases expand bioinformatics pipelines for -Omic studies of “neglected” microbial metazoan groups, focusing on tool development as well as community building efforts.

← Back

STRAINS, FUNCTIONS, AND DYNAMICS IN THE EXPANDED HUMAN MICROBIOME PROJECT

Jason Lloyd-Price[1,2], Anup Mahurkar[3], Gholamali Rahnavard[1,2], Jonathan Crabtree[3], Joshua Orvis[3], A. Brantley Hall[2], Arthur Brady[3], Heather H Creasy[3], Carrie McCracken[3], Michelle G Giglio[3], Daniel McDonald[4], Eric A Franzosa[1,2], Rob Knight[4,5], Owen White[3], Curtis Huttenhower[1,2]

[1]Harvard T. H. Chan School of Public Health, Biostatistics Department, Boston, MA, [2]Broad Institute, Medical and Population Genetics, Cambridge, MA, [3]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [4]University of California San Diego, Pediatrics, La Jolla, CA, [5]University of California San Diego, Computer Science and Engineering, La Jolla, CA

The NIH Human Microbiome Project (HMP) has provided one of the broadest characterizations of the baseline human microbiome, serving as a reference in studies of disease, microbial population diversity, biogeography, and molecular function. Here, we present new findings and a dramatic expansion of shotgun metagenomes (now ~2,400 samples) from the HMP, termed HMP1-II, expanding both the number of subjects with sequenced metagenomes and the number of analyzed samples from multiple longitudinal visits. De novo assembly of this expanded set of samples yielded a new body-wide gene catalog totaling 1-8 million families per site; gene family discovery was not yet saturated for any of the targeted body sites, despite the ~3 fold increase in metagenomes and high-quality co-assemblies from multiple samples per subject. Strain identification revealed distinct sub-species clades genetically specialized within body sites, including Haemophilus parainfluenzae and Rothia mucilaginosa. This has also enabled the identification of species with phylogenetic diversity under-represented in reference genomes, to be prioritized for future isolation. The study also identified uniquely human-enriched microbial processes across the body, indicative of core functions defining the human microbiome. Other pathways were classified as niche-specific within body sites, or universal to microbial life (e.g. housekeeping). Finally, the longitudinal sampling enabled a characterization of the temporal dynamics of microbiome taxa and metabolic processes; specifically, we employed Gaussian processes to identify host-specific, temporally-variable, and rapidly-variable organisms and pathways. We found that species abundances in the gut were most individualized, with the Bacteroides genus in particular exhibiting highly stable and personal abundances. Meanwhile, members of the Firmicutes, as well as most metabolic pathway abundances, tended to be shared among individuals but vary in abundance over time. This detailed characterization of microbial diversity advances our understanding of personalized microbiome function and dynamics.

← Back

K-MER COMPARISON METHODS IN METAGENOMICS, APPLICATIONS AT THE COMMUNITY LEVEL

David C Molik[1,2], Michael E Pfrender[1], Scott Emrich[2,3]

[1]University of Notre Dame, Department of Biological Sciences, Notre Dame, IN, [2]University of Notre Dame, Integrated Biomedical Sciences Program, Notre Dame, IN, [3]University of Notre Dame, Department of Computer Science and Engineering, Notre Dame, IN

K-mer comparison methods, such as alignment-free approaches, like those provided by mash (1) or CAMERA (2) generate metagenomic sample to sample distances, which in turn can be clustered via hierarchical methods. Some of these comparison methods, like those that leverage sketching, suffix trees, or de Bruijn graphs, show significant speed up while reducing data footprint; an important feature as both sequence depth as well as the breadth of available data increase. These tools show promising application towards understanding microbial communities because unsupervised clustering can help determine interrelated communities among individuals or sampling sites. The methods also can be used in species detection and surveillance, especially among low abundance species.

De novo clustering of k-mer derived distances describe structure within the data that may correlate with environmental drivers, such as the environment in which the organism lives, i.e., a soil, marine, or host organism environment. Equally as promising is work in environmental DNA, or metabarcoding looking for multicellular life, which shows that this still holds true. K-mer similarity can also sort samples into sampling location if the sampled community is clearly distinct, as in the case with water samples from shipping ports, desert soil, or streams.

Minhash sketching as implemented by mash can also be used for species detection and surveillance using eukaryotic barcoded sequences. Coarse (sample-wide) and fine-grained (each sequence is sketched) strategies were can be used for determining the presence or absence of a targeted species in a total of 12634 metagenomic datasets, which at the time of analysis were the entire collection of 18S metagenomic datasets on NCBI's SRA (short read archive). We show that available methods in k-mer analysis can be used in metabarcoding studies for the determination of significant subsets and for the presence absence of particular species.

References

[1] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Sergey Koren, and Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology 17(1): 132.
[2] Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: A Community Resource for Metagenomics. PLOS Biology 5(3): e75

← Back

A WHOLE-GENOME PHYLOGENETIC HYPOTHESIS ACROSS THE THREE DOMAINS OF LIFE

Rebecca B Dikow[1], Katrina M Pagenkopp Lohan[2], Paul B Frandsen[1]

[1]Smithsonian Institution, Data Science Lab, Washington, DC, [2]Smithsonian Institution, Smithsonian Environmental Research Center, Edgewater, MD

The phylogenetic relationships among Archaea, Bacteria, and Eukaryota provide the context for understanding diversification and adaptation within and across these three major groups. This is a challenging question for a number of reasons: (1) they are anciently diverging, leading to nucleotide saturation and extinction, (2) genetic material has been transferred horizontally, and (3) there are significant numbers of undiscovered taxa (including higher taxa such as phyla). Previous hypotheses of these relationships have been based on few, highly conserved, loci. Here a hypothesis is presented based on complete or draft whole genome alignments for more than 3,000 species representing all available taxonomic groups. Three separate sets of genome alignments, one with a Bacteria reference species, one with an Archaea reference species, and one with a Eukaryote reference species, are considered. We present Maximum Likelihood trees based on analysis of complete alignments, partitioned by gene, gene tree analyses, and species tree analyses. Beyond the phylogenetic results, strategies for data mining, quality control, and visualization for large comparative genomics datasets are presented.

A STATISTICAL TEST FOR STRUCTURAL COVARIATIONS IN RNA AND PROTEINS

Elena Rivas[1], Sean R Eddy[1,2,3,4]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, [2]Howard Hughes Medical Institute, Harvard University, Cambridge, MA, [3]FAS Center for Systems Biology, Harvard University, Cambridge, MA, [4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

Pairwise covariations observed in RNA and protein alignments provide a powerful means of deducing evolutionarily conserved structures. However, confounding covariations can appear for other reasons than a conserved RNA or protein structure. We have introduced a method, called R-scape, for calculating the significance of base pair covariation in an alignment under a null hypothesis that considers spurious covariations that arise by phylogenetic correlation instead of structural constraints.

For RNAs, R-scape finds significantly covarying base pairs in several novel RNAs, including autoregulatory ribosomal protein mRNA leaders in γ-proteobacteria, and noncoding RNAs in α-proteobacteria, a noncoding RNA in the ciliate *Oxytricha*. R-scape also helps improve the structural annotation of known structural RNAs such as those in the database Rfam. On the other hand, R-scape finds no significant covariation for several secondary structures that have been proposed for the long noncoding RNAs (lncRNAs) HOTAIR, SRA, or Xist, nor for any alternative structure of these RNAs. R-scape can be used as a systematic tool to identify new structural RNAs.

For proteins, R-scape can help elucidate how many of the observed protein contacts can be traced back to covariations due to structural constraints versus induced by the particular phylogeny of the alignment. R-scape can also help elucidate the relative merit of different covariation measures. The need for a statistical test to distinguish structural covariations from phylogenetic ones is independent of which particular covariation measure one uses. R-scape implements several different covariation measures, some (such as mutual information or the G-test) are simple statistics calculated directly from the alignment, whereas others (such as those derived from statistical Potts models) depend on many parameters which require training. We will present results comparing different covariation measures.

← Back

LARGE-SCALE ANALYSIS OF GENOME-WIDE ENHANCER AND GENE ACTIVITY REVEALS A NOVEL ENHANCER-PROMOTER MAP

Tom A Hait[1], David Amar[1,2], Ran Elkon[3], Ron Shamir[1]

[1]Tel-Aviv University, Computer Science, Tel-Aviv Yafo, Israel, [2]Stanford University, Stanford Center for Inherited Cardiovascular Disease, Stanford, CA, [3]Tel-Aviv University, Human Molecular Genetics and Biochemistry, Tel-Aviv Yafo, Israel

Massive international efforts have recently documented hundreds of thousands of putative enhancers in the human genome. A pressing challenge is to identify which of these candidate elements are actually functional and pinpoint for each gene which enhancers regulate it. Here we present FOCS, a novel method for inferring enhancer-promoter (E-P) links based on correlated activity patterns across many samples from heterogeneous genomic sources. FOCS uses a rigorous statistical validation pipeline tailored for zero-inflated enhancer activity data, and optimizing each gene model to derive the most important E-P links.

We applied FOCS to four massive data sets, spanning together 2,630 samples from diverse cell types and conditions. The data originated from ENCODE, Roadmap Epigenomics, FANTOM5, and our novel GRO-seq based compendium of eRNA and gene expression profiles. We collectively inferred ~300,000 cross-validated E-P links spanning ~16K known genes. When tested against gold standards of E-P interactions derived from ChIA-PET and eQTL data, FOCS made far more predictions than extant methods and at the same time achieved higher true positive rate. The new extensive resource of statistically validated E-P interactions can greatly assist the functional interpretation of the non-coding genome.

MODELING METHYL-SENSITIVE TRANSCRIPTION FACTOR MOTIFS WITH AN EXPANDED EPIGENETIC ALPHABET

Coby Viner[1,2], James Johnson[3], Charles A Ishak[2], Nicolas Walker[4], Hui Shi[4], Marcela Sjöberg[5], Shu Yi Shen[2], David J Adams[5], Anne C Ferguson-Smith[4], Daniel D De Carvalho[2,6], Timothy L Bailey[7], Michael M Hoffman[1,2,6]

[1]University of Toronto, Department of Computer Science, Toronto, Canada, [2]Princess Margaret Cancer Centre, Toronto, Canada, [3]The University of Queensland, Institute for Molecular Bioscience, Brisbane, Australia, [4]University of Cambridge, Department of Genetics, Cambridge, United Kingdom, [5]Wellcome Trust Sanger Institute, Cambridge, United Kingdom, [6]University of Toronto, Department of Medical Biophysics, Toronto, Canada, [7]University of Nevada, Reno, Department of Pharmacology, Reno, NV

Introduction. Many transcription factors (TFs) initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a TF might encounter. Cytosine is often present in the modified forms: 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC). TFs have been shown to distinguish unmodified from modified bases. Modification-sensitive TFs provide an additional epigenetic mechanism to modulate gene regulation and downstream expression.

Methods. To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites in DNA with covalent modifications. Our models expand the standard `A`/`C`/`G`/`T` alphabet, adding `m` (5mC) and `h` (5hmC). We adapted the position weight matrix (PWM) formulation of TF binding affinity to this expanded alphabet.
We engineered several tools to work with expanded-alphabet sequence and PWMs. First, we developed a program, Cytomod, to create a modified sequence, using data from bisulfite and oxidative bisulfite sequencing experiments. Second, new versions of MEME, DREME, and MEME-ChIP enable *de novo* discovery of modification-sensitive motifs. A new version of CentriMo enables central motif enrichment analysis to infer direct DNA binding in an expanded-alphabet context. These versions permit users to specify new alphabets, anticipating future alphabet expansions. Finally, we are collaborating with the authors of the recently-developed RSAT `matrix-clustering` software to add support for our alphabet, enabling clustering of modified PWMs.

Results. We created expanded-alphabet genomes using whole-genome maps of 5mC and 5hmC in naive *ex vivo* mouse T cells and 5mC in K562 leukemia cells. Using this sequence, expanded-alphabet PWMs, and ChIP-seq data from Mouse and Human ENCODE and others, we identified *cis*-regulatory modules active only in the presence or absence of cytosine modifications. We reproduced various known methylation binding preferences, including the preference of ZFP57 and C/EBPβ for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. Using these known binding preferences as controls, we discovered novel preferences for 5 TFs, as well as numerous new 5mC and 5hmC motifs. We are currently validating our top predictions, planning to conduct ChIP-BS-seq and CUT&RUN. We expect that using `matrix-clustering`, we will be able to elucidate motifs groups with distinct methylation preferences, reconciling recent indications of TFs with bidirectional or ambiguous methylation preferences.

NEAR-NUCLEOTIDE MAPPING OF R-LOOPS SHOWS THAT PROMOTER-ASSOCIATED R-LOOPS ARE BOUNDED AT FIRST EXON-INTRON JUNCTIONS

Jason G Dumelie, Samie R Jaffrey

Weill Cornell Medicine, Cornell University, Pharmacology, New York City, NY

Human genomes contain R-loops, which are structures that consist of a strand of DNA hybridized to RNA and a complementary DNA strand expelled by the RNA. R-loops are enriched within promoter regions where R-loops have been shown to impact gene expression. However, the location of R-loops within promoter regions has not been well defined. As a result, it is not clear what genomic domains are perturbed by the formation of promoter-associated R-loops. To resolve the location of promoter-associated R-loops, I established a novel R-loop mapping method, bisDRIP-seq, that uses bisulfite to map R-loops throughout the genome at near-nucleotide resolution. Through this method I discovered that the location of promoter-associated R-loops is determined by the intron structure of genes. In genes that contain introns, R-loops are typically restricted to the region between the transcription start site and the first exon-intron junction. In genes that lack introns, I frequently observed very prominent R-loop formation. R-loops in these genes lacking introns are also often bounded, but there is gene-specific heterogeneity in the boundaries of these genes. This study therefore generates a near-nucleotide map of R-loops and demonstrates that R-loop formation is impacted by gene structure.

← Back

NEW METHODS FOR MEASURING NATURAL SELECTION AND PREDICTING DELETERIOUS VARIANTS IN NONCODING REGIONS OF THE HUMAN GENOME

Adam Siepel

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Many genetic variants that influence phenotypes of interest are located outside of protein-coding genes, yet existing methods for identifying such variants have poor predictive power. I will describe a new computational method, called LINSIGHT, that substantially improves the prediction of noncoding nucleotide sites at which mutations are likely to have deleterious fitness consequences, and which, therefore, are likely to be phenotypically important. LINSIGHT combines a generalized linear model for functional genomic data with a probabilistic model of molecular evolution. The method is fast and highly scalable, enabling it to exploit the 'big data' available in modern genomics. I will show that LINSIGHT outperforms the best available methods in identifying human noncoding variants associated with inherited diseases. Finally, I will describe an extension of LINSIGHT that considers the full site frequency spectrum and allows for the estimation of position- and allele-specific selection coefficients. So far, we have applied this method to coding sequences in the human genome, where it reveals surprisingly strong selection on synonymous sites, classes of genes that have undergone relaxed and enhanced selection in recent human evolution, and other aspects of natural selection on coding regions. Work is underway to extend this method to noncoding regions.

SELECTIVE CONSTRAINTS ON ENHANCER AND PROMOTER SEQUENCES ACROSS HUMAN CELL-TYPES

Max Schubach, Martin Kircher

Berlin Institute of Health (BIH), Computational Genome Biology Group, Berlin, Germany

The ENCODE consortium profiled DNase-accessible regions and numerous histone marks genome-wide and in a large panel of different human cell types. This data makes it possible to compile a catalog of tissue-specific regulatory elements and ENCODE recently released a registry of candidate Regulatory Elements (cREs). The registry entails more than 1.31 million elements in over 618 human cell types and 45 different tissues.

For about one fifth (126/618) of all cell types, individual cREs are annotated as enhancer-like and promoter-like based on histone marks. We analyzed these cREs and detected large variation in the number of active enhancer or promoter cREs across cell types and tissues. We show that the presence or absence of such elements provides a cell-type ontogeny, similar to that obtained from gene expression data.

We overlapped cREs with human-derived variants, identified as differences between a 1000 Genome consensus and the Ensembl Compara human ancestor sequence, as well as with variants present in present-day human populations. Like coding exons, candidate promoter and enhancer regions are under strong functional constraint (i.e. purifying selection). The constraint is larger on promoters than it is for enhancers; it is generally weaker compared to coding regions. Further, the depletion of human-derived variants (i.e. amount of constraint) shows a clear positive correlation with the number of active promoters across cell-types (Pearson's rho 0.87, p-value < 2.2e-16). This suggests that a high number of active promoters in a cell type is compensated by higher functional constraint on those promoters. Interestingly, we do not observe a similar effect for enhancers.

Our analysis outlines a promising route for understanding functional constraints on regulatory sequences and studying their relationships across cell-types. While still not comprehensive for all cell-types studied in the ENCODE efforts, the cREs registry provides a rich resource for our and similar studies.

← Back

THE CONSEQUENCES OF PROMOTER BIRTH AND DEATH IN THE HUMAN POPULATION

Robert S Young, Martin S Taylor

University of Edinburgh, MRC Human Genetics Unit, MRC IGMM, Edinburgh, United Kingdom

Promoters are the site at which gene regulatory signals are integrated and the site of transcription initiation during gene expression. Gene expression changes are thought to underlie much phenotypic variation between species and individuals, and may be responsible for individual variation in disease susceptibility and prognosis. We have used Cap Analysis of Gene Expression (CAGE) data to identify functional promoters across a range of tissues in both human and mouse [1]. Our evolutionary analysis of these sequences has shown that their complete birth and death has been a common occurrence since the divergence of these two species [2]. Promoter birth and death is particularly frequent in the testis and the immune system, suggesting a potential selective pressure driving these events. Furthermore, they are associated with genes whose coding sequence is undergoing positive selection across the mammalian clade. However, using derived allele frequency tests, we show that these evolutionarily volatile promoters are not consistently experiencing selection within the human population. We integrated these evolutionary records with the data of the GTEx consortium and discovered a specific enrichment of eQTLs within those human-specific promoters that show sequence turnover (inserted or deleted promoter sequence) but not functional turnover (where the underlying sequence is conserved). These insertion- and deletion-associated eQTLs tend to have a greater magnitude of effect than other promoter-overlapping eQTLs. These eQTLs also have a lower minor allele frequency than other GTEx variants suggesting those human-specific promoters that are detectably influencing genic transcription are subject to purifying selection in the human population. By analysing the transcriptomic and potential phenotypic consequences of evolutionary volatile promoters we hope to better understand the effect of these common evolutionary events in subsequently driving biological variation within the human population.

1. Forrest, A.R., et al., *A promoter-level mammalian expression atlas*. Nature, 2014. **507**(7493): p. 462-70.
2. Young, R.S., et al., *The frequent evolutionary birth and death of functional promoters in mouse and human*. Genome Research, 2015. **25**: p. 1546-1557.

← Back

META-ANALYSIS OF CHROMATIN ACCESSIBILITY TO DETERMINE MEANINGFUL VARIATION

Jayon Lihm[1], Sara Ballouz[1], Sandra Ahrens[2], Hayan Lee[3], Shane McCarthy[4], W. Richard McCombie[1], Bo Li[1,2], Jesse Gillis[1]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY, [2]Cold Spring Harbor Laboratory, Neurosicence, Cold Spring Harbor, NY, [3]Stanford University, Genetics Department, Stanford, CA, [4]Regeneron, Statistical Genetics at Regeneron Genetics Center, Tarrytown, NY

Chromatin accessibility provides an important window into the regulation of gene expression. Recently, the Assay of Transposase Accessible Chromatin with sequencing (ATAC-seq) was developed to profile genome-wide chromatin accessibility. Due to its simple protocols and low requirement in the amount of cells, ATAC-seq have been widely used in recent epigenetics studies. However, there has been little comparative or aggregate evaluation of technical bias and background distribution of ATAC-seq data. This is critical both to determine appropriate methodologies, controls, and efficacy, as well as to determine the global biological landscape of chromatin accessibility across diverse conditions. One major technical problem to address is that the counts of ATAC-seq reads underlying each peak vary substantially within a single sample and also between samples. Such variation makes the comparison to determine presence and absence of peaks, i.e. the open and closed state of chromatin, more difficult and less statistically well grounded. In this work we analyzed 197 ATAC-seq mouse samples from 13 studies in order to test the robustness of peak calling results and their specificity across studies. We find that peaks are promiscuously identified, with approximately 34K peaks per sample on average. These peaks overlap substantially with transcription start sites (TSS), covering 11K genes on average. Bootstrapped assessment of the underlying reads identifies these peaks the most robust. We also find a set of 451 genes that are commonly accessible, regardless of sample specifics. We evaluate the properties of these genes in detail, including mean expression across a diverse corpus of data and functional enrichment. Finally, we propose a novel approach to evaluate the robustness of peak signals and sensitivity by subsampling reads and re-calling peaks for each sampling, whose calls are then aggregated. This yields peak calls that are highly robust to variation in noise as a source of peaks within data itself. We applied this approach to our own ATAC-seq data on amygdala and cortex mouse samples under two different experiments; fear-conditioning vs control and ErbB4 knock-out vs wildtype. The subsampling approach reduced the number of TSS-accessible genes to nearly half, increasing the specificity. We identify 22 genes that are preferentially accessible in cortex samples.

FAST GENOME ALIGNMENTS FROM PSEUDOALIGNED RNA-SEQ DATASETS USING KALLISTO

Páll Melsted[1], Harold Pimentel[2], Nicolas Bray[3,4], Lior Pachter[5]

[1]University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, Reykjavik, Iceland, [2]Stanford University, Departments of Genetics, Palo Alto, CA, [3]University of California, Berkeley, Innovative Genomics Institute, Berkeley, CA, [4]University of California, Berkeley, Department of Molecular and Cell Biology, Berkeley, CA, [5]California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA

Visualizing RNA-Seq data requires aligning reads to a genome or transcriptome. Most read visualizers work with genome-sorted BAM files, thus making transcript visualization cumbersome. We introduce a new feature into kallisto for producing fast and approximate genome alignments of RNA-Seq reads post-quantification. In this mode reads are pseudoaligned to the reference transcriptome, projected down to genome coordinates, and written directly to a coordinate-sorted binary BAM file. This binary file is indexed and ready for viewing using Integrated Genome Viewer (IGV) or similar software. The advantage of doing alignments post quantification is that alignment records can be tagged with the transcripts they pseudoaligned to and their posterior alignment probabilities. This allows for new developments in visualizing RNA-Seq alignments in a genome context at the transcript level.

DEVELOPING A SCOTTISH VARIANT REPOSITORY

Alison M Meynert[1], Javier Santoyo Lopez[2], Timothy J Aitman[3], Colin Semple[1]

[1]University of Edinburgh, MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom, [2]Edinburgh Genomics, Edinburgh, United Kingdom, [3]University of Edinburgh, MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom

The Scottish Genomes Partnership (SGP) is a major funding initiative between Scottish universities and NHS Scotland (NHSS) to sequence human genomes for research and to build clinical whole genome sequencing capacity. The SGP is funding the whole genome sequencing of population isolates (Viking study of Shetland islanders), rare diseases in a research context (eye malformations, motor neuron disease, microencephaly, sex differentiation), cancer (pancreatic, oesophageal, and ovarian) and clinical patient sequencing within the NHSS. The Edinburgh arm of the partnership is developing a securely hosted Scottish variant repository to warehouse and provide the variants for use within the SGP and longer term. Variants from previous sequencing projects, primarily exome based, will also be stored in the repository.

The variant repository will offer both genome browser and sample-centric views, including familial context for samples sequenced as part of trio or family based studies. Multiple software packages will be deployed to serve the differing needs of users. The gnomAD[1] browser code base will be adapted to provide functional annotations and summary allele frequencies. The underlying individual genotypes will be accessed separately via a sample-centric browser with links out to the functional and frequency annotations.

[1]Lek et al 2016. Nature 536:285-291.

← Back

IDENTIFICATION OF CANDIDATE GENES UNDERLYING NODULATION-SPECIFIC PHENOTYPES IN MEDICAGO TRUNCATULA THROUGH INTEGRATION OF GENOME-WIDE ASSOCIATION STUDIES AND CO-EXPRESSION NETWORKS

Jean-Michel Michno[1,2], Liana Burghardt[3], Junqui Liu[2], Joseph R Jeffers[4], Peter Tiffin[3], Robert M Stupar[1,2], Chad L Myers[1,4]

[1]University of Minnesota, Bioinformatics and Computational Biology Program, Minneapolis, MN, [2]University of Minnesota, Department of Agronomy and Plant Genetics, Saint Paul, MN, [3]University of Minnesota, Department of Plant and Microbial Biology, Saint Paul, MN, [4]University of Minnesota, Department of Computer Science and Engineering, Minneapolis, MN

Genome-wide association studies (GWAS) are valuable for identifying genetic intervals associated with phenotypic variation. Specific GWAS intervals vary in size, depending on the historical local recombination within and around each associated locus. Typically, significant intervals span numerous gene models, limiting the ability to resolve high-confidence candidate genes underlying the trait of interest. However, additional data, such as gene expression profiles, can be combined with the genetic mapping information to successfully identify candidate genes. Co-expression network analysis provides information about the functional relationship of each gene through its similarity of expression patterns to other well defined clusters of genes. Though co-expression networks have been effective in identifying genes associated with a specific phenotype, they are rarely used to prioritize specific genes that contribute to highly quantitative traits. In this study, we integrated data from GWAS and co-expression networks to pinpoint candidate genes that may be associated with nodule-related phenotypes in Medicago truncatula. We further investigated a subset of these genes and discovered several with annotations linked to nodulation, including MEDTR2G101090 (PEN3-like), a previously validated gene associated with nodule number.

DESIGNING CANCER VACCINES FOR TRIALS OF PERSONALIZED IMMUNOTHERAPY

Christopher A Miller[1,2], Jasreet Hundal[1], Susanna Kiwala[1], Aaron Graubert[1], Joshua McMichael[1], Jason Walker[1], Amber Wollam[1], Jonas Neichin[1], Megan Neveau[1], Obi L Griffith[1,2], Elaine R Mardis[3], Malachi Griffith[1,2]

[1]Washington University in St Louis, McDonnell Genome Institute, St Louis, MO, [2]Washington University in St Louis, Div of Oncology, Dept of Medicine, St Louis, MO, [3]Nationwide Children's Hospital, Institute for Genomic Medicine, Columbus, OH

Immunomodulatory drugs, such as pembrolizumab, have been shown in recent clinical trials to be remarkably effective agents against cancers with a high mutation load. By counteracting tumor-mediated suppression, they allow the immune system to recognize neoantigens formed by mutations specific to cancer cells. Personalized cancer vaccines are a parallel and likely complementary approach, in which tumor-specific neoantigens are identified in a patient and then a vaccine is tailored specifically to their disease.

Designing effective cancer vaccines is a cross-disciplinary challenge, involving genomics, proteomics, immunology, and computational approaches. We have built a computational framework called pVAC-Tools that, when paired with a well-established genomics pipeline, produces an end-to-end solution for vaccine design. Key steps include: 1) identification of altered peptides from different mechanisms (i.e. point mutations, indels, gene fusions, or neo-ORFs). 2) Prediction of peptide accessibility via an ensemble of MHC Class I and II binding algorithms. 3) Prioritization by integrating data like mutant allele expression, peptide binding affinities, and the clonal/subclonal nature of a mutation 4) Interactive visualization via a web interface that allows clinical collaborators to curate results. 5) For DNA-vector vaccines, a procedure that minimizes junctional epitopes and produces near-optimal vaccine ordering through a simulated annealing approach.

We have used this package to design and produce vaccines for ongoing clinical trials (triple-negative breast cancer) as well as for compassionate use patients (glioblastoma, pediatric ependymoma) and mouse models. Though these studies are still in progress, we show data from several patients that demonstrate enhanced immune responses post-vaccination, as well as sequencing data from residual tumors that reveal patterns of tumor evolution and the possible emergence of resistant subclones.

PSEUDOGENES IN THE MOUSE LINEAGE: TRANSCRIPTIONAL ACTIVITY AND STRAIN-SPECIFIC HISTORY

Cristina Sisu[1,2], Paul Muir[2], Mark Gerstein[2]

[1]Brunel University London, Biosciences, London, United Kingdom, [2]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT

Pseudogenes are ideal markers of genome remodeling. In turn, the mouse is an ideal platform for studying them, particularly with the availability of transcriptional time course data during development (just completed in phase 3 of ENCODE) and the sequencing of 18 strains (completed by the Mouse Genome Project). Here we present a comprehensive genome-wide annotation of the pseudogenes in the mouse reference genome and associated strains. We compiled this from combining manual curation of over 10,000 pseudogenes with results from automatic annotation pipelines. By comparing the human and mouse, we annotated 271 unitary pseudogenes in mouse, and 431 unitaries in human. We collected all our annotation and analysis into a comprehensive resource that is freely available online at mouse.pseudogene.org. The overall mouse pseudogene repertoire (in the reference and strains) is similar to human in terms of overall size, biotype distribution (~80% processed, ~20% duplicated) and top family composition (with many GAPDH and ribosomal pseudogenes). However, notable differences arise in the age distribution of pseudogenes, with multiple retro-transpositional bursts in mouse evolutionary history and only one burst in human. Furthermore, in each strain ~20% of the pseudogenes are unique, reflecting strain-specific functions and evolution; for instance, the differences observed in the evolution of taste receptors associated pseudogenes in the NZO mice can be related to their change in the diet. Additionally, we find that ~15% of the pseudogenes are transcribed, a fraction similar to that for human and that pseudogene transcription exhibits greater tissue and strain specificity compared to their protein coding counterparts. Finally, we show that processed pseudogenes are commonly associated with highly transcribed genes.

← Back

IDENTIFICATION, REGULATION, AND FUNCTION OF ANTISENSE TRANSCRIPTION IN THE ESTROGEN RESPONSE IN BREAST CANCER CELLS

Tulip Nandu[1,2,3], Rui Li[1,2,3], Miao Sun[1,2,3], Shrikanth Gadad[1,2,3], Minho Chae[1,2,3], W Lee Kraus[1,2,3]

[1]The Laboratory of Signaling and Gene Regulation, Department of Obstetrics and Gynecology, Dallas, TX, [2]Cecil H. & Ida Green Center for Reproductive Biology Science, Department of Obstetrics and Gynecology, Dallas, TX, [3]The Division of Basic Sciences, Department of Obstetrics and Gynecology, Dallas, TX

Transcriptome profiling studies suggest that greater than 70% of the human genome is transcribed, with about 20% of transcribed regions exhibiting antisense transcription from the opposite DNA strand. Antisense transcription generates antisense RNAs (asRNAs), which are complimentary to their cognate sense RNAs, including messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs). Previous studies suggest that sense/antisense transcript ratios are globally altered in breast cancers and specific asRNAs, including H19 and TFPI-2, are overexpressed in breast cancers. We are interested in characterizing the antisense transcriptome in breast cancer cells. To that end, we have used genomic approaches, including global run-on sequencing (GRO-Seq), which provides the location of actively transcribing RNA polymerases across the genome, and RNA-seq, which indicates the steady-state levels of transcripts genome-wide, to generate comprehensive lists of antisense transcription units and asRNAs in MCF-7 human breast cancer cells. As expected, the promoters driving antisense transcription exhibit chromatin features found at active promoters, such as enrichment for DNAseI hypersensitivity and H3K4me3, indicating proper identification of the antisense genes. We are now assessing the roles of antisense transcription, which may generate transcriptional interference, and asRNAs, which may cause RNA interference, in estrogen-regulated transcriptional responses. We found that antisense transcription running through the transcription start sites of sense estrogen-regulated genes is associated with enhanced estrogen-dependent transcription, increased RNA polymerase II loading, and a more open chromatin architecture relative to those sense genese that do not have antisense transcription. We are now exploring the molecular and functional consequences of these outcomes. These studies are revealing new facets of the estrogen-regulated transcriptome and suggest that antisense transcription and/or asRNAs may play a role in estrogen-dependent signaling outcomes. This work is supported by a grant from the NIH/NIDDK to W.L.K.

← Back

COPY NUMBER VARIATION ANALYSIS USING A TARGETED NEXT-GENERATION SEQUENCING AMPLICON PANEL: NEXTFLEX® DMD AMPLICON PANEL FOR DUCHENNE MUSCULAR DYSTROPHY

Dylan Fox, Jiri Nehyba, Radmila Hrdlickova, Lucas Akin, Masoud Toloue

PerkinElmer, Bioo Scientific, Austin, TX

Duchenne muscular dystrophy (DMD) is the most common fatal genetic disorder diagnosed in childhood. With 20,000 new cases appearing each year, it affects one in 3,500 males across the globe. DMD codes for the protein dystrophin and greater than 1,000 genetic mutations have been identified. The most common mutations in the DMD gene are deletions (65%) and, less frequently, missense mutations (29%) and erroneous duplications (6%). In the past, labs have had to defer to the use of both non-sequencing methods such as MLPA or microarrays and either Sanger or next-generation sequencing as it was not possible to detect deletions and duplications (copy number variation - CNV) by targeted sequencing alone.

The new NEXTflex® DMD Amplicon Panel and accompanying analysis solution detects both CNV and SNVs across the entire coding region of DMD (~25.7 kb of genomic sequence) in a single sequencing run. Reagent and analysis algorithm performance was verified using samples known to contain CNV alterations via an alternative method. Using the panel's robust coverage uniformity across 167 amplicons that cover DMD as well as other select sequences as a negative control for CNV calculations, the analysis solution uses a combination of field-accepted open-source tools for alignment with a proprietary software component for sequence grooming, CNV detection, and reporting. Furthermore, a Taylor series-based CNV detection algorithm was compared with a method based on internal controls to establish an optimal algorithm for CNV detection. This analysis strategy may be applicable to other genes where CNV is a commonly found mutation in different disease states.

← Back

SCAFF10X: A RELATIONAL MATRIX BASED ALGORITHM FOR GENOME SCAFFOLDING USING 10X DATA

Zemin Ning, Francesca Giordano

Wellcome Trust Sanger Institute, Sequencing Informatics, Cambridge, United Kingdom

Scaff10x is a pipeline for genome scaffolding using 10x data. Barcoded tags are extracted from raw sequencing reads and appended to read names for further processing. Alignments are carried out either with BWA or SMALT. Barcodes are processed and sorted together with contigs as well as mapping coordinates. A relational matrix is constructed to record the shared barcodes among the contigs which may be linked. Order and orientation are conducted after nearest neighbours are searched. We report a number of applications such as improving supernova assemblies directly, adding 10X data to scaffold PacBio assemblies and comparisons with another 10x scaffolder ARCS. In the human genome, an assembly with scaffold N50 = 1.0 Mb has been improved to N50 = 30.1Mb (30 times) by Scaff10x. For the 3.10 Gb Tasmanian devil genome, we have produced an assembly of N50 at 105Mb with only 64 scaffolds using Bionano and 10x reads.

The pipeline can be downloaded at

https://sourceforge.net/projects/phusion2/files/scaff10x/

HIGHLY SCALABLE GENOME ANALYSIS USING ADAM, CANNOLI, AND AVOCADO

Frank A Nothaft[1], Alyssa Morrow[1], Devin Petersohn[1], Michael Heuer[1], Justin Paschall[1], Ryan Williams[2], Taner Dagdelen[1], Uri Laserson[2], David Haussler[3], Benedict Paten[3], David A Patterson[1], Anthony D Joseph[1]

[1]UC Berkeley, AMPLab, Berkeley, CA, [2]Mount Sinai, School of Medicine, New York, NY, [3]UC Santa Cruz, Biomedical Engineering, Santa Cruz, CA

The rapidly increasing amount of genomic data in public and private repositories currently rivals the scale of data collected and processed by Internet-scale companies. This scale of data necessitates the development of parallel genome analysis software that is optimized for both cloud computing and on-premises cluster computer. In this talk, we describe ADAM, an open source library that enables the use of the popular Apache Spark data processing framework to parallelize genomic analyses across large compute clusters. By using ADAM and cloud computing services, we can align and call variants on a 60x coverage whole genome in under an hour at a cost of less than $15.

Using the ADAM library, we have built a variant caller (Avocado) that achieves accuracy competitive with state-of-the-art tools (>99% SNP accuracy, >97% INDEL accuracy on NIST Genome-in-a-Bottle truth sets) while reducing end-to-end latency for alignment through variant calling to less than an hour for a 60x WGS dataset on an 800 core compute cluster. To perform the alignment phase of this pipeline, we developed the Cannoli tool to parallelize the BWA MEM aligner across an Apache Spark cluster. Beyond BWA MEM, Cannoli can be used to parallelize other common aligners such as Bowtie, Bowtie2, GSNAP, GMAP, and STAR, as well as variant callers like FreeBayes, and annotation tools like SnpEff. While several prior projects such as CloudBurst and CrossBow have used MapReduce platforms to parallelize genomics tools, Cannoli provides a general approach to automatically parallelizing single-node tools, and a new tool can typically be parallelized on top of Cannoli using less than 10 lines of code.

Beyond variant calling, ADAM presents a clean interface for querying against genomic data in R, Python, SQL, Scala, or Java, and can operate on genomic reads, variants/genotypes, and features. The ADAM library integrates with tools for performing statistical genetics tests, training machine learning models, visualizing genomic data, and performing genomic arithmetic/set theory operations. ADAM is released as open source software under a permissive Apache 2 license and can be deployed both in on-premises clusters, or on multiple cloud computing providers.

NETWORK ANALYSIS OF TRANSCRIPTOME IDENTIFIES PREDICTIVE BIOLOGICAL PATHWAYS IN HYPERTENSIVE AFRICAN AMERICANS

Cihan Oguz, Adam R Davis, Gary H Gibbons

National Human Genome Research Institute, Cardiovascular Disease Section, Bethesda, MD

Hypertension (HTN) is a cardiovascular disease (CVD) risk factor that substantially increases the risk for heart attack, congestive heart failure, chronic kidney disease, and stroke, if left untreated. In U.S., HTN is significantly more prevalent among African Americans in comparison with other ethnicities. While increased age and obesity are the strongest risk factors leading to treatment-resistant HTN, the differences between controlled and treatment-resistant HTN in terms of their underlying molecular pathophysiology remain unclear, especially among African Americans.

We implemented a predictive modeling approach to study hypertensive cases and normotensive controls (with optimal blood pressure of <= 120/80 mmHg without any antihypertensive medication) among 180 African-American patients (67 females and 113 males) from the Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID) Network Study. Model inputs included clinical data (from serum and urine samples) and RNA-Seq based gene expression data (from peripheral whole blood samples). By integrating these data sets into machine learning models and interaction networks, we identified anatomical, adaptive, neural, hemodynamic, endocrine, and humoral processes predictive of HTN, and its treatment-resistant and controlled subphenotypes. The presence of several stress response related pathways and processes within the molecular signature of treatment-resistant HTN suggested that the resistance against anthihypertensive treatment could be a result of advanced "vascular age". In contrast, a far less complex picture emerged from the molecular signature of controlled HTN.

Understanding the biological processes driving the pathophysiologies of different subphenotypes of complex diseases is one of the goals of personalized medicine. Our findings, which were consistent with the "Mosaic Theory of HTN", shed light into distinct molecular signatures and biological processes associated with the controlled and treatment-resistant subphenotypes of HTN. Biological pathways enriched within the most predictive network modules suggested that different mechanisms and groups of genes were predictive of these two subphenotypes. This study describes a machine learning based framework that uses genomic data to generate biological insights, which can guide future clinical and pharmacogenomic studies in targeting specific pathways to improve treatment of HTN in African Americans.

← Back

SREVED - SPLICING REGULATORY ELEMENT VARIANT EFFECT DETERMINATION

Gavin R Oliver, Naresh Prodduturi, Klee Eric

Mayo Clinic, Center for Individualized Medicine, Rochester, MN

Genomic variants that affect the normal pattern of mRNA splicing are widespread and play central roles in the pathogenesis of disease. These variants can affect a variety of splicing regulatory elements (SREs) including splice sites, splicing enhancers and splicing silencers. Multiple algorithms exist to predict these elements within DNA and identify variants within them, however the manifestation of their effects at the mRNA level is highly variable and difficult to predict, resulting in a high burden of false positive variants incorrectly predicted as pathogenic. Conversely, RNA-based methods exist to predict aberrant splicing but these solutions similarly generate large numbers of candidate events amongst which false positive results predominate. The shortcomings of these methods severely limited their applicability in genomics studies.

To address these shortfalls we have developed an approach to identify genuine splicing aberrations caused by variants in SREs based on RNA-Seq and optionally supporting DNA-Seq data. By manually annotating the mRNA-level effects of variants on thousands of SREs we have developed a pipeline that measures highly specific local neighborhood criteria in the vicinity of SRE affecting variants to accurately characterize the ultimate effect of a variant at the RNA level in terms of common effects including whole or partial intron retention or exon exclusion, and abolition or introduction of splicing sites. Where reference expression data exist, the method can optionally predict effects of the splicing aberration’s effects on relative exon or gene level transcript expression.

The resulting tool, SREVED (Splicing Regulatory Element Variant Effect Determination) is a high specificity solution with practical applicability to genomics and transcriptomics studies within both basic sciences and translational research. We have applied the method to undiagnosed diagnostic odyssey patient data to identify candidate causative genomic variants.

THE GEAR PORTAL – SHARING AND DISPLAYING GENE EXPRESSION NOW SIMPLIFIED AND DIVERSIFIED

Joshua D Orvis[1], Dustin J Olley[1], Amiel Dror[2], Michael Kelly[3], Anup Mahurkar[1], Ronna Hertzano[1]

[1]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [2]Tel Aviv University, ., Tel Aviv, Israel, [3]National Institutes of Health, NIDCD, Bethesda, MD

Next generation sequencing as a tool for recording gene expression has increased in popularity over the past decade, now extending also to single cell RNA-seq. While tissue processing, library preparation and sequencing are now both faster and more affordable, data sharing and visualization remain challenging. We previously introduced the gEAR portal (gene Expression Analysis Resource) as a website for dissemination, presentation and cross-comparison of cell type-specific gene expression in the public domain. Since its conception, the gEAR has evolved to allow many new features, including: 1) multiple forms of data visualization (graph, SVG-coloring and violin plots); 2) a user-friendly data uploader; 3) uploaded data can now remain private and viewed in the context of the public datasets, or be posted for sharing publically; 4) private data sharing with collaborators is now possible through unique links; 5) monitoring privacy of shared data; 6) customized layouts; 7) online manual and many additional features and updates.

The gEAR is a home and a portal for all members of the hearing research community, but is expanding into other fields, including brain research. For those working on gene expression it allows visualization of their data in the context of other private and public datasets, including single cell RNA-seq and soon to come complex data-query. For all other researchers, the gEAR is an ever updating portal answering questions from as simple as ‘where is this gene expressed in the ear’ to ‘how does it change in published datasets’ and also offers direct links to a variety of resources.

The gEAR portal is publically available and can be accessed through umgear.org

← Back

GENE TREE GUIDED SEARCH AND VISUALIZATION AT GRAMENE

Andrew Olson[1], Demitri Muna[1], James Thomason[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, USDA-ARS, Ithaca, NY

The Ensembl Compara gene tree pipeline reconstructs the evolutionary history of protein-coding gene families, documenting speciation and duplication events in ancestral genomes. This foundational resource is used by the Gramene project (www.gramene.org) to identify gene annotation errors (split gene models), project curated pathways from rice to other plant species, and suggest functional annotation for less well-annotated orthologs.

Gramene's search engine (data.gramene.org) indexes gene models with their associated ontology terms, InterPro domains, pathways, and homology relationships. For gene models that lacking a meaningful name or description, the search interface displays the closest homolog in a model species. When a search only matches genes in well-annotated model organisms, the homology pane can be used to modify the search to show homologs.

The homology pane contains an integrated visualization component that displays the gene tree, expanded to show the gene of interest and the closest well-annotated homolog. Additional information is displayed as a track next to each leaf node or collapsed subtree. The default display mode shows an overview of the multiple sequence alignment, color-coded by InterPro domain. Users can navigate the view or switch to a scrollable sequence level mode. A new neighborhood conservation view has been added that shows local syntenic relationships across the gene family. A gene expression display mode that would show expression profiles for comparable tissues/conditions is currently under development.

Navigating a gene tree with hundreds of genes from dozens of species is overwhelming and unnecessary for most users. This is of particular importance in light of Gramene's plans to more than double the number of hosted reference genomes. To help users cope with large gene families, we have implemented a filter to limit search results and prune species and gene trees on the fly to only display genes from a user-defined subset of genomes. The gene tree pruning code also removes excess gaps from the multiple sequence alignment which result from removing less closely related species.

These developments add value beyond the standard Ensembl views by integrating diverse data into a fast gene tree interface while giving scientists the flexibility to focus on the species most relevant to their research.

All code developed for the web service and the search interface are maintained on GitHub (github.com/warelab)

Gramene is supported by NSF grant IOS-1127112, and partially from USDA-ARS (1907-21000-030-00D).

← Back

RAPID, LARGE-SCALE ANNOTATION USING THE CLOUD-ENABLED GENOMIC ANNOTATION LOGIC AND EXECUTION SYSTEM (GALES)

Joshua D Orvis[1,2], Heather Creasy[1], Anup Mahurkar[1], Owen White[1], Michelle Giglio[1]

[1]University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, [2]Johns Hopkins University, Advanced Academic Programs, Baltimore, MD

It is now possible to generate a quality structural and functional annotation of an entire bacterial genome in minutes on a single laptop, with result visualization and full evidence attribution. To achieve this, we implemented flexible pipelines using the Common Workflow Language which runs within a Docker image locally or on cloud resources.

Wishing to accommodate a variety of projects, we added pipeline templates into GALES for prokaryotic, eukaryotic and metagenomic annotation, though users can modify these or build their own. The system uses an ordered hierarchy to apply functional evidence based on the user's settings, along with pre-compiled reference data sources and annotation attribute collections. Users can easily change scoring cutoffs, choose/exclude evidence tiers, or prioritize the order in which evidence is applied. All required software and reference databases are stored within the Docker container, so users can download, install, and generate an annotation locally using the system in as few as two commands.

GALES also contains a user interface which can be used to browse and interrogate the resulting annotation of each genome. It displays aggregate information such as GO term clouds, overall statistics, full genes lists and genomic context views as well as display of all the evidence which contributed to the annotated function. Annotations can be exported in GFF3, GBK or TBL format for direct GenBank submission.

Informing its development methodologies, and to illustrate its scalability, GALES was used to annotate thousands of sequenced reference bacterial isolates as part of the Human Microbiome Project.

← Back

RNA SEQUENCING AND PROTEOMICS APPROACHES REVEAL NOVEL DEFICITS IN THE CORTEX OF *MECP2*-DEFICIENT MICE, A MODEL FOR RETT SYNDROME

Natasha L Pacheco[1], Michael R Heaven[2], Leanne M Holt[1,3], David K Crossman[4], Kristin J Boggio[5], Scott A Shaffer[5], Daniel L Flint[2], Michelle L Olsen[1,3]

[1]University of Alabama at Birmingham, Cellular, Developmental, and Integrative Biology, Birmingham, AL, [2]Vulcan Analytical, LLC, Proteomics, Birmingham, AL, [3]Virginia Polytechnic and State University, School of Neuroscience, Blacksburg, VA, [4]University of Alabama at Birmingham, Heflin Center for Genomic Science, Birmingham, AL, [5]University of Massachusetts Medical School, Proteomics and Mass Spectrometry Facility, Biochemistry and Molecular Pharmacology, Shrewsbury, MA

Rett syndrome (RTT) is an X-linked neurodevelopmental disorder caused by mutations in the transcriptional regulator MeCP2. RTT is characterized by having apparently normal development until 6-18 months, when a progressive decline in motor and language functions begins and breathing abnormalities and seizures present. Despite intense research, the molecular targets of MeCP2 and their contribution to the disease are unknown. Here we present the first comprehensive transcriptomic and proteomic analysis in a RTT mouse model. Examining whole cortex tissue in symptomatic males (*Mecp2*-null) with wildtype littermates, we have identified 391 genes (FDR < 0.05), and 465 proteins ($p < 0.1$) considered to be significantly, differentially expressed. These data indicate RNA metabolism, proteostasis, monoamine metabolism, and cholesterol synthesis are disrupted in the RTT proteome. Hits common to both data sets indicate disrupted cellular metabolism, calcium signaling, protein stability, DNA binding, and cytoskeletal cell structure. Finally, in addition to confirming disrupted pathways and identifying novel hits in neuronal function, our data indicate aberrant myelination, inflammation and vascular disruption. Intriguingly, in opposition to what is typically observed in most neurological diseases, there is no evidence of reactive gliosis. Instead, gene, protein, and pathway analyses suggest astrocytic maturation and morphological deficits. To further investigate this finding, we performed transcriptomic analyses on acutely isolated cortical astrocytes throughout postnatal development of *Mecp2*-null and wildtype littermates. Preliminary analysis of transcriptomic data from symptomatic *Mecp2*-null male cortical astrocytes (postnatal day 60+) also indicates disrupted genes and pathways associated with astrocytic maturation. Transcriptomic analyses of the remaining cortical astrocytic developmental time points are ongoing. Collectively, these analyses support previous works indicating widespread CNS dysfunction and may serve as a valuable resource for those interested in cellular dysfunction in RTT.

← Back

zUMIs: A FAST AND FLEXIBLE PIPELINE TO PROCESS RNA SEQUENCING DATA WITH UMIS

Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, Ines Hellmann

Ludwig-Maximilians University, Munich, Anthropology & Human Genomics, Department of Biology II, Munich, Germany

RNA-seq methods have become more and more sensitive, so that today RNA-seq libraries are prepared from single cells or even single nuclei. The throughput of such scRNA-seq protocols is rapidly increasing, enabling the profiling of tens of thousands of cells and opening exciting possibilities to analyse cellular identities. In this context, unique molecular identifiers (UMIs) are used to reduce amplification noise and sample-specific barcodes are used to increase throughput. Here, we present zUMIs, a fast and flexible pipeline to process data from RNA-seq protocols with barcodes and UMIs. zUMIs is a pipeline that processes paired fastq files containing the UMI, barcode and cDNA sequence, filters out reads with bad barcodes or UMIs based on sequence quality, maps reads to the genome and summarizes results in count tables and descriptive statistics plots. Additionally, for cell types such as neurons, it has proven to be more feasible to isolate RNA from single nuclei which decreases mRNA amounts further. In such cases, it has been suggested to count intron-mapping reads as part of nascent RNAs. Thus, zUMIs also allows the quantification of intronic reads that are generated from unspliced RNA. zUMIs also helps to make the often hugely varying library sizes of single cell data comparable by implementing a downsampling function for cells with excessive number of reads. zUMIs is flexible with respect to the length and sequences of the barcode and UMIs, making it compatible with all major scRNA-seq protocols featuring UMIs, including single-nuclei sequencing techniques, droplet based methods where the barcode is unknown as well as plate-based UMI-methods with known barcodes.
zUMIs is open source and available at https://github.com/sdparekh/zUMIs.

SCALING UP THE TREEFAM RESOURCE IN ENSEMBL

Mateus Patricio, Matthieu Muffato, Uma Maheswari, Nishadi De Silva, Paul Kersey, Bronwen Aken, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom, United Kingdom

The Ensembl and Ensembl Genomes projects generate and distribute large scale genome annotations for a wide range of genomes, including model organisms and several different mouse strains, allowing free access to genomic data annotation. The rapidly increasing number of publicly available genomes offers a great opportunity to better understand the evolution of species across the different kingdoms of life. We provide an automated method [1] to infer phylogenetic trees and homologies that has been regularly used across all Ensembl websites. The team is now expanding the TreeFam method, initially focused on Metazoans, to reach the different divisions of the tree of life available on Ensembl Genomes (fungi, plants and protists) and unify them under a common resource.

The TreeFam method places profile Hidden Markov Models (HMMs) at its heart. HMMs provide a scalable and sensitive way of classifying genes into families. We will present the effort undertaken to build a comprehensive library of HMM profiles representing gene families across all eukaryotes. It is based on two existing exhaustive classifications - Panther v12 (www.panther.org) and TreeFam v9 (www.treefam.org) - that we expand by creating new HMMs to fill gaps in gene coverage. Alongside, we are also deploying quality control checks to evaluate the gene-families based on their presence across the tree of life, properties of their multiple-sequence alignment and of their phylogenetic tree. Additionally, in order to evaluate the degree of changes that gene families are undertaking with the introduction of this new classification system, we compute the Jaccard index, which compares the similarity and diversity of the families, and the Gini coefficient to measure their statistical dispersion. Finally, we are also controlling that the resulting gene-trees and homologies still exhibit the same quality as the existing Ensembl methods in terms of properties of the sequence alignment, conservation of protein domains, and orthology predictions for model organisms. Here we will present an overview of the TreeFam approach with details on the quality controls, and statistics on the gene-families.

[1] - Ensembl comparative genomics resources - Javier Herrero, Matthieu Muffato et al. Database (Oxford). 2016; 2016: bav096.

LARGE-SCALE DISTRIBUTED GENOMIC ANALYSIS USING LIME AND GNOCCHI

Devin Petersohn, Taner Dagdelen, Frank A Nothaft, Nathaniel Parke, Patrick Yang, Gunjan Baid, Anthony D Joseph

UC Berkeley, AMPLab, Berkeley, CA

As a result of the widespread adoption of high throughput sequencing technologies, genomic datasets have rapidly grown both in number of samples and amount of data generated per sample, and this trend is accelerating. However, our ability to cost-efficiently process and analyze genomic data has not kept pace. To address the gap between high throughput data generation platforms and today's slow single-node processing tools, we present two software tools built for distributed computing environments, Lime and Gnocchi, to enable large scale genomic set arithmetic and Genome Wide Association Studies (GWAS), respectively. Both tools are implemented using Apache Spark and the open-source large-scale genomic processing platform ADAM and for each we demonstrate efficient scalability to large datasets and concordance with existing tools on small datasets.

Genomic set theory and arithmetic are an important part of the overwhelming majority of genomics data analysis pipelines. Existing tools, such as Bedtools, minimize their memory and CPU footprints to accomplish set-theory and arithmetic for genomics data. Most of the tools require that the data be pre-sorted before most of the computation to take advantage of the indexing for memory optimization. This is very effective for small workloads that run on a researcher's personal machine or in a resource limited system. However, performance breaks down as workloads grow in size. With our tool, Lime, we aim to reduce the computing cost and runtimes of genomic set-theory and provide a platform that scales, allowing users to pay more to get faster results for time-sensitive applications. We also provide these functionalities as both command-line tools and extensible, open-source libraries for users to add to existing high-throughput pipelines.

GWAS is a critical analysis tool for identifying associations between single nucleotide polymorphisms (SNPs) and phenotypic traits. While using whole genome sequences (WGS) rather than these tag SNPs should increase the power of GWAS to identify significant variants using available sample sizes, conducting a GWAS on WGS has not been possible.This is currently due to the fact that existing tools are unable to accommodate large cohorts of samples (hundreds of terabytes of data). In order to solve this problem, we built Gnocchi, to enable GWAS over large amounts of WGS data. Using Gnocchi and dbGaP datasets, we demonstrate concordance with current tools that can run GWAS on SNPs, and also demonstrate Gnocchi's ability to scale to large sets of whole genome sequences using UK10K datasets which resulted in increased ability to identify significant associations between genetic variants and phenotypic traits.

dbSNP 2.0

Lon Phan, Brad Holmes, Eric Moyer, Evgeny Ivanchenko, Damon Revoe, Hua Zhang, Wang Qiang, Eugene Shekhtman, David Shao, Anna Glodek, Rama Maiti, Ming Ward
NIH/NLM/NCBI, dbSNP, Bethesda, MD

NCBI dbSNP house variation and frequency data from large scale projects including HapMap, 1000Genomes, GO-ESP, ExAC, TOPMED, and HLI to more focused studies such as locus-specific databases (LSDB) and clinical sources. The database is used world-wide in the fields of personal genomics, medical genetics, and for managing, annotating, and analysis of variation data. dbSNP aggregate genetic variation data from multiple submitters and assigned stable Reference SNP (rs) identifiers that can be used for citing in publication and for integrating with other data sources. The rs records are annotated on RefSeq genome, mRNA, and protein sequences and integrated with other NCBI resources (Assembly, Gene, RefSeq, PubMed, and BioProject) and disseminated to the scientific community.

Current trends suggest millions of new variations will be discovered in the next few years from large-scale WGS and WES projects that will lead the explosive growth of dbSNP. For instance, within the last few months, dbSNP human data have double in size to 325M Reference SNP (rs) record with the addition of more than 170M novel RS in the latest build 150 release (April 2016). As the data volume continue to grow it become increasing challenging to process, annotate, and exchange with the community quickly and efficiency. To address these challenges and to improve data quality and accuracy, dbSNP has undergone a redesign and provided some preview products for community feedbacks that include:

1) A new algorithm and data model to describe sequence changes and variant normalization called Sequence Position Deletion Insertion (SPDI) that is available as service accessible programmatically (https://www.ncbi.nlm.nih.gov/projects/variation/services/v0/). The features are:

-- Produce a contextual allele that corrects left/right-shifting
-- Right-shifted HGVS
-- Left-shifted VCF fields
-- Remap (or lift-over) locations based on the alignment dataset used by ClinVar and dbSNP
-- Cluster to a canonical representative at a single location

2) Migration from SQL data to the JSON data object. Descriptions, examples, tutorials are available on dbSNP FTP site (ftp://ftp.ncbi.nlm.nih.gov/snp/.redesign/) and GitHub (https://github.com/ncbi/dbsnp).

-- Complete RS data struct without complex SQL queries
-- Faster parsing than XML
-- RS retrieval API (Coming soon)

3) A new RefSNP Page design with a modern UI and backend (https://www.ncbi.nlm.nih.gov/snp/rs328)

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

← Back

GENEY: A DATA ECOSYSTEM THAT ENABLES BIOLOGISTS TO EFFICIENTLY SUBSET, VISUALIZE, AND ANALYZE GENOMIC DATA

PJ Tatlow[1], Jonathan B Dayton[1], Zachary Ence[1], Stephen R Piccolo[1,2]
[1]Brigham Young University, Department of Biology, Provo, UT,
[2]University of Utah, Biomedical Informatics, Salt Lake City, UT

Modern technologies are generating vast troves of genomic data, and many of these data are are publicly available. Many non-computational biologists would like to work with relatively small subsets of these data—for example, to examine expression levels of a particular gene in breast tumors or to compare tumors that do or do not harbor a particular genomic aberration. However, these researchers often have trouble performing such analyses because the data are large in size and are stored in a variety of arbitrary formats. Consequently, these researchers must often rely on collaborators with computational skills, thus slowing and complicating the research process—and preventing computational biologists from working on more interesting problems. To address this challenge, we have developed a data ecosystem called *Geney*. From a user's perspective, *Geney* is a Web-based application through which the user can select datasets, filter the data based on sample-level criteria, slice the data based on desired features (e.g., particular genes or proteins), and download the data in their format of choice. In addition, the web interface can be used to generate and export publication-quality graphics and perform statistical analyses—graphing options are displayed dynamically depending on the data types available in a given data set. From a developer's perspective, *Geney* is a continuous-integration system that extracts data sets from public repositories (e.g., The Cancer Genome Atlas, Gene Expression Omnibus, LINCS), stores the data in a consistent, compressed format, and provides a REST API that can be queried by the front end (or programmatically, if desired). These data-preparation steps are encoded using reproducible scripts and monitored by a continuous-integration server. In our presentation, we will emphasize the computational aspects of this architecture and will describe evaluations we performed to optimize the storage format, filtering process, and user-interface design.

← Back

BENCHMARKING 50 CLASSIFICATION ALGORITHMS ON 45 TRANSCRIPTIONAL BIOMARKER DATASETS

Nathan P Golightly[1], Avery Bell[1], Stephen R Piccolo[1,2]

[1]Brigham Young University, Biology, Provo, UT, [2]University of Utah, Biomedical Informatics, Salt Lake City, UT

Researchers use transcriptional profiles to predict biomedical outcomes such as disease development, prognosis, and treatment response. When such predictions are sufficiently accurate, they have potential to guide precision-medicine efforts. Due to the size and complexity of these data, machine- and statistical-learning algorithms promise to help increase the accuracy of such predictions. However, due to the vast array of algorithms and hyperparameter combinations, it is difficult for a given researcher to know a priori which algorithm(s) and hyperparameter combination(s) will lead to the highest accuracy on a given dataset. In practice, researchers often choose algorithms and hyperparameters arbitrarily, which may result in suboptimal performance and/or bias. A better alternative would be to make such decisions in a data-driven manner. To address this need, we have compiled a collection of 45 transcriptional biomarker datasets from the public domain and have performed an extensive benchmark analysis. To minimize confounding effects, we renormalized the datasets using a standard preprocessing pipeline, removed low-quality samples, and custom-curated the clinical annotations. Using these data, we evaluated the performance of 50 different classification algorithms. These algorithms came from various open-source machine-learning libraries—including scikit-learn, weka, and mlr. Initially, we used default hyperparameter values and found that algorithm performance varied dramatically for a given dataset and across datasets. In general, kernel-based and ensemble-based algorithms outperformed other types of algorithm. Next we repeated the analysis but optimized hyperparameter choice and found that the performance often improved, sometimes considerably. Importantly, even the algorithms that performed best on average performed quite poorly in some cases, thereby confirming the need to choose algorithms empirically on an external reference set. Finally, we used feature-selection algorithms to reduce dimensionality, but this generally resulted in worse performance, thereby demonstrating the need to develop and test more sophisticated methods for selecting and constructing features. Our analysis provides additional insights on optimizing algorithm performance, which may not be obvious from other benchmark studies, such as the DREAM challenges, because we held potential confounding factors constant and focused solely on algorithm performance. Our data, scripts, and results are all freely available for others to examine and reuse (see https://osf.io/ssk3t/).

ASSEMBLY OPTIMIZATION IN BOTH SPACE AND TIME OF THE LARGEST GENOME TO DATE.

Sean Powell[1], Martin Pippel[1], Sergej Nowoshilow[2], Elly Tanaka[2], Siegfried Schloissnig[1]

[1]Heidelberg Institute of Theoretical Studies, CBI, Heidelberg, Germany,
[2]Institute of Molecular Pathology, Tanaka Group, Vienna, Austria

Discounting filtering, the all-against-all comparison performed during the assembly using long noisy reads results in excessive runtime and storage requirements due to the polynomial scaling of the number of repeat induced alignments dependent on the read-mass. Brute-forcing this comparison becomes quickly infeasible when moving beyond genomes of the size and repetitiveness of the human genome. We assembled the 32GB genome of the *Ambystoma mexicanum*, commonly referred to as the Axolotl, using the MARVEL assembler. Without runtime and storage optimizations we estimated runtime and storage requirements in excess of 1M CPU hours and 2PB respectively. We developed an on-the-fly repeat masking strategy to reduce those requirements by an order of magnitude, at the price of a decrease in assembly contiguity. The assembly provides valuable insights into the forcing responsible for the genome's expansion and proves that giant genomes are now amenable for assembly.

← Back

FORGe: PRIORITIZING VARIANTS IN GRAPH GENOMES

Jacob Pritt[1,2], Ravi Gaddipati[3], Ben Langmead[1,2]

[1]Johns Hopkins, Computer Science, Baltimore, MD, [2]Johns Hopkins, Center for Computational Biology, Baltimore, MD, [3]Johns Hopkins, Biomedical Engineering, Baltimore, MD

Read alignment is a central computational problem in genomics. There is growing interest in using known genetic variants to augment the reference into a ‘graph genome’ to improve accuracy and reduce allelic bias. While adding a variant has the positive effect of removing an undesirable alignment-score penalty, it can also have negative effects, such as increasing the ambiguity and repetitiveness of the reference genome and increasing the cost of storing and querying the reference index.

We introduce new methods and software for prioritizing genetic variants according to their positive and negative effects on alignment accuracy. We propose simple, efficient models for these effects and show that analysis objectives can be optimized by prioritizing variants suggested by the models. We also explore the effectiveness of the graph genome on various genomic regions and read types and assess the relative benefit of a population-specific vs a global set of variants. These methods are implemented in FORGe, a toolkit designed for finding the optimal graph genome. Users provide a linear reference genome and set of variants, and FORGe constructs the optimal graph genome according to specified accuracy and blowup constraints.

Interestingly, we find that when including around 10% of the 84 million variants from the 1000 Genomes project, we reach a peak in increased alignment accuracy while also reaching a point of diminishing returns in reduced allelic bias at HET sites. These results support the idea that the emerging graph alignment paradigm should be accompanied by careful analysis of the pros and cons of including genetic variants in the reference. In particular, FORGe, together with HISAT2, enables advantageous tradeoffs between accuracy and computational overhead.

MICROBIOMEDB: A WEB-BASED DATA-MINING PLATFORM FOR INTERROGATING MICROBIOME EXPERIMENTS

Jane A Pulman[1], Kathryn Crouch[2], Sufen Hu[3], Jessica C Kissinger[4]

[1]University of Liverpool, Center for Genomic Research, Liverpool, United Kingdom, [2]University of Glasgow, Wellcome Trust Centre for Molecular Parasitology, Glasgow, United Kingdom, [3]University of Pennsylvania, Biology, Philadelphia, PA, [4]University of Georgia, Institute of Bioinformatics, Athens, GA

MicrobiomeDB.org, a data discovery and analysis platform developed in collaboration between hostmicrobe.org and EuPathDB, allows researchers to explore microbiome datasets based on experimental metadata. Microbiology has been revolutionized by high-throughput sequencing allowing culture-independent profiling of complex microbial communities, complementing culture-based approaches. Microbiome experiments are often accompanied by sample attributes (metadata) such as information about the source from which the sample was derived, quantitative or qualitative biometrics from clinical studies, technical comments about sample processing and sequencing, respondent survey data, and much more. MicrobiomeDB allows for a better understanding of how such characteristics affect the structure and function of microbial communities.

Experimental datasets are loaded into MicrobiomeDB in the standard Biological Observation Matrix (.biom) format with taxonomic assignment based on the GreenGenes database and a custom MIxS-compliant ontology. By utilizing the infrastructure and user interface of EuPathDB, which allows users to construct in silico experiments using an intuitive graphical ‘strategy’ approach, datasets can be filtered based on any of the ontolology terms while at the same time visualizing the structure of the data including relationships between terms. Users can look at individual samples via a sample record page and download all metadata and data for individual samples or in bulk based on the search query results.

This powerful search interface feeds naturally into a set of interactive apps that enable users to statistically analyze the query results. The database has a bi-monthly release cycle with new apps in active development. The current release houses four R shiny tool apps:

1) Relative taxonomic abundance
2) Beta-diversity
3) Alpha-diversity
4) Difference in taxonomic abundance

The data and any visualizations resulting from the analyses can be downloaded by the user. As with all EuPathDB databases, strategies and queries can be saved and shared with colleagues and collaborators.

Currently, MicrobiomeDB is a 'first-pass' example of microbiome data mining. We envision significantly expanding our pipeline to include loading additional 16S rRNA databases, metadata that describe taxa (i.e. basic microbiological properties), as well as bacterial metabolic pathway databases (e.g. KEGG), and much more.

Presented on behalf of the EuPathDB and MicrobiomeDB teams.

← Back

ABSENCE OF RECEPTOR FOR HYALURANON-MEDIATED MOTILITY (RHAMM) ALTERS GENOME-WIDE MUTATIONAL LANDSCAPES ASSOCIATED WITH TUMORIGENESIS AND METASTASIS.

Freda W Qi[1], Cornelia Toelg[2], Bin Luo[3], Maja Milojevic[1], Majorie Elizabeth Osborne Locke[4], Mark Daley[4], Charmaine B Dean[3], Reg Kulperger[3], Eva Turley[2], Kathleen A Hill[1,4]

[1]The University of Western Ontario, Biology, London, Canada, [2]London Health Sciences Centre, Victoria Hospital, London Regional Cancer Program, London, Canada, [3]The University of Western Ontario, Statistical and Actuarial Sciences, London, Canada, [4]The University of Western Ontario, Computer Science, London, Canada

The dynamic interactions in genetic and microenvironment changes with tumorigenesis and metastasis are poorly understood. Here, we use single nucleotide polymorphic (SNP) loci as sentinels to investigate mutagenesis arising in the MMTV-PyMT transgenic mouse model of primary mammary tumorigenesis and lung metastasis. Microenvironment differences are achieved by knockout of Receptor for Hyaluronan-mediated Motility (RHAMM). Phenotypically, RHAMM loss is associated with unchanged tumorigenesis but more aggressive metastasis. We used the Mouse Diversity Genotyping Array to examine *de novo* mutations at 220,615 SNP loci in the primary tumor and metastatic tissue of three *Rhamm*$^{+/+}$ and three *Rhamm*$^{-/-}$ mice. Based on the number of unique mutations in each tissue, *Rhamm*$^{+/+}$ mice showed higher inter-animal variation, whereas *Rhamm*$^{-/-}$ mice showed greater inter-animal homogeneity. The RHAMM loss microenvironment produced a 13.8-fold increase in the number of unique mutations in the metastatic tissue compared to the primary tumor, consistent with the more aggressive metastatic phenotype observed in *Rhamm*$^{-/-}$ mice. We identified 342 mutations in 124 candidate genes that may be underlying aggressive metastasis in *Rhamm*$^{-/-}$ mice. Of the mutated genes, four were growth factor receptors identified as cancer driver genes (*Kdr, Fgfr2, Ppp2r5c, and Ntrk2*) that act in the same cell growth and division pathway, and interact with the hub molecule ERK1,2. We also investigated the spatial distribution of mutations using a nonparametric statistical test to identify potential contributing mutational mechanisms. Clustering of *de novo* mutations was observed in at least five chromosomes in the metastatic tissue of each *Rhamm*$^{-/-}$ mouse, whereas mutations in the primary tumors did not deviate from spatial randomness. Heterozygous SNPs, which have previously been associated with higher mutation rates, were generally not spatially associated with *de novo* copy number variants in both *Rhamm*$^{+/+}$ and *Rhamm*$^{-/-}$ mice. However, a 2-fold increase in proximal associations was observed in *Rhamm*$^{-/-}$ mice. In conclusion, altered microenvironments arising with RHAMM loss elevate *de novo* mutations in metastatic tissue and elevated heterozygosity was not associated with structural mutations.

← Back

A New Method Of Hisone-Modification Guided Genome Assembly

Meifang Qi, Zijuan Li, Luhuan Ye, Yijing Zhang

Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

De novo assembly of large genomes is a very difficult task, such as assembly for wheat. In genome research, we usually only consider the sequences of promoter and gene body. Here we utilize the feature of histone modification which usually located in the promoter and gene body to assembly the sequences of promoter and gene body to get the core genomes. We find that combine H3K27me3 and H3K4me3 in wheat can get a relatively good result, which recovers more than 70% genes. And we find that genes with high expression and dramatic changes in expression have better results. And in different ecotypes, we find that assembly is relatively stable, and we can use this method to compare SNPs within a population.

← Back

MANAGING THE ANALYSIS OF HIGH-THROUGHPUT SEQUENCING DATA

Javier Quilez[1,2], Enrique Vidal[1,2], Francois Le Dily[1,2], Francois Serra[1,2,3], Yasmina Cuartero[1,2,3], Ralph Stadhouders[1,2], Thomas Graf[1,2], Marc A Marti-Renom[1,2,3,4], Miguel Beato[1,2], Guillaume Filion[1,2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science, Gene Regulation, Stem Cells and Cancer Program, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), NA, Barcelona, Spain, [3]CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), NA, Barcelona, Spain, [4]ICREA, NA, Barcelona, Spain

In the last decade we have witnessed a tremendous rise in sequencing throughput as well as an increasing number of genomic assays based on high-throughput sequencing (HTS). As a result, the management and analysis of the growing amount of sequencing data present several challenges with consequences for the cost, the quality and the reproducibility of research. Most common issues include poor description and ambiguous identification of samples, lack of a systematic data organization, absence of automated analysis pipelines and lack of tools aiding the interpretation of the results. To address these problems, we suggest to structure HTS data management by automating the quality control of the raw data, establishing metadata collection and sample identification systems and organizing the HTS data with an human-friendly hierarchy. We insist on reducing metadata field entries to multiple choices instead of free text, and on implementing a future-proof organization of the data on storage. These actions further enable the automation of the analysis and the deployment of web applications to facilitate data interpretation. Finally, a comprehensive documentation of the procedures applied to HTS data is fundamental for reproducibility. To illustrate how these recommendations can be implemented we present a didactic dataset. This work seeks to clearly define a set of best-practices for managing the analysis of HTS data and provides a quick start guide for implementing them into any sequencing project.

← Back

BENCHMARKING RNA-SEQ IN PLANT SPECIES

Srividya Ramakrishnan[1], Fritz Sedlazeck[3], Michael Schatz[1,2]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Cold Spring Harbor Laboratory, Quantitative biology, Cold Spring Harbor, NY, [3]Baylor College of Medicine, Houston, TX

High-throughput RNA sequencing is well-established as a versatile platform with the potential to study the complex transcriptomics of many organisms across the tree of life. Most RNA-Seq experiments aim to build a catalogue of species transcripts, assemble them and quantify the changing expression levels of transcripts during different stages of development or under different conditions. Despite the fact that a large number of mapping algorithms have been developed for RNA-seq read mapping, transcriptome assembly and differential analysis in recent years, accurate alignment of RNA-seq reads is still challenging and yet unsolved problem because of exon-exon spanning junction reads, relatively short read lengths and the ambiguity of multiple-mapping reads. In addition to these inherent complexities, RNA-seq within plant species is especially problematic for a variety of reasons. One of the biggest challenges is plant gene annotations which are used during the aligning step and the assembly step, are far from complete and the annotations of many genes that are particularly unique to plants, are still of poor quality. Furthermore, the genomes of many plant species are lower quality draft sequences with gene families collapsed or entirely missing. These factors significantly impact the downstream gene expression estimates in plants. In this study, we evaluate the performance of several of the most popular RNA-Seq aligners, including Tophat2, HiSat2 and STAR and the quantification of alignments from these aligners using Cufflinks and StringTie on both real and simulated RNA-Seq data from plants. Alternatively we also compare alignment independent quantification algorithms like Salmon, Sailfish and kallisto on several plant datasets. Specifically we assess these tool’s performance on gene abundance estimation and differential gene expression based on the presence/absence of complete gene annotations, mutated reference genome and propose a comprehensive RNA-Seq pipeline for the different plant RNA-Seq experiments.

← Back

IDENTIFICATION OF DNA METHYLATION DRIVEN TRANSCRIPTOMIC ALTERATIONS AND CLINICAL OUTCOMES IN AFRICAN AMERICAN MEN WITH PROSTATE CANCER

Swathi Ramakrishnan, Xuan Peng, Qianya Qi, Qiang Hu, Gissou Azabdaftari, Elena Pop, James Mohler, Kristopher Attwood, Li Yan, Jianmin Wang, Anna Woloszynska-Read

Roswell Park Cancer Institute, Buffalo, NY

Prostate cancer health disparities between African American (AA) and European American (EA) men are attributed to socioeconomic as well as biological differences between the two groups. Despite lower frequencies of genetic mutations, copy number alterations and TMPRSS2-ERG fusions, AA men are diagnosed with more aggressive prostate cancer than EA men. The purpose of our study is to investigate how epigenetic modifications, specifically DNA methylation, contribute to aggressive prostate cancer in AA men. To address this, we conducted Infinium DNA methylation array and RNA sequencing on 12 radical prostatectomy tumors from AA men treated at Roswell Park Cancer Institute (RPCI). Integrative MethylMix approach was used to identify genes whose expression may be regulated by DNA methylation. To increase analytical power, we included additional AA (n=22)/EA (n=5) men treated at RPCI and AA (n=22)/EA (n=172) patients from The Cancer Genome Atlas (TCGA). Gleason score <3+4 and >4+3 were classified as low and high aggressive tumors, respectively. MethylMix revealed three genes potentially negatively regulated by DNA methylation: ANGPTL4, GALNT5 and ZNF750. TCGA data confirmed negative correlation between DNA methylation and gene expression for all three genes and revealed lower mRNA z-scores for all three genes in AA men as compared to EA men. We further examined if global alterations in DNA methylation alone segregates tumors based on clinicopathological parameters. Unsupervised hierarchical clustering on combined RPCI and TCGA cohorts grouped tumors into several DNA methylation clusters. We focused on 2 largest clusters, Cluster 1 (n=133) and Cluster 3 (n=73). Cluster 1 consisted predominantly of low aggressive disease (p=0.00002) compared to Cluster 3. Following this trend, AA but not EA patients, in Cluster 1 had better overall survival (57 vs. 50 months, p=0.48) and disease free time (47 vs. 22 months, p=0.01) compared to Cluster 3. Our results suggest that DNA methylation potentially contributes to differential gene expression of ANGPTL4, a negative regulator of invasion, GALNT5 that glycosylates mucin-family proteins, and ZNF750, involved in terminal differentiation, in EA and AA prostate cancer patients. DNA methylation of these genes will be analyzed in EA and AA prostate cancer samples with functional validation in in vitro prostate cancer models. Based on the clinical outcomes, our results suggest that altered DNA methylation can contribute to disease aggressiveness in AA men.

← Back

ALTERNATIVE SPLICING OF NEUROFIBROMIN 1 IS ASSOCIATED WITH ELEVATED MAPK ACTIVITY AND POOR PROGNOSIS IN GLIOMA

Robert Siddaway[1], Arun K Ramani[2], Man Yu[1], Michael Brudno[2], Cynthia Hawkins[1]

[1]Hospital for SickKids, Cell Biology, Toronto, Canada, [2]Hospital for SickKids, Genetics and Genome Biology, Toronto, Canada

High-grade gliomas (HGG) are invasive with poor prognosis regardless of age: diffuse intrinsic pontine gliomas (DIPG), arising in the brainstem, are almost universally fatal and the leading cause of brain-tumor death in children; while adult anaplastic astrocytoma and glioblastoma multiforme (GBM) have median survivals of 1-3 years. The mutational spectra of adult and pediatric HGG differ, with pediatric tumours containing recurrent mutations of H3F3A and HIST1H3B. However, alterations leading to RAS/MAPK/PI3K pathway activation, including PDGFRA amplification, EGFRvIII, BRAF-V600E, NF1 deletion, are frequently found, although not all tumours will have mutations in this pathway. The neurofibromin 1 (NF1) gene negatively regulates RAS signalling by stimulating RAS-GTP turnover, thereby leading to RAS-inactivation. The two major isoforms, NF1-I and NF1-II, differ only by inclusion of the 21 aa exon23a in the GAP-related domain of NF1-II. Exon23a-inclusion has been shown to render NF1 10 times less active towards RAS, leading to elevated MAPK signalling. The brain expresses predominantly NF1-I, while the major isoform elsewhere is NF1-II. Here we used RNA-Seq to identify genes alternatively spliced between DIPG and normal brain, identifying an isoform switch from NF1-I in normal brain to NF1-II; we additionally found the same isoform switch in the TCGA adult GBM and LGG cohorts. For both GBM and LGG, RAS/MAPK/PI3K wild-type tumors with elevated NF1-II conferred significantly reduced patient survival compared to RAS/MAPK/PI3K mutant tumors. NF1-exon23a inclusion is known to be repressed in cell model systems by the CELF and ELAV-like families of splice regulators. We further show that members of these gene families are downregulated in HGG. Together, our results indicate a novel mechanism by which gliomas can activate signaling downstream from RAS independent of mutations and tumor grade, which promotes tumorigenesis by regulating pathways such as proliferation and invasion.

← Back

DISCOVERY : A FAST AND LIGHTWEIGHT METHOD TO ISOLATE Y CHROMOSOME-SPECIFIC SEQUENCES

Samarth Rangavittal, Marta Tomaszkiewicz, Kateryna Makova, Paul Medvedev

Pennsylvania State University, University Park, PA

The Y chromosome plays an important role in sex determination and male fertility, but low coverage due to its haploid nature and the presence of large palindromic repeats complicate Y-assembly [1]. Techniques such as Sanger sequencing of bacterial artificial chromosomes - used to assemble human and chimpanzee Y [2,3], and enrichment by flow sorting - used to assemble the gorilla Y [4], are either expensive or technically complicated methods that remain inaccessible to many laboratories. With the advent of relatively inexpensive whole-genome sequencing by Illumina, PacBio and 10x Genomics, it has become possible to use whole-genome sequencing of a male to assemble the Y chromosome. However, the identification of Y contigs from a whole-genome assembly remains a challenge. Existing approaches have demonstrated that it is possible to leverage female genome references, low-coverage female sequence data, or depth of coverage of male sequencing data to isolate Y contigs. One such method is the Y-Genome Scan (YGS) [5], which uses k-mer proportion sharing with the female to isolate Y-contigs. This approach has been tested on human simulated data and real Drosophila contigs.

We build on the YGS method by using Bloom Filters and k-mer set subtraction for space and memory gains. We tested our method (called DiscoverY) by successfully isolating 23.6 Mb of Y-contigs (out of expected 27 Mb) on a simulated human dataset. Subsequently, we used DiscoverY on a novel great ape dataset generated in our lab. From an assembly of bonobo male whole-genome Illumina data, we isolated 11.9 Mb of Y chromosome in 692 contigs with an N50 of 43.8 kb. On average, DiscoverY completes within 5 hours of runtime and 4 GB of memory usage on a desktop computer. The assembly of such novel Y chromosomes will enable a broader analysis of Y chromosome evolution in great apes.

References:

[1] M. Tomaszkiewicz, P. Medvedev, K.D. Makova, Y and W chromosome assemblies: approaches and discoveries. Trends Genet. Apr;33(4):266-282. (2017)
[2] H. Skaletsky, et al., The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423, 825–837 (2003)
[3] J.F. Hughes, et al., Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.Nature 463, 536–539 (2010)
[4] M. Tomaszkiewicz, S. Rangavittal, M. Cechova, et al., A time- and cost-effective strategy to sequence mammalian Y chromosomes: an application to the de novo assembly of gorilla Y. Genome Res. 26, 530–540 (2016)
[5] A.B. Carvalho, A.G. Clark, Efficient identification of Y chromosome sequences in the human and Drosophila genomes. Genome Res. 23, 1894-1907 (2013)

← Back

AN ALGORITHM FOR CELLULAR REPROGRAMMING

Scott Ronquist[1], Geoff Patterson[2], Lindsey Muir[3], Stephen Lindsly[1], Haiming Chen[1], Markus Brown[4], Max Wicha[5], Anthony Bloch[6], Roger Brockett[7], Indika Rajapakse[1,6]

[1]University of Michigan, Bioinformatics, Ann Arbor, MI, [2]IXL Learning, Raleigh, NC, [3]University of Michigan, Pediatrics and Communicable Diseases, Ann Arbor, MI, [4]University of Maryland, Biological Sciences, College Park, MD, [5]University of Michigan, Hematology/Oncology, Ann Arbor, MI, [6]University of Michigan, Mathematics, Ann Arbor, MI, [7]Harvard, School of Engineering and Applied Sciences, Cambridge, MA

The day we understand the time evolution of subcellular events at a level of detail comparable to physical systems governed by Newton's laws of motion seems far away. Even so, quantitative approaches to cellular dynamics add to our understanding of cell biology. With data-guided frameworks we can develop better predictions about, and methods for, control over specific biological processes and system-wide cell behavior. Here we describe an approach for optimizing the use of transcription factors (TFs) in cellular reprogramming, based on a device commonly used in optimal control. We construct an approximate model for the natural evolution of a cell cycle-synchronized population of human fibroblasts, based on data obtained by sampling the expression of 22,083 genes at several time points during the cell cycle. In order to arrive at a model of moderate complexity, we cluster gene expression based on division of the genome into topologically associating domains (TADs) and then model the dynamics of TAD expression levels. Based on this dynamical model and additional data, such as known TF binding sites and activity, we develop a methodology for identifying the top TF candidates for a specific cellular reprogramming task. Our data-guided methodology identifies a number of TFs previously validated for reprogramming and/or natural differentiation and predicts some new potentially useful combinations of TFs. Our findings highlight the immense potential of dynamical models, mathematics, and data-guided methodologies for improving strategies for control over biological processes.

SINGLE-CELL RNA SEQUENCING IN SPERM FROM FATHERS OF AUTISTIC CHILDREN

Delia Tomoiaga[4], Jonathan Foox[4], Shristi Shrestha[3], Shawn Levy[3], Christopher E Mason[4], Jeffrey A Rosenfeld[1,2]

[1]Rutgers Cancer Institute of NJ, Cancer, New Brunswick, NJ, [2]Robert Wood Johnson Medical School, Pathology, New Brunswick, NJ, [3]Hudson Alpha Institute for Biotechnology,, Hudson Alpha, Huntsville, AL, [4]Weill Cornell Medical College, Physiology, New York, NY

The incidence of autism has been shown by numerous studies to be correlated with an increase in the age of the father at the time of conception. This connection is thought to be related to the increase in de novo mutations or copy number variations in sperm. In addition to autism, there are many other disorders associated with advanced paternal age that are collectively known as paternal age effect (PAE) disorders. For disorders, such as Apert syndrome, the causative variant is found in a small percentage of sperm, but counterintuitively, the sperm that cause the disorder have a selective advantage for fertilization. Less than 1% of the sperm will have the causative mutation and therefore, a large number of sperm would need to be tested to screen for the presence of mutated sperm or dis-regulated genes. There is currently no comprehensive method to test individual sperm for the presence of mutations related to PAE disorders. The current guidance from clinical geneticists is for patients to avoid having children when the male is older, but there is no clear way to screen out those younger men who could have mutations or dysregulated genes in their sperm.

Until recently, the only way to look for rare mutations in sperm would be to take a bulk sample of cells, sequence them to a very high depth and then to look for variants identified by a very small number of reads. This type of assay is expensive, and it is difficult to differentiate the very low frequency variants from those that are a result of a sequencing errors or allelic drop-out (ADO). This problem has been alleviated with the introduction of systems such as the 10X Genomics Chromium instrument. Using a per-cell, molecular barcoding scheme and the separation of individual cells, here we show that the Chromium system has enabled the direct sequencing of single sperm cells' RNA.

Using the Chromium system, we have examined the individual sperm of 8 donors (5 normal, 3 autistic) spanned over 240,000 individual sperm cells in the donors and patients and determined that germline-specific gene expression patterns are specifically enriched in these samples, including distinct profiles showing a shift of spermatogonia stem cell profiles (PRM1, VGL, TIGAR, DAG1, CRISP2, ATF). Moreover, we observe significant differences in the developmental pathways between the cases and controls, indicating this may be a method to reveal baseline functional genomics differences in sperm donors. Ongoing work in a larger replication cohort of donors (n=100) can open these methods as a potential screen for sperm to determine which patients are at a higher risk of having a child affected with autism or another PAE disorder.

← Back

POPDEL: POPULATION-SCALE DETECTION OF GENOMIC DELETIONS

Sebastian Roskosch[1], Bjarni V Halldórsson[2,3], Birte Kehr[1]

[1]Berlin Institute of Health, JRG Genome Informatics, Berlin, Germany, [2]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland, [3]Reykjavik University, School of Science and Engineering, Reykjavik, Iceland

An untapped potential for the detection of structural variants (SVs), which are major contributors to genetic variation with a notable impact on phenotypes, lies within the utilization of high sample numbers in a single calling step. Calling of SVs, as compared to SNVs and indels, requires tailored approaches because small variant callers as *GATK* are not designed for this purpose. Here we focus on one of the most common and basic SV classes, genomic deletions. Various approaches have been developed for detecting deletions and other SVs. They have been implemented in tools like *Delly*, *Lumpy* or *Breakdancer* to name only a few, but most existing methods do not yet harness the potential lying in the growing number of available samples. We are developing *PopDel*, a light-weight tool for fast and accurate detection of deletions ranging from a few hundred to ten-thousands of base-pairs in a single- or multi-sample joint calling approach on short-read paired-end data. While *PopDel* can run on single samples, it is designed to exploit a high number (i.e. thousands) of samples for improved precision and recall while maintaining good scaling of running time and memory requirements. It applies a read-pair-distance-based approach on all samples simultaneously for calculating the most likely positions and sizes of deletions found in any of the samples. The minimum size of deletions found by *PopDel* does not rely on hard-thresholds. Instead, *PopDel* computes a likelihood ratio based on the empirical insert size distribution and iteratively estimated deletion lengths. *PopDel* does not repeatedly read the whole BAM files but it preprocesses the mappings once and stores only the most relevant information in small profile files, which are used in place of the BAM files for all subsequent steps. This strategy reduces the I/O overhead – a crucial point when operating on a compute cluster. These features give *PopDel* a competitive edge when compared to existing tools. To show its suitability for these tasks, we are applying *PopDel* on Icelandic data generated by deCODE Genetics. Preliminary results suggest that a tool like *PopDel* can be a valuable asset for saving computational resources and improving the quality of the calls, especially for the increasingly large variant calling efforts performed by, for example, deCODE Genetics or Genomics England. In the future, *PopDel* will greatly simplify the inclusion of deletions in population-scale studies that search for variants associated with disease.

TRACKATURE: RECONSTRUCTING MUTATIONAL SIGNATURES THROUGH TIME TO TRACK TUMOUR EVOLUTION

Yulia Rubanova[1], Roujia Li[2], Jeff Wintersinger[1], Amit Deshwar[3], Nil Sahin[4], Quaid Morris[1,2,3,4]

[1]University of Toronto, Department of Computer Science, Toronto, Canada, [2]Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada, [3]University of Toronto, Department of Electrical and Computer Engineering, Toronto, Canada, [4]University of Toronto, Department of Molecular Genetics, Toronto, Canada

Mutations in cancer arise from many sources, including both environmental factors (e.g., smoking and exposure to UV light) and internal ones (e.g., DNA replication and damage-repair errors). Some of these sources exhibit characteristic mutational signatures, consisting of a unique distribution over types of mutations. The contribution of a signature to a cancer's mutational burden is deemed the signature exposure. Changes in signature exposures may alter a patient's prognosis and responsiveness to treatment. Our goal is to find changepoints in signature exposures that mark these potential evolutionary transitions.

We have developed Trackature, a method to reconstruct signature exposure trajectories over time from single cancer samples. We sort mutations by relative time of occurrence based on variant allele frequency and divide them into bins of four hundred. Each bin corresponds to one time point. We assign each mutation within a timepoint to one of 96 mutation types and fit the mutational signatures to those counts using mixture of multinomials. Derived mixture coefficients correspond to exposures of the mutational signatures in each bin. To find changepoints in signature trajectories, we use a greedy search algorithm, then use the Bayesian Information Criterion to determine the optimal number of changepoints. Finally, we evaluate uncertainty by bootstrapping sets of mutations and recomputing exposure estimates.

We applied our approach to 2435 whole-genome sequencing samples of 40 cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. Of these samples, 69% have an absolute, additive change in exposures >10%. Esophageal adenocarcinoma (ESAD), bone leiomyosarcoma (SARC), and lung squamous cell carcinoma (LUSC) have the greatest fraction of samples with exposure changes >10% (96.9% of ESAD, 93.9% of SARC, 93.6% of LUSC).

We see consistent trends in signature activity. Signature 1 (correlated with patient age) contributes a larger fraction of mutations early in cancers' evolutionary histories. This same trend occurs with carcinogen-related signatures, including signature 4 (smoking) and signature 7 (exposure to UV light). Signature 3 (associated with BRCA1 and BRCA2 mutations) occurs early in breast cancer but late in ovarian.

In addition, we examined whether the emergence of subclonal cancer populations corresponded to changes in signature exposures, which would suggest that signature changes help drive cancer development. We observe that 26% of clonal-subclonal transitions and 41% of subclonal-subclonal transitions coincide with exposure boundaries. Increasing subclonal variegation is also associated with greater exposure change.

AN R CLIENT FOR THE CANCER IMAGING ARCHIVE REST API

Pamela Russell[1], Kelly Fountain[2], Dulcy Wolverton[2], Debashis Ghosh[1]

[1]Colorado School of Public Health, Department of Biostatistics and Informatics, Aurora, CO, [2]University of Colorado School of Medicine, Department of Radiology, Aurora, CO

The Cancer Imaging Archive (TCIA) hosts de-identified medical images of cancer available for public download. TCIA includes 72 collections of images, typically organized by cancer type, with a total of almost 35,000 patients. In particular, there are 21 collections comprising 1,777 patients whose genomic data are also available in The Cancer Genome Atlas (TCGA). A variety of imaging modalities are represented. TCIA also includes rich metadata and clinical information for each set of images.

TCIA provides a web interface as well as a REST API for programmatic access to the data. Currently, developers using the REST API must understand the structure of API endpoints, construct HTTP requests, and parse the responses. We present TCIApathfinder, the first R client for the TCIA REST API. TCIApathfinder wraps API functionality in simple functions that can be easily incorporated into scripts by users familiar with R. Downloaded DICOM images can be processed with the existing oro.dicom package followed by appropriate image processing packages. TCIApathfinder will soon be available as a package on CRAN.

MACHINE LEARNING AND COMPUTER VISION APPROACHES FOR PHENOTYPIC PROFILING IN YEAST

Nil Sahin[1,2], Mojca Mattiazzi-Usaj[2], Erin Styles[2], Charlie Boone[1,2,3], Brenda Andrews[1,2,3], Quaid Morris[1,2,4]

[1]University of Toronto, Department of Molecular Genetics, Toronto, ON, Canada, [2]University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, Toronto, ON, Canada, [3]University of Toronto, Banting and Best Department of Medical Research, Toronto, ON, Canada, [4]University of Toronto, Department of Computer Science, Toronto, ON, Canada

A powerful method to study the genotype-to-phenotype relationship is the systematic assessment of subcellular phenotypes using high-content screening, and coupling it with high-throughput microscopy can increase the depth of information about the defects occurring inside the cell that are missed by fitness-based assays. In our labs, we imaged genome-wide genetic perturbations on subcellular compartments to identify many possible morphological defects. In order to answer biological questions via studying these massive generated datasets, computational approaches such as image recognition, feature extraction and machine learning can provide quantitative and reproducible results. My overarching goal is to develop new computational methods for cellular image analysis in order to explore a comprehensive list of subcellular morphological defects on both cellular and population levels in response to thousands of genetic perturbations using Synthetic Genetic Array (SGA) analysis in *Saccharomyces cerevisiae*. Here, I describe analyses for calculating the penetrance of gene defects in a population level through single cell measurements and automatically identifying mutant phenotypes using machine learning. While the central players of many important biological processes have been discovered, there remain numerous gaps in our understanding of the regulation of cellular morphology. This analysis will allow for the identification of connections between discrete biological processes, the prediction of novel gene function, and the generation of a clearer understanding of basic eukaryotic cell biology.

← Back

THE RAINFALL PLOT: ITS MOTIVATION, CHARACTERISTICS AND PITFALLS

Diana Domanska[1], Daniel Vodak[4], Christin Lund-Andersen[4], Stefania Salvatore[1], Eivind Hovig[1,2,3,4], Geir Kjetil Sandve[1]

[1]University of Oslo, Department of Informatics, Oslo, Norway, [2]Statistics for Innovation, Norwegian Computing Center, Oslo, Norway, [3]Oslo University Hospital, Institute for Medical Informatics, Oslo, Norway, [4]Oslo University Hospital, Department of Tumor Biology, Oslo, Norway

Background: A visualisation referred to as rainfall plot has recently gained popularity in genome data analysis. The plot is mostly used for illustrating the distribution of somatic cancer mutations along a reference genome, typically aiming to identify mutation hotspots. In general terms, the rainfall plot can be seen as a scatter plot showing the location of events on the x-axis versus the distance between consecutive events on the y-axis. Despite its frequent use, the motivation for applying this particular visualisation and the appropriateness of its usage have never been critically addressed in detail.

Results: We show that the rainfall plot allows visual detection even for events occurring at high frequency over very short distances. In addition, event clustering at multiple scales may be detected as distinct horizontal bands in rainfall plots. At the same time, due to the limited size of standard figures, rainfall plots might suffer from inability to distinguish overlapping events, especially when multiple datasets are plotted in the same figure. We demonstrate the consequences of plot congestion, which results in obscured visual data interpretations.

Conclusions: This work provides the first comprehensive survey of the characteristics and proper usage of rainfall plots. We find that the rainfall plot is able to convey a large amount of information without any need for parameterisation or tuning. However, we also demonstrate how plot congestion and the use of a logarithmic y-axis may result in obscured visual data interpretations. To aid the productive utilisation of rainfall plots, we demonstrate their characteristics and potential pitfalls using both simulated and real data, and provide a set of practical guidelines for their proper interpretation and usage.

HIERARCHICAL GSUITE HYPERBROWSER: ANALYSIS ACROSS MULTIPLE DIMENSIONS OF EPIGENOMIC VARIATION

Diana Domanska[1], Chakravarthi Kanduri[1], Stefania Salvatore[1], Boris Simovski[1], Florian Krull[2], Sveinung Gunderson[1], Eivind Hovig[1,3,4,5], Geir K Sandve[1]

[1]University of Oslo, Department of Informatics, Oslo, Norway, [2]Norment part University of Oslo, Faculty of Medicine, Oslo, Norway, [3]Institute for Cancer Research, Oslo University Hospital, Dept. of Tumor Biology, Oslo, Norway, [4]Statistics For Innovation, Norwegian Computing Center, Oslo, Norway, [5]Oslo University Hospital, Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo, Norway

We have recently developed GSuite HyperBrowser, the first comprehensive solution for integrative analysis of dataset collections across the genome and epigenome. The GSuite HyperBrowser is an open-source system for streamlined acquisition and customizable statistical analysis of large collections of genome-wide datasets. The system is based on new computational and statistical methodologies that permit comparative and confirmatory analyses across multiple disparate data sources.
Looking ahead, with a variety of omics data becoming available for ever more conditions, flat collections can not reflect the myriad dimensions of cellular state in a satisfactory manner. There will be an increasing need for data representation and analytical methodology that can incorporate variation across multiple dimensions, such as individuals, tissues, cell types and cell conditions. We are currently pursuing a variety of ideas on how to constructively approach such structured collections of omics datasets.

THE SIMILARITY DISTRIBUTION OF GENE PAIRS CREATED BY RECURRENT ALTERNATION OF WHOLE GENOME DUPLICATION AND FRACTIONATION IN PLANTS

David Sankoff, Yue Zhang, Chunfang Zheng

University of Ottawa, Mathematics and Statistics, Ottawa, Canada

We solve modeling and inference problems around the mechanism of fractionation, the genome-wide process of losing one gene per duplicate pair following whole genome doubling, tripling, etc. (WGD), motivated by the evolution of plants over many tens of millions of years, with their repeated cycles of genome doubling and fractionation. We focus on the frequency distribution of similarities between the two genes, over all the duplicate pairs in the genome. Our model is fully general, accounting for repeated duplication, triplication or other k-tupling events, as well as a general fractionation rate in any time period among multiple progeny of a single gene. It also has a biologically and combinatorially well-motivated way of handling the tendency for at least one sibling to survive fractionation. We show how the method reduces to models we previously proposed for special cases, and settles unresolved questions about the expected number of gene pairs tracing their ancestry back to each WGD event. The parameters of the model, such as event time and fractionation rates, may be inferred directly from the empirical distribution of duplicate pair similarities. Because of a trade-off between fractionation rates and multiplicity of WGD events: doubling versus tripling, etc,, however, in many instances the multiplicity of the various polyploidy events giving rise to WGD in the evolution of a genome cannot be determined uniquely, which is a severe problem for understanding its history. We propose a way to remedy this shortcoming by combining the syntenic approach pioneered ten years ago in the publication of the grapevine genome by Jaillon et al. with our model to to produce a method capable of estimating the multiplicity of the WGD events, as well as the fractionation parameters. We focus on two important instances, one where a hexaploidization (whole genome triplication) precedes a tetraploidization (whole genome duplication), and the other where the triplication follows the duplication. The non-uniqueness of ploidy inference is characterized mathematically and is illustrated with data from the turnip, or Napa cabbage (Brassica rapa) genome. Our method distinguishes whole genome triplication from whole genome duplication, given the distribution of duplicate gene similarity, and we apply this to confirm the known sequence of events in the ancestral history of this species.

PREDICTING PREFERENCES OF RNA BINDING PROTEINS FROM PROTEIN SEQUENCE

Alexander Sasse[1,2], Quaid D Morris[1,2,3]

[1]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [2]University of Toronto, Donnelly Centre, Toronto, Canada, [3]University of Toronto, Department of Computer Science, Toronto, Canada

RNA binding proteins (RBPs) play important roles in co- and post-transcriptional processes; influencing mRNA biogenesis, modification, stability, transport, and cellular localization (Glisovic, T., et al, 2008). RBP target sites can be identified from in vivo crosslinking immunoprecipitation and subsequent sequencing (CLIP-seq) or in vitro selection assays, such as RNAcompete (Ray, D., et al, 2013) or SELEX. To identify the underlying binding motifs, various algorithms have been developed and successfully applied (Alipanahi, B., et al, 2015). Although the binding preferences of many RBPs have been identified this way, it seems impossible that RBP binding preferences can be determined experimentally for all biologically studied organisms due to the expensive and time consuming nature of these experiments. However, studies investigating post-transcriptional regulation rely on knowing the binding sites for each RBP. Protein domain sequence identity can be used to infer motifs for RBPs lacking experimental data (Ray, D., et al, 2013). Other measures of sequence identity, such as amino acid 4-mer compositions, also seem to be sufficient for predicting binding specificity (Pelossof, R., et al, 2015).

To investigate the ability to infer binding motifs for uncharacterized RBPs, we have implemented and assessed different machine learning approaches, such as regularized linear regression and bi-crossvalidation, to predict binding preference similarity from protein sequences. A variety of representations of the protein sequence were used as input, along with in silico predicted structural features. The performance of each algorithm was assessed on a set of 207 RBPs of variable composition (different types, amounts, and combinations of RNA binding domains) with experimentally determined sequence motifs. We found that pairwise whole sequence alignments outperformed linear regression and bi-crossvalidation on this set. However, on a subset of 46 single RRMs bi-crossvalidation using gapped residue 4mers showed the best performance on average. Next, we plan to combine different methods to obtain optimal predictive performance. Moreover, we think that the learned parameters from our methods can be used to distinguish biologically meaningful patterns which determine motif specific protein-RNA interactions.

STRELKA2: FAST AND ACCURATE SMALL VARIANT CALLING FOR GERMLINE AND CANCER SEQUENCING APPLICATIONS

Sangtae Kim[1], Konrad Scheffler[1], Eunho Noh[1], Aaron L Halpern[1], Mitchell A Bekritsky[2], Peter Krusche[2], Christopher T Saunders[1]

[1]Illumina, Inc., Bioinformatics, San Diego, CA, [2]Illumina Cambridge Ltd., Bioinformatics, Little Chesterford, United Kingdom

We present Strelka2, a fast and accurate small variant caller for clinical germline and cancer sequencing applications. Strelka2 introduces rapid analysis of germline variation for small cohorts, and improves on Strelka's original analysis of somatic variation in tumor/normal sample pairs. The germline caller employs an efficient tiered haplotype model to improve accuracy and provide read-backed phasing, adaptively selecting between assembly and a faster alignment-based haplotyping approach at each variant locus. In addition, the germline caller incorporates a novel mixture model for indel error estimation from the input sequencing data, improving robustness to indel noise without requiring any prior sample variation data. The somatic calling model is improved for liquid and late-stage tumor analysis by accounting for possible tumor cell contamination in the normal sample. A final empirical variant rescoring step using random forest models trained on various call quality features has been added to both callers to further improve precision.

Strelka2 germline calling results were compared with submissions to the recent PrecisonFDA challenge, showing an average indel F-score 1.6% higher than the best challenge submission, while giving comparable results for SNVs. These results are notable in that all Strelka2 results are generated with a default configuration, whereas the PrecisionFDA challenge submissions are acquired using customized pipelines. To assess runtime, we benchmarked Strelka2 against the Sentieon DNAseq Haplotyper pipeline, a fast reimplementation of the GATK HaplotypeCaller. For the 4 PrecisionFDA datasets, Strelka2 was on average 2.2x faster than the Sentieon pipeline, while also outperforming it in accuracy, with an average F-score improvement of 2.1% for indels and 0.27% for SNVs (Genome in a Bottle v3.3.2).

Strelka2's somatic calling was assessed using mixtures of unrelated individuals (NA12878 and NA12877 as tumor and normal, respectively) to simulate differing purity levels. We generated 3 such datasets representing different levels of tumor purity (20, 50 and 80%) and 1 representing a liquid tumor sample with 90% normal purity. Performance was evaluated against NA12878 variants where the corresponding NA12877 genotype is homozygous reference. We compared Strelka2 with the Sentieon TNhaplotyper pipeline (a fast reimplementation of Mutect2) using these in-silico mixtures as the tumor sample and a separate NA12877 dataset as the normal sample. We found that Strelka2 calls had substantially higher F-scores in every dataset, with an average F-score improvement over Sentieon of 28% for SNVs and 34% for indels, while also completing 3.2x faster on average.
Strelka2 is freely available under the GNU General Public License v3.0 at https://github.com/Illumina/strelka.

BEHIND THE VEIL: USING VISUALIZATION TOOLS TO EXAMINE GENOME CURATION

Valerie A Schneider[1], Anatoliy Kuznetsov[1], Victor Ananiev[1], Eric Weitz[1], Terence D Murphy[1], Kerstin Howe[2], Tina Graves-Lindsay[3], Paul Flicek[4], Paul A Kitts[1]

[1]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, [2]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [3]The McDonnell Genome Institute, Washington University, St. Louis, MO, [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

Visualization tools such as genome browsers are typically thought of as vehicles that facilitate the evaluation of our accumulated knowledge of the genome. Images of annotation tracks containing information about genes and regulatory regions are common features in scientific publications. However, annotation qualities are dependent upon the characteristics of the underlying assembly, which is itself a data construct warranting its own critical evaluation. Fortunately, visualization tools are also an effective means of appraising the genomic substrate upon which annotations are placed, exposing regions in need of and under curation, and offering insights into global and local assembly quality.

As a consequence of its central role in data management for the Genome Reference Consortium (GRC), the group responsible for curating the high-quality reference assemblies for human, mouse and other organisms, NCBI has emphasized the development of tools that facilitate the visualization of genomic data curation. These include specialized features of Genome Data Viewer (GDV), its browser for RefSeq annotated assemblies, the Assembly Support track set that promotes GDV-based evaluation of assembly structure, as well as several graphical displays on the GRC website for the exploration of genome regions under curation. We will present examples and recent developments in each of these that illustrate the value and importance of using visualization tools to assess the characteristics of the curated reference assembly as part of evaluating variations, gene annotations and even other genome assemblies.

We will also examine what these visualization tools reveal about ongoing efforts to curate the human reference genome assembly. GRCh38, the last coordinate changing update to the reference, was released in late 2013. Since the release, the GRC has continued its curation to correct errors and expand genomic diversity with non-coordinate changing patch releases. We will show how to use these tools to find regions that have undergone significant improvements as part of this ongoing work, as well as problematic assembly regions that remain largely unchanged since before the release of GRCh38, and how both contribute to considerations and timelines surrounding future coordinate changing assembly releases.

DEBACTER - HIGH-RESOLUTION DECONTAMINATION OF GENOMES USING DEEP SEQUENCING DATA

Gunnar Schulze[1], Eivind Valen[1,2], David Fredman[1]

[1]University of Bergen, Computational Biology Unit, Informatics Institute, Bergen, Norway, [2]University of Bergen, Sars International Centre for Marine Molecular Biology, Bergen, Norway

Genome sequencing datasets often contain traces of untargeted organisms from the sampling environment, commensal organisms or lab contaminants. Untargeted sequences, especially when integrated into an assembly, can cause severe problems in downstream analyses and ultimately lead to a pollution of reference databases with mis-assembled or mis-labelled sequences. Current *in-silico* methods for the decontamination of sequencing projects often still rely on the validity of existing reference databases to identify non-target sequences, are not suited to analyze genomes with a complex, heterogeneous sequence composition, or lack the resolution to detect mis-assembled, chimeric contigs. To overcome these issues, we have developed Debacter, a computational pipeline that exploits high-throughput sequencing datasets to detect contamination in both *de-novo* assemblies and established references sequences. In contrast to other methods, Debacter does not rely on previous sequence annotation from databases, is not affected by heterogeneous sequence composition and can discern between target and non-target sequences at the contig and sub-contig level. We applied Debacter to several genomes of model and non-model organisms and verified its ability to detect foreign sequences erroneously placed in reference assemblies.

ACCURATE AND FAST DETECTION OF COMPLEX AND NESTED STRUCTURAL VARIATIONS USING LONG READ TECHNOLOGIES.

Fritz J Sedlazeck[1], Philipp Rescheneder[2], Arndt von Heaseler[2], Michael C Schatz[3]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna, Vienna, Austria, [3]Johns Hopkins University, Department of Computer Science, Baltimore, MD

The impact of structural variations (SVs) is becoming more prominent within a variety of organisms and diseases, especially human cancers. Short-read sequencing has proved invaluable for recognizing copy number variations and other simple SVs, although has been highly limited for detecting most other SVs because of repetitive elements and other limitations of short reads. The advent of long-read technologies, such as PacBio or Nanopore sequencing that now routine produce reads over 10,000bp, offer a more powerful way to detect SVs. However, currently available methods often lack precision and sensitivity when working with highly erroneous reads, especially for more complex or nested SVs.

Here we present NGMLR and Sniffles, two novel methods to align long-read sequencing data and detect all types of SVs. NGMLR includes novel alignment and split read algorithms to more precisely align long reads. One of the key advances is the use of a convex gap scoring scheme to accurately align these noisy long reads. Subsequently, Sniffles finds SVs using split-read alignments as well as alignment events (e.g. indels). A unique feature of Sniffles is detecting nested SVs such as inversions flanked by deletions, which we now commonly detect in several samples.

Using real and simulated data, we demonstrate the enhanced ability of Sniffles to detect SVs over existing long-read methods (e.g. PBHoney) or short read methods (e.g. Lumpy, Delly, Manta) and find thousands of variants missed by short read approaches. We demonstrate its robustness with a Mendelian recall rate of 99.98% in Arabidopsis and similar in Human (GiaB) data. In addition, we measured a validation rate of 72% (Sniffles) vs. 30% based on short reads on a challenging breast cancer cell line, and enabling us to discover several novel gene fusions. Furthermore, we could identify severe biases in short read based SV calling especially for translocations (>80% FDR) and inversions (>50% FDR) due to variants in low complexity regions that we can now reliable detect using Sniffles.

Working with genuine PacBio and Nanopore reads with human cancer samples, healthy human samples, and other species, we show how Sniffles combined with NGMLR reduces the coverage, and therefore cost, required per sample for highly sensitive and specific SV detection. Sniffles and NGM-LR are available open-source at Github, and are already being used by multiple researchers around the world.

A COMPREHENSIVE PIPELINE FOR ANALYSIS OF COMPLEX-SETUP GENOME-SCALE AND FOCUSED POOLED CRISPR-SCREENS

Vitaly Sedlyarov[1], Ulrich Goldmann[1], Adrian Cesar-Razquin [1], Manuele Rebsamen[1], Enrico Girardi[1], Giulio Superti-Furga[1,2]

[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria, [2]Medical University of Vienna, Center for Physiology and Pharmacology, Vienna, Austria

Understanding gene functions is an important task of molecular biology. Clustered regularly interspaced short palindromic repeat (CRISPR)-Cas9 system provides increasing opportunities to study gene function in various cellular contexts and assays in a high-throughput manner. Loss-of-function and gain-of-function CRISPR-screens can be performed to respectively abolish or increase expression of genes. The complexity of the experimental setup of a screen can be markedly variable. While rather simple setups include only two conditions, more complex screens involve multiple conditions, reference points and time courses.

We established a versatile pipeline to analyze CRISPR-screens of different levels of complexity. It includes three major steps: (1) matching reads to the sgRNA library and counting sgRNA representation, (2) analysis of differential sgRNA representation between conditions using a linear model, (3) aggregation of sgRNA level analysis to corresponding genes. We implemented a fast and memory efficient algorithm in Python which performs matching of reads to the library and counts sgRNA frequency. At the second step, differential representation of sgRNAs is analyzed using DESeq2. To aggregate sgRNAs to gene level, we employed robust rank aggregation (RRA) as well as gene set enrichment analysis (GSEA).

We apply the pipeline for analysis of CRISPR-screens elucidating functions of solute carrier proteins (SLC) – a large family of transporter proteins. Despite a profound interest in specific members of the SLC family, functions and cargos for many members remain unknown. We have embarked on a broad survey of the potential involvement of SLC genes in immune cell function. In order to identify SLC transporters involved in regulation of phagocytosis, we introduced a SLC-focused pooled CRISPR-Cas9 library into human monocytic U937 cells. We performed a phagocytosis assay in phorbol myristate acetate (PMA)-differentiated U937 cells using dual-colored latex beads (pH-insensitive and pH-sensitive staining) to simultaneously track phagocytic uptake of particles and acidification of phagosomal content. Using the presented analysis pipeline, we identified the sodium bicarbonate cotransporter SLC4A7 to be absolutely required for the phagocytic process. Current studies are now focusing on elucidating the precise mechanism by which SLC4A7 may be regulating acidification of the phagosomal environment.

← Back

ASSEMBLY OF INDIVIDUAL CHROMOSOMES AT MULTI-MEGABASE SCALE USING LINKED-READS

Preyas N Shah[1], Neil I Weisenfeld[1], Vijay Kumar[1], Stephen R Williams[2], Claudia Catalanotti[2], Nikka Keivanfar[3], Deanna M Church[2], David B Jaffe[1]

[1]10x Genomics, Computational Biology, Pleasanton, CA, [2]10x Genomics, Applications, Pleasanton, CA, [3]10x Genomics, Molecular Biology, Pleasanton, CA

Routinely probing the true biology of individual genomes at population scale is necessary to unlock applications such as precision medicine, understanding disease vectors, crop improvement, and variation within species. Performing *de novo* assembly of the sample and elucidating novel haplotype-level sequence is a direct and powerful way to probe biology. However, most available genome assemblies, often generated painstakingly and at a great cost, represent an arbitrary consensus haploid genome where allelic sequences from the haplotypes are collapsed together, losing biological meaning when the sequences differ. There is an imminent need for technologies that enable the low cost generation of diploid assemblies, where the alleles are phased as separate sequences.

We describe here algorithmic improvements to a *de novo* assembler, Supernova^{TM} (PMID 28381613) that creates low cost, diploid assemblies using barcoded short reads (10x Linked-Reads) sequenced from a single library generated from 1 ng of high molecular weight DNA. Our algorithms are powered by the fact that reads sharing a barcode are localized to a handful of long molecules from randomly distributed segments of genome. Starting from a de Bruijn graph (K=48), Supernova progressively assembles the data into a draft assembly with greater degrees of resolution and contiguity (K=200) using read pairs. Next, the algorithm uses the barcodes to create $\sim 10^5$ local assemblies, which are then merged with the draft assembly to fill gaps and extend scaffolds. Collapsed distant regions and mis-assemblies are detected and resolved, and the assembly is finally phased using barcodes. This push-button assembly process runs locally or in the cloud.

We tested our method on data from six human cell lines, blood from a Human Genome Project (HGP) donor, and a double hydatidiform mole synthetic diploid. These assemblies are highly contiguous as validated by long scale (~25 kb) perfect matches between the assembly for the HGP sample and ~340 Mb sequence of finished clones for the same. Our assemblies are also phased at multi-megabase distances through complex regions like the 5 Mb MHC region. For example, 97% of the MHC region in the long read assemblies of individual moles is contained in one phased scaffold of our synthetic diploid. We demonstrate the simplicity and power of our method by assembling genomes for some diverse nonhuman species.

← Back

NUCLEOPROTEIN OF INFLUENZA A VIRUS REGULATES HOST TRANSLATION MACHINERY BY TARGETING MTOR-EIF4E PATHWAY PROTEINS AND ITS CONTROLLING MICRORNAS

Shipra Sharma[1], Anirvan Chatterjee[1], Sunil K Lal[2], Kiran Kondabagil[1]

[1]IIT-Bombay, Biosciences and Bioengineering, Mumbai, India, [2]Monash University, Microbiology, Selangor Darul Ehsan, Malaysia

The interplay between influenza virus and host factors to support the viral life cycle is well documented. Influenza A virus (IAV) infection triggers intracellular signaling of host organism. These pathways carry out various cellular functions which are commonly hijacked by the infecting virus to make its own copies and increase the viral load on host. The multipronged PI3K/Akt pathway plays a pivotal role in prolonging cell survival through mTOR and its downstream proteins. We have provided evidence that indicates the activation of Akt through a Ras family protein, N-Ras in X-31 infected lung epithelial A549 cells. Additionally, IAV stimulates the increase in levels of phosphorylated protein such as mTOR and S6K1, as the duration of infection and MOI increases. The levels of phospho-4EBP1 are decreased by the inhibitory effects of mTOR, which in turn activates phospho-eIF4E and hence, the synthesis of protein increases in the infected cells. Here, we have also shown for the first time, that on transiently expressing IAV nucleoprotein in A549 cells, N-Ras and phospho-mTOR levels are upregulated, resulting in reduction of p-4EBP1 levels, which leads to release of eIF4E and therefore, increase in host and viral protein translation. In concordance with these observations, it can be concluded that NP of IAV is controlling the N-Ras/Akt/mTOR pathway and thus regulating the protein synthesis of host cell. However, total mTOR and S6K1 transcripts alleviated at high dose of infection, although phospho levels of both the proteins increased at the same dose of virus, suggesting that microRNAs (miRNAs) may be regulating the signaling pathway during IAV infection. To examine it further, 24 hour post-infection, at MOI of 5, RNA was extracted from X-31 infected A549 cells or mock infected cells, empty vector expressed and NP expressed cells and then used for small RNA library construction. All libraries with different indexes were pooled together with equal concentration, followed by high throughput sequencing based on MiSeq platform. The identified miRNA species were classified and compared into up and downregulated miRNAs across the four samples. miRNAs specifically involved in regulating the mTOR-S6K1-eIF4E during IAV infection and NP expressed cells are being further investigated, which may further give insights in influenza A virus-host interaction for effective therapeutic interventions.

← Back

ENVIRONMENTAL HEALTH SCIENCES DATA COMMONS (EDAC) - A RESEARCH DATA MANAGEMENT AND DATA WORKFLOW AUTOMATION SYSTEM.

Maria Shatz[1], Adam Burkholder[3], Brian Papas[3], Charles Schmitt[1], David Fargo[2], Stephanie Holmgren[1]

[1]NIEHS, Office of Data Science, Research Triangle Park, NC, [2]NIEHS, Office of Scientific Information Officer, Research Triangle Park, NC, [3]NIEHS, Integrative Bioinformatics Support Group, Research Triangle Park, NC

Biomedical research organizations face significant challenges in managing individual and organizational scientific data. To address these challenges, we have developed the Environmental Health Sciences Data Commons (EDAC), a research data management and data workflow automation system. In order to prioritize functionality and demonstrate the utility of this resource, we have initially focused on solutions to challenges derived from the Next Generation Sequencing core laboratory and Integrated Bioinformatics Support Group given their complex data workflows, formats, structures, and management priorities and broad use of the data across the institute. Specific use cases include a scientist searching or browsing private or widely shared organizational data; a core laboratory manager generating a report; and the automation of data workflows developed by a collaborative bioinformatics group. These diverse examples demonstrate the utility and flexibility of our approach and provide a foundation for further development that is scalable, modular, and that can be integrated with existing independently developed LIMS or API systems. EDAC has been developed using iRODS as a highly scalable, storage resource agnostic platform that enables file and data collection tagging, provides microservices supporting data workflow automation, and has a rule engine that permits automated and triggered actions. EDAC includes command line and graphical web-based access to data browsing and retrieval; the creation, updating, and sharing of file collections; and metadata generation and management to provide a broad data lifecycle management framework. The resource code, metadata terms and schemas, and documentation are available at the NIEHS GitHub repository under the MIT License (MIT).

MCSPLICER: A PROBABILISTIC MODEL FOR ALTERNATIVE SPLICING.

Heejung Shim[1,2], Israa Al-Qassem[3], Yash Sonthalia[3], Maria Spletter L Spletter[4], Stefan Canzar[5]
[1]University of Melbourne, School of Mathematics and Statistics, Melbourne, Australia, [2]Purdue University, Department of Statistics, West Lafayette, IN, [3]Purdue University, Department of Computer Science, West Lafayette, IN, [4]Ludwig-Maximilians-University of Munich, Department of Physiological Chemistry, Biomedical Center, Planegg-Martinsried, Germany, [5]Ludwig-Maximilians-Universität München, Gene Center, Munich, Germany

Alternative splicing of pre-mRNA allows a single gene to produce different proteins in eukaryotes, and they have been shown to affect various gene functions, and eventually disease. Most widely used approaches to alternative splicing analysis using RNA-seq data are based on either local splicing events (e.g., exon skipping) or full-length isoforms. While the isoform-based approach provides a global picture of splicing patterns along transcripts, it poses major challenges because 1) each RNA-seq read has information on only a small part of a transcript, and 2) alternative isoforms from the same gene often share a large amount of sequence [1]. To overcome these challenges, we propose to build a probabilistic model as an approximation to the underlying splicing processes, rather than modeling individual outcomes of the processes such as exon skipping, individual isoforms, etc. Our model is based on gene-wide usage of splice sites. Specifically, we first assume that potential 5' and 3' splice sites are given. This information can be obtained from annotation databases or estimated from RNA-seq data by using other methods. Those potential splice sites partition a gene into a sequence of segments. We introduce a sequence of hidden variables, each of which indicates whether a corresponding segment is part of an isoform. We model the splicing process by assuming that this sequence of hidden variables follows an inhomogeneous Markov chain. The parameters in the model are interpreted as splice site usage. Using those parameters, we can describe the splicing process, and estimate the relative frequency of isoforms or local splicing events. We performed two types of experiments to assess the performance of our method, McSplicer. In a simulation study we verified the accuracy of McSplicer with respect to a known set of expressed human transcripts. Single-end and paired-end reads were simulated using the Flux Simulator. We observed a high agreement of our estimation with the ground truth, even for complex genes expressing multiple (overlapping) isoforms. Additionally, we used McSplicer to compare the splicing pattern between fibrillar flight muscles (IFM) and tubular muscles in Drosophila. When run on RNA-seq samples obtained from IFMs, jump muscles, and entire legs at different time points during pupal development, gene-wide usage of splice sites inferred by McSplicer was consistent with previously reported IFM-specific splicing but also included many relevant cases of splice site usages that seemingly differed between muscle types. Currently we are extending the proposed methods to identify differential splicing between multiple groups of sample. Our work is related to a method proposed in [1] that uses probabilistic splicing graphs to model splicing events.

[1] LeGault and Dewey (2013). Bioinformatics.

CHARACTERIZATION OF BACKGROUND ERRORS IN TARGETED DEEP SEQUENCING DATA SPECIFICALLY ASSOCIATED WITH PLASMA DNA.

Seung-Ho Shin[1,2], Gahee Park[1,3], Hyo-Jeong Jeon[1], Yeon Jeong Kim[1], Danbi Lee[1], Dae-Soon Son[1], Woong-Yang Park[1,2,4], Donghyun Park[1]

[1]Samsung Genome Institute, Samsung Medical Center, Seoul 06351, South Korea, [2]Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul 06351, South Korea, [3]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, South Korea, [4]Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Suwon 16419, South Korea

Cell-free DNAs (cfDNAs) released from apoptotic and/or necrotic cells circulate in the bloodstream. Although cfDNAs in the circulation system was studied for various aspects related to their origin, DNA damages specifically associated with creation of plasma DNA has not been identified yet. Here, we attempted to identify cfDNA-specific DNA damages by analyzing background errors in targeted deep sequencing data. For the purpose, we first compared error rates of cfDNA across substitution classes with their matched genomic DNA isolated from peripheral blood leukocytes (PBLs). To eschew complications by technical background noise related to DNA fragmentation, we sheared genomic DNA under the mild condition that we recently reported. The most significant elevation of plasma-specific errors was found in C:G>T:A (C>T: $p < 7.89e\text{-}03$; G>A: $p < 4.08e\text{-}33$), which was followed by A:T>T:A substitution. When two groups were compared at every position across the entire target regions, only 0.6% (4,221 / 613,006) displayed *p*-value less than 0.05 indicating that plasma-specific errors were primarily random rather than position-specific. Our result indicated that cytosine deamination were either not properly repaired in the cells releasing cfDNAs or incurred during the generation of cfDNA, which resulted in C to T on the damaged strand and G to A on the other strand.

PISCES: A PACKAGE FOR QUANTITATION AND QC OF BIG mRNA-SEQ DATASETS

Matthew Shirley, Joshua Korn

Novartis Institutes for Biomedical Research, Oncology, Cambridge, MA

PISCES is a package that eases the burden of processing large numbers of mRNA-seq libraries, and subsequently reducing errors in parameter selection and QC validation and consisting of three analysis modules: 1) single sample analysis of individual mRNA-seq libraries including species detection, SNP fingerprinting, library geometry detection, and quantitation using salmon, 2) summarization, TMM normalization, and differential expression analysis of multiple libraries to produce data formats ready for visualization and further analysis, 3) aggregation and visualization of mRNA-seq library QC metrics. PISCES improves the specificity of read mapping by masking repetitive regions of the transcriptome, increasing accuracy of abundance measures. PISCES is implemented as a python3 package, and is bundled with all necessary dependencies to enable reproducible analysis and easy deployment. Configuration files are specified to build transcriptome indices, supply sample metadata, define contrasts for differential expression analysis using DEseq2, and define default program parameters. PISCES builds on many existing open-source tools, and development versions of PISCES are available on Bitbucket, with python packages installable using pip.

HIT'nDRIVE: PATIENT-SPECIFIC MULTI-DRIVER GENE PRIORITIZATION FOR PRECISION ONCOLOGY

Raunak Shrestha*[1,2], Ermin Hodzic*[3], Thomas Sauerwald[4], Phuong Dao[5], Kendric Wang[1], Jake Yeung[1], Shawn Anderson[1], Fabio Vandin[6], Gholamreza Haffari[7], Colin Collins[1,2], Cenk Sahinalp[1,3,8]

[1]Vancouver Prostate Centre, LAGA, Vancouver, Canada, [2]University of British Columbia, Urologic Sciences, Vancouver, Canada, [3]Simon Fraser University, Computing Science, Burnaby, Canada, [4]University of Cambridge, Computer Laboratory, Cambridge, United Kingdom, [5]NIH, NCBI, Bethesda, MD, [6]University of Padova, Information Engineering, Padova, Italy, [7]Monash University, Information Technology, Melbourne, Australia, [8]Indiana University, Informatics & Computing, Bloomington, IN

* authors contributed equally

Prioritization molecular alterations that act as drivers of cancer remain as a crucial challenge and a bottleneck in the therapeutic development. The problem is particularly complicated by extensive mutational heterogeneity observed in the cancer (sub)types, yielding a long-tailed distribution of mutated genes across the patients, possibly implying the existence of many private drivers.

In order to address this problem we have developed HIT’nDRIVE, a combinatorial algorithm that measures the potential impact of genomic aberrations to expression of other genes which are in close proximity in a gene/protein-interaction network, and prioritizes those aberrations with the highest impact as drivers. HIT’nDRIVE formulates the driver prioritization problem as a “random-walk facility location” (RWFL) problem, which differs from the standard facility location problem by its use of “hitting time”, the expected number of hops to reach a “target” (expression-altered) gene from a “source” (i.e. potential drivers) gene, as a distance measure in an interaction network. HIT’nDRIVE uses “inverse” hitting time as a measure of influence of a source gene over a target gene to identify the subset of sequence-wise altered/source genes whose overall influence over expression altered/target genes is maximum possible. HIT’nDRIVE reduces RWFL to a weighted multi-set cover problem, which it solves as an integer linear program (ILP).

When applied to large tumor cohort, HIT'nDRIVE revealed many potentially clinically actionable driver genes. We demonstrate that drivers predicted by HIT’nDRIVE seed network modules/pathways that can effectively classify cancer phenotypes and sub-types as well as predict drug-efficacy and survival time. Overall, HIT'nDRIVE may help clinicians contextualize massive multi-omics data in therapeutic decision making, enabling widespread implementation of precision oncology.

RARE SPLICE SITES IN PLANT PROTEIN-CODING GENES

Shengqiang Shu, David M Goodstein, Daniel S Rokhsar

DOE JGI, Plant Program, Walnut Creek, CA,

Many gene predictions and post processing programs support only 3 known splice sites: GT/AG, GC/AG and AT/AC, resulting in a few inaccurate gene structures. We implemented an intron scoring scheme based on matches to consensus sequences of 2 splice sites and their respective branch point of U2-type and U12-type splicing and incorporate it into PASA to allow rare splice site introns supported by transcriptome in protein-coding genes. We annotated or re-annotated 6 plant genomes using IGC (JGI integrated gene call pipeline) or GMI (JGI gene model improvement pipeline) with the improved PASA. Using rare splice site protein-coding genes as homology seeds, we predicted rare splice site genes from 61 Viridiplantae genomes using Exonerate. Predicted rare splice sites and genes are verified by aligning respective transcriptome short reads to CDS sequences. Some genes with rare splice genes are conserved (same intron counting from C-terminal and same rare splice sites) widely among Angiosperm genomes while a few are specific to a clade. Based on sequences of splice sites and best scoring branch point in 30 conserved genes, AT/AC intron in these genes is likely handled by U2-type spliceosome unlike many others by U12-type spliceosome.

← Back

INCREASING THE LOWER LIMIT OF DETECTION FOR MUTATIONS WITH LIMITED NUMBER OF READS USING UNIQUE MOLECULAR IDENTIFIERS AND CONSENSUS BUILDING

Angad P Singh, Matt Hims, Alina Raza, Markus Riester, Katie D'Aco, Michael Morrissey, Rebecca Leary, Wendy Winckler, Derek Chiang

Novartis Institutes for Biomedical Research, NGDx, Cambridge, MA

Unique Molecular Identifiers (UMIs) have established their utility for accurate identification of low allelic mutations. UMIs contribute to increased performance for two reasons: 1) eliminating pseudo-duplicates from PCR 2) suppression of PCR/Sequencing errors by consensus building. Tools are now available that accomplish a combination of the above two tasks to varying extents. Most protocols though work on small targeted regions and place huge requirements on library diversity and coverage. We describe a tool that does both, eliminate pseudo-duplicates and suppress errors lacking consensus support without requiring large number of consensus reads. By requiring just two consensus reads per molecule, we show that it is possible to significantly improve the sensitivity and specificity of an assay without sequencing any additional reads. We demonstrate these results on a PanCancer gene panel containing ~ 3million bp of genomic region.

We demonstrate results on cell-free DNA samples for allele fractions between 0.25% and 0.8%. Mutations are called using MuTect in both the standard and consensus-based pipelines, and we polish the outputs using a pool of normals. While we sequence 300 – 350million reads per library across the panel, performance is demonstrated for libraries down-sampled to as low as 100m reads. Even at <1000X mean coverage, UMI-aware reads improve the sensitivity and specificity of mutation calling, especially at lower allelic fractions. At full coverage levels we show not only a significant improvement in sensitivity but also a >50% drop in the false positive count. By eliminating pseudo-duplicates, we are able to increase the unique coverage of our libraries by roughly 50%. In short we show that it is possible to benefit from UMIs across a wide spectrum of coverage levels and allele frequency ranges. Our pool of normal samples were prepared without using UMIs and thus contain errors at levels higher than those present in the consensus libraries. In the future we plan to sequence a pool of normals using UMIs for more accurate polishing of the consensus libraries.

← Back

RETROSUITE, AN INTEGRATED PIPELINE FOR THE GENOME-WIDE ANALYSIS OF TRANSPOSABLE ELEMENTS

Nicholas J Skvir, Steven Criscione, Andrew Leith, Nicola Neretti

Brown University, CCMB, Providence, RI

Mobilization of transposable elements in somatic tissue has been recently implicated in a host of disease pathologies, but their genome-wide study is complicated by the fact that many sequencing reads originating from the many copies of elements do not map uniquely. Here, we present *RetroSuite*, a collection of software that allows users 1) to identify novel transposition events, and 2) to quantify the expression level of, or the enrichment of epigenetic marks at transposable elements. *RetroFind* is a pipeline to detect transposition events from high-throughput, paired-end sequencing reads by using discordant and split reads of evidence for either side of the insertion. To assist in the experimental validation of novel insertions, *RetroFind* performs de novo assembly of all supporting reads and aligns the contigs to the genome to display putative transposition events. *RepEnrich2* is an upgraded and extended version of our popular, previously-published RepEnrich pipeline, which identifies transcriptional enrichment of repetitive elements. It addresses the difficulties inherent to mapping reads to these repetitive regions by utilizing a separate alignment of multi-mapping reads to repetitive 'pseudogenome' assemblies representing individual repetitive element subfamilies. We show applications of these methods on both *in silico* and real data, and compare their performance to other existing methodologies.

RGD: DATA AND TOOLS FOR PRECISION MODELS OF HUMAN DISEASE

Jennifer R Smith[1], Stanley J Laulederkind[1], G Thomas Hayman[1], Shur-Jen Wang[1], Matthew J Hoffman[1,2], Elizabeth R Bolton[1], Yiqing Zhao[1], Omid Ghiasvand[1], Jyothi Thota[1], Monika Tutaj[1], Marek A Tutaj[1], Jeffrey L De Pons[1], Melinda R Dwinell[1,2], Mary E Shimoyama[1]

[1]Medical College of Wisconsin and Marquette University, RGD/Biomedical Engineering, Milwaukee, WI, [2]Medical College of Wisconsin, Department of Physiology, Milwaukee, WI

A major challenge for preclinical research is finding, or establishing, a good model for the human disease of interest--one that, more or less, faithfully recapitulates the phenotypic and genetic profile of that disease in the human system. In many cases, canonical model organisms such as rat or mouse are acceptable models, but this is not always the case. As such, the Rat Genome Database (RGD, http://rgd.mcw.edu) has undertaken to incorporate additional mammalian species to allow researchers to leverage a rich dataset across multiple species to find the best model for their needs. In addition to rat, RGD has always offered data for human and mouse for the purpose of cross-species comparisons. Now these have been enhanced with data for long-tailed chinchilla (*Chinchilla lanigera*), 13-lined ground squirrel (*Ictidomys tridecemlineatus*), bonobo (*Pan paniscus*, also known as pygmy chimpanzee), and dog (*Canis lupus familiaris*). In each case, these species are used as models for human disease, including diseases of the inner and middle ear, retinal diseases, cancer, heart disease, arthritis, autoimmune dysfunction and hypoxia-reperfusion injury. Utilizing the existing robust and adaptable infrastructure, RGD has imported gene records, genomic data and ortholog assignments for these species from NCBI, as well as protein information and Gene Ontology (GO) annotations where available from UniProtKB. Further functional information is being added to these records via the assignment of GO, disease and pathway annotations based on sequence similarity to human, rat and mouse genes, and work is well underway to expand RGD's suite of analysis tools to include genes from all of these species wherever possible. Complementing the functional data, RGD is incorporating a substantial set of genomic variants, currently for rat and human but with an eye toward expanding the dataset to other species. The Variant Visualizer tool has been updated to present both rat strain-specific variants and human clinical (e.g. ClinVar) variants. This expanded offering of data for multiple species and the analysis tools to easily and efficiently leverage this data gives researchers an excellent resource for discovering precision models for their diseases of interest.

← Back

VALIDATION AND IMPLEMENTATION OF KIDNEYSEQ^{TM}: A COMPREHENSIVE GENE PANEL FOR GENETIC RENAL DISEASES

Ramakrishna Sompallae[1,2], Adela Mansilla[1], Sara Mason[1], Mycah Kimble[1], Carla Nishimura[1], Anne Kwitek[1,3], Colleen Campbell[1,4], Thomas Christie[4,5,6], Richard Smith[1,4,5]

[1]University of Iowa, Iowa Institute of Human Genetics, Iowa City, IA, [2]University of Iowa, Department of Pathology, Iowa City, IA, [3]University of Iowa, Department of Pharmacology, Iowa City, IA, [4]University of Iowa, Department of Internal Medicine, Iowa City, IA, [5]University of Iowa, Department of Pediatrics, Iowa City, IA, [6]6Veterans Affairs Medical Center, Iowa City, IA

Next-generation sequencing (NGS) technology is increasingly used as a clinical tool to facilitate genetic diagnoses and optimize patient care. To expedite the diagnosis of genetic kidney diseases, we developed KidneySeq^{TM}, an NGS platform that simultaneously screens 179 genes implicated in 75 renal diseases. Sequence data are processed through a customized bioinformatics workflow that includes open source software for the genome alignment of reads and the detection of a wide range of genetic variations, including single nucleotide variants (SNVs) and indels, as well as single- and multi-exon copy number variations (CNVs). Detected variants are annotated and filtered using in-house software. Due to the occurrence of multiple pseudogenes and the complexity of unambiguous PKD1 variant calling, a parallel workflow is used to identify variants in PKD1. A KidneySeq^{TM} sequencing output of 5 million reads per sample equates to coverage of >99% of target bases with at least 30 reads (30x) and the detection of >99.9% of variants. Validation was completed using two control CEPH samples (NA12287 and NA12878) to compare KidneySeq^{TM} and high-density SNP arrays genotype calls. In addition, we compared results from the control sample, NA12878 against high confidence variant calls from the Genome in a Bottle (GIAB) consortium, a widely used dataset for sequence validation. With both comparisons, concordance in variant calling was 100%, with a sensitivity of 100% and specificity of 99.8%. We then tested 292 variants from 31 control samples by Sanger sequencing and calculated the positive predictive value and negative predictive value to be 98.5% and 100%, respectively. KidneySeq^{TM} has now been used as a clinical test in over 100 patients with a variety of kidney diseases. The diagnostic rate is ~40%; in one-third of these cases, the genetic diagnosis has led to a change in the clinical diagnosis. Pathogenic variants include SNVs (82%%), indels (9.8%) and CNVs (8.2%). Our findings support the use of comprehensive genetic testing using platforms like KidneySeq^{TM} to improve the clinical care of patients with renal disease.

CLASSX: SCALABLE SIMULTANEOUS TRANSCRIPT ASSEMBLY OF MULTIPLE RNA-SEQ DATA SETS

Li Song[1], Liliana Florea[1,2]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Johns Hopkins School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

High-throughput sequencing of RNA (RNA-seq) has become the experimental method of choice for analyzing the transcriptome of an organism or species. The low cost and fast turnaround is making it possible to sequence cohorts of hundreds and even thousands of samples as part of clinical or genetic variation studies. There have already been significant advances in designing tools for analyzing a single RNA-seq sample. However, challenges remain owing in part to the different expression levels of genes, as many transcripts do not have enough reads to allow full-length reconstruction. The current approach, which stitches together partial transcripts from multiple samples, improves the coverage but inherits the errors of individual data sets and is prone to further misassemblies. Therefore, analyzing each sample individually limits both the accuracy and the potential to identify splice variants, in particular rare or low expression events.

We propose a novel transcript assembly method, implemented in the tool ClassX, that can simultaneously analyze hundreds of RNA-seq data sets. Instead of first assembling transcripts in each sample and then stitching them to predict full-length models, ClassX selects a high-confidence set of features (exons and introns) from the full set of samples using statistical methods and connects them into a subexon splice graph, a compact data structure that represents the gene and its splice variants in a comprehensive way. To select a subset of transcripts from among those encoded in the graph, ClassX employs a variation of the weighted SET-COVER problem to select a parsimonious set of transcripts that can explain all the input alignments in each data set. ClassX does not process each data set independently, rather it uses the read patterns and abundance information across data sets to improve the accuracy. ClassX is both sensitive, taking advantage of the latent information in multiple samples to identify rare or low expression variations, and precise, leveraging redundancy across multiple samples to select a high confidence set of transcripts. While still in its beta version, ClassX is efficient and scalable, and can be effectively used for comprehensive transcriptomic analyses of tens to hundreds of RNA-seq data sets.

← Back

KEVLAR: REFERENCE-FREE VARIANT DISCOVERY IN HUMAN GENOMES AND BEYOND

Daniel S Standage[1], C. Titus Brown[1,2], Fereydoun Hormozdiari[2,3]

[1]UC Davis, Population Health and Reproduction, School of Veterinary Medicine, Davis, CA, [2]UC Davis, Genome Center, Davis, CA, [3]UC Davis, MIND Institute, Davis, CA

Discovery of small-scale variation in genome composition and structure, and characterizing this variation in the context of human disease, is an area of intense research interest. Evidence from whole-exome sequencing of genetic disease samples suggests a significant contribution from rare de novo or somatic variants to diseases such as autism and cancer. Strategies based on aligning short reads to a reference genome dominate variant discovery methods. This approach suffers from several related deficits. First, many reads align to the reference genome poorly, or not at all, due to repetitive DNA, novel sequence, or misassemblies in the reference. These reads contain interesting and potentially critical data that is discarded de facto by reference-based methods. Also, mapping-based methods are insensitive to certain classes of variants such as 5-200 nucleotide indels and many structural variants. And finally, there are many research contexts in agriculture, veterinary medicine, and related fields where reference genomes are either unavailable or of insufficient quality for reference-based variant discovery.

We have developed a novel alignment-free k-mer based method, **Kevlar**, for discovery of de novo and somatic variants. Based on simulations we have shown that novel mutations generally produce many k-mers not present in the reference genome. Accordingly, our method is based on analysis of k-mer abundances directly from raw reads, which we can achieve in very low memory using a novel k-mer banding strategy. k-mers unique to a diseased individual, or (more generally) of differential abundance in case samples versus controls, point directly to loci of probable interest. Reads containing novel k-mers are loaded into an assembly graph, which can be partitioned into disconnected components representing distinct variants, and subsequently filtered, refined, assembled, and used for variant calling and annotation. Tests on simulated data have confirmed kevlar's ability to correctly call SNVs and INDELs of various lengths, and preliminary tests on real data suggest that there may be between 10 and 100Mbp of novel, non-erroneous sequence in samples from the 1000 Genomes YR1 trio. Initial results have confirmed many high-confidence variants including an experimentally validated de novo Alu insertion.

The Kevlar method is being developed as an open source research software project, and is freely available at https://github.com/dib-lab/kevlar. The initial implementation is optimized for simplex studies, but Kevlar's k-mer banding strategy supports scaling to very large cohort studies (such as cancer case/control studies) even when available memory is limited.

GENOME REARRANGEMENT TRIGGERED BY THERMOSTABLE RESTRICTION ENZYME INDUCING MULTIPLE DNA DOUBLE-STRAND BREAKS

Hidenori Tanaka[1], Nobuhiko Muramoto[1], Arisa Oda[2,3], Takahiro Nakamura[2], Kazuto Kugou[2], Kunihiro Ohta[2]

[1]Toyota Central R&D Labs., Inc., Genome Engineering Program, Nagakute, Japan, [2]The University of Tokyo, Department of Life Sciences, Graduate School of Arts and Sciences, Meguro-ku, Japan, [3]The University of Tokyo, Universal Biology Institute, Meguro-ku, Japan

DNA double strand breaks (DSBs) induced by interruption in replicating and chemicals generate chromosomal rearrangement. During reduction division stage, most meiotic recombination is initiated by the formation of DSBs made by Spo11. In recent years, genome editing technologies using endonucleases, such as TALEN and CRISPR-Cas9, are developing rapidly. These nucleases create site-specific DSBs, however, multiple DSBs' effect on genome in mitotic cells is unclear.

In this study, we developed an experimental tool for introducing multiple DSBs in *Arabidopsis thaliana* plants (diploids/tetraploids) using a heat-activated endonuclease. When the enzyme was activated in vivo, the frequency of homologous recombination (HR) is strongly elevated and the expression of gene involved in HR was induced. To check whether induced DSBs in mitotic cells affect plant genome in the progeny, we conducted whole-genome sequencing (100-bp pair end analysis). The numbers of mutations, such as SNV, deletion, insertion, were elevated in the progeny genome. Moreover, genome-wide analyses at single nucleotide resolution indicated that multiple DSBs led to diverse chromosomal rearrangements (CRs) in tetraploid plants, whereas CRs were rarely detected in diploids. These data suggest that *de novo* mutations induced by DSBs can be inherited overcoming reproductive lineage. Our research tool may be useful for facilitating the generation of polyploid crops, allowing the faster development of new varieties.

← Back

QUALITY ASSESSMENT AND LARGE-SCALE INTEGRATION OF CHROMOSOME CONFORMATION CAPTURE DATASETS.

Michael E Sauria, James Taylor

Johns Hopkins University, Biology and Computer Science, Baltimore, MD

Chromatin architecture is recognized as an integral component of gene regulation and cell differentiation. Chromosome conformation capture (3C) and derived techniques have revealed many features about the structure of chromatin. However, making full use of the data produced by these experiments is challenging for many researchers because of computational limitations, bioinformatics knowledge, and a lack of user-friendly tools. To address these challenges, we present a set of new methods, tools, and databases to make chromosome conformation data more widely usable.

QuASAR (Quality Assessment of Spatial Arrangement Reproducibility) is a score to assess the intrinsic quality of Hi-C and related datasets. QuASAR is built on the idea of spatial consistency: that as the 3D distance between two genomic regions approaches zero, the correlation between those regions distance to every other region should approach one. QuASAR quantifies the extent to which this consistency holds across an entire dataset and aggregates it into a single score. The QuASAR QC measure is an interpretable score that can perfectly rank simulated datasets according to noise levels and distinguish low quality real Hi-C experiments from high quality ones.

Interaction hub: we have compiled a comprehensive database of Hi-C data, supported by Galaxy, and integrated with analysis and visualization tools allowing truly open access to more than 1,500 Hi-C datasets. We have created a uniform processing and analysis pipeline, executed using CWL workflows and run in containerized environments. We also have developed quality metrics for Hi-C samples to help evaluate sample quality and replicate reproducibility. Each processing step is made available rather than simply endpoint data, including quality metrics for each phase. Data were all processed using HiFive, a Hi-C analysis suite available on Galaxy main.

Visualization in Galaxy Trackster: We have also created a 2-dimensional genome browser connected to Trackster for easy data exploration within Galaxy. Samples can be directly loaded from the data library into Trackster-2D for visual assessment, comparison to one-dimensional genomic annotations, or for Hi-C inter-dataset comparison. In order to support fast browsing and compact of these sparse datasets, we also have developed a multi-resolution 2-dimensional binary tree file format, allowing easy access to any level of resolution and the random access to data necessary for real-time browsing.

← Back

GRAMENE: COMPARATIVE GENOMICS, GENE EXPRESSION AND PATHWAY REFERENCE RESOURCES FOR PLANT COMMUNITIES

Marcela K Tello-Ruiz[1], Joshua Stein[1], Sharon Wei[1], Parul Gupta[2], Sushma Naithani[2], Justin Preece[2], Andrew Olson[1], Yinping Jiao[1], Sunita Kumari[1], Kapeel Chougule[1], Bo Wang[1], Young K Lee[1], James Thomason[1], Peter D'Eustachio[3], Robert Petryszak[4], Paul Kersey[4], Pankaj Jaiswal[2], Doreen Ware[1,5]

[1]Cold Spring Harbor Laboratory, Plant Genomics, C, NY, [2]Oregon State University, Dept Botany & Plant Pathology, Corvallis, OR, [3]NYU School of Medicine, Dept Biochemistry & Molecular Pharmacology, New York, NY, [4]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, [5]USDA ARS NEA Plant,, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

Understanding the relationships among different plant genomes and their constituent gene sets facilitates the identification of DNA sequences preserved over time in different organisms, *e.g.*, genes that are essential to life and genomic signals that control gene function across species. Gramene (http://www.gramene.org) is a powerful online resource that provides easy access to plant comparative genomics, gene expression, and pathway data through search, visualization, and analytical tools.

Gramene integrates software components such as Ensembl's genome browser, Reactome's pathways database infrastructure, and Expression Atlas widgets to provide a robust informatics platform for over 1.7 million genes from 44 reference genomes. The majority of genes are organized in 62,367 GeneTree families, with orthologous and paralogous gene classification, whole-genome alignments, and synteny. In addition, it hosts over 240 curated rice pathways and orthology-based projections for 66 plant species. Our integrated search database and modern user interface leverage these diverse annotations to facilitate finding genes using interactive viewing modes, including genomic context, expression anatograms, associated pathways, and customized gene trees. Besides phylogeny-derived data, Gramene incorporates genetic, structural, and phenotypic diversity data. Our BioMart data mining utility enables complex queries and bulk download of sequence, functional annotation, homology, and variation data. In addition to hosting data for comparative studies, Gramene also provides unified access to diverse plant community resources and gives communities the abilities to upload, display, and analyze private data sets in multiple standard formats.

Gramene is supported by NSF grant IOS-1127112, and partially from USDA-ARS (1907-21000-030-00D).

← Back

THE *IN VIVO* TRANSCRIPTION START SITE AND ENHANCER LANDSCAPE OF INFLAMMATORY BOWEL DISEASE ENABLES DISEASE CLASSIFICATION AND INTERPRETATION OF NON-CODING SNPs.

Malte Thodberg*[1,2], Mette Boyd*[1,2], Morana Vitezic*[1,2], Jette Bornholdt*[1,2], Kristoffer Vitting-Seerup[1,2], Anders G Pedersen[4], Yun Chen[1,2], Ole H Nielsen[3], Jacob T Bjerrum*[3], Albin Sandelin*[1,2]

[1]University of Copenhagen, Department of Biology, Copenhagen, Denmark, [2]University of Copenhagen, Biotech Research and Innovation Centre, Copenhagen, Denmark, [3]Herlev Hospital, Department of Gastroenterology, Herlev, Denmark, [4]Technical University of Denmark, Center for Biological Sequence Analysis, Lyngby, Denmark

Traditionally, much of disease genomics and transcriptomics have focused on disease-specific changes of protein-coding transcripts. However, a slew of recent scientific work has highlighted the importance of changes in intergenic regions (DNAse hypersensitive sites, chromatin marks, methylation, enhancers, etc.), which may affect the transcriptional regulation of genes and thereby transcript levels.

Cap-Analysis of Gene Expression (CAGE) is a technique that captures the first 30 basepairs of transcribed 5’-capped RNAs. While originally developed to detect and measure gene transcription start site (TSS) activity, we have shown that active enhancers transcribe short bidirectional so-called enhancer RNAs (eRNAs), and they can therefore also be detected by CAGE. This is promising from a medical genomics perspective, as CAGE can be used on small, frozen tissue biopsies, which is challenging with chromatin-based techniques.

To show this utility, we have profiled gut biopsies from a large number of patients of a common chronic gut disease, inflammatory bowel disease (IBD). Despite having a strong genetic component, the molecular pathogenesis of IBD is poorly understood, but believed to be a complex interplay of both genetic, luminal, and environmental factors that trigger an abnormal mucosal immune response to the gut microbiome.

We used CAGE data obtained from intestinal tissue biopsies from 94 IBD patients to build an IBD-specific TSS and enhancer atlas. We show that TSSs and enhancers share transcription factor binding enrichments and can distinguish both the severity of inflammation and the type of IBD (Ulcerative Colitis or Crohn’s Disease) of patients. We identify several antibacterial peptides as extreme outliers in terms of expression variance across patients, which may be interesting for patient stratification and precision medicine. Utilizing GWAS data we show that enhancers are more enriched for the heritability of IBD than promoters around active TSSs, underscoring their role in IBD pathogenesis.

In addition to providing insights into IBD biology and potential IBD biomarkers, this study also illustrates the general usefulness of CAGE for building disease-specific TSS and enhancer atlases.

COMPARISON OF RAT STRAIN-SPECIFIC VARIANTS COLLECTION FOR ALL RAT GENOME REFERENCES

Monika Tutaj, Jennifer R Smith, Marek A Tutaj, Jeffrey L De Pons, Stanley J Laulederkind, Thomas G Hayman, Shur-Jen Wang, Mary E Shimoyama

Medical College of Wisconsin, Biomedical Engineering, Milwaukee, WI

As a part of a large-scale of genomic, genetic, phenotype and disease data collection the Rat Genome Database (RGD, http://rgd.mcw.edu) provides information about strain-specific variants, including copy number variations, single nucleotide variants (SNVs) and indels, generated from whole genome sequencing (WGS) of rat strains used as models for a variety of common human diseases. These variants were originally analyzed against previous rat genomic assemblies (RGSC 3.4 and RGSC 5.0). To bring this up to date, RGD re-analyzed available rat strain whole-genome sequences using the newest version of the rat reference assembly (RGSC 6.0, released in 2014). We identified over 12 million genomic variants (SNVs and indels) among the 25 rat strains using GATK Best Practices recommendations (DePristo 2011, Van der Auwera 2013). Difficulties with correct variant identification and developing reliable calling methods require up-to-date evaluation of known and new algorithms, their accuracy and performance. To accomplish this, we are developing an analysis pipeline which utilizes state-of-the-art tools for variant calling and effect prediction. In addition, we are expanding our current search tools to include the ability to use the combination of a variant position in the genome with the functional annotations assigned to RGD genes and strains. The combination of improved variant data and disease, phenotype, pathway, function and chemical interaction information will help researchers find appropriate models for disease research and drug testing.

← Back

RAPID IDENTIFICATION OF MHC ALLELES AND HAPLOTYPES IN GENETICALLY DIVERGENT CATTLE POPULATIONS USING NGS

Deepali Vasoya*[1], Andy Law[1], Laura-Agundez Muriel[2], Paolo Motta[1], Mingyan Yu[3], Adrian Muwonge[1], Elizabeth Cook[3], Xiaoying Li[2], Karen Bryson[2], Amanda MacCallam[2], Tatjana Sitt[4], Philip Toye[3], Barend Bronsvoort[1], Mick Watson[1], Ivan Morrison[2], Timothy Connelley*[2]

[1]The Roslin Institute, University of Edinburgh, Genetics and Genomics, Edinburgh, United Kingdom, [2]The Roslin Institute, University of Edinburgh, Infection and Immunity, Edinburgh, United Kingdom, [3]International Livestock Research Institute, Livestock genetics, Nairobi, Kenya, [4]The University of Vermont, Animal and veterinary sciences, Burlington, VT

The MHC (Major Histocompatibility Complex) region contains many genes that are key regulators of both innate and adaptive immunity including the polymorphic MHCI and MHCII genes. Consequently, the characterisation of the repertoire of MHC genes is critical to understanding the variation that determines the nature of immune responses. Our current knowledge of the bovine MHC repertoire is limited with only the Holstein-Friesian breed having been characterised to an appreciable extent. Traditional methods of MHC genotyping are of low resolution and laborious and this has been a major impediment to a more comprehensive analysis of the MHC repertoire of other cattle breeds. Next-generation sequencing (NGS) technologies have been used to enable high throughput and much higher resolution MHCI and MHCII typing in a number of species. In our method, we designed the pan-MHCI and MHCII (BoLA-DRB3) primers that allowed amplification and sequencing of all known and many novel bovine MHC alleles using Illumina MiSeq platform. A bioinformatics pipeline has been established that can comprehensively analyse the resulting data and filters out the redundant sequences and defined the known and novel MHC genes and haplotypes. It was validated initially on a cohort of Holstein-Friesian animals (European *Bos taurus*) and then demonstrated to enable characterisation of MHCI repertoires in cohort of Boran (*Bos indicus*) cattle from Kenya, Fulani cohort (African *Bos taurus* x *Bos indicus*) cattle from Cameroon and Brazilian cattle breeds (European *Bos taurus* and *Bos indicus*), for which there was limited or no available data. During the course of these studies we identified >200 novel classical MHCI genes, >20 novel MHCII genes and defined >80 novel MHCI haplotypes, dramatically expanding the known bovine MHC repertoire. These identified novel genes help to study diversity and evolutionary biology of bovine MHC system across different bovine populations and provides data fundamental to development of vaccines targeting T-cells.

← Back

COPY NUMBER AND EXPRESSION VARIATION IN AMPLICONIC GENES ON THE Y CHROMOSOMES OF GREAT APES.

Rahulsimham Vegesna, Marta Tomaszkiewicz, Paul Medvedev, Kateryna Makova

Pennsylvania State University, University Park, PA

The male-specific region of the human Y chromosome harbors nine multi-copy ampliconic gene (AG) families [1]. The gene copies within each family are frequently 99.9% identical to each other, because most of them occupy opposite arms of palindromes, or massive inverted repeats, on the Y. AGs are expressed exclusively in testis and encode proteins functioning during spermatogenesis [2]. Experimental studies demonstrated that there is variation in the number of AGs per family within and across great ape species [3,4,5]. However, how the variation in AG copy numbers affects male fertility and expression levels in testis of humans and other great apes has remained understudied.

Here we present a novel method, Ampliconic Copy Number Estimator (AmpliCoNE), that utilizes read depth information to estimate ampliconic gene copy number (CN) per family. We estimated ampliconic gene CN and expression levels in 23 human samples from the Genotype-Tissue Expression (GTEx) project using AmpliCoNE and Kallisto, respectively. Across the ampliconic gene families, the more copious ones (*TSPY* and *RBMY*) had higher expression levels than less copious ones, but, surprisingly, no significant relationship between CN estimates and gene expression was found within individual gene families.

For the more copious gene families, differences in both gene CN and expression levels were partially determined by Y haplogroups. In particular, the *TSPY* family in haplogroup E (African) had higher CN estimates but lower expression levels than in haplogroup R (European). In contrast, the *RBMY* family in haplogroup R had higher CN estimates and lower gene expression levels than in haplogroup E. We also observed that the *VCY* family, which is deleted in most of great apes species, has lost one of its two copies in most of the human samples.

We assembled the transcriptomes [3] of bonobo, gorilla and orangutan testis to obtain transcripts representing AG families across these three great ape species. Complete transcripts of more copious families were obtained in all three ape species, which implies a consistently high gene expression level for these gene families. However, the transcripts of the less copious gene families were fragmented likely due to low expression levels. We are in the process of comparing the AG expression levels among the great apes species.

1. Skaletsky H., et al. Nature (2003).
2. Vallender, E., Bruce L. BioEssays (2004).
3. Tomaszkiewicz M, et al. Genome Res. (2016).
4. Oetjens, M., et al. GBE (2016).
5. Massaia, A., Yali X. Human genetics (2017).

← Back

POWSIMR: POWER ANALYSIS FOR BULK AND SINGLE CELL RNA-SEQ EXPERIMENTS

Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, Ines Hellmann
LMU, Biology II, Munich, Germany

Recent development of very sensitive RNA-seq protocols, such as Smart-seq2 and CEL-seq allows transcriptional profiling at single-cell resolution and droplet devices make single cell transcriptomics high-throughput, allowing to characterize thousands or even millions of single cells.
In powsimR, we have implemented a flexible tool to assess power and sample size requirements for differential expression (DE) analysis of single cell and bulk RNA-seq experiments. For our read count simulations, we (1) reliably model the mean, dispersion and dropout distributions as well as the relationship between those factors from the data. (2) Simulate read counts from the empirical mean-variance and dropout relations, while offering flexible choices of the number of differentially expressed genes, effect sizes and DE testing method. (3) Finally, we evaluate the power over various sample sizes.

The number of replicates required to achieve the desired statistical power is mainly determined by technical noise and biological variability and both are considerably larger if the biological replicates are single cells.
powsimR can not only estimate sample sizes necessary to achieve a certain power, but also informs about the power to detect DE in a data set at hand. We believe that this type of posterior analysis will become more and more important, if results from different studies are to be compared. Often enough researchers are left to wonder why there is a lack of overlap in DE-genes across similar experiments. PowsimR will allow the researcher to distinguish between actual discrepancies and incongruities due to lack of power.
The R package and associated tutorial are freely available at https://github.com/bvieth/powsimR.
doi: 10.1093/bioinformatics/btx435

LINKED-READS VS LONG READS: BALANCING COST AND CONTIGUITY IN THE *VITIS CINEREA* GENOME.

George L Wang[1,7], Michael S Campbell[1], Deanna M Church[4], Fred Gouker[5], Jason Londo[2], Mike Regulski[1], Bruce Reich[5], Qi Sun[3], Tim Smith[6], Will Thompson[6], Stephen R Williams[4], Xia Xu[2], Lance Cadle-Davidson[2], Doreen Ware[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]USDA-ARS, Geneva, NY, [3]Cornell University, Ithaca, NY, [4]10X Genomics, Pleasanton, CA, [5]Cornell University, Geneva, NY, [6]USDA-ARS, Clay Center, NE, [7]Yale University, New Haven, CO

Grape breeders are introgressing disease resistance from *Vitis cinerea* into highly susceptible cultivated grapes. However, given the high genetic diversity in *Vitis*, many molecular markers used in marker-assisted breeding do not work well across breeding germplasm. To identify a set of trait-linked molecular markers that will work across multiple grape breeding programs we generated a reference assembly of the Vitis cinerea genome using Single Molecule Real Time sequencing. We also collaborated with 10X Genomics to generate a reference assembly from Linked-Reads. Both sequencing approaches are designed to preserve long-range information, but the linked read approach is less expensive by nearly an order of magnitude, and provides long range haplotype information as a product of the assembly process. Here we present a comparison between the two technologies, including assembly contiguity and completeness, transposable element content, protein coding gene content, and cost. These analyses were done using the National Science Foundation cloud computing service JetStream. We created a system images, permitting anyone with a JetStream allocation to access the exact computing environment used for this project. The image for the computing environment used in this project can be found at https://use.jetstream-cloud.org/application/images/451.

← Back

COMPUTATIONAL MODELING LONG NON-CODING RNAS MEDIATED TRANSCRIPTION REGULATION

Haozhe Wang[1,2], Jiawei Gu[1,2], Zhenyu Xuan[1,2]

[1]The University of Texas at Dallas, Department of Biological Sciences, Richardson, TX, [2]The University of Texas at Dallas, Center for Systems Biology, Richardson, TX

RNA-protein interactions are essential for understanding many important cellular processes. In particular, long noncoding RNAs(lncRNAs) together with ribonucleoprotein (RNP) complexes and numerous chromatin regulators can target appropriate locations in the genome to regulate gene expression. However, the experimental discovery and validation of lncRNA-protein-DNA interactions remain time-consuming and labor-expensive, and only a few theoretical approaches are available for predicting potential lncRNA–protein-DNA associations. We have developed a machine learning based method to computationally model the lncRNA-protein-DNA interaction. Using support vector machine (SVM) method, we integrated genome-wide protein-DNA interaction and RNA-DNA interaction profiles to infer the set of the most probable proteins involved in lncRNA recognizing its genomic target regions. We applied this method in studying lncRNA NEAT1 and its targets in MCF7 breast cancer cell line, and achieved the prediction accuracy of 90%. We also identified a group of transcription factors that may potentially be involved in NEAT1-DNA interaction. This method provides novel insights to our understanding how lncRNA regulates gene expression through the specific interactions with chromatin.

GTD: ESTIMATING GENOTYPE LIKELIHOOD BY DEEP NEURAL NETWORKS.

Jiayao Wang[1,4], Hongjian Qi[1,3], Yufeng Shen[1,2]

[1]Columbia University Medical Center, Department of system biology, New York, NY, [2]Columbia University Medical center, JP Sulzberger Columbia Genome Center, New York, NY, [3]Columbia Univesity, Department of Applied physics and Applied Mathematics, New York, NY, [4]Columbia University Medical center, Department of Pediatrics, New York, NY

Accurate estimation of genotype likelihood of single nucleotide variants (SNVs) and small insertions/deletions (indels) is the foundation of modern genetic studies and clinical genetic diagnosis. Conventional methods are mostly based on a similar likelihood model that assumes errors in reads that support the same alternative alleles are nearly independent. However, this assumption is not valid due to systematic issues such as context-dependent sequencing errors, alignment errors due to repeats or imperfection of reference genomes. As a result, genotype quality estimate is far from adequate in downstream genetic analysis, and a range of ad hoc filters are required to minimize false positives. In practice, manually inspecting variants in IGV has become a de facto standard to vet candidates for validation or follow-up analysis. In large scale sequencing projects, this process is too laborious and not consistent. Here we describe GTD, a new method to estimate genotype quality by deep neural network approaches. We aim to use GTD to replace IGV manual inspection process and provide basis for calling de novo mutations using probabilistic approaches. Given a candidate variant (SNVs or indels) called by any methods, GTD extracts the information of all reads that cover the candidate variant site from original read alignment, and use a tensor to represent both DNA sequence and base quality. GTD captures the intuition of identifying false positive calls by their inconsistency with local haplotypes. We designed the model based on a residual network (ResNet) and implemented the method using TensorFlow. We trained the method using whole genome sequencing data from a trio on Genome in a Bottle (GIAB). We show that genotype likelihood estimated by GTD is much more accurate than GATK and other conventional methods based on comparison with gold standard call sets from GIAB. As a result, using genotype quality alone, GTD achieves better precision-recall performance than GATK GQ or VQSR, especially for rare SNVs. Additionally, we implement the method to be able to perform on-the-fly training to account for sample-specific error rates. Finally, we tested the utility in calling de novo mutations using a large set of trio exome sequencing data, and demonstrate that GTD has better performance than published ad hoc filtering methods.

Defusion: a tool to improve predictions of tandemly duplicated genes created by the MAKER annotation pipeline.

Jie Wang[1,2], Dongyan Zhao[1], C. Robin Buell[1], Kevin L Childs[1,2]

[1]Michigan State University, Department of Plant Biology, East Lansing, MI, [2]Michigan State University, Center for Genome-Enabled Plant Science, East Lansing, MI

Genome sequencing of nonmodel organisms is growing rapidly, and it has become increasingly important to correctly annotate genes in an automated fashion. The MAKER pipeline provides a robust and scalable solution for genome annotation. However, accurate gene prediction by MAKER is dependent on high-quality and correctly aligned protein and transcript evidence. Exonerate is used by MAKER for aligning protein and transcript evidence. Unfortunately, with tandemly duplicated genes, exonerate often stretches protein and transcript alignments across the duplicated loci. MAKER interprets this evidence to mean that the tandemly duplicated genes are in fact a single locus, and one gene model is created that crosses the two loci. This is particularly problematic for plant species as genes involved in producing secondary metabolites are often tandemly arrayed. Here we developed a Python-based tool called deFusion to identify and locally annotate fused gene models generated by MAKER. Fused genes are recognized by finding sequence similarities between the 5’ and 3’ ends in MAKER gene models. By default, the midpoint of the fused gene model is suggested as the breakpoint, but human-curated customized breakpoints are also accepted. Fused loci are split, and each separated locus is locally re-annotated. Newly aligned evidence, new ab initio predictions and new MAKER gene models replace the evidence and models from the fused locus, and deFusion outputs a corrected transcript, protein and MAKER gff files. The deFusion tool has been successfully applied to fix fused gene models from two medicinal plant genomes: *Catharanthus roseus* and *Camptotheca acuminata*.

← Back

METAGENOMICS STUDY OF INDIVIDUALS WITH SEASONAL INFLUENZA

Yinpeng Andy Wang, Benjamin Cowling, Joseph Wu, Malik Peiris, Herbert Pang

The University of Hong Kong, School of Public Health, Hong Kong, Hong Kong

Background: Next-generation sequencing technologies have given us an unprecedented opportunity for studying the microbiome of complex infectious diseases like influenza. The objective of this study is to explore the underlying microbiome profiles of patients with seasonal influenza type A.

Methods: Nasal swabs were collected from ten healthy subjects at baseline. These subjects might become infected at the second time point. Each sample was subjected to whole-genome shotgun sequencing. The microbiota of subjects who remained healthy at time point two were compared with those who got infected at phylum and genus levels.

Results: At the second visit, five subjects remained healthy and five got infected. The nasal bacterial communities were dominated by phyla Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria. We observed that the change of the abundance of Proteobacteria in infected subjects was significantly different from healthy subjects. At genus rank, the dominant bacteria were Corynebacterium, Moraxella, and Staphylococcus and the change of Corynebacterium in infected subjects differed significantly from healthy subjects.

Summary: Our study represents the first survey of the nasal microbiota before and during seasonal influenza infection.

← Back

CONVERGENCE OF LIGHT, STRESSES AND CIRCADIAN RHYTHM ON NUCLEAR-ENCODED CHLOROPLAST-LOCALIZED GENES (NECGS) IN ARABIDOPSIS

Yuejun Wang*, Meifang Qi*, Jian Liu, Fei Zhao, Jingfei Cheng, Yijing Zhang

National Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

* Authors contributed equally to this work.

Internal rhythms and external light synchronize plant responses to the environment. Thousands of environmental response genes display periodic expression, potentially resulting in time-of-day-dependent responses, but how this large-scale rhythmic expression is orchestrated by light and the circadian clock is unclear. Here, based on systematic analysis of large-scale biological data, we observed a trade-off between light-induced and stress-induced gene expression, which is balanced via endogenous circadian expression. Furthermore, central clock components antagonize the effect of the transcription factors (TFs) in the hub of light responses by both blocking TF expression and repressing TF target expression. Together, these findings systematically revealed the convergence of the central oscillator and external light and stresses through direct co-occupancy of TFs in the regulatory hub.

← Back

SPARKLE: FINDING MISSING GENES IN DRAFT GENOMES

Kourosh Banaeianzadeh, Matthew L Workentine, James D Wasmuth

University of Calgary, Faculty of Veterinary Medicine, Calgary, Canada

Sparkle is a suite of bioinformatic tools that accelerate and improve the complex task of annotating a species' genome.

The well reported reduction in the cost of DNA sequencing is leading to a rapid increase in the number species with genome projects. The assembly and annotation of these genomes is increasingly fully automated. However, our ability to make sense of these data are often confounded by errors in the genome annotation. With respect to genome annotation, e.g. gene finding, there is insufficient funding or technical expertise to perform detailed manual curation. The Wasmuth lab attempted to reconstruct various signalling pathways in species of nematodes. Strikingly, several genes, that had been cloned in targeted studies, were absent from the genome assembly. We were able to recover these genes from the raw sequence reads. Further screens revealed this to be a problem common across many metazoan genomes.

We have developed Sparkle, part of our ShinyR genomes project. In brief, an organism's sequence reads are aligned against a curated set of protein sequences. We use the number and distribution of the read-to-protein alignments to determine the likelihood that a given protein is encoded in the genome based on coverage from the sequence reads. The results are presented with a web-based graphic interface, which enables the user to explore different parameter settings and to approve or discard read-to-protein alignments.

Sparkle is written in the R programming language and employs the Shiny framework for web applications. The codebase is open access to encourage community input for contributions. We are using Sparkle to improve the annotation of the genome of *Haemonchus contortus*, a helminth species that parasitizes small ruminants. We have already found a previously missing gene that completes a signalling pathway known to control parasitism related behaviour.

← Back

HOW SHAPES OF NUCLEOSOMAL DNA CAN REGULATE GENE EXPRESSION

Sergiusz Wesolowski[1], Jorge Martinez[1], Wei Wu[2], Daniel Vera[3]

[1]Florida State University, Mathematics, Tallahassee, FL, [2]Florida State University, Statistics, Tallahassee, FL, [3]Florida State University, Center of Genomics and Personalized Medicine, Tallahassee, FL

Nucleosomes positioning, histone modifications, and general nucleosome make-up are the underlying factors in affecting gene transcription, thus, understanding how nucleosomes are distributed can advance the field of genomics. Despite of recent advances in technology of sequencing and experimental design, the full comprehension of how nucleosome are arrangement influences gene regulation remains elusive. Untapping the full potential of the next generation sequencing experiments, and drawing sound inferences from such experiments, relies on appropriate mathematical methods that can capture the complexity of the structure of the underlying biological processes.

In this talk we describe a new elastic shape analysis framework based on Square Root Slope Functions to analyze MNase-seq experiments. The new model redefines experimental results as elastic shapes over reference genome. The shape interpretation allows us to establish the connection between elastic changes of DNA arrangement around the nucleosome near TSS and the changes in the following gene's expression. The model explains how nucleosome shape and position can regulate gene activity.

SNAPTRON: A TOOL AND SERVICE FOR STUDYING SPLICING IN TENS OF THOUSANDS OF INDIVIDUALS

Christopher Wilks[1], Phani Gaddipati[2], Abhinav Nellore[3,4], Jonathan Ling[5], Ben Langmead[1]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [3]Oregon Health & Science University, Biomedical Engineering, Portland, OR, [4]Oregon Health & Science University, Surgery, Portland, OR, [5]Johns Hopkins University, Neuroscience, Baltimore, MD

As more and larger genomics studies appear, there is a growing need for comprehensive and queryable cross-study summaries. These enable researchers to leverage vast datasets that would otherwise be too difficult to obtain or too computationally unwieldy to analyze from scratch. We present Snaptron[1], a search engine for summarized RNA sequencing data. It serves data from over 70,000 human RNA-seq samples, analyzed using Rail-RNA[2,3] and also served in a more raw form by recount2[4]. Snaptron's computational core is a query planner that leverages R-tree, B-tree and inverted indexing strategies to rapidly execute queries over 146 million exon-exon splice junctions from over 70,000 human samples.

The easiest way to use Snaptron is via its RESTful web service interface (http://snaptron.cs.jhu.edu), which allows researchers to immediately start posing queries (e.g. simply starting with a gene name) with little or no software installation. Most queries take only a few seconds and can be tailored by constraining which junctions and samples to consider. Snaptron can score junctions according to tissue specificity or other criteria. Importantly, Snaptron can also score samples according to alternative splicing patterns by calculating the "percent spliced in" of individual exons. Using this framework, we have identified hundreds of previously unannotated cell type-specific exons and the splicing factors that regulate these exons. We further highlight several case studies relevant to human disease to illustrate the versatility of Snaptron.

[1] Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. bioRxiv doi: 10.1101/097881.
[2] Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. Bioinformatics. 2016 btw575.
[3] Nellore A, Wilks C, Hansen KD, Leek JT, Langmead B. Rail-dbGaP: analyzing dbGaP-protected data in the cloud with Amazon Elastic MapReduce. Bioinformatics. 2016 Aug 15;32(16):2551-3.
[4] Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using recount2. Nature Biotechnology. 2017 Apr 11;35(4):319-321.

← Back

GENE FUSION INFORMATION MANAGEMENT

Paul D Williams

Thermo Fisher Scientific, Life Sciences Solutions Group, Ann Arbor, MI

Chromosomal aberrations generating gene fusions are important to the process of cancer progression. The expression of fusion genes can lead to abnormally high levels of cellular growth signaling, due to overexpression or through constitutive kinase domain activation. These events may be drivers of the cancer phenotype, and their detection in tumor cells is important for translational research. The number of reported gene fusion isoforms is large (~10,000) and continues to grow. Many different methods to describe fusion isoforms can be found in the literature, which, combined with exon numbering ambiguity, make interpretation of fusion breakpoint sequences challenging. Here we describe the development of a proprietary, internal database and web interface to store information about fusion isoforms and the assays used to measure them. The database, built using the Django framework, contains detailed information about individual isoforms, literature references, and the genes, exons, and transcripts involved. This allows for rapid traceability from assay design to evidence sources and additionally provides an interface to manage additional metadata about the isoforms in a consistent manner, such as whether the isoform is expected to generate a functional product or which gene partner, if any, is likely to be the driver. This database has been used to assist the development of targeted sequencing assays that can accurately detect the presence of gene fusions. Examples of isoforms that were translated into research assays to detect rare fusion isoforms will be presented.

← Back

CORRECTING DISCORDANCE BETWEEN SIX COPY-NUMBER-CALLING METHODS FOR 2778 TUMOURS

Jeff Wintersinger[1,2]

[1]University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada, [2]University of Toronto, Department of Computer Science, Toronto, Canada

Changes in genomic copy number play a critical role in cancer. As part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, we created consensus clonal copy-number-aberration (CNA) predictions across 2778 tumours by combining six different copy-number-calling methods. Though the six selected methods represented the state of the art, they nevertheless exhibited a surprising degree of disagreement. To address this, we developed a robust consensus strategy to maximize the comprehensiveness of our predictions without compromising their accuracy. Our work suggests that smaller-scale studies relying on a single tool for determining copy number may obtain inaccurate copy-number estimates for many tumours. However, by using our consensus strategy, it is possible to recover from mistakes made by one or more algorithms on part or all of a tumour genome.

Disagreement between methods arose from two sources. Firstly, methods made differing predictions concerning whether clonal whole-genome duplications (WGDs) occurred, affecting more than one-third of tumours. To correct such cases, we used a combination of heuristics and manual review to determine the most probable ploidy.

Secondly, methods disagreed on how to segment each tumour's genome into regions of different copy number, complicating our consensus efforts. Of the six methods, two were “liberal” segmenters, calling on average an order of magnitude more segments per genome. When we compared every pair of the six methods head-to-head, the liberal segmenters had higher recall but lower precision. Conversely, the four “conservative” segmenters exhibited better precision at the cost of lower recall. These differences stemmed from the design of each algorithm, with liberal segmenters relying heavily on read-depth variation to segment the genome, while conservative methods primarily used allelic imbalance in heterozygous germline SNPs. To capitalize on the strengths of both conservative and liberal segmenters, we developed a novel interval-based consensus algorithm that establishes a “majority vote” for each segment, while also incorporating the uncertainty inherent in different input segmentations. By augmenting these results with structural variants, which often demarcate changes in copy number status, we created consensus segmentations for each genome.

By resolving WGD uncertainty and creating consensus segmentations for each tumour, we achieved a substantial consensus between the six CNA-calling methods on PCAWG's 2778 tumours. Beyond PCAWG, our work illustrates a surprising discordance between CNA-calling methods, while also providing a strategy for combining their predictions to produce consensus profiles drawing upon the strengths of individual methods.

GENERATING FULL-LENGTH, HIGH-QUALITY HUMAN TRANSCRIPTOMES FROM PACBIO ISO-SEQ DATA

Dana Wyman[1,2], Gabriela Balderrama-Gutierrez[1,2], Shan Jiang[1,2], Weihua Zeng[1,2], 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

Conventional short-read RNA sequencing has been widely used to quantify gene expression in a variety of applications. However, short reads on their own lack the ability to resolve full-length isoforms, which can be several kilobases in length. Furthermore, computational methods developed to reconstruct isoforms from short read data are plagued by challenges, and results from different algorithms tend to be inconsistent. While long read sequencing technologies such as PacBio Iso-seq and Oxford Nanopore have a higher error rate than Illumina sequencing, they have great potential for isoform discovery and characterization of the 90% of multi-exon human genes that are thought to undergo alternative splicing. To take advantage of these properties, we develop a computational pipeline to process long reads into cleaned isoforms and generate a high-quality, full-length transcriptome. We demonstrate this process on PacBio Iso-seq data from human cell lines K562, GM12878, and HepG2 and show that the technology is mature enough to produce full-length transcriptomes by comparing the results to existing ENCODE data.

← Back

AN ITERATIVE APPROACH FOR RECONSTRUCTING FULL-LENGTH RIBOSOMAL GENES FROM WHOLE METATRANSCRIPTOMIC DATA

Yaxin Xue[1], Inge Jonassen[1], Anders Lanzén[2]

[1]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway, [2]Soil Microbial Ecology Group, Department for Conservation of Natural Resources of NEIKER-Tecnalia, Derio, Spain

Motivation:
Technological advances in metatranscriptome sequencing (MTS) have enabled a deep understanding of the structure and function of microbial communities. Whole MTS (total RNA) provides a unique opportunity to investigate active microbial community from all three domains of life (rRNA) and function (mRNA) simultaneously. A critical step for the former involves 16S rRNA reconstruction. However, current tools are developed for amplicon and metagenomic data, and do not handle the data volumes of total RNA data sets and also struggle with the high complexity of the data sets.

Results:
In this work, we introduce a novel Iterative Approach for Reconstructing Ribosomal genes (IARR) in MTS. 16S rRNA of highly abundant species can be reconstructed from sequence subsamples, and our iterative approach takes advantage of this to reduce computational costs. We apply the approach to several simulated microbial communities, shown that our tool can recover more rRNA genes with less false positive results and time/memory usage, compared with several specially designed rRNA reconstruction tools.

← Back

MODULARITY ANALYSIS OF ENHANCER-PROMOTER INTERACTION NETWORKS

Chengfei Yan, Shaoke Lou, Mark Gerstein

Yale University, Molecular Biophysics and Biochemistry, New Haven, CT

The 3D organization of human genome plays an important role in regulating gene expressions through bringing distal functional elements like enhancers and promoters into spatial proximity to form physical interactions. In this study, we build the regulatory element interaction networks including the enhancer-promoter network, the promoter-promoter network and the mixed network including both enhancers and promoters for human GM12878 cell line by connecting these regulatory elements with various types of high resolution genome interaction data. Modularity analysis of these networks reveals certain number of communities. Community-based functional and gene expression analysis are performed. The relationships between these modules and genome architecture units including TADs and compartments are also investigated.

JULIP++: FAST AND ULTRA-SENSITIVE IDENTIFICATION OF DIFFERENTIAL SPLICING EVENTS FROM LARGE RNA-SEQ DATA COLLECTIONS

Guangyu Yang[1], Liliana Florea[1,2]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Johns Hopkins School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

Alternative splicing is essential to characterize gene regulation and to understand development and disease. In recent years, the high-throughput RNA sequencing (RNA-Seq) technology has become a powerful tool for quantitative profiling of RNA splicing and alternative splicing. It has become increasingly common to carry out RNA-Seq experiments on a large number of samples. However, determining the full extent of splicing variation and comparing gene splicing profiles among biological states at a global scale remains challenging. Current reference methods only analyze signals from one sample at a time, which limits transcript reconstruction and fails to detect a complete set of alternative splicing events. We introduce JULiP++, a novel method that simultaneously analyzes information across multiple samples to reliably and comprehensively identify alternative splicing variations and differential isoform usage at the level of splice junctions (introns). JULiP++ can accurately discover novel introns in samples and precisely assess the statistical significance of splicing changes between different condition groups. The performance of JULiP++ was evaluated on simulated and real RNA-Seq data. For the intron selection task, JULiP++ detected thousands more introns at higher or comparable precision (>98%) compared to tested assemblers including Cufflinks, CLASS2, StringTie, FlipFlop and ISP. For the differential splicing task, JULiP++ outperformed two existing methods (rMATS and JunctionSeq) in the simulation setting, yielding very high sensitivity (>85%) and precision (>86%). JULiP++ is multi-threaded and parallelized, taking just a few minutes to analyze up to 100 data sets on a multicomputer cluster, and can easily scale up to allow analyses of hundreds and thousands of RNA-seq samples.

← Back

IDENTIFICATION OF PREDICTORS FOR THERAPEUTIC RESPONSE TO IMMUNOTHERAPY

Ping Ye

Avera Cancer Institute, Avera Cancer Institute, Sioux Falls, SD

Checkpoint inhibitors have demonstrated clinical efficacy for cancers ranging from melanoma to non-small cell lung cancer, renal cell cancer, and breast cancers. Despite the efficacy, adverse effects of immunotherapy are noted in more than half of the patients. Further, the majority of patients do not respond to the therapy, and a subset of patients develop “hyper-progressive” disease. Therefore, it has become evident that predictive biomarkers of response are needed to choose patients who are most likely to respond to these novel immune agents. Several biomarkers have been identified, including tumor mutational burden (TMB), microsatellite instability, PDL1 and PDL2, and tumor-infiltrating lymphocytes. However, none has reliably predicted response in a manner for routine use. The objective of this study is to assess the counts of four types of mutations, single nucleotide variant (SNV), indel, copy number variant (CNV), and gene fusion, detected by tissue next generation sequencing (NGS) and reported from FoundationOne, to analyze their predictive value of the immunotherapy at Avera Cancer Institute. We assessed 43 patients with solid malignancies who received checkpoint inhibitor-based immunotherapy and tissue NGS testing prior to initiation of immunotherapy, and were evaluated for clinic outcome (CR, PR, SD, PD, or Deceased). Median patient age was 66 years (range: 28 to 82 years). Fourteen patients (33%) were men. The most common tumor types were breast cancer (11 cases), non-small cell lung cancer (10 cases), and gynecologic cancer (10 cases). When we divided the patients into two groups (CR+PR+SD vs PD+Deceased), we found greater numbers of SNV and indel in the responders (mean=14) than non-responders (mean= 12). However, the numbers of CNV and gene fusion did not reveal any difference between the two groups. Consistent with the trend of mutation counts, we observed a significant difference in TMB between responders and non-responders (P-value=0.03). Further, mutation counts exhibit a significant positive correlation with TMB (Correlation coefficient=0.83, P-value=1.635e-11), suggesting the number of SNV and indel could potentially approximate TMB and predict outcome. Finally, mutations enriched among responders and non-responders were identified and their contributions to TMB were characterized.

GENOMICS AND BIOINFORMATICS APPROACH TO INVESTIGATING THE ROLE OF REPETITIVE ELEMENTS IN THE CHEMOPREVENTION OF COLORECTAL CANCER

Suman Lee, Jayarama B Gunaje, Wenfeng An, Ping Ye

South Dakota State University, College of Pharmacy and Allied Health Professions, Brookings, SD

Epidemiological and clinical data show that long-term aspirin use is the most consistent example of a chemopreventive agent against colon cancer. However, the mechanisms of aspirin chemoprevention are not entirely clear. Long interspersed elements 1 (LINE1) are the only autonomous retrotransposons that can produce new insertions in human genomes. The link between LINE1 and colon cancer is supported by novel insertions in the tumor suppressor gene APC where mutations are frequent in colorectal patients. To systematically investigate LINE1 activity in the chemoprevention of colorectal cancer, we performed an RNA-Seq study on HCT116 (human colorectal carcinoma) cell lines treated with aspirin at three doses (0mM, 5mM, 10mM). RNA-seq libraries were constructed using Illumina's TruSeq Stranded mRNA Library Preparation Kit, and sequenced on an Illumina NextSeq 500. Mapping sequencing libraries to the integrated luciferase reporter gene showed that L1HS (the youngest LINE1 family in the human) promoter driving the reporter transgene was suppressed at 10mM of aspirin treatment. This result is consistent with luciferase enzymatic assays. Mapping sequencing libraries to the L1HS consensus sequence indicated that sense transcripts were suppressed at 10mM aspirin treatment. Mapping sequencing libraries to the genome yielded a similar outcome. Interestingly, sense transcript suppression by aspirin significantly correlates with the age of 26 LINE1 families. However, genome-wide LINE1 expression suppression cannot be solely explained by LINE1 promoter inhibition, as promoter-less LINE1 elements also exhibited sense transcript suppression. Moreover, the LINE1 expression suppression is not due to gene expression. In fact, the expression of L1 neighboring genes was elevated upon 10mM aspirin treatment. We are in the process of investigating how other factors would influence LINE1 expression during chemoprevention of colorectal cancer, such as epigenetic modifications, long non-coding RNAs, and transcriptional factor binding. Experimental effort is also underway to characterize whether aspirin suppresses LINE1 insertion. Our studies may reveal a global role of LINE1 suppression during chemoprevention of colon cancer by aspirin.

← Back

LEARNING RNA BINDING PROTEIN MOTIFS USING CONVOLUTIONAL NEURAL NETWORKS

Adamo J Young[1,2], Quaid D Morris[1,2,3], Timothy R Hughes[2,3]

[1]University of Toronto, Department of Computer Science, Toronto, Canada, [2]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [3]University of Toronto, Terrence Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada

RNA binding proteins (RBPs) regulate most aspects of RNA processing in the cell and play a critical role in gene expression. Understanding the behaviour of RBPs is essential for developing treatment for diseases that are associated with abnormal RBP activity. RBPs work by biochemically recognizing and binding to specific RNA sequence and structure patterns, which are commonly called binding motifs. We developed a deep-learning based model, inspired by DeepBind (Alipanahi *et al.* 2015), that can learn an RBP's binding motif from *in vitro* data. Our model utilizes a convolutional neural network (CNN) structure, an architecture that is often used for computer vision applications because it can accurately capture and combine key image features (like eyes, mouths, and noses for facial recognition) in a manner that is resistant to variation in position and orientation. These qualities are advantageous when learning RBP binding preferences, as RBP motifs are often composed of smaller sub-motifs that reflect the modular properties of the RNA binding domains in their binding sites. We used a type of *in vitro* RBP-RNA binding data (RNAcompete-S, Cook *et al.* 2017) that captures information about the sequences and structures that an RBP will bind. For each protein, we trained a model to accurately identify which RNAcompete-S sequences would be bound. After training, we extracted the learned binding motifs from the model's convolution filters and compared them with motifs from the literature. Our approach offers an alternative to existing models for RBP motif elicitation and may be particularly useful for RBPs with strong structural binding preferences. The motifs generated by our model could potentially be used to identify regulatory regions of transcripts and causal disease variants.

← Back

THE CANCER GENOME COLLABORATORY

Christina K Yung[1], George L Mihaiescu[1], Bob Tiernay[1], Junjun Zhang[1], Michelle D Brazas[1], Francois Gerthoffert[1], Andy Yang[1], Jared Baker[1], Guillaume Bourque[2], Paul C Boutros[1], Bartha M Knoppers[2], Francis Ouellette[1], Cenk Sahinalp[4], Sohrab P Shah[5], Vincent Ferretti[1], Lincoln D Stein[1,3]
[1]Ontario Institute for Cancer Research, Toronto, Canada, [2]McGill University, Montreal, Canada, [3]University of Toronto, Toronto, Canada, [4]Simon Fraser University, Vancouver, Canada, [5]BC Cancer Agency, Vancouver, Canada

The Cancer Genome Collaboratory is an academic compute cloud designed to enable computational research on the world's largest and most comprehensive cancer genome dataset, the International Cancer Genome Consortium (ICGC). The ICGC is on target to categorize the genomes of 25,000 tumors by 2018. A subproject of ICGC, the PanCancer Analysis of Whole Genomes (PCAWG) alone has generated over 800TB of harmonized sequence alignments, variants and interpreted data from over 2,800 cancer patients. A dataset of this size requires months to download and significant resources to store and process. By making the ICGC data available in cloud compute form in the Collaboratory, researchers can bring their analysis methods to the cloud, yielding benefits from high availability, scalability and economy offered by cloud services, avoiding large investment in compute resources and eliminating time for download.

To facilitate the computational analysis on the ICGC data, the Collaboratory has developed software solutions that are optimized for typical cancer genomics workloads, including well tested and accurate genome aligners and somatic variant calling pipelines. We have developed a simple to use, but fast and secure, data transfer tool that imports genomic data from cloud object storage into the user's compute instances. Because cancer datasets have restrictions on their storage locations, it is important to have software solutions that are interoperable across multiple cloud environments. We have successfully demonstrated interoperability across The Cancer Genome Atlas (TCGA) dataset hosted at University of Chicago's Bionimbus Protected Data Cloud, the ICGC dataset hosted at the Collaboratory, and ICGC datasets stored in the Amazon Web Services (AWS) S3 storage.

The Collaboratory is actively growing, with a target hardware infrastructure of over 3000 CPU cores and 15 petabytes of raw storage. As of August 2017, the Collaboratory holds information on 2,000 ICGC PCAWG donors (550TB total). The Collaboratory has been successfully utilized by multiple research groups, most notably PCAWG project researchers who analyzed thousands of genomes at scale over a few weeks' time. The Collaboratory is now open to the public and we invite cancer researchers to learn more about our cloud resources at cancercollaboratory.org, and apply for access to the Collaboratory.

← Back

JTRACKER – WORKFLOW MANAGEMENT AND EXECUTION BACKED ON GIT REPOSITORY WITH FULL PROVENANCE

Junjun Zhang, Linda Xiang, Brice Aminou, Chen Chen, Lincoln Stein, Vincent Ferretti

Ontario Institute for Cancer Research, Dept. of Informatics, Toronto, Canada

A persistent challenge in bioinformatics has been the reproducibility of complex computational workflows. One of the major factors that contributes to this challenge is the difficulty of representing the full execution details of the workflows, including command-line parameters, software versions, and the representation of operations that distribute computation across multiple machines and then gather their outputs for further steps. Efforts have been undertaken to promote standardization and transparency of workflow definitions. However the computational details may not be sufficiently recorded for individual workflow execution to allow result replication by other researchers. Here we present JTracker - a new Git repository based solution for workflow management and execution. Using the inherent capability of revision control provided by Git, JTracker records all aspects of every workflow execution allowing full history to be replayed for large-scale multi-batch computational analysis.

Leveraging the Git server's role as a remote repository, a Git server is set up to record workflow tasks and their states. Distributed task execution workers directly communicate with the Git server to pick up queued tasks and report back execution states and parameters needed by subsequent tasks. Git merge conflict exceptions occur if a second worker picks up a task that is being executed by another worker. As such, without writing a single line of server side code, JTracker is able to use Git server to perform the critical role of orchestrating the execution of large amount of tasks and independent workers.

JTracker is completely compute environment agnostic, can be deployed in traditional HPC clusters, academic/commercial clouds. Additionally, JTracker plays well with container technologies (such as Docker), individual tools in a JTracker workflow can either be dockerized in an image or pre-installed on compute nodes, no special settings needed in JTracker workflow definition to enable containerized tools.

In Cancer Genome Collaboratory, a cloud platform for collaborative research, we use JTracker to perform workflows to transfer large amount of ICGC datasets between EGA and Collaboratory. To date, over 80 TB of data was transferred and QC'd using ~32,000 jobs.

Although JTracker currently offers basic functionality, we aim to develop it as a full-featured generic solution for workflow authoring, management and execution, which will be able to run workflows written in other languages, such as CWL and WDL. We also plan to develop a web UI to allow Collaboratory users to run their workflows in a fully automated and tracked fashion.

JTracker source code available at: https://github.com/icgc-dcc/jtracker

AN IMPROVED ASSEMBLY IDENTIFIES NEW FEATURES OF CUCUMBER GENOME

Zhonghua Zhang

Chinese Academy of Agricultural Sciences, Institute of Vegetables and Flowers, Beijing, China

The new single-molecule sequencing technologies can generate reads with the average lengths of more than 10,000bp, thus providing the opportunity to improve the complicated reference genomes. Here, the improved cucumber genome (226.4 Mb) was assembled de novo into 879 contigs with an N50 length of 2.8 Mb from 50X single-molecular sequencing data. The error rate for each base was estimated to be 0.0001 using ~100X Illumina reads. These contigs were linked into 703 scaffolds with an N50 length of 11.6 Mb combining with the mate pair reads from different insert size libraries, fosmid ends and BAC ends. Using about 2000 genetic markers from three maps, 205.5 Mb were anchored onto the seven chromosomes. The genetic and physical positions show a high consistency, suggesting the high-quality accuracy of the contiguity. Compared with the previous assembly, 29.2 Mb novel sequences, which include a lot of repetitive sequences and novel genic sequences, were generated. This high-quality genome assembly will serve as a valuable resource for comprehensive analysis of genomic organization in cucumber as well as plant comparative genomics.

← Back

THE GENOME ORGANIZATION OF AN AUTOTETRAPLOID POTATO（*SOLANUM TUBEROSUM* L.）

Qian Zhou[1,2], Wu Huang[1], Chunzhi Zhang[2], Zhonghua Zhang[1], Sanwen Huang[1,2]

[1]Institute of vegetables and flowers,Chinese Academy of Agricultural Sciences, Department of Biotechnology, Beijing, China, [2]Agricultural genomics Institute At Shenzhen,Chinese Academy of Agricultural Sciences, Department of Biotechnology, Shenzhen, China

Potato (*Solanum tuberosum* L.), the most important tuber food crop in the world, has been facing many difficulties in its genetic improvement during the past hundred years. The haploid potato genome has significantly accelerated the scientific research in potato, but its contribution to molecular breeding of potato is limited. The major reason is that the cultivated potato is autotetraploid, which is highly heterozygous and hard to be detailedly explored. Here we reported a chromosome-scale 2.6 gigabase draft genome of a potato cultivar, C88. The assembly covered ~87% of the tetraploid genome and was anchored onto 48 pseudochromosomes assisted by a genetic map generated from 1000 individuals. We identified ~1.6 gigabase repeated sequences accounting for 65% of the assembly and annotated 133,733 protein-coding genes. The analyses of the RNA-seq data and Methyl-seq data uncovered the allele-specific expression and modification among the homolog genes in this autopolyploid genome. The assembly, the arrangement of homolog sequences and the fates of multiple alleles provide insights into evolution and domestication of autoplolyploid crop.

← Back

AN INTEGRATIVE ROADMAP TO PAX3 TARGET GENE NETWORKS IN MELANOCYTES AND MELANOMA

Kirby A Ziegler, Alan Underhill

University of Alberta, Oncology, Edmonton, Canada

Melanoma accounts for 70% of skin cancer-associated deaths, despite representing a small fraction of all skin cancers. The developmental origin of melanocytes is thought to be a key driver of this aggressive behaviour. In this context, melanocyte identity is determined during embryogenesis by the hierarchical action of transcription factors, exemplified by MITF, SOX10 and PAX3. Each of these proteins has key roles in melanoma, reflecting their capacity to control pathogenic gene expression programs. Within this scheme, PAX3 controls cell division and differentiation by recognizing specific target sequences in the genome and altering expression of associated genes. To this end, PAX3 contains two DNA-binding domains, the homeodomain and the paired domain, which itself comprises two subdomains. In theory, the inherent modularity in this architecture permits the recognition of multiple DNA-binding patterns, yet we have not had a comprehensive view of PAX3 DNA-binding specificity or associated targets. *We hypothesize that PAX3 utilizes multiple modes of DNA recognition that can contribute to distinct functional outputs during melanoma progression.* To address this, we have repurposed data derived from cyclic amplification and selection of targets to statistically model DNA-binding specificities for human and mouse PAX3 proteins. Significantly, these profiles represent the first set of optimal motifs described for full-length PAX3. The robustness of this library was validated *in situ* by calculating its enrichment across published ChIP-seq datasets for exogenous PAX3 and the PAX3-FOXO1 pathogenic variant in muscle. This provided a foundation for predicting PAX3 occupancy in putative cell-specific regulatory regions defined using epigenomic signatures. Notably, PAX3 motifs were significantly enriched in human melanocyte enhancer regions derived from the Roadmap Epigenomics Project, as well as mouse melanocyte enhancers demarcated by H3K4me1 and adjacent EP300 occupancy. To identify downstream targets of PAX3, we used RNA-sequencing to profile differential gene expression following PAX3 attenuation across syngeneic melanocyte and melanoma cell lines. Putative targets were subsequently integrated with predicted PAX3 occupancy to connect distinct DNA-binding profiles to transcriptional pathways across cell types. Collectively, these analyses provide novel insight into the discrete target gene networks associated with the differential use of PAX3 DNA-binding modules and how these programs may be altered during progression from untransformed melanocyte to metastatic melanoma. In addition, they provide a framework for defining PAX3 regulatory networks across a broad range of PAX3-expressing cell types.

HYBRID ASSEMBLY OF CHALLENGING GENOMES WITH MASURCA MEGA-READS.

Aleksey V Zimin[1], Daniela Puiu[1,2], Steven L Salzberg[1,2]

[1]University of Maryland, IPST, College Park, MD, [2]Johns Hopkins University, IGM, Baltimore, MD

The third generation (PacBio SMRT and Oxford Nanopore) genome sequencing data opened a large new realm of possibilities in de novo genome assembly due to long (10kb+) read lengths and no sequencing bias. However the inherent high error rates of about 15% present a challenge to using these data to assemble highly repetitive or heterozygous plant genomes. I will describe a hybrid technique that is capable of overcoming the assembly challenges by effectively combining the third generation long PacBio or Nanopore reads with the second generation short but accurate Illumina reads. We have successfully applied this technique to create complete and accurate assemblies of several challenging plant genomes such as ancestral wheat A. tauschii, and the hexaploid wheat T. aestivum. The latest Nanopore MinION data sets worked extremely well on the NA12878 human sequence.

The technique is implemented in publicly available MaSuRCA assembler. One can learn about the assembler and download the code from http://masurca.blogspot.com.

THE SPECTRUM OF LOSS OF FUNCTION TOLERANCE IN THE HUMAN GENOME

Konrad J Karczewski[1,2], Laurent Francioli[1,2], Kaitlin E Samocha[3], Beryl Cummings[1,2], Daniel Birnbaum[1,2], Mark J Daly[1,2], Daniel G MacArthur[1,2]

[1]Broad Institute, Medical and Population Genetics, Cambridge, MA, [2]Massachusetts General Hospital, Analytic and Translational Genetics Unit, Boston, MA, [3]Wellcome Trust Sanger Institute, Human Genetics, Cambridge, United Kingdom

Deciphering the function and essentiality of genes in the genome is a central problem in human genetics. While knockout generation is a workhorse of elucidating gene function in model organism genetics, obvious ethical and technical limitations prevent its use in humans. Fortunately, large-scale exome and genome sequencing panels have provided a glimpse into standing genetic variation, including naturally-occurring loss-of-function (LoF) variation. The presence of LoF variants at high rates suggests a gene's redundancy, while a significant depletion against an expectation suggests strong selective pressures against these variants, and thus, the gene's essentiality. Understanding exactly where each human gene lies along the spectrum between these extremes is important for prioritizing variants both as candidate disease genes and for the development of inhibitory therapeutics.

Using a mutational model to generate expected numbers of variants for each gene in the genome, we previously found that in a dataset of 60,706 individuals with exome sequencing data from the Exome Aggregation Consortium (ExAC), 3,230 genes were found to be significantly depleted (constrained) for LoF variation, even in the heterozygous state. Here, we have updated our mutational model using data from 15,496 genomes from the Genome Aggregation Database (gnomAD) and incorporated CpG methylation status for each base in the genome. We present improvements to the constraint model by incorporating variant prioritization algorithms such as LOFTEE, PolyPhen, and CADD, as well as transcript expression. Furthermore, we incorporate allele frequency information to assess constraint with respect to zygosity. We apply this model to high-confidence LoF variants to exome sequencing data from 123,136 exomes from gnomAD to define a set of genes constrained against heterozygous and homozygous variation. Finally, we investigate genes at the other end of the spectrum and define a set of confidently LoF-tolerant genes.

← Back

PASSENGER MUTATIONS IN 2500 CANCER GENOMES: OVERALL MOLECULAR FUNCTIONAL IMPACT AND CONSEQUENCES

Mark Gerstein[1,2,3], Sushant Kumar[1,2], Jonathan Warrell[1,2], William Meyerson[2], Patrick Mcgillivray[2], Leonidas Salichos[1,2], Shantao Li[2], Arif Harmanci[1,2], Calvin Chan[6], Alexander Fundichely[4], Carl Herrmann[6], Morten Nielsen[7], Yan Zhang[5], Xiaotong Li[2], Ekta Khurana[4]

[1]Yale University, Molecular Biophysics and Biochemsitry, New Haven, CT, [2]Yale University, Program in computational biology, New Hven, CT, [3]Yale University, Computer Science, New Haven, CT, [4]Weill Cornell Medical College, Meyer Cancer Center, New York, NY, [5]Ohio State University, Department of Biomedical Informatic, Coloumbus, OH, [6]German Cancer Research Center, Department of theoretical bioinformatics, Heidelberg, Germany, [7]Aarhus University Hospital, Clinical Medicine, Aarhus, Denmark

Classical models of cancer progression posit that only a small number of these mutations strongly drive tumor progression and that the remaining ones (termed “nominal passengers”) are inconsequential for tumorigenesis. In this study, we leveraged the comprehensive variant data from Pan-cancer Analysis of Whole Genomes (PCAWG) project to ascertain the molecular functional impact of each variant, including nominal passengers. This allowed us to decipher their overall impact uniformly over different genomic elements, both coding and non-coding. The functional impact distribution of PCAWG mutations shows that, in addition to high and low impact variants, there is a group of medium-impact nominal passengers predicted to influence gene expression or activity. Moreover, we found that functional impact relates to the underlying mutational signature: different signatures confer contrasting impact, differentially affecting distinct regulatory subsystems and categories of genes. Also, we find that functional impact varies based on subclonal architecture (i.e. early vs. late mutations) and can be related to patient survival. Furthermore, we adapted an additive effects model derived from complex trait studies to show that aggregating nominal passenger variants provide significant predictability for cancer phenotypes beyond the characterized driver mutations. We further used the additive effects model to provide a conservative estimate on the number of weak drivers and deleterious passengers in different cancer cohorts. Finally, we delineate multiple lines of evidence that correlate the overall burdening of cancer mutations with the existence of both weak positive and negative selection during tumor evolution.

UTILIZATION OF LINKED-READ, WHOLE GENOME, WHOLE EXOME AND TRANSCRIPTOME SEQUENCING IN THE COMPREHENSIVE MOLECULAR PROFILING OF PEDIATRIC BRAIN TUMORS

Ben Kelly[1], James Fitch[1], Catherine Cottrell[1,2], Vincent Magrini[1,3], Daniel Koboldt[1,3], Julie Gastier-Foster[1,2,3], Jeff Leonard[4,5], Richard Wilson[1,3], Elaine Mardis[1,3], Peter White[1,3]
[1]Nationwide Children's Hospital, The Institute for Genomic Medicine, Columbus, OH, [2]Nationwide Children's Hospital, Department of Pathology, Columbus, OH, [3]The Ohio State University, Department of Pediatrics, Columbus, OH, [4]Nationwide Children's Hospital, Department of Pediatric Neurosurgery, Columbus, OH, [5]Nationwide Children's Hospital, Center for Childhood Cancer and Blood Diseases, Columbus, OH

In this study we performed a comprehensive genomic characterization of brain tumors in a pediatric cohort while simultaneously vetting novel NGS approaches and bioinformatics workflows for utility in clinical care. By bridging expertise in both clinical and research genomic profiling, research results with clinical confirmation will be returned to the care provider to inform diagnosis, prognosis and eligibility for targeted therapeutics and clinical trials.

To identify clinically relevant somatic and germline variants, each subject enrolled underwent comprehensive genomic profiling using DNA and RNA extracted from tumor and DNA from normal blood through a combination of WES, WGS and RNA-Seq, along with 10X Genomics' (10XG) linked-read technology. Each method of sequencing used in this study was chosen for its ability to accurately capture clinically relevant variations which may impact care. High-depth WES identifies low variant allele fraction somatic coding variants and germline susceptibility loci while WGS enabled us to detect structural and non-coding variants. RNA-Seq identifies both potential gene fusions and expressed somatic variants and 10XG proved to be particularly powerful for identifying the structural variants underpinning our gene fusions.

This approach led to significant computational challenges, with ten distinct analyses and pipelines needing to be performed. Elastic computational infrastructure utilizing Amazon Web Services enabled us to effectively handle this workflow, with the ability to scale up resources to process large volumes of sequence data and scale down as analyses complete. We developed a translational bioinformatics pipeline which integrated these diverse analyses, and evaluated the utility of incorporating orthogonal data analysis approaches to improve accurate clinically relevant variant identification.

To date, we have enrolled 13 pediatric brain cancer patients. Six patients have thus far been fully processed from sequencing through to targeted clinical verification. Three had somatic gene fusions involving known cancer genes *BRAF* and *FGFR1*. One had a somatic copy number loss of *TP53* and two others had somatic changes in *PIK3CA*, *FGFR1* and *PTPN11*. All variants were successfully confirmed by clinically-validated assays and have the strong potential to impact patient diagnosis, prognosis and eligibility for targeted therapeutics and clinical trials.

COMPLETING A HUMAN GENE KNOCKOUT CATALOG THROUGH ACCURATE PHASING OF 15K RARE, DELETERIOUS COMPOUND HETEROZYGOUS MUTATIONS IN 61K EXOMES

Jeffrey Staples, Evan K Maxwell, Lukas Habegger, Jeffrey G Reid

Regeneron Genetics Center, Genome Informatics, Tarrytown, NY

A primary goal of human genetics is to better understand the function of every gene in the genome. Homozygous loss-of-function mutations (LoFs) are a powerful tool to gain insight into gene function by analyzing the phenotypic effects of these "human knockouts" (KOs). Rare (MAF <1%) homozygous LoFs have been highlighted in recent large-scale sequencing studies and have been critical in identifying many gene-phenotype interactions. While rare compound heterozygous mutations (CHMs) of two heterozygous LoFs are functionally equivalent to a rare homozygous KO, they are rarely interrogated in large sequencing studies. Accurate identification of rare CHMs of LoFs is valuable: (1) rare CHMs substantially increase the number of human gene KOs, improving statistical power; (2) rare CHMs KOs may involve extremely rare heterozygous mutations which may lack homozygous carriers; and (3) rare CHMs provide a more complete set of KOs for a "human KO catalog".

We performed a survey of rare CHMs among 61K whole exome sequenced individuals from the DiscovEHR cohort. First, we identified 39,459 high-quality putative CHMs (pCHMS) consisting of pairs of rare heterozygous variants that are either LoFs (i.e., nonsense, frameshift, or splice-site mutations) or missense variants with strong evidence of being deleterious. Second, we phased all pCHMs using a combination of allele-frequency-based phasing (EAGLE) and pedigree-based phasing. EAGLE phased all of the pCHMs with 91% accuracy based on trio validation. DiscovEHR cohort has >35K 1st and 2nd degree relatives involving 53% of the cohort that we used to phase nearly a third of the pCHMs with ~100% accuracy, reducing inaccurate phasing by 31%.

In total, 38% of the pCHMs were phased in trans, yielding a high-confidence set of 15,040 rare, deleterious CHMs distributed among >11K individuals. Over 3,000 genes contain >=1 CHMs. When combined with 12,554 rare homozygous LoF and deleterious missense carriers, CHMs increased the number of genes with >=10 carriers from 181 to 629 (~250% increase). Only considering the 3,915 homozygous LoFs and the 1,307 LoF-LoF CHMs resulted in a 54% increase in the number of genes with >=10 carriers of gene KOs.

In conclusion, a large number of rare deleterious CHMs can be accurately phased using population allele frequencies and cryptic relationships to significantly augment the number of human gene KOs. When coupled to phenotypic data, these KOs may inform on our understanding of gene function in humans.

(We expect to have these results updated to our 90K person freeze by the time of presentation.)

← Back

SEQSPARK: AN ANALYSIS TOOL FOR LARGE SCALE SEQUENCE-BASED GENETIC EPIDEMIOLOGICAL STUDIES.

Di Zhang[1], Linhai Zhao[1], Biao Li[1], Zongxiao He[1], Gao T Wang[2], Dajiang J Liu[3], Suzanne M Leal[1]

[1]Baylor College of Medicine, Center for Statistical Genetics, Department of Molecular and Human Genetics, Houston, TX, [2]University of Chicago, Department of Human Genetics and Statistics, Chicago, IL, [3]Pennsylvania State University College of Medicine, Department of Public Health Sciences, Hershey, PA

Massively parallel sequencing technologies provide great opportunities for discovering rare susceptibility variants involved in complex disease etiology via large-scale imputation, exome and whole genome sequence based association studies. Due to modest effect sizes, large sample sizes of tens to hundreds of thousands of individuals are required for adequately powered studies. Current analytical tools are obsolete when it comes to handling these large datasets. To facilitate the analysis of large-scale sequence-based studies, we developed SEQSpark which implements parallel processing based on Spark to increase the speed and efficiency of performing data quality control, annotation and association analysis. To demonstrate the versatility and speed of SEQSpark, we analyzed whole genome sequence data from the UK10K, testing for associations with waist-to-hip ratio. The analysis which was completed in 1.5 hours, included loading data, annotation, principal component analysis, single variant and rare variant aggregate association analysis of >9 million variants. For rare variant aggregate analysis, an exome-wide significant association (P<2.5x10-6) was observed with *CCDC62* [SKAT-O (P=6.89x10-7), Combined Multivariate Collapsing (P=1.48x10-6) and Burden of Rare Variants (P=1.48x10-6)]. SEQSpark was also used to analyze 50,000 simulated exomes and it required 1.75 hours for the analysis of a quantitative trait using several rare variant aggregate association methods. Additionally, the performance of SEQSpark was compared to Variant Association Tools and PLINK/SEQ. SEQSpark was always faster and in some situations computation was reduced to a hundredth of the time. SEQSpark will empower large sequence-based epidemiological studies to quickly elucidate genetic variation involved in the etiology of complex traits.

← Back

THE BIOINFORMATICS OF LIQUID BIOPSIES: CELL-FREE DNA AS A BIOMARKER OF DISEASE AND AGING.

Yee Voan Teo[1], Miriam Capri[2,3], Cristina Morsiani[2,3], Claudio Franceschi[2,3], Nicola Neretti[1,4]

[1]Brown University, Department of Molecular Biology, Cell Biology, and Biochemistry, Providence, RI, [2]Universita' di Bologna, Department of Experimental, Diagnostic and Specialty Medicine (DIMES), Bologna, Italy, [3]Universita' di Bologna, Interdepartmental Center "L. Galvani" (C.I.G.), Bologna, Italy, [4]Brown University, Center for Computational Molecular Biology, Providence, RI

Liquid biopsies refer to the analysis of material obtained via minimally invasive techniques, such as blood extraction. The presence of fragments of cell-free DNA (cfDNA), combined with the latest sequencing technologies, can provide biomarkers with great translational potential. However, cfDNA provides unique analytical challenges as it typically originates from the apoptosis of many different cell types. Here we present a novel methodology to estimate 1) the tissue of origin and abundance of the different cell types, and 2) features linked to the epigenetic state of the cells of origin such as DNA accessibility. We rely on a key feature of cfDNA fragments, namely their ability to provide an in vivo genome-wise nucleosome footprint. We build an inference model linking the nucleosome relative spacing at gene sites with gene expression to determine the tissue of origin of the signal. We train our model on available tissue-specific RNA-seq datasets such as ENCODE and the Genotype-Tissue Expression (GTEx) project. We also define chromosome accessibility at different locations, in both euchromatin (e.g. gene regions) and heterochromatin (e.g. transposable elements). We demonstrate our method on cfDNA extracted from blood in a cohort of aging individuals, which includes extremely long-lived subjects (centenarians) and different health statuses, and provide the first in vivo evidence of global and local chromatin changes in human aging. Our results show that nucleosome signals inferred from cfDNA are consistent with the redistribution of heterochromatin observed in vitro and in other model systems. We also characterize the presence of apoptotic cells from tissues that can discriminate between ages and health status. Finally, we describe how the nucleosome footprint derived from cfDNA relates to the methylation state of CpG sites.

CHROMATIN LOOP ANCHORS ARE ASSOCIATED WITH GENOME INSTABILITY IN CANCER AND RECOMBINATION HOTSPOTS IN THE GERMLINE.

Vera B Kaiser, Colin A Semple

University of Edinburgh, IGMM, MRC HGU, Edinburgh, United Kingdom

Chromatin loops form a basic unit of interphase nuclear organisation, providing contacts between regulatory regions and target promoters, and forming higher level patterns defining self interacting domains. Recent studies have shown that mutations predicted to alter chromatin loops and domains are frequently observed in tumours and can result in the upregulation of oncogenes, but the combinations of selection and mutational bias underlying these observations remains unknown. Here, we explore the unusual mutational landscape associated with chromatin loop anchor points (LAPs), which are located at the base of chromatin loops and form a kinetic trap for cohesin. We show that LAPs are strongly depleted for single nucleotide variants (SNVs) in tumours, which is consistent with their relatively early replication timing. However, despite low SNV rates, LAPs emerge as sites of evolutionary innovation showing enrichment for structural variants (SVs). They harbour an excess of SV breakpoints in cancers, are prone to double strand breaks in somatic cells, and are bound by DNA repair complex proteins. Recurrently disrupted LAPs are often associated with genes annotated with functions in cell cycle transitions. An unexpectedly large fraction of LAPs (16%) also overlap known meiotic recombination hotspot (HSs), and are enriched for the core PRDM9 binding motif, suggesting that LAPs have been foci for diversity generated during recent human evolution. We suggest that the unusual chromatin structure at LAPs underlies the elevated SV rates observed, marking LAPs as sites of regulatory importance but also genomic fragility.

← Back

AN ULTRA-HIGH RESOLUTION CAPTURE-C PROMOTER 'INTERACTOME' IMPLICATES CAUSAL GENES AT SLE GWAS LOCI

Alessandra Chesi, Matthew E Johnson, Elisabetta Manduchi, Carole Le Coz, Michelle E Leonard, Sumei Lu, Kenyaita M Hodge, Neil D Romberg, Struan F Grant, Andrew D Wells

The Children's Hospital of Philadelphia, Philadelphia, PA

Genome Wide Association Studies (GWAS) have been successful in yielding >60 loci for Systemic Lupus Erythematosus (SLE). However, it is known that GWAS just reports genomic signals and not necessarily the precise localization of culprit genes, with eQTL efforts only able to infer causality to a minority of such loci. Thus, we sought to carry out physical and direct 'variant to gene mapping'. Chromatin conformation capture-based techniques that detect contacts between distant regions of the genome offer a powerful opportunity to understand GWAS signals that principally reside in non-coding regions, and thus likely acting as regulatory elements for neighboring genes. To move beyond analyzing one locus at a time and to improve on the low resolution of available Hi-C data, we developed a massively parallel, ultra-high resolution Capture-C based method to simultaneously characterize the genome-wide interactions of all human promoters in any cell type. We applied this method to study the promoter 'interactome' of primary human T Follicular Helper (TFH) cells from tonsils of healthy volunteers, a model relevant to SLE as TFH operate upstream of the activation of pathogenic autoantibody-producing B cells during the disease. We designed a custom Agilent SureSelect library targeting both ends of *DpnII* restriction fragments that overlap promoters of protein-coding, noncoding, antisense, snRNA, miRNA, snoRNA and lincRNA transcripts, totaling 36,691 RNA baited fragments. Each library was sequenced on multiple lanes of an Illumina HiSeq 4000 and then significant interactions were determined. We also generated ATAC-seq open chromatin maps from the same TFH samples to determine informative proxy SNPs for each of the 63 SLE GWAS loci (*Nat Genet* 48, 940-946, 2016). By intersecting our sub-1kb promoter 'interactome' data with SNPs from 33 candidate loci provided by the ATAC-seq experiments, we observed consistent contacts for at least 20 loci. Some 'nearest' genes to the sentinel SNP were supported e.g. *ARID5B* and *IKZF3*, while at other loci more distant genes were implicated e.g. *LCLAT1* at the '*LBH*' locus and the master TFH transcription factor *BCL6* at the '*LPP-TPRG1*' locus. In conclusion, we observed consistent contacts to at least 30% of SLE GWAS loci using the highest resolution promoter 'interactome' to date in a single, disease-relevant cell type. Only by establishing which genes such loci regulate in the correct cellular context can one truly translate GWAS findings.

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Dr. Aaron Adamson
City of Hope
aadamson@coh.org

Mr. Ricky Adkins
Institute for Genome Sciences
sadkins@som.umaryland.edu

Dr. Michael Agostino
Pfizer
michael.agostino@pfizer.com

Mr. Lin An
Pennsylvania State University
lua137@psu.edu

Mr. Jatin Arora
Max Planck Institute for Evolutionary Biology
arora@evolbio.mpg.de

Mr. Meharji Arumilli
University of Helsinki
meharji.arumilli@helsinki.fi

Mr. Hossein Asghari
Simon Fraser University
hasghari@sfu.ca

Mr. Gurnit Atwal
University Of Toronto
gurnit.atwal@mail.utoronto.ca

Dr. Xiaodong Bai
Regeneron Pharmaceuticals
xiaodong.bai@regeneron.com

Dr. Tracy Ballinger
University of Edinburgh
tracy.ballinger@igmm.ed.ac.uk

Prof. Armand Bankhead III
University of Michigan
bankhead@med.umich.edu

Mr. Mario Banuelos
University of California, Merced
mbanuelos4@ucmerced.edu

Dr. Sofia Barreira
National Human Genome Research Institute
sofia.barreira@nih.gov

Dr. Boris Bartholdy
Albert Einstein College of Medicine
boris.bartholdy@einstein.yu.edu

Dr. Philipp Berninger
Evolva
philippb@evolva.com

Dr. Holly Bik
University of California Riverside
hbik@ucdavis.edu

Ms. Dana Bis
University of Miami
dmb107@miami.edu

Dr. Daniel Blankenberg
Cleveland Clinic / CCL College of Medicine of CWRU
blanked2@ccf.org

Dr. Hidemasa Bono
Research Organization of Information and Systems
bono@dbcls.rois.ac.jp

Mr. Arthur Brady
Institute for Genome Sciences
ABrady@som.umaryland.edu

Dr. Stephen Brooks
NIAMS/NIH
brookss1@mail.nih.gov

Dr. Michele Busby
Gritstone Oncology
mbusby@gritstone.com

Dr. Scott Cain
Ontario Institute for Cancer Research
scott@scottcain.net

Dr. Michael Campbell
Cold Spring Harbor Laboratory
mcampbel@cshl.edu

Dr. Lesley Chapman
National Institute of Standards and Technology
lesley.chapman@nist.gov

Dr. Anirvan Chatterjee
Indian Institute of Technology Bombay
anirvan.chatterjee@iitb.ac.in

Mr. Youdinghuan Chen
Dartmouth College
youdinghuan.chen.gr@dartmouth.edu

Dr. Yibu Chen
University of Southern California
yibuchen@usc.edu

Dr. Zelin Chen
NIH
zelin.chen@nih.gov

Dr. Li Chen
Leidos Biomedical Research, Inc
li.chen2@nih.gov

Ms. Jingfei Cheng
Chinese Academy of Sciences
jfeicheng@sibs.ac.cn

Mr. Steve Chervitz
Personalis, Inc.
steve.chervitz@personalis.com

Dr. Alessandra Chesi
CHOP
chesia@email.chop.edu

Mr. Kashyap Chhatbar
University of Edinburgh
kashyap.c@ed.ac.uk

Ms. Manjusha Chintalapati
Max Planck Institute for Evolutionary Anthpology
m_chintalapati@eva.mpg.de

Ms. Eunji Choi
Yonsei University
yatookk@gmail.com

Mr. Kapeel Chougule
Cold Spring Harbor Laboratory
kchougul@cshl.edu

Dr. Chi-Yeh Chung
Salk Institute for Biological Studies
cchung@salk.edu

Ms. Katarzyna Chyzynska
University of Bergen
katarzyna.chyzynska@uib.no

Ms. Laura Clarke
EBI
laura@ebi.ac.uk

Dr. Manuel Corpas
Repositive Ltd
manuel@repositive.io

Ms. Paula Costa
Agilent Technologies
paula_costa@agilent.com

Mr. Jonathan Crabtree
University of Maryland School of Medicine
jcrabtree@som.umaryland.edu

Dr. Kathryn Crouch
University of Glasgow
kathryn.crouch@glasgow.ac.uk

Mr. Fabio Cumbo
The Pennsylvania State University
fabio.cumbo@iasi.cnr.it

Dr. Carla Cummins
EMBL-EBI
carlac@ebi.ac.uk

Mr. Taner Dagdelen
University of California, Berkeley
tkdagdelen@berkeley.edu

Dr. Ryan Dale
NIH
dalerr@niddk.nih.gov

Ms. Charlotte Darby
Johns Hopkins University
cdarby@jhu.edu

Dr. Carrie Davis
Stanford University
carried@stanford.edu

Dr. Sean Davis
NIH
seandavi@gmail.com

Dr. Niantao Deng
Garvan Institute of Medical Research
n.deng@garvan.org.au

Dr. Adnan Derti
Gritstone Oncology
aderti@gritstone.com

Mr. Nicolas Dierckxsens
Universite Libre de Bruxelles
nicolasdierckxsens@hotmail.com

Dr. Rebecca Dikow
Smithsonian Institution
DikowR@si.edu

Dr. Egor Dolzhenko
Illumina, Inc
edolzhenko@illumina.com

Dr. Diana Domanska
Oslo University
dianadom@ifi.uio.no

Ms. Marion Dubarry
Genoscope / CEA
mdubarry@genoscope.cns.fr

Dr. Jorge Duitama
Universidad de los Andes
ja.duitama@uniandes.edu.co

Dr. Jason Dumelie
Weill Cornell Medicine
jad2033@med.cornell.edu

Ms. Sara Dunaj
Translate Bio
sdunaj@translate.bio

Dr. Anindita Dutta
Illumina Inc
adutta1@illumina.com

Dr. Michael Eberle
Illumina, Inc
meberle@illumina.com

Dr. Zachary Ence
Brigham Young University
zac.ence@gmail.com

Mr. Karoly Erdos
EMBL-EBI
karoly@ebi.ac.uk

Dr. Yue Fan
Johnson & Johnson (China) Investment Ltd.
yfan40@its.jnj.com

Mr. Christopher Faulk
University of Minnesota
cfaulk@umn.edu

Mr. James Fitch
Nationwide Children's Hospital
James.Fitch@nationwidechildrens.org

Dr. Paul Flicek
EMBL-EBI
flicek@ebi.ac.uk

Dr. Liliana Florea
Johns Hopkins School of Medicine
florea@jhu.edu

Mr. Jerry Fong
Washington University School of Medicine
fongj@wustl.edu

Dr. Paul Frandsen
Smithsonian Institution
FrandsenP@si.edu

Dr. Ken Frazer
University of Oregon
ksf@uoregon.edu

Dr. Donald Freed
Sentieon Inc
donfreed12@gmail.com

Dr. Fengli Fu
Ciphergene
fujuli@gmail.com

Dr. Jeff Gaither
Nationwide Children's Hospital
Jeffrey.Gaither@nationwidechildrens.org

Dr. Molly Gale Hammell
Cold Spring Harbor Labs
mhammell@cshl.edu

Mr. Kiran Garimella
University of Oxford
kiran@well.ox.ac.uk

Dr. Bernat Gel Moreno
Fund. Inst. Germans Trias i Pujol
bernatgel@gmail.com

Dr. Mark Gerstein
Yale University
pi@gersteinlab.org

Dr. Francesca Giordano
Wellcome Trust Sanger Institute
fg6@sanger.ac.uk

Ms. Aishwarya Gogate
UT Southwestern Medical Center
aishwarya.gogate@utsouthwestern.edu

Dr. Madelaine Gogol
Stowers Institute for Medical Research
mcm@stowers.org

Dr. Giorgio Gonnella
University of Hamburg, Germany
gonnella@zbh.uni-hamburg.de

Mr. David Gordon
Nationwide Children's Hospital
david.gordon@nationwidechildrens.org

Ms. Steffi Grote
Max-Planck-Institute for Evolutionary Anthropology
steffi_grote@eva.mpg.de

Dr. Jose Afonso Guerra-Assuncao
University College London
a.guerra@ucl.ac.uk

Mr. Toby Gurran
University of Edinburgh
s1582371@sms.ed.ac.uk

Dr. Harendra Guturu
Ancestry.com LLC
hguturu@ancestry.com

Dr. Tanwir Habib
Sidra Medical and Research Center
thabib@sidra.org

Mr. Tom Hait
Tel Aviv University
sthait@gmail.com

Mr. Gisli Halldorsson
deCODE Genetics
gislih@decode.is

Mr. John Hamilton
Michigan State University
jham@msu.edu

Dr. Nathan Hammond
Stanford Health Care
nathanhammond@stanfordhealthcare.org

Dr. Oliver Hampton
Memorial Sloan Kettering Cancer Center
hamptono@mskcc.org

Dr. Nancy Hansen
National Human Genome Research Institute
nhansen@mail.nih.gov

Dr. Jason Harris
Personalis
jason.harris@personalis.com

Mr. Christopher Harris
Memorial Sloan Kettering Cancer Center
harrisc2@mskcc.org

Mr. James Havrilla
University of Utah
semjaavria@gmail.com

Dr. Tim Hefferon
NIH
theffero@mail.nih.gov

Dr. Javier Herrero
UCL Cancer Institute
javier.herrero@ucl.ac.uk

Ms. Angie Hinrichs
UC Santa Cruz
angie@soe.ucsc.edu

Mr. Phuc Hoang
The Institute of Cancer Research
Phuc.Hoang@icr.ac.uk

Dr. Guillaume Holley
University of Iceland
guillaumeholley@gmail.com

Ms. Jessica Holmes
University of Illinois at Urbana-Champaign
jholmes5@illinois.edu

Ms. Sufen Hu
University of Pennsylvania
sufenhu@upenn.edu

Dr. Toby Hunt
EMBL-EBI
toby@ebi.ac.uk

Ms. Elizabeth Hutton
Cold Spring Harbor Laboratory
ehutton@cshl.edu

Dr. Yinping Jiao
Cold Spring Harbor Laboratory
yjiao@cshl.edu

Dr. Kevin Johnson
The Jackson Laboratory for Genomic Medicine
kevin.c.johnson@jax.org

Mr. Patrick Jongeneel
Personalis, Inc.
patrick.jongeneel@personalis.com

Dr. Isaac Joseph
Agilent Technologies, Inc.
isaac.joseph4Aagilent.com

Prof. Anthony Joseph
University of California Berkeley
adj@cs.berkeley.edu

Dr. Andre Kahles
ETH Zurich
andre.kahles@inf.ethz.ch

Dr. Vera Kaiser
University of Edinburgh
vera.kaiser@igmm.ed.ac.uk

Dr. Chakravarthi Kanduri
University of Oslo
skanduri@ifi.uio.no

Dr. Maricel Kann
University of Maryland
mkann@umbc.edu

Dr. Konrad Karczewski
Massachusetts General Hospital/Broad Institute
konradk@broadinstitute.org

Dr. Ulya Karpuzcu
University of Minnesota Twin Cities
ukarpuzc@umn.edu

Dr. Birte Kehr
Berlin Institute of Health
birte.kehr@bihealth.de

Mr. Keffy Kehrli
Stony Brook University
keffy.kehrli@stonybrook.edu

Mr. Ben Kelly
Nationwide Children's Hospital
ben.kelly@nationwidechildrens.org

Dr. Janet Kelso
Max Planck Institute for Evolutionary Anthropology
kelso@eva.mpg.de

Dr. Sam Khalouei
Hospital for Sick Children
sam.khalouei@sickkids.ca

Mr. S. Karen Khatamifard
University of Minnesota
khatami@umn.edu

Dr. Seok-Won Kim
RIKEN
seokwon.kim@riken.jp

Ms. April Kim
Broad Institute
aprilkim@broadinstitute.org

Ms. Young Kim
Stony Brook University
young.c.kim@stonybrook.edu

Dr. Carl Kingsford
Carnegie Mellon University
carlk@cs.cmu.edu

Dr. Manjari Kiran
University of Virginia
mk9ua@eservices.virginia.edu

Dr. Jessica Kissinger
University of Georgia
jkissing@uga.edu

Dr. Paul Kitts
NIH/NLM/NCBI
kitts@ncbi.nlm.nih.gov

Mr. Antonin Klima
Norwegian University of Science and Technology
antonink@idi.ntnu.no

Mr. Sree Rohit Raj Kolora
University of Leipzig; iDiv
rohit@bioinf.uni-leipzig.de

Dr. Sergey Koren
National Institutes of Health
sergey.koren@nih.gov

Dr. Prachi Kothiyal
Inova Fairfax Hospital
prachi.kothiyal@gmail.com

Mr. Sam Kovaka
Johns Hopkins University
skovaka1@jhu.edu

Dr. Alper Kucukural
University of Massachusetts Medical School
alper.kucukural@umassmed.edu

Mr. Naveen Kumar
RIKEN Center for Integrative Medical Sciences
naveen.kumar@riken.jp

Dr. Vivek Kumar
CSHL
vkumar@cshl.edu

Mr. Grant Lammi
Nationwide Childrens Hospital
Grant.Lammi@nationwidechildrens.org

Dr. Ben Langmead
Johns Hopkins University
langmea@cs.jhu.edu

Dr. Delphine Lariviere
Penn State University
delphinel@galaxyproject.org

Ms. Kaitlin Laverty
University of Toronto
kaitlin.laverty@mail.utoronto.ca

Dr. Ryan Layer
University of Utah
ryan.layer@gmail.com

Dr. Suzanne Leal
Baylor College of Medicine
sleal@bcm.edu

Mr. Jinyoung Lee
Yonsei University
hd00ljy@naver.com

Dr. Jing Leng
Illumina
lengjingworld@gmail.com

Ms. Lixia Li
Merck
lixia_li@merck.com

Ms. Xiaotong Li
Yale University
xiaotong.li@yale.edu

Dr. Jayon Lihm
Cold Spring Harbor Laboratory
jlihm@cshl.edu

Mr. Hee-Woong Lim
University of Pennsylvania
heewlim@mail.med.upenn.edu

Dr. Chiao-Feng Lin
Harvard Medical School
cflin@bwh.harvard.edu

Mr. Xiaoxuan Lin
Cancer Science Institute of Singapore
lin.xx@u.nus.edu

Dr. Kuan-Ting Lin
Cold Spring Harbor Laboratory
klin@cshl.edu

Dr. Jonathan Ling
Johns Hopkins University
jling@jhu.edu

Dr. Zhi Liu
National Institutes of Health
zhi.liu@nih.gov

Dr. Jason Lloyd-Price
Broad Institute
jasonlp@broadinstitute.org

Dr. Lucas Lochovsky
Yale University
lucas.lochovsky@yale.edu

Dr. Jinfeng Lu
Columbia University College of P&S
jl5103@cumc.columbia.edu

Mr. Zhenyuan Lu
CSHL
luj@cshl.edu

Ms. Jennifer Lu
Johns Hopkins University
jlu26@jhmi.edu

Dr. David Lukatsky
Ben Gurion University of the Negev
lukatsky@bgu.ac.il

Mr. Richard Lupat
Peter Maccallum Cancer Centre
richard.lupat@petermac.org

Ms. Cong Ma
Carnegie Mellon University
congm1@andrew.cmu.edu

Dr. Anup Mahurkar
Institute for Genome Sciences
amahurkar@som.umaryland.edu

Dr. Guillaume Marcais
Carnegie Mellon University
gmarcais@cs.cmu.edu

Prof. Gabor Marth
University of Utah
gmarth@genetics.utah.edu

Dr. Alvaro Martinez Barrio
10x Genomics
ambarrio@10xgenomics.com

Dr. Sergio Martinez Cuesta
University of Cambridge
sermarcue@gmail.com

Dr. David Mayhew
GlaxoSmithKline
david.n.mayhew@gsk.com

Mr. Calvin McCarter
Carnegie Mellon University
calvinm@cmu.edu

Dr. Shane McCarthy
Wellcome Trust Sanger Institute
sm15@sanger.ac.uk

Ms. Carrie McCracken
University of Maryland, Baltimore
cmccracken@som.umaryland.edu

Mr. Warren McGee
Northwestern University
warren-mcgee@fsm.northwestern.edu

Dr. Adam McLain
SUNY Polytechnic Institute
mclaina@sunyit.edu

Mr. Will Mclaren
EMBL-EBI
wm2@ebi.ac.uk

Mr. Cory McLean
Google Inc.
cym@google.com

Prof. Paul Medvedev
Penn State
pzm11@psu.edu

Dr. Pall Melsted
University of Iceland
pmelsted@gmail.com

Dr. Vincent MEYER
CEA
vmeyer@cng.fr

Dr. Alison Meynert
University of Edinburgh
alison.meynert@igmm.ed.ac.uk

Mr. Jean-Michel Michno
University of Minnesota
mich0391@umn.edu

Dr. Christopher Miller
Washington University School of Medicine
c.a.miller@wustl.edu

Mr. David Molik
University of Notre Dame
dmolik@nd.edu

Dr. Scott Mottarella
Stanford Health Care
smottarella@stanfordhealthcare.org

Dr. Khyobeni Mozhui
University of Tennessee Health Science Center
kmozhui@uthsc.edu

Mr. Paul Muir
Yale University
paul.muir@yale.edu

Ms. Lauren Murray
Eli Lilly and Company
murrayln@mit.edu

Mr. Tulip Nandu
UT Southwestern Medical Center
tulip.nandu@utsouthwestern.edu

Dr. Jiri Nehyba
PerkinElmer
jiri.nehyba@perkinelmer.com

Dr. Anton Nekrutenko
Penn State
anton@nekrut.org

Dr. Nicola Neretti
Brown University
nicola_neretti@brown.edu

Dr. Zemin Ning
Wellcome Trust Sanger Institute
zn1@sanger.ac.uk

Mr. Frank Nothaft
UC Berkeley
fnothaft@berkeley.edu

Dr. Barbara Novak
Agilent Technologies
barbara_a_novak@agilent.com

Dr. Cihan Oguz
National Institutes of Health
cihan.oguz@nih.gov

Mr. Gavin Oliver
Mayo Clinic
oliver.gavin@mayo.edu

Mr. Dustin Olley
University of Maryland School of Medicine
jorvis@gmail.com

Mr. Andrew Olson
Cold Spring Harbor Laboratory
olson@cshl.edu

Mr. Joshua Orvis
Univ. of Maryland School of Medicine
jorvis@gmail.com

Ms. Natasha Pacheco
University of Alabama at Birmingham
npacheco@uab.edu

Prof. Lior Pachter
UC Berkeley
lpachter@math.berkeley.edu

Dr. Ashutosh Pandey
GlaxoSmithKline
ashutosh.k.pandey@gsk.com

Ms. Swati Parekh
Ludwig-Maximilians University Munich
parekh@bio.lmu.de

Dr. Donghyun Park
Samsung Medical Center
eastwise37@gmail.com

Mr. Nathaniel Parke
Unite Genomics
nate@unitegenomics.com

Mr. Mateus Patricio
EMBL-EBI
mateus@ebi.ac.uk

Dr. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Mr. Devin Petersohn
UC Berkeley
devin@eecs.berkeley.edu

Dr. Lon Phan
NIH
lonphan@mail.nih.gov

Dr. Adam Phillippy
National Human Genome Research Institute
adam.phillippy@nih.gov

Dr. Stephen Piccolo
Brigham Young University
stephen_piccolo@byu.edu

Mr. Thomas Powell
Heidelberg Institute for Theoretical Studies
sean.powell@h-its.org

Mr. Jacob Pritt
Johns Hopkins University
jacobpritt@gmail.com

Dr. Jane Pulman
The University of Liverpool
jane.pulman@liverpool.ac.uk

Meifang Qi
Chinese Academy of Sciences
qimeifang@sibs.ac.cn

Ms. Freda Qi
University of Western Ontario
fqi2@uwo.ca

Dr. Javier Quilez Oliete
CRG-Centre for Genomic Regulation
javier.quilez@crg.eu

Dr. Aaron Quinlan
University of Utah
aaronquinlan@gmail.com

Dr. Srividya Ramakrishnan
Johns Hopkins University
srividya.ramki@gmail.com

Dr. Swathi Ramakrishnan
Roswell Park Cancer Institute
swathi.ramakrishnan@roswellpark.org

Dr. Arun Ramani
Hospital for Sick Children
arun.ramani@sickkids.ca

Mr. Satishkumar Ranganathan Ganakammal
Missions Hospital / Clemson University
satishk@clemson.edu

Mr. Samarth Rangavittal
The Pennsylvania State University
samarthrvittal@gmail.com

Dr. Rozita Razavi
University of Toronto
rozy.razavi@gmail.com

Dr. Kris Richardson
Whitehead Institute
krichard@wi.mit.edu

Dr. Elena Rivas
Harvard University
elenarivas@fas.harvard.edu

Dr. Nicolas Robine
New York Genome Center
nrobine@nygenome.org

Mr. Scott Ronquist
University of Michigan
scotronq@umich.edu

Dr. Jeffrey Rosenfeld
Rutger Cancer Institute of NJ
jeffrey.rosenfeld@rutgers.edu

Mr. Sebastian Roskosch
Berlin Institute of Health
sebastian.roskosch@bihealth.de

Ms. Yulia Rubanova
University of Toronto
rubanova@cs.toronto.edu

Ms. Pamela Russell
Colorado School of Public Health
pamela.russell.ucdenver@gmail.com

Ms. Nil Sahin
University of Toronto
nil.sahin@mail.utoronto.ca

Dr. Kristoffer Sahlin
Pennsylvania State University
krsahlin@gmail.com

Dr. Stefania Salvatore
University of Oslo
stefasal@ifi.uio.no

Mr. Florian Sandron
CEA
sandron@cng.fr

Dr. David Sankoff
University of Ottawa
sankoff@uottawa.ca

Dr. Alexander Sasse
University of Toronto
alexander.sasse@mail.utoronto.ca

Dr. Christopher Saunders
Illumina
csaunders@illumina.com

Dr. Michael Schatz
CSHL and JHU
mschatz@cshl.edu

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@mail.nih.gov

Mr. Max Schubach
Berlin Institute of Health (BIH)
max.schubach@bihealth.de

Mr. Gunnar Schulze
University of Bergen
Gunnar.Schulze@uib.no

Dr. Fritz Sedlazeck
Baylor College of Medicine
Fritz.Sedlazeck@bcm.edu

Mr. Vitaly Sedlyarov
CeMM GmbH
office@cemm.oeaw.ac.at

Dr. Chris Seidel
Stowers Institute For Medical Research
cws@stowers.org

Mr. Fayaz Seifuddin
National Heart Lung and Blood Institute
hillary.flowers@nih.gov

Dr. Colin Semple
University of Edinburgh
colin.semple@igmm.ed.ac.uk

Dr. Preyas Shah
10x Genomics
preyas.shah@10xgenomics.com

Mr. Ronak Shah
Northwell Health
rshah22@northwell.edu

Dr. Shipra Sharma
IIT, Bombay
shiprasharma10@gmail.com

Dr. Maria Shatz
National Institute of Environmental Healts
shatzm@niehs.nih.gov

Ms. Rachel Sherman
Johns Hopkins University
rsherman@jhu.edu

Dr. Heejung Shim
University of Melbourne
heejung.shim@unimelb.edu.au

Mr. SeungHo Shin
Samsung Medical Center
sin12ho@gmail.com

Dr. Matthew Shirley
Novartis Institute for Biomedical Research
matt_d.shirley@novartis.com

Mr. Raunak Shrestha
University of British Columbia
rshrestha@prostatecentre.com

Dr. Shengqiang Shu
DOE-JGI
sqshu@lbl.gov

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Jared Simpson
Ontario Institute for Cancer Research
Jared.Simpson@oicr.on.ca

Mr. Angad Singh
Novartis Institutes for Biomedical Research
angad.singh@novartis.com

Ms. Smruthy Sivakumar
UT MD Anderson Cancer Center
ssivakumar@mdanderson.org

Mr. Nicholas Skvir
Brown University
nicholas_skvir@brown.edu

Ms. Jennifer Smith
Medical College of Wisconsin
jrsmith@mcw.edu

Dr. Ramakrishna Sompallae
Iowa Institute of Human Genetics
rama-sompallae@uiowa.edu

Dr. Dae-Soon Son
Samsung Medical Center
biostat.sait@gmail.com

Dr. Jawon Song
University of Texas at Austin
jawon@tacc.utexas.edu

Mr. Li Song
Johns Hopkins University
lsong10@jhu.edu

Dr. Daniel Standage
University of California, Davis
dsstandage@ucdavis.edu

Dr. Jeffrey Staples
Regeneron Genetics Center
jeffrey.c.staples@gmail.com

Dr. Oliver Stegle
European Molecular Biology Laboratory
oliver.stegle@ebi.ac.uk

Dr. Frederick Tan
Carnegie Institution
tan@ciwemb.edu

Dr. Hidenori Tanaka
Toyota Central RD Labs., Inc.
e1613@mosk.tytlabs.co.jp

Dr. Lin Tang
Springer Nature
lin.tang@nature.com

Dr. Todd Taylor
RIKEN IMS
taylor@riken.jp

Dr. James Taylor
Johns Hopkins University
james@jamestaylor.org

Dr. David Tegay
NYIT College of Osteopathic Medicine
dtegay@nyit.edu

Dr. Marcela Tello-Ruiz
Cold Spring Harbor Laboratory
mmonaco@cshl.edu

Mr. Malte Thodberg
University of Copenhagen
malte.thodberg@bio.ku.dk

Dr. Ilker Tunc
National Heart Lung and Blood Institute
hillary.flowers@nih.gov

Dr. Stephen Turner
University of Virginia
sdt5z@virginia.edu

Dr. Monika Tutaj
Medical College of Wisconsin
motutaj@mcw.edu

Mr. Arjun Vadapalli
Agilent Technologies
arjun.vadapalli@agilent.com

Ms. Deepali Vasoya
University of Edinburgh
Deepali.Vasoya@roslin.ed.ac.uk

Mr. Rahulsimham Vegesna
Pennsylvania State University
v.rahul.simham@gmail.com

Dr. Daniel Vera
Florida State University
vera@genomics.fsu.edu

Mr. Neil Versel
GenomeWeb
nversel@genomeweb.com

Ms. Beate Vieth
LMU Munich
vieth@bio.lmu.de

Mr. Coby Viner
University of Toronto
cviner@cs.toronto.edu

Dr. Kristoffer Vitting-Seerup
University of Copenhagen, Denmark
kristoffer.vittingseerup@bio.ku.dk

Mr. Justin Wagner
University of Maryland
jwagner@cs.umd.edu

Mr. Yinpeng Andy Wang
The University of Hong Kong
yinpeng@hku.hk

Dr. Paul Wang
University of Pennsylvania
zhipwang@upenn.edu

Dr. Jie Wang
Michigan State University
wangjie6@msu.edu

Ms. Haozhe Wang
The University of Texas at Dallas
hxw111830@utdallas.edu

Mr. Jiayao Wang
Columbia University
jw3514@cumc.columbia.edu

Dr. Qingyu Wang
Agilent Technologies
qingyu.wang@agilent.com

Ms. Yuejun Wang
Institute of Plant Physiology & Ecology, SIBS, CAS
yjwang02@sibs.ac.cn

Dr. Daifeng Wang
Stony Brook University
daifeng.wang@stonybrookmedicine.edu

Mr. George Wang
Cold Spring Harbor Laboratory
georgeleewang@gmail.com

Dr. James Wasmuth
University of Calgary
jwasmuth@ucalgary.ca

Mr. Sergiusz Wesolowski
Florida State University
wesserg@protonmail.com

Dr. Peter White
Nationwide Children's Hospital
peter.white@nationwidechildrens.org

Mr. Christopher Wilks
Johns Hopkins University
cwilks3@jhu.edu

Dr. Paul Williams
Thermo Fisher Scientific
paul.d.williams@thermofisher.com

Dr. Melissa Wilson Sayres
Arizona State University
Melissa.Wilsonsayres@asu.edu

Mr. Jeff Wintersinger
University of Toronto
jeff@wintersinger.org

Mr. Adam Wright
Ontario Institute for Cancer Research
adam.j.wright82@gmail.com

Dr. Thomas Wu
Genentech, Inc.
twu@gene.com

Ms. Dana Wyman
University of California, Irvine
dwyman@uci.edu

Dr. Guanjue Xiang
Pennsylvania State University
gzx103@psu.edu

Mr. Yaxin Xue
University of Bergen
yaxin.xue@uib.no

Dr. Chengfei Yan
Yale University
chengfei.yan@yale.edu

Dr. Hsih-Te Yang
The Feinstein Institute for Medical Research
hyang4@northwell.edu

Mr. Guangyu Yang
Johns Hopkins University
gyang22@jhu.edu

Dr. Ping Ye
Avera Cancer Institute
ping.ye@avera.org

Dr. Robert Young
University of Edinburgh
robert.young@igmm.ed.ac.uk

Mr. Adamo Young
University of Toronto
adamo.young@mail.utoronto.ca

Mr. Luke Zappia
The University of Melbourne
luke.zappia@mcri.edu.au

Prof. Zhonghua Zhang
Chinese Academy of Agricultural Sciences
zhangzhonghua_79@163.com

Ms. Yue Zhang
University of Ottawa
yzhan481@uottawa.ca

Dr. Junjun Zhang
Ontario Institute for Cancer Research
junjun.zhang@oicr.on.ca

Dr. Fei Zheng
St Jude Children's Hospital
fei.zheng@stjude.org

Ms. Qian Zhou
Chinese Academy of Agricultural Sciences
zhouqian_solab@163.com

Ms. Kirby Ziegler
University of Alberta
kziegler@ualberta.ca

Dr. Aleksey Zimin
University of Maryland
aleksey.zimin@gmail.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	**631-351-2000** **(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# 61360** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail and printing in the business center area
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 61360 (then enter #)**

Meetings & Courses Front Office

Hours during meetings: 8am – 7pm, until 9pm on arrival day
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**

Local Interest

Fish Hatchery	631-692-6758
Sagamore Hill	516-922-4788
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**)
Executive Limo Service	631-696-8000
Super Shuttle	800-957-4533 (**1033**)
US Limousine Service	800-962-2827,ext:3 **(1047)**

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	877-690-5114
New Jersey Transit	973-275-5555

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
British Airways	800-247-9297
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Southwest Airlines	800-435-9792
United	800-241-6522
Virgin American	877-359-9792