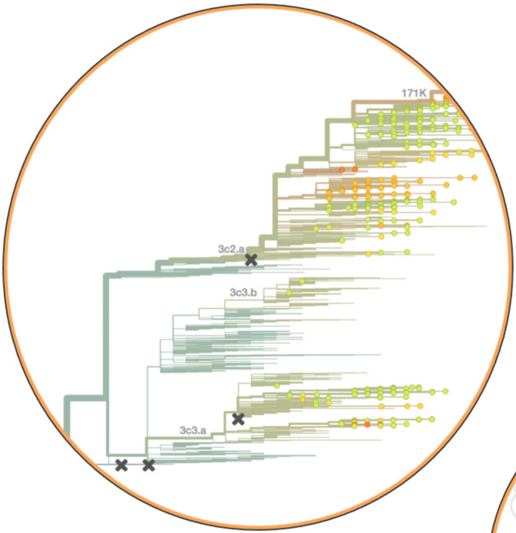
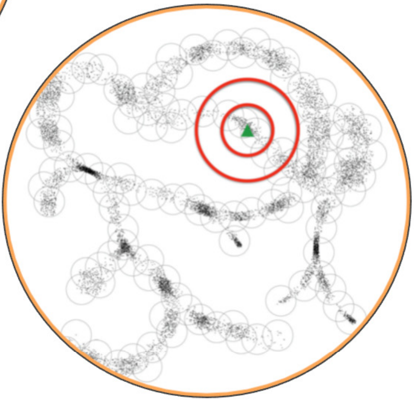
Abstracts of papers presented
at the 2016 meeting on

BIOLOGICAL DATA SCIENCE

October 26–October 29, 2016





$$H_A \;:\; \log_2(FPKM_i + 1) = \beta_0^* + \sum_{t=1}^{4} \beta_t^* \mathrm{spline}_t(RIN_i) + \sum_{p=1}^{5} \gamma_p^* 1(Pop_i = p) + \eta^* q75_i + \epsilon_i^*$$

$$H_0 \;:\; \log_2(FPKM_i + 1) = \beta_0 + \sum_{p=1}^{5} \gamma_p 1(Pop_i = p) + \eta q75_i + \epsilon_i$$

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2016 meeting on

BIOLOGICAL DATA SCIENCE

October 26–October 29, 2016

Arranged by

Bonnie Berger, *Massachusetts Institute of Technology*
Jeff Leek, *Johns Hopkins University*
Michael Schatz, *Cold Spring Harbor Laboratory*

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 Affiliate

Ionis Pharmaceuticals

BIOLOGICAL DATA SCIENCE

Wednesday, October 26 – Saturday, October 29, 2016

Wednesday	7:30 pm	**1** Software for Biologists
Thursday	9:00 am	**2** Machine Learning I
Thursday	11:00 am	**Master Lecture**
Thursday	1:30 pm	**3** Machine Learning II
Thursday	2:30 pm	**Education Forum**
Thursday	3:30 pm	**4** Poster Session I
Thursday	4:30 pm	*Wine and Cheese Party*
Thursday	7:30 pm	**5** Compute Infrastructure
Friday	9:00 am	**6** Algorithmics
Friday	1:30 pm	**7** Human Biology
Friday	4:30 pm	**Keynote Speaker**
Friday	5:30 pm	**8** Poster Session II and Cocktails
Friday	7:00 pm	Banquet
Saturday	9:00 am	**9** Data Infrastructure

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

Abstracts are the responsibility of the author(s) and publication of an abstract does not imply endorsement by Cold Spring Harbor Laboratory of the studies reported in the abstract.

These abstracts should not be cited in bibliographies. Material herein should be treated as personal communications and should be cited as such only with the consent of the author.

Please note that ANY photography or video/audio recording of oral presentations or individual posters is strictly prohibited except with the advance permission of the author(s), the organizers, and Cold Spring Harbor Laboratory.

Printed on 100% recycled paper.

PROGRAM

WEDNESDAY, October 26—7:30 PM

Opening Remarks

SESSION 1 SOFTWARE FOR BIOLOGISTS

Chairpersons: **Rafael Irizarry,** Dana-Farber Cancer Institute, Boston, Massachusetts
Martin Morgan, Roswell Park Cancer Institute, Buffalo, New York

Overcoming bias and systematic error in high-throughput technologies are key for the success of data science
Rafael Irizarry.
Presenter affiliation: Dana Farber Cancer Institute, Boston, Massachusetts. 1

How To Train Your DragoNN (Deep Regulatory Genomic Neural Network)
Johnny Israeli, Michael Wainberg, Avanti Shrikumar, Nasa Sinnott-Armstrong, Anna Scherbina, Irene Kaplow, Rahul Mohan, Chuan-Sheng Foo, Anshul Kundaje.
Presenter affiliation: Stanford University, Stanford, California. 2

Let the right ones in—The cost and benefit of including alternate alleles in the reference genome
Jacob Pritt, Ravi Gaddipati, Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland; Johns Hopkins University, Baltimore, Maryland. 3

Processing terabyte-scale genomics datasets with ADAM and Toil
Frank A. Nothaft, Uri Laserson, David Haussler, Benedict Paten, David A. Patterson, Anthony D. Joseph.
Presenter affiliation: University of California, Berkeley, Berkeley, California. 4

Traditional approaches and contemporary challenges to production of useful scientific software—Lessons from *Bioconductor*
Martin Morgan.
Presenter affiliation: Roswell Park Cancer Institute, Buffalo, New York. 5

MetaR—Towards an R notebook with composable languages
Fabien Campagne, Alexander Pann, William E. Digan, Manuele Simi.
Presenter affiliation: Weill Cornell Medicine, New York, New York. 6

Apollo—Collaborative and scalable manual genome annotation
Nathan A. Dunn, Monica Muñoz-Torres, Deepak Unni, Eric Yao, Colin Diesch, Ian Holmes, Chris Elsik, Suzanna Lewis.
Presenter affiliation: Lawrence Berkeley National Laboratory, Berkeley, California. 7

Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation
Han Fang, Yifei Huang, Aditya Radhakrishnan, Max Doerfel, Adam Siepel, Rachel Green, Gholson Lyon, Michael Schatz.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 8

THURSDAY, October 27—9:00 AM

SESSION 2 MACHINE LEARNING I

Chairpersons: **Barbara Engelhardt,** Princeton University, New Jersey.
Michael Hoffman, University of Toronto, Canada

Exploiting transcriptional time-series data to understand the drivers of cellular response to exposure
Barbara Engelhardt.
Presenter affiliation: Princeton University, New Jersey.

Estimation of nucleotide- and allele-specific selection coefficients in the human genome using deep learning and population genetics
Yifei Huang, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 9

Winner’s Curse in quantitative genomics studies
Gregory Darnell, Jenny Tung, Christopher Brown, Sayan Mukherjee, Barbara Engelhardt.
Presenter affiliation: Princeton University, Princeton, New Jersey. 10

Not just a black box—Interpretable deep learning for genomics
Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, Johnny Israeli, Nasa Sinnott-Armstrong, Anshul Kundaje.
Presenter affiliation: Stanford University, Stanford, California. 11

Reconstructing changes in mutational processes during tumour evolution
Yulia Rubanova, Jeff Wintersinger, Amit Deshwar, Nil Sahin, Quaid Morris.
Presenter affiliation: University of Toronto, Toronto, Canada; Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada. 12

THURSDAY, October 27—11:00 AM

MASTER LECTURE

Olga Troyanskaya
Princeton University

THURSDAY, October 27—1:30 PM

SESSION 3 MACHINE LEARNING II

Chairpersons: **Barbara Engelhardt,** Princeton University, New Jersey.
Michael Hoffman, University of Toronto, Canada

Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet
Coby Viner, James Johnson, Nicolas Walker, Hui Shi, Marcela Sjöberg, David J. Adams, Anne C. Ferguson-Smith, Timothy L. Bailey, Michael M. Hoffman.
Presenter affiliation: University of Toronto, Toronto, Canada. 13

CERES—A model for inferring genetic dependencies in cancer cell lines from CRISPR knockout screens
Jordan G. Bryan, Robin M. Meyers, Aviad Tsherniak.
Presenter affiliation: The Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 14

Combining phenome and genome to uncover the genetic basis for naturally occurring differences in development and behavior
Tiffany A. Timbers, Catrina Loucks, Stephane Flibotte, Don G. Moerman, Michel R. Leroux.
Presenter affiliation: University of British Columbia, Vancouver, Canada. 15

THURSDAY, October 27—2:30 PM

EDUCATION FORUM

Moderated by: **Michael Gerstein,** Yale University

THURSDAY, October 27—3:30 PM

SESSION 4 POSTER SESSION I

Universal microbial diagnostics using random DNA probes
Amirali Aghazadeh, Adam Y. Lin, Mona A. Sheikh, Allen L. Chen, Lisa M. Atkins, Coreen L. Johnson, Joseph F. Petrosino, Rebekah A. Drezek, Richard G. Baraniuk.
Presenter affiliation: Rice University, Houston, Texas. 16

A computational framework to predict chromatin interaction using genomic and epigenomic data
Lin An, Tyler Derr, Yanli Wang, Feng Yue.
Presenter affiliation: Pennsylvania State University, State College, Pennsylvania. 17

A convolutional neural network framework for modelling cis-regulatory elements with application to RNA stability
Žiga Avsec, Julien Gagneur.
Presenter affiliation: Technical University of Munich, Munich, Germany; Ludwig Maximilian University of Munich, Munich, Germany. 18

Investigating how the transcription factor *fkh-8* is expressed in BAG neurons and the dopaminergic pathway
Mohammed O. Awan, Bryan E. Cawthon, Brian L. Nelms.
Presenter affiliation: Fisk University, Nashville, Tennessee. 19

Transposable element-mediated gene regulation in the parasite *Trichomonas vaginalis*
Martina Bradic, Sally Warring, Jonathan Bermeo, Jane Carlton.
Presenter affiliation: Center for Genomics and Systems Biology, New York, New York. 20

Fast annotation of metagenomic shotgun sequences with a microbial gene catalog and Bowtie2
Stuart M. Brown, Konstantinos Krampis, Hao Chen, Yuhan Hao.
Presenter affiliation: New York University School of Medicine, New York, New York. 21

Planemo—A scientific workflow SDK
John M. Chilton, Aysam Guerler, the Galaxy Team.
Presenter affiliation: The Pennsylvania State University , University Park, Pennsylvania. 22

Methods for rapid and scalable quality assessment of RNA structure probing data
Krishna Choudhary, Huan Chen, Luyao Ruan, Nathan P. Shih, Fei Deng, Sharon Aviran.
Presenter affiliation: University of California Davis, Davis, California. 23

Benchmarking alignment and quantification tools on plant RNA-seq data with Cyverse cyberinfrastructure
Kapeel M. Chougule, Liya Wang, Joshua Stein, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; CyVerse, Tucson, Arizona. 24

Integrative approaches for variant interpretation in coding regions
Declan Clarke, Sushant Kumar, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 25

Pollen counting and analysis with the Attune acoustic flow cytometer
Schuyler Corry, Rachel K. Smith, Barb Seredick.
Presenter affiliation: Thermo Fisher Scientific, Eugene, Oregon. 26

Computational analysis of endogenous retroviral element expression during the development of human embryos
Engin Cukuroglu, Jonathan Goke.
Presenter affiliation: Genome Institute of Singapore, Singapore. 27

Evolinc—A computational pipeline for comparative genomic and transcriptomic analyses of long non-coding RNAs from large RNA-Seq datasets
Upendra Kumar Devisetty, Andrew D. Nelson, Asher K. Haug-Baltzell, Eric Lyons, Mark A. Beilstein.
Presenter affiliation: CyVerse, University of Arizona, Tucson, Arizona. 28

A whole-genome phylogenetic hypothesis across the three domains of life
Rebecca B. Dikow, Katrina M. Pagenkopp Lohan, Paul B. Frandsen.
Presenter affiliation: Smithsonian Institution, Washington, DC. 29

Building a biological data science infrastructure at the Smithsonian Institution
Paul B. Frandsen, Rebecca B. Dikow.
Presenter affiliation: Smithsonian Institution, Washington, DC. 30

Community computational challenges in biological data science—An optimistically cautionary tale
Iddo Friedberg.
Presenter affiliation: Iowa State University, Ames, Iowa. 31

Complete site saturation of *Bacillus subtilis* Lipase A to study the influence of single amino acids on the overall stability in surfactants
Alexander Fulton, Josiane Frauenkron-Machedjou, Pia Skoczinski, Susanne Wilhelm, Uli Schwaneberg, Karl-Erich Jaeger.
Presenter affiliation: Novozymes A/S, Bagsvaerd, Denmark. 32

***Arid1a* mutations in an *Apc/Pten*-deficient mouse model of ovarian tumors profoundly alter DNA methylation and transcription**
Nicholas Giangreco, Hanna M. Petrykowska, Alexandra Scott, Valer Gotea, Gennady Margolin, Cho R. Kathleen, Laura Elnitski.
Presenter affiliation: National Human Genome Research Institute, Rockville, Maryland; Columbia University Medical Center, New York, New York. 33

Towards "dry side" reproducibility—A technical survey of the challenges to and building blocks for reproducible, secure, and usable (microbial) genomics data infrastructure
Nicholas Greenfield, Samuel Minot, Roderick Bovee.
Presenter affiliation: One Codex, San Francisco, California. 34

In vivo characterization of Linc-p21 reveals functional cis-regulatory DNA elements
Abigail F. Groff, Diana B. Sanchez-Gomez, Marcela M. Soruco, Chiara Gerhardinger, James C. Lee, John L. Rinn.
Presenter affiliation: Harvard University, Cambridge, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 35

The quantitative relationship between histone modifications and gene expression across different individuals
Kipper Fletez-Brant, Kasper D. Hansen.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 36

Use of low coverage short read sequence data for quality control and familial relationship identification
Nancy F. Hansen, Benjamin Heil, Miriam K. Minsk, Sean Conlan, Pamela J. Thomas, Julia A. Segre, James C. Mullikin.
Presenter affiliation: National Human Genome Research Institute, Bethesda, Maryland. 37

Prediction of essential coding and noncoding elements using CRISPR
Elizabeth Hutton, Chris Vakoc, Junwei Shi, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 38

Shell scripting-based mathematical modeling for transcriptome data with bootstrapping and parallel computing
Kazuo Ishii, Kazutaka Nakamura, Nobuaki Tounaka.
Presenter affiliation: Tokyo University of Agriculture and Technology, Tokyo, Japan. 39

Analysis Tool for Annotated Variants—A comprehensive platform for population-scale genomic analyses
Zhong Ren, Petrovski Slave, Cirulli T. Elizabeth, Quanli Wang, Copeland Brett, Bridgers Joshua, Gussow Ayal, Erin Heinzen-Cox, Andrew Allen, David B. Goldstein, Sitharthan Kamalakaran.
Presenter affiliation: Columbia University, New York, New York. 40

FANTOM5 SSTAR—A challenge to establish a continuously extensible database for a large scale data production project
Imad Abugessaisa, Hisashi Shimoji, Serkan Sahin, Atsushi Kondo, Jayson Harshbarger, Marina Lizio, Yoshihide Hayashizaki, Piero Carninci, Alistair Forrest, the FANTOM5 Consortium, Takeya Kasukawa, Hideya Kawaji.
Presenter affiliation: RIKEN, Yokohama, Japan. 41

Hi-C data compression
Minji Kim, Olgica Milenkovic.
Presenter affiliation: University of Illinois at Urbana-Champaign, Urbana, Illinois. 42

pVAC-Seq—Development of an open-source software pipeline for tumor neoantigen discovery and personalized immunotherapy
Susanna Kiwala, Jasreet Hundal, Aaron Graubert, Joshua McMichael, Adam Coffman, Jason Walker, Christopher A. Miller, Obi L. Griffith, Elaine R. Mardis, Malachi Griffith.
Presenter affiliation: Washington University School of Medicine, St. Louis, Missouri. 43

Dolphin—A graphical user interface for the analysis and processing of high throughput genomics
Alper Kucukural, Nicholas Merowsky, Manuel Garber.
Presenter affiliation: University of Massachusetts Medical School, Worcester, Massachusetts. 44

Analysis and synthesis of gene regulatory networks via formal reasoning—Main applications and future challenges
Hillel Kugler.
Presenter affiliation: Bar-Ilan University, Ramat-Gan, Israel; Microsoft Research, Cambridge, United Kingdom. 45

Gtracks—A framework for creating and maintaining UCSC track databases using google spreadsheets
Kathleen E. Kyle, Hank W. Bass, Daniel L. Vera.
Presenter affiliation: Florida State University, Tallahassee, Florida. 46

Comparison of protein similarity network topologies using sequence- and active site-similarity edge metrics
Janelle B. Leuthaeuser, Julia Hayden, Nicholas Biffis, Jacquelyn S. Fetrow.
Presenter affiliation: University of Richmond, Richmond, Virginia. 47

Optimized contigs assembly and SNP discovery by overlapping paired-end RAD sequencing in roughskin sculpin (*Trachidermus fasciatus* Heckel)
Yulong Li, Dongxiu Xue, Jinxian Liu.
Presenter affiliation: Institute of Oceanology, Qingdao, China. 48

The TCGA data analysis for the function and regulation of the protein tyrosine phosphatase superfamily in human cancers
Suh-Yuen Liang, Tzu-Hao Chang, Tsung-Hsien Pu, Chia-Cheng Chou, Tzu-Ching Meng.
Presenter affiliation: Academia Sinica, Taipei, Taiwan. 49

Genome-wide SNP discovery and preliminary population genetic analysis in Japanese eel (*Anguilla japonica*) based on RAD sequencing
Bingjian Liu.
Presenter affiliation: Institute of Oceanology, Qingdao, China. 50

Simultaneous detection of SNPs and Indels using a 16-genotype probabilistic model
Ruibang Luo, Steven L. Salzberg, Michael C. Schatz.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 51

The CellFinder on-line data resource and its applications for stem cell research
Nancy Mah, Khadija El Amrani, Fritz Lekschas, Stefanie Seltmann, Marie Bittner, Harald Stachelscheid, Miguel Andrade-Navarro, Andreas Kurtz.
Presenter affiliation: Charite University Medicine Berlin, Berlin, Germany. 52

GRASS—Graph regularized annotation via semi-supervised learning
Laraib I. Malik, Rob Patro.
Presenter affiliation: Stony Brook University, New York. 53

Improved repeat alignment by simultaneous estimation of variation and alignment
Wilson H. McKerrow, Charles E. Lawrence.
Presenter affiliation: Brown University, Providence, Rhode Island. 54

Rapid, accurate, and actionable identification of pathogenic and commensal microorganisms, antibiotic resistance profiles, and virulence factors from metagenomic samples
Kelly Moffat, Nur A. Hasan, Poorani Subramanian, Huai Li, Christopher E. Mason, Rita R. Colwell.
Presenter affiliation: CosmosID, Inc, Metagenomics, Rockville, Maryland. 55

Deployment of a bioinformatics analytics platform in the cloud
David Molik, Yu-Jui Ho, Ying Jin, Molly Hammell.
Presenter affiliation: Cold Spring Harbor Lab, Cold Spring Harbor, New York. 56

Conseq—A metadata-driven tool for executing bioinformatic pipelines
Philip G. Montgomery.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts. 57

THURSDAY, October 27—4:30 PM

Wine and Cheese Party

THURSDAY, October 27—7:30 PM

SESSION 5 COMPUTE INFRASTRUCTURE

Chairpersons: **Nancy Cox,** Vanderbilt University Medical Center, Nashville, Tennessee
Jeremy Goecks, George Washington University, Ashburn, Virginia

Novel discovery from a gene by medical phenome catalog
Nancy J. Cox.
Presenter affiliation: Vanderbilt University Medical Center, Nashville, Tennessee. 58

Experiences migrating large scale bioinformatics data pipelines to a cloud infrastructure
Thomas Eastman, Claire Gu, Steven Herrin, Tulasi Paradarami, Kiran Singh, Andrew Stiles, Neil Thomas, Patrick Yee.
Presenter affiliation: 23andMe, Mountain View, California. 59

Reactome pathway analysis—A high-performance in-memory approach
Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, Pablo Marin-Garcia, Peter D’Eustachio, Lincoln Stein, Henning Hermjakob.
Presenter affiliation: European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom. 60

Thousands of public epigenomic datasets can be exploited through the Galaxy-GenAP project and the GeEC tool
Jonathan Laperle, David A. Morais, Michel Barrette, Charlotte Bastin, Marc-Antoine Robert, Marie Harel, Alexei N. Markovits, David Bujold, Alain Veilleux, Guillaume Bourque, Pierre-Étienne Jacques.
Presenter affiliation: Universite de Sherbrooke, Sherbrooke, Canada. 61

Real-time data analysis in the Galaxy platform
Jeremy Goecks.
Presenter affiliation: George Washington University, Ashburn, Virginia.

How to quantify expression for thousands of RNA-Sequencing samples in a day for a minimal cost
PJ Tatlow, Stephen R. Piccolo.
Presenter affiliation: Brigham Young University, Provo, Utah. 62

The KnowEnG platform for biological data analytics on the Cloud—A use case in gene prioritization for pharmacogenomics
Amin Emad, Junmei Cairns, Liewei Wang, Saurabh Sinha.
Presenter affiliation: University of Illinois at Urbana-Champaign, Urbana, Illinois. 63

Advancing systems biology using an open, extensible and scalable KBase platform
Vivek Kumar, Sunita Kumari, Srividya Ramakrishnan, James Thomason, Doreen Ware, Michael Schatz, Shinjae Yoo, Priya Ranjan, Samuel Seaver, Nomi Harris, Christopher Henry, Robert Cottingham, Adam Arkin.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 64

FRIDAY, October 28—9:00 AM

SESSION 6 ALGORITHMICS

Chairpersons: **Ben Langmead,** Johns Hopkins University, Baltimore, Maryland
Lior Pachter, University of California, Berkeley

A tandem simulation framework for predicting mapping quality
Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 65

CONVEX—Fast and accurate de novo transcriptome recovery from long reads
Meisam Razaviyayn, Elizabeth Tseng, David Tse.
Presenter affiliation: University of Southern California, Los Angeles, California. 66

VarMatch—Robust matching of small variant datasets using flexible scoring schemes
Chen Sun, Paul Medvedev.
Presenter affiliation: The Pennsylvania State University, University Park, Pennsylvania. 67

Metagenome annotation with distributed reference graphs
Andre Kahles, Gunnar Rätsch.
Presenter affiliation: ETH Zurich, Zurich, Switzerland; MSKCC, New York, New York. 68

Algorithms for metagenome analysis
Lior Pachter.
Presenter affiliation: University of California, Berkeley.

Fast approximate mapping of single molecule sequences using MinHash
Chirag Jain, Alexander Dilthey, Sergey Koren, Srinivas Aluru, Adam Phillippy.
Presenter affiliation: Georgia Institute of Technology, Atlanta, Georgia. 69

Accurate and fast detection of complex and nested structural variations using long read technologies
Fritz J. Sedlazeck, Philipp Rescheneder, Arndt v Heaseler, Michael C. Schatz.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 70

Quantification of sensitive information leakage from genomic linking attacks
Arif Harmanci, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 71

FRIDAY, October 28—1:30 PM

SESSION 7 HUMAN BIOLOGY

Chairpersons: **Alexis Battle,** Johns Hopkins University, Baltimore, Maryland
Ben Neale, Massachusetts General Hospital, Boston,

Understanding regulation of transcription and splicing through transcriptome-wide networks
Alexis Battle.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland

Predicting tissue-specific effects of rare genetic variants
Farhan N. Damani, Yungil Kim, Xin Li, Emily K. Tsang, Joe R. Davis, Colby Chiang, Zachary Zappala, Benjamin J. Strober, Alexandra J. Scott, Ira M. Hall, Stephen B. Montgomery, Alexis Battle.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 72

Computational proteogenomic identification and functional interpretation of translated fusions and micro structural variations in cancer
Yen-Yi Lin, Alexander R. Gawronski, Faraz Hach, Sujun Li, Ibrahim Numanagic, Iman Sarrafi, Swati Mishra, Andrew McPherson, Colin Collins, Milan Radovich, Haixu Tang, Cenk Sahinalp.
Presenter affiliation: Simon Fraser University, Burnaby, Canada. 73

A systematic study on the distribution and occurrence of mutations within regulatory motifs from cancer genome
Ambarnil Ghosh, Subhra Pradhan, Kyeong Kyu Kim.
Presenter affiliation: Sungkyunkwan University School of Medicine, Suwon, South Korea. 74

Opportunities for scaling up genomic analysis
Ben Neale.
Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts.

Cell type specific gene expression variation in differentiated induced pluripotent stem cells
YoSon Park, Evanthia E. Pashos, Mayda Hernandez, Stacey Lyttle, Dawn Marchadier, Juan Arbelaez, Wenjun Li, Zhaorui Lian, Jianting Shi, Katherine J. Slovik, Ruilan Yan, Kiran Musunuru, Wenli Yang, Edward E. Morrisey, Stephen Duncan, Daniel J. Rader, Christopher D. Brown.
Presenter affiliation: Perelman School of Medicine University of Pennsylvania (PSoM UPenn), Philadelphia, Pennsylvania. 75

Distant regulatory effects of genetic variation in multiple human tissues
Brian Jo, Yuan He, Benjamin J. Strober, Princy Parsana, Francois Aguet, Andrew A. Brown, Ariel Gewirtz, Eun Yong Kang, Ian C. McDowell, Ashis Saha, Kristin Ardlie, Don Conrad, Emmanouil T. Dermitzakis, Eric Gamazon, Eleazar Eskin, The GTEx Consortium, Barbara Engelhardt, Alexis Battle.
Presenter affiliation: John Hopkins University, Baltimore, Maryland. 76

Functionalizing human disease variants through systems modeling and network analysis
Song Yi, Limei Hu, Nidhi Sahni.
Presenter affiliation: The University of Texas MD Anderson Cancer Center, Houston, Texas. 77

FRIDAY, October 28—4:30 PM

KEYNOTE SPEAKER

Xiaole Shirley Liu
Harvard School of Public Health

“Inference of tumor immunity and T-cell receptor repertoire from TCGA RNA-seq data” 78

FRIDAY, October 28—5:30 PM

SESSION 8 POSTER SESSION II and COCKTAIL PARTY

Interactive genomic visualization tools for long read sequencing, assembly, and cancer
Maria Nattestad, Chen-Shan Chin, Michael C. Schatz.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 79

Enhancing pre-defined workflows with ad hoc analytics in a single environment—Unlocking Jupyter and RStudio for biologists
Bjorn Gruning, Eric Rasche, Torsten Houwaart, John Chilton, James Taylor, Anton Nekrutenko.
Presenter affiliation: Penn State University, University Park, Pennsylvania. 80

Gpeg—Lossy and lossless compression of sequencing read alignments across many samples
Abhinav Nellore, Jacob Pritt, Rachel Ward, Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 81

Recount—A large-scale resource of analysis-ready RNA-seq expression data
Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E. Ellis, Margaret A. Taub, Kasper D. Hansen, Andrew E. Jaffe, Ben Langmead, Jeffrey T. Leek.
Presenter affiliation: Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland; Johns Hopkins University, Baltimore, Maryland. 82

TACO—Multi-sample transcriptome assembly for RNA-seq
Yashar S. Niknafs, Balaji Pandian, Hariharan K. Iyer, Arul M. Chinnaiyan, Matthew K. Iyer.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 83

Integrative and predictive modeling of severe controlled and severe resistant hypertension among African-Americans in the MH-GRID Network
Cihan Oguz, Adam R. Davis, Gary H. Gibbons.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 84

Searching and exploring Gramene's comparative genomics datasets on the web
Andrew Olson, Joseph Mulvaney, James Thomason, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 85

Analysis of quantitative data of nuclear division dynamics from single gene knockdown embryos for all essential embryonic genes in *C. elegans*
Shuichi Onami.
Presenter affiliation: RIKEN Quantitative Biology Center, Kobe, Japan. 86

dbSNP in the era of next-generation sequencing and big data
Lon Phan, Hua Zhang, Juliana Feltz, Wang Qiang, Rama Maiti, David Shao, Ming Ward.
Presenter affiliation: National Center for Biotechnology Information, Bethesda, Maryland. 87

ShinyLearner—Enabling biologists to perform robust machine-learning classification
Stephen R. Piccolo, Terry J. Lee.
Presenter affiliation: Brigham Young University, Provo, Utah. 88

EuPathDB—Integrating eukaryotic pathogen genomics data with advanced search capabilities
Jane A. Pulman, Wei Li.
Presenter affiliation: University of Liverpool, Liverpool, United Kingdom. 89

RNA-seq expression analysis made easy in KBase
Srividya Ramakrishnan, Michael Schatz, Vivek Kumar, Sunita Kumari, Jim Thomason, Doreen Ware, Shinjae Yoo, Priya Ranjan, Samuel Seaver, Nomi Harris, Christopher Henry, Robert Cottingham, Adam Arkin.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 90

A state-of-the-art compression method for ChIP-seq data
Vida Ravanmehr, Zhiying Wang, Olgica Milenkovic.
Presenter affiliation: University of Illinois at Urbana-Champaign, Urbana, Illinois. 91

Computational determination of tumor type for cancers of unknown primary using mutation data
Jeffrey A. Rosenfeld, Greg Riedlinger, Anshuman Panda, Wenjin Chen, Gyan Bhanot.
Presenter affiliation: Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey. 92

Transcriptome-wide networks reveal candidate splicing regulatory relationships
Ashis Saha, Yungil Kim, Benjamin J. Strober, Barbara E. Engelhardt, The GTEx Consortium, Alexis Battle.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 93

SCENA—Secure compressed gENomic data analysis in a cloud environment
Feng Chen, Ibrahim Numanagic, Sean Simmons, Can Kockan, Bonnie Berger, Xiaoqian Jiang, Tianyi Zhu, Weijia Wu, Shuang Wang, Cenk Sahinalp.
Presenter affiliation: Indiana University, Bloomington, Indiana; Simon Fraser University, Burnaby, Canada. 94

***In silico* analysis of viral diversity obtained through viral metagenomics of chicken respiratory tract**
Manisha R. Sajnani, Tejas G. Oza, Ramesh J. Pandit, Prakash G. Koringa, Subhash J. Jakhesara, Prakash B. Banakar, Chaitanya G. Joshi, D Sudharsanam, Vaibhav D. Bhatt.
Presenter affiliation: Bharathiar University, Coimbatore, India. 95

Optimizing de novo assembly for *Meloidogyne indica* transcriptome—A root-knot nematode
Manisha Sajnani, Prakash Banakar, Uma Rao.
Presenter affiliation: Indian Agricultural Research Institute, New Delhi, India. 96

Joint probabilistic model for multiple steps of gene regulation
Hirak Sarkar, Yi-Fei Huang, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; Stony Brook University, Stony Brook, New York. 97

Explaining cancer microenvironments via automated literature exploration
Adam A. Butchy, Khaled Sayed, Kai-Wen Liang, Bryce Aidukaitis, Cheryl Telmer, Natasa Miskov-Zivanov.
Presenter affiliation: University of Pittsburgh, Pittsburgh, Pennsylvania. 98

Interactive visualization and analysis of large-scale sequencing datasets with the ZENBU genome browser system
Jessica Severin, Marina Lizio, Hideya Kawaji, Nicolas Bertin, Alistair Forrest, Yoshihide Hayashizaki.
Presenter affiliation: RIKEN , Yokohama, Japan. 99

C-terminome—An application to investigate C-terminal minimotifs in human proteins
Surbhi Sharma, Oniel Toledo, Kenneth F. Lyon, Steven Brooks, Roxainne P. David, Justin Limtong, Jackyln Newsome, Nemanja Novakovic, Sanguthevar Rajasekaran, Vishal Thapar, Sean Williams, Martin R. Schiller.
Presenter affiliation: University of Nevada, Las Vegas, Las Vegas, Nevada. 100

Deep learning regulatory sequence drivers of chromatin accessibility dynamics during cellular reprogramming
Anna Shcherbina, Avanti Shrikumar, Glenn Markov, Thach Mai, David M. Burns, Helen Blau, Anshul Kundaje.
Presenter affiliation: Stanford University, Stanford, California. 101

Multi-Omic Network Analysis (MONA), a new tool for integrative analysis
Paul A. Stewart, Adam L. Borne, Muhaimen Shamsi, Steven A. Eschrich, Ann Chen, Eric B. Haura.
Presenter affiliation: Moffitt Cancer Center, Tampa, Florida. 102

GenoGAM—Genome-wide generalized additive models for ChIP-seq analysis
Georg Stricker, Alexander Engelhardt, Daniel Schulz, Matthias Schmid, Achim Tresch, Julien Gagneur.
Presenter affiliation: Gene Center, Ludwig-Maximilians-University Munich, Munich, Germany; Technical University Munich, Munich, Germany. 103

Compendium-wide analysis of public data with eADAGE reveals novel regulatory mechanisms
Jie Tan, Georgia Doing, Kimberley A. Lewis, Courtney E. Price, Deborah A. Hogan, Casey S. Greene.
Presenter affiliation: Geisel School of Medicine at Dartmouth, Hanover, New Hampshire. 104

Novel roles for RNA revealed from *in vivo* RNA structuromes
Yin Tang, Anton Nekrutenko, Philip C. Bevilacqua, Sarah M. Assmann.
Presenter affiliation: Pennsylvania State University, University Park, Pennsylvania. 105

Making research software more robust
Morgan Taschuk, Greg Wilson.
Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada. 106

Plant Reactome—A resource for comparative plant pathway analysis
Sushma Naithani, Justin Preece, Peter D'Eustachio, Justin L. Elser, Parul Gupta, Antonio F. Mundo, Joel Weiser, Marcela K. Tello-Ruiz, Lincoln Stein, Doreen Ware, Pankaj Jaiswal.
Presenter affiliation: Cold Spring Harbor Laboratory, Plant Bioinformatics, New York. 107

Infrastructure and development of the exRNA virtual biorepository
William Thistlethwaite, Sai Lakshmi Subramanian, Bob Carter, Brittney Miller, Javier Figueroa, Fred Hochberg, Ryan Kim, Johnny Akers, Douglas Galasko, Matt Huentelman, Kendall Jensen, Rebecca Reiman, Jorge Arango, Yashar Kalani, Julie Saugstad, Theresa Lusardi, Joseph Quinn, Aleksandar Milosavljevic. Andrew Jackson, Neethu Shah, Aaron Baker, Sameer Paithankar, Matthew Roth, Betty Lind.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 108

Training a somatic mutations caller with deep learning and semi-simulated data
Remi C. Torracinta, Fabien Campagne.
Presenter affiliation: Weill Cornell Medicine, New York, New York. 109

Volt—Cloud base pure java NGS Mapping software and analysis pipeline
Hiroki Ueda, Hiroyuki Aburatani.
Presenter affiliation: Fujitsu Limited, Tokyo, Japan. 110

Super-scaffolding of large eukaryotic genomes with single molecule maps
Davide Verzotto, Audrey S. Teo, Luka Sterbic, Burton K. Chia, Mile Sikic, Axel M. Hillmer, Niranjan Nagarajan.
Presenter affiliation: Genome Institute of Singapore, Singapore. 111

kmer2vec—Unsupervised embedding of regulatory DNA sequences
Rahul Mohan, Michael Wainberg, Nasa Sinnott-Armstrong, Anshul Kundaje.
Presenter affiliation: Stanford University, Stanford, California. 112

SciApps—A distributed CyVerse system for cloud computing
Liya Wang, Zhenyuan Lu, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 113

Snaptron—Enabling flexible querying and exploratory analysis of human RNA splicing
Christopher Wilks, Phani Gaddipati, Abhinav Nellore, Ben Langmead.
Presenter affiliation: Department of Computer Science, Baltimore, Maryland; Center for Computational Biology, Baltimore, Maryland. 114

Required parameters—What does it take to bring bioinformatics into the classroom at the national level?
Jason Williams, Mona Spector, Cornel Ghiban, Joslynn Lee, David Micklos.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 115

Characterization of regulatory roles of CtBP in breast cancer using integrated genomic and proteomic analysis
Tingfen Yan, Jung S. Byun, Dae Ik Yi, Samson Park, Genqing Liang, Mohamed Kabbout, Kevin Gardner.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 116

Identification of robust metabolomics causal networks in observational study using data integration
Azam Yazdani.
Presenter affiliation: UTHealth , Houston, Texas. 117

Tumor-suppressive and -promoting functions of Sufu in the SHH subgroup of medulloblastoma
Wen-Chi Yin, Thevagi Satkunendran, Rong Mo, Huayun Hou, Hayden Selvadurai, Sorana Morrissy, Michael Wilson, Peter B. Dirks, Michael D. Taylor, Chi-chung Hui.
Presenter affiliation: University of Toronto, Toronto, Canada; Hospital for Sick Children, Toronto, Canada. 118

The critical assessment of protein function annotation—Improving on the “state-of-the-art”
Naihui Zhou, Yuxiang Jiang, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, Iddo Friedberg.
Presenter affiliation: Iowa State University, Ames, Iowa. 119

Using a top-down approach to understand the human transcriptome
Weizhuang Zhou, Russ B. Altman.
Presenter affiliation: Stanford University, Stanford, California. 120

MSKCC Paperless Lab Initiative (MSK-PLI)—Software for analysis of fragment and qualitative PCR assays
John Ziegler, Jack Birnbaum, Paulo Salazar, Ivelise Rijo, Nana Mensah, Georgi Lukose, Lissette Fernandez, Kelly Rios, Ahmet Zehir, Maria Arcila.
Presenter affiliation: MKSCC, New York, New York. 121

FRIDAY, October 28—7:00 PM

BANQUET

SATURDAY, October 29—9:00 AM

SESSION 9 DATA INFRASTRUCTURE

Chairpersons: **Benedict Paten,** University of California, Santa Cruz
Valerie Schneider, National Center for Biotechnology Information, Bethesda, Maryland

Reproducible, portable, scalable and open genomics with Toil, ADAM and Dockstore
Benedict Paten, John Vivian, Arjun Rao, Frank Nothaft, Jacob Pfeil, Audrey Musselman-Brown, Hannes Schmidt, Christopher Ketchum, Anthony Joseph, David Haussler, Brian O'Connor.
Presenter affiliation: UC Santa Cruz Genomics Institute, Santa Cruz, California. 122

IHEC Data Portal—A resource for discovering, analysing and sharing epigenomics data
David Bujold, David A. Morais, Carol Gauthier, Catherine Côté, Maxime Caron, Tony Kwan, Kuang C. Chen, Jonathan Laperle, Alexei N. Markovits, Tomi Pastinen, Bryan Caron, Alain Veilleux, Pierre-E´tienne Jacques, Guillaume Bourque.
Presenter affiliation: McGill University, Montréal, Canada; McGill University, Montréal, Canada 123

Choice of reference genome can introduce massive bias in bisulfite sequencing data
Kasper D. Hansen.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 124

The exRNA Atlas—Linking data, tools, and computable pathway knowledge to interpret the first 1000 uniformly processed profiles of extracellular RNA from human bodily fluids
Sai Lakshmi Subramanian, William Thistlethwaite, Andrew R. Jackson, Neethu Shah, Aaron Baker, Sameer Paithankar, Robert Kitchen, Kristina Hanspers, Ginger Tsueng, Roger P. Alexander, David J. Galas, Joel Rozowsky, Alexander Pico, Andrew I. Su, Matthew E. Roth, Mark B. Gerstein, Aleksandar Milosavljevic, Andrew Jackson, Neethu Shah, Aaron Baker, Sameer Paithankar, Matthew Roth, Betty Lind.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 125

The evolving human reference genome assembly drives changes in data management and representation
V A. Schneider, P A. Kitts, A Kimchi, K D. Pruitt, K Clark, T D. Murphy.
Presenter affiliation: National Center for Biotechnology Information, Bethesda, Maryland. 126

Evaluating population structure for subjects in the dbGaP Database
Yumi Jin, Stephanie Pretel, Anne Sturcke, Moira Lee, Natilia Popova, Wenyu Wu, Stephen Sherry, Michael Feolo.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 127

Functional annotation and visualization of large-scale biological networks
Anastasia Baryshnikova.
Presenter affiliation: Princeton University, Princeton, New Jersey. 128

StruMs—A flexible and information-rich representation of DNA motifs
Peter M. DeFord, James Taylor.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 129

Closing Remarks

AUTHOR INDEX

Abugessaisa, Imad, 41
Aburatani, Hiroyuki, 110
Adams, David J., 13
Aghazadeh, Amirali, 16
Aguet, Francois, 76
Aidukaitis, Bryce, 98
Akers, Johnny, 108
Alexander, Roger P., 125
Allen, Andrew, 40
Altman, Russ B., 120
Aluru, Srinivas, 69
An, Lin, 17
Andrade-Navarro, Miguel, 52
Arango, Jorge, 108
Arbelaez, Juan, 75
Arcila, Maria, 121
Ardlie, Kristin, 76
Arkin, Adam, 64, 90
Assmann, Sarah M., 105
Atkins, Lisa M., 16
Aviran, Sharon, 23
Avsec, Žiga, 18
Awan, Mohammed O., 19
Ayal, Gussow, 40

Bailey, Timothy L., 13
Baker, Aaron, 108, 125
Banakar, Prakash, 95, 96
Baraniuk, Richard G., 16
Barrette, Michel, 61
Baryshnikova, Anastasia, 128
Bass, Hank W., 46
Bastin, Charlotte, 61
Battle, Alexis, 72, 76, 93
Beilstein, Mark A., 28
Berger, Bonnie, 94
Bermeo, Jonathan, 20
Bertin, Nicolas, 99
Bevilacqua, Philip C., 105
Bhanot, Gyan, 92
Bhatt, Vaibhav D., 95
Biffis, Nicholas, 47
Birnbaum, Jack, 121
Bittner, Marie, 52
Blau, Helen, 101
Borne, Adam L., 102
Bourque, Guillaume, 61, 123
Bovee, Roderick, 34
Bradic, Martina, 20
Brett, Copeland, 40
Brooks, Steven, 100
Brown, Andrew A., 76
Brown, Christopher, 10, 75
Brown, Stuart M., 21
Bryan, Jordan G., 14
Bujold, David, 61, 123
Burns, David M., 101
Butchy, Adam A., 98
Byun, Jung S., 116

Cairns, Junmei, 63
Campagne, Fabien, 6, 109
Carlton, Jane, 20
Carninci, Piero, 41
Caron, Bryan, 123
Caron, Maxime, 123
Carter, Bob, 108
Cawthon, Bryan E., 19
Chang, Tzu-Hao, 49
Chen, Allen L., 16
Chen, Ann, 102
Chen, Feng, 94
Chen, Hao, 21
Chen, Huan, 23
Chen, Kuang C., 123
Chen, Wenjin, 92
Chia, Burton K., 111
Chiang, Colby, 72
Chilton, John, 22, 80
Chin, Chen-Shan, 79
Chinnaiyan, Arul M., 83
Chou, Chia-Cheng, 49
Choudhary, Krishna, 23
Chougule, Kapeel M., 24
Clark, K, 126
Clarke, Declan, 25
Coffman, Adam, 43
Collado-Torres, Leonardo, 82
Collins, Colin, 73
Colwell, Rita R., 55

Conlan, Sean, 37
Conrad, Don, 76
Corry, Schuyler, 26
Côté, Catherine, 123
Cottingham, Robert, 64, 90
Cox, Nancy J., 58
Cukuroglu, Engin, 27

D'Eustachio, Peter, 60
Damani, Farhan N., 72
Darnell, Gregory, 10
David, Roxainne P., 100
Davis, Adam R., 84
Davis, Joe R., 72
DeFord, Peter M., 129
Deng, Fei, 23
Dermitzakis, Emmanouil T., 76
Derr, Tyler, 17
Deshwar, Amit, 12
D'Eustachio, Peter, 107
Devisetty, Upendra Kumar, 28
Diesch, Colin, 7
Digan, William E., 6
Dikow, Rebecca B., 29, 30
Dilthey, Alexander, 69
Dirks, Peter B., 118
Doerfel, Max, 8
Doing, Georgia, 104
Drezek, Rebekah A., 16
Duncan, Stephen, 75
Dunn, Nathan A., 7

Eastman, Thomas, 59
El Amrani, Khadija, 52
Elizabeth, Cirulli T., 40
Ellis, Shannon E., 82
Elnitski, Laura, 33
Elser, Justin L., 107
Elsik, Chris, 7
Emad, Amin, 63
Engelhardt, Alexander, 103
Engelhardt, Barbara, 10, 76, 93
Eschrich, Steven A., 102
Eskin, Eleazar, 76

Fabregat, Antonio, 60
Fang, Han, 8
Feltz, Juliana, 87
Feolo, Michael, 127
Ferguson-Smith, Anne C., 13
Fernandez, Lissette, 121
Fetrow, Jacquelyn S., 47
Figueroa, Javier, 108
Fletez-Brant, Kipper, 36
Flibotte, Stephane, 15
Foo, Chuan-Sheng, 2
Forrest, Alistair, 41, 99
Frandsen, Paul B., 29, 30
Frauenkron-Machedjou, Josiane, 32
Friedberg, Iddo, 31, 119
Fulton, Alexander, 32

Gaddipati, Phani, 114
Gaddipati, Ravi, 3
Gagneur, Julien, 18, 103
Galas, David J., 125
Galasko, Douglas, 108
Gamazon, Eric, 76
Garber, Manuel, 44
Gardner, Kevin, 116
Gauthier, Carol, 123
Gawronski, Alexander R., 73
Gerhardinger, Chiara, 35
Gerstein, Mark, 25, 71, 125
Gewirtz, Ariel, 76
Ghiban, Cornel, 115
Ghosh, Ambarnil, 74
Giangreco, Nicholas, 33
Gibbons, Gary H., 84
Goke, Jonathan, 27
Goldstein, David B., 40
Gotea, Valer, 33
Graubert, Aaron, 43
Green, Rachel, 8
Greene, Casey S., 104, 119
Greenfield, Nicholas, 34
Greenside, Peyton, 11
Griffith, Malachi, 43
Griffith, Obi L., 43
Groff, Abigail F., 35
Gruning, Bjorn, 80
Gu, Claire, 59
Guerler, Aysam, 22

Gupta, Parul, 107

Hach, Faraz, 73
Hall, Ira M., 72
Hammell, Molly, 56
Hansen, Kasper D., 36, 82, 124
Hansen, Nancy F., 37
Hanspers, Kristina, 125
Hao, Yuhan, 21
Harel, Marie, 61
Harmanci, Arif, 71
Harris, Nomi, 64, 90
Harshbarger, Jayson, 41
Hasan, Nur A., 55
Haug-Baltzell, Asher K., 28
Haura, Eric B., 102
Haussler, David, 4, 122
Hayashizaki, Yoshihide, 41, 99
Hayden, Julia, 47
He, Yuan, 76
Heil, Benjamin, 37
Heinzen-Cox, Erin, 40
Henry, Christopher, 64, 90
Hermjakob, Henning, 60
Hernandez, Mayda, 75
Herrin, Steven, 59
Hillmer, Axel M., 111
Ho, Yu-Jui, 56
Hochberg, Fred, 108
Hoffman, Michael M., 13
Hogan, Deborah A., 104
Holmes, Ian, 7
Hou, Huayun, 118
Houwaart, Torsten, 80
Hu, Limei, 77
Huang, Yifei, 8, 9, 97
Huentelman, Matt, 108
Hui, Chi-chung, 118
Hundal, Jasreet, 43
Hutton, Elizabeth, 38

Irizarry, Rafael, 1
Ishii, Kazuo, 39
Israeli, Johnny, 2, 11
Iyer, Hariharan K., 83
Iyer, Matthew K., 83

Jackson, Andrew R., 108, 125
Jacques, Pierre-Étienne, 61, 123
Jaeger, Karl-Erich, 32
Jaffe, Andrew E., 82
Jain, Chirag, 69
Jaiswal, Pankaj, 107
Jakhesara, Subhash J., 95
Jensen, Kendall, 108
Jiang, Xiaoqian, 94
Jiang, Yuxiang, 119
Jin, Ying, 56
Jin, Yumi, 127
Jo, Brian, 76
Johnson, Coreen L., 16
Johnson, James, 13
Joseph, Anthony, 4, 122
Joshi, Chaitanya G., 95
Joshua, Bridgers, 40

Kabbout, Mohamed, 116
Kahles, Andre, 68
Kalani, Yashar, 108
Kamalakaran, Sitharthan, 40
Kammers, Kai, 82
Kang, Eun Yong, 76
Kaplow, Irene, 2
Kasukawa, Takeya, 41
Kathleen, Cho R., 33
Kawaji, Hideya, 41, 99
Ketchum, Christopher, 122
Kim, Kyeong Kyu, 74
Kim, Minji, 42
Kim, Ryan, 108
Kim, Yungil, 72, 93
Kimchi, A, 126
Kitchen, Robert, 125
Kitts, P A., 126
Kiwala, Susanna, 43
Kockan, Can, 94
Kondo, Atsushi, 41
Koren, Sergey, 69
Koringa, Prakash G., 95
Krampis, Konstantinos, 21
Kucukural, Alper, 44
Kugler, Hillel, 45
Kumar, Sushant, 25
Kumar, Vivek, 64, 90
Kumari, Sunita, 64, 90
Kundaje, Anshul, 2, 11, 101, 112

Kurtz, Andreas, 52
Kwan, Tony, 123
Kyle, Kathleen E., 46

Lakshmi Subramanian, Sai, 108
Langmead, Ben, 3, 65, 81, 82, 114
Laperle, Jonathan, 61, 123
Laserson, Uri, 4
Lawrence, Charles E., 54
Lee, James C., 35
Lee, Joslynn, 115
Lee, Moira, 127
Lee, Terry J., 88
Leek, Jeffrey T., 82
Lekschas, Fritz, 52
Leroux, Michel R., 15
Leuthaeuser, Janelle B., 47
Lewis, Kimberley A., 104
Lewis, Suzanna, 7
Li, Huai, 55
Li, Sujun, 73
Li, Wei, 89
Li, Wenjun, 75
Li, Xin, 72
Li, Yulong, 48
Lian, Zhaorui, 75
Liang, Genqing, 116
Liang, Kai-Wen, 98
Liang, Suh-Yuen, 49
Limtong, Justin, 100
Lin, Adam Y., 16
Lin, Yen-Yi, 73
Lind, Betty, 108
Liu, Bingjian, 50
Liu, Jinxian, 48
Liu, Xiaole Shirley, 78
Lizio, Marina, 41, 99
Loucks, Catrina, 15
Lu, Zhenyuan, 113
Lukose, Georgi, 121
Luo, Ruibang, 51
Lusardi, Theresa, 108
Lyon, Gholson, 8
Lyon, Kenneth F., 100
Lyons, Eric, 28
Lyttle, Stacey, 75

Mah, Nancy, 52
Mai, Thach, 101
Maiti, Rama, 87
Malik, Laraib I., 53
Marchadier, Dawn, 75
Mardis, Elaine R., 43
Margolin, Gennady, 33
Marin-Garcia, Pablo, 60
Markov, Glenn, 101
Markovits, Alexei N., 61, 123
Mason, Christopher E., 55
McDowell, Ian C., 76
McKerrow, Wilson H., 54
McMichael, Joshua, 43
McPherson, Andrew, 73
Medvedev, Paul, 67
Meng, Tzu-Ching, 49
Mensah, Nana, 121
Merowsky, Nicholas, 44
Meyers, Robin M., 14
Micklos, David, 115
Milenkovic, Olgica, 42, 91
Miller, Brittney, 108
Miller, Christopher A., 43
Milosavljevic, Aleksandar, 108, 125
Minot, Samuel, 34
Minsk, Miriam K., 37
Mishra, Swati, 73
Miskov-Zivanov, Natasa, 98
Mo, Rong, 118
Moerman, Don G., 15
Moffat, Kelly, 55
Mohan, Rahul, 2, 112
Molik, David, 56
Montgomery, Philip G., 57
Montgomery, Stephen B., 72
Mooney, Sean D., 119
Morais, David A., 61, 123
Morgan, Martin, 5
Morris, Quaid, 12
Morrisey, Edward E., 75
Morrissy, Sorana, 118
Mukherjee, Sayan, 10
Mullikin, James C., 37
Mulvaney, Joseph, 85
Mundo, Antonio F., 107

Muñoz-Torres, Monica, 7
Murphy, T D., 126
Musselman-Brown, Audrey, 122
Musunuru, Kiran, 75

Nagarajan, Niranjan, 111
Naithani, Sushma, 107
Nakamura, Kazutaka, 39
Nattestad, Maria, 79
Nekrutenko, Anton, 80, 105
Nellore, Abhinav, 81, 82, 114
Nelms, Brian L., 19
Nelson, Andrew D., 28
Newsome, Jackyln, 100
Niknafs, Yashar S., 83
Nothaft, Frank, 4, 122
Novakovic, Nemanja, 100
Numanagic, Ibrahim, 73, 94

O'Connor, Brian, 122
Oguz, Cihan, 84
Olson, Andrew, 85
Onami, Shuichi, 86
Oza, Tejas G., 95

Pagenkopp Lohan, Katrina M., 29
Paithankar, Sameer, 108, 125
Panda, Anshuman, 92
Pandian, Balaji, 83
Pandit, Ramesh J., 95
Pann, Alexander, 6
Paradarami, Tulasi, 59
Park, Samson, 116
Park, YoSon, 75
Parsana, Princy, 76
Pashos, Evanthia E., 75
Pastinen, Tomi, 123
Paten, Benedict, 4, 122
Patro, Rob, 53
Patterson, David A., 4
Petrosino, Joseph F., 16
Petrykowska, Hanna M., 33
Pfeil, Jacob, 122
Phan, Lon, 87
Phillippy, Adam, 69
Piccolo, Stephen R., 62, 88
Pico, Alexander, 125
Popova, Natilia, 127
Pradhan, Subhra, 74
Preece, Justin, 107
Pretel, Stephanie, 127
Price, Courtney E., 104
Pritt, Jacob, 3, 81
Pruitt, K D., 126
Pu, Tsung-Hsien, 49
Pulman, Jane A., 89

Qiang, Wang, 87
Quinn, Joseph, 108

Rader, Daniel J., 75
Radhakrishnan, Aditya, 8
Radivojac, Predrag, 119
Radovich, Milan, 73
Rajasekaran, Sanguthevar, 100
Ramakrishnan, Srividya, 64, 90
Ranjan, Priya, 64, 90
Rao, Arjun, 122
Rao, Uma, 96
Rasche, Eric, 80
Rätsch, Gunnar, 68
Ravanmehr, Vida, 91
Razaviyayn, Meisam, 66
Reiman, Rebecca, 108
Ren, Zhong, 40
Rescheneder, Philipp, 70
Riedlinger, Greg, 92
Rijo, Ivelise, 121
Rinn, John L., 35
Rios, Kelly, 121
Robert, Marc-Antoine, 61
Rosenfeld, Jeffrey A., 92
Roth, Matthew, 108, 125
Rozowsky, Joel, 125
Ruan, Luyao, 23
Rubanova, Yulia, 12

Saha, Ashis, 76, 93
Sahin, Nil, 12
Sahin, Serkan, 41
Sahinalp, Cenk, 73, 94
Sahni, Nidhi, 77
Sajnani, Manisha, 95, 96
Salazar, Paulo, 121
Salzberg, Steven L., 51

Sanchez-Gomez, Diana B., 35
Sarkar, Hirak, 97
Sarrafi, Iman, 73
Satkunendran, Thevagi, 118
Saugstad, Julie, 108
Sayed, Khaled, 98
Schatz, Michael C., 8, 51, 64, 70, 79, 90
Scherbina, Anna, 2
Schiller, Martin R., 100
Schmid, Matthias, 103
Schmidt, Hannes, 122
Schneider, V A., 126
Schulz, Daniel, 103
Schwaneberg, Uli, 32
Scott, Alexandra, 33, 72
Seaver, Samuel, 64, 90
Sedlazeck, Fritz J., 70
Segre, Julia A., 37
Seltmann, Stefanie, 52
Selvadurai, Hayden, 118
Seredick, Barb, 26
Severin, Jessica, 99
Shah, Neethu, 108, 125
Shamsi, Muhaimen, 102
Shao, David, 87
Sharma, Surbhi, 100
Shcherbina, Anna, 11, 101
Sheikh, Mona A., 16
Sherry, Stephen, 127
Shi, Hui, 13
Shi, Jianting, 75
Shi, Junwei, 38
Shih, Nathan P., 23
Shimoji, Hisashi, 41
Shrikumar, Avanti, 2, 11, 101
Sidiropoulos, Konstantinos, 60
Siepel, Adam, 8, 9, 38, 97
Sikic, Mile, 111
Simi, Manuele, 6
Simmons, Sean, 94
Singh, Kiran, 59
Sinha, Saurabh, 63
Sinnott-Armstrong, Nasa, 2, 11, 112
Sjöberg, Marcela, 13
Skoczinski, Pia, 32
Slave, Petrovski, 40
Slovik, Katherine J., 75
Smith, Rachel K., 26
Soruco, Marcela M., 35
Spector, Mona, 115
Stachelscheid, Harald, 52
Stein, Joshua, 24
Stein, Lincoln, 60, 107
Sterbic, Luka, 111
Stewart, Paul A., 102
Stiles, Andrew, 59
Stricker, Georg, 103
Strober, Benjamin J., 72, 76, 93
Sturcke, Anne, 127
Su, Andrew I., 125
Subramanian, Poorani, 55
Subramanian, Sai Lakshmi, 125
Sudharsanam, D, 95
Sun, Chen, 67

Tan, Jie, 104
Tang, Haixu, 73
Tang, Yin, 105
Taschuk, Morgan, 106
Tatlow, PJ, 62
Taub, Margaret A., 82
Taylor, James, 80, 129
Taylor, Michael D., 118
Tello-Ruiz, Marcela K., 107
Telmer, Cheryl, 98
Teo, Audrey S., 111
Thapar, Vishal, 100
Thistlethwaite, William, 108, 125
Thomas, Neil, 59
Thomas, Pamela J., 37
Thomason, James, 64, 85, 90
Timbers, Tiffany A., 15
Toledo, Oniel, 100
Torracinta, Remi C., 109
Tounaka, Nobuaki, 39
Tresch, Achim, 103
Tsang, Emily K., 72
Tse, David, 66
Tseng, Elizabeth, 66
Tsherniak, Aviad, 14
Tsueng, Ginger, 125
Tung, Jenny, 10

Ueda, Hiroki, 110

Unni, Deepak, 7

Vakoc, Chris, 38
Veilleux, Alain, 61, 123
Vera, Daniel L., 46
Verzotto, Davide, 111
Viner, Coby, 13
Viteri, Guilherme, 60
Vivian, John, 122
von Heaseler, Arndt, 70

Wainberg, Michael, 2, 112
Walker, Jason, 43
Walker, Nicolas, 13
Wang, Liewei, 63
Wang, Liya, 24, 113
Wang, Quanli, 40
Wang, Shuang, 94
Wang, Yanli, 17
Wang, Zhiying, 91
Ward, Ming, 87
Ward, Rachel, 81
Ware, Doreen, 24, 64, 85, 90, 107, 113
Warring, Sally, 20
Weiser, Joel, 107
Wilhelm, Susanne, 32
Wilks, Christopher, 114
Williams, Jason, 115
Williams, Sean, 100
Wilson, Greg, 106
Wilson, Michael, 118
Wintersinger, Jeff, 12
Wu, Weijia, 94
Wu, Wenyu, 127

Xue, Dongxiu, 48

Yan, Ruilan, 75
Yan, Tingfen, 116
Yang, Wenli, 75
Yao, Eric, 7
Yazdani, Azam, 117
Yee, Patrick, 59
Yi, Dae Ik, 116
Yi, Song, 77
Yin, Wen-Chi, 118
Yoo, Shinjae, 64, 90
Yue, Feng, 17

Zappala, Zachary, 72
Zehir, Ahmet, 121
Zhang, Hua, 87
Zhou, Naihui, 119
Zhou, Weizhuang, 120
Zhu, Tianyi, 94
Ziegler, John, 121

OVERCOMING BIAS AND SYSTEMATIC ERROR IN HIGH-THROUGHPUT TECHNOLOGIES ARE KEY FOR THE SUCCESS OF DATA SCIENCE

Rafael Irizarry

Dana Farber Cancer Institute, BCB, Boston, MA

The unprecedented advance in digital technology during the second half of the 20th century has produced a measurement revolution that is transforming science. In the life sciences, data analysis is now part of practically every research project. Genomics, in particular, is being driven by new measurement technologies that permit us to observe certain molecular entities for the first time. These observations are leading to discoveries analogous to identifying microorganisms and other breakthroughs permitted by the invention of the microscope. An examples of this are the many application of next generation sequencing.

Biases, systematic errors and unexpected variability are common in biological data. Failure to discover these problems often leads to flawed analyses and false discoveries. As datasets become larger, the potential of these biases to appear to be significant actually increases. In this talk I will describe several examples of these challenges using very specific examples from gene expression microarrays, RNA-seq, and single-cell assays. I will describe data science solution to these problems.

← Back

HOW TO TRAIN YOUR DRAGONN (DEEP REGULATORY GENOMIC NEURAL NETWORK)

Johnny Israeli[1], Michael Wainberg[2], Avanti Shrikumar[2], Nasa Sinnott-Armstrong[3], Anna Scherbina[4], Irene Kaplow[2], Rahul Mohan[5], Chuan-Sheng Foo[2], Anshul Kundaje[2,3]

[1]Stanford University, Biophysics, Stanford, CA, [2]Stanford University, Computer Science, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA, [4]Stanford University, Biomedical Informatics, Stanford, CA, [5]Bellarmine College Preparatory School, Math and Science, San Jose, CA

Deep learning models have been recently applied to several key problems in regulatory genomics including prediction of protein-DNA interactions, context-specific chromatin state and non-coding regulatory variants. However, model design, parameter selection and training Deep RegulAtory GenOmics Neural Networks (DragoNNs) remains something of a black art. When is a DragoNN good choice for a learning problem in genomics? How does one design a high-performance model? And more importantly, can we interpret these models to discover novel patterns in the input data and the induced features to gain new biological insights? To demystify these questions, we developed the dragonn toolkit (http://kundajelab.github.io/dragonn/) - an interactive, cloud-based framework to allow users with inter-disciplinary backgrounds to learn and experiment with strategies for training and interpretting DragoNNs that model regulatory DNA sequence data. The dragonn toolkit provides a customizable simulation engine for regulatory DNA sequences; instructive built-in simulations capturing key properties of regulatory DNA; interactive IPython notebook tutorials for novice users; a command-line interface for simple applications of DragoNNs to custom user-defined sequence data; an interpretation toolkit for model exploration, pattern discovery and visualization and cloud resources for easy access to software and hardware. We will showcase the dragonn toolkit on simulated and real regulatory genomic data, demystify popular DragoNN architectures and provide guidelines for modeling and interpreting regulatory sequence using DragoNN models. We have used dragonn in several introductory workshops and tutorials on deep learning for genomics. We plan to continue its development into a community resource with guidelines for best practices and support for a model zoo allowing rapid access to and development of high performance, interpretable deep learning models for genomics.

LET THE RIGHT ONES IN: THE COST AND BENEFIT OF INCLUDING ALTERNATE ALLELES IN THE REFERENCE GENOME

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

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

Accurate analysis of sequencing data relies on precise mapping of reads to a reference genome. When the sequenced genome differs from the reference genome, e.g. at polymorphic loci, reads overlapping these regions are less likely to be mapped to the correct location. This has downstream effects, especially for applications such as allele-specific expression and isoform quantitation. The graph genome has recently grown in popularity as a means of incorporating much of the population variation into a single reference genome. A graph genome contains the variants of a large number of representative individuals, reducing the chance that a read overlapping some variants should fail to align (false negative rate). To date, most graph genome construction strategies focus on incorporating all known alleles into the graph to minimize this false negative rate. However, a more complex graph brings an exponential blowup in size and mapping time. A larger graph also increases the chance that a read from elsewhere in the genome will incorrectly align to a region with newly-added variation (false positive rate). It is likely that there is some optimal strategy to determine which SNPs to include in the graph genome, but little research has been done to date on this question.

Designing the graph genome requires a clear understanding of the tradeoffs between false negative rate, false positive rate, and blowup. We define these measures in objective terms and present a heuristic-free method to compute them for a particular graph genome. We explore the effects of various SNP stratification methods and inclusion strategies on these measures. Finally, we show how these results both (a) shed light on the utility of different graph genome proposals, and (b) allow users to mix the best properties of both non-graph and graph aligners.

PROCESSING TERABYTE-SCALE GENOMICS DATASETS WITH ADAM AND TOIL

Frank A Nothaft[1], Uri Laserson[3], David Haussler[2], Benedict Paten[2], David A Patterson[1], Anthony D Joseph[1]

[1]University of California, Berkeley, Electrical Engineering and Computer Sciences, Berkeley, CA, [2]University of California, Santa Cruz, Biomedical Engineering, Santa Cruz, CA, [3]Mount Sinai, Icahn School of Medicine, New York, NY

The detection and analysis of rare genomic events requires integrative analysis across large cohorts with terabytes to petabytes of genomic data. Contemporary genomic analysis tools have not been designed for this scale of data-intensive computing. This talk presents ADAM, an Apache 2 licensed library built on top of the popular Apache Spark distributed computing framework. ADAM is designed to allow genomic analyses to be seamlessly distributed across large clusters, and presents a clean API for writing parallel genomic analysis algorithms. In this talk, we'll look at how we've used ADAM to achieve a 3.5× improvement in end-to-end variant calling latency and a 66% cost improvement over current toolkits, without sacrificing accuracy. We will talk about a recent recompute effort where we have used ADAM to recall the Simons Genome Diversity Dataset against GRCh38. We will also talk about using ADAM alongside Apache Hbase to interactively explore large variant datasets.

← Back

TRADITIONAL APPROACHES AND CONTEMPORARY CHALLENGES TO PRODUCTION OF USEFUL SCIENTIFIC SOFTWARE: LESSONS FROM *BIOCONDUCTOR*

Martin Morgan

Roswell Park Cancer Institute, Biostatistics and Bioinformatics, Buffalo, NY

Bioconductor is a long-standing, successful and well-respected software project for the statistical analysis and comprehension of high-throughput genomic data. Several themes and practices established at the start of the project contribute to our success. Examples include and emphasis on traditional and 'vignette' documentation, nightly cross-platform builds, fully versioned software, a release structure that allows innovation without overly disrupting user work flows, and support for our developer and user communities. Rapid changes in contemporary computational environments pose challenges to our approach, and to development of reproducible software more generally. Many changes have an ironic component, e.g., cloud-based computation reduces challenges of providing software that runs on user computers but substantially increases the ability to provide reproducible analytic environments, and emergence of public version control repositories and continuous integration systems provide access to much better development practices without providing users with confidence in stable long-term software availability. Identifying and arriving at creative and robust solutions to the modern computational environment represent a significant challenge, for *Bioconductor* and for other projects hoping to make lasting contributions to biological data science.

← Back

MetaR: TOWARDS AN R NOTEBOOK WITH COMPOSABLE LANGUAGES

Fabien Campagne[1,2,3], Alexander Pann[1], William E Digan[1], Manuele Simi[1]

[1]Weill Cornell Medicine, The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, New York, NY, [2]Weill Cornell Medicine, Clinical Translational Science Center, New York, NY, [3]Weill Cornell Medicine, Department of Physiology and Biophysics, New York, NY

Data analysis tools have become essential to the study of biology. Here, we applied language workbench technology (LWT) to create data analysis languages tailored for biologists with a diverse range of experience: from beginners with no programming experience to expert bioinformaticians and statisticians.

A key novelty of our approach is its ability to blend user interface with scripting in a single platform with better language integration than possible with electronic notebooks. This new approach has several advantages over the state of the art: experts can design simplified data analysis languages that require no programming experience, and behave like graphical user interfaces, yet have the advantages of scripting. We report on such a simple language, called MetaR [1], which we have used to teach complete beginners how to call differentially expressed genes and build heatmaps. We found that beginners can complete this task in less than 2 hours with MetaR, when more traditional teaching with R and its packages would require several training sessions (6-24hrs). MetaR also seamlessly integrates with docker to enable reproducibility of analyses and simplified R package installations during training sessions.

We used the same approach to develop the first composable R language [1]. A composable language is a language that can be extended with micro-languages. We illustrate this capability with a Biomart micro-language designed to compose with R and help R programmers query Biomart interactively to assemble specific queries to retrieve data, (The same micro-language also composes with MetaR to help beginners query Biomart.)

Our teaching experience suggests that language design with LWT can be a compelling approach for developing intelligent data analysis tools and can accelerate training for common data analysis task. LWT offers an interactive environment with instant change detection and refreshing of plots and tables. Instant refresh was added to MetaR 2.0 and makes using popular R packages (e.g., tidyr, dplyr, ggplots2) much more interactive. The talk will provide an introduction to LWT and MetaR and describe our latest interactive notebook results [1, 2, 3].

Software is distributed under the Apache 2.0 license and available at https://github.com/CampagneLaboratory/MetaR and http://metaR.campagnelab.org.

References: [1] http://dx.doi.org/10.1101/030254 [2] Documentation http://tinyurl.com/zx9wvjw [3] BOSC 2016 Talk https://youtu.be/bWYhuQCSaqU

← Back

APOLLO: COLLABORATIVE AND SCALABLE MANUAL GENOME ANNOTATION

Nathan A Dunn[1], Monica Muñoz-Torres[1], Deepak Unni[2], Eric Yao[3], Colin Diesch[2], Ian Holmes[3], Chris Elsik[2], Suzanna Lewis[1]

[1]Lawrence Berkeley National Laboratory, Berkeley Bioinformatics Open-source Projects, Environmental Genomics and Systems Biology Division, Berkeley, CA, [2]University of Missouri, Division of Plant Sciences and Department of Animal Sciences, Columbia, MO, [3]University of California, Department of Bioengineering, Berkeley, CA

Manual curation is crucial to improving the quality of the annotations of a genome. It enables curators to refine automated gene predictions using experimental data and aligned predictions from closely related organisms to more accurately represent the underlying biology. Apollo is a web-based genome annotation editor that allows curators to manually revise and edit the structure and function of predicted genomic elements.

Apollo, built on top of the JBrowse genome browser, offers an 'Annotator Panel' that allows users to efficiently navigate the genome and its annotations. Changes are reflected in real-time to all users (similar to Google Docs) and aggregated in a revertible, visual history of structural edits. Apollo allows the export of sequences and metadata associated with each annotated genomic element in FASTA, GFF3, or Chado. A single Apollo server can be scaled to support multiple genome projects and curators. Access to genomes is controlled with fine-grained permissions (e.g. administrator, curator, public). To support integration into larger workflows, we expose the suite of web services that drives user-interface functionality. These web-services have been leveraged to integrate with Docker and the Galaxy platform.

Striving to increase Apollo's repertoire of visual exploration and exploratory analytics tools, two major undertakings are currently under development. First, the ability to visualize variant data and to annotate their predicted effects, primarily on coding regions. New technology trends and scientific paradigms point to new needs in genomic analytic tools to leverage information about variants that impact human health. Driven by a growing need to identify disease causing variants across diverse groups, we are working towards providing full functionality in genomic variant analysis and curation. Second, is the transformation of separate genomic coordinates into a single, synthetic region. This will allow the visualization of two or more genomics regions, from the length of entire chromosomes to just a few exons, within an artificially constructed genomic region. Artificially joining scaffolds facilitates annotation of genomic features split across two or more regions of a fragmented assembly (e.g, scaffolds), likely informing potential improvements to the genome assembly in the process. Additionally, this will allow hiding (visual genome folding) intra- and intergenic regions to provide a more information-rich visualization of the genome. For example, bringing exons closer together will facilitate annotating gene models with long introns, as sequences at the edge of exons separated by thousands of base-pairs will be shown adjacently.

Apollo is currently being used in over one hundred genome annotation projects around the world, ranging from annotation of a single species to lineage-specific efforts supporting the annotation of dozens of species at a time.

← Back

SCIKIT-RIBO REVEALS PRECISE CODON-LEVEL TRANSLATIONAL CONTROL BY DISSECTING RIBOSOME PAUSING AND CODON ELONGATION

Han Fang[1], Yifei Huang[1], Aditya Radhakrishnan[2], Max Doerfel[1], Adam Siepel[1], Rachel Green[2], Gholson Lyon[1], Michael Schatz[1,2]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Johns Hopkins University, Baltimore, MD

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

After improving the ribosome A-site resolution to 3bp, we built a negative binomial mixture model to identify and analyze ribosome pausing sites. From this we discovered the commonly used RPKM-based TE calculation is very sensitive to ribosome pausing events, thus negatively skewing the TE distributions in almost all previous studies and limiting their ability to differentiate translation efficiency and codon optimality. To solve this, we built a generalized linear model to simultaneously infer protein TE and codon elongation rates, while accounting for mRNA abundance and secondary structure. We show our prediction method has much higher accuracy for identifying A-site location than previous methods (0.90 vs. 0.64, 10-fold CV). Scikit-ribo's predicted genome-wide codon usage fraction also has a significant correlation with published estimates (ρ=0.87). We also successfully identified nearly 100 genes with over 100 ribosome pausing sites in wild-type yeast. Subsequently we discovered mRNA with stronger secondary structure tend to have pausing ribosomes (p-value<2×10-16). Scikit-ribo almost perfectly reproduced relative codon dwell time from Weinberg et al (ρ=0.99) and found significant correlation between tRNA abundance and codon elongation rates (ρ=0.53). We also showed a balanced log2(TE) distribution after accounting for mRNA secondary structure and codon elongation rates, revealing the systematic bias in typical Riboseq analysis. Together, these results show that scikit-ribo provides robust methods for Riboseq analysis and better understanding of translational control.

← Back

ESTIMATION OF NUCLEOTIDE- AND ALLELE-SPECIFIC SELECTION COEFFICIENTS IN THE HUMAN GENOME USING DEEP LEARNING AND POPULATION GENETICS

Yifei Huang, Adam Siepel

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology,
Cold Spring Harbor, NY

Recently, several computational methods have been developed to estimate the strength of negative selection on genomic sequences, either explicitly or implicitly. Methods of this kind, such as fitCons, CADD, and FunSeq2, not only provide insights into the functional constraints on genomic sequences but they can also help to prioritize putative disease variants for follow-up study. However, most existing methods are simple linear classifiers. Furthermore, even those based on models of evolution, such as fitCons, are unable to estimate selection coefficients, the most interpretable measures of natural selection. Here we describe a novel statistical model, DeepINSIGHT, that directly estimates selection coefficients for all possible mutations in the human genome while also accounting for potential nonlinear relationships between informative genomic features and selective effects. We formulate the estimation of selection coefficients as a regression problem in which the covariates are genomic features and the response is the observed derived allele frequency. DeepINSIGHT uses a deep-learning strategy to solve this regression problem with a cost function derived from the Poisson random field model in population genetics. First, DeepINSIGHT uses a mixture model based on population genetics to estimate the genome-wide distribution of selection coefficients across all possible mutations in the human genome while accounting for the demographic changes in the human evolutionary history. Given likelihood terms from this population genetic model, an extension of the mixture density network is then used to characterize the unknown relationship between variant-specific selection coefficients and genomic features. We apply DeepINSIGHT to a large number of genomic features and the high-coverage 1000 Genomes data, and show that it produces highly accurate estimates of variant-specific selection coefficients, unmatched by any existing computational methods. Using disease variants from the ClinVar and HGMD databases, we show that DeepINSIGHT is a powerful method both for obtaining insights into natural selection and for the prioritization of disease variants.

← Back

WINNER'S CURSE IN QUANTITATIVE GENOMICS STUDIES

Gregory Darnell[1], Jenny Tung[2], Christopher Brown[3], Sayan Mukherjee[4], Barbara Engelhardt[1,5]

[1]Princeton University, Lewis-Sigler Institute, Princeton, NJ, [2]Duke University, Department of Evolutionary Anthropology, Durham, NC, [3]University of Pennsylvania, Genetics, Philadelphia, PA, [4]Duke University, Departments of Statistics, Mathematics, Computer Science, Durham, NC, [5]Princeton University, Computer Science, Center for Statistics and Machine Learning, Princeton, NJ

Winner's Curse is a phenomenon characterized for common value auctions in economics to describe the winner of the auction tending to overpay for the item, and has been used to explain the overestimation of effect sizes and lack of reproducibility of associations that have plagued genomics studies [Zöllner et al., 2007]. The bias in effect size estimation traces back to agricultural QTL mapping studies as the Beavis effect, but has been widely overlooked in the human genetics community [Göring et. al., 2001]. Standard association mapping techniques in modern eQTL studies are mostly concerned with the trade-off between computational tractability and statistical power. However, little focus has been placed on accurate and reproducible effect size estimates, even when false positives are properly controlled and statistical power is high. A corollary to inflated effect sizes is that the causal association is rarely identified, but, instead, a non-functional genetic variant often has the most significant test statistic value. Even conditioning on detection of an association between the truly causal variant and gene expression level of interest, estimation of the strength of association is almost always inflated. We demonstrate with a suite of previously proposed and novel methods that trade off computational and statistical properties, that each method suffers from one or more of four consequences: (1) increasing false positives and decreasing true positives, (2) enrichment of discoveries at low minor allele frequency, (3) false discoveries that are not in high LD with causal variants, and (4) underestimation of locus heterogeneity. In addition, we show lack of reproducibility and lack of correspondence of effect sizes in real data between gene expression microarrays of the same individuals under baseline conditions and drug responses. We propose new directions for robust and accurate effect size estimates in attempt to increase reproducibility in eQTL studies and mitigate Winner's Curse.

NOT JUST A BLACK BOX: INTERPRETABLE DEEP LEARNING FOR GENOMICS

Avanti Shrikumar*[1], Peyton Greenside*[2], Anna Shcherbina[2], Johnny Israeli[3], Nasa Sinnott-Armstrong[4], Anshul Kundaje[1,4]

[1]Stanford University, Department of Computer Science, Stanford, CA, [2]Stanford University, Department of Biomedical Data Science, Stanford, CA, [3]Stanford University, Department of Biophysics, Stanford, CA, [4]Stanford University, Department of Genetics, Stanford, CA

* co-first authors

Deep neural networks have emerged as powerful learning engines for modeling sequence determinants of functional genomic signals [1] [2] [3] [4] [5]. However, methods for interpreting these models leave much room for improvement. In-silico mutagenesis [1] [2] is a commonly-used technique to infer the relevance of individual bases in the input sequence. However, it is computationally expensive, and (as we show) can even provide misleading results when the input contains redundant signals that buffer each other. Gradient-based methods are computationally efficient alternatives to in-silico mutagenesis for computing importance scores [4], but we show that they have distinct failure modes. For instance, an importance of zero may not always be appropriate even though the gradient is zero. Finally, individual neurons are often visualized and have been found to have some qualitative similarity to position weight matrix representations of regulatory sequence 'motifs' [1] [3] [4] [5]. However, this approach ignores the fact that multiple neurons may cooperatively describe a single pattern. Here, we present DeepLIFT (Deep Learning Important FeaTures), a family of computationally efficient techniques that address the aforementioned limitations. DeepLIFT compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference, thereby circumventing pitfalls of gradient-based methods. Per-base contribution scores are then clustered to reveal recurring patterns such as regulatory sequence motifs and DNAse footprints. DeepLIFT is also, to our knowledge, the first method to identify patterns by integrating the combined effects of multiple cooperating neurons. We apply DeepLIFT to models of transcription factor binding and chromatin state to detect novel sequence motif representations, infer high resolution point binding events of TFs, dissect regulatory sequence grammars, learn predictive nucleosome architectural features and unravel the context-specific heterogeneity of regulatory elements.

[1] DeepBind (Nat. Biotechnol. 33,831–838, 2015)
[2] DeepSEA (Nat. Methods 12, 931–934, 2015)
[3] Basset (Genome Research, 10.1101/gr.200535.115, 2016)
[4] DeepMethyl (Scientific Reports, 10.1038/srep19598, 2016)
[5] DeepMotif (ArXiv, 1605.01133, 2016)

RECONSTRUCTING CHANGES IN MUTATIONAL PROCESSES DURING TUMOUR EVOLUTION

Yulia Rubanova[1,4], Jeff Wintersinger[1,4], Amit Deshwar[3,4], Nil Sahin[2,4], Quaid Morris[1,2,4,5]

[1]University of Toronto, Department of Computer Science, Toronto, Canada, [2]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [3]University of Toronto, Department of Electrical and Computer Engineering, Toronto, Canada, [4]Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada, [5]Ontario Institute of Cancer Research, Toronto, Canada

Cancer is caused by somatic mutations that accumulate in cells throughout life. Mutations arise due to different processes occurring in a cell, associated with internal (e.g., failure in DNA repair, copy errors during replication) or external factors (e.g., smoking, exposure to UV light). Each process creates a unique pattern of mutations, known as a mutational signature. Signature's rate of generating mutations (signature exposure) changes over time and differs across tumors. Investigating variations in signatures exposures over time provides deeper understanding of cancer development and enables for more accurate estimation of patient survival.

We present a method to estimate temporal changes in signatures based on mutations from a single cancer sample. First, we separate mutations into 96 types based on their three-nucleotide context. Then we divide mutations into bins based on a pseudo-time estimate of their relative order of occurrence derived from PhyloWGS. We apply mixture of multinomials on each bin to estimate signature exposures. Signatures are represented as multinomial distributions over 96 types of mutations. We use signatures defined by Alexandrov et al. Obtained mixtures coefficients represent signature exposures. Finally, we evaluate uncertainty by bootstrapping set of mutations and recomputing exposure estimates.

Time points when signature profiles change substantially represent loss or gain of different mutational processes. To find a new change point, we iterate through all time points and recompute mixtures of multinomials in time slices formed by a potential change point. A point with maximum likelihood is considered a new checkpoint. We use Bayesian Information Criterion to estimate optimal number of change points.

We applied our approach to hundreds of samples of a diverse range of cancer types. Over 30% of samples in three cancer types have a signature that changes by more than 20% of overall exposure. To investigate if there are changes in signature exposures within subclonal populations, we compared our mutation groups based on exposure change point to mutation groups in tumor evolution tree reconstructed by PhyloWGS. Changes in signature exposures occur within pseudo-time boundaries that correspond to subclones.

Estimating future behaviour of the signatures allows to predict formation of new subclones, evaluate efficacy of treatment and estimate patient survival more accurately.

← Back

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

Coby Viner[1], James Johnson[2], Nicolas Walker[3], Hui Shi[3], Marcela Sjöberg[4], David J Adams[4], Anne C Ferguson-Smith[3], Timothy L Bailey[2], Michael M Hoffman[1]

[1]Univ of Toronto, Toronto, Canada, [2]Univ of Queensland, Brisbane, Australia, [3]Univ of Cambridge, Cambridge, United Kingdom, [4]Sanger Institute, Cambridge, United Kingdom

Motivation. 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 a mechanism by which widespread changes in DNA methylation and hydroxymethylation can dramatically shift active gene expression programs.
Methods. To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites (TFBSs) in DNA with covalent modifications. Our models expand the standard A/C/G/T alphabet, adding m (5mC) and h (5hmC). We additionally add symbols to encode guanine complementary to these modified cytosine nucleobases and represent states of ambiguous modification. We adapted the position weight matrix model of TFBS affinity to an expanded alphabet. We developed a program, Cytomod, to create a modified sequence. We also enhanced the MEME Suite to be able to handle custom alphabets. We created an expanded-alphabet sequence using whole-genome maps of 5mC and 5hmC in naive *ex vivo* mouse T cells.
Results. Using this sequence and ChIP-seq data from Mouse ENCODE and others, we identified modification-sensitive *cis*-regulatory modules. We elucidated 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. We demonstrated that our method is robust to parameter perturbations, with TF sensitivities for methylated and hydroxymethylated DNA broadly conserved across a range of modified base calling thresholds. Hypothesis testing across different threshold values was used to determine cutoffs most suitable for further analyses. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.
Discussion. Hypothesis testing of motif central enrichment provides a natural means of differentially assessing modified versus unmodified binding affinity. This approach can be readily extended to other DNA modifications. As more high-resolution epigenomic data becomes available, we expect this method to continue to yield insights into altered TFBS affinities across a variety of modifications.

← Back

CERES: A MODEL FOR INFERRING GENETIC DEPENDENCIES IN CANCER CELL LINES FROM CRISPR KNOCKOUT SCREENS

Jordan G Bryan*, Robin M Meyers*, Aviad Tsherniak

The Broad Institute of MIT and Harvard, Cancer Program Data Science, Cambridge, MA

One of the major goals in cancer research is to find genes that are essential to the proliferation of cancer cells, but are non-essential in normal cells. In recent years, the CRISPR-Cas9 system has furthered this goal by enabling biologists to systematically knock out individual genes in cancer cell lines and measure the resulting effect on cell viability. To increase the signal-to-noise ratio, experimentalists often use multiple reagents to target each gene and combine the reagent viability effects into one gene essentiality score. However, we and others have recently shown that, in addition to the gene knockout effect, the measured viability effect is also made up of a DNA cutting effect that is directly proportional to the number of cuts made by the Cas9 protein complex. Because cancer cells carry many genetic aberrations, the cutting effect presents a serious difficulty in identifying true dependencies that lie in amplified regions of the genome.

To address this problem, we introduce CERES, a model that estimates gene essentiality while accounting for the toxicity effect due to DNA cutting by Cas9. CERES assumes that each measurement is a linear combination of cell-specific gene effects and cell-specific cutting effects, scaled by the cutting efficacy of the RNA strand that guides the gene knockout. To test CERES, we use data from genome-scale CRISPR knockout screens generated by Project Achilles at The Broad Institute. The model is fit to measurements from all screened cell lines by maximum likelihood using stochastic gradient descent. Its performance is evaluated by prediction error on a held-out test set. We assess the biological validity of CERES gene solutions by benchmarking against known dependencies in cancer cell lines. Further, we show that the genetic dependency scores estimated by CERES have enriched correlation structure within annotated protein complexes, suggesting that the model uncovers functional relationships between genes. Importantly, we find that the number of false-positive dependencies in amplified regions is vastly reduced in the CERES gene solutions when compared to existing methods of estimating gene dependencies from multiple-reagent measurements.

*authors contributed equally

← Back

COMBINING PHENOME AND GENOME TO UNCOVER THE GENETIC BASIS FOR NATURALLY OCCURRING DIFFERENCES IN DEVELOPMENT AND BEHAVIOUR

Tiffany A Timbers[1], Catrina Loucks[2], Stephane Flibotte[3], Don G Moerman[3], Michel R Leroux[2]

[1]University of British Columbia, Statistics, Vancouver, Canada, [2]Simon Fraser University, Molecular Biology and Biochemistry, Burnaby, Canada, [3]University of British Columbia, Zoology, Vancouver, Canada

What is the genetic basis for naturally occurring differences in development and behaviour? To address this complex question we used a microscopic soil-dwelling nematode, *Caenorhabditis elegans* as a model. First we used an automated computer vision system, the Multi-worm Tracker, to collect high-dimensional and high-throughput behavioural and morphological characteristics (phenotypes) for each strain. A machine learning approach, iterative de-noising, is now being implemented to create a phenotypic profile for each strain collected from various locations around the world. We will then combine this phenotypic profile with the available whole-genome sequences for these strains via multivariate rare-variant analysis (e.g., MURAT) to identify naturally-occurring genetic variants responsible for behavioural and morphological phenotypes. Finally, genetic engineering via an innovative new technology, CRISPR, will be used to confirm the causality of the identified naturally occurring genetic variants on the identified behavioural and morphological phenotypes. This study will identify genetic elements that are important for naturally occurring differences in development and behaviour in *C. elegans* and utilizes a novel approach to identify such elements that may be useful in other organisms.

← Back

UNIVERSAL MICROBIAL DIAGNOSTICS USING RANDOM DNA PROBES

Amirali Aghazadeh[1], Adam Y Lin[2], Mona A Sheikh[1], Allen L Chen[2], Lisa M Atkins[3], Coreen L Johnson[3], Joseph F Petrosino[3], Rebekah A Drezek[2], Richard G Baraniuk[1]

[1]Rice University, Electrical and Computer Engineering, Houston, TX, [2]Rice University, Department of Bioengineering, Houston, TX, [3]Baylor College of Medicine, Department of Molecular Virology and Microbiology, Houston, TX

Early identification of pathogens is essential for limiting development of therapy-resistant pathogens and mitigating infectious disease outbreaks. Most bacterial detection schemes use target-specific probes to differentiate pathogen species, creating time and cost inefficiencies in identifying newly discovered organisms. Here, we present a novel universal microbial diagnostic (UMD) platform to screen for microbial organisms in an infectious sample using a small number of random DNA probes that are agnostic to the target DNA sequences. Our platform leverages the theory of sparse signal recovery (compressive sensing) to identify the composition of a microbial sample that potentially contains novel or mutant species. We validated the UMD platform in vitro using five random probes to recover eleven pathogenic bacteria. We further demonstrated in silico that UMD can be generalized to screen for common human pathogens in different taxonomy levels. UMD's unorthodox sensing approach opens the door to more efficient and universal molecular diagnostics.

← Back

A COMPUTATIONAL FRAMEWORK TO PREDICT CHROMATIN INTERACTION USING GENOMIC AND EPIGENOMIC DATA

Lin An[1], Tyler Derr[2], Yanli Wang[1], Feng Yue[1,2]

[1]Pennsylvania State University, Huck Institutes of the Life Sciences, State College, PA, [2]Pennsylvania State University, Department of Biochemistry and Molecular Biology, Hershey, PA

3-Dimensional genome organization is essential for the tissue-specific and developmental stage-specific gene expression, as it can reveal the physical interactions between distal regulatory elements and their target genes. Several recent high-throughput technologies based on chromatin conformation capture have emerged and given us an unprecedented opportunity to study the higher-order genome organization. Among them, Hi-C technology shows the greatest potential due to its unbiased genomewide coverage. Unfortunately, due to the sequencing cost and the complex nature of the experimental procedure, the application of Hi-C has only been limited to a small number of cell types. At the same time, thousands of epigenomic datasets have been generated through the ENCODE and Roadmap Epigenomics projects.

Here, we present a supervised machine-learning framework based on random forest method to impute Hi-C interactions, by taking advantage of the published epigenomic data from ENCODE and Roadmap Epigenomics projects. Our prediction model is based on the combination of genomic features (such as GC content and mappability) and epigenomic features (such as histone modifications and transcription factors). The result shows that our computational model can accurately predict Hi-C interaction matrices when compared with experimentally generated data, with Pearson correlation efficiency $r = 0.80$ (corrected by distance between contacts) when trained and predicted in the same cell type (GM12878) and $r = 0.77$ when trained and predicted in different cell lines (GM12878 vs K562). Moreover, we observe that different features possess different predicting powers and even for the same feature, it has different predicting power at different distance. For example, CTCF carries more weight when predicting long-range interaction at 1.25Mb distance than 400Kb.

This work provides an extremely useful resource in tissue/cell types where Hi-C data is not available. It will also be a huge boost to the study of tissue-specific gene expression by filling in the gaps between functional genome annotation by the Roadmap Epigenomics Project and 3D structure in hundreds of cell types. Moreover, our machine learning strategy can also discover and rank the importance of each genomic/epigenomic feature and suggest novel biological insights underlying chromatin interactions.

← Back

A CONVOLUTIONAL NEURAL NETWORK FRAMEWORK FOR MODELLING CIS-REGULATORY ELEMENTS WITH APPLICATION TO RNA STABILITY

Žiga Avsec[1,2], Julien Gagneur[1]

[1]Technical University of Munich, Department of Informatics, Munich, Germany, [2]Ludwig Maximilian University of Munich, Graduate School of Quantitative Biosciences (QBM), Munich, Germany

Despite their importance in controlling gene expression, we currently lack good models for predicting mRNA half-life and other levels of post transcriptional regulation (PTR) given a gene sequence. Recent studies have shown that convolutional neural networks are effective tools to model sequence elements, including in vitro DNA and RNA binding efficiencies and transcriptional enhancers [1,2]. Here, we developed a statistical model using convolutional neural networks to predict mRNA half-life from its sequence and quantify the effects of single nucleotide variants (SNVs). Our model is a neural network that models i) the contributions of CREs across a whole sequence using convolutional layers [1,2], ii) positional biases of CREs using smooth splines, iii) additional contributions of covariates such as codon-usage. Moreover the model is able to encode the cost function of the biophysical model featureREDUCE [3] by using exponential activation functions. Performance of the model was assessed on a genome-wide datasets of mRNA half-life data in S. Pombe [4]. The model shows improved prediction accuracy over other state-of-the-art methods and similar performance in quantifying the effect of single-nucleotide-variants. We provide an open-source implementation of our model in TensorFlow as a user-friendly python package. It can be directly used for predicting general categorical or continuous response variables from genomic sequence, together with quantifying SNV effects, assessing CRE positional biases and visualizing single base contributions.

[1] Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8), 831–838.
[2] Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods, 12(10), 931–4.
[3] Riley, T. R., Lazarovici, A., Mann, R. S., & Bussemaker, H. J. (2015). Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. eLife, 4(December), e06397.
[4] Eser, P., Wachutka, L., Maier, K. C., Demel, C., Boroni, M., Iyer, S., … Gagneur, J. (2016). Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. Molecular Systems Biology, 12(2), 857–857.

INVESTIGATING HOW THE TRANSCRIPTION FACTOR *FKH-8* IS EXPRESSED IN BAG NEURONS AND THE DOPAMINERGIC PATHWAY.

Mohammed O Awan, Bryan E Cawthon, Brian L Nelms

Fisk University, Biology, Nashville, TN

Caenorhabditis elegans are incredibly useful nematodes whose entire genome has been sequenced. *C. elegans* are commonly used in research, much due to their rapid life cycle, transparency, and because they are genetically conserved with humans. *C. elegans*, move to areas with optimum levels CO_2 with CO_2-aroused Calcium responses in BAG and AFD neurons. Theses neurons encourage movement as CO_2 levels increase and inhibit movement as CO_2 levels fall. This works to keep the worm at an optimum level of CO_2.

The transcription factor forkhead-8 (*fkh-8*) is hypothesized as a negative regulator for BAG neurons due to quantified fluorescence in guanylyl cyclase (*gcy-33*) in *C. elegans* mutants (both with and without the transcription factor *fkh-8*). The *gcy-33* gene is found on chromosome V that uses fluorescent proteins wherever BAG neurons are being expressed. In images under a fluorescent microscope, worms without the *fkh-8* transcription factor would have a higher intensity of expression compared to the wild-type worms with the transcription factor *fkh-8*.

We have used ImageJ, an open-source image processing tool that can be used to analyze pictures for scientific purposes, for our image analysis. The relative fluorescence value obtained is calculated from the difference of integrated density of the area of expression in the picture, and the product of the area of the expression with the mean gray value of a selection of the background. The two-tailed p-value from the t test result equals 0.0014, suggesting an accurately measured fluorescence of the worms.

The *fkh-8* transcription factor is also highly expressed in dopaminergic neurons, as well as many other sensory neurons. The phenotypes expressed in mutants missing the transcription factor *fkh-8*, seem to imply that there is an overflow of dopamine in the synapse. This excess dopamine spills into the dopamine receptors of surrounding cholinergic motor neurons, causing the worms to induce Swimming Induced Paralysis (SWIP) faster in worms missing the transcription factor *fkh-8*.

In the future, we will continue to investigate how *fkh-8* is expressed in other sensory neurons and how it affects the dopaminergic pathways.

← Back

TRANSPOSABLE ELEMENT-MEDIATED GENE REGULATION IN THE PARASITE TRICHOMONAS VAGINALIS

Martina Bradic, Sally Warring , Jonathan Bermeo, Jane Carlton

Center for Genomics and Systems Biology, Department of Biology, New York University, New York , NY

Transposable elements (TEs) have been characterized as “junk DNA” and their variation and function largely neglected. With the development of genome-wide technologies, there is growing evidence that TEs co-evolved with the genome of their hosts, and represent an important force shaping genomic evolution and plasticity of almost all organisms. *Trichomonas vaginalis* is a parasite that causes the most common non-viral sexually transmitted disease worldwide. A remarkable 1/3 of the genome is composed of DNA transposons, including a family of ~1,000 Mariner elements and ~3,000 Maverick elements. The likely role of TE activity in the massive genome expansion in the parasite raises compelling questions regarding the impact that thousands of TEs have on the biology and evolution of parasitism. We developed Transposon Display sequencing (TD-seq), a genome-wide sequencing method that selectively amplifies TE families and their adjacent regions in order to characterize TE insertion variation in different isolates of *T. vaginalis*. Using this method, we identified 1,385 Mariner (Tvmar1) insertions in 16 *T. vaginalis* geographically-distinct isolates. Our analysis suggests that Tvmar1 inserts more frequently next to genes than gene-free regions and that the majority of elements are present in low frequencies, underlining the possible deleterious effect of insertions and their purging by purifying selection. We found some of the TE insertions are domesticated and have differing impacts on host gene expression, which we are confirming by whole transcriptome analysis. Moreover, TEs seem to regulate expression of abundant gene families (i.e., BspA family) related to the parasite’s virulence. We are currently developing computational pipelines to identify non-annotated insertions across different *T. vaginalis* regions. This is the first genome-wide *T. vaginalis* study to show the impact of Tvmar1 population dynamics and its contribution to transcriptional regulation that could potentially shed light on the role of TEs in medically important traits.

FAST ANNOTATION OF METAGENOMIC SHOTGUN SEQUENCES WITH A MICROBIAL GENE CATALOG AND BOWTIE2

Stuart M Brown[1], Konstantinos Krampis [2], Hao Chen[1], Yuhan Hao[3]

[1]New York University School of Medicine, Center for Health Informatics and Bioinformatics, New York, NY, [2]Hunter College, Biology, New York, NY, [3]Fordham University, Computer Science, New York, NY

The collection of large scale "shotgun" DNA sequence data sets from microbes associated with the environment, the human body (human microbiome), or associated with other animals has become very common. A key goal for the analysis of metagenomic shotgun (MGS) data is to estimate the functional metabolic capacity of the microbial community by mapping the DNA sequence fragments to a reference database. BLAST is the most commonly used (and the most sensitive) method to compare DNA sequences to a database, but it requires hundreds of CPU hours to analyze a typical MGS sample FASTQ data file with hundreds of millions of reads.

We have created a computationally efficient pipeline for MGS analysis, which combines the MetaPhlAn tool for rapid and accurate profiling of taxonomic composition and Bowtie2 (vs. a catalog of microbial genes) for rapid functional profiling of microbial sequence fragments. Our pipeline can process a typical MGS sample (two paired-end FASTQ files) in about an hour on an 8 CPU Linux computer to produce QC report, taxonomic abundance, and gene function abundance (by KEGG ortholog ID). Over 60% of (cleaned) reads from a human gut microbiome sample can be mapped witht his system.We have implemented this pipeline as a Virtual Machine (VM) in a Docker container which can be freely downloaded and run on any computer, or run as a cloud server on the Amazon EC2.

PLANEMO – A SCIENTIFIC WORKFLOW SDK

John M Chilton[1], Aysam Guerler[2], the Galaxy Team [1,2,3]

[1]The Pennsylvania State University , Biochemistry and Molecular Biology, University Park, PA, [2]Johns Hopkins University, Departments of Biology and Computer Science, Baltimore, MD, [3]The George Washington University, Computational Biology Institute, Asburn, VA

A novel approach to building, refining, and running scientific workflows leveraging Galaxy through Planemo will be presented. The Galaxy workflow editor and workflow extraction interface are great tools enabling any Galaxy user to easily build workflows. However, tool authors using Planemo and sophisticated bioinformaticians may prefer driving workflow development through their existing tool chains such as programming text editors, command-line testing, and revision control. The approach presented leverages YAML-based workflow descriptions as plain files allowing exactly this.

The Planemo-driven approach will be used as a lens to highlight these workflows formats (Format 2 Galaxy workflows and Common Workflow Language (CWL) workflows) as well as important highlights from the myriad of recent Galaxy workflow enhancements that have made them dramatically more usable, powerful, and performant.

Available today, Format 2 Galaxy workflows map directly to existing Galaxy tool and workflow concepts and are described in a very concise and readable YAML format. CWL specifications for tools and workflows are developed in an open fashion by many organizations with the aim of creating truly portable descriptions. The execution of CWL workflows in Galaxy is being actively worked on and progress will be discussed.

Underlying all of this is core Galaxy enhancements that will be demonstrated. The user interface for workflows has been overhauled and improved. Additionally, workflows now allow nesting, labels, non-data inputs, implicit connections between steps, and many new operations over collections - greatly increasing the expressive power of Galaxy workflows. Finally, recent performance enhancements allow Galaxy workflows to scale to thousands of datasets.

METHODS FOR RAPID AND SCALABLE QUALITY ASSESSMENT OF RNA STRUCTURE PROBING DATA

Krishna Choudhary, Huan Chen, Luyao Ruan, Nathan P Shih, Fei Deng, Sharon Aviran

University of California Davis, Biomedical Engineering and Genome Center, Davis, CA

RNA is a biomolecule that plays an integral role in many biological processes, ranging from adaptive immunity to biocatalysis. In fact, RNAs have been implicated in diseases as well as utilized for their potential in medicine and biotechnology. To serve numerous functional roles, RNA must fold into specific structures. RNA structure has been traditionally studied with crystallography, NMR spectroscopy or phylogenetic analysis but these are costly, labor-intensive, and of limited applicability. Computational methods that make predictions based on sequence information alone are scalable but display poor accuracy. The recent advance of new structure probing techniques coupled with high-throughput sequencing has helped RNA studies expand in scope and depth to *in vivo* genome-wide capabilities. These techniques generate data that have also been utilized to significantly improve accuracy of prediction algorithms.

Numerous probing techniques that differ in protocols and analysis platforms have been recently developed. Despite their differences, most experiments face similar challenges in assessing reproducibility due to the stochastic nature of chemical probing and sequencing. To date, quality of such data is assessed through visual inspection or simple statistical tests, which are often applicable only to specific techniques. However, as protocols expand to genome-wide studies, quality control becomes a daunting task. General and efficient methods are needed to quantify variability and quality in the broad range of existing and emerging techniques.

We recently developed a new method to rapidly and quantitatively evaluate data reproducibility in probing experiments. We used a signal-to-noise ratio concept to evaluate replicate agreement, which has the capacity to identify high-quality data as well as to screen for potential structure differences. We demonstrated efficacy and utility of the method on numerous recently published small and large-scale datasets (Choudhary *et al.*, *Bioinformatics*, 2016). We have found theoretical relationships between noise in structural data and controllable design parameters in probing experiment. Such relationships can guide rational design of experiments to meet desired quality criteria. Additionally, we developed realistic noise models to evaluate sensitivity of secondary structure prediction, which is one of the important applications of structure probing data, to noise in data. Conclusions from our modeling studies can help biologists set quality criteria for their probing data to meet desired accuracy in structure prediction. Finally, we have refined our methods and integrated them with novel quality summaries in an interactive visualization tool to facilitate smooth transfer to community of biologists.

← Back

BENCHMARKING ALIGNMENT AND QUANTIFICATION TOOLS ON PLANT RNA-SEQ DATA WITH CYVERSE CYBERINFRASTRUCTURE

Kapeel M Chougule[1,3], Liya Wang[1,3], Joshua Stein[1], Doreen Ware[1,2,3]

[1]Cold Spring Harbor Laboratory, Ware Laboratory, Cold Spring Harbor, NY, [2]USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY, [3]CyVerse, Bio5 Institute, Tucson, AZ

RNA-seq has become a workhorse in transcriptomic studies. This is primarily because improvement in sequencing technology over the years has made it more affordable to perform full transcriptomic experiments with deeper coverage as compared to initial days of RNA sequencing. As a results this has manifested in a data deluge from RNA-seq experiments. The current ecosystem of RNA seq tools have more efficient and scalable options, this involves inclusion of tools like HISAT2, StringTie and Ballgown as an alternate to the existing Tuxedo protocol which has been the preferred suite of tools used for RNA-seq analysis.There is also emergence other pseudo-alignment based RNA-seq tools like Kallisto, Sailfish and Salmon which provide an alternate to the existing aligners. Most tools have been benchmarked for scalability and sensitivity using human datasets, in this study we will be using a powerful computational infrastructure named Cyverse in benchmarking public plant dataset and simulated dataset with new and existing tools in Cyverse. Plant genomes although being less complex tends to usually have larger genomes compared to animal, with genes differing in terms of intron size, type of alternative splicing and showing low compactness in gene configuration for highly expressed genes. These differences raise questions on the effectiveness and strategy of using the new tools on analyzing plant data set. In this study we will benchmark these tools for accuracy and speed, providing a guideline for using them on plant RNA seq studies.

INTEGRATIVE APPROACHES FOR VARIANT INTERPRETATION IN CODING REGIONS

Declan Clarke[1,2], Sushant Kumar[1,2], Mark Gerstein[1,2,3]

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

The pace of data generation by next-generation sequencing is presenting considerable challenges in terms of variant interpretation. Though deep sequencing is unearthing large numbers of rare single-nucleotide variants (SNVs), the rarity of these variants makes it difficult to evaluate their potential deleteriousness with conventional phenotype-genotype associations. Furthermore, many disease-associated SNVs act through mechanisms that remain poorly understood. 3D protein structures may provide valuable substrates for addressing these challenges. We present two general frameworks for doing so. In our first, we employ models of conformational change to identify key allosteric residues by predicting essential surface pockets and information-flow bottlenecks (a new software tool that enables this analysis is also described). In our second approach, we use localized frustration, which quantifies unfavorable residue interactions, as a metric to investigate the local effects of SNVs. In contrast to this metric, previous efforts have quantified the global impacts of SNVs on protein stability, despite the fact that local effects may impact functionality without disrupting global stability (e.g. in relation to catalysis or allostery). Importantly, although these two frameworks are fundamentally structural in nature, they are computationally efficient, thereby making analyses on large datasets accessible. We detail how these database-scale analyses shed light on signatures of conservation, as well as known disease-associated variants, including those involved in cancer.

← Back

POLLEN COUNTING AND ANALYSIS WITH THE ATTUNE ACOUSTIC FLOW CYTOMETER

Schuyler Corry, Rachel K Smith, Barb Seredick

Thermo Fisher Scientific, R&D, Eugene, OR

Pollen counting is currently performed manually by a trained analyst, severely limiting the amount of data that can be collected. In the United States, the National Allergy Bureau requires certification through the American Academy of Allergy Asthma & Immunology program for all pollen counters. This certification includes a multiple-choice exam and a slide identification exam, so has a significant barrier to entry for budding palynologists. Additionally, the lack of certified pollen counters means that geographic data is very coarse in some regions—in Oregon, for example, there is only a single pollen counting station, which is located in the pollen-heavy Willamette Valley. Flow cytometry is a widely used tool to count and characterize small particles, but has been used infrequently for pollen analysis, and in most of those instances the researchers wanted to characterize polyploidy of particular species. Autofluorescence of phytochemicals decreases the utility of many types of cellular staining and analysis for pollen particles. However, this seemingly negative property of autofluorescence can be exploited to differentiate pollen populations. Flow cytometers collect up to ~50 parameters for each triggering event, including forward / side-scatter, as well as multi-color fluorescence for each excitation laser, yielding a rich data set for clustering of the pollen sub-populations. Furthermore, the precision volumetric delivery system of the Attune flow cytometer means that absolute particle counts can be obtained without using reference beads. And although pollen is responsible for allergy and asthma responses, this approach would also quantify other aerosol particulates such as from motor vehicle exhaust. More granular data on environmental aerosols would be useful not only for people suffering from allergies and asthma, but also for policy makers and permit issuers.

← Back

COMPUTATIONAL ANALYSIS OF ENDOGENOUS RETROVIRAL ELEMENT EXPRESSION DURING THE DEVELOPMENT OF HUMAN EMBRYOS

Engin Cukuroglu, Jonathan Goke

Genome Institute of Singapore, Computational and Systems Biology, Singapore, Singapore

Improvements in sequencing techniques let us study transposable elements with high amount of data. In order to do a systematic analysis of transposable elements, fast and accurate tools have to be developed to help understanding the data. Here, we generated a pipeline to analyse differential expression of transposable elements in the human genome. We run our pipeline on single cell RNA-Seq data from pre-implantation embryos and pluripotent cells to confirm transcription of endogenous retroviral elements (ERV). We found that expression of ERVs is sufficient to cluster cells according to developmental stage, indicating a highly specific pattern during embryonic development. Interestingly, we found that expression of a specific subset of ERV elements is important for each different stage of early human embryonic development. Especially HERVH and LTR7Y are indicative of naïve pluripotency, suggesting that their activity dynamically reacts to changes in regulatory networks. We are currently applying the pipeline to identify transposable elements in human disease samples. In summary, our pipeline is a useful tool to study the contribution of transposable elements in genome wide data.

← Back

EVOLINC: A COMPUTATIONAL PIPELINE FOR COMPARATIVE GENOMIC AND TRANSCRIPTOMIC ANALYSES OF LONG NON-CODING RNAS FROM LARGE RNA-SEQ DATASETS

Upendra Kumar Devisetty[1], Andrew D Nelson[2], Asher K Haug-Baltzell[1], Eric Lyons[1], Mark A Beilstein[2]

[1]CyVerse, University of Arizona, Bio5, Tucson, AZ, [2]University of Arizona, Plant Sciences, Tucson, AZ

Transcriptomic analyses from across eukaryotes have led to the conclusion that most of the genome is transcribed at some point in the developmental trajectory of an organism. One class of these transcripts is termed long noncoding RNAs (lncRNAs) which lack coding potential and have a length >200 base pairs (bp). Reported lncRNA repertoires in mammals vary, but are commonly in the thousands to tens of thousands of transcripts, accounting for ~90% of the genome. With the recent advances in sequencing technology that have opened a new era in RNA and DNA studies and subsequent generation of large datasets, attention has focused on understanding the evolutionary dynamics of lncRNAs, particularly their conservation within genomes. To facilitate lncRNA discovery and comparative analyses at the genomic and transcriptomic level from large RNA-Seq datasets, we present Evolinc, a computational pipeline that identifies long non-coding RNAs from transcriptome assembly files and then searches for homologs in other species. Using reiterative and reciprocal BLAST based approaches, Evolinc reconstructs families of homologous lncRNAs, aligns the constituent sequences, builds gene trees, and uses gene tree / species tree reconciliation to infer evolutionary processes. The novelty of this computational approach is that it allows the user to investigate factors affecting lincRNA diversity within a large number of species. This approach is scaleable, highly memory efficient working on one to thousands of lncRNAs and can perform comparisons between both large and small genomes. Evolinc is useful not only for inferring mechanisms affecting lncRNA diversity, but also for identifying lncRNAs in non-model systems where genome data is available. For ease of use we have pre-packaged Evolinc into an image available in the CyVerse’s Atmosphere cloud computing service, as well as in CyVerse's Discovery Environment, a Graphical User Interface web browsing service.

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, Office of Research Information Services, Washington, DC, [2]Smithsonian Institution, Smithsonian Environmental Research Center, Washington, DC

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 3,000 taxa representing all available taxonomic groups. Two separate sets of genome alignments, one with a Bacteria reference species and one with an Archaea reference species are considered. Beyond the phylogenetic results, strategies for data mining, quality control, and visualization for large comparative genomics datasets are presented.

← Back

BUILDING A BIOLOGICAL DATA SCIENCE INFRASTRUCTURE AT THE SMITHSONIAN INSTITUTION

Paul B Frandsen, Rebecca B Dikow

Smithsonian Institution, Office of Research Information Services,
Washington, DC

The Smithsonian Institution is one of the most public, yet least understood government institutions. It boasts 20 museums and galleries, 9 research centers, and a zoo. The recent boom in the amount of research data being generated by traditionally “small data” fields has resulted in the formation of the Smithsonian’s first ever research computing and data science department. From its inception, a major focus of this effort has been to support the computational initiatives of the Smithsonian Institute for Biodiversity Genomics. Research utilizing genome resources across the tree of life can be challenging due to a paucity of computational tools purpose-built to account for genomic variation among diverse organisms and a lack of high quality reference genomes. Here we focus on insights gained from our efforts to enable this research by (1) providing computational training, (2) improving access to high performance computing, (3) tracking the research lifecycle through the Smithsonian’s data management system, and (4) working with industry partners to enable and refine biodiversity genomics research tools in the cloud.

← Back

COMMUNITY COMPUTATIONAL CHALLENGES IN BIOLOGICAL DATA SCIENCE: AN OPTIMISTICALLY CAUTIONARY TALE

Iddo Friedberg

Iowa State University, Veterinary Microbiology and Preventive Medicine, Ames, IA

While biological data of many types are proliferating, we are witnessing a renaissance of community efforts forming around different data types, trying to better understand and analyze them. One type of community effort is the community challenge, in which groups compete to create methods to better analyze the data and understand the underlying biology. Data analysis by a community challenge holds many hopes to improving our understanding of data, but also poses questions about how well a purely method-oriented endeavor with strict guidelines can capture our understanding of the data.

In my talk I shall review the history and evolution of community challenges in computational biology, starting with the various “Critical Assessments”, and through today’s DREAM challenges. I will discuss the motivation that lead to the challenges, the results, and how certain fields have been advanced or even transformed by challenges. I will also discuss the issues that arise from the tension between collaboration and competition, setting assessment metrics, over-interpreting challenge results, data sharing and confidentiality, manuscript authorship, funding, and other aspects of a large community endeavor. Examples will be provided from the Critical Assessment of Protein Function Annotation (CAFA) to illustrate some of my points, and best practices as we have learned them.

← Back

COMPLETE SITE SATURATION OF *BACILLUS SUBTILIS* LIPASE A TO STUDY THE INFLUENCE OF SINGLE AMINO ACIDS ON THE OVERALL STABILITY IN SURFACTANTS

Alexander Fulton[1], Josiane Frauenkron-Machedjou[3], Pia Skoczinski[2], Susanne Wilhelm[2], Uli Schwaneberg[3], Karl-Erich Jaeger[2]

[1]Novozymes A/S, Lipid Stain Removal, Bagsvaerd, Denmark, [2]Heinrich-Heine-University Duesseldorf, Research Center Juelich, Insitute for molecular Enzyme Technology, Juelich, Germany, [3]RWTH Aachen University, Lehrstuhl fuer Biotechnologie, Aachen, Germany

A model enzyme was used to determine the effect of every single amino acid substitution on surfactant tolerance. *Bacillus subtilis* lipase A (BSLA) is a minimal α/β-hydrolase of 181 amino acids with a known crystal structure. Site saturation mutagenesis resulted in a library of 3439 variants, each with a single amino acid exchange as confirmed by Sanger sequencing. The library was tested against four surfactants, namely SDS, CTAB, Tween 80, and sulfobetaine. From a first analysis, surface remodeling emerged as an effective engineering strategy to increase tolerance towards surfactants. The results show that single amino acid exchanges can significantly affect the tolerance for each of the four surfactants. This systematic analysis provides an experimental dataset to help derive novel protein engineering strategies as well as to direct modeling efforts.

ARID1A MUTATIONS IN AN *APC*/*PTEN*-DEFICIENT MOUSE MODEL OF OVARIAN TUMORS PROFOUNDLY ALTER DNA METHYLATION AND TRANSCRIPTION

Nicholas Giangreco[1,2], Hanna M Petrykowska[1], Alexandra Scott[1], Valer Gotea[1], Gennady Margolin[1], Cho R Kathleen[3], Laura Elnitski[1]

[1]Translational and Functional Genomics Branch, National Human Genome Research Institute, Rockville, MD, [2]Columbia University Medical Center, Systems Biology, New York, NY, [3]University of Michigan, Pathology, Ann Arbor, MI

ARID1A is a DNA binding subunit of the BRG1-associated factor (BAF) chromatin-remodeling complex, which regulates gene transcription by repositioning nucleosomes and thus possibly altering the DNA methylation landscape. Inactivation of the tumor suppressor ARID1A is frequently observed in ovarian carcinomas with deregulated canonical WNT/β-catenin and PI3K/AKT signaling pathways[1,2]. Prior work has shown homozygous inactivation of the *Apc* and *Pten* tumor suppressor genes, which deregulate the WNT/β-catenin and PI3K/AKT pathways, in the mouse ovarian surface epithelium promotes tumor formation with histology and gene expression traits similar to human ovarian endometrioid carcinomas (OECs)[2]. Inactivation of the tumor suppressor gene *Arid1a* in these *Apc*/*Pten*-deficient tumors leads to formation of ovarian tumors that differ in epithelial differentiation and tumor aggressiveness[3].

To address the molecular effects of mutations in the gene encoding the chromatin-modifying enzyme Arid1a, we compared existing knockout mouse models of ovarian endometrioid carcinoma with and without the mutation. Whereas homozygous inactivation of *Apc* and *Pten* resulted in net genome hypomethylation relative to normal mouse ovarian tissue, the additional homozygous deletion of *Arid1a* resulted in a switch to net genome hypermethylation. This divergent methylation pattern was enriched for genes that regulate distinct oncogenic processes, such as mesenchymal stem cell differentiation and TGFβ signaling in *Apc*$^-$;*Pten*$^-$ ovarian tumors, and non-canonical Wnt receptor signaling and epithelial differentiation in *Apc*$^-$;*Pten*$^-$;*Arid1a*$^-$ ovarian tumors. Furthermore, differential DNA methylation associated with *Arid1a* loss occurred most prominently at CpG islands and promoters, where aberrant promoter methylation inversely correlated with gene expression and significantly inhibited binding of transcription factors such as C-Myc and Max. Thus, we provide a list of methylation and expression alterations that are directly related to mutation of *Arid1a* in a mouse model of human ovarian endometrioid carcinoma. These altered sites implicate direct targets of the SWI/SNF complex, which contains Arid1a. In the future, this information on aberrant methylation and differential gene regulation may contribute to improved models of human tumors, fostering precision medicine.

References

1. Mao et al. (2013) *Journal of Gynecological Cancer* 24: 376-381.
2. Wu et al. (2007) *Cancer Cell* 11: 321-333.
3. Zhai et al. (2016) *Journal of Pathology* 238: 21-30

← Back

TOWARDS "DRY SIDE" REPRODUCIBILITY: A TECHNICAL SURVEY OF THE CHALLENGES TO AND BUILDING BLOCKS FOR REPRODUCIBLE, SECURE, AND USABLE (MICROBIAL) GENOMICS DATA INFRASTRUCTURE

Nicholas Greenfield, Samuel Minot, Roderick Bovee

One Codex, Reference Genomics, Inc., San Francisco, CA

The number of microbial genomics sequencing projects has exploded in recent years, consisting of both ever-larger consortia efforts (e.g., HMP, MetaHIT; PathoMAP, MetaSUB) as well as thousands of smaller studies. At the same, a rapidly growing ecosystem of bioinformatics tools continues to develop, facilitating myriad primary and secondary analyses of these datasets. Despite these advances, however, both exploratory and inferential meta-analyses of these datasets remain extremely challenging – confounded by both wet lab and “dry side” factors including: different sample preparation techniques and sequencing technologies; inconsistent, incomplete metadata and phenotype information; highly variable code and documentation quality across bioinformatics tools; and nonstandard, non-reproducible computational environments.

Here we offer a deep technical dive into the infrastructure of the One Codex platform for microbial genomics, and our own modest attempt to realize some of the opportunities presented by the move from physical specimen biobanks to genome sequencing (and other -omics) "databanks". Specifically, we describe – and attempt to unpack the jargon surrounding – the following key components of the One Codex platform:

- (1) Infrastructure for scheduling and processing individual analyses while ensuring the "bitwise reproducibility" of results (relevant jargon: containerization, Docker, deterministic workflows);

- (2) The design of a simple, easy-to-use application programming interface (API) for analysis of 100s or 1000s of multi-GB metagenomic samples (relevant jargon: API versioning, HATEOAS and REST, object schemas); and

- (3) Secure notebook environments that encourage exploratory, yet reproducible, analyses building on top of #1 and #2 (relevant jargon: Jupyter, Kubernetes, multitenant architectures).

Finally, we outline how these parts integrate to provide reproducible, extensible, and secure building blocks for microbial genomics workflows, while also being easy-to-use for a wide range of both technical and less bioinformatically-savvy end users.

IN VIVO CHARACTERIZATION OF LINC-P21 REVEALS FUNCTIONAL CIS-REGULATORY DNA ELEMENTS

Abigail F Groff[1,2,3], Diana B Sanchez-Gomez[1], Marcela M Soruco[1], Chiara Gerhardinger[1], James C Lee[1,4], John L Rinn[1,2,3]

[1]Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, [2]Systems Biology, Harvard University, Cambridge, MA, [3]Broad Institute of MIT and Harvard, Cambridge, MA, [4]Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom

The Linc-p21 locus, encoding a long non-coding RNA, plays an important role in p53 signalling, cell cycle regulation, and tumour suppression. However, despite extensive study, confusion exists regarding its mechanism of action: is activity driven by the transcript acting in trans, in cis, or by an underlying functional enhancer? Here, using a knockout mouse model and a massively parallel enhancer assay, we delineate the functional elements at this locus. We observe that even in tissues with no detectable Linc-p21 transcript, deletion of the locus significantly affects local gene expression, including of the cell cycle regulator Cdkn1a. To characterize this RNA-independent regulatory effect, we systematically interrogated the underlying DNA sequence for enhancer activity at nucleotide resolution, and confirmed the existence of multiple enhancer elements. Together, these data suggest that, in vivo, the cis-regulatory effects mediated by Linc-p21, in the presence or absence of transcription, are due to DNA enhancer elements.

← Back

THE QUANTITATIVE RELATIONSHIP BETWEEN HISTONE MODIFICATIONS AND GENE EXPRESSION ACROSS DIFFERENT INDIVIDUALS

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

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

Histone modifications are known to mark functionally important regions such as promoters and enhancers. Typically, histone modifications are profiled in cell lines and much work has been done understanding the relationship between histone modifications and cell type differences. Here, we ask to what extent a quantitative measure of histone modifications are related to phenotype across individuals, within a cell type. As phenotype we use gene expression, a functionally relevant readout. We use publicly available data on lymphoblastoid cell lines from different individuals profiled using ChIP-seq for H3K4me3, H3K4me1 and H3K27ac and show that these measures are significantly correlated with gene expression in some, but not all, genes. For example, H3K4me3 at promoters is significantly correlated with gene expression for 3,388 loci. We show that the input channel, a commonly used control experiment, is significantly correlated with gene expression. We show that using the input channel as a control results in a 5-50x reduction in the number of significant correlations. We replicate our results for H3K4m3 in rats.

← Back

USE OF LOW COVERAGE SHORT READ SEQUENCE DATA FOR QUALITY CONTROL AND FAMILIAL RELATIONSHIP IDENTIFICATION

Nancy F Hansen[1], Benjamin Heil[1], Miriam K Minsk[1], Sean Conlan[2], Pamela J Thomas[1], Julia A Segre[2], James C Mullikin[1]

[1]National Human Genome Research Institute, Cancer Genetics and Comparative Genomics Branch, Bethesda, MD, [2]National Human Genome Research Institute, Translational and Functional Genomics Branch, Bethesda, MD

Recently published algorithms allow the inference of individual ancestry and sample identity from very low coverage, short read sequence data. These statistical methods make use of sparse sequence data at the individual read level, bypassing the need for variant and genotype prediction. In this way, users can leverage the large size of the genome to make accurate predictions about samples from datasets containing as little as 0.1x sequence coverage.

In this work, we extend the published methods to estimate the accuracy of individual predictions of identity, and to predict parental versus sibling relationships from very low coverage datasets. We perform a large-scale quality control check on sequence data from a large metagenomic study of the skin microbiome, detecting potential sample contamination and/or misidentification in a set of over 1,000 sequence libraries, each containing only small amounts of human DNA.

In addition, we test these methods on low-coverage subsets of publicly available whole genome datasets from family trios and large pedigrees, attempting to use patterns of identity by descent to predict whether two datasets are derived from the same individual, parent and child, a sibling pair, or more distantly or unrelated individuals.

← Back

PREDICTION OF ESSENTIAL CODING AND NONCODING ELEMENTS USING CRISPR

Elizabeth Hutton, Chris Vakoc, Junwei Shi, Adam Siepel

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Deleterious genomic variants are difficult to distinguish from the vast majority of benign mutations in the average human genome. Recent development of CRISPR-based viability screening assays allows quantitative evaluation of large sets of specific sites in their native genomic context. These screens cover a limited fraction of the genome, however, requiring computational prediction to comprehensively evaluate unannotated variants. By training a machine learning model with CRISPR screens of coding and noncoding regions, our method generates a score for the genome-wide identification of regions essential for cell growth and proliferation.

In the initial phase of this project, we have created a logistic regression model to predict the impact of CRISPR-guided deletions in protein coding regions. Our model is informed by DNA, amino acid, and structural features, and benchmarked with established coding variant function prediction tools. We are in the process of applying our method to noncoding regions, and continue to refine our machine learning methods. Guided by our insight from CRISPR screening assays, this work’s genome-wide prediction of variant function aids in the continuing refinement of personalized genomics and the ranking of causal variants in disease.

← Back

SHELL SCRIPTING-BASED MATHEMATICAL MODELING FOR TRANSCRIPTOME DATA WITH BOOTSTRAPPING AND PARALLEL COMPUTING

Kazuo Ishii[1], Kazutaka Nakamura[2], Nobuaki Tounaka[2]

[1]Tokyo University of Agriculture and Technology, Department of Applied Biological Science, Tokyo, Japan, [2]Universal Shell Programming Laboratory, Ltd, Tokyo, Japan

Shell scripting-based mathematical modeling for transcriptome data using bootstrapping and parallel computing is a flexible and rapid method for mathematical modeling to create the phenotype prediction system. The choice of combination of variables is important in the mathematical modeling for predicting phenotypes. A large number of explanatory variables cause multicollinearity and over-fitting, and cause performance deterioration in the mathematical model. Therefore, the choice of a small number of variables is demanded. The most effective variable selection method is to select the optimal combination by evaluating all possible combination of variables, called "brute-force method". In the case of large number of candidate genes, it is very difficult to evaluate all possible combination of variables by a brute-force method because its number of combinations increases exponentially when candidate genetic numbers increase. Random extraction using bootstrapping from all possible combination, and evaluation of discrimination performance using parallel computing is an effective method for optimization of a mathematical model. We have performed optimization of a mathematical model by bootstrapping and parallel computing with R. In this study, we implemented the Shell scripting-based data modeling approach by developing a data analysis command set to achieve more rapid and flexible data analysis.

← Back

ANALYSIS TOOL FOR ANNOTATED VARIANTS – A COMPREHENSIVE PLATFORM FOR POPULATION-SCALE GENOMIC ANALYSES

Zhong Ren[1], Petrovski Slave[1], Cirulli T Elizabeth[2], Quanli Wang[1], Copeland Brett[1], Bridgers Joshua[1], Gussow Ayal[1], Erin Heinzen-Cox[1], Andrew Allen[2], David B Goldstein[1], Sitharthan Kamalakaran[1]

[1]Columbia University, Institute for Genomic Medicine, New York, NY, [2]Duke University, Durham, NC

Diagnostic and cohort sequencing studies benefit from large samples of similarly processed control data. Analysis on such data involves laborious processing to create a single joint-genotyped VCF and necessitates complex data wrangling and computing from diverse storage environments. Furthermore, parameters used for any downstream analyses are complex making the analyses difficult to reproduce.

We have developed Analysis Tool for Annotated Variants (ATAV) to allow analysts with minimal programming experience to executive complex genomic analyses by abstracting data processing, computing and statistical analyses at the application level. The ATAV platform written in Java and built on a MySQL database (AnnoDB), which hosts sample sequencing data (variant calls, qualities and read coverage across all sequenced bases) and external data (ExAC, GERP, SNPEff annotations). AnnoDB hosts data from 25K exomes/1.5K whole genomes and contains 30 billion variant calls from 125 million genome positions.

The ATAV command line tool is the interface to AnnoDB and has three modules. (i) The query engine translates user’s sample list (in ped format) and analysis parameters into an efficient SQL query for AnnoDB (ii) A runtime variant object creator parses SQL output into a a “variant” object collection and stores all variant-level information. (iii) The stat module performs analysis at variant or region level, and all statistical functions iterate over the variant object. ATAV currently supports tests for identifying putative de novo and inherited compound-het genotypes in trios, performing region-level rare-variant collapsing analyses for identifying genes associated with phenotype (binary, discrete or quantitative).

The modularized ATAV framework using “variant object” abstraction allows us to continuously develop new functions operating on variant objects without the need for understanding data provenance. We have incorporated analytical tools for site-level read coverage comparisons, principal component analysis using ancestry informative markers, to estimate genomic inflation with Q-Q plots, identify parental mosaic transmissions, and many others.

All ATAV analyses produce an auditable log of software version and all parameters to ensure reproducibility. We have used ATAV to discover novel genes involved in epilepsy (PMID: 23934111, 25262651), ALS (25700176), AHC (22842232) and in diagnostic interpretation of over 400 familial trios (25590979, 27148561, 26138499). ATAV source code is freely available at: https://redmine.igm.cumc.columbia.edu/projects/atav/wiki

FANTOM5 SSTAR: A CHALLENGE TO ESTABLISH A CONTINUOUSLY EXTENSIBLE DATABASE FOR A LARGE SCALE DATA PRODUCTION PROJECT

Imad Abugessaisa[1], Hisashi Shimoji[1], Serkan Sahin[1], Atsushi Kondo[1], Jayson Harshbarger[1], Marina Lizio[1], Yoshihide Hayashizaki[2], Piero Carninci[1], Alistair Forrest[1,3], the FANTOM5 Consortium[1], Takeya Kasukawa[1], Hideya Kawaji[1,2,4]

[1]RIKEN, Center for Life Science Technologies, Yokohama, Japan, [2]RIKEN, Preventive Medicine and Diagnosis Innovation Program, Wako, Japan, [3]the University of Western Australia, QEII Medical Centre and Centre for Medical Research, Nedlands, Australia, [4]RIKEN, Advanced Center for Computing and Communication, Yokohama, Japan

We have been organizing FANTOM (Functional Annotation of Mammalian Genome) projects since 2000, which is an international collaborative project focusing mainly on elucidating functions of transcriptome encoded in the mammalian genomes (http://fantom.gsc.riken.jp/). In the recent 5th term of the FANTOM project (FANTOM5), we analyzed transcriptional start sites and their activities in ~1,800 human and ~1,000 mouse samples including primary cells, cancer cell lines, tissues and time course series in transition between different cell types, and obtained large-scale promoter and enhancer atlas in mammals. The dataset contains various kind of information like coordination and annotation of promoter/enhancer regions, their further bioinformatics analysis results, and metadata of samples and sequencing.

In handling these large-scale datasets and developing a database system to handle them during the collaborative studies, we faced several challenges. One challenge is adaptation to new types of data being generated as research grows. New data, which is often obtained from the novel ideas during collaborative research activities, do not always fit to the current data models, and its adaptation often requires changes of database schemas. Another one is that visual representation has to be designed for manual inspection of the growing data. These requires frequent incremental changes both in schema and interface.

Here, we show a web-based database named SSTAR (Semantic catalog of Samples, Transcription initiation And Regulators) as a platform to deliver FANTOM5 sample information and analysis results to the research community. The system is based on Semantic MediaWiki (SMW) platform. SMW is an extension to MediaWiki to store additional semantic data (termed semantic properties) along with wiki content. SMW has functionality to query on semantic properties, termed semantic query, and the queries can be embedded in wiki content. We utilized semantic properties and queries for adopting new data without destructive schematic change on existing data by implement associations between properties by properties and in-line queries. We also added several visual extensions for genomic view, quantitative value display, and biological ontology classes. The database is publicly accessible at http://fantom.gsc.riken.jp/5/. SSTAR can be a new approach to efficiently handle large-scale research data

Currently, we started a new 6th term of FANTOM project now focusing on systematically elucidating functions of long non-coding RNA by adopting various sequencing technologies. We also would like to discuss about data management for the new project including systematic collection of meta data, and efficient management of databases.

← Back

HI-C DATA COMPRESSION

Minji Kim, Olgica Milenkovic

University of Illinois at Urbana-Champaign, Electrical and Computer Engineering Department, Urbana, IL

Recent years have witnessed the development of a myriad of new methods for detecting physical interactions between genomic regions based on high throughout sequencing technologies: examples include the chromosome conformation capture (3C) method, and its extension, the Hi-C method. Emerging 3C and Hi-C datasets may enable a paradigm shift in our understanding and modeling of three-dimensional genome structures, their time dynamics, and role in transcriptional regulation.

Hi-C probes can currently achieve resolutions as small as 1 kilobase, thereby measuring contact frequencies between genomic regions with very high resolution. Consequently, one faces the challenge of processing, transferring and storing large contact maps, which are usually of the form of 2-dimensional matrices containing integer-valued contact frequencies. For example, a Human 5 kb resolution genome-wide map can result in as much as 720 GB of data. To address this emerging storage problem, we have developed a specialized Hi-C data compression algorithm. Our scheme exploits the unique structure of contact maps: proximal regions are likely to interact, while distant locations are less likely to be in proximity of each other. This leads to dense block substructures placed along the diagonal, and sparse sub matrices appearing at off-diagonal entries. Furthermore, the contact matrices are nearly-symmetric. We hence first cluster the entries of the upper triangular matrix into dense and sparse parts using new community detection paradigms. We subsequently read the dense components using specialized space-filling curves and implemented suitable encoding methods for the obtained 1-dimensional strings such as Huffman and Golomb encoding. The sparse block components were treated differently: we encoded only the non-zero values of the sub matrices using compressed sparse row encoding.

On a subset of mouse stem cell data GSE35156, our algorithm compressed the contact map into merely 10% of the original file size in a few minutes. Likewise, the decompression algorithm losslessly recovered the original file in minutes. Similar results were observed for a number of other datasets.

← Back

pVAC-SEQ: DEVELOPMENT OF AN OPEN-SOURCE SOFTWARE PIPELINE FOR TUMOR NEOANTIGEN DISCOVERY AND PERSONALIZED IMMUNOTHERAPY

Susanna Kiwala[1], Jasreet Hundal[1], Aaron Graubert[1], Joshua McMichael[1], Adam Coffman[1], Jason Walker[1], Christopher A Miller[1], Obi L Griffith [1,2,4], Elaine R Mardis[1,3,4,5], Malachi Griffith[1,4]

[1]Washington Univ. School of Medicine, McDonnell Genome Institute, St. Louis, MO, [2]Washington Univ. School of Medicine, Department of Medicine, Division of Oncology, St. Louis, MO, [3]Washington Univ. School of Medicine, Department of Medicine, Division of Genomics and Bioinformatics, St. Louis, MO, [4]Washington Univ. School of Medicine, Department of Genetics , St. Louis, MO, [5]Washington Univ. School of Medicine, Department of Molecular Microbiology, St. Louis, MO

*These authors contributed equally

We have developed an improved, in silico based sequence analysis method for identification and refinement of personalized variant antigens by cancer sequencing (pVAC-Seq). This flexible and streamlined computational workflow integrates tumor mutation and expression data (DNA- and RNA-Seq) to shortlist candidate neoantigen peptides for a personalized vaccine.

Cancer immunotherapy enables a programmed immune response to attack cancerous cells that are biologically flagged by tumor-specific mutant antigens (‘TSMAs’ or neoantigens). TSMAs arise from numerous genetic changes, acquired somatically that are present exclusively in tumor and not in normal cells. The current regimen for predicting and screening neoantigens from sequencing data is laborious and involves a large number of intermediate steps to identify potentially high-quality neoantigens. Our open-source software, pVAC-Seq, provides higher efficiency and faster turnaround by rapidly streamlining the screening and identification of a smaller number of potentially immunogenic neoepitopes within the landscape of all neoepitopes, thereby increasing its applicability for clinical use.

pVAC-Seq has evolved from a collection of standalone, independent commands to a managed and packaged workflow of sub-commands with a unified interface. Software design best practices have been employed to improve both the quality and agility of the software development. Design decisions for this development project focused on user accessibility and ease of installation. New features since the original publication include VCF and VEP support, enabling predictions for indels, supporting multiple prediction algorithms as well as integrated addition of read-counts and expression estimates. Future development goals include neoantigen predictions for fusions, support for MHC class II binding predictions, as well as an interactive user-interface for visualizations and reporting of filtered candidates. pVAC-Seq is available at https://github.com/griffithlab/pVAC-Seq.

← Back

DOLPHIN: A GRAPHICAL USER INTERFACE FOR THE ANALYSIS AND PROCESSING OF HIGH THROUGHPUT GENOMICS

Alper Kucukural[1,2], Nicholas Merowsky[1], Manuel Garber[1,2,3]

[1]University of Massachusetts Medical School, Bioinformatics Core, Worcester, MA, [2]University of Massachusetts Medical School, Program in Molecular Medicine, Worcester, MA, [3]University of Massachusetts Medical School, Bioinformatics and Integrative Biology, Worcester, MA

High throughput sequencing methods are now routinely being applied to discover key differences between cell types and/or conditions. Availability of high throughput methods together with inexpensive sequencing costs has enabled complex experimental designs that include tens and even hundreds of different conditions and replicates. The bottleneck now is the processing and analysis of this ever increasing data. Sequence data processing usually involves multiple programs to perform analysis, e.g. quality filtering, read alignment, gene quantification and differential gene expression.

Existing programs are not designed to process data from “end to end” and take raw input to usable results; instead they are designed and optimized for specific steps in the process. Approaches such as Galaxy, GenePattern and GeneProf attempt to solve this problem by allowing users to build “pipelines” that string specialized programs into end-to-end processes that take raw data into a form that is suitable for analysis. Current solutions were designed when sequencing throughput was lower and users had only a handful of samples to process. As a result they are designed to handle a single sample at a time and make no effort to keep experimental details (i.e. metadata). Consequently, they are not well suited to handle the large datasets that are now commonplace.

To address these issues we have created Dolphin, a parallel platform designed to process raw sequence data with the specific goal of handling large datasets. Dolphin keeps metadata information about the experimental conditions and provides an integrated processing and analysis platform. It allows users with limited bioinformatics experience to analyze large numbers of samples on High Performance Computing (HPC) systems through a user-friendly web interface. This interface allows searching, viewing metadata and controlling high-throughput analysis pipeline (re)execution. Coupled with DEBrowser, an interactive differential expression (DE) analysis toolbox, dolphin can do differential expression analysis. DEBrowser is a R based tool to address the most important features of gene expression analysis. DEBrowser;

a) groups samples and tests differentially expressed genes with various DE methods between the defined groups and multiple group comparisons in a single table b) with All-to-All scatter, heatmaps and Principal Component Analysis (PCA), outliers and batch effects can be detected c) can be used for visualization of up or down regulated genes using interactive scatter, MA, and volcano plots d) has interactive gene ontology and pathway analysis.

ANALYSIS AND SYNTHESIS OF GENE REGULATORY NETWORKS VIA FORMAL REASONING: MAIN APPLICATIONS AND FUTURE CHALLENGES

Hillel Kugler[1,2]

[1]Bar-Ilan University, Faculty of Engineering, Ramat-Gan, Israel, [2]Microsoft Research, Biological Computation, Cambridge, United Kingdom

Formal reasoning methods make use of mathematical and algorithmic techniques to prove the correctness of a model with respect to properties specified in an appropriate type of logic. Significant progress in the scalability of formal reasoning methods and tools over the past decades enabled the successful application of these methodologies in the hardware and software industries. We discuss the application of formal reasoning methods to analyzing and synthesizing Gene Regulatory Networks (GRNs), focusing on developmental and stem cell systems [1,2,3,4]. The approach allows to make new predictions of complex mutant phenotypes and uncover potential new interactions and network components. Despite recent success stories, significant limitations of the current state-of-the-art methods still prevent wider applications of the reasoning methods. We discuss the main limitations, including dealing with stochastic behavior, noisy data sets, single-cell data and tissue-level phenotypes, and outline some ideas and research strategies that may be explored to extend the utility of the approach and enable more mainstream application for experimental biologists.

[1] S.J. Dunn et al. Defining an essential transcription factor program for naive pluripotency. Science 2014.

[2] I. S. Peter et al. Predictive computation of genomic logic processing functions in embryonic development. Proc. of the National Academy of Sciences, 2012.

[3] Y. Shavit et al. Automated synthesis and analysis of switching gene regulatory networks. BioSystems, In Press., 2016.

[4] B. Yordanov et al. A method to identify and analyze biological programs through automated reasoning. npj Systems Bioilogy and Applications, 2016.

GTRACKS: A FRAMEWORK FOR CREATING AND MAINTAINING UCSC TRACK DATABASES USING GOOGLE SPREADSHEETS.

Kathleen E Kyle[1], Hank W Bass[2], Daniel L Vera[1]

[1]Florida State University, Center for Genomics and Personalized Medicine, Tallahassee, FL, [2]Florida State University, Biological Science, Tallahassee, FL

Visualization is a critical process for the assessment and interpretation of genomic data. The UCSC Genome Browser is a powerful web-based platform that offers a rapid, intuitive, and centralized means of viewing and sharing various types of genomic data mapped to a reference genome. The browser also allows users to maintain large catalogs of private data tracks in "track hubs". Despite its great utility, the browser hosted at UCSC is largely limited to vertebrate genomes. However, "assembly hubs" are track hubs that allow the use of user-provided genomes. In addition, users with large numbers of custom genomes may benefit from the installation of a self-hosted instance of the browser. Track and assembly hubs require the use of configuration files that are difficult to setup and maintain when large amounts of custom datasets are used. Installing a self-hosted instance of the browser is even more challenging, requiring extensive use of the unix command line and knowledge in system administration. To address these challenges, we have developed user-friendly tools to easily and quickly setup and maintain track hubs, assembly hubs, and genome browser installations. The need to use complex configuration files and databases are replaced by easily-maintainable google spreadsheets. A series of scripts and dockerfiles facilitate the rapid and automated deployment of genome browser installations and loading of custom genomes and datasets. As the cost and time required to generate sequencing data continues to decrease, the need for tools to setup and maintain large numbers of genomic data tracks grows. These tools can be used to quickly and easily deploy track hubs, assembly hubs, and browser installations, enabling any user to utilize the more advanced functions of the genome browser for exploring their own genomes and data sets.

COMPARISON OF PROTEIN SIMILARITY NETWORK TOPOLOGIES USING SEQUENCE- AND ACTIVE SITE-SIMILARITY EDGE METRICS

Janelle B Leuthaeuser, Julia Hayden, Nicholas Biffis, Jacquelyn S Fetrow

University of Richmond, Chemistry Department, Richmond, VA

The elucidation of protein molecular function lags far behind the rate of high-throughput sequencing technology; thus, it is essential to develop accurate and efficient computational methods to define functional relationships. Clustering using sequence similarity networks has emerged as a simple, high-throughput method for defining relationships between proteins, but it is difficult to identify a single score threshold that identifies protein clusters which correlate with molecular function without over-clustering some functional groups while simultaneously under-clustering others. Previous work analyzing structurally characterized proteins has shown that similarity networks using active site similarity as the edge metric may better balance the over- and under-clustering, but identifying a single score threshold to extract isofunctional clusters remains a challenge. To identify isofunctional clusters of protein structures accurately and efficiently without relying on a single score threshold, we developed an iterative clustering process, TuLIP. An active site profile was created for each TuLIP-identified cluster and these profiles were input for searches of GenBank using the search tool DASP, which identifies protein sequences containing active site features similar to those in the query profile. Network topologies were evaluated using two different similarity networks constructed for the functionally related sequences identified in each DASP search. In the first, sequence similarity scores were used as the edge metric; in the second, active site similarity scores were calculated as the edge metric. Analysis of network topology demonstrates the difficulty in identifying a single score threshold to define isofunctional groups, regardless of the edge metric used. Additionally, visualizing TuLIP-identified isofunctional proteins within sequence similarity networks highlights areas where evolutionary sequence-based relationships differ from functional relationships. Ultimately, this work suggests how similarity networks can be used to evaluate relationships among proteins, as well as the limitations in defining functional relationships with such networks.

← Back

OPTIMIZED CONTIGS ASSEMBLY AND SNP DISCOVERY BY OVERLAPPING PAIRED-END RAD SEQUENCING IN ROUGHSKIN SCULPIN (*TRACHIDERMUS FASCIATUS* HECKEL)

Yulong Li, Dongxiu Xue, Jinxian Liu

Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China

Next generation sequencing technology based on restriction site associated DNA sequencing is revolutionizing studies in ecological, evolutionary and conservation genomics. However, the assembly for paired-end RAD data is still challenging, especially for non-model species with high genetic variations. We present here an optimized approach of assembly by using overlapping paired-end RAD sequencing to generate high quality contigs (~600bp) for downstream genotyping and annotations. Roughskin sculpin is selected for RAD reference assembly due to the high and significant genetic divergence among populations. Both bioinformatics and experimental verifications suggest the high quality of the assembled contigs. RAD contigs provide sufficient flanking sequences for gene annotation and primers design. Based on the RAD contigs, a total of 43,707 SNPs are generated using both the paired reads after filtering. The F_{ST} value (0.0705, $P < 0.001$) and cluster analyses suggest a high and significant genetic divergence between the two populations. In conclusion, the generated SNPs and contigs should be a valuable genome resource for future population genetics studies, and the approach of assembly should provide an optional tool for dealing with RAD assembly complexity for non-model species.

THE TCGA DATA ANALYSIS FOR THE FUNCTION AND REGULATION OF THE PROTEIN TYROSINE PHOSPHATASE SUPERFAMILY IN HUMAN CANCERS

Suh-Yuen Liang[1], Tzu-Hao Chang[2], Tsung-Hsien Pu[1], Chia-Cheng Chou[3], Tzu-Ching Meng[1]

[1]Academia Sinica, Institute of Biological Chemistry, Taipei, Taiwan, [2]Taipei Medical University Hospital, Graduate Institute of Biomedical Informatics, Taipei, Taiwan, [3]National Center for High-performance Computing, National Applied Research Laboratories, Hsinchu, Taiwan

Accumulated evidences have indicated that human protein tyrosine phosphatases (PTP) are involved in tumorigenesis and hold promise for cancer therapeutic targets. This study aims to identify the prognostic PTPs for overall survival and tumor progression, further elucidating the regulatory mechanisms which command the expression of such PTPs based on pan-cancer analysis of the Cancer Genomic Atlas (TCGA) deep sequencing genomic data.

We analyzed the expression of 102 members in PTP superfamily in relation to the overall survival and between normal to various tumor stages. The cancer prognostic PTPs were subjected to further analysis with IPA (Ingenuity® Pathway Analysis) for functional enrichment and pathfinding of upstream regulators. Pearson's correlation analysis was applied to elucidate potential regulation of gene expression by methylation, microRNA, and copy number variations.

The top ten prognostic PTPs with oncogenic effect were found, including four cell cycle regulators CDKN3, CDC25A, CDC25B, and CDC25C; two MAPK signaling regulator, DUSP4 and DUSP12; two non-receptor PTPs PTPN2 and PTPN12; and two PTPs whose function in cancers was mostly uncharacterized PTPDC1, and STYXL1. The top ten prognostic PTPs with tumor suppressive effect include four phosphoinositide phosphatases, PTEN, MTMR10, MTM1, and INPP5A; two PTPs with Src homology 2 (SH2) domain TNS1 and TENC1; two MAPK signaling regulator, DUSP19 and DUSP26; one membrane receptor-like PTP, PTPRT; and EPM2A which is associated with glucose metabolism.

The IPA pathfinding analysis has assigned multiple upstream regulators, particularly for oncogenic CDKN3, CDC25A, CDC25B, CDC25C, DUSP4, and PTPN12. Interestingly, the upstream regulators were consistent in line with the corresponding oncogenic PTP among affected cancers. The results from Pearson's correlation analysis identified a number of microRNAs which are mostly negative regulators of PTP expression. The DNA methylation, however, caused either negative or positive effect on expression. In addition, some PTPs were found to be highly regulated by variations of gene copy number. Together with the prognostic PTP profile and the results of regulatory mechanisms in control of their expression, an online PTPome database was built up to facilitate the endeavor of the scientific community to find new prognostic markers or therapeutic targets in human cancers.

← Back

GENOME-WIDE SNP DISCOVERY AND PRELIMINARY POPULATION GENETIC ANALYSIS IN JAPANESE EEL (*ANGUILLA JAPONICA*) BASED ON RAD SEQUENCING

Bingjian Liu

Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China

Reduced representation genome sequencing such as restriction-site-associated DNA (RAD) sequencing is finding increased use to identify and genotype large numbers of single-nucleotide polymorphisms (SNPs) in model and nonmodel species. Here, we generated a resource of novel SNP markers for the Japanese eel (Anguilla japonica) using the RAD sequencing approach from 24 individuals. The identified 73,557 SNPs were widely distributed across the eel genome, aligning to 28,407 different contigs. No differentiation between the two populations was detected based on all SNPs or neutral loci. However, results were highly differentiated based on SNP data set of outlier SNPs. Moreover, signature of local adaptation was highlighted by F_{ST}-based outlier tests implemented in ARLEQUIN and a total of 250 potentially locally selected SNPs were identified. BLAST2GO annotation of contigs containing the outlier SNPs yielded hits for 61 (72%) of 85 significant BLASTX matches. The KEGG pathway approach implemented using the tool DAVID showed that some of the putative targets of local selection including genes in several important pathways, such as calcium signaling pathway, Intestinal immune network for IgA production, Cytokine-cytokine receptor interaction and MAPK signaling pathway. The generated SNP resource provides a valuable tool for future population genetics and genomics studies and allows for targeting specific genes and particularly interesting regions of the eel genome.

← Back

SIMULTANEOUS DETECTION OF SNPs AND INDELS USING A 16-GENOTYPE PROBABILISTIC MODEL

Ruibang Luo[1,2], Steven L Salzberg[1,2], Michael C Schatz[1,3]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Johns Hopkins University School of Medicine, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [3]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, New York, NY

Single nucleotide polymorphisms (SNPs) and indels that occur at a genome locus are interdependent; i.e., evidence that elevates the probability of one variant type should decrease the probability of other possible variant types, and the probability of all possible alleles should sum up to 1. However, widely-used tools such as GATK's UnifiedGenotyper[1] and SAMtools[2] use separate models for SNP and indel detection. The model for SNP calling in these two tools is nearly identical: both assume all variants are biallelic and use a probabilistic model of 10 genotypes (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT). For indel calling, the GATK UnifiedGenotyper uses a model initially derived from Dindel's model[3], while Samtools' model is derived from BAQ[4].

In order to merge the detection of SNPs and indels, we propose a new 16-genotype probabilistic model. Using X and Y to denote the indels with the highest (X) and second highest (Y) supports, we add 6 new genotypes AX, CX, GX, TX, XX, XY to the traditional 10-genotype probabilistic model to accommodate four possibilities: 1) one homozygous indel (XX); 2) one reference allele plus one heterozygous indel (AX, CX, GX, TX); 3) one heterozygous SNP plus one heterozygous indel (AX, CX, GX, TX); and 4) two heterozygous indels (XY). We exclude the 5 possible combinations AY, CY, GY, TY, YY because they are illegitimate. By unifying SNP and indel calling in a single variant calling algorithm, the new model not only consumes less computational resources, but also demonstrates improved sensitivity and accuracy for variant detection. The new model also improves the sensitivity on detecting SNPs and indels for somatic mutations found in paired normal-tumor samples, where the sequences include complex mixtures with more than two haplotypes.

1. McKenna, Aaron, et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research 20.9 (2010): 1297-1303.
2. Li, Heng, et al. "The sequence alignment/map format and SAMtools." Bioinformatics 25.16 (2009): 2078-2079.
3. C.A. Albers et al. (2011). Dindel: Accurate indel calls from short-read data. Genome Res. 2011 Jun;21(6):961-73.
4. Li, Heng. "Improving SNP discovery by base alignment quality." Bioinformatics 27.8 (2011): 1157-1158.

← Back

THE CELLFINDER ON-LINE DATA RESOURCE AND ITS APPLICATIONS FOR STEM CELL RESEARCH

Nancy Mah[1], Khadija El Amrani[1], Fritz Lekschas[1], Stefanie Seltmann[1], Marie Bittner[1], Harald Stachelscheid[1], Miguel Andrade-Navarro[2], Andreas Kurtz[1]

[1]Charite University Medicine Berlin, Berlin-Brandenburg Center for Regenerative Therapies, Berlin, Germany, [2]Johannes Gutenberg University of Mainz, Institute of Molecular Biology, Mainz, Germany

CellFinder is a freely available on-line resource that aims to simplify access to diverse kinds of data associated with in vivo cells and in vitro cell lines. CellFinder has three main entry points for queries: the Semantic Body Browser, developmental tree or a simple text search. Use of the CELDA ontology provides a framework to organize relations between ontology terms (e.g. cell types, developmental trees) and data associated to the cell types, such as molecular data or images. Molecular data is manually curated with ontology terms, pre-processed for large-scale analysis, and stored in a PostgreSQL database. We also develop analysis methods with the aim to provide these methods as user-friendly web-based tools in CellFinder, using the expertly curated molecular database. Our suite of analysis tools currently includes a tool to generate differentially regulated genes from a stem cell dataset (CompareTool) and to generate tissue-specific markers from normal tissues (MarkerTool). A method to evaluate cell identity from transcriptome profiles will be presented (CellScore), for future implementation in the CellFinder framework. CellFinder is available at: http://www.cellfinder.org/.

GRASS: GRAPH REGULARIZED ANNOTATION VIA SEMI-SUPERVISED LEARNING

Laraib I Malik, Rob Patro

Student, Computer Science, Stony Brook, NY, [2]Assistant Professor, Computer Science, Stony Brook, NY

De novo transcriptome assembly is the first major step of many RNA-seq analyses in non-model organisms. Though we can run typical RNA-seq pipelines on such data (e.g. quantify transcript abundances and assess differential expression), to meaningfully interpret the results, we must have some notion of the transcripts and genes that our assembled contigs represent. Thus, accurately annotating the contigs of a *de novo* assembly using information from related organisms is an important step in *de novo* transcriptome analysis.

We present a method, GRASS, that employs annotation transfer from previously annotated species, in conjunction with a graph representing contig-level similarity in the *de novo* assembly, to improve annotation quality. We consider a weighted, contig-level similarity graph, G, in which contigs are vertices, and where each pair of contigs is connected by an edge if RNA-seq reads multimap between them. The weight of each edge is proportional to the fraction of fragments multimapping between the adjacent contigs. This graph can be efficiently built from previously generated fragment equivalence classes.

We initially label the nodes of G (i.e. the contigs of our assembly) using a traditional approach, such as a BLAST search. Subsequently, a semi-supervised learning method, label propagation, is used to transfer these annotations to unannotated nodes in accordance with the topology of the graph. We build an iterative algorithm atop label propagation, which alternates between probabilistically labeling the nodes via label propagation, and modifying the topology and edge weights of the graph to conform to the current labeling. This process is repeated until the graph topology converges. The final result of our approach is a collection of annotations for the contigs in the *de novo* assembly.

We compare our results against a number of tools used for integrating information from annotated species with *de novo* assemblies. We show that GRASS significantly increases the number of correctly-annotated contigs as compared to existing tools. We also demonstrate that incorporating the annotation information improves the quality of contig clustering, allowing for improved "gene-level" differential analysis. In addition to considerably improving annotation accuracy, GRASS is very fast, taking only a few minutes to run on typical *de novo* assemblies.

← Back

IMPROVED REPEAT ALIGNMENT BY SIMULTANEOUS ESTIMATION OF VARIATION AND ALIGNMENT.

Wilson H McKerrow, Charles E Lawrence

Brown University, Applied Mathematics, Providence, RI

High-throughput short-read sequencing provides large data sets, but alignment of short reads to repetitive sequence is often inaccurate. As repetitive elements tend to be longer than single reads, reads are likely to have multiple potential alignments. Inference in repetitive sequence can be improved by building a probabilistic model of read alignment. In such a model values inferred from the alignment can pass information back, allowing for an updated, more accurate alignment.

I will focus on the inference of RNA hyper editing in transposable elements. RNA hyper editing occurs when the enzyme ADAR converts many of the adenosines in a dsRNA target into inosines – molecules that are recognized as guanine by cellular machinery. Certainty about an alignment increases when a read overlaps more informative positions – nucleotides that are specific to the repeat. As we learn about editing, the A to G changes provide new informative positions that we can use to refine the alignment.

I will describe how the expectation maximization algorithm can be used to provide accurate repeat alignments by simultaneously estimating alignment and variation. 94% of reads were aligned correctly in a simulated RNA editing experiment. The aligner, bwa correctly aligns only 63%.

← Back

RAPID, ACCURATE, AND ACTIONABLE IDENTIFICATION OF PATHOGENIC AND COMMENSAL MICROORGANISMS, ANTIBIOTIC RESISTANCE PROFILES, AND VIRULENCE FACTORS FROM METAGENOMIC SAMPLES

Kelly Moffat[1], Nur A Hasan[1,2], Poorani Subramanian[1], Huai Li[1], Christopher E Mason[4], Rita R Colwell[1,2,3]

[1]CosmosID, Inc, Metagenomics, Rockville, MD, [2]University of Maryland, Center for Bioinformatics and Computational Biology, College Park, MD, [3]Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD, [4]Weill Cornell, Medical College, New York, NY

Versatility, reduced cost, and faster turnaround time of sequencing suggest high-throughput and high-resolution metagenomic analysis can be used in clinical management of disease, pathogen and commensal identification, antibiotic resistance profiling, and for characterization of microbiota in many environments. While 16S rRNA analysis has been widely used for microbial identification in a metagenomic sample, significantly more information can be obtained from whole genome shotgun sequencing analysis, including strain level identification, antimicrobial resistance gene and virulence factor profiles, characterization of non-bacterial organisms such as viruses, parasites and fungi, and functional analysis of the microbes present in the sample. CosmosID has used its high-resolution metagenomic platform for the analysis of >20,000 samples, including studies to identify polymicrobial infections, pathogenic agents in heart valve infections, and for characterization of nosocomial reservoirs in sink trap biofilms. Also in the clinical setting, a retrospective case-control study of 65 fecal samples was done that included patients of known and unknown disease etiology as well as healthy individuals. Samples were collected during a systematic surveillance at the National Institute of Cholera and Enteric Diseases, Kolkata, India and analyzed by CosmosID. The analysis showed that the intestinal microbiome could differentiate healthy, diseased, asymptomatic carriers, and individuals in early stages of disease. Lastly, metagenomic exploration of the built environment has been of increasing interest globally as recognized in the MetaSUB consortium. CosmosID performed strain-level metagenomic analyis and antibiotic resistance profiling of 1,572 MetaSUB samples from New York City to help in the exploration of microbial diversity and identification and surveillance of antimicrobial resistance markers in the built environment.

← Back

DEPLOYMENT OF A BIOINFORMATICS ANALYTICS PLATFORM IN THE CLOUD

David Molik[1,2], Yu-Jui Ho[2], Ying Jin[1,2], Molly Hammell[2]

[1]Cold Spring Harbor Lab, Bioinformatics Shared Resource, Cold Spring Harbor, NY, [2]Cold Spring Harbor Lab, Molly Hammell Lab, Cold Spring Harbor, NY

Deployment of web-based bioinformatics applications requires physical or virtual web servers that must accomplish several tasks including: handling web traffic, data upload/download, backend computations, and return of results. For deployment of new informatics applications, it can be difficult to predict the number of users that will be accessing these servers and the size of the data files users submit. For this reason, even small bioinformatics software labs might be faced with decisions about the best type of computational infrastructure to support their needs for web demos or full web servers. Typically, this involves either purchasing dedicated physical or virtual web servers or purchasing compute time from commercial cloud providers. These each have benefits; physical servers can be cheaper to run in the long term, as long as the server is carefully calibrated to user needs, but require constant maintenance and cannot be quickly scaled if a web server attains considerably more traffic or much larger data file sizes than estimated. Deploying in the cloud may be able to maximize resources dedicated to a bioinformatics analytics platform, especially if the deployment can scale larger or smaller according to the amount of users, data file sizes, or the performance of the server.

In order to test the deployment of bioinformatics analytics platforms in the cloud, Yabi[1], a web-based bioinformatics analysis platform, was installed onto Amazon Web Services (AWS). Yabi can be heavily customized to host a variety of bioinformatics software tools, much like the Galaxy analysis platform[2]. However, unlike Galaxy, it does not require centralized data storage. To provide high availability, two redundant AWS web servers were allocated. Both of them receive traffic and can update each other; they share the same database, and backend services for redundant handling of web traffic. The deployment scales the number of compute servers according to the performance of each currently running server for dynamic allocation of resources. AWS provides a layer of abstraction that makes development and coding updates to the core functions of Yabi easier, greatly reducing the need for redundant development servers. Additionally, computational servers are created in the cloud for each application and then quickly deleted after analysis completes, providing rapid and efficient control of resources. Here we present a review of best practices for reliable and resource-conscious deployment of bioinformatics platforms on a commercial cloud server.

[1]. Hunter, Adam A., Andrew B. Macgregor, Tamas O. Szabo, Crispin A. Wellington, and Matthew I. Bellgard. "Yabi: An online research environment for grid, high performance and cloud computing." Source code for biology and medicine 7, no. 1 (2012): 1.

[2]. Afgan, Enis, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update." Nucleic acids research (2016): gkw343.

← Back

CONSEQ: A METADATA-DRIVEN TOOL FOR EXECUTING BIOINFORMATIC PIPELINES

Philip G Montgomery

Broad Institute, Cancer Program Data Science, Cambridge, MA

Bioinformatics workflows often compose multiple tools into a pipeline, and many tools exist for encoding a pipeline in terms of input and output files. However, filenames are a limited mechanism to organize heterogenous collections of results.

Conseq takes a metadata-centric approach by maintaining a schemaless database of all results and their metadata, which is updated as a pipeline runs. Pipelines are modeled as a query to execute and an associated process to execute on that query's results. After processes successfully complete, their results are published back to the database, triggering any downstream processing which has been waiting for those results. By specifying the processes to execute using a lightweight template syntax, Conseq generates scripts using the metadata from results, making it trivial to parameterize complex workflows. Lastly, by allowing for parallel local execution, as well as parallel submission to remote hosts and submission to batch queues such as OGE, Conseq provides a central point from which a complex workflow can be orchestrated, results collected, and provenance recorded.

Using Conseq, we are able to coordinate preprocessing steps for our production pipelines as well as automatically execute downstream analyses as new data is arrives. The database of results can then be published for collaborators who can browse the provenance of all data as well as associated metadata and the final result files, streamlining data handoff and communication.

← Back

NOVEL DISCOVERY FROM A GENE BY MEDICAL PHENOME CATALOG

Nancy J Cox

Vanderbilt University Medical Center, Medicine, Nashville, TN

Large-scale biobanks permit new kinds of investigations into how genetic variation impacts human phenotypes. Rather than considering a single disease entity with a few related quantitative phenotypes, with data from biobanks that link to electronic health records (EHR), it is possible to discover how genome variation relates to the entire medical phenome simultaneously. We have used an approach that integrates functional genomics information in the form of mRNA transcript levels measured through RNA-Seq with genome variation. Using RNA-Seq data and genome variation data from the Genotype Tissue Expression (GTEx) project, we use predicted expression scanning, or PrediXcan (Gamazon et al., 2015), to build SNP-based predictors of gene expression using the elastic net for each gene expressed at high enough levels and with sufficient interindividual variability to have the possibility of prediction. We then apply these predictors to genotype data obtained from sample in BioVU, the biobank at Vanderbilt University, with DNA samples on more than 218,000 subjects linked to a de-identified and continuously updated image of the EHR at Vanderbilt University Medical Center. By applying PrediXcan to BioVU, we have created the first gene x medical phenome catalog. This is analogous to having conducted a knock-down experiment on each gene in the human genome, and then reading out the consequences of the knock-down experiment across the entire medical phenome, and similarly doing an up-regulation experiment on each gene, and reading out the consequences across the entire medical phenome. Rather than manipulating the expression of human genes, we are using natural variation that influences gene expression to test the association of genetically predicted gene expression with the medical phenome. Among the big-picture results are that genes with coordinate regulation contribute to the polygenic architecture of some common human diseases: there are many genes with measured expression highly correlated within GTEx, and the genetically predicted expression of these genes are also correlated even though we include only local SNPs (within 1 Mb of transcription start/stop) in prediction equations. The largest set of genes with correlated expression accounts for a major axis of phenotypic association across BioVU in which the same sets of phenotypes are observed. Among the genes most strongly associated with phenotypes on this axis include several that are characterized as pivots on the innate immunity / wound healing axis, while others have been implicated in vascular biology, skin formation, and immune processes. This phenotypic axis includes a number of human diseases long characterized as having a strong polygenic architecture.

← Back

EXPERIENCES MIGRATING LARGE SCALE BIOINFORMATICS DATA PIPELINES TO A CLOUD INFRASTRUCTURE

Thomas Eastman, Claire Gu, Steven Herrin, Tulasi Paradarami, Kiran Singh, Andrew Stiles, Neil Thomas, Patrick Yee

23andMe, Engineering, Mountain View, CA

23andMe has collected over 1 million genotypes in our database. At this scale, it has become difficult to operate a traditional high-performance scientific cluster that can handle peak load without maintaining excessive capacity or causing long processing delays. In order to address this, we have begun moving our computationally intensive pipelines and workflows from our cluster to a cloud-based infrastructure. We will present our experiences migrating genotype imputation and detection of IBD segments to this new infrastructure. This will include discussions of overall architecture, workflow management, scalability, fault-tolerance, and the engineering involved.

← Back

REACTOME PATHWAY ANALYSIS: A HIGH-PERFORMANCE IN-MEMORY APPROACH

Antonio Fabregat[1], Konstantinos Sidiropoulos[1], Guilherme Viteri[1], Pablo Marin-Garcia[2], Peter D'Eustachio[3], Lincoln Stein[4,5], Henning Hermjakob[1,6]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, United Kingdom, [2]Fundación Investigación INCLIVA, Universidad de Valencia, Valencia, Spain, [3]NYU Langone Medical Center, New York University, New York, NY, [4]Ontario Institute for Cancer Research, OICR, Toronto, Canada, [5]Cold Spring Harbor Laboratory, Bioinformatics, Cold Spring Harbor, NY, [6]National Center for Protein Sciences, NCPSB, Beijing, China

Reactome (http://www.reactome.org) is a free, open-source, open-data, curated and peer-reviewed knowledge-base of biomolecular pathways. It aims to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. Pathway analysis methods have a broad range of applications in physiological and biomedical research, helping researchers to discover which areas of biology and biomolecules are crucial to understand the phenomena under study. These methods are mainly used to analyse Omics data obtained from high-throughput technologies. One of the main problems is the constantly-increasing size of the data samples.

Reactome offers a set of pathway analysis tools which address this scenario and yet provide reliable and accurate results with interactive (seconds) response time for genome-wide datasets. A high-performance in-memory approach of the well established over-representation analysis method can be achieved by dividing it into four different steps so that specific data structures can be used in each to improve performance and minimise the memory footprint. The first step, determining whether an identifier in the sample corresponds to an entity in Reactome, is addressed using a radix tree as a lookup table. The second step, modeling the proteins, chemicals, their orthologous in other species and their composition in complexes and sets, is covered with a graph. Finally, the third and fourth steps, which aggregate the results and calculate the statistics, utilise a double-linked tree.

Through this highly optimised process, Reactome has achieved a stable, high performance pathway analysis service, enabling the analysis of genome-wide datasets within seconds, allowing interactive exploration and analysis of high throughput data. The proposed pathway analysis approach is available via a web service (http://www.reactome.org/AnalysisService/) for programmatic access or via the data submision interface integrated in the PathwayBrowser (http://www.reactome.org/PathwayBrowser/#/TOOL=AT). The code implementing the described anlaysis method is freely available at https://github.com/reactome/AnalysisTools.

THOUSANDS OF PUBLIC EPIGENOMIC DATASETS CAN BE EXPLOITED THROUGH THE GALAXY-GENAP PROJECT AND THE GEEC TOOL

Jonathan Laperle[1,2], David A Morais[3,4], Michel Barrette[3], Charlotte Bastin[2], Marc-Antoine Robert[2], Marie Harel[2], Alexei N Markovits[1,2] , David Bujold[4], Alain Veilleux[3], Guillaume Bourque[4,5], Pierre-Étienne Jacques[1,2,3]

[1]Universite de Sherbrooke, Informatique, Sherbrooke, Canada, [2]Universite de Sherbrooke, Biologie, Sherbrooke, Canada, [3]Universite de Sherbrooke, Centre de calcul scientifique, Sherbrooke, Canada, [4]McGill University, McGill University and Génome Québec Innovation Center, Montréal, Canada, [5]McGill University, Department of Human Genetics, Montréal, Canada

Public databases such as GEO and ArrayExpress are giving access to raw epigenomic data, while repositories such as the International Human Epigenome Consortium (IHEC) Data Portal (epigenomesportal.ca/ihec) are offering processed datasets for discovery, visualization and download. The Genomic Efficient Correlator (GeEC) tool (Laperle et al., in preparation), aimed at efficiently performing pairwise correlation of thousands of epigenomic datasets, has been used over the last year to pre-calculate static correlation matrices that are incorporated to the IHEC Data Portal. GeEC has proven useful for members of IHEC to demonstrate that despite the non-uniform experimental and processing procedures applied by the different consortia, the IHEC datasets are overall highly comparable as they mainly cluster based on the assay type and sample cell type, rather than on the producing consortium. Moreover, the correlation data were used to identify potentially mislabeled or problematic datasets, as part of a quality control pipeline implemented in the IHEC Data Portal.

Through a public version of GeEC, integrated to the Galaxy framework of the Genetics and genomics Analysis Platform (GenAP) project (galaxy-public.genap.ca), users working with human or yeast data can now compare their own epigenomic datasets to >7000 datasets from IHEC, and to >3500 yeast data downloaded from GEO/ArrayExpress and uniformly processed. This could be useful, for instance, to help in the characterization or validation of private datasets. The various features of GeEC integrated to Galaxy include the support of many genomic file formats (bigWig, WIG, bedGraph, BAM) submitted by the user, the possibility to compute correlations at different resolutions (from 1 Kb to 10 Mb), using different metrics, and on different subsets of regions (e.g. whole genome, only genes, TSS, user-defined), as well as the post-processing of the generated correlation matrices. We also provide a user-friendly interface facilitating the selection of the desired datasets among the thousands available. We plan to also offer processed datasets from model organisms generated by international consortia such as modENCODE, as well as data from organisms other than yeast, downloaded from GEO/ArrayExpress and uniformly processed. We will present the design and implementation of GeEC as well as a performance comparison with other tools and some of the key results obtained so far.

← Back

HOW TO QUANTIFY EXPRESSION FOR THOUSANDS OF RNA-SEQUENCING SAMPLES IN A DAY FOR A MINIMAL COST

PJ Tatlow, Stephen R Piccolo

Brigham Young University, Biology, Provo, UT

The Cancer Genome Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE) are publicly available resources that contain molecular data for thousands of cancer samples. Many research studies have relied on these data. However, some data types in these resources were processed using bioinformatics tools and annotations that have now become obsolete. Fortunately, raw data are available, so the data can be reprocessed using newer tools and annotations. Yet due to the massive size of the data (>60 terabytes), such an effort would require computational resources that exceed what is accessible at most academic computing environments.

Recently, the National Cancer Institute initiated a series of cloud-computing pilots to explore the potential to process cancer data using commercial cloud-computing services. We participated in one of these pilots and used Google Cloud to reprocess 12,307 RNA-Sequencing samples from TCGA and CCLE. To quantify transcript-expression levels, we used kallisto, a popular quantification tool that is much faster than previous-generation tools. We evaluated two distinct configurations: 1) a Kubernetes-based cluster and 2) an approach that used preemptible virtual machines to scavenge resources. As the computations were being performed, we collected detailed performance metrics and cost information. With the preemptible-node configuration, we were able to process RNA-Sequencing samples for as little as $0.09 per sample ($1,065.49 total for 11,373 TCGA samples). Due to the scalability of the cloud infrastructure, this approach is capable of processing thousands of RNA-Sequencing samples in a single day.

As expected, the kallisto's speed helped us to reach this level of performance. However, we also noted that preprocessing steps—converting files between formats, sorting data, etc.—required more processing time than transcript quantification. Accordingly, we recommend that tool developers take these factors into account to obtain further optimizations.

We have created open-source Docker containers that include all the software and scripts necessary to replicate our analysis and to collect detailed performance metrics (see https://osf.io/gqrz9). The processed data are available in tabular format and in Google's BigQuery database. We hope these resources—and our experiences with the Cloud Pilots—will serve as a guide to other scientists who wish to reprocess public data or to process their own data in the cloud.

THE KNOWENG PLATFORM FOR BIOLOGICAL DATA ANALYTICS ON THE CLOUD: A USE CASE IN GENE PRIORITIZATION FOR PHARMACOGENOMICS

Amin Emad[1], Junmei Cairns[2], Liewei Wang[2], Saurabh Sinha[1]

[1]University of Illinois at Urbana-Champaign, BD2K Center of Excellence (KnowEnG), Urbana, IL, [2]Mayo Clinic, Center for Individualized Medicine, Rochester, MN

We are developing the KnowEnG platform to facilitate scalable and ‘knowledge-guided’ analysis of genomics data on a cloud infrastructure. The KnowEnG user will upload their genomics data, e.g., a collection of transcriptomic profiles, in the form of spreadsheets, and perform machine learning and data mining tasks such as classification, regression, feature selection, clustering, and dimensionality reduction, through an easy-to-use interface. An important feature of the analysis algorithms will be the ability to exploit prior knowledge about genes and their relationships (e.g., Gene Ontology and pathway annotations, protein-protein interactions, etc.) available from public domain knowledge-bases.

In addition to introducing this knowledge engine for scalable genomics on the Cloud, we will report on our research into a specific analytic task: the identification and prioritization of genes whose expression levels are predictive of a quantitative phenotype. We developed a computational method, called ProGENI, to identify genes associated with sensitivity to cytotoxic treatments, by leveraging basal gene expression profiles as well as prior knowledge in the form of an experimentally derived network of protein-protein interactions (PPI) and genetic interactions. The method is based on identifying genes whose expression is correlated with drug response, followed by the identification of their neighbors in an interaction network using random walk techniques.

Application of ProGENI to a dataset comprising approximately 300 lymphoblastoid cell lines, and including their responses to 24 cytotoxic treatments as well as their basal gene expression profiles, revealed a significant improvement in predicting drug response over other methods that do not consider network information. A significant improvement was also observed on another dataset from the Genomics of Drug Sensitivity in Cancer (GDSC) database, containing approximately 600 cell lines from 13 tissue types and responses to 139 drugs. In addition, the literature confirmed that the knockdown of many of the genes identified using ProGENI have been shown to affect drug sensitivity. These results suggest ProGENI to be a powerful computational technique in identifying genes that play a key role in determining drug response. A user-friendly cloud-based implementation of ProGENI, allowing its use for gene prioritization for any quantitative phenotype, will be included as part of the KnowEnG platform and will be made available for public use.

← Back

ADVANCING SYSTEMS BIOLOGY USING AN OPEN, EXTENSIBLE AND SCALABLE KBASE PLATFORM

Vivek Kumar[1], Sunita Kumari[1], Srividya Ramakrishnan[1], James Thomason[1], Doreen Ware[1,2], Michael Schatz[1,3], Shinjae Yoo[4], Priya Ranjan[5], Samuel Seaver[6], Nomi Harris[7], Christopher Henry[6], Robert Cottingham[5], Adam Arkin[7]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]USDA-ARS, Ithaca, NY, [3]Johns Hopkins University, Baltimore, MD, [4]Brookhaven National Laboratory, Upton, NY, [5]Oak Ridge National Laboratory, Oak Ridge, TN, [6]Argonne National Laboratory, Argonne, IL, [7]Lawrence Berkeley National Laboratory, Berkeley, CA

The U.S. Department of Energy Systems Biology Knowledgebase (KBase, http://kbase.us) aims to provide a computational environment to meet the key challenges of systems biology: predicting and ultimately designing biological function. KBase distinguishes itself as a knowledgebase that supports the sharing and integration of biological data and any related analysis, modeling and simulation. It is not simply a database or a workbench that serves data or canned analyses.

The KBase hardware infrastructure is a distributed, private cloud with logical groupings of virtual machines that serve production services and workflows as well as active development. In addition, KBase leverages access to DOE high-performance computing (HPC) resources that can support certain large-scale workflows. The production infrastructure is currently distributed between two primary sites at LBNL/National Energy Research Scientific Computing Center (NERSC) and ANL/Magellan to provide redundancy and failover.

KBase has a Jupyter-based, socially-aware user interface that supports a persistent and provenanced environment enabling experimental and computational biologists to work together to share and publish data, approaches, workflows, and conclusions, leading to transparent and reproducible computational experiments that give credit where credit is due.

KBase provides open access to quality-controlled data and high-performance modeling and simulation tools that enable researchers to build new knowledge, interpret missing information necessary for predictive modeling, test hypotheses, design experiments, and share findings such that they can be reproduced and extended by others.

Another distinguishing feature of KBase is its integrated data model that brings together diverse biological datasets and represents them as rich data types that describe relationships among data. This integration enables comparison across biological domains and interoperability with both standard and next-generation tools.

KBase provides a number of apps for large-scale analyses across genome assembly and annotation, expression analysis, metabolic modeling, phylogenetics, comparative genomics and microbial community analysis. These are supported by a variety of data types including sequence reads and assemblies, annotations, genomes and genome features, metabolic models, media, biochemistry, metabolic pathways and pangenomes. Users can also import their own reads, genomes, plant transcripts, media, flux balance analysis models, phenotype sets, and expression matrices for analysis and visualization.

A TANDEM SIMULATION FRAMEWORK FOR PREDICTING MAPPING QUALITY

Ben Langmead

Johns Hopkins University, Computer Science, Baltimore, MD

Read alignment is the first step in most sequencing data analyses. It is also a source of errors and interpretability problems. Repetitive genomes, algorithmic shortcuts, and genetic variation impede the aligner's ability to determine a read's true point of origin. Aligners therefore report a mapping quality: the probability the reported point of origin for a read is incorrect. I will describe an accurate, aligner-agnostic framework for predicting mapping quality called "tandem simulation," which works by simulating a set of tandem reads, similar to the input reads in important ways, but for which the true point of origin is known. Alignments of tandem reads are used to build a model for predicting mapping quality, which is then applied to the input-read alignments. The model is automatically tailored to the alignment scenario at hand, allowing it to make accurate mapping-quality predictions across a range of read lengths, alignment parameters, genomes, and read aligners

I will also describe an efficient implementation of the tandem simulation framework in a new open source software tool called Qtip. Qtip is compatible with popular read aligners such as Bowtie 2, BWA-MEM and SNAP. Qtip's mapping-quality predictions are superior to those produced by the read aligners themselves, both on average and for most specific MAPQ cutoffs tested.

CONVEX: FAST AND ACCURATE DE NOVO TRANSCRIPTOME RECOVERY FROM LONG READS

Meisam Razaviyayn[1], Elizabeth Tseng[2], David Tse[3]

[1]University of Southern California, Industrial and Systems Engineering, Los Angeles, CA, [2]Pacific Biosciences, Menlo Park, CA, [3]Stanford University, Electrical Engineering, Stanford, CA

The complexity of higher eukaryotic genomes imposes significant limitations on the assembly of transcript and splicing discrimination. In particular, it is known that in the presence of certain repeat structures, the transcriptome de novo assembly and splice product discrimination from short reads are impossible even when all constituent elements are identified. These limitations promote the use of long read isoform sequencing (Iso-Seq) technology to discover novel splicings.

We consider an unsupervised clustering problem motivated by the de novo processing of the long sequencing reads. Despite combinatorial nature of the problem, we propose an iterative convex reformulation of the problem using a greedy procedure **with order optimal sample complexity**. Based on the proposed greedy procedure, we develop an algorithm, dubbed CONVEX, which has linear computational complexity in the number of reads and can be implemented on parallel multi-core machines. We compare the performance of the algorithm with the state-of-the-art software, ICE, on various datasets such as ERCC2.0 and heart/liver tissue PacBio datasets. Our numerical experiments show that CONVEX results in up to 20% improvement in the number of denoised reads in addition to being multiple times faster than ICE on moderate and large size datasets

← Back

VARMATCH: ROBUST MATCHING OF SMALL VARIANT DATASETS USING FLEXIBLE SCORING SCHEMES

Chen Sun, Paul Medvedev

The Pennsylvania State University, University Park, PA

Small variant (<30bp) calling is widely used in medical and genetic research. Variant matching is the problem of determining which variants are in common between different sets of variant calls. Variant matching can be done to (1) compare the performance of different tools with respect to each other or to a ground truth, (2) extract high-confidence variants by taking the intersection of calls from multiple callers, and (3) find variants that are shared or unique across different individuals.

A set of small variants is typically represented as a collection of VCF entries, and the most straightforward variant matching algorithm is to directly match identical VCF entries. However, it can fail to match two different VCF entries that nevertheless result in the same diploid donor genome. Normalization and decomposition have been proposed to alleviate these problems, however, there are still alternate representations for the same variant that are not matched. On the other hand, an exhaustive search approach can find all matches but suffers from large non-polynomial running times and large memory usage. Additionally, these approaches can only support maximizing the number of total matched VCF entries. However, this is sensitive to whether a tool represents complex variants as a single entry or as multiple, decomposed, entries. A more representation-invariant optimization criteria would be to maximize the number of matched nucleotides. When comparing multiple callers to a ground truth set, it is instead desirable to maximize the total number of matched ground truth entries.

To address these problems, we introduce a new tool VarMatch (https://github.com/medvedevgroup/varmatch). VarMatch contains an exact algorithm for variant matching that is guaranteed to find matching variants under a wide variety of optimization criteria. VarMatch employs a provably optimal divide and conquer strategy to partition the set of variants into small disjoint subproblems which can then be solved independently using an exact algorithm. We tested VarMatch on real published data sets (CHM1 and GIAB), where it detected more matches than either normalization or decomposition. Our experiments also illustrated that under different optimization criteria, variant matching results on the same data set are different. Our algorithm performs very fast in practice and uses an order of magnitude less memory than exhaustive search. This can be crucial for applications in medical settings, where the software may be run on embedded processors or portable devices. VarMatch is also a parallel algorithm that scales over multiple processors and/or threads.

← Back

METAGENOME ANNOTATION WITH DISTRIBUTED REFERENCE GRAPHS

Andre Kahles[1,2], Gunnar Rätsch[1,2]

[1]ETH Zurich, Biomedical Informatics Group, Zurich, Switzerland,
[2]MSKCC, Computational Biology Center, New York, NY

The accurate and comprehensive annotation but yet sparse representation of a sample of mixed sequences remains an unsolved problem for metagenome and metatranscriptome sequencing projects. The characterization of microbial communities becomes an increasingly relevant task in clinical research. Currently established methods for the analysis of microbiota assign variants of 16S ribosomal RNA or reads from whole genome shotgun sequencing (WGS) to single entities in a given taxonomy or to functional databases. These approaches are limited by incomplete taxonomies and annotation biases and, importantly, waste a large fraction of the raw sequence data that cannot be assigned to existing references. To address these shortcomings, we have implemented a new, highly sensitive approach to combine, represent and identify the microbial and functional composition of a large set of metagenome samples with a major focus on taking previous knowledge into account. Building on techniques from genome assembly and text compression we use succinct data structures to efficiently represent all sequence information in a k-mer based assembly graph, which not only represents single species and their individual relationships but also captures intra-species variability. A set of 2785 different bacterial genome sequences is compressed by over 50% when stored in the graph instead as raw sequences. The graph is structured as a dynamic self-index that allows for efficient extension and can be used for alignment and annotation of reads arising from metagenome sequencing experiments. Hence, our representation is not only sparse but also efficiently searchable and allows for efficient extension with new sequences without re-indexing. We also developed a concept to distribute the index over a set of computers for fast alignment. The nodes of the graph are colored using hashes of binary annotation vectors, encoding information such as species, functional elements or other associated metadata. The graph's full utility is tailored for the use on WGS data, where not only unknown species can be represented in their correct relationship but also single functional entities, e.g., single genes, can be identified. The reference graph leverages information from known genomes as well as from the many previous studies, giving access to rare observations not yet present in reference databases. It will integrate further knowledge over time and to accumulate information, e.g., over many patients and studies. Thus, it will have a greater sensitivity to detect unseen or rarely seen species and inherently represents nearest neighbors with less bias towards species overrepresented in existing databases.

← Back

FAST APPROXIMATE MAPPING OF SINGLE MOLECULE SEQUENCES USING MINHASH

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

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

Enabled by recent advances in nanopore sequencing, real-time, single-molecule DNA sequencing is poised to become an important tool for clinical diagnostics. Single-molecule reads, however, are much longer and more noisy than Illumina reads, posing challenges for established mapping algorithms. Further, many mapping applications do not require a full alignment, which can be costly for noisy sequences. We propose a fast, lightweight read mapping tool based on minimizers and similarity approximation using the MinHash technique. This tool computes the positional origin of a sequence in the reference and estimates nucleotide identity under an assumed probabilistic model of errors. Although we do not compute a full gapped alignment, our method is >100x faster than alignment-based tools like BWA-mem and BLASR, and sufficient for applications such as copy number estimation, metagenomic read classification, and controlling Oxford Nanopore's “read until” feature. For mapping human PacBio reads to the hg38 reference, we achieve a recall rate of 96% when measured against BWA-mem alignments, and a precision of 90%, as tested by Smith-Waterman validation. Our two-stage mapping strategy combined with identity estimation enables a much higher precision than minimap (an alternative approximate read mapper), while requiring similar runtime and significantly less memory. This has allowed us to perform real-time, approximate mapping of unknown nanopore reads against a complete RefSeq database containing >60,000 genomes. We also discuss the theoretical and empirical correlations of read accuracy and length on performance, and discuss potential applications to high-throughput, streaming sequence data as produced by nanopore sensors.

← Back

ACCURATE AND FAST DETECTION OF COMPLEX AND NESTED STRUCTURAL VARIATIONS USING LONG READ TECHNOLOGIES.

Fritz J Sedlazeck[1], Philipp Rescheneder[2], Arndt v Heaseler[2], Michael C Schatz[1]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna, Vienna, Austria

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 Sniffles, a method for detecting all types of SVs from long-read sequencing data. A unique feature of Sniffles is detecting nested SVs (e.g. chromothripsis), such as inversions flanked by deletions, which we now commonly detect in several samples. Sniffles finds SVs from split-read alignments as well as the analysis of "noisy regions", where sequencing errors mask the presence of biological differences. A self-balancing interval tree enables a fast runtime so that whole human genome datasets can be analyzed in minutes. Furthermore, Sniffles offers read-level phasing to study complex breakpoints, as in chromoplexy. Using real and simulated data, we demonstrate the enhanced ability of Sniffles to detect SVs over existing methods like PBHoney or short read methods such as Lumpy, Delly, or Manta. We further introduce a new long-read mapping method called NGM-LR to enhance the accuracy of Sniffles and reduce the false discovery of SVs even further. NGM-LR is uniquely capable to detect and react if a read overlaps with a disturbed region (e.g. SVs) during the mapping phase and thus provides a more accurate alignment.

Working with genuine PacBio and Oxford Nanopore reads with human cancer samples (SKBR3), healthy human samples (GIAB), and other species, we show how Sniffles combined with NGM-LR 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 institutes around the world.

← Back

QUANTIFICATION OF SENSITIVE INFORMATION LEAKAGE FROM GENOMIC LINKING ATTACKS

Arif Harmanci[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

Genomic privacy has recently gained much attention. Most studies on privacy focus on identification of scientific study participants by protection of genetic variants. The extent to which phenotypic data can be utilized in breach of privacy is just recently being recognized. Basically, an adversary can use phenotypic information and predict genotypic information utilizing the publicly available datasets that contain phenotype-genotype correlations, e.g. quantitative trait loci (QTL). The predicted genotypes can be cross-referenced with genotype datasets and be linked to the phenotype datasets. These links can then be used to reveal sensitive information, such as disease status, that is stored within genotype/phenotype datasets. The availability of large scale raw and processed sequencing datasets increases the likelihood of a successful linking attack. We present methods for quantifying leakage of sensitive information that can be used to characterize individuals. We also present a risk assessment and prevention protocol for instantiation of linking attacks. This enables systematic analysis and avoidance of these attacks. We finally present several instances of these attacks that are simple yet highly effective. We focus on assessing the risks when the attacker uses a heterogeneous set of data types in linking attack.

← Back

PREDICTING TISSUE-SPECIFIC EFFECTS OF RARE GENETIC VARIANTS

Farhan N Damani[1], Yungil Kim[1], Xin Li[2], Emily K Tsang[3], Joe R Davis[4], Colby Chiang[5], Zachary Zappala[4], Benjamin J Strober[6], Alexandra J Scott[5], Ira M Hall[5], Stephen B Montgomery[2,4], Alexis Battle[1]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Stanford University, Pathology, Stanford, CA, [3]Stanford University, Biomedical Informatics, Stanford, CA, [4]Stanford University, Genetics, Stanford, CA, [5]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [6]Johns Hopkins University, Biomedical Engineering, Baltimore, MD

Despite the abundance of rare variants in human populations, the functional role of rare regulatory and non-coding variants has been largely uncharacterized. Population-level test statistics have had limited success in characterizing the effects from rare single nucleotide variants (SNVs) on gene regulation due to limited statistical power. We describe a hierarchical Bayesian model that captures tissue-specific cis-regulatory effects of rare variants using a transfer learning framework. This approach presents the opportunity to learn reliable, tissue-specific effects despite observing only a limited number of tissues or samples for each individual. To infer the regulatory impact of each rare SNV, our probabilistic model capitalizes on a growing body of rich epigenetic data to inform the deleteriousness of a SNV in specific cell-types, using factors such as conservation scores, transcription factor binding sites, and chromatin marks. These annotations are integrated with individual, tissue-specific gene expression data across 44 tissues from the Genotype-Tissue Expression (GTEx) project. Our framework explicitly models effects on gene expression from common variants, rare structural variants, and short insertions and deletions, in addition to rare SNVs. We compared our method to state-of-the-art variant prediction tools and observed a significant performance boost in predicting held-out expression levels. Our model also predicted tissue-specific allele-specific expression significantly better than existing models or cross-tissue models. Together this provides a framework for prioritizing rare variants likely to affect gene regulation, an important step toward personal genomics.

← Back

COMPUTATIONAL PROTEOGENOMIC IDENTIFICATION ANDFUNCTIONAL INTERPRETATION OF TRANSLATED FUSIONS ANDMICRO STRUCTURAL VARIATIONS IN CANCER

Yen-Yi Lin[1], Alexander R Gawronski[1], Faraz Hach[1,3], Sujun Li[2], Ibrahim Numanagic[1], Iman Sarrafi[1,3], Swati Mishra[5], Andrew McPherson[1], Colin Collins[3], Milan Radovich[5], Haixu Tang[2], Cenk Sahinalp[1,2,3]

[1]Simon Fraser University, Computing Science, Burnaby, Canada, [2]Indiana University, Informatics and Computing, Bloomington, IN, [3]Vancouver Prostate Centre, Vancouver, Canada, [4]University of British Columbia, Urologic Sciences, Vancouver, Canada, [5]Indiana University, Surgery, Indianapolis, IN

Rapid advancement in high throughput genome and transcriptome sequencing (HTS) and mass spectrometry (MS) technologies has enabled the acquisition of the genomic, transcriptomic and proteomic data from the same tissue sample. Here we introduce a novel computational framework which, for the first time, can integratively analyze all three types of omics data to obtain a complete molecular profile of a tissue sample, in normal and disease conditions. Our framework includes MiStrVar, an algorithmic method we developed to identify micro structural variants (microSVs) on genomic HTS data. Coupled with deFuse, a popular gene fusion detection method we developed earlier, MiStrVar can provide an accurate profile of structurally aberrant transcripts in cancer samples. Given the breakpoints obtained by MiStrVar and deFuse, our framework can then identify all relevant peptides that span the breakpoint junctions and match them with unique proteomic signatures in the respective proteomics data sets. Our framework's ability to observe structural aberrations at three levels of omics data provides means of validating their presence, potentially implying functional significance, and, in principle, novel drug targets in the tumor tissues we analyzed.

We have applied our framework to all The Cancer Genome Atlas (TCGA) breast cancer Whole Genome Sequencing (WGS) and/or RNA-Seq data sets, spanning all four major subtypes, for which proteomics data from Clinical Proteomic Tumor Analysis Consortium (CPTAC) have been released. A recent study on this dataset focusing on SNVs has reported many that lead to novel peptides [1]. Complementing and significantly broadening this study, we detected 363 novel peptides from 568 candidate genomic or transcriptomic sequence aberrations. Many of the fusions and microSVs we discovered have not been reported in the literature. Interestingly, the vast majority of these translated aberrations (in particular, fusions) were private, demonstrating the extensive inter-genomic heterogeneity present in breast cancer. Many of these aberrations also have matching out-of-frame downstream peptides, potentially indicating novel protein sequence and structure. Moreover, the most significantly enriched genes involved in translated fusions are cancer-related. Furthermore a number of the somatic, translated microSVs are observed in tumor suppressor genes.

1. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62 (2016)

← Back

A SYSTEMATIC STUDY ON THE DISTRIBUTION AND OCCURRENCE OF MUTATIONS WITHIN REGULATORY MOTIFS FROM CANCER GENOME.

Ambarnil Ghosh, Subhra Pradhan, Kyeong Kyu Kim

Sungkyunkwan University School of Medicine, Department of Molecular Cell biology, Samsung Biomedical Research Institute, Suwon, South Korea

Tremendous advancement of DNA sequencing technology enabled large scale identification of mutations specific to a particular disease or patient in TCGA pan-cancer analyses era. Up to date, most of the significant mutations are counted on the basis of protein coding regions and consequent potential to manipulate open reading frame. Deciphering the exact functional role of the mutations occurred within noncoding regions remains a monumental challenge. Several studies were published to focus noncoding mutations within the genome but prioritizing important target region is a challenging task. In the current computational biology based data mining work, authors made a thorough mapping of the mutations in the regulatory DNA motifs of the promoter region from more than 1200 gastrointestinal cancer patient genome. Total five types of TCGA cancer types (CHOL, ESCA, LIHC, PAAD and STAD) and several regulatory motifs (including TATABox, initiator, important protein binding motifs, etc.) from whole genome promoters are considered for this study. The percentage of genes involved in core promoter motif mutation varies among types of cancer: ~1 to 5% and also for different motif types. Interestingly, top scoring genes (or a single gene) (sample densities) involved in a particular type of cancer emerges significantly through the analysis; as example gene HMGN1 holds highest number of mutations among esophageal cancer genomes' for TATAbox containing motif region. Likewise MIR88, BCKDHA, etc. hold as leading candidates in other gastrointestinal cancers. Only ~0.5 to 5% of the genes with promoter motif mutations are found with significant high sample densities. Therefore, the analyses of these high frequency mutations those are occurring in the non-protein coding regulatory regions can enhance our understanding of the unusual transcriptional regulatory networks of the cancer genome. This study will also increase the possibility of finding new candidates involving in cancer progression. This pipeline of genome-wide analysis of mutations can be applied separately to various group of cancer to find out the novel recurrent mutations for each type. Further analysis would lead to the identification of clinically significant mutations that can be used as novel drug-targets for therapeutic purpose.

CELL TYPE SPECIFIC GENE EXPRESSION VARIATION IN DIFFERENTIATED INDUCED PLURIPOTENT STEM CELLS

YoSon Park*[1], Evanthia E Pashos*[1,2], Mayda Hernandez[1], Stacey Lyttle[1], Dawn Marchadier[1,2], Juan Arbelaez[4], Wenjun Li[3], Zhaorui Lian[4], Jianting Shi[4], Katherine J Slovik[4], Ruilan Yan[4], Kiran Musunuru[1,3], Wenli Yang[4], Edward E Morrisey[3,4], Stephen Duncan[5], Daniel J Rader[1,2,3], Christopher D Brown[1]

[1]Perelman School of Medicine University of Pennsylvania (PSoM UPenn), Dept of Genetics, Philadelphia, PA, [2]PSoM UPenn, Division of Translational Medicine and Human Genetics, Dept of Medicine, Philadelphia, PA, [3]PSoM UPenn, Cardiovascular Institute, Philadelphia, PA, [4]PSoM UPenn, Institute of Regenerative Medicine, Philadelphia, PA, [5]Medical University of South Carolina, Dept of Regenerative Medicine and Cell Biology, Charleston, SC

Genome-wide association studies (GWAS) have identified thousands of loci associated with complex traits, but few causal genes and variants have been validated. Efforts such as the Genotype-Tissue Expression (GTEx) project have identified genes whose expression levels are correlated with genetic variants. However, these studies rely on scarce primary tissues from postmortem donors. In contrast, induced pluripotent stem cells (iPSCs) offer an alternative, convenient, and potentially unlimited resource for molecular trait mapping and other functional studies. They can be obtained from living donors non-invasively, are expandable in vitro, and can be differentiated to the desired cell type. We recruited a cohort of 90 donors and generated 89 iPSCs and 86 iPSC-differentiated hepatocytes (iPSC-Heps). We generated genotype and whole transcriptome (RNA-seq) data for all samples. Further, we contrasted findings from our cohort with RNA-seq data from publicly available GTEx tissue samples. We developed an analytic pipeline facilitating study of multiple cell lineages. All raw data are processed via identical pipelines and stringent quality control measures were applied. Analysis of gene expression demonstrated that iPSC-Heps form a distinct cluster from either iPSCs, hepatocytes, or liver. These results suggest that incorporating iPSCs and iPSC-Heps provides additional insight that whole-liver samples (i.e. non-specific to hepatocytes) may not. We identified eQTLs within each tissue and characterized their pattern of cell type specificity by meta-analysis. We identified a total of 6,231, 3,993 and 3,055 eGenes at FDR<5% for iPSCs, iPSC-Heps and GTEx-Livers, respectively. Considering the importance of liver in lipid metabolism, we applied Mendelian randomization to identify potential causal associations between eQTLs identified in our study and loci associated with blood lipid levels, identified by the Global Lipids Genetics Consortium (GLGC). In additional to numerous novel discoveries, we replicated functional associations at known lipid loci such as *PCSK9* and *SORT1* via MR. We note that incorporating diverse cohorts including publicly available resources such as GTEx enhanced our power to profile liver-specific regulatory mechanisms. Further, we emphasize the importance of establishing standardized computational pipelines minimizing technical batch effects between data processed from multiple sources.

← Back

DISTANT REGULATORY EFFECTS OF GENETIC VARIATION IN MULTIPLE HUMAN TISSUES

Brian Jo*[1], Yuan He*[2], Benjamin J Strober*[2], Princy Parsana*[3], Francois Aguet[4], Andrew A Brown[5], Ariel Gewirtz[1], Eun Yong Kang[6], Ian C McDowell[7], Ashis Saha[3], Kristin Ardlie[8], Don Conrad[9], Emmanouil T Dermitzakis[5], Eric Gamazon[10], Eleazar Eskin[11], The GTEx Consortium[12], Barbara Engelhardt*[1], Alexis Battle*[3]

[1]Princeton Univ., Quantitative and Computational Biology, Princeton, NJ, [2]John Hopkins Univ., Biomedical Engineering, Baltimore, MD, [3]John Hopkins Univ., Computer Science, Baltimore, MD, [4]Harvard Univ., Cell Biology, Cambridge, MA, [5]University of Geneva, Genève, Genetic Medicine and Development, Geneva, Switzerland, [6]California State University, Computer Science, Los Angeles, CA, [7]Duke University, Computational Biology and Bioinformatics, Durham, NC, [8]Broad Institute of MIT and Harvard, Broad, Cambridge, MA, [9]Washington University School of Medicine, Genetics, Saint Louis, MO, [10]Vanderbilt University, Medicine, Nashville, TN, [11]University of California, Computer Science; Human Genetics, Los Angeles, CA, [12] MD

Expression quantitative trait loci, or eQTLs, are enriched for polymorphisms that have been found to be associated with disease risk. Distal or trans-eQTLs may have broader effects on the transcriptome and large phenotypic consequences, necessitating a comprehensive study of trans-effects on gene regulation. In this work, we identify trans-eQTLs in the Genotype-Tissue Expression (GTEx) project data, consisting of 449 individuals with RNA-sequencing across 44 tissue types. We find 81 genes with a trans-eQTL in at least one tissue, and demonstrate greater differences between tissues for trans-eQTLs than cis-eQTLs. We evaluate the genomic and functional properties of trans-eQTL variants, identifying strong enrichment in enhancer elements. Also, we investigate three-dimensional contact between trans-eQTL genes and corresponding variants, and the motif enrichment between trans-eQTLs and the associated genes' neighborhood in the 44 tissues. Finally, we describe two tissue-specific regulatory loci underlying relevant disease associations: 9q22 in thyroid and 5q31.1 in skeletal muscle. These analyses provide a comprehensive characterization of trans-eQTLs across human tissues, which contributes to an improved understanding the tissue-dependent cellular mechanisms of regulatory genetic variation.

FUNCTIONALIZING HUMAN DISEASE VARIANTS THROUGH SYSTEMS MODELING AND NETWORK ANALYSIS

Song Yi, Limei Hu, Nidhi Sahni

The University of Texas MD Anderson Cancer Center, Department of Systems Biology, Houston, TX

In the past decade, genome and exome sequencing projects have identified thousands of genetic variants in patients across a large number of Mendelian disorders, complex traits and cancer types. However, the explosion of genomic information has left many fundamental questions regarding genotype-phenotype relationships unresolved. One critical challenge is to distinguish causal disease mutations from non-pathogenic polymorphisms. Even when causal mutations are identified, the functional consequence of such mutations is often elusive. Classical "one gene, one function, one disease" models can not reconcile with the complexity that different mutations of the same gene often lead to different phenotypes. The extent to which interactome network perturbations are involved in disease malfunction and how distinct interaction perturbation patterns can distinguish disease mutations are largely unknown. Here we report a systematic approach to investigate genetic variant-specific effects on binary molecular interactions at large scale across diverse human diseases. Remarkably, in comparison to non-disease polymorphisms, disease mutations are more likely to associate with interaction perturbations. Over 60% of missense disease mutations are found to cause protein interaction alterations. While about ~30% result in loss of all their interactions (null-like), the other 30% exhibit selective elimination of specific interactions (edgetic). Different mutations of the same gene give rise to different interaction profiles, accounting for distinct disease outcomes. Edgetic mutations perturb interactions through disrupting specific interaction interfaces, and the perturbed partners are more likely expressed in relevant disease tissue. Together, our approach is insightful in prioritizing disease-causing variants, and uncovering patient mutation-specific disease mechanisms at a base-pair resolution, a critical step towards personalized precision medicine. Furthermore, our results suggest distinct interaction perturbations as a widespread mechanism underlying genetic heterogeneity, providing a fundamental link between genotype and phenotype in human disease.

← Back

INFERENCE OF TUMOR IMMUNITY AND T-CELL RECEPTOR REPERTOIRE FROM TCGA RNA-SEQ DATA

Xiaole Shirley Liu

Harvard University and Dana-Farber Cancer Institute, Biostatistics and Computational Biology, Center for Functional Cancer Epigenetics, Boston, MA

We developed a computational approach to study tumor-infiltrating immune cells and their interactions with cancer cells. Analysis of over ten thousand RNA-seq samples from the Cancer Genome Atlas (TCGA) identified strong association between immune infiltrates and patient clinical features, viral infection status, and cancer genetic alterations. We found that melanomas with high levels of CTLA4 as well as kidney tumors with high TIM3 separate into two distinct groups with respect to CD8 T-cell infiltration. We also developed a computational method to infer the complementarity determining region 3 (CDR3) sequences of tumor infiltrating T-cells in TCGA RNA-seq samples. CDR3 sequence length distribution and amino acid conservation, as well as variable gene usage of infiltrating T-cells in many tumors, except brain and kidney cancers, resembled those in the peripheral blood of healthy donors. We identified 3 potential immunogenic somatic mutations based on their co-occurrence with CDR3 sequences. The analysis of immune infiltrates and T-cell receptor repertoire might lead to useful insights on the clinical responses to immune checkpoint blockade.

← Back

INTERACTIVE GENOMIC VISUALIZATION TOOLS FOR LONG READ SEQUENCING, ASSEMBLY, AND CANCER

Maria Nattestad[1], Chen-Shan Chin[2], Michael C Schatz[1,3]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Pacific Biosciences, Bioinformatics, Menlo Park, CA, [3]Johns Hopkins University, Departments of Computer Science and Biology, Baltimore, MD

The complexity and scale of genome sequencing data creates challenges in interpretation, analysis, and algorithm development. Here we share two new interactive visualization tools created to address major challenges in studying complex variations.

First, most of the information in contig-to-contig or read-to-reference alignments is lost in the one-dimensional views of most genome browsers showing only reference coordinates. Structural variation can be contained entirely within long reads or assembled contigs, creating specific alignment signatures that are lost when looking at only the reference perspective. We overcome this by creating Ribbon (genomeribbon.com), an online visualization tool that displays alignments along both reference and query sequences. Ribbon enables understanding of complex variants, detection of sequencing and library preparation issues, testing of aligners and variant-callers, and rapid curation of structural variant candidates. In addition to SAM and BAM files, Ribbon can also load simple coordinate files from whole genome aligners. Therefore, Ribbon can be used to test assembly and scaffolding algorithms or inspect the similarity between species.

The second visualization challenge is that most genome browsers are insufficient for showing distant or interleaved rearrangements, especially the complexity of connections in a cancer genome. Widespread genomic rearrangements in many cancers can only be understood by looking at more than one chromosome at a time and seeing all the connections. To address this, we built SplitThreader (splitthreader.com), an online visualization tool focusing on long-range variants and their associated copy number changes. SplitThreader highlights all rearrangements between any pair of chromosomes and makes it possible to investigate concordance between variant calls and copy number changes. Moreover, SplitThreader builds a sequence graph of genomic rearrangements, and can query it to find connections between genomic loci, including those spanning multiple chromosomes, using a modified priority queue breadth-first search algorithm. We apply this graph search algorithm to test for genomic evidence of gene fusions identified using transcriptome sequencing and discover gene fusions that took place through multiple genomic rearrangements.

These tools open up new ways of looking at complex biological data, from the full view of alignments with Ribbon to the intertwining of long-range variants in cancer with SplitThreader.

← Back

ENHANCING PRE-DEFINED WORKFLOWS WITH AD HOC ANALYTICS IN A SINGLE ENVIRONMENT: UNLOCKING JUPYTER AND RSTUDIO FOR BIOLOGISTS

Bjorn Gruning[1], Eric Rasche[2], Torsten Houwaart[1], John Chilton[4], James Taylor[3], Anton Nekrutenko[4]

[1]University of Freiburg, BioInf, Freiburg, Germany, [2]Texas A&M, Phage Technology Center, College Station, TX, [3]Johns Hopkins University, Biology, Baltimore, MD, [4]Penn State University, BMB, University Park, PA

Trees, rivers, and the analysis of next generation sequencing (NGS) data are examples of branching systems so ubiquitous in nature: numerous types of NGS applications (i.e., variation detection, ChIP-seq, RNA-seq) 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 imaginary tree. Each of these subsequently gives off smaller offshoots (variant calling, RNA-, ChIP- and other ‘seqs’), that, in turn, split further as analyses become focused towards specific goals of an experiment. As we traverse the tree, established analysis tools are becoming 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) representing conclusive, publishable results. Consider transcriptome analysis as an example. Initial steps of RNA-seq analysis (in our tree allegory these are trunk and main branches), such as quality control, read mapping, 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 (smaller branches) ranging from simple format conversion to various statistical tests and visualizations. Thus every NGS analysis can be divided into two stages. (1) The first stage involves processing of raw data using a finite set of common generic tools. This stage can be scripted and automated and also lends itself to building user interfaces. (2) 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 cannot be coerced into a handful of automated routines or well defined user interfaces. The main motivation for this work was development of a system where biomedical researchers can perform both stages of data analysis: initial steps using established tools and exploratory and data interpretation steps with ad hoc approaches. To achieve this goals we augmented Galaxy system to integrate with Jupyter and RStudio interactive analysis platforms. The examples we use to highlight this new functionality demonstrate that it lowers entry barriers for individuals interested in data analysis, significantly improves reproducibility of published results, eases collaborations, and enables straightforward dissemination of best analysis practices.

← Back

GPEG: LOSSY AND LOSSLESS COMPRESSION OF SEQUENCING READ ALIGNMENTS ACROSS MANY SAMPLES

Abhinav Nellore[1,2,3], Jacob Pritt[1,3], Rachel Ward[4], Ben Langmead[1,3]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Johns Hopkins University, Department of Biostatistics, Baltimore, MD, [3]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [4]University of Texas, Department of Mathematics, Austin, TX

Sequencing projects including GTEx's RNA sequencing (RNA-seq) of normal tissue and ExAC's exome sequencing of unrelated individuals now span thousands to tens of thousands of samples, with several million reads per sample. So there is an increasing need to reduce storage of reads while maintaining queryability across samples at scale. Alignments of sequencing reads may be represented as a matrix, where reads are rows, columns are genome positions, and the (i, j)th element is nonzero if and only if read i maps across position j. This matrix, in turn, may be thought of as an image and thus admits compression strategies employed in image processing. We describe selected strategies used in a new compressed format for sequencing reads, gpeg. We offer fine-grained control over lossiness, providing options to discard read names, quality information, and poorly covered variants. Our approach for compression of RNA-seq reads in particular is inspired by Pritt and Langmead's Boiler, a radically lossy approach based on approximate recovery of read alignments from genome coverage vectors. We also describe how to extend our compressed format to index reads across many samples and rapidly obtain approximate results for cross-sample queries of sequencing depth and coverage of variants as well as junctions for RNA-seq reads. As a demonstration, we use gpeg to index RNA-seq reads from the GEUVADIS project, which spans lymphoblastoid cell line samples of 465 individuals from five populations of the 1000 Genomes Project.

RECOUNT: A LARGE-SCALE RESOURCE OF ANALYSIS-READY RNA-SEQ EXPRESSION DATA

Leonardo Collado-Torres[1,2,3], Abhinav Nellore[1,2,4], Kai Kammers[1,2], Shannon E Ellis[1,2], Margaret A Taub[1,2], Kasper D Hansen[1,2,6], Andrew E Jaffe[1,2,3,5], Ben Langmead[1,2,4], Jeffrey T Leek[1,2]

[1]Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD, [2]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [3]Johns Hopkins Medical Campus, Lieber Institute for Brain Development, Baltimore, MD, [4]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [5]Johns Hopkins University, Department of Mental Health, Baltimore, MD, [6]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

recount is a resource of processed and summarized expression data spanning nearly 60,000 human RNA-seq samples from the Sequence Read Archive (SRA). The associated recount Bioconductor package provides a convenient API for querying, downloading, and analyzing the data. Each processed study consists of meta/phenotype data, the expression levels of genes and their underlying exons and splice junctions, and corresponding genomic annotation. We also provide data summarization types for quantifying novel transcribed sequence including base-resolution coverage and potentially unannotated splice junctions. We present workflows illustrating how to use recount to perform differential expression analysis including meta-analysis, annotation-free base-level analysis, and replication of smaller studies using data from larger studies. recount provides a valuable and user-friendly resource of processed RNA-seq datasets to draw additional biological insights from existing public data. The resource is available at https://jhubiostatistics.shinyapps.io/recount/.

← Back

TACO: MULTI-SAMPLE TRANSCRIPTOME ASSEMBLY FOR RNA-SEQ

Yashar S Niknafs[1], Balaji Pandian[1], Hariharan K Iyer[2], Arul M Chinnaiyan[1,3], Matthew K Iyer[1]

[1]University of Michigan, Michigan Center for Translational Pathology, Ann Arbor, MI, [2]Colorado State University, Department of Statistics, Fort Collins, CO, [3]University of Michigan, Howard Hughes Medical Institute, Ann Arbor, MI

Recent efforts to sequence eukaryotic transcriptomes have made available thousands of datasets. Accurate inference of transcript structure and abundance from these data is foundational for molecular discovery and investigation of disease mechanisms. Towards this end, we present TACO, a computational method to reconstruct a consensus transcriptome from multiple RNA-Seq datasets. TACO employs change point detection to delineate transcript start and end sites and split gene-dense loci into individual genes, and features a novel dynamic programming assembly algorithm that leverages splicing pattern information. When applied to RNA-Seq data from the CCLE, TACO reconstructed reference genes with markedly higher precision and sensitivity than either Cuffmerge or StringTie-merge. For example, TACO correctly assembled 14,080 transcripts from CCLE breast cancer cell line data, compared to 11,535 and 9,672 assembled by Cuffmerge and StringTie-merge, respectively, achieving this performance with higher precision (2.2-fold and 1.7-fold greater than Cuffmerge and StringTie-merge, respectively). We anticipate that when incorporated into existing RNA-Seq analysis protocols, TACO will produce high quality transcriptome reconstructions and lead to accurate downstream analyses of gene abundance, conservation, variation, and functionality. The tool is available for download at http://tacorna.github.io.

← Back

INTEGRATIVE AND PREDICTIVE MODELING OF SEVERE CONTROLLED AND SEVERE RESISTANT HYPERTENSION AMONG AFRICAN-AMERICANS IN THE MH-GRID NETWORK

Cihan Oguz, Adam R Davis, Gary H Gibbons

National Institutes of Health, Cardiovascular Disease Section, GMCIDB, National Human Genome Research Institute, Bethesda, MD

Hypertension (HTN), or persistently high blood pressure (BP), is a medical condition that substantially increases the risk for heart attack, congestive heart failure, chronic kidney disease, and stroke, if left untreated. African-Americans are disproportionately affected by HTN and are more prone to its earlier onset compared to other ethnicities in the United States. In this study, we used random forests (RF) and neural networks (NN) for predictive modeling of HTN with clinical data and RNA-Seq based gene expression data from the peripheral whole blood samples of 180 African-American patients in the Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID) Network. This cohort was composed of healthy controls, severe controlled hypertensive (SCH) cases with well controlled BP under two or more antihypertensives, and severe resistant hypertensive (SRH) cases with inadequately controlled BP despite three or more antihypertensives.

With the aim of identifying the differences between the predictive signatures of SRH and SCH, we first built RF models of HTN by using 28 clinical variables and expression levels of 440 genes (Gene Panel 1) previously identified as putative blood pressure regulators in the literature. Combining clinical and expression data led to improved predictive performance as opposed to using each layer of data in isolation. We then determined SRH-specific and SCH-specific genes, as well as genes highly predictive of both HTN phenotypes. Biological processes linked to inflammation, including cytokine and collagen production were enriched among SRH-specific genes, whereas fundamental blood pressure regulation processes, such as the regulation of cytosolic calcium and vasoconstriction were enriched among SCH-specific genes. In contrast, transcriptional regulatory processes were highly enriched among genes robustly predictive of both HTN phenotypes. Next, we derived an alternative set of genes (Gene Panel 2) that generated significantly more predictive RF models than Gene Panel 1 and compared the predictive processes enriched in the two gene panels to derive biological insights. Finally, we used NN models to verify the performances of the RF-based gene subsets predictive of either HTN phenotype. Our systems-level approach illustrates the potential of multiple machine learning methods as diagnostic and informative tools for modeling HTN. The derived biological insights and the identified phenotype-specific predictive genes have potential implications within the context of HTN treatment for African-Americans.

← Back

SEARCHING AND EXPLORING GRAMENE'S COMPARATIVE GENOMICS DATASETS ON THE WEB

Andrew Olson*[1], Joseph Mulvaney*[1], James Thomason[1], Doreen Ware[1,2]

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

Gramene is a unique resource that comprises genomic, variation, pathway, and expression data for economically important crops and model organisms across the plant kingdom. These data are maintained in widely adopted platforms such as Ensembl, Reactome, and EBI Atlas. We perform comparative analyses across these datasets, populate MongoDB and Solr document databases, and provide access to the integrated data through a performant web service (http://data.gramene.org).

We have also developed a search interface (http://search.gramene.org) that uses this service to let researchers easily find genes and explore the results. The application proactively suggests appropriate terms as you type; shows an interactive distribution of search results across all genomes; and allows users to select details of a result to focus or expand a search. Each gene in the result list includes interactive views of the gene structure, gene family evolution, relevant pathways, expression profiles, and links within Gramene or to third-party databases.

In order to model the evolution of gene families, Gramene and Ensembl Plants have included many annotated genomes of crops, model organisms, and more distantly related species. This can lead to large gene trees where it is impossible to focus on a specific subset of genomes in the default visualization. In Gramene's search interface, users can choose a subset of genomes of interest which automatically filters search results and trims the species tree and gene trees to only show branches that contain genomes of interest.

In plants, the model organism Arabidopsis thaliana has many well annotated genes, so we use the gene trees to select a well annotated homolog as a guide for genes that don't have a descriptive name.

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

Development of the Gramene database is supported by an NSF award (IOS-1127112).

* Authors contributed equally

← Back

ANALYSIS OF QUANTITATIVE DATA OF NUCLEAR DIVISION DYNAMICS FROM SINGLE GENE KNOCKDOWN EMBRYOS FOR ALL ESSENTIAL EMBRYONIC GENES IN *C. ELEGANS*

Shuichi Onami

RIKEN Quantitative Biology Center, Laboratory for Developmental Dynamics, Kobe, Japan

In animal development, cells are genetically controlled to generate the three-dimensional structures of bodies and organs. A collection of quantitative data of morphological dynamics when a wide variety of individual genes are perturbed provides a rich resource for understanding animal development. Here we created a collection of quantitative data of nuclear division dynamics in early *C. elegans* embryos when each of all 351 essential embryonic genes was silenced individually by RNA interference (RNAi). The collection consists of 33 sets of quantitative data for wild-type embryos and 1,142 sets for RNAi-treated embryos, which were obtained by combining four-dimensional differential interference contrast microscopy and image processing. We applied computational phenotype analysis to this collection by mathematically defining 408 phenotypic characters, and found 9,251 RNAi-induced phenotypic alterations, which included many three-dimensional spatial and temporal alterations. We then applied hierarchical clustering to the result of computational phenotype analysis, and found seven clusters each of which contained genes functioning in a specific cellular process such as polarity/asymmetric division, DNA replication and chromosome maintenance/segregation. We also developed a novel computational method for inferring a causal network of phenotypic expression by finding correlations between phenotypic characters. By using this method, we deduced a causal network of the expressions of the 408 phenotypic characters during early *C. elegans* embryogenesis. Furthermore, we developed a method for inferring genes involved in this causal network. We created a model of *C. elegans* embryogenesis by integrating the causal network of phenotypic expression and a gene network deduced by using multi-omics data.

← Back

dbSNP IN THE ERA OF NEXT-GENERATION SEQUENCING AND BIG DATA

Lon Phan, Hua Zhang, Juliana Feltz, Wang Qiang, Rama Maiti, David Shao, Ming Ward

National Center for Biotechnology Information, NLM/NIH, Bethesda, MD

dbSNP, despite its name, is a database for all short (<=50bp) genetic variations that include SNV, small indels, microsatellites, and non-polymorphic germline and somatic variants. The primary roles of the database are to process and archive submissions, assign stable accessions, aggregate information from multiple submitters, map and annotate variants on the latest genome assembly and RefSeq sequences (genomic, mRNA, and protein), and distribute them for general use. dbSNP data are used in diverse fields including personal genomics, medical genetics, and variant analysis in many different organisms. The data are also integrated with other NCBI resources including ClinVar, Gene, PubMed, Nucleotide, Protein, Structure, BioSample, and BioProject. dbSNP data are made available in many ways: Entrez searches, annotated on the genome assemblies and RefSeqs, in Sequence Viewers, and via ftp downloads as BED, VCF, and other formats. In addition, many bioinformatics centers including EBI and UCSC host dbSNP data and the data is incorporated into popular open-source and commercial bioinformatics tools and pipelines. dbSNP house over 1.7 Billion Submitted SNP (ss) records that cluster to 800 Million non-redundant Reference SNP (rs) from over 360 organisms. Human is the largest organism by data volume, with 540 Million ss and 154 Million rs including large datasets from WES and WGS projects such as 1000Genomes, GO-ESP, and ExAC. These and other large-scale NGS submissions have enriched dbSNP with over 50 million rare human variants (MAF < = 0.001) as well as WES (3X over WGS) coding and splicing variants. Current NGS trends suggest thousands to millions of new samples will be sequenced in the next few years that will require new tools to annotate, prioritize, and interpret variants. To facilitate analysis and interpretation of variants from these genomes and promote development of new tools, dbSNP aims to enrich rs annotation to include information from VAAST variant priority score (VVP) (Flygare et al.), protein features such as post-translational modification sites, Conserved Domain Database (CDD), protein 2D and 3D structures, binding sites for interaction with protein, drug targets, and small molecules. This is in addition to the functional consequence, allele frequency, ClinVar clinical significance, and many other rs attributes already reported by dbSNP. This presentation will discuss the incorporation of the new annotations and tools for analysis of dbSNP and ClinVar contents.

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

← Back

SHINYLEARNER: ENABLING BIOLOGISTS TO PERFORM ROBUST MACHINE-LEARNING CLASSIFICATION

Stephen R Piccolo, Terry J Lee

Brigham Young University, Biology, Provo, UT

Machine-learning classification is becoming an invaluable tool for biologists. For example, scientists can derive biomarkers that predict whether patients will respond to a particular drug or belong to a specific disease subtype. When high accuracy has been attained, they can elucidate mechanisms underlying such outcomes. Although many machine-learning algorithms and software libraries have been developed, considerable barriers remain for biologists—especially those with limited computational background—to take advantage of these tools. Different algorithms may be written in different programming languages and may require different input formats. Software libraries may require dependencies that are difficult to install or may fail if the wrong versions are installed. To employ algorithms implemented in multiple software libraries, a biologist would need to learn multiple programming languages and be careful to ensure that valid comparisons were made across the algorithms. Support for combining evidence across classification algorithms—especially across multiple libraries—is limited in the current software ecosystem.

We developed ShinyLearner, a software framework that reduces these barriers. ShinyLearner integrates several popular machine-learning libraries (e.g., scikit-learn, mlr, weka) within a Docker container and includes all necessary software dependencies. Accordingly, ShinyLearner can be installed with relative ease, and research results can be reproduced more readily. ShinyLearner supports Monte Carlo cross validation and k-fold cross validation and provides an option for feature selection. When multiple classification algorithms are used, ShinyLearner dynamically selects the best algorithm using nested validation. Biologists (and other scientists) can design ShinyLearner analyses via either a command-line interface or a Web interface. Upon completion of an analysis, ShinyLearner produces output files in “tidy data” format so the results can be imported easily into external analysis tools, such as R, Python, or Excel. In addition, it produces an HTML report that summarizes the results in a manner suitable for non-computational scientists.

ShinyLearner provides flexibility for integrating new machine-learning libraries. Scientists who wish to do so must only 1) modify the Dockerfile to account for new dependencies, 2) create a bash script that supports specific parameters, and 3) perform a GitHub pull request. The source code is released under an MIT license (https://github.com/srp33/ShinyLearner).

← Back

EUPATHDB: INTEGRATING EUKARYOTIC PATHOGEN GENOMICS DATA WITH ADVANCED SEARCH CAPABILITIES.

Jane A Pulman[1], Wei Li[2]

[1]University of Liverpool, Center for Genomic Research, Liverpool, United Kingdom, [2]University of Pennsylvania, Department of Biology, Philadelphia, PA

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

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

Recent developments include a richer ontological representation of meta-data (including clinical meta-data) describing experiments and samples. A graphical UI aids users in identifying samples of interest by allowing them to visualize the structure of the underlying metadata as they make their selections.

Authors present on behalf of the EuPathDB team.

← Back

RNA-SEQ EXPRESSION ANALYSIS MADE EASY IN KBASE

Srividya Ramakrishnan[1], Michael Schatz[1,2], Vivek Kumar[1], Sunita Kumari[1], Jim Thomason[1], Doreen Ware[1,3], Shinjae Yoo[4], Priya Ranjan[5], Samuel Seaver[6], Nomi Harris[7], Christopher Henry[6], Robert Cottingham[5], Adam Arkin[7]

[1]Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [3]USDA ARS, Ithaca, NY, [4]Brookhaven National Laboratory, Upton, NY, [5]Oak Ridge National Laboratory, Oak Rigde, TN, [6]Argonne National Laboratory, Argonne, IL, [7]Lawrence Berkeley National Laboratory, Berkeley, CA

The U.S Department of Energy Systems Biology Knowledgebase (KBase, http://kbase.us) provides an open, web-accessible system for systems biology research focused on microbes, plants and their communities. It offers access to a range of integrated biological data types and a variety of analysis tools that include modeling, simulation methods and visualizations.

KBase allows users to integrate new analysis tools and data, and supports a rich set of computational methods and curated datasets for performing gene expression analysis from RNA-seq reads. This includes a selection of preprocessed high-quality reference genomes and a wide variety of analytical methods including algorithms for short-read mapping, identification of splice junctions, transcript and isoform detection and quantitation, differential expression and visualization. More specifically, KBase supports the Tuxedo suite of tools, including Bowtie, Tophat, Cufflinks, Cuffdiff, and CummeRbund. Newer tools such as HISAT2, StringTie and Ballgown will be available in the near future. KBase also provides other useful services that are directly integrated with gene expression profiles to do downstream analysis including clustering of expression profiles using different clustering algorithms.

The RNA-seq analysis services are available from within a highly interactive and dynamic user interface called the Narrative Interface. Within a Narrative, short reads from an RNA-seq experiment can be uploaded into KBase to perform gene expression analysis and the results can be shared, reproduced, and the research extended by others in the KBase community. We demonstrate the utility of the Narratives by performing a point-and-click, yet detailed analysis of public RNA-seq data from several species and tissue types, including experiments in A. thaliana and E. coli.

A State-of-the-art Compression Method for ChIP-seq Data

Vida Ravanmehr[1], Zhiying Wang[2], Olgica Milenkovic[1]
[1]University of Illinois at Urbana-Champaign, Coordinated Science Laboratory, Urbana, IL, [2]University of California, Irvine, The Henry Samueli School of Engineering, Irvine, CA

Chromatin immunoprecipitation-sequencing (ChIP-seq) is a technique that enables analyzing the interactions between protein and DNA using next generation sequencing technologies. ChIP-seq is an inexpensive DNA sequencing technique which combines ChIP with enormously parallel DNA sequencing to identify the binding sites of DNA-associated proteins. Specifying protein-DNA interactions and their role in regulating gene expression is of crucial importance in many biomedical applications. However, the vast volume of ChIP-seq data has caused challenges in terms of data storage, data transfer and exchange and hence increased the overall cost of data maintenance. To tackle this problem, different data compression techniques have been proposed to reduce the size of ChIP-seq files.

ChIP-seq data is usually represented in the wiggle (Wig) format. Wig format is composed of declaration lines and data lines. The ChIP-seq Wig files start with a declaration line and is followed by two columns containing chromosome positions and data values. Chromosome positions are integers while data values are continuous (non-negative real data) displaying the estimate of affluence.

We propose a new low-rate compression method especially designed for ChIP-seq data. Our approach is based on source coding techniques which include transform coding, differential coding and arithmetic coding for integers and correlated real numbers.

We tested our compression method on different ChIP-seq data generated by the ENCODE project. The results reveal that our method offers on average a 5-6 fold decrease in file size compared to bigWig, almost 2-fold decrease in file size compared to cWig and almost 14-fold decrease in file size compared to the original Wig files. The average size of the tested files was 1027. 75 MB in the original Wig format, 403.3125 MB in bigWig and 73.0437 MB under our compression method. Compression and decompression times of the proposed algorithm are comparable to those of bigWig.

Computational Determination of Tumor Type for Cancers of Unknown Primary Using Mutation Data

Jeffrey A Rosenfeld, Greg Riedlinger, Anshuman Panda, Wenjin Chen, Gyan Bhanot

Rutgers Cancer Institute of New Jersey, Pathology, New Brunswick, NJ

For cancer patients, determining the type of cancer is important for proper treatment. Cancer can form in any tissue in the body and the primary cancer can metastasize to other parts of the body. In order to determine the type of cancer and whether it is primary or metastatic a variety of laboratory tests including histologic analysis, immunohistochemistry, reverse transcription-PCR, cytogenetic analysis, and electron microscopy may be employed. Despite the use of these tests, occasionally cancer cells are found but the type of cancer is not clear; such as with cancer of unknown primary. Additionally, some patients have more than one primary cancer and if metastases develop it can be difficult to determine the site of origin. The recent implementation of clinical next-generation sequencing of cancer patients is an additional tool which can be useful for diagnostic purposes. Using the TCGA data, we investigated whether tumor type can be accurately determined using mutations identified through targeted sequencing. Specific mutations are very often found in particular types tumors, so anecdotally, it would be expected that this classification could work. We coded the presence or absence of a mutation as a binary term and applied computational algorithms to the classification task. We started by using Principal Components analysis and the results were of minimal quality. This is most likely due to the sparseness of the data where each tumor will mostly be normal and only have a small number of the possible mutations. To achieve greater power, we utilized an SVM to classify 4 types of cancer in women (breast,lung,skin,brain) and men (prostate,lung,skin,brain). For each group, we were able to achieve > 90% cross-validation accuracy for determining tumor type. This data was validated in an independent set of tumors we sequenced in-house.

← Back

TRANSCRIPTOME-WIDE NETWORKS REVEAL CANDIDATE SPLICING REGULATORY RELATIONSHIPS

Ashis Saha[1], Yungil Kim[2], Benjamin J Strober[3], Barbara E Engelhardt[4], The GTEx Consortium[5], Alexis Battle[6]

[1]Graduate Student, Department of Computer Science, Johns Hopkins University, Baltimore, MD, [2]Postdoctoral Researcher, Department of Computer Science, Johns Hopkins University, Baltimore, MD, [3]Graduate Student, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [4]Assistant Professor, Department of Computer Science & Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, [5]-, -, -, MD, [6]Assistant Professor, Department of Computer Science, Johns Hopkins University, Baltimore, MD

Alternative splicing is a complex but essential mechanism to code multiple proteins from a gene influencing downstream phenotypes. It plays a causal role in many human diseases including cancers. However, the regulation of splicing is not yet completely understood. Traditional network inference methods mainly capture co-regulation patterns via gene expression alone to characterize transcriptional regulatory relationships, overlooking regulatory effects on alternative splicing. RNA-sequencing has enabled us to simultaneously quantify a diverse range of transcription phenotypes, including alternative splicing, non-coding transcripts, and allele-specific expression. Leveraging these additional quantities, we can discover new regulatory relationships and elucidate the interplay between splicing and transcription. Here, we build transcriptome-wide networks (TWNs) over total gene expression and alternative splicing together from RNA-seq data using sparse Gaussian Markov Random Fields. We learn TWNs for 16 different human tissues using RNA-sequencing data from the GTEx project. Candidate splicing regulatory hubs in these networks are significantly enriched with known splicing regulators and RNA-binding proteins. These enrichments are even stronger in candidate regulatory hubs shared across multiple tissues. Our TWNs significantly capture regulatory relationships among genes within same pathways reflecting they are consistent with our current knowledge, and we are able to replicate our network relationships for whole blood in an independent dataset. In addition, we observe tissue-specific splicing regulator genes which may help explain context-specific regulatory behavior. We also investigate the ability of our networks to improve the power to detect trans splicing quantitative trait loci. Thus our transcriptome-wide networks can provide more comprehensive snapshots of gene regulation in multiple human tissues, crucial for understanding complex mechanisms in various contexts including human diseases.

← Back

SCENA: SECURE COMPRESSED GENOMIC DATA ANALYSIS IN A CLOUD ENVIRONMENT

Feng Chen[1], Ibrahim Numanagic[3], Sean Simmons[3,4], Can Kockan[2], Bonnie Berger[4], Xiaoqian Jiang[1], Tianyi Zhu[1], Weijia Wu[1], Shuang Wang[1], Cenk Sahinalp[2,3]

[1]UCSD, Biomedical Informatics, La Jolla, CA, [2]Indiana Univ., Sch. Informatics & Computing, Bloomington, IN, [3]Simon Fraser Univ, Sch. Computing Sci., Burnaby, Canada, [4]MIT, Computer Sci., Cambridge, MA

Background: Genomic data are highly sensitive: risks include -but are not limited to- re-identification of patients and inference of family ancestry. However, sharing and analyzing human genomic data offers many opportunities for scientific and medical discoveries. Encryption, especially when combined with proper data compression is essential for outsourcing genomic data in an untrusted cloud environment. Existing secure solutions (e.g., multiparty and homomorphic encryption or garbled circuits) for computation on encrypted data typically do not scale up well. Thus we present Secure Compressed gENomic data Analysis (SCENA) framework to enable both efficient and secure genomic sequence comparison and search across private databases managed in a cloud. SCENA builds on the recent breakthrough in software and hardware based secure computation architecture using Intel Software Guard Extensions (SGX). In SGX, a protected area inside the CPU, which is referred to as the secure enclave, is dedicated for executing sensitive codes on private data securely. Data confidentiality and integrity is achieved through proper systematic design choices in SGX applications.
Method: We extended the DeeZ SAM/BAM compression into a secure framework to enable random access and analysis (e.g., sequence alignment and sequence similarity search) over encrypted and compressed genomic data. SCENA is based on four key innovative steps to ensure the data confidentiality and data integrity. 1) The remote attestation protocol allows both the data owner and an enclave to verify each other's authenticity and integrity. 2) After attestation, the data owner can send encrypted and compressed data to a trusted enclave. 3) Once the data received, the enclave can use data sealing to securely store data outside the enclave (i.e., secure storage outsourcing). 4) The enclave can securely retrieve arbitrary gene regions from the sealed data followed by a decryption and decompression procedures within the enclave for further data analysis.
Experiment: Using data from the 1000 Genomes project, we demonstrated that **SCENA** can efficiently and securely compute pairwise sequence similarity across genomic regions of interest. On a pair of DeeZ compressed BAM files of ~200MB, SCENA requires 4.548s for encryption and 6.614s for data sealing in the enclave. An additional 12.527s is needed for for data unsealing, decompression and pairwise sequence similarity computation on genomic regions of 10Kbp.

← Back

IN SILICO ANALYSIS OF VIRAL DIVERSITY OBTAINED THROUGH VIRAL METAGENOMICS OF CHICKEN RESPIRATORY TRACT

Manisha R Sajnani[1], Tejas G Oza[2], Ramesh J Pandit[3], Prakash G Koringa[3], Subhash J Jakhesara[3], Prakash B Banakar[5], Chaitanya G Joshi[3], D Sudharsanam[4], Vaibhav D Bhatt[2]

[1]Bharathiar University, Bioinformatics, Coimbatore, India, [2]Saurashtra University, Pharmaceutical Sciences, Rajkot, India, [3]Anand Agricultural University, Animal Biotechnology, Anand, India, [4]Loyola College, Advanced Zoology and Biotechnology, Medical Lab-technology & Bio Medical Instrumentation, Chennai, India, [5]Indian Agricultural Research Institute, Nematology, New Delhi, India

Background: The incidence and severity of respiratory disease in commercial broiler chicken flocks have increased recently in India caused by viruses. However, the roles of viruses, singly or jointly, in recent outbreaks of respiratory disease are not clear. Recent advancement in the sequencing technologies, bioinformatics tools, genome databases and supercomputing facilities has overcome the challenges of analyzing large genome datasets generated from High Throughput Sequencer (HTS). In the present study, we have used Ion Torrent Personal Genome Machine (PGM) to sequence viral metagenome present in the respiratory tract in poultry. Furthermore, to understand the diversity of viruses, abundance and their functions, we have used various bioinformatics tools for quality check (QC), host specific sequence screening (HSS), de novo assembly and reference based mapping with viral sequences available in genomic databases.

Results: Deep sequencing resulted in 1.6 gigabases (Gb) of DNA virus sequences and 1.1 Gb of RNA virus sequences of which 1.1 Gb and 0.8 Gb of datasets passed QC and HSS, respectively. The downstream analysis was performed using supercomputing facility of 2 TB RAM and 100 core processor. The majority of the sequences analyzed were novel suggesting that the viral community in poultry, as well as the animal and plant hosts they feed on, is highly diverse and largely uncharacterized. Virus family identified included alloherpesviridae, herpesviridae, phycodnaviridae, gemycircularvirus, mimiviridae and nudiviridae, present in high proportion in DNA sequences whereas retroviridae, baculoviridae, paramyxoviridae, coronaviridae and bunyaviridae, present predominantly in RNA sequences, amongst which herpesviridae and retroviridae are reported previously to be inducing cancer in poultry. Numerous bacteriophages and unknown viruses were also detected.

Conclusion: This study had enabled broad survey of viral diversity, utilizing various bioinformatics software and databases and has significantly increased our understanding of the viruses associated with respiratory infections.

← Back

OPTIMIZING DE NOVO ASSEMBLY FOR *MELOIDOGYNE INDICA* TRANSCRIPTOME: A ROOT-KNOT NEMATODE

Manisha Sajnani, Prakash Banakar, Uma Rao

Indian Agricultural Research Institute, Nematology, New Delhi, India

Meloidogyne indica is commonly known to be citrus root-knot nematode which possess an extraordinary ability of infecting woody perennials like citrus and BT-cotton. The management control of this nematode is essential to reduce the economic losses caused to plants. Thus to understand the biology of the pre-parasitic juveniles, we performed transcriptome sequencing using Illumina technology. The sequencing resulted in 84 million reads of 76 read length, 6.42 giga bases and 71.3X coverage data. As no genome resource is available for this organism, an optimum assembly for this transcriptome was required for its downstream analysis. To address this challenge, we used three different de novo transcriptome assemblers: Velvet-Oases pipeline, Trinity and rnaSPAdes at different k-mers and observed that Trinity assembler showed better results at k-mer 31 generating 39,649 contigs. The quality of the these assemblies was assessed by the number of contigs generated, contig length distribution, N50 contig size, percent paired-end read mapping, and gene model representation via BLASTX. This research will assist in the study of the genes, targeted proteins, orthologos, repeats, gene expressions, effectors that are involved in its parasitism.

← Back

JOINT PROBABILISTIC MODEL FOR MULTIPLE STEPS OF GENE REGULATION

Hirak Sarkar[1,2], Yi-Fei Huang[1], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for quantitative biology, Cold Spring Harbor, NY, [2]Stony Brook University, Computer Science, Stony Brook, NY

Eukaryotic gene expression is a complicated, multistage process that is regulated at transcriptional, post-transcriptional, and translational levels. Various high-throughput assays, including PRO-seq/GRO-seq, RNA-seq, and Ribo-seq, have been developed to quantify different aspects of gene regulation, but they are rarely applied in combination. Here, we propose a novel probabilistic graphical model that jointly describes high-throughput data representing multiple stages of gene regulation. We focus in particular on the estimation of degradation rates of RNA, by contrasting rates of transcription with steady-state RNA concentrations. For each gene, we model the transcription rate and the degradation rate as two latent random variables. The observed data from a GRO-seq experiment are interpreted as noisy readouts of the transcription rate. In addition, assuming a birth-death equilibrium, the observed RNA-seq data indicate the ratio of transcription rate to degradation rate. To study the change of transcription and degradation rates in a treatment/control setting, we use Bayesian model comparison to select the optimal model. Then, given the best model, we estimate the transcription and degradation rates for each condition. As a proof of concept, we have applied our model to a publicly available GRO-seq and RNA-seq dataset from a MCF7 cell line treated with 17β-estradiol (E2). Preliminary results indicate that we can identify a set of genes regulated at the degradation step, providing new insights into the response of MCF7 cell line to E2 treatment.

← Back

EXPLAINING CANCER MICROENVIRONMENTS VIA AUTOMATED LITERATURE EXPLORATION

Adam A Butchy[1], Khaled Sayed[2], Kai-Wen Liang[3], Bryce Aidukaitis[4], Cheryl Telmer[5], Natasa Miskov-Zivanov[6]

[1]University of Pittsburgh, Bioengineering, Pittsburgh, PA, [2]University of Pittsburgh, Electrical and Computer Engineering, Pittsburgh, PA, [3]Carnegie Mellon University, Electrical and Computer Engineering, Pittsburgh, PA, [4]University of Virginia, Biomedical Engineering, Charlottesville, VA, [5]Carnegie Mellon University, Biological Sciences, Pittsburgh, PA, [6]University of Pittsburgh, Electrical and Computer Engineering, Pittsburgh, PA

Simulating complex biological processes with executable models lends insight into how they function and can lead to focused disease treatments. With biomedical research results being published at a high rate and using existing search engines, the vast amount of published work is easily accessible and can be translated into an executable model. Here, we present an automated framework called LEaRn (Literature Exploration and Reasoning) which allows for rapid analysis of the existing knowledge in literature. Although the technology that we are developing can be applied to other domains, thus far, our work has been focused on the cancer microenvironment; specifically, interactions between pancreatic cancer cells and macrophages.

Creating an executable model from published literature in an automated way starts with reading engines that select published papers according to a set of search terms, then processes these texts, and finally outputs the extracted information using formats that are easily understood by computers. An example format is a JSON object representing an event as an "interaction" between a regulator and a regulated element. Each new element extracted from literature is described in terms of its element type (e.g. protein, chemical, gene, etc.), the cell type in which it is found, its location in the cell, its activators and/or inhibitors, mechanisms of interactions, and references to articles that demonstrate the interactions. From these extracted interactions, we can either assemble new models or extend an existing model. As only a few search terms can result in thousands of selected papers and hundreds of thousands of extracted interactions, we developed several methods to group these new interactions. Next, we incorporate these grouped interactions into our model, the Dynamic Cancer Environment (DyCE) model, and create multiple new model versions. Our framework then applies stochastic simulation and statistical model checking to study the behavior of extended models under different conditions, and to compare them with the baseline model. In simulation, initial conditions are set for each element to describe its initial state in the system (e.g., not present, low presence/activity, high presence/activity). The next value for each element is determined by an update rule and a stochastic update scheme is used to execute these rules and to obtain element trajectories outlining the behavior of the system over time.

By representing system ground truths and hypotheses as formal logic properties, our framework can determine model extensions or perturbations that satisfy these properties. This way, using existing knowledge in published literature, the LEaRn framework quickly tests hypotheses and provides a technology to rapidly identify suitable system interventions and novel targets for treatment design.

← Back

INTERACTIVE VISUALIZATION AND ANALYSIS OF LARGE-SCALE SEQUENCING DATASETS WITH THE ZENBU GENOME BROWSER SYSTEM

Jessica Severin[1], Marina Lizio[1], Hideya Kawaji[1,2], Nicolas Bertin[1,3], Alistair Forrest[1,4], Yoshihide Hayashizaki[1,2]

[1]RIKEN , Center for Life Science Technologies, Yokohama, Japan, [2]RIKEN, Preventive Medicine and Diagnosis Innovation Program, Wako, Japan, [3]National University of Singapore, Cancer Science Institute, Singapore, Singapore, [4]Harry Perkins Institute of Medical Research, Cancer and cell biology division, Nedlands WA, Australia

Recent genome-wide compendium studies, such as FANTOM5, Encode, Epigenome Roadmap, TCGA, and more recently single-cell studies are providing new challenges for visualization and analysis due to their unprecedented breadth and scale. To address this need, we developed the ZENBU system (Severin et al., /Nature Biotechnology/, 2014). ZENBU extends the genome browser concept by integrating advanced, on-demand data processing and analysis with interactive visualization optimized for comparison across 1000s of experimental samples.

A key difference between ZENBU and previous tools is the ability to dynamically combine thousands of experimental datasets in an interactive visualization environment through linked genome location and experimental signal and expression views. This allows scientists to compare their own experiments against the over 6000 ENCODE and FANTOM consortium datasets currently loaded into the system. ZENBU also provides a data security and collaboration system, allowing users to upload 100s-1000s of experimental data files (BAM, GFF, BED, tab-text files), create analysis views, and share them with a selected group of collaborators.

ZENBU can easily work with both transcriptomics and epigenomics data sets providing raw data visualization, signal clustering, filtering, overlap analysis, gene expression analysis, and enrichment analysis as some examples. Currently we are expanding ZENBU’s capabilities into interactive whole genome analysis by leveraging ZENBU's rich experimental metadata abilities along with ZENBU’s abilities to manipulate large collections of experimental data files simultaneously.

ZENBU is freely available for use on the web and for installation in individual laboratories. The tool can be accessed or downloaded from http://fantom.gsc.riken.jp/zenbu/.

← Back

C-TERMINOME: AN APPLICATION TO INVESTIGATE C-TERMINAL MINIMOTIFS IN HUMAN PROTEINS

Surbhi Sharma[1,2], Oniel Toledo[1,2], Kenneth F Lyon[1,2], Steven Brooks[1,2], Roxainne P David[1,2], Justin Limtong[1,2], Jackyln Newsome[1,2], Nemanja Novakovic[1,2], Sanguthevar Rajasekaran[3], Vishal Thapar[4], Sean Williams[1,2], Martin R Schiller[1,2]

[1]University of Nevada, Las Vegas,, School of Life Sciences, Las Vegas, NV, [2]University of Nevada, Las Vegas,, Nevada Institute of Personalized Medicine, Las Vegas, NV, [3]University of Connecticut, Department of Computer Science and Engineering, Storrs, CT, [4]Massachusetts General Hospital, Department of Pathology, Boston, MA

The C-termini of proteins often possess minimotifs (also known as short linear motifs) that help in regulating protein functions. Minimotifs are 2-15 amino acids long contiguous peptide sequences with a known function in at least one protein. These functions include binding to other molecules, trafficking proteins to specific sub-cellular compartments, and covalently attaching small molecules to the proteins. Our analysis of ~550,000 minimotifs from the Minimotif Miner 3.0 database revealed 3,593 C-terminal minimotifs present in the human proteome, representing ~13% of all human genes. We asked if the remaining 87% of human genes also express proteins with a functional C-terminus. We designate this area of research as the C-terminome. Our hypothesis is that most human proteins have a functional C-terminus. To test our hypothesis, we used used sequence based prediction of new functions on the C-termini of proteins. Functions were predicted for 27,546 sequences based on minimotif consensus sequences, and 867 minimotifs were inferred based on the experiments done in the rodent proteome. We identified highly over represented novel consensus sequences in the human proteome. A subset of these consensus sequences were experimentally tested, identifying potential binding partners. The information has been consolidated into the C-terminome database and web-system where users can mine the C-termini for function in proteins of interest. The websystem can be browsed and searched for C-terminal minimotifs and proteins for their functionality. Weblink: http://cterminome.bio-toolkit.com/

DEEP LEARNING REGULATORY SEQUENCE DRIVERS OF CHROMATIN ACCESSIBILITY DYNAMICS DURING CELLULAR REPROGRAMMING

Anna Shcherbina[1], Avanti Shrikumar[2], Glenn Markov[3], Thach Mai[3], David M Burns[3], Helen Blau[3], Anshul Kundaje[3]

[1]Stanford University, Biomedical Data Science, Stanford, CA, [2]Stanford University, Computer Science, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA

Cellular reprogramming to a pluripotent state has garnered intense interest in recent years. However, there is a paucity of insights into the earliest molecular regulatory factors activated within hours of reprogramming. Although much has been learned from several time courses of reprogramming to pluripotency using the iPSC reprogramming model, limited mechanistic insight derive from inefficient asynchronous reprogramming, clonal expansion, cell death, protocols lasting several weeks, or a selection of only a subset of cells. To observe earlier events in reprogramming, we use cell fusion of human fibroblasts and mouse embryonic stem cells resulting in synchronous, high efficiency (70%) and rapid (48 hours) reprogramming of the fused heterokaryon cells. We performed ATAC-seq experiments at four early time points and observed extensive changes in chromatin accessibility during reprogramming.

In order to decipher key regulatory factors and their combinatorial binding patterns that drive chromatin dynamics, we developed a multi-task, deep convolutional neural network that accepts raw DNA sequence underlying chromatin accessible sites as an input to predict chromatin accessibility changes across the time course. We used a novel hybrid approach that allows for discrimination of accessible sites from inaccessible sites at each time point (binary classification) while also regressing on the continuous ATAC-seq signal at accessible sites. The deep learning architecture described above achieved mean auPRG=0.77 (auROC=0.85) on the classification tasks. The algorithm was benchmarked against the Basset deep learning architecture [2], which achieved a lower mean auPRG = 0.70. The performance (Spearman correlation) of the regression tasks ranged from 0.5 to 0.7.

Finally, we developed a novel interpretation method known as DeepLIFT[3] to compute the predictive importance of each nucleotide in every ATAC-seq peak across the time course. We used DeepLIFT to identify predictive time-point specific sequence patterns representing putative transcription factor binding motifs as well as complex homotypic and heterotypic motif grammars involving several known and novel reprogramming factors including AP1, OCT4, SOX2, NANOG, TEAD1 driving chromatin dynamics. Our analyses, suggest several novel factors that could be used to improve efficiency and fine control of cellular reprogramming.

References

[1] Brady JJ, Li M, Suthram S, Jiang H, Wong WH, Blau HM. (2013) Early role for IL-6 signalling during generation of induced pluripotent stem cells revealed by heterokaryon RNA-Seq, Nature Cell Biology, doi:10.1038/ncb2835.

[2] Kelley DR, Snoek J, Rinn J (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res doi:10.1101/gr.200535.115.

[3] Shrikumar A, et al (2015). Not just a black box: learning important features through propagating activation differences. Biorxiv (arXiv:1605.01713).

← Back

MULTI-OMIC NETWORK ANALYSIS (MONA), A NEW TOOL FOR INTEGRATIVE ANALYSIS

Paul A Stewart[1], Adam L Borne[1], Muhaimen Shamsi[1], Steven A Eschrich[2], Ann Chen[2], Eric B Haura[1]

[1]Moffitt Cancer Center, Thoracic Oncology, Tampa, FL, [2]Moffitt Cancer Center, Biostatistics and Bioinformatics, Tampa, FL

Integrative analyses can give a synergistic view of biological processes since they allow for the simultaneous utilization of multi-omic datasets. However, complexity, technological and computational limitations, small sample size, and a large number of biological variables make integrative analyses very challenging. Here, we present Multi-Omic Network Analysis (MONA), a network-based tool for integrative omic analyses. MONA uses a greedy search algorithm to build differentially expressed subnetworks using experimental data from different molecular types combined with known molecular interactions mined from the STRING and STITCH databases. The MONA search algorithm can be weighted by the confidence of the molecular interactions so that false positive networks are less likely to be identified. MONA uses nonparametric, rank-based scoring, allowing it to handle multi-omics data from separate experiments that differ in scale. MONA assigns p-values to differentially expressed networks by random sample permutations, and false discovery rate is controlled by the Benjamini-Hochberg procedure. MONA is implemented in Galaxy, a widely used open source and web-based platform for biomedical research. MONA can output resulting differentially expressed subnetworks as tables or as XGMML for direct import into Cytoscape. We have used MONA to analyze proteomics and metabolomics data on different histological types of lung cancer, and MONA was able to identify significantly differentially expressed subnetworks related to glycolysis and nucleotide metabolism, which may represent novel vulnerabilities in small cell lung cancer. The MONA framework was designed with extensibility in mind, so additional databases and new molecular types can be added with minimal effort. Future directions include support for different search algorithms (co-expression; Bayesian) and support for post-translational modifications from proteomic experiments.

← Back

GENOGAM: GENOME-WIDE GENERALIZED ADDITIVE MODELS FOR CHIP-SEQ ANALYSIS

Georg Stricker*[1,2], Alexander Engelhardt*[1], Daniel Schulz[1], Matthias Schmid[3], Achim Tresch[4,5], Julien Gagneur[1,2]

[1]Gene Center, Ludwig-Maximilians-University Munich, Biochemistry, Munich, Germany, [2]Technical University Munich, Bioinformatics, Munich, Germany, [3]University Hospital Bonn, Institut für Medizinische Biometrie, Informatik und Epidemiologie, Bonn, Germany, [4]University of Cologne, Institute for Genetics, Cologne, Germany, [5]Max Planck Institute for Plant Breeding Research, Department of Plant Breeding and Genetics, Cologne, Germany

Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) is a widely used approach to study protein-DNA interactions. To analyze ChIP-Seq data, practitioners are required to combine tools based on different statistical assumptions and dedicated to specific applications such as calling protein occupancy peaks or testing for differential occupancies. Here, we present GenoGAM (Genome-wide Generalized Additive Model), which brings the well-established and flexible generalized additive models framework to genomic applications using a data parallelism strategy. We model ChIP-Seq read count frequencies as products of smooth functions along chromosomes. Smoothing parameters are estimated from the data eliminating ad-hoc binning and windowing needed by current approaches. We derived a peak caller based on GenoGAM with performance matching state-of-the-art methods. Moreover, GenoGAM provides significance testing for differential occupancy. Application to a ChIP-Seq dataset showed increased sensitivity by more than 5-fold over existing methods with controlled type I error rate. By analyzing a set of DNA methylation data, we further demonstrate the potential of GenoGAM as a generic analysis tool for genome-wide assays.

COMPENDIUM-WIDE ANALYSIS OF PUBLIC DATA WITH EADAGE REVEALS NOVEL REGULATORY MECHANISMS

Jie Tan[1], Georgia Doing[2], Kimberley A Lewis[2], Courtney E Price[2], Deborah A Hogan[2], Casey S Greene[3]

[1]Geisel School of Medicine at Dartmouth, Department of Molecular and Systems Biology, Hanover, NH, [2]Geisel School of Medicine at Dartmouth, Department of Microbiology and Immunology, Hanover, NH, [3]University of Pennsylvania, Department of Systems Pharmacology and Translational Therapeutics, Philadelphia, PA

Public data provide a rich but challenging-to-analyze source of information about biological systems from diverse aspects. In public compendia, curated metadata can be incomplete or incorrect. The underlying data are complex, noisy, and may contain technical artifacts in addition to biological signals. We need computational approaches that, without reliable metadata, extract biological features while being robust to the types of noise seen in the data. Our recently developed algorithm, ADAGE (Analysis using Denoising Autoencoders for Gene Expression), successfully extracts biological features from public data. With *Pseudomonas aeruginosa*, ADAGE extracted features representing biological states of *P. aeruginosa* (such as strain variation and response to oxygen abundance) from a complete collection of unlabeled public transcriptomic data generated by different labs under various experimental conditions. Our new ensemble-ADAGE (eADAGE) technique improves the algorithm's robustness. Despite being built without curated pathway information, models built by eADAGE cover significantly more KEGG pathways than ADAGE models and capture these pathways more precisely. eADAGE models constructed from a compendium can be used for individual-experiment analyses, cross-experiment comparisons, and compendium-wide analyses. We performed a compendium-wide analysis of media response. Most experiments use a single medium, so the only large-scale approach to identify the effect of media in these data is to look across multiple experiments. Our analysis identified a gene expression signature of the response to low-phosphate in *P. aeruginosa* grown in multiple media that cut across experiments. This signature revealed a previously undiscovered role of KinB: at certain phosphate concentrations, the data suggested that it regulated phosphate acquisition by modifying the activity of the canonical effector PhoB. We experimentally validated KinB's role in the regulation of PhoB. We also performed a kinase screen for this function to evaluate its specificity and found that no other kinases shared this functional role. Our eADAGE method performs unsupervised integration of noisy public data, enables cross-cutting analyses that identify transcriptional responses common across multiple experiments, and these cross-compendium analyses reveal regulatory mechanisms that are only evident when multiple distinct experiments are examined simultaneously.

← Back

NOVEL ROLES FOR RNA REVEALED FROM *IN VIVO* RNA STRUCTUROMES

Yin Tang[1,2], Anton Nekrutenko[1,3], Philip C Bevilacqua[1,3,4,5], Sarah M Assmann[1,2,5]

[1]Pennsylvania State University, Bioinformatics and Genomics Graduate Program, University Park, PA, [2]Pennsylvania State University, Department of Biology, University Park, PA, [3]Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA, [4]Pennsylvania State University, Department of Chemistry, University Park, PA, [5]Pennsylvania State University, Center for RNA Molecular Biology, University Park, PA

RNA can fold into secondary and tertiary structures, which are important for RNA regulation of gene expression. We have developed Structure-seq (Ding *et al*., 2014; Ding *et al*., 2015), which can provide genome-wide RNA structural information at nucleotide resolution in living cells, and applied this methodology to the model plant species Arabidopsis. Several interesting biological observations have been revealed from our *in vivo* RNA structurome, including a triplet structural reactivity pattern throughout the CDS region and an RNA secondary structural pattern related to alternative polyadenylation and splicing (Ding *et al*., 2014). In addition, we found a relationship between RNA structure and the encoded protein structure (Tang *et al*., 2016), wherein regions of individual mRNAs that code for protein domains (or ordered regions) generally have significantly higher structural reactivity than regions that encode protein domain junctions (or disordered regions). We have developed StructureFold (Tang *et al*., 2015) to facilitate the bioinformatic analysis of high-throughput RNA structure profiling data, which is now a component of the PSU Galaxy platform (https://usegalaxy.org).

We have recently applied Structure-seq to rice (*Oryza sativa*), the most widely consumed staple food for humans, to investigate changes in RNA structure induced by heat stress. Significantly elevated average structural reactivity was observed at 42 °C on all mRNA regions as compared to 22 °C. Among these regions, 3' UTRs saw the most significant elevation of average structural reactivity at high temperature, and 5' UTRs saw the largest change in average structural reactivity between temperatures, indicating regulatory roles of RNA structure within these regions.

References

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C. and Assmann, S.M. ***Nature*** 2014;505(7485):696-700.
Ding, Y., Kwok, C.K., Tang, Y., Bevilacqua, P.C. and Assmann, S.M. ***Nat Protoc*** 2015;10(7):1050-1066.
Tang, Y., Assmann, S.M. and Bevilacqua, P.C. ***J Mol Biol*** 2016;428(5 Pt A):758-766.
Tang, Y., Bouvier, E., Kwok, C.K., Ding, Y., Nekrutenko, A., Bevilacqua, P.C. and Assmann, S.M. ***Bioinformatics*** 2015;31(16):2668-2675.

← Back

MAKING RESEARCH SOFTWARE MORE ROBUST

Morgan Taschuk[1], Greg Wilson[2]

[1]Ontario Institute for Cancer Research, Informatics and Biocomputing, Toronto, Canada, [2]Software Carpentry, Toronto, Canada

Bioinformatics software is often written by a single person who is both developer and scientist, intent on producing insightful, novel results for their own projects. Other people may want to use their methods, but much of academic software is written in an ad hoc, exploratory style that is incompatible with reuse. Often, the end user will need to spend a considerable amount of time either reading code or in communication with the original developer(s) in order to get the software running. More times than we like to admit, we will simply give up on the software and either write our own or find something else that accomplishes the same purpose. This practice is detrimental to creating reproducible results, to the pace of scientific research, and to sharing practical knowledge within our community. This familiar problem has spawned a range of suggested solutions, from virtual machines and containerization [1], to more training for bioinformaticians [2], to reform of scientific reward [3]. All of these solutions require time and money.

As an alternative, we present a set of guidelines to encourage code reuse that can be applied to any language, library, or environment for both open-source and closed-source software. The recommendations include: creating a minimum set of documentation with a good README and command-line usage information; avoiding the use of root privileges; guidelines on hard-coding, config files, and command line arguments; structuring test sets and scripts; setting version numbers; and making dependencies and old releases available. By following these guidelines, bioinformaticians can smooth the transition from exploratory code to software that is intended for publication, production environments, and general reuse by the scientific community.

[1]: Howe, B. *Computing in Science & Engineering* 14:4, 36-41 http://dx.doi.org/10.1109/MCSE.2012.62 (2012).
[2]: Lawlor, B. & Walsh, P. *Bioengineered* 6:4, 193-203 http://dx.doi.org/10.1080/21655979.2015.1050162 (2015).
[3]: Prins, P. et al. *Nature Biotech.* 33:7 686-687 http://dx.doi.org/10.1038/nbt.3240 (2015).

PLANT REACTOME: A RESOURCE FOR COMPARATIVE PLANT PATHWAY ANALYSIS

Sushma Naithani[1], Justin Preece[1], Peter D'Eustachio[2], Justin L Elser[1], Parul Gupta[1], Antonio F Mundo[3], Joel Weiser[4], Marcela K Tello-Ruiz[5], Lincoln Stein[4], Doreen Ware[4,5], Pankaj Jaiswal[1]

[1]Oregon State University, Department of Botany & Plant Pathology, Corvallis, OR, [2]New York University, School of Medicine, New York, NY, [3]European Molecular Biology Laboratory - European Bioinformatics Institute, EMBL-EBI, Hinxton, United Kingdom, [4]Ontario Institute of Cancer Research, OICR, Toronto, Canada, [5]Cold Spring Harbor Laboratory, Plant Bioinformatics, Cold Spring Harbor, NY, [6]USDA-ARS, Robert Holley Center, Ithaca, NY

The Plant Reactome database (http://plantreactome.gramene.org) hosts metabolic, genetic and signaling pathways in several model and crop plant species. The Reactome data model organizes gene products, small molecules and macromolecular interactions into reactions and pathways in the context of their subcellular location to build a systems-level framework of a plant cell. A pathway browser offers visualizations of these pathways, reactions, and other components in a dynamic network of navigable diagrams. The Plant Reactome database features *Oryza sativa* (rice) as a reference species, built by importing the RiceCyc metabolic network and curating new metabolic, signaling and genetic pathways. The database now contains 222 rice reference pathways and orthology-based pathway projections for 62 plant species. Plant Reactome allows users to i) compare pathways across various plant species; ii) query and visualize curated baseline and differential expression data available in the EMBL-EBI's Expression Atlas in the context of pathways in the Plant Reactome; and iii) analyze genome-scale expression data and conduct pathway enrichment analysis to enable researchers to identify pathways affected by the stresses or treatments studied in their data sets. Plant Reactome links out to numerous external reference resources, including the gene-pages of Gramene, Phytozome, SoyBase, Legume Information System, PeanutBase, UniProt, as well as ChEBI for small molecules, PubMed for literature supported evidences, and GO for molecular function and biological processes. Users can access and download Plant Reactome data in various formats from the web site and via APIs. This project is supported by the Gramene database award (NSF IOS-1127112) and the Human Reactome award (NIH: P41 HG003751, ENFIN LSHG-CT-2005-518254, Ontario Research Fund, and EBI Industry Programme).

← Back

INFRASTRUCTURE AND DEVELOPMENT OF THE exRNA VIRTUAL BIOREPOSITORY

William Thistlethwaite[1], Sai Lakshmi Subramanian[1], Bob Carter[2], Brittney Miller[2], Javier Figueroa[2], Fred Hochberg[2], Ryan Kim[2], Johnny Akers[2], Douglas Galasko[3], Matt Huentelman[4], Kendall Jensen[4], Rebecca Reiman[4], Jorge Arango[5], Yashar Kalani[6], Julie Saugstad[7], Theresa Lusardi[8], Joseph Quinn[9], Aleksandar Milosavljevic[1], Andrew Jackson[1], Neethu Shah[1], Aaron Baker[1], Sameer Paithankar[1], Matthew Roth[1], Betty Lind[1]

[1]Baylor College of Medicine, Molecular and Human Genetics, Houston, TX, [2]UC San Diego, Neurosurgery, San Diego, CA, [3]UC San Diego, Neurosciences, San Diego, CA, [4]Translational Genomics Research Institute, Neurogenomics, Phoenix, AZ, [5]Barrow Neurological Institute, Neurosurgery, Phoenix, AZ, [6]University of Utah School of Medicine, Department of Neurosurgery, Salt Lake City, UT, [7]Oregon Health and Science University, Anesthesiology and Perioperative Medicine, Portland, OR, [8]Oregon Health and Science University, Computational Biology, Portland, OR, [9]OHSU School of Medicine, Neurology, Portland, OR

The Data Management and Resource Repository (DMRR) and Resource Sharing Working Group, as part of the Extracellular RNA Communication Consortium (ERCC), are currently developing an exRNA virtual biorepository (EVB) for convenient biosample tracking and sharing within the scientific community. The EVB follows a hub-node model, where the hub functions as a central entity that communicates with each participating institution and returns information to the researcher about available biosamples. Genboree REST API infrastructure will allow the hub to seamlessly gather sample metadata from nodes using Linked Data technologies. All front end UIs will be implemented using a web-based Redmine (Ruby on Rails) plugin.

The EVB Hub's portal page will contain a Sample Request UI where researchers can search and filter available samples by various metadata properties like biofluid type, condition, and donor demographics. The page will then display a results grid that shows all matching samples, and each listing will have a link to a full page report where users can read more about a given sample, including information about its associated study and contact information for the sample submitter. Users can then place requests for samples of interest.

All metadata documents are stored in the nodes. Node members will add, edit, and view documents using the node management tools. Members will add new documents by either using the node's web-based Data Entry UI or by uploading documents in bulk (in tabbed format).

The first use case for the EVB is currently underway and involves the evaluation of cerebrospinal fluid (CSF) for the diagnosis of various brain-related conditions. Involved institutions include UCSD, TGen, OHSU, and BNI, and each institution will have its own node in the hub-node infrastructure described above.

TRAINING A SOMATIC MUTATIONS CALLER WITH DEEP LEARNING AND SEMI-SIMULATED DATA.

Remi C Torracinta[1], Fabien Campagne[1,2,3]

[1]Weill Cornell Medicine, The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, New York, NY, [2]Weill Cornell Medicine, Clinical Translational Science Center, New York, NY, [3]Weill Cornell Medicine, Department of Physiology and Biophysics, New York, NY

A number of approaches have been developed to rank and filter genomic sites. Ad-hoc methods are initially useful to quickly identify sites with sought-after characteristics, but they are often supplanted by probabilistic methods that provide more robust performance across ranges of datasets. Probabilistic methods have been traditionally developed from statistical principles, by modeling the sources of noise in an assay and the expected null distribution of the signal. The process of developing a statistical model for a new assay can require months of effort from a skilled statistician. Here, we explored whether probabilistic ranking and filtering models can be trained from data, rather than developed from first principles.

We tested this paradigm by developing a model to estimate the probability that a genomic site harbors a somatic mutation in a sample, given aligned RNA-Seq data for matched germline and somatic samples.

We trained deep feed-forward neural networks with semi-simulated data. Semi-simulated datasets are constructed by planting somatic mutations in real datasets where no mutations are expected. Using semi-simulated data provides the millions of training examples necessary for training deep neural networks with data representative of the sources of variability encountered in these datasets. Models were trained with early stopping and frozen. When evaluated on independent datasets the frozen model achieved an AUC of 0.95.

When tested on semi-simulated exome DNA datasets produced with the Haloplex assay, we find that the models trained on RNA-Seq data remain predictive (sens ~4.0 & spec ~0.9 at cutoff of P>=0.9), albeit with lower overall performance (AUC=0.68). Careful examination of the largest errors reveals that the model learned characteristics of the RNA-Seq assay that does not translate to the Haloplex enzymatic cleavage protocol, which have also confused human analysts when they first analyzed the data from the Haloplex assay (read de-duplication must not be performed for this protocol because most reads stack at sites of enzymatic cleavage). Furthermore, training models directly on Haloplex data, for matched germline/somatic or trio somatic experimental designs and testing these models on independent data strongly suggests that trained probabilistic models can adapt to specific assays and experimental designs (test AUC=0.99 for both matched design and trio design).

The key advantage of this new paradigm is that models can be trained and specialized for specific assays and/or experimental designs, while retaining the ability of experts to contribute to the model by informing the design of the data simulator.

Deep learning probabilistic models for calling somatic variations will be offered in version 3 of the open-source Goby framework (http://goby.campagnelab.org).

VOLT: CLOUD BASE PURE JAVA NGS MAPPING SOFTWARE AND ANALYSIS PIPELINE.

Hiroki Ueda[1], Hiroyuki Aburatani[2]

[1]Fujitsu Limited, Bio-IT R&D Office, Healthcare Systems Unit, Tokyo, Japan, [2]The University of Tokyo, Research Center for Advanced Science and Technology, Tokyo, Japan

We present the software “Volt” the pure Java mapping software for next generation sequencing (NGS) analysis, and genotyper for germline and somatic mutation call which run on cloud environment. As the NGS data size has expanded, data analysis usually requires distributed processing in large scale, However, most software has no mean of direct integration to the standard distributed computing framework Hadoop/YARN, so, most of the conventional NGS pipeline exchange data between software via disk IO, that creates a large bottleneck in data processing, in addition to the unbalanced or unautomated distributed processing. To address this problem, Volt is made using pure Java and integrated directory to the Hadoop/YARN using Java API. By this software design, Volt runs on any Hadoop/Yarn cluster without additional environmental setting, nor software installation, and achieving scalable computation up to 300 nodes per sample. Also, each process is optimized using multi-threads process that enables effective use of many core CPU environments. The algorithms consist of five parts that are, 1) read mapping, 2) realignment, 3) recalibration, 4)pileup, 5) annotation. For read mapping, we have implemented new software which is based on space seed and hash DP algorithm furthered with alt region aware graph reference. The speed and accuracy is comparable to BWAMEM and novoalign (when compared in standalone environment). For post processing of mapped reads, we have and combined in house software and Pure Java Open source software, SRMA for realignment, SnpEff for annotation, ADAM for Columnar data storage. Currently, the system is tested using 100 core computer nodes where 15GB Exome analysis takes 40minitue, and 100GB Whole Genome analysis takes 8 hours to finish. Additional performance testing result will be shown at the meeting. Volt is available as open-source software at http://sourceforge.net/projects/voltmr/files/

SUPER-SCAFFOLDING OF LARGE EUKARYOTIC GENOMES WITH SINGLE MOLECULE MAPS

Davide Verzotto[1], Audrey S Teo[2], Luka Sterbic[1,3], Burton K Chia[1], Mile Sikic[3,4], Axel M Hillmer[2], Niranjan Nagarajan[1]

[1]Genome Institute of Singapore, Computational and Systems Biology, Singapore, Singapore, [2]Genome Institute of Singapore, Cancer Therapeutics & Stratified Oncology, Singapore, Singapore, [3]University of Zagreb, Department of Electronic Systems and Information Processing, Zagreb, Croatia, [4]Bioinformatics Institute, Biomolecular Modeling and Design Division, Singapore, Singapore

Obtaining highly accurate and contiguous sequence assemblies is a key goal for improved characterization of individual and reference genomes, particularly in uncovering complex rearrangements and amplifications, and in understanding the function of repetitive sequences and the non-coding genome.

We introduce CROM, a super-scaffolding approach that directly utilizes raw single molecule optical maps to cost-effectively reconstruct arm-level assemblies of large eukaryotic genomes. CROM is based on the scaffold overlap-extension-bridge paradigm and combines a sensitive and precise map-to-sequence alignment with a local error minimizing consensus stage and a combinatorial approach to scaffold bridging. CROM's flexible framework also allows for the detection of alternative haplotypes.

Benchmarking CROM against state-of-the-art methods showed that it provides significant improvements in contiguity (20–212%, NG50) while maintaining high-quality assemblies. We show that the application of CROM enables spanning of human genome hotspots, improved placement of previously unlocalized human reference scaffolds and the best *de novo* human assembly to date (43 Mb scaffold NG50 corresponding to 73% of GRCh38 NG50) using only two single-molecule clone-free technologies. CROM is the first method that can successfully exploit different genome-mapping technologies, allowing for the reuse of datasets for improving draft genomes.

← Back

KMER2VEC: UNSUPERVISED EMBEDDING OF REGULATORY DNA SEQUENCES

Rahul Mohan[1], Michael Wainberg[2], Nasa Sinnott-Armstrong[3], Anshul Kundaje[2,3]

[1]Bellarmine College Preparatory School, San Jose, CA, [2]Stanford University, Computer Science, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA

The human genome is richly structured, but it is unclear how much of this structure is captured by existing functional annotations and these annotations are not consistently available across cell types and tissues. We present kmer2vec, an unsupervised learning method based on the concept of 'word embeddings' for understanding human language, which can leverage the unannotated structure encoded in the genome to improve performance on a wide variety of supervised training tasks, including functional variant detection and peak classification. kmer2vec assigns a vector to each k-mer of a given size, based on the k-mer's pattern of co-occurrences with other k-mers in a list of input sequences. Supervised classifiers can then be trained on this high-level vector representation of the sequence instead of the raw sequence. As a result of the rich and compact representation the embedding provides, it particularly excels at tasks where only a small number of annotated examples are available, such as functional variant detection. Across a wide variety of variant tasks, kmer2vec consistently exceeds the performance of both deltaSVM, a state-of-the-art supervised method, and Eigen, a state-of-the-art unsupervised method, with deltaSVM only performing better on the dsQTL task it was trained on and Eigen performing worse on every task. Finally, we develop an interpretation system for models trained with kmer2vec that indicates how each subregion in an input sequence contributes to the final prediction. We provide genome-wide kmer2vec scores for numerous common functional genomics tasks, along with the framework necessary to train and interpret new models on any task of interest.

← Back

SCIAPPS: A DISTRIBUTED CYVERSE SYSTEM FOR CLOUD COMPUTING

Liya Wang[1], Zhenyuan Lu[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,
Ithaca, NY

To overcome the challenges of moving massive amount data among remote storage/compute nodes, we built a federated CyVerse platform, which is constituted by several storage and compute systems physically located at CSHL, and fully integrated with CyVerse's Cyber-infrastructure through Agave API and iRODS. On top of the federated system, we designed and implemented a web portal using ReactJS, SciApps.org, that automatically populates the interfaces of Agave apps, handles data uploading and browsing into remote and local data storage system, supports job submission and monitoring, automates complicate scientific workflows, and provides various ways to visualize analysis results. For example, we integrated genome browsers, Biodalliance and JBrowse, for visualizing alignments, variants, and genome annotation results. The platform allows us to keep large amount of data processed locally with existing CyVerse apps and pipelines, thus facilitate handling massive amount of data efficiently (by reducing national wide data transfer) and easy exchanging of apps/pipelines across systems. The platform is designed to facilitate cloud computing and sharing of data, apps, storage, and computing systems, and it supports fast exchange of data with the CSHL hosted Gramene/Ensembl plants database.

Development of the federated system is supported by an NSF award (DBI-1265383).

← Back

SNAPTRON: ENABLING FLEXIBLE QUERYING AND EXPLORATORY ANALYSIS OF HUMAN RNA SPLICING

Christopher Wilks[1,2], Phani Gaddipati[3], Abhinav Nellore[1,2,4], Ben Langmead[1,2,4]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, [2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

As more and larger genomics studies appear, public archives are filling with large summarized datasets. A summarized dataset often consists of a vast number of measurements at individual bases (coverage, genotype), across genomic intervals (gene count, junction count), or across sets of genomic intervals (transcript abundance). Users will cross-reference these measurements with annotations that associate metadata with genomic positions and intervals, e.g. gene annotations, mappability tracks and repetitive element tracks. Further, users will simultaneously query many datasets at once, often using free-text sample metadata to first arrive at a relevant subset (e.g. all brains) or set of subsets (e.g. all colon cancer/normal pairs) of samples. Enabling scientists to query this data requires systems that combine and index diverse kinds of data, including numeric and free text data, data associated with genomic positions, intervals, or sets of intervals, annotation data, and sample metadata.

Snaptron is a genomic search engine that leverages R-tree (Tabix) and inverted indices (Lucene) to facilitate low-latency queries over more than 81 million splicing events (introns) across 49,849 publicly available human RNA-Seq samples aligned by the Rail-RNA pipeline. Further it allows large result sets from these data to be filtered by additional search predicates in conjunction with genomic interval and metadata specific queries. Snaptron provides a flexible set of interfaces which can be run locally, via a set of REST web services, through the use of a graphical user interface, or by means of a provided command line client all of which utilize the same search interface. Snaptron also provides a simple, general ranking function, the junction inclusion ratio, to compare splicing patterns across samples for user-defined sets of junctions. These features enable the targeted exploration of important biological phenomena at the same time requiring minimal investment from the user in terms of software dependencies and query knowledge. We outline examples of pertinent biological questions which can be explored using Snaptron, including the effects of an alternative transcription start site in the ALK gene recently uncovered by Wiesner et al. in melanoma and the investigation of repetitive element loci spliced into existing transcripts.

REQUIRED PARAMETERS – WHAT DOES IT TAKE TO BRING BIOINFORMATICS INTO THE CLASSROOM AT THE NATIONAL LEVEL?

Jason Williams, Mona Spector, Cornel Ghiban, Joslynn Lee, David Micklos

Cold Spring Harbor Laboratory, DNA Learning Center, Cold Spring Harbor, NY

Bioinformatics is as essential to modern biology as the microscope. At the same time, it is widely acknowledged that most biologists (and biology students) face a training gap. Survey data and published articles make it clear that the majority of biologists don't have access to skills needed to analyze data, or to be equal and informed collaborators with bioinformatics domain experts. This training gap impacts undergraduate education and consequently influences researchers over the course of their academic preparation.

While individual (and usually well-resourced) institutions have capabilities to develop their own programs, we address the problem of scaling training and resources to effect national impact. We will share results from the "RNA-Seq for the Next Generation Project" (http://www.rnaseqforthenextgeneration.org/), a program that trained faculty (mostly at small colleges/primarily undergraduate institutions) to explore bioinformatics by engaging students in development and analysis of RNA-Seq experiments.

We will outline our strategy for developing a national-scale project, including:
1) development of robust, classroom-usable cyberinfrastructure tools for large datasets (using CyVerse, DNA Subway, and the Agave API); 2) Development of student-centered workflows that include both wet-lab and bioinformatics applications; 3) Faculty training strategy and development of course-based research lesson plans.

Examples drawn from the 84 participating faculty will illustrate the challenges and successes of incorporating substantial bioinformatics training into the classroom. Taken together, these experiences will be highly informative to efforts to bring bioinformatics into classrooms at a variety of institutional settings.

← Back

CHARACTERIZATION OF REGULATORY ROLES OF CTBP IN BREAST CANCER USING INTEGRATED GENOMIC AND PROTEOMIC ANALYSIS

Tingfen Yan[1], Jung S Byun[2], Dae Ik Yi[2], Samson Park[2], Genqing Liang[2], Mohamed Kabbout[1], Kevin Gardner[1,2]

[1]National Institutes of Health, NIHMD, Bethesda, MD, [2]National Institutes of Health, NCI, Bethesda, MD

The C-terminal Binding Proteins (CtBP1 and CtBP2) coordinate the assembly of multiple different epigenetic regulatory complexes. The CtBPs are NADH-dependent transcriptional repressors that are post-transcriptionally stabilized in the presence of carbohydrate excess, thus providing a linkage between metabolism and epigenetic regulation. CtBP expression is elevated in breast and many other epithelial cancers. Genome-wide analysis of CtBP binding by ChIP-seq in the estrogen positive breast cancer cell line, MCF7, revealed that CtBP-bound promoters and enhancers are associated with numerous genes that are involved in epithelial to mesenchymal transition, stemness and genome instability pathways. To fully understand CtBP's regulatory roles in breast cancer, we are currently conducting an integrated genomic and proteomic analysis of CtBP expression in breast cancer. To this, we are employing gene set enrichment analysis (GSEA) that distinguishes patient tumors or tumor cell lines with high expression levels of CtBP from those with low levels of CtBP. The results in which CtBP1 and CtBP2 levels are defined by RNA-seq and protein abundance by immunohistochemistry, respectively, are compared. These results are then integrated with a parallel analysis of breast epithelial and breast cancer cell lines (MCF-7, MDA-MB-231, and MCF10A) in which CtBP1 or CtBP2 is either overexpressed or silenced. To determine which model of CtBP over-expression in cell lines most closely resembles what is seen in tumors, the GSEAs are compared and correlated. This data is then combined with cellular network analysis using the ARACANe algorithm to infer gene-gene direct interactions based on expression profiles in both patients tumor and breast cancer cell lines stratified by differential expression of CtBP1 and CtBP2. The GSEA comparisons among three cell lines with over-expressed CtBPs indicate that CtBPs play important regulatory roles in the progression and aggressiveness of breast cancer by promoting epithelial to mesenchymal transition and inflammatory response, as well as in drug resistance by activating transcription factors responsible for multidrug resistance and repressing DNA repair. The GSEA overlaps between all the three cell lines and tumor patient samples with overexpressed CtBPs reveal that the non-tumorigenic MCF10A breast epithelial cell line most closely resembles patient tumors that overexpress CtBPs, suggesting that the MCF10A cell line may be a more appropriate model to study CtBP's function in breast cancer progression.

← Back

IDENTIFICATION OF ROBUST METABOLOMICS CAUSAL NETWORKS IN OBSERVATIONAL STUDY USING DATA INTEGRATION

Azam Yazdani

UTHealth , Human Genetics Center , Houston, TX

A metabolomics causal network is identified using the genome granularity directed acyclic graph (GDAG) algorithm. The GDAG algorithm extracts information across the genome in a deeper level of granularity using principal component analysis to create strong instrumental variables and identify causal relationships among metabolites in an upper level of granularity. Information from 1,034,945 genetic variants distributed across the genome was used to identify a metabolomics causal network among 122 serum metabolites reliably measured by a combination of gas and liquid chromatography and mass spectrometry.

Causal effect measurement is discussed while confounders and mediators are identified given the network. Good targets for intervention and prediction are identified based on individual properties. Four nodes corresponding to the metabolites leucine, arichidonoyl-glycerophosphocholine, N-acyelyalanine, and glutarylcarnitine may have a high impact on the entire network by virtue of having multiple arrows pointing out and propagated long distances. Five modules are identified through analyzing data using the GDAG algorithm which largely correspond to structural sub-classes of metabolites. The module boundaries are determined using directionality and causal effect sizes.

← Back

TUMOR-SUPPRESSIVE AND -PROMOTING FUNCTIONS OF SUFU IN THE SHH SUBGROUP OF MEDULLOBLASTOMA

Wen-Chi Yin[1,3], Thevagi Satkunendran[1,3], Rong Mo[3], Huayun Hou[1,3], Hayden Selvadurai[3], Sorana Morrissy[3], Michael Wilson[1,3], Peter B Dirks[1,3], Michael D Taylor[2,3], Chi-chung Hui[1,3]

[1]University of Toronto, Department of Molecular Genetics, Toronto, Canada, [2]University of Toronto, Department of Laboratory Medicine and Pathobiology, Toronto, Canada, [3]Hospital for Sick Children, Program for Developmental and Stem Cell Biology, Toronto, Canada

Medulloblastoma (**MB**) is the most common malignant childhood brain tumor. Approximately 30% of MBs arise due to aberrant activity of the Sonic Hedgehog (**SHH**) signaling pathway. In SHH MBs, GLI proteins, particularly the major transcriptional activator GLI2, are highly expressed. Mutations in transmembrane signaling proteins PTCH1 and SMO, believed to unleash GLI activity, are found in 60% of human SHH MBs. The exact roles of the intracellular GLI regulator SUFU in MB tumorigenesis remain poorly understood. *SUFU* mutations are identified in ~30% of infant SHH MBs which represent the difficult to treat subgroup, because radiation and non-specific chemotherapy are both extremely harmful to infants and young children. Limited progress to date is due to the lack of faithful preclinical models and a poor molecular understanding of infant SHH MB.

Strikingly, despite the Gli-inhibitory functions of Sufu, we found that loss of Sufu alone is insufficient to drive MB tumorigenesis in mice due to unsustained Gli activity. We showed that Sufu and Spop, an adaptor of E3 ubiquitin ligase, play central roles in suppressing MB tumorigenesis in mice through regulation of Gli2 protein stability. Using genetic mutants of *Sufu* and *Spop*, we established a novel preclinical mouse model of infant SHH MB with aggressive early-onset tumor formation. By ChIP-Seq and RNA-Seq analyses, we identified more than 800 putative direct transcriptional targets of Gli2 in the *Sufu;Spop* infant MB mouse model, representing the first comprehensive dataset of Gli2 target genes. Our study here provides important insight to the molecular understanding of the transcriptional events underlying SHH MB tumorigenesis and reveals potential new therapeutic targets. We also found loss of Sufu in Ptch1/Smo MB models leads to significantly prolonged survival, reduced Gli2 protein, and low overall pathway activity. Thus, Sufu acts positively to promote MB formation in Ptch1/Smo MB models, and represents a potential therapeutic target for MBs harboring *PTCH1/SMO* mutations.

Further mining of the ChIP-Seq and RNA-Seq data generated in this study and comparison with available human MB data will provide insight to the tumorigenesis of SHH MB and potential therapeutic targets. Since GLI2 is overexpressed in various cancer types, analysis of its target genes may also help define the role of GLI2 in other cancers.

← Back

THE CRITICAL ASSESSMENT OF PROTEIN FUNCTION ANNOTATION: IMPROVING ON THE “STATE-OF-THE-ART”

Naihui Zhou[1], Yuxiang Jiang[2], Sean D Mooney[3], Casey S Greene[4], Predrag Radivojac[2], Iddo Friedberg[5]

[1]Iowa State University, Program of Bioinformatics and Computational Biology, Ames, IA, [2]Indiana University, Department of Computer Science and Informatics, Bloomington, IN, [3]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA, [4]University of Pennsylvania, Department of Systems Pharmacology and Translational Therapeutics, Philadelphia, PA, [5]Iowa State University, Veterinary Microbiology and Preventive Medicine, Ames, IA

The increasing volume and variety of genotypic and phenotypic data is a major defining characteristic of modern biomedical sciences. Assigning functions to biological macromolecules, especially proteins, turn out to be one of the major challenges to understand life on a molecular level. While molecular experiments provide the most reliable annotation of proteins, their relatively low throughput and restricted purview have led to an increasing role for computational function prediction. However, properly assessing methods for protein function prediction and tracking progress in the field remain challenging as well.

The Critical Assessment of Functional Annotation (CAFA) is a timed challenge to assess computational methods that automatically assign protein function. Here we report the results of the second CAFA challenge, and outline the new changes that will take place in the third CAFA this year.

One hundred and twenty six methods from 56 research groups were evaluated for their ability to predict biological functions for 3,681 proteins from 11 species in the second CAFA challenge. These functions are described by the Gene Ontology (GO) and the Human Phenotype Ontology (HPO). CAFA2 featured increased data size as well as improved assessment metrics, especially the additional assessment metrics based on semantic similarity. Comparisons between top-performing methods in CAFA1 and CAFA2 showed significant improvement in prediction accuracy, demonstrating the general improvement of automatic protein function prediction algorithms. It also showed that the performance of different metrics is ontology-specific, which revealed that the different evaluation metrics can be used to probe the nature of protein functions in different biological processes and human phenotypes.

CAFA3 will start September 2016, featuring expanded protein sets for predictions. We are using whole genome screens to generate term-centric tracks for Drosophila melanogaster and Pseudomonas aeruginosa. Additionally, we will use sets of moonlighting proteins, disordered proteins, and prediction of binding sites to further challenge function prediction methods.

← Back

USING A TOP-DOWN APPROACH TO UNDERSTAND THE HUMAN TRANSCRIPTOME

Weizhuang Zhou[1], Russ B Altman[1,2,3]

[1]Stanford University, Bioengineering, Stanford, CA, [2]Stanford University, Genetics, Stanford, CA, [3]Stanford University, Medicine, Stanford, CA

The advent of microarrays and next-gen sequencing technologies has facilitated the simultaneous study of multiple genes in the human transcriptome. While this has provided detailed characterization of cellular behavior, the high dimensionality of the data is not well suited for traditional statistical analysis. Interpreting the information from such experiments has mostly relied on curated gene interaction networks and pathways. Leveraging on the fact that many genes are highly correlated, there are also many who propose the use of dimensionality reduction techniques such as low-rank matrix factorizations. These top-down approaches free us from the limits of current biological knowledge and allow us to study genes in tandem. Principal component analysis (PCA) is a method that is frequently used due to its simplicity and well-understood properties. It has also been argued in recent years that most of the microarray data can in fact be represented by just the first few principal components, suggesting that these data are actually low rank. We present a related approach based on independent component analysis (ICA), and show that the gene modules obtained by our method capture many biological and physiological relationships. Unlike PCA, ICA does not maximize the variance explained in each component, but instead maximizes the statistical independence of each component. This allows the obtained components to be more biologically relevant, while still allowing for dimension reduction. We also show how drug-gene databases, such as the Connectivity Map, can be analyzed with higher statistical power in the lower dimension space defined by our functional modules.

← Back

MSKCC PAPERLESS LAB INITIATIVE (MSK-PLI)– SOFTWARE FOR ANALYSIS OF FRAGMENT AND QUALITATIVE PCR ASSAYS

John Ziegler, Jack Birnbaum, Paulo Salazar, Ivelise Rijo, Nana Mensah, Georgi Lukose, Lissette Fernandez, Kelly Rios, Ahmet Zehir, Maria Arcila

MKSCC, Dept. of Pathology, New York, NY

Diagnostic laboratories are inundated with paper records. Every test, sample, and record increases complexity and makes it more difficult to meet pressing deadlines with accuracy. Furthermore, keeping paper records makes data mining virtually impossible. While the Next-Generation Sequencing (NGS) based assays are inherently software-driven, the traditional non-NGS assays have not benefited from digitization. The Paperless Lab initiative in the Molecular Diagnostics Service (MDS) in the Department of Pathology at Memorial Sloan Kettering Cancer Center (MSKCC) is designed to address this deficiency, creating a set of software systems that store, analyze, and visualize data from thousands of samples across many different assays.

We present a set of web-based applications that enable gathering, mining, and analysis of data from a series of non-NGS assays performed in MDS. The application suite includes tools for organizing experiments, generating sequencer input files, result processing, and data visualization/aggregation. Assays supported include STR/CTTV engraftment testing, fragment testing, and quantitative RNA tests. Laboratory Information Management System (LIMS) is used to track samples after acquisition in the laboratory. Meta-data collected in LIMS is sent to MSK-PLI via RESTful web services where, the samples are gathered and assembled onto a virtual PCR Plate. Users can arrange the samples on the plate and automatically generate the corresponding input file for the pertinent sequencer. After the completion of sequencing, the user uploads the sample results to the web applications, records notes, and views quality-control (QC) results before moving the samples forward for review by pathologists. Result calculations are performed where applicable, and data visualizations are built to allow faster and more accurate analysis. Clinical reports are also generated from the applications. All data is maintained and indexed in a MongoDB implementation. The front-end web view is built utilizing React and Flux architecture. The data retrieval layer is built in Python.

We have used MSK-PLI since February 2016 in the MDS to process over 2500 samples with various assays. Utilizing the software has resulted in a 76% reduction in paper records, decreased manual errors, and increased lab efficiency. These tools and applications may be used independently or in conjunction with existing software solutions. Next steps include the addition of more assays, advanced data visualizations, and the exploration of machine learning and additional data aggregation/classification techniques.

REPRODUCIBLE, PORTABLE, SCALABLE AND OPEN GENOMICS WITH TOIL, ADAM AND DOCKSTORE

Benedict Paten[1], John Vivian[1], Arjun Rao[1], Frank Nothaft[2], Jacob Pfeil[1], Audrey Musselman-Brown[1], Hannes Schmidt[1], Christopher Ketchum[1], Anthony Joseph[2], David Haussler[1], Brian O'Connor[1]

[1]UC Santa Cruz Genomics Institute, Computational Genomics Lab, Santa Cruz, CA, [2]UC Berkeley, AMP Lab, Berkeley, CA

Genomics is in transition. The growth in data—driven by the need for vast sample sizes to gain statistical significance and the explosion of clinical sequencing—is far outpacing Moore’s law. Large projects like The Cancer Genome Atlas have generated petabyte scale datasets that very few groups have the capacity to analyze independently. Looking forward, improvements in the computing technologies will be insufficient to satisfy the community's exponentially growing needs for computing throughput and storage capacity. The cost of computing on the rapidly growing data is compounded by the expanding complexity of genomic workflows. Typically, dozens of programs must be precisely configured and run to reproduce an analysis. The drastic increases in data volume and workflow complexity have created a serious threat to scientific reproducibility. If groups can neither afford to run the computation nor reconstitute the analysis steps, then the genomics community will become crippled.

In collaboration with groups in the Global Alliance for Genomics and Health (GA4GH), we have developed Toil, ADAM and Dockstore to address these concerns. Toil is a system designed to orchestrate massive computational pipelines at a scale previously unheard of within genomics. We describe the analysis of 20,000 RNA-seq samples across 32,000 cores in under four days on AWS at a cost level that makes such analysis possible. Toil supports the emerging workflow standards CWL and WDL ADAM is a genomics platform built in the Apache Spark ecosystem to enable distributed algorithm execution. It enables order of magnitude improvement in speed and cost for data intensive analysis genomics algorithms. Dockstore in a system that enables the sharing of genomic workflows and containers open and precisely. The efforts, and the ecosystem being built by the GA4GH will change big data genomics.

← Back

IHEC DATA PORTAL: A RESOURCE FOR DISCOVERING, ANALYSING AND SHARING EPIGENOMICS DATA

David Bujold[1], David A Morais[2], Carol Gauthier[2], Catherine Côté[1], Maxime Caron[1], Tony Kwan[1], Kuang C Chen[3], Jonathan Laperle[4], Alexei N Markovits[5], Tomi Pastinen[1,6], Bryan Caron[3], Alain Veilleux[2], Pierre-Étienne Jacques[2,4,5], Guillaume Bourque[1,6]

[1]McGill University, McGill University and Génome Québec Innovation Center, Montréal, Canada, [2]Université de Sherbrooke, Centre de calcul scientifique, Sherbrooke, Canada, [3]McGill University, McGill High Performance Computing Centre, Montréal, Canada, [4]Université de Sherbrooke, Département d'Informatique, Sherbrooke, Canada, [5]Université de Sherbrooke, Département de Biologie, Sherbrooke, Canada, [6]McGill University, Department of Human Genetics, Montréal, Canada

The International Human Epigenome Consortium (IHEC) coordinates the production of reference epigenome maps through the characterization of the regulome, methylome and transcriptome from a wide range of tissues and cell types. Some of the initial challenges of the consortium were to define conventions ensuring the compatibility of datasets generated by different IHEC members and to establish an infrastructure enabling data integration, analysis and sharing. The IHEC Data Portal (epigenomesportal.ca/ihec) was developed to address these issues and become the official source to navigate through IHEC datasets. The Portal provides access to more than 7000 reference epigenomic datasets generated from over 600 tissues, which have been contributed by 7 international consortia.

To help maximize the utility of these reference maps, the Portal facilitates the discovery, visualization, analysis, download and sharing of epigenomics data. Datasets quality assessment features have been implemented, including a correlation tool between user selections. A quality control pipeline was also developed, verifying datasets quality by evaluating multiple metrics. An IHEC-wide analysis using this pipeline has demonstrated that even with large inter-consortia variability in methods for library preparation and downstream analysis, and differences in cell types, diseases and other factors, a useful qualitative assessment can be offered to data producers and users.

Other features include bi-yearly persistent releases and publicly-accessible tracks that are now being served directly from the Portal, enabling reliable navigation sessions. Users can also save their Portal sessions permanently, allowing them to obtain an ID that links to their datasets selection and filtering options. These IDs can be shared and are directly citable in papers making use of the IHEC Data Portal datasets. Lastly, a publicly accessible Web API allows users to query available datasets metadata, in both JSON and human-readable formats.

← Back

CHOICE OF REFERENCE GENOME CAN INTRODUCE MASSIVE BIAS IN BISULFITE SEQUENCING DATA

Kasper D Hansen

Johns Hopkins University, Biostatistics, Baltimore, MD

The use of a reference genome is required for using existing short read alignment tools. Such a reference genome is typically different from the sample genome and this difference may introduce mapping bias. Mapping bias have been studied in the context of splicing and gene expression and its impact range from high (splicing) to low (gene expression). Here we study mapping bias in analysis of whole-genome bisulfite sequencing (WGBS) data, the most comprehensive experimental method for profiling DNA methylation. Using data from WGBS of inbred mouse strains we show that the choice of reference genome has a massive impact on the inferred methylation landscape of the samples. It can result in wrongly inferring thousands of small differentially methylated regions and up to gigabases of large differentially methylated regions. We show that these differentially methylated regions can be completely reversed by changing the choice of reference genome. We show that the bias is predominantly, but not exclusively, caused by single nucleotide variants in CpG sites. We develop two strategies for removing the bias using either (1) alignment to personal genomes or (2) knowledge of single nucleotide variants over CpG sites. Finally, we produce a set of 976 bias-free genomic regions which are differentially methylated between strains and show that these regions are enriched for functionally relevant marks such as H3K4me1, H3K4me3 and H3K27ac and depleted in CpG islands.

THE exRNA ATLAS: LINKING DATA, TOOLS, AND COMPUTABLE PATHWAY KNOWLEDGE TO INTERPRET THE FIRST 1000 UNIFORMLY PROCESSED PROFILES OF EXTRACELLULAR RNA FROM HUMAN BODILY FLUIDS.

Sai Lakshmi Subramanian[1], William Thistlethwaite[1], Andrew R Jackson[1], Neethu Shah[1], Aaron Baker[1], Sameer Paithankar[1], Robert Kitchen[2], Kristina Hanspers[3] , Ginger Tsueng[5], Roger P Alexander[4], David J Galas[4], Joel Rozowsky[2], Alexander Pico[3], Andrew I Su[5], Matthew E Roth[1], Mark B Gerstein[2], Aleksandar Milosavljevic[1]

[1]Baylor College of Medicine, Houston, TX, [2]Yale University, New Haven, CT, [3]Gladstone Institutes, San Francisco, CA, [4]Pacific Northwest Diabetes Research Institute, Seattle , WA, [5]The Scripps Research Institute, La Jolla, CA

The Extracellular RNA Communication Consortium, a new NIH Common Fund initiative has created the first exRNA Atlas, a compendium of uniformly processed extracellular RNA profiles obtained by RNA-seq from human plasma, serum, saliva, bile, and cerebrospinal fluids. The profiles include both vesicular (exosomal) RNA as well extra-vesicular protein- and lipid-stabilized RNAs of human and non-human (microbial, animal, fungal, plant) origin. The large diversity and volume of extracellular RNA (exRNA) data, multiplicity of processing, normalization, and analysis methods, and recent but fast-growing pathway knowledge of exRNAs posed a substantial integration challenge. We employed a strategy that combines metadata modeling, biomedical ontologies, cloud computing and crowd-sourcing to process and integrate a diverse set of exRNA profiles into the exRNA Atlas and enable integrative analysis in the context of computable, collaboratively developed exRNA network and pathway knowledge. To enable metadata and knowledge modeling, we developed GenboreeKB, a free open-source Linked Data (RDF, JSON-LD) document-oriented modeling and distributed storage system backed by a MongoDB database and Redmine collaborative framework. GenboreeKB is integrated with Genboree Workbench, a collaborative distributed hybrid (commercial and private) cloud computing front-end that integrates a multiplicity of tools for exRNA analysis in the context of exRNA networks and the computable exRNA community-curated exRNA section of WikiPathways and exRNA gene reports in BioGPS. We demonstrate how this integrative framework enables detection and interpretation of extracellular miRNA perturbations in Alzheimer’s disease and cancer for the purpose of biomarker discovery, understanding etiology, as well as biogenesis and downstream effects of exosomal miRNAs taken up by target cells.

← Back

THE EVOLVING HUMAN REFERENCE GENOME ASSEMBLY DRIVES CHANGES IN DATA MANAGEMENT AND REPRESENTATION

V A Schneider, P A Kitts, A Kimchi, K D Pruitt, K Clark, T D Murphy

National Center for Biotechnology Information, Bethesda, MD

As a modeled representation of the DNA in a cell or an organism, a genome assembly is by its very nature “biological data”. The human reference genome assembly, envisaged as a pan-human genome representation, plays a central role in basic and clinical research, and is the backbone upon which we annotate additional forms of biological data. In order to succeed in these things, not only must the reference genome assembly itself be comprised of high quality data, but it is essential that its information content be represented in such a way that it can be communicated to and accessed by its diverse user base. The high quality standards of the Human Genome Project (HGP) and ongoing curation of the reference genome by the Genome Reference Consortium (GRC) helps ensure the former. However, this curation also makes the reference genome assembly a dynamic data set, which in turn creates challenges in data management, representation and usability. As a member of the GRC, the National Center for Biotechnology Information (NCBI) provides bioinformatics support for this curation effort, and as part of its basic mission, has developed data models, databases and tools for assembly evaluation, representation and use. We will talk about how the dynamic nature of the human reference genome assembly and advances in genomic biology, sequencings and assembly technology are shaping these efforts at NCBI, and present recent developments. We will also describe the impact that changes in assembly data content have for consumers of the reference genome and related NCBI resources. Finally, as genome representations are continuing to evolve, we will discuss how new assemblies and assembly models may impact the data content of the reference and the resources needed to ensure the accessibility of the reference to a broad user community.

← Back

EVALUATING POPULATION STRUCTURE FOR SUBJECTS IN THE DBGAP DATABASE

Yumi Jin, Stephanie Pretel, Anne Sturcke, Moira Lee, Natilia Popova, Wenyu Wu, Stephen Sherry, Michael Feolo

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

Genome-wide association studies (GWAS) have been widely used to identify genetic variants associated with human diseases and other phenotypes. For GWAS analyses, it is essential to take population stratification into consideration to avoid spurious associations resulting from allele frequency differences between different subpopulations.

As of August, 2016, 1,561,367 subjects have been submitted to dbGaP. However, ancestry information is disparate as some dbGaP submissions do not provide subject ancestry information, and furthermore, when the information is provided, it is submitted in many different ways. We have created computer programs to check the variable names and values in the submitted datasets to find the putative variables that contain subject ancestry information. By using these programs, as well as manually checking the datasets, documents, and other information, the dbGaP curatorial staff have identified ancestry values for 816,142 subjects, and marked them for tracking in a relational database.

Many of these submitted subject ancestry values are provided as free text instead of using standard terms. In total there are 444 unique ancestry values for these subjects. We categorized these values into 34 population groups, with 94.1% of the subjects with ancestry values in dbGaP belonging to the following four population groups: European: 575,601; African: 115,565; Hispanic: 41,784; Asian: 34,935.

For the subjects with SNP genotypes submitted to dbGaP, we are able to use genotype data to compute population structure across all samples in dbGaP, and compare it to the study reported ancestry backgrounds. In order to find the subject overlap among different dbGaP studies, genotypes of 10,000 well-separated SNPs were extracted from the submitted genotype datasets and loaded into a genotype fingerprint database. Using the genotype data in the fingerprint database, we resolved the population structures of 610,589 subjects from 230 dbGaP studies using the software fastSTRUCTURE. We also evaluated the prediction precisions of the software to separate different populations with the self-reported ancestry backgrounds using separation power, defined as the percentages of correctly separated subjects out of the maximum separable subjects. The results show that fastSTRUCTURE can distinguish the major population groups from each other very well. Specifically, we observed the following separation powers: 98.1% for Asian and Hispanic, 97.0% for African and European, 95.7% for Asian and African, 84.4% for African and Hispanic, 72.7% for Asian and European, and 38.6% for European and Hispanic.

← Back

FUNCTIONAL ANNOTATION AND VISUALIZATION OF LARGE-SCALE BIOLOGICAL NETWORKS

Anastasia Baryshnikova

Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ

Large-scale biological networks map functional connections between most genes in the genome and can potentially uncover high level organizing principles governing cellular functions. Despite the availability of an incredible wealth of network data, our current understanding of their functional organization is very limited and essentially opaque to biologists. To facilitate the discovery of functional structure and advance its biological interpretation, I developed a systematic quantitative approach to determine which functions are represented in a network, which parts of the network they are associated with and how they are related to one another. This method, named Spatial Analysis of Functional Enrichment (SAFE), detects network regions that are statistically overrepresented for a functional group or a quantitative phenotype of interest, and provides an intuitive visual representation of their relative positioning within the network. Using SAFE, I examined the most recent genetic interaction network from budding yeast Saccharomyces cerevisiae, which was derived from the quantitative growth analysis of over 20 million double mutants. By annotating the genetic interaction network with GO biological process, protein localization and protein complex membership data, SAFE showed that the network is structured hierarchically and reflects the functional organization of the yeast cell at many different levels of resolution. In addition, we analyzed the network using a large-scale chemical genomics dataset and generated a global view of the yeast cellular response to chemical treatment. This view recapitulated the known modes-of-action of chemical compounds and identified a potentially novel mechanism of resistance to the anti-cancer drug bortezomib. These observations demonstrate that systematic integrative annotation of biological networks is a powerful tool for investigating the global wiring diagram of the cell.

← Back

STRUMS: A FLEXIBLE AND INFORMATION-RICH REPRESENTATION OF DNA MOTIFS

Peter M DeFord, James Taylor

Johns Hopkins University, Biology, Baltimore, MD

The position weight matrix (PWM) has long been a useful tool for describing DNA sequence variation in regions such as transcription factor (TF) binding sites. It is difficult, however, to relate the sequence-based representation of a DNA motif to biological features of the interaction of a TF with its binding site. Here, we present an alternative strategy for representing DNA motifs – called Structural Motif (StruM) – that can easily represent different sets of structural features. Here we infer structural features from dinucleotide properties listed in the Dinucleotide Property Database. These StruMs show distinct improvements over standard PWMs in several ways. For example, we show that StruMs capture biologically relevant features of a TF binding motif as determined by crystal structures of DNA-protein complexes. Additionally, in DNase footprinting experiments performing binary classification of bound vs. unbound regions StruMs achieve improved precision and recall over PWMs. StruMs provide a flexible and adaptable approach that can accommodate and integrate diverse features to investigate their contribution to DNA-protein interactions.

🡐 Back

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Mr. Amirali Aghazadeh
Rice University
amirali@rice.edu

Ms. Thahmina Ali
Hunter College
thahminaali@gmail.com

Mr. Mohammad Ruhul Amin
Stony Brook University
moamin@cs.stonybrook.edu

Mr. Lin An
Pennsylvania State University
lua137@psu.edu

Gladys Andino
Purdue University
gandino@purdue.edu

Mr. Ziga Avsec
Technical University Munich
avsec@in.tum.de

Mr. Mohammed Awan
Fisk University
m.omar.awan@gmail.com

Dr. Anastasia Baryshnikova
Princeton University
abarysh@princeton.edu

Dr. Alexis Battle
Johns Hopkins University
ajbattle@cs.jhu.edu

Dr. Bonnie Berger
MIT CSAIL
bab@mit.edu

Mr. Evan Biederstedt
New York Genome Center
ebiederstedt@NYGENOME.ORG

Dr. Stefan Boeing
The Francis Crick Institute
stefan.boeing@crick.ac.uk

Dr. Gerard Bouffard
NIH/NHGRI
bouffard@mail.nih.gov

Prof. Guillaume Bourque
McGill University
guil.bourque@mcgill.ca

Dr. Martina Bradic
New York University
mb3188@nyu.edu

Dr. Stuart Brown
NYU School of Medicine
stuart.brown@nyumc.org

Mr. Jordan Bryan
The Broad Institute of MIT and Harvard
jbryan@broadinstitute.org

Dr. Ben Busby
NIH/NLM/NCBI
busbybr@mail.nih.gov

Mr. Adam Butchy
University of Pittsburgh
aab133@pitt.edu

Mr. Andrew Butler
New York Genome Center
andrew.butler33@gmail.com

Dr. Fabien Campagne
Weill Cornell Medicine
fac2003@campagnelab.org

Dr. Zach Charlop-Powers
The Rockefeller University
zcharlop@rockefeller.edu

Mr. John Chilton
Galaxy Project
jmchilton@gmail.com

Mr. Krishna Choudhary
University of California Davis
kchoudhary@ucdavis.edu

Mr. Kapeel Chougule
Cold Spring Harbor Laboratory
kchougul@cshl.edu

Dr. Declan Clarke
Yale University
declan.clarke@yale.edu

Mr. Michael Considine
Johns Hopkins University
mconsid3@jhmi.edu

Mr. Schuyler Corry
Thermo Fisher Scientific
schuyler.corry@thermofisher.com

Dr. Nancy Cox
Vanderbilt University
nancy.j.cox@vanderbilt.edu

Dr. Engin Cukuroglu
Genome Institute of Singapore
cukuroglue@gis.a-star.edu.sg

Dr. Farhan Damani
Johns Hopkins University
farhand7@gmail.com

Mr. Gregory Darnell
Princeton University
gdarnell@princeton.edu

Mr. Peter DeFord
Johns Hopkins University
pdeford@jhu.edu

Dr. UPENDRA KUMAR DEVISETTY
CyVerse
upendra@cyverse.org

Dr. John Didion
NIH
john.didion@nih.gov

Ms. Valentina Difrancesco
National Human Genome Research Institute
vdifrancesco@mail.nih.gov

Dr. Rebecca Dikow
Smithsonian Institution
DikowR@si.edu

Dr. Alexander Dilthey
NIH
alexander.dilthey@nih.gov

Mr. Igor Dolgalev
NYU Langone Medical Center
igor.dolgalev@nyumc.org

Ms. Changsu Dong
Hunter College
abbysu8@gmail.com

Dr. Friederike Duendar
Weill Cornell Medicine
frd2007@med.cornell.edu

Dr. Nathan Dunn
Lawrence Berkeley National Labs
nathandunn@lbl.gov

Dr. Amin Emad
University of Illinois at Urbana-Champaign
emad2@illinois.edu

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Mr. Antonio Fabregat Mundo
EMBL-EBI
fabregat@ebi.ac.uk

Mr. Jeff Fagerlund
Camas Incorporated
jfagerlund@camasinc.net

Mr. Han Fang
Cold Spring Harbor Laboratory
hanfang.cshl@gmail.com

Dr. Andrew Fant
National Institutes of Health
andrew.fant@nih.gov

Mr. Hao Feng
Emory University
hfeng5@emory.edu

Dr. Paul Frandsen
Smithsonian Institution
FrandsenP@si.edu

Dr. Iddo Friedberg
Iowa State University
idoerg@iastate.edu

Dr. Alexander Fulton
Novozymes
afu@novozymes.com

Mr. Alexander Gawronski
Simon Fraser University
agawrons@sfu.ca

Dr. Mark Gerstein
Yale University
pi@gersteinlab.org

Dr. Ambarnil Ghosh
Sungkyunkwan University
ambarnilghosh@gmail.com

Dr. Nicholas Giangreco
Columbia University
npg2108@cumc.columbia.edu

Dr. Andrew Giessel
Moderna Therapeutics
agiessel@modernatx.com

Dr. Jeremy Goecks
The George Washington University
jgoecks@gwu.edu

Mr. Nick Greenfield
One Codex
nick@onecodex.com

Ms. Abigail Groff
Harvard University
agroff@fas.harvard.edu

Ms. Kristen Gulino
New York University
kmg549@nyu.edu

Mr. John Hamilton
Michigan State University
jham@msu.edu

Dr. Kasper Hansen
Johns Hopkins University
kasperdanielhansen@gmail.com

Dr. Nancy Hansen
NHGRI/NIH
nhansen@mail.nih.gov

Dr. Yuan Hao
Cold Spring Harbor Laboratory
mhammell@cshl.edu

Dr. Arif Harmanci
Yale University
arif.harmanci@yale.edu

Ms. Yuan He
Johns Hopkins Univ.
yhe23@jhu.edu

Dr. R Antonio Herrera
Stony Brook University
dr@ranton.io

Dr. Steve Herrin
23andMe
sherrin@23andme.com

Dr. Michael Hoffman
University of Toronto
michael.hoffman@utoronto.ca

Dr. Yifei Huang
Cold Spring Harbor Laboratory
yihuang@cshl.edu

Mr. Hong Hur
The Rockefeller University
hhur@rockefeller.edu

Dr. Sajid Hussain
Fisk University
shussain@fisk.edu

Ms. Elizabeth Hutton
Watson School of Biological Sciences, CSHL
ehutton@cshl.edu

Mr. Niel Infante
West Virginia University
aminfante@hsc.wvu.edu

Dr. Rafael Irizarry
Dana-Farber Cancer Institute
rafa@jimmy.harvard.edu

Dr. Kazuo Ishii
Tokyo University of Agriculture and Technology
kazuoishii2014@gmail.com

Mr Johnny Israeli
Stanford University
jisraeli@stanford.edu

Dr. Pierre-Etienne Jacques
Université de Sherbrooke
Pierre-Etienne.Jacques@USherbrooke.ca

Mr. Chirag Jain
Georgia Institute of Technology
cjain7@gatech.edu

Dr. Yumi Jin
NIH/NLM/NCBI
jinyu@mail.nih.gov

Dr. Andre Kahles
Eidgenoessische Technische Hochschule Zuerich
andre.kahles@inf.ethz.ch

Dr. Sitharthan Kamalakaran
Columbia University
sk3031@cumc.columbia.edu

Mr. Venkateswara Kanaparthi
Thermo Fisher Scientific
VENKATESWARA.KANAPARTHI@LIFETECH.COM

Dr. Peer Karmaus
St. Jude Children's Research Hospital
peer.karmaus@stjude.org

Dr. Takeya Kasukawa
RIKEN Clst
takeya.kasukawa@riken.jp

Ms. Jenna Kefeli
Columbia University
jnk2127@columbia.edu

Mr. Keffy Kehrli
Stony Brook University
keffy.kehrli@stonybrook.edu

Mr. Stephen Kelly
NYU Langone Medical Center
stephen.kelly@nyumc.org

Dr. Radhika Khetani
Harvard T. H. Chan School of Public Health
rkhetani@hsph.harvard.edu

Mr. Alireza Khodadadi-Jamayran
Applied Bioinformatics Center
alireza.khodadadi.j@gmail.com

Ms. Minji Kim
University of Illinois at Urbana-Champaign
mkim158@illinois.edu

Mr. Baekdoo Kim
Hunter College, BioIT Core Lab
baegi7942@gmail.com

Ms. Susanna Kiwala
Washington University School of Medicine
ssiebert@wustl.edu

Dr. Michael Knudsen
Aarhus University Hospital
michaelk@clin.au.dk

Mr. Ilya Korsunsky
New York University
ilya.korsunsky@gmail.com

Dr. Alper Kucukural
University of Massachusetts Medical School
alper.kucukural@umassmed.edu

Dr. Hillel Kugler
Bar-Ilan University
hkugler@outlook.com

Dr. Vivek Kumar
Cold Spring Harbor Laboratory
vkumar@cshl.edu

Ms. Kathleen Kyle
Florida State University
kyle@genomics.fsu.edu

Dr. Ben Langmead
Johns Hopkins University
langmea@cs.jhu.edu

Mr. Jonathan Laperle
Universite de Sherbrooke
jonathan.laperle@usherbrooke.ca

Christian Le Cocq
Agilent Technologies Inc.
christian_lecocq@agilent.com

Dr. Joslynn Lee
Cold Spring Harbor Laboratory
jolee@cshl.edu

Dr. Jeff Leek
Johns Hopkins University
jtleek@jhu.edu

Dr. Janelle Leuthaeuser
University of Richmond
jleuthae@gmail.com

Dr. Lixia Li
Merck
lixia_li@merck.com

Dr. Wei Li
University of Pennsylvania
weili1@upenn.edu

Dr. Yulong Li
Institute of Oceanology
liyulong12@mails.ucas.ac.cn

Dr. Suh-yuen Liang
Academia Sinica
syliang@gate.sinica.edu.tw

Dr. Yupu Liang
The Rockefeller University
liangy@rockefeller.edu

Dr. Christophe Liseron-Monfils
Bayer Cropscience
christophe.liseron-monfils@bayer.com

Dr. Fenglong Liu
Monsanto
fenglong.liu@monsanto.com

Mr. Bingjian Liu
Institute of Oceanology
442753420@qq.com

Dr. Shirley Liu
Dana Farber Cancer Institute
xsliu@jimmy.harvard.edu

Dr. Michael Lovci
Intrexon
michaeltlovci@gmail.com

Mr. Zhenyuan Lu
CSHL
luj@cshl.edu

Mr. Ruibang Luo
The Johns Hopkins University
rluo5@jhu.edu

Dr. Nancy Mah
Charite University Medicine Berlin
nancy-lynne.mah@charite.de

Ms. Laraib Malik
Stony Brook University
lmalik@cs.stonybrook.edu

Dr. David Mayhew
GlaxoSmithKline
david.n.mayhew@gsk.com

Mr. Wilson McKerrow
Brown University
willmckerrow@gmail.com

Prof. Paul Medvedev
Penn State
pashadag@cse.psu.edu

Dr. Aleksandar Milosavljevic
Baylor College of Medicine
amilosav@bcm.edu

Mr. Siddhartha Mitra
The Rockefeller University
smitra@rockefeller.edu

Ms. Kelly Moffat
CosmosID
kelly.moffat@cosmosid.com

Mr. David Molik
Cold Spring Harbor Lab
dmolik@cshl.edu

Mr. Philip Montgomery
Broad institute
pmontgom@broadinstitute.org

Dr. Martin Morgan
Roswell Park Cancer Institute
martin.morgan@roswellpark.org

Dr. Adriana Munoz Jimenez
CSHL
amunoz@cshl.edu

Ms. Sushma Nagaraj
University of Maryland, School of Medicine
sparankush@som.umaryland.edu

Mr. Matt Nash
Q2 Solutions
matthew.nash@q2labsolutions.com

Dr. Maria Nattestad
Cold Spring Harbor Laboratory
mnattest@cshl.edu

Dr. Benjamin Neale
Massachusetts General Hospital
neale@atgu.mgh.harvard.edu

Dr. Anton Nekrutenko
Penn State/galaxyproject.org
anton@nekrut.org

Dr. Abhinav Nellore
Johns Hopkins University
anellore@gmail.com

Mr. Yashar Niknafs
University of Michigan
yniknafs@med.umich.edu

Mr. Frank Nothaft
UC Berkeley
fnothaft@alumni.stanford.edu

Dr. Cihan Oguz
National Institutes of Health
cihan.oguz@nih.gov

Mr. Andrew Olson
Cold Spring Harbor Laboratory
olson@cshl.edu

Dr. Shuichi Onami
RIKEN Quantitative Biology Center
sonami@riken.jp

Prof. Lior Pachter
UC Berkeley
lpachter@math.berkeley.edu

Mr. Aspyn Palatnick
Cold Spring Harbor Laboratory
paluter@gmail.com

Dr. Alex Pankov
Thermo Fisher Scientific
aleksandr.pankov@thermofisher.com

Dr. YoSon Park
University of Pennsylvania
ypar@upenn.edu

Dr. Benedict Paten
UCSC
benedict@soe.ucsc.edu

Dr. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Dr. Lon Phan
NIH/NLM/NCBI
lonphan@ncbi.nlm.nih.gov

Dr. Stephen Piccolo
Brigham Young University
stephen_piccolo@byu.edu

Mr. Mark Pritt
Johns Hopkins University
jacobpritt@gmail.com

Dr. Jane Pulman
University of Liverpool
parsonsl@liverpool.ac.uk

Dr. Srividya Ramakrishnan
Cold Spring Harbor Laboratory
srividya.ramki@gmail.com

Dr. Gunnar Ratsch
ETH Zurich
raetsch@inf.ethz.ch

Dr. Alexander Ratushny
Celgene
aratushny@celgene.com

Dr. Vida Ravanmehr
University of Illinois at Urbana-Champaign
vidarm@illinois.edu

Dr. Meisam Razaviyayn
Stanford University
meisamr@stanford.edu

Dr. Jeffrey Rosenfeld
Rutger Cancer Institute of NJ
jeffrey.rosenfeld@rutgers.edu

Ms. Yulia Rubanova
University of Toronto
yul.rubanova@gmail.com

Ms. Pamela Russell
Colorado School of Public Health
pamela.russell@gmail.com

Mr. Ashis Saha
Johns Hopkins University
ashis@jhu.edu

Dr. Cenk Sahinalp
Indiana University, Bloomington
cenksahi@indiana.edu

Ms. Manisha Sajnani
Indian Agricultural Research Institute
mani.sajnani@gmail.com

Prof. Steven Salzberg
Johns Hopkins University
salzberg@jhu.edu

Dr. Hirak Sarkar
Cold Spring Harbor Lab
hsarkar@cshl.edu

Mr. Khaled Sayed
University of Pittsburgh
kss60@pitt.edu

Dr. Michael Schatz
Cold Spring Harbor
mschatz@cshl.edu

Dr. Valerie Schneider
NIH/NLM/NCBI
schneiva@ncbi.nlm.nih.gov

Dr. Fritz Sedlazeck
Johns Hopkins University
fritz.sedlazeck@gmail.com

Ms. Jessica Severin
RIKEN Yokohama
jessica.severin@riken.jp

Mr. Ronak Shah
Memorial Sloan-Kettering Cancer Center
shahr2@mskcc.org

Ms. Surbhi Sharma
University of Nevada, Las Vegas
SHARMAS6@UNLV.NEVADA.EDU

Ms. Anna Shcherbina
Stanford University
annashch@stanford.edu

Ms. Avanti Shrikumar
Stanford University
avanti.shrikumar@gmail.com

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Eric Solis
Yumanity Therapeutics
esolis@yumanity.com

Dr. Paul Stewart
Moffitt Cancer Center
paul.stewart@moffitt.org

Mr. Georg Stricker
Technical University Munich
georg.stricker@in.tum.de

Mr. Chen Sun
The Pennsylvania State University
bbsunchen@outlook.com

Dr. Christopher Tabone
FlyBase - Harvard University
ctabone@morgan.harvard.edu

Ms. Jie Tan
Geisel School of Medicine at Dartmouth
jie.tan.gr@dartmouth.edu

Mr. Yin Tang
Pennsylvania State University, University Park
yxt148@psu.edu

Ms. Morgan Taschuk
Ontario Institute for Cancer Research
morgan.taschuk@oicr.on.ca

Dr. Marcela Tello-Ruiz
Cold Spring Harbor Laboratory
mmonaco@cshl.edu

Mr. William Thistlethwaite
Baylor College of Medicine
thistlew@bcm.edu

Dr. John Thompson
Bristol-Myers Squibb
john.thompson@bms.com

Dr. Tiffany Timbers
University of British Columbia
tiffany.timbers@stat.ubc.ca

Mr. Remi Torracinta
Weill Cornell Medicine
remi.torracinta@campagnelab.org

Dr. Olga Troyanskaya
Princeton University
ogt@princeton.edu

Mr. Aviad Tsherniak
Broad Institute
aviad@broadinstitute.org

Dr. Hiroki Ueda
Bio-IT R&D Office
ueda.hiroki-02@jp.fujitsu.com

Ms. Meghana Vemulapalli
NIH
meghana.vemulapalli@nih.gov

Dr. Daniel Vera
Florida State University
vera@genomics.fsu.edu

Dr. Davide Verzotto
Genome Institute of Singapore, A*STAR
d.verzotto@gmail.com

Mr. Jan Vogel
Genentech Inc
vogelj4@gene.com

Mr. Michael Wainberg
Stanford University
wainberg@stanford.edu

Dr. Liya Wang
Cold Spring Harbor Labs
wangli@cshl.edu

Dr. Li Wang
Broad Institute
wangli@broadinstitute.org

Mr. Christopher Wilks
Johns Hopkins University
cwilks3@jhu.edu

Mr. Jason Williams
Cold Spring Harbor Laboratory
williams@cshl.edu

Mr. Adam Wright
Ontario Institute for Cancer Research
adam.wright@oicr.on.ca

Dr. Tingfen Yan
NIH/NIMHD
yant@mail.nih.gov

Dr. Li Yang
CAS-MPG Partner Institute for Computational Biol
liyang@picb.ac.cn

Dr. Azam Yazdani
University of Texas Health Science Center
azam.yazdani@uth.tmc.edu

Dr. Song Yi
MD ANDERSON CANCER CENTER
yisong2008@gmail.com

Ms. Wen-Chi Yin
The Hospital for Sick Children
wenchi.yin@utoronto.ca

Ms. Miu ki Yip
Cold Spring Harbor Lab
myip@cshl.edu

Dr. Ying Zhang
Hua Zhong University of Science and Technology
ying_zh@hust.edu.cn

Mr. Zhancheng Zhang
GeneDx, Inc.
zzhang@genedx.com

Mr. Weizhuang Zhou
Stanford University
wzchew@stanford.edu

Ms. Naihui Zhou
Iowa State University
nzhou@iastate.edu

Ms. Sai Zhou
Stony Brook University
sai.zhou@stonybrook.edu

Mr. John Ziegler
Memorial Sloan Kettering Cancer Center
zieglerj@mskcc.org

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers

BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers

Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
***Helpful tips* – Use PIN# 63345** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access

Grace Auditorium
Upper level: E-mail and printing
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar

Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late (Cash Only)
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes, ATM

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m.
Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, east wing, lower level
PIN#: **Press 63345 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a “single stream waste management” system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($9.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
US Limousine Service	800-962-2827,ext:3 **(1047)**
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322