Abstracts of papers presented
at the 2023 meeting on

GENOME INFORMATICS

December 6–December 9, 2023

GGAAGCATGCTTGATGAAAATATTTGTTGCCACTGTGGA
GC
GCCCA
GGAGAGAAGAAACA
AG
GGGAGACTGGAAAACGAATATAAGAGGGGGCAAGGACC
GGAATA
GAGGGAAAGCCCA
AA
GAAAAAAAAAAAGAAGAAATAAAAAAA
AATGGAAAAACACCACCAAACA
GATAAAAAAGGAGAAA
AGCAAGAAACCACACCA

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2023 meeting on

GENOME INFORMATICS

December 6–December 9, 2023

Arranged by

Joanna Kelley, *University of California, Santa Cruz*
Páll Melsted, *University of Iceland*
Nicola Mulder, *University of Cape Town, South Africa*
Oliver Stegle, *German Cancer Research Center, Germany*

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Support for this meeting was provided in part by the **National Human Genome Research Institute (NHGRI),** a branch of the **National Institutes of Health; PacBio;** and the **James P. Taylor Foundation for Open Science Scholarship Fund.**

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

Corporate Benefactors

Estée Lauder Companies
Regeneron

Corporate Sponsors

Agilent Technologies
Biogen
Bristol-Myers Squibb Company
Calico Labs
Genentech *A member of the Roche Group*
Merck & Co., Inc.
New England Biolabs
Novartis Institutes for Biomedical Research

Corporate Partners

Alexandria Launch Labs

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

Cover: Páll Melsted, with generous assistance from DALLE-3.

GENOME INFORMATICS

Wednesday, December 6 – Saturday, December 9, 2023

Wednesday	7:30 pm – 10:30 pm	**1** Microbial and Metagenomics
Thursday	9:00 am – 12:00 pm	**2** Pangenome
Thursday	1:30 pm – 2:15 pm	Keynote Lecture by Sean Eddy
Thursday	2:30 pm – 5:30 pm	**3** Single Cell and Spatial Omics
Thursday	5:30 pm	*Wine and Cheese Party*
Thursday	7:30 pm – 10:30 pm	**Poster Session I**
Friday	8:30 am – 11:00 am	**4** Functional Genomics
Friday	11:45 am – 12:30 pm	**NIH Discussion Early Stage Investigators**
Friday	1:30 pm – 4:30 pm	**5** Variant Discovery
Friday	4:45 pm – 5:30 pm	Keynote Lecture by Karen Miga
Friday	5:30 pm – 7:00 pm	**Poster Session II**
Friday	7:00 pm	*Cocktails and Banquet*
Saturday	9:00 am – 12:00 pm	**6** Genome Assembly and Sequencing Algorithms

---

Workshop (immediately following morning sessions)
PacBio, Thursday, December 7

All times shown are US Eastern: Time Zone Converter

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

Cold Spring Harbor Laboratory is committed to maintaining a safe and respectful environment for all meeting attendees, and does not permit or tolerate discrimination or harassment in any form. By participating in this meeting, you agree to abide by the Code of Conduct.

For further details as well as Definitions and Examples and how to Report Violations, please see the back of this book.

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(s).

Please note that 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.

Any discussion via social media platforms of material presented at this meeting requires explicit permission from the presenting author(s).

Printed on 100% recycled paper.

PROGRAM

WEDNESDAY, December 6—7:30 PM

SESSION 1 MICROBIAL AND METAGENOMICS

Chairperson: **Eran Segal,** Weizmann Institute of Science, Rehovot, Israel

Major data analysis errors invalidate cancer microbiome findings
Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, Steven L. Salzberg.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 1

Empowering global disease surveillance—The CURED tool for rapid identification of unique clonal biomarkers
Erin Theiller, Elizabeth Qian, Andries Feder, Alice Slotfeld Viana, Agnes Marie Sá Figueiredo, Paul J. Planet, Ahmed M. Moustafa.
Presenter affiliation: Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania. 2

Strainy—Assembly-based metagenomic strain phasing using long reads
Ekatarina Kazantseva, Ataberk Donmez, Mihai Pop, Mikhail Kolmogorov.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland; University of Maryland, College Park, Maryland. 3

GC-bias aware species abundance estimation from metagenomic data with GuaCAMOLE increases accuracy and comparability
Laurenz Holcik, Florian Pflug, Arndt von Haeseler.
Presenter affiliation: University of Vienna, Vienna, Austria. 4

Personalized medicine based on deep human phenotyping
Eran Segal.
Presenter affiliation: Weizmann Institute of Science, Rehovot, Israel. 5

Ultrafast, coverage-corrected genome similarity queries for metagenomic shotgun samples with sylph
Jim Shaw, Yun William Yu.
Presenter affiliation: University of Toronto, Toronto, Canada. 6

< Back

Centrifuger—Lossless compression of microbiome genomes for efficient and accurate metagenomic sequence classification
Li Song, Daehwan Kim, Ben Langmead.
Presenter affiliation: Dartmouth College, Hanover, New Hampshire. 7

Leveraging large language models for metagenomic analysis
Mohammadsaleh Refahi, Bahrad Sokhansanj, Gail Rosen.
Presenter affiliation: Drexel University, Philadelphia, Pennsylvania. 8

THURSDAY, December 7—9:00 AM

SESSION 2 PANGENOME

Chairperson: **Veli Makinen,** University of Helsinki, Finland

Pangenomics with founder sequences and graphs
Veli Makinen.
Presenter affiliation: University of Helsinki, Helsinki, Finland. 9

Compressed linear pangenome indexes for more robust read classification
Omar Ahmed, Massimiliano Rossi, Travis Gagie, Christina Boucher, Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 10

Graph-based and gene-based pangenome of *Lactococcus lactis* and *Lactococcus cremoris*
Paulo J. Dias, José M. Santos, Diogo L. Antunes, Sofia O. Duarte, Leonilde M. Moreira, Gabriel A. Monteiro.
Presenter affiliation: iBB - Institute for Bioengineering and Biosciences, Lisbon, Portugal; Associate Laboratory i4HB - Institute for Health and Bioeconomy, Lisbon, Portugal. 11

Co-linear chaining on pangenome graphs
Jyotshna Rajput, Ghanshyam Chandra, Chirag Jain.
Presenter affiliation: Indian Institute of Science, Bengaluru, India. 12

MONI-Align—An r-index based pangenomics aligner
Rahul Varki, Eddie Ferro, Massmiliano Rossi, Marco Oliva, Ben Langmead, Christina Boucher.
Presenter affiliation: University of Florida, Gainesville, Florida. 13

< Back

Panagram—Alignment-free and interactive pan-genome visualization
Katharine Jenike, Sam Kovaka, Matthias Benoit, Srividya Ramakrishnan, Shujun Ou, James Saterlee, Stephen Hwang, Iacopo Gentile, Anat Hendelman, Michael Passalacqua, Xingang Wang, Michael Alonge, Hamsini Suresh, Ryan Santos, Blaine Fitzgerald, Gina Robitaille, Edeline Gagnon, Melissa Kramer, Sara Goodwin, W.Richard McCombie, Jaime Prohens, Tiina E. Särkinen, Amy Frary, Jesse Gillis, Joyce Van Eck, Ben Langmead, Zachary B. Lippman, Michael C. Schatz.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 14

Giggles—Pangenome-based genome inference using long reads
Samarendra Pani, Jana Ebler, Tobias Marschall.
Presenter affiliation: Heinrich Heine University, Düsseldorf, Germany. 15

Real-time nanopore adaptive sampling with Movi
Mohsen Zakeri, Nathaniel Brown, Omar Ahmed, Travis Gagie, Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 16

THURSDAY, December 7—1:30 PM

KEYNOTE SPEAKER

Computational analysis of RNA structure and function
Sean R. Eddy.
Presenter affiliation: HHMI & Harvard University, Cambridge, Massachusetts. 17

THURSDAY, December 7—2:30 PM

SESSION 3 SINGLE CELL AND SPATIAL OMICS

Chairperson: **Amos Tanay,** Weizmann Institute, Rehovot, Israel
Anne Carpenter, Broad Institute of Harvard and MIT, Cambridge, Massachusetts

Spatio-temporal quantitative models for embryonic development
Amos Tanay.
Presenter affiliation: Weizmann Institute, Rehovot, Israel. 18

< Back

PerturbDecode, a probabilistic analysis framework to recover regulatory circuits and predict genetic interactions from large-scale perturbation screens
Basak Eraslan, Katie Geiger-Schuller, Kelvin Chen, Romain Lopez, Payman Yadollahpour, Olena Kuksenko, Pratiksha Thakore, Ozge Karayel Eren, Andrea Yung, Anugraha Rajagopalan, Ana Meireles de Sousa, Karren Dai Yang, Nir Hacohen, Caroline Uhler, Orit Rozenblatt-Rosen, Shimon Sakaguchi, Aviv Regev.
Presenter affiliation: Stanford University, Stanford, California; Genentech Research and Early Development, South San Francisco, California; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 19

Assessing performance of supervised and unsupervised cell type labeling algorithms for cancer scRNA-seq data
Erik Christensen, Ping Luo, Andrei Turinsky, Mia Husic, Alaina Mahalanabis, Alaine Naidas, Javier Diaz-Mejia, Michael Brudno, Trevor Pugh, Arun Ramani, Parisa Shooshtari.
Presenter affiliation: Western University, London, Canada. 20

Isoform quantitative trait loci analysis of neuropsychiatric disorders in adult brains at single-cell resolution
Eric Nguyen, Matthew Jensen, Diego Garrido Martin, Declan Clarke, Prashant Emani, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 21

Functional genomics using image-based profiling—From variant impact to drug screening
Anne E. Carpenter.
Presenter affiliation: Broad Institute of Harvard and MIT, Cambridge, Massachusetts. 22

Cell type-specific interaction analysis using doublets in scRNA-seq (CIcADA)
Courtney Schiebout, Hannah Lust, Yina Huang, H. Robert Frost.
Presenter affiliation: Dartmouth College, Hanover, New Hampshire. 23

Robust and scalable intratumor heterogeneity and tumor progression tree inference and assessment through single-cell RNA sequencing data
Farid Rashidi Mehrabadi, Salem Malikic, Eva Perez-Guijarro, Kerrie L. Marie, Glenn Merlino, Chi-Ping Day, S. Cenk Sahinalp.
Presenter affiliation: National Cancer Institute, National Institutes of Health, Bethesda, Maryland. 24

< Back

SpaceTree—Deciphering Tumor microenvironments by joint modeling of cell states and genotype-phenotype relationships in spatial omics data
Olga Lazareva, Elyas Heidari, Omer Bayraktar, Oliver Stegle.
Presenter affiliation: German Cancer Research Center, Heidelberg, Germany; European Molecular Biology Laboratory, Heidelberg, Germany. 25

THURSDAY, December 7—5:30 PM

Wine and Cheese Party

THURSDAY, December 7—7:30 PM

POSTER SESSION I

See p. xv for List of Posters

FRIDAY, December 8—8:30 AM

SESSION 4 FUNCTIONAL GENOMICS

Chairpersons: **Eleftheria Zeggini,** Helmholtz Zentrum München, Germany
Ben Lehner, Centre for Genomic Regulation, Barcelona, Spain

Translational genomics of complex disease
Eleftheria Zeggini.
Presenter affiliation: Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; Technical University of Munich (TUM) and Klinikum Rechts der Isar, Munich, Germany. 26

Reference-free differential isoform analysis using short-read RNA-seq data
Carlos F. Buen Abad Najar, Dongyue Xie, Matthew Stephens, Yang I. Li.
Presenter affiliation: University of Chicago, Chicago, Illinois. 27

< Back

TransferQTL expands existing eQTL catalogs across human tissues
Yuhang Chen.
Presenter affiliation: Yale University, New Haven, Connecticut. 28

Uncovering novel targets in human tuberous sclerosis and non-alcoholic fatty liver disease models by integrating multiomics and automation with pooled optical screening
Max R. Salick, Srinivasan Sivanandan, Bobby Leitmann, Saradha Venkatachalapathy, Atieh Givmanesh, Rahul Atmaramani, Shengjiang Tu, Owen Chen, Zack Phillips, John Bisognano, Alicia Lee, Lexie Ewer, Bay Johnson, Navpreet Ranu, Eilon Sharon, David Lloyd, Ajamete Kaykas, Ci Chu.
Presenter affiliation: insitro, South San Francisco, California. 29

Finding all our switches mutating everything to understand allostery
Ben Lehner.
Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom; CRG, Barcelona, Spain. 30

3D genomic features across >50 diverse cell types reveal insights into the differing genomic architectures of BMD determination and osteoporotic fracture pathogenesis
Khanh B. Trang, Matthew C. Pahl, James A. Pippin, Alessandra Chesi, Yadav Wagley, Babette S. Zemel, Kurt D. Hankenson, Andrew D. Wells, Struan F. Grant.
Presenter affiliation: The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania. 31

DNABERT-2—Efficient foundation model for multi-species genomes
Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Han Liu, Ramana V. Davuluri.
Presenter affiliation: Stony Brook University, Stony Brook, New York. 32

Recovering hidden layers of information in single-cell data
Mor Nitzan.
Presenter affiliation: Hebrew University of Jerusalem, Jerusalem, Israel. 33

< Back

FRIDAY, December 8—11:45 AM

PANEL DISCUSSION
NIH Early Stage Investigators

Moderator: **Anton Nekrutenko,** Pennsylvania State University

Panelists: **Sean Eddy,** Harvard University / HHMI
Shurjo Sen, NHGRI, National Institutes of Health

FRIDAY, December 8—1:30 PM

SESSION 5 VARIANT DISCOVERY

Chairpersons: **Jared Simpson,** University of Toronto, Canada
Ida Moltke, University of Copenhagen, Denmark

Calling somatic mutations from long read tumors without matched normal samples
Jared T. Simpson.
Presenter affiliation: University of Toronto, Ontario Institute for Cancer Research, Toronto, Canada. 34

Minimizing reference bias with an impute-first approach
Naga Sai Kavya Vaddadi, Taher Mun, Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 35

Severus—A tool for automatic characterization of complex germline and somatic rearrangements in cancer using long-read sequencing.
Ayse Keskus, Tanveer Ahmad, Ataberk Donmez, Asher Bryant, Isabel Rodriguez, Nicole M. Rossi, Yi Xie, Byunggil Yoo, Rose Milano, Hong Lou, Jimin Park, Joshua Gardner, Brandy McNulty, Karen Miga, Midhat Farooqi, Benedict Paten, Michael Dean, Mikhail Kolmogorov.
Presenter affiliation: CCR, National Cancer Institute, Bethesda, Maryland. 36

< Back

Long-read sequencing of 1000 genomes samples to build a comprehensive catalog of human genetic variation
Nikhita Damaraju, J. (Gus) Gustafson, Sophia B. Gibson, Miranda Galey, Kendra Hoekzema, Joy Goffena, Maisha Sinha, The 1000 Genomes ONT Sequencing Consortium, Fritz Sedlazeck, Matt Loose, Miten Jain, Evan E. Eichler, Danny E. Miller.
Presenter affiliation: University of Washington, Seattle, Washington. 37

Insights into the genetic architecture of diabetes and other complex traits in the Greenlandic population
Ida Moltke.
Presenter affiliation: University of Copenhagen, Copenhagen, Denmark. 38

STIX—A novel approach for comprehensive somatic structural variation detection and gene fusion identification
Christopher E. Ojukwu, Murad Chowdhury, Ryan Layer.
Presenter affiliation: University of Colorado Boulder, Boulder, Colorado. 39

The molecular environment of osteoarthritis risk genes in primary cartilage
Georgia Katsoula, John E G Lawrence, Mauro Tutino, Ana Luiza Arruda, Petra Balogh, Lorraine Southam, Diane Swift, Sam Behjati, Sarah Teichmann, J Mark Wilkinson, Eleftheria Zeggini.
Presenter affiliation: Technical University of Munich (TUM) and Klinikum Rechts der Isar, Munich, Germany; Institute of Translational Genomics, Munich, Germany. 40

Querying biobank-scale genomes to rapidly identify genetically matched cohorts
Kristen E. Schneider, Murad Chowdhury, Mariano Tepper, Mark Hildebrand, Jawad Khan, Marteinn Hardarson, Bjarni Halldorsson, Chris Gignoux, Ryan Layer.
Presenter affiliation: University of Colorado Boulder, Boulder, Colorado. 41

< Back

FRIDAY, December 8—4:45 PM

KEYNOTE SPEAKER

Complete genomes to expand studies of genetic and epigenetic inheritance of centromeres
Monika Cechova**,** Sergey Koren, Julian K. Lucas, Rebecca Serra Mari, Mobin Asri, David Porubsky, Jordan M. Eizenga, Brandy McNulty, Andrey Bzikadze, Shloka Negi, Christopher Markovic, Tamara Potapova, Jennifer L. Gerton, Pavel A. Pevzner, Evan E. Eichler, Benedict Paten, Adam M. Phillippy, Ting Wang, Nathan O. Stitziel , Robert S. Fulton, Tobias Marschall, Karen H. Miga.
Presenter affiliation: University of California, Santa Cruz, Santa Cruz, California. 42

FRIDAY, December 8—5:30 PM

POSTER SESSION II

See p. xxvii for List of Posters

FRIDAY, December 8—7:00 PM

COCKTAILS and BANQUET

SATURDAY, December 9—9:00 AM

SESSION 6 GENOME ASSEMBLY AND SEQUENCING ALGORITHMS

Chairpersons: **Rayan Chikhi,** Institut Pasteur, Paris, France
Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Current tools for peta-scale sequence exploration
Rayan Chikhi.
Presenter affiliation: Institut Pasteur, Paris, France. 42

Verkko's approaches for one-button T2T resolution for diploid human-sized genome
Dmitry Antipov, Adam Phillippy, Sergey Koren.
Presenter affiliation: NHGRI, Bethesda, Maryland. 43

< Back

Haplotype-resolved characterization of repeat expansions and patterns of methylation from 1000 Genomes ONT Consortium data
Sophia B. Gibson, J. (Gus) Gustafson, Nikhita Damaraju, Miranda Galey, Kendra Hoekzema, Joy Goffena, Jordan Knuth, Maisha Sinha, 1000 Genomes ONT Sequencing Consortium, GREGoR Consortium, Fritz Sedlazeck, Matt Loose, Miten Jain, Lea Starita, Evan Eichler, Danny E. Miller.
Presenter affiliation: University of Washington, Seattle, Washington. 44

Adapting analysis tools to a workflow-centric world
Nate Coraor, John M. Chilton, Nuwan A. Goonasekera, Anton Nekrutenko, The Galaxy Team.
Presenter affiliation: Pennylvania State University, University Park, Pennsylvania. 45

Understanding genetic variation in modern and archaic human genomes
Janet Kelso.
Presenter affiliation: Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. 46

DNA and RNA modification detection via rapid nanopore signal alignment, analysis, and visualization with Uncalled4
Sam Kovaka, Paul W. Hook, Katharine Jenike, Luke Morina, Winston Timp, Michael C. Schatz.
Presenter affiliation: Johns Hopkins University, Balitmore, Maryland. 47

Multi-haplotype curation for high ploidy level plant genomes—The *Urtica dioica* story
Dominic Absolon, Ksenia Krasheninnikova, Shane A. McCarthy, Jonathan M. Wood.
Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom. 48

Enabling petabase-scale search of millions of metagenomes with branchwater
Luiz Irber, N Tessa Pierce-Ward, Suzanne M. Fleishman, Adam R. Rivers, C Titus Brown.
Presenter affiliation: UC Davis, Davis, California. 49

< Back

POSTER SESSION I

rMATS-turbo—An efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data
Jenea I. Adams, Yuanyuan Wang, Zhijie Xie, Eric Kutschera, Kathryn E. Kadash-Edmondson, Yi Xing.
Presenter affiliation: The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; University of Pennsylvania, Philadelphia, Pennsylvania. 51

Haplotype-specific karyotypes reconstruction and copy number aberrations/profiling from long reads sequencing data
Tanveer Ahmad, Mikhail Kolmogorov.
Presenter affiliation: Center for Cancer Research, Bethesda, Maryland. 52

Composition, function and strain-sharing between maternal breast milk and the infant gut microbiome
Mattea Allert, Pamela Ferretti, Kelsey Johnson, Timothy Heisel, Cheryl A. Gale, Ellen W. Demerath, Dan Knights, David A. Fields, Frank W. Albert, Ran Blekhman.
Presenter affiliation: University of Minnesota, Minneapolis, Minnesota. 53

Long-read sequencing-based pipeline for Neo-epitope candidates for immunotherapeotic targeting of U1 snRNA mutation in cancer
Fatemeh Almodaresi, Ander Diaz-Navarro, Andrea Senff-Ribeiro, Marie-Pierre Hardy, Quang Trinh, Sachin Kumar, Shimin Shuai, Craig Daniels, Xose S. Puente, Elias Campo, Michael Taylor, Claude Perreault, Lincoln Stein.
Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada. 54

Integrated analysis of imbalanced allelic expression to infer gene regulatory patterns in cancer
Mona Arabzadeh, Amartya Singh, Hossein Khiabanian, Shridar Ganesan.
Presenter affiliation: Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey. 55

Longitudinal single-cell genomic and transcriptomic analysis of relapsed pediatric AML
Byron Avihai, Amartya Singh, Hossein Khiabanian, Daniel Herranz.
Presenter affiliation: Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey. 56

< *Back*

Enhancing pandemic preparedness—A novel partial matching approach for identifying similar genetic material from diverse sources for pathogen surveillance
Morteza Baradaran, Ryan Layer, Kevin Skadron.
Presenter affiliation: University of Virginia, Charlottesville, Virginia. 57

Enhanced genome-based taxonomy for precise identification of Fusarium pathogens
Kassaye H. Belay, Reza Mazloom, Lenwood Heath, Boris A. Vinatzer.
Presenter affiliation: VVirginia Tech, Blacksburg, Virginia. 58

Quantum computing comes to the Galaxy
Bryan Raubenolt, Fabio Cumbo, Jayadev Joshi, Daniel Blankenberg.
Presenter affiliation: Cleveland Clinic, Cleveland, Ohio; Case Western Reserve University College of Medicine, Cleveland, Ohio. 59

What about B.O.B.? Bio-Ontology-Biases—The effects of study bias on less studied groups and classes in knowledge graph embedding methods
Michael S. Bradshaw.
Presenter affiliation: University of Colorado Boulder, Boulder, Colorado. 60

Benchmarking long-read somatic structural variant callers on a collection of tumor/normal cell lines
Asher Bryant, Ayse Keskus, Tanveer Ahmad, Ataberk Donmez, Isabel Rodriguez, Nicole M. Rossi, Yi Xie, Byunggil Yoo, Rose Milano, Hong Lou, Jimin Park, Joshua Gardner, Brandy McNulty, Karen Miga, Midhat Farooqi, Benedict Paten, Michael Dean, Mikhail Kolmogorov.
Presenter affiliation: Center for Cancer Research, NCI, Bethesda, Maryland. 61

mEnrich-seq—Methylation-guided enrichment sequencing and novel informatic methods to investigate specific bacterial taxa of interest directly from microbiome
Lei Cao, Yimeng Kong, Yu Fan, Mi Ni, Alan Tourancheau, Magdalena Ksiezarek, Edward A. Mead, Tonny Koo, Melissa Gitman, Xue-Song Zhang.
Presenter affiliation: Icahn School of Medicine at Mount Sinai, New York, New York. 62

< Back

Single-cell, long-read sequencing of the mouse hippocampus reveals learning-induced alternative splicing patterns and transcript isoform expression across cell types
Sheridan H. Cavalier, Paul Hook, Winston Timp, Richard Huganir.
Presenter affiliation: Johns Hopkins SOM, Baltimore, Maryland. 63

Splam—A deep-learning-based splice site predictor that improves spliced alignments
Kuan-Hao Chao, Alan Mao, Steven L. Salzberg, Mihaela Pertea.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 64

Scalable telomere-to-telomere assembly for diploid, polyploid and cancer genomes with double graph
Haoyu Cheng, Mobin Asri, Julian Lucas, Sergey Koren, Heng Li.
Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts. 65

Somatic driver gene alterations are associated with predictive anticancer response and prognostic assessment in pancreatic ductal adenocarcinoma
Eunwoo Choi, Jiyoon Hong, Seungmin Bang, HeeSeung Lee, Sangwoo Kim.
Presenter affiliation: Yonsei Univesrsity College of Medicine, Seoul, South Korea. 66

QuadST—A powerful and robust approach for identifying cell-cell interaction changed genes on spatially resolved transcriptomics
Jinmyung Choi, Pei Wang, Guo-Cheng Yuan, Xiaoyu Song.
Presenter affiliation: Icahn School of Medicine at Mount Sinai, New York, New York. 67

Swift pan-genomic methods for comprehensive genome annotation in crop genomes
Kapeel Chougule, Sharon Wei, Zhenyuan Lu, Andrew Olson, Doreen Ware.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 68

GenArk—Towards a million UCSC genome browsers
Hiram Clawson, Brian T. Lee, Brian J. Raney, Bogdan M. Kirilenko, Jonathan Casper, Michael Hiller, Robert M. Kuhn, Jairo N. Gonzalez, Angie S. Hinrichs, Christopher M. Lee, Luis R. Nassar, Gerardo Perez, Brittney Wick, Joel Armstrong, Matthew L. Speir, David Haussler, W. James Kent, Maximilian Haeussler.
Presenter affiliation: University of California, Santa Cruz, California. 69

< Back

RHEA—Recovering horizontal gene transfer events from metagenome assembly graphs
Kristen D Curry, Rayan Chikhi, Eduardo Rocha, Todd J Treangen.
Presenter affiliation: Rice University, Houston, Texas. 70

Development of a haplotype-aware assembly pipeline for analysis of rearrangements at the human CYP2D6 locus
Daisy Dahiya, Benjamin Alleva, Florencia Pratto, R. Daniel Camerini-Otero.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 71

Diversity and representation of South Asian genomes
Arun Das, Michael C. Schatz.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 72

Data access architecture at "Galactic" scale—Lessons learned (so far)
John Davis.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 73

Capturing and functional annotating the immunoglobulin loci of various species with thrid-generation sequencing
Dori Z. Deng, Will Seligmann, Helen M. Dooley, Richard E. Green, Russ Corbett-Detig, Christopher Vollmers.
Presenter affiliation: University of California Santa Cruz, Santa Cruz, California. 74

Comparison of 16s rRNA gene sequencing and shotgun metagenomic sequencing for rumen microbiome analysis
Gerardo R. Diaz Ortiz, Noelle N. Noyes.
Presenter affiliation: University of Minnesota, Saint Paul, Minnesota. 75

A metagenomic investigation of the work-related microbiome in US swine workers
Gerardo R. Diaz Ortiz, Ilya B. Slizovskiy, Montserrat Torremorell, Noelle R. Noyes
Presenter affiliation: University of Minnesota, Saint Paul, Minnesota. 76

Phylogenetic diversity patterns among gastrointestinal bacterial strains
Veronika Dubinkina.
Presenter affiliation: The Gladstone Institute of Data Science and Biotechnology, San Francisco, California. 77

< Back

Detecting differential transcript usage in complex diseases with SPIT
Beril Erdogdu, Ales Varabyou, Stephanie C. Hicks, Steven L. Salzberg, Mihaela Pertea.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland; Johns Hopkins School of Medicine and Whiting School of Engineering, Baltimore, Maryland. 78

Investigating the origins and impacts of structural variation and DNA methylation in high-grade serous ovarian cancer
Edward Esiri-Bloom, Stuart Aitken, Graeme Grimes, Ailith Ewing, Alison Meynert, Ryan Silk, Stuart Brown, Michael Churchman, C Simon Herrington, Patricia Roxburgh, Charlie Gourley, Colin A. Semple.
Presenter affiliation: MRC Human Genetics Unit, Edinburgh, United Kingdom. 79

Strain-resolution long read metagenome assembly
Xiaowen Feng, Heng Li.
Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts. 80

Sensitive and specific detection of mosaic chromosomal alterations from large-scale RNA-seq datasets
Teng Gao, Maria E. Kastriti, Viktor Ljungström, Andreas Heinzel, Arthur S. Tischler, Rainer Oberbauer, Po-Ru Loh, Igor Adameyko, Peter J. Park, Peter V. Kharchenko.
Presenter affiliation: Harvard Medical School, Boston, Massachusetts. 81

Re-analysis of microbial content found in tumors sequenced by The Cancer Genome Atlas Project
Peter Ge, Jennifer Lu, Daniela Puiu, Mahler Revsine, Amanda Xu, Mihaela Pertea, Steven L. Salzberg.
Presenter affiliation: Johns Hopkins School of Medicine, Baltimore, Maryland. 82

Integrative computational framework, *Dyscovr*, links mutated driver genes to metabolic dysregulation across 22 cancer types
Sara E. Geraghty, Jacob Boyer, Matthew McBride, Joshua Rabinowitz, Mona Singh.
Presenter affiliation: Princeton University, Princeton, New Jersey. 83

< Back

Characterizing host-pathogen interaction dynamics for *Toxoplasma gondii* with single-cell RNA sequencing—A pilot study
Yomna Gohar, Veronica Raba, Tobias Lautwein, Daniel Wind, Lisanna Hülse, karin Buchholz, Raba Katharina, Daniel Degrandi, klaus Pfeffer, Alexander Dilthey.
Presenter affiliation: Heinrich Heine University Düsseldorf, Düsseldorf, Germany. 84

Multi-sample Nanopore sequencing provides insights into melanoma heterogeneity and evolution
Anton Goretsky, Yuelin Liu, Ayse Keskus, Salem Malikic, Glenn Merlino, Chi-Ping Day, Erin Molloy, S. Cenk Sahinalp, Mikhail Kolmogorov.
Presenter affiliation: National Cancer Institute, Bethesda, Maryland; University of Maryland, College Park, Maryland. 85

Landscape of differentiation induced oncogenesis regulated by pseudogenes—A neural network based study of gastrointestinal tract
Pravallika Govada, Rajasekaran Ramalingam.
Presenter affiliation: Vellore Institute of Technology, Vellore, India. 86

Speeding up whole genome alignment by increasing hardware utilization
A. Burak Gulhan, Mahmut Kandemir, Maximilian Haeussler, Anton Nekrutrenko.
Presenter affiliation: Penn State, State College, Pennsylvania. 87

Evaluation of haplotype-aware long-read error correction with hifieval
Yujie Guo, Xiaowen Feng, Heng Li.
Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts. 88

For the development of a PCR primer design pipeline for detection of contamination in foods
Yoritaka Harazono, Keisuke Soga, Masahiro Kasahara.
Presenter affiliation: the University of Tokyo, Kahiwa-shi, Japan. 89

Toward telomere-to-telomere Felid genomes
Andrew J. Harris, Leslie A. Lyons, Wesley C. Warren, Kendra Hoekzema, Evan E. Eichler, William J. Murphy.
Presenter affiliation: Texas A&M University, College Station, Texas. 90

< Back

TEM-seq—An ultrasensitive multiomic platform for epitope-targeted DNA methylation mapping
Allison Hickman, Vishnu U. Sunitha Kumary, Bryan Venters, Jennifer Spengler, Anup Vaidya, Ryan Ezell, Jonathon Burg, Zu-wen Sun, Martis Cowles, Hang Geong Chin, Pierre Esteve, Chaithanya Ponnaluri, Isaac Meek, Sriharsa Pradhan, Michael-Christopher Keogh.
Presenter affiliation: EpiCypher, Inc, Durham, North Carolina. 91

Interpretable text-based machine learning for inferring systematic tissue and disease annotations of public transcriptome samples
Parker Hicks, Hao Yuan, Mansooreh Ahmadian, Arjun Krishnan.
Presenter affiliation: University of Colorado Anschutz Medical Campus, Aurora, Colorado. 92

Compleasm—A faster and more accurate reimplementation of BUSCO
Neng Huang, Heng Li.
Presenter affiliation: Dana-Farber Cancer Institute, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts. 93

Computational analysis of copy number variations in spatial transcriptomics data
Rongting Huang, Xianjie Huang, Ajit J. Nirmal, Yuanhua Huang.
Presenter affiliation: The University of Hong Kong, Hong Kong; Harvard Medical School, Boston, Massachusetts; Brigham and Women’s Hospital, Boston, Massachusetts. 94

Environmental and genetic insights into carcinogenesis—An approach using passive sampling and CHIP analysis in the companion dog
Christopher Husted, Kate Megquier, Adam Harris, Diane Genereux, Kim Anderson, Alexander Bick, Frances Chen, Elinor Karlsson.
Presenter affiliation: University of Massachusetts Chan Medical School, Worcester, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 95

Compressed indexing for pangenome substring queries
Stephen Hwang, Omar Y. Ahmed, Ben Langmead.
Presenter affiliation: Johns Hopkins School of Medicine, Baltimore, Maryland. 96

< Back

Investigating RNA splicing as a source of cellular diversity using a binomial mixture model
Keren Isaev, David A. Knowles.
Presenter affiliation: Columbia University , New York, New York; New York Genome Center, New York, New York. 97

Detection, characterization, and prevention of MMEJ deletions
Aditee Kadam, Shay Shilo, Hadas Naor, Mark Minden, Nathali Kaushansky, Noa Chapal, Liran Shlush.
Presenter affiliation: Weizmann Institute of Science, Rehovot, Israel. 98

CCSA—Concurrent force-based position solving and quantization for multiple sequence alignment
Daniel Kim.
Presenter affiliation: Horace Greeley High School, Chappaqua, New York. 99

Revealing hidden transcripts with a complete reference and personalized transcriptome graph
Juhyun Kim, Elizabeth Tseng, Adam M. Phillippy, Arang Rhie.
Presenter affiliation: NIH, Bethesda, Maryland; Seoul National University, Seoul, South Korea. 100

MOSCAL—Detection of mosaic variants using linked-read sequencing
Yongjun Kim, Shinwon Hwang, Hyeonju Son, Sangwoo Kim.
Presenter affiliation: Yonsei University College of Medicine, Seoul, South Korea. 101

Leveraging public datasets to understand Parkinson's disease progression
Rohit Kolora, Anna Rychkova.
Presenter affiliation: Alector Therapeutics, South San Francisco, California. 102

Gene expression prediction from histopathology images of colorectal cancer
Jonas Lehmitz, Philip Bischoff, Alexander Sudy, Johannes Liebig, Christian Conrad, Teresa G. Krieger.
Presenter affiliation: Charité, Berlin, Germany; Berlin Institute of Health, Berlin, Germany. 103

< Back

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)
Natalie Kucher, Michael C. Schatz, Anthony Philippakis.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 104

Transcriptome analysis of *Helicoverpa armigera* (Hubner) (Lepidoptera: Noctuidae) infected with *Metarhizium anisopliae*
Pankaj Kumar, Malyaj R. Prajapati.
Presenter affiliation: Sardar Vallabhbhai Patel University of Agriculture and Technology, Meerut, Indiana. 105

Automated reference genome assembly on public infrastructure with Galaxy
Delphine Lariviere, Linelle Abueg, Nadolina Brajuka, Cristóbal Gallardo-Alba, Bjorn Grüning, Byung June Ko, Alex Ostrovsky, Marc Palmada-Flores, Brandon D. Pickett, Keon Rabbani, Erich D. Jarvis, Adam M. Phillippy, Anton Nekrutenko, Michael Schatz, Giulio Formenti.
Presenter affiliation: Pennsylvania State University, University Park, Pennsylvania. 106

Exploring functional divergence in paralogs using embeddings from protein language models
Denise Le, Alan M. Moses.
Presenter affiliation: University of Toronto, Toronto, Canada. 107

A fully assembled, phased yucatan reference genome enables accurate on-target and off-target analysis
Feng Li, Xiaoyun Guo, Shiyi Yin, Owen Pearce, Wing-On Ng, Chris Chao, Josh Stirba, Matthew Pandelakis, Jacob V. Layer, Wenning Qin, Ranjith P. Anand, Sagar Chhangawala.
Presenter affiliation: eGenesis, Inc, Cambridge, Massachusetts. 108

Analysis of pancreatic cancer risk variants using long read sequencing
Qiuhui Li, Carolina Montano, Jessica Hosea, Luke Morina, Bohan Ni, Justin Paschall, Beth Marosy, Michelle Kokosinski, Jessica Gearhart, Brian Craig, Alan Scott, David Mohr, Michelle Mawhinney, David McKean, Nicholas Roberts, Zhanmo Ni, Alexis Battle, Kimberly Doheny, Winston Timp, Michael Schatz, Alison Klein.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 109

< Back

Genetic and dietary modulators of the inflammatory response in the gastro-intestinal tract of the BXD mouse genetic reference population
Xiaoxu Li, Jean-David Morel, Giorgia Benegiamo, Johanne Poisson, Alexis Bachmann, Alexis Rapin, Jonathan Sulc, Evan Williams, Alessia Perino, Kristina Schoonjans, Maroun Bou Sleiman, Johan Auwerx.
Presenter affiliation: École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. 110

Systems genetics of metabolic health in the BXD mouse genetic reference population
Xiaoxu Li, Jean-David Morel, Alessia De Masi, Amélia Lalou, Jonathan Sulc, Giorgia Benegiamo, Johanne Poisson, Yasmine Liu, Arwen W. Gao, Maroun Bou Sleiman, Johan Auwerx.
Presenter affiliation: École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. 111

Measuring, visualizing and diagnosing reference bias with biastools
Mao-Jan Lin, Sheila Iyer, Nae-Chyun Chen, Ben Langmead.
Presenter affiliation: Johns Hopkins, Baltimore, Maryland. 112

Using Better Base Quality (BBQ) to detect low-frequency somatic mutations accurately
Yixin Lin, Carmen Oroperv, Claus L. Andersen, Mikkel H. Schierup, Asger Hobolth, Thomas Bataillon, Kristian Almstrup, Søren Besenbacher.
Presenter affiliation: Aarhus University, Aarhus, Denmark. 113

DNA bendability regulates transcription factor pioneer binding to nucleosomes
Xiao Liu, Luca Mariani, Martha L. Bulyk.
Presenter affiliation: Harvard Medical School, Boston, Massachusetts; Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts. 114

PathwayGAT—A method to trace back biological interactions from microbes to host phenotypes
Shaoke Lou, Weihao Zhao, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 115

< Back

Spatial transcriptomics data analyses revealed cancer-endothelial cell communication in hepatocellular carcinoma
Chenyue Lu, Amaya Pankaj, Michael Raabe, Cole Nawrocki, Ann Liu, Nova Xu, Bidish K. Patel, Matthew J. Emmett, Avril K. Coley, Cristina R. Ferrone, Vikram Deshpande, Irun Bhan, Yujin Hoshida, David T. Ting, Martin A. Aryee, Joseph W. Franses.
Presenter affiliation: Massachusetts General Hospital, Boston, Massachusetts; Harvard-MIT, Cambridge, Massachusetts; Dana-Farber Cancer Institute, Boston, Massachusetts; Broad Institute, Cambridge, Massachusetts. 116

The Kraken protocol in action—Identifying potential pathogens in Ugandan individuals with unexplained acute febrile illness
Jennifer Lu, Abraham J. Kandathil, Raghavendran Anantharam, Kenneth Kobba, Paul W. Blair, Matthew L. Robinson, Edgar C. Ndawula, Francis Kakooza, Mohammed Lamorde, David L. Thomas, Martin Steinegger, Yukari C. Manabe, Steven L. Salzberg.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 117

Towards a cloud-agnostic scalable ecosystem for open genomic data science with Bioconductor and Galaxy
Alexandru Mahmoud, Enis Afgan, Dirk Eddelbuettel, Marcel Ramos, Ludwig Geistlinger, Jen Wokaty, Lori S. Kern, Davide Risso, Sean Davis, Levi Waldron, Vincent J. Carey.
Presenter affiliation: Harvard Medical School/Mass General Brigham, Boston, Massachusetts. 118

Leveraging representations of multi-species networks and ontologies to improve gene classification
Keenan Manpearl, Remy Liu, Christopher Mancuso, Arjun Krishnan.
Presenter affiliation: University of Colorado-Anschutz, Aurora, Colorado. 119

Tools and workflows enabling scaling of genome assembly across the Tree of Life
Shane A. McCarthy, on behalf of Tree of Life Programme, and Darwin Tree of Life Project.
Presenter affiliation: Wellcome Sanger Institute, Hinxton, United Kingdom; University of Cambridge, Cambridge, United Kingdom. 120

< Back

Understanding and mitigating amplification biases in single-cell DNA sequencing for accurate genotype calling
Aleksei Mikhalchenko, Nuria Marti Gutierrez, Daniel Frana, Paula Amato, Shoukhrat Mitalipov.
Presenter affiliation: Oregon Health & Science University, Portland, Oregon. 121

Quality assessment of human splice site annotation based on conservation in 470 species
Ilia Minkin, Steven L. Salzberg.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 122

OTB (Only The Best genome assembly tools)—A phased genome assembly nextflow pipeline
David C. Molik, Amanda Stahlke.
Presenter affiliation: USDA Agricultural Research Service, Manhattan, Kansas. 123

Cell-type deconvolution with long-read, single molecule methylation
Luke B. Morina, Courtney Hall, Jessica Hosea, Roham Razaghi, Winston Timp.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 124

Redcarpet—A tool for rapid recombination detection amidst expanding genomic databases
Ahmed M. Moustafa, Erin Theiller, Arnav Lal, Andries Feder, Apurva Narechania, Paul J. Planet.
Presenter affiliation: Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; University of Pennsylvania, Perelman School of Medicine, Philadelphia. 125

Improving the accuracy of miro-variant detection in whole genome sequencing data
Shandukani Mulaudzi, Maximillian Marin, Maha Farhat.
Presenter affiliation: Harvard Medical School, Boston, Massachusetts. 126

Lancet2—Improved performance and genotyping of somatic variants using localized genome graphs
Rajeeva Lochan Musunuri, Bryan Zhu, Dickson Chung, Shreya Sundar, Adam Novak, Timothy Chu, Jennifer Shelton, Nicolas Robine, Giuseppe Narzisi.
Presenter affiliation: New York Genome Center, New York, New York. 127

< Back

Exploring gene expression properties of the human brain with BITHub
Urwah Nawaz, Kieran Walsh, Gavin Sutton, Jozef Gecz, Irina Voineagu.
Presenter affiliation: The University of Adelaide, Adelaide, Australia. 128

CADD v1.7—Using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions
Max Schubach, Thorben Maass, Lusine Nazaretyan, Sebastian Röner, Martin Kircher.
Presenter affiliation: Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany. 129

Genomic characterization and global contextualization of ESBL-producing *E. coli* from pediatric patients in Qatar
Matthew Nguyen, Clement Tsui, Patrick Tang, Andres Perez-Lopez, William Hsiao.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland; University of British Columbia, Vancouver, Canada; Simon Fraser University, Vancouver, Canada. 130

POSTER SESSION II

Evolutionary insights into primate sex chromosomes—Sharing and gene content of palindromes
Karol Pal, Robert Harris, Monika Cechova, Sergey Koren, Sergey Nurk, Huiqing Zeng, Arang Rhie, Melissa A. Wilson, Brandon D. Pickett, Brendan J. Pinto, Prajna Hebbar, Mark Diekhans, Benedict Paten, Evan Eichler, Adam M. Phillippy, Kateryna D. Makova.
Presenter affiliation: Pennsylvania State University, University Park, Pennsylvania. 131

Diagnosis of ocular infection using Nanopore metagenomic sequencing
Dongwoo Park, Junwon Lee, Hyun Goo Kang, Han Jeong, Sangwoo Kim, Min Kim.
Presenter affiliation: Yonsei University College of Medicine, Seoul, South Korea. 132

ContextSV—A novel computational method for calling structural variants and integrating information across sequencing platforms
Jonathan E. Perdomo, Kai Wang.
Presenter affiliation: Drexel University, Philadelphia, Pennsylvania; Children's Hospital of Philadelphia, Philadelphia, Pennsylvania. 133

< Back

The effect of dynamic pressure on the gene expression of *Deinococcus radiodurans* R1
Cesar A. Perez Fernandez, Lily Zhao, K.T. Ramesh, Jocelyne DiRuggiero.
Presenter affiliation: Universidad Privada Boliviana, Santa Cruz, Bolivia; Johns Hopkins University, Baltimore, Maryland. 134

Sequence-associated mechanistic insights of DNA fragility
Patrick Pflughaupt, Aleksandr B. Sahakyan.
Presenter affiliation: University of Oxford, Oxford, United Kingdom. 135

Prioritization of fluorescence in situ hybridization (FISH) probes for differentiating primary sites of neuroendocrine tumors with machine learning
Lucas Pietan, Hayley Vaughn, James Howe, Andrew Bellizzi, Ben Darbro, Terry Braun, Tom Casavant.
Presenter affiliation: University of Iowa, Iowa City, Iowa. 136

Integrative modeling of activity, responsiveness and contact (ARC) reveals enhancer-gene connections using single-cell data
Wei-Lin Qiu, Maya Sheth, Rosa Ma, Andreas Gschwind, Anthony Tan, Hjörleifur Einarsson, Danilo Dubocanin, Evelyn Jagoda, Lars Steinmetz, Anshul Kundaje, Jesse Engreitz, Robin Andersson.
Presenter affiliation: University of Copenhagen, Copenhagen, Denmark; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 137

Genomic analysis of novel hydrocarbonoclastic *Chryseobacterium oranimense* strain COTT, a putative bioremediation agent with multi-drug resistance and enzymes for industry
Amanda C. Ramdass, Sephra N. Rampersad.
Presenter affiliation: The University of the West Indies, St. Augustine, Trinidad And Tobago. 138

Deducing the evolution of allorecognition and primordial immunity in Cnidarians
Alberto M. Rivera, Andy Baxevanis.
Presenter affiliation: National Institutes of Health, Bethesda, Maryland. 139

On-going sequencing and analysis of a new tumor cell line for development of a genome in a bottle tumor/normal benchmark
Gail Rosen, Justin Wagner, Jennifer McDaniel, Andrew Liss, Justin Zook.
Presenter affiliation: Drexel University, Philadelphia, Pennsylvania. 140

< Back

A novel structural variant detection pipeline in cancer genomes—The personalized matched-control reference-based approach
Yoshitaka Sakamoto, Masahiro Sugawa, Ai Okada, Yotaro Ochi, Yosuke Tanaka, Yasunori Kogure, Kenichi Chiba, Wataru Nakamura, Junji Koya, Hiroyuki Mano, Seishi Ogawa, Keisuke Kataoka, Yuichi Shiraishi.
Presenter affiliation: National Cancer Center Research Institute, Tokyo, Japan. 141

Integrating multiple transcriptome-based methods to repurpose drugs for infectious diseases
Kewalin Samart, Amy Tonielli, Arjun Krishnan, Janani Ravi.
Presenter affiliation: University of Colorado Anschutz Medical Campus, Aurora, Colorado. 142

Identifying niche-specific genetic adaptations in *Acinetobacter baumannii*
Sydney Sanchez, Gisela Di Venanzio, Mario Feldman, Federico Rosconi, Juan C. Ortiz-Marquez.
Presenter affiliation: Boston College, Boston, Massachusetts. 143

Biomedical and biological applications of machine learning using Galaxy Project
Michelle Savage, Michael Schatz, Galaxy Project.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 144

Cell lineage inference using patterns of DNA damage
Lucy C. Scott, Cameron Wyatt, Elizabeth Patton, Martin S. Taylor.
Presenter affiliation: University of Edinburgh, Edinburgh, United Kingdom. 145

A surrogate modeling framework for interpreting deep neural networks in functional genomics
Evan Seitz, Justin Kinney, Peter Koo.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 146

Democratizing access to genomics and data science education through the cloud—The GDSCN success story
Shurjo K. Sen, Genomic Data Science Community Network.
Presenter affiliation: NHGRI, Bethesda, Maryland. 147

< Back

Classification of antibiotic resistance of *Mycobacterium tuberculosis* via linear and non-linear machine learning
Mohammadali Serajian, Simone Marini, Jarno N. Alanko, Noelle R. Noyes, Mattia Prosperi, Christina Boucher.
Presenter affiliation: University of Florida, Gainesville, Florida. 148

EASTR—Identifying and eliminating systematic alignment errors in multi-exon genes
Ida Shinder, Richard Hu, Hyun Joo Ji, Kuan-Hao Chao, Mihaela Pertea.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 149

Systematic discovery of splice-site creating variants from massive publicly available transcriptome sequencing data
Yuichi Shiraishi, Naoko Iida, Ai Okada, Kenichi Chiba.
Presenter affiliation: National Cancer Center Research Institute, Tokyo, Japan. 150

Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv
Yuichi Shiraishi, Junji Koya, Kenichi Chiba, Ai Okada, Yasuhito Arai, Yuki Saito, Tatsuhiro Shibata, Keisuke Kataoka.
Presenter affiliation: National Cancer Center Research Institute, Tokyo, Japan. 151

Sigmoni—Classification of nanopore signal with a compressed pangenome index
Vikram S. Shivakumar, Omar Y. Ahmed, Sam Kovaka, Mohsen Zakeri, Ben Langmead.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 152

Claspy—Cell line authentication with STRs in Python
Alaina G. Shumate, Rebecca N. Mitchell, Daniel S. Standage.
Presenter affiliation: National Bioforensic Analysis Center, Frederick, Maryland. 153

AI approach for de novo genome assembly
Lovro Vrcek, Xavier Bresson, Thomas Laurent, Martin Schmitz, Kenji Kawaguchi, Mile Šikic.
Presenter affiliation: Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (ASTAR), Singapore; University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia. 154

< Back

Mitochondrial mutation and dysfunction in high grade serous ovarian cancer
Ryan P. Silk, Alison M. Meynert, Ailith Ewing, Brian Dougherty, Patricia Roxburgh, Charlie Gourley, Colin A. Semple.
Presenter affiliation: University of Edinburgh, Edinburgh, United Kingdom. 155

TreeVal—Data generation for the curation of chromosome-scale genomes
Damon Pointon, Ying Sims, William Eagles, Jonathan Wood, Shane McCarthy.
Presenter affiliation: Wellcome Sanger Institute, Cambridge, United Kingdom. 156

Simplifying and improving single-cell gene expression analysis with Piccolo
Amartya Singh, Hossein Khiabanian, Daniel Herranz.
Presenter affiliation: Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey. 157

TrioKala—A trio co-assembly approach for *de novo* variant detection
Steven J. Solar, Carlos R. Ferreira, Sergey Koren, Dmitry Antipov, Mikko Rautiainen, Adam Phillippy.
Presenter affiliation: National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland. 158

MicroHapDB—A comprehensive catalog of human microhaplotype variation
Daniel Standage.
Presenter affiliation: National Bioforensic Analysis Center, Frederick, Maryland. 159

CRISPRsc—Pooled CRISPR screening with single-cell transcriptome resolution
Tiana Stastny, Maximilian Blanck, Matthew Perkett, John McGonigle, Simon Scrace, Carlos le Sage, Anja Smith.
Presenter affiliation: Revvity, Lafayette, Colorado. 160

Innovation, constraint, and the evolution of genetic networks in major eukaryotic lineages
Jacob L. Steenwyk, Maxwell C. Coyle, Noah Bradley, Chris T. Hittinger, Antonis Rokas, Nicole King.
Presenter affiliation: UC-Berkeley / HHMI, Berkeley, California. 161

< Back

SomaMutDB—A database of somatic mutations in normal human tissues
Shixiang Sun, Yujue Wang, Alexander Maslov, Xiao Dong, Jan Vijg.
Presenter affiliation: Albert Einstein College of Medicine, Bronx, New York. 162

Investigation of tissue-specific transcriptome at isoform-level in GTEx and TCGA datasets
Pallavi Surana, Pratik Dutta, Chirayush Patel, Jayendra Kantipudi, Roshan R Yedla, Sankara Kota, Ramana V Davuluri.
Presenter affiliation: Stony Brook University, Stony Brook, New York. 163

ModDotPlot—A rapid and interactive visualization of tandem repeats
Alexander P. Sweeten, Michael C. Schatz, Adam M. Phillippy.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland; National Human Genome Research Institute, Bethesda, Maryland. 164

Estimating background protein signals to enhance data normalization in CITE-Seq
Cerag Oguztuzun, Tamas Ryszard Sztanka-Toth, Alex Javidi.
Presenter affiliation: Data Science and Digital Health, Janssen R&D, Neuss, Germany. 165

New computational methods for the analysis of transcription factor CUT&RUN data
Luomeng Tan, Coby Viner, Michael M. Hoffman.
Presenter affiliation: University Health Network, Toronto, Canada. 166

Less is more—Trimming long massively parallel sequence reads can improve mapping rates and depth of coverage
Jamie K. Teer, Jane C. Figueiredo, Stephanie L. Schmit.
Presenter affiliation: H. Lee Moffitt Cancer Center, Tampa, Florida. 167

Guiding single-cell RNA-seq clustering with rank-based metrics
Christopher V. Thai, Hossein Khiabanian, Daniel Herranz.
Presenter affiliation: Rutgers University, New Brunswick, New Jersey. 168

< Back

Variant analysis in an inbred rat population—A lesson from the Hybrid Rat Diversity Panel
Monika Tutaj, Akiko Takizawa, Lynn Malloy, Rebecca Schilling, Kent C. Brodie, Jeffrey L. De Pons, Wendy M. Demos, Thomas G. Hayman, Mary L. Kaldunski, Stan J. Laulederkind, Jennifer R. Smith, Marek A. Tutaj, Mahima Vedi, Shur-Jen Wang, Anne E. Kwitek, Melinda R. Dwinell.
Presenter affiliation: Medical College of Wisconsin, Milwaukee, Wisconsin. 169

Accurate comparison of insertion and deletion mutation rates using sequence composition correction with novel sequence ambiguity scoring
Jan C. Verburg, Martin S. Taylor.
Presenter affiliation: Medical Research Council Human Genetics Unit, Edinburgh, United Kingdom. 170

Pbfusion—Detecting gene fusions and other transcriptional abnormalities using PacBio HiFi data
Roger Volden, Daniel Baker, Zev Kronenberg, Aaron Gillmor, Ted Verhey, Michael Monument, Donna Senger, Harsharan Dhillon, Jason Underwood, Elizabeth Tseng, Primo Baybayan, Michael A. Eberle, Jonas Korlach, Sorana Morrissy.
Presenter affiliation: PacBio, Menlo Park, California. 171

FAIR Bioheaders—Fair Header Reference genome (FHR)
Adam J. Wright, David C. Molik.
Presenter affiliation: Ontario Institute for Cancer Research, Toronto, Canada. 172

Examining chromatin heterogeneity through PacBio long-read sequencing of M.EcoGII methylated genomes—An m^{6}A detection efficiency and calling bias correcting pipeline
Zhuwei Xu, Allison F. Dennis, David J. Clark.
Presenter affiliation: NIH, Bethesda, Maryland. 173

scReadSim—S single-cell RNA-seq and ATAC-seq read simulator
Guan'ao Yan, Dongyuan Song, Jingyi Jessica Li.
Presenter affiliation: University of California, Los Angeles, Los Angeles, California. 174

A metagenomics genome-phenome association (MetaGPA) study reveals 2-aminoadenine (dZ) biosynthetic pathway in unculturable phage metagenome
Weiwei Yang, Shuangyong Xu, Laurence Ettwiller.
Presenter affiliation: New England Biolabs, Ipswich, Massachusetts. 175

< Back

***Stator*---High order expression dependencies finely resolve cryptic states and subtypes in scRNA-seq data**
Yuelin Yao, Abel Jansma, Jareth Wolfe, Luigi Del Debbio, Sjoerd Beentjes, Chris Ponting, Ava Khamseh.
Presenter affiliation: University of Edinburgh, Edinburgh, United Kingdom. 176

Model-based characterization of the equilibrium dynamics of transcription initiation and promoter-proximal pausing in human cells
Yixin Zhao, Lingjie Liu, Rebecca Hassett, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 177

Investigating mosaic structural variations across thousands of genomes with STIX
Xinchang Zheng, Ryan M. Layer, Fritz J. Sedlazeck.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 178

Data beats machine learning for genome annotation
Aleksey V. Zimin.
Presenter affiliation: Johns Hopkins University, Baltimore, Maryland. 179

< Back

AUTHOR INDEX

Absolon, Dominic, 49
Abueg, Linelle, 106
Adameyko, Igor, 81
Adams, Jenea I., 51
Afgan, Enis, 118
Ahmad, Tanveer, 36, 52, 61
Ahmadian, Mansooreh, 92
Ahmed, Omar, 10, 16, 96, 152
Aitken, Stuart, 79
Alanko, Jarno N., 148
Albert, Frank W., 53
Allert, Mattea, 53
Alleva, Benjamin, 71
Almodaresi, Fatemeh, 54
Almstrup, Kristian, 113
Alonge, Michael, 14
Amato, Paula, 121
Anand, Ranjith P., 108
Anantharam, Raghavendran, 117
Andersen, Claus L., 113
Anderson, Kim, 95
Andersson, Robin, 137
Antipov, Dmitry, 44, 158
Antunes, Diogo L., 11
Arabzadeh, Mona, 55
Arai, Yasuhito, 151
Armstrong, Joel, 69
Arruda, Ana Luiza, 40
Aryee, Martin A., 116
Asri, Mobin, 42, 65
Atmaramani, Rahul, 29
Auwerx, Johan, 110, 111
Avihai, Byron, 56

Bachmann, Alexis, 110
Baker, Daniel, 171
Balogh, Petra, 40
Bang, Seungmin, 66
Baradaran, Morteza, 57
Bataillon, Thomas, 113
Battle, Alexis, 109
Baxevanis, Andy, 139
Baybayan, Primo, 171
Bayraktar, Omer, 25
Beentjes, Sjoerd, 176
Behjati, Sam, 40
Belay, Kassaye H., 58
Bellizzi, Andrew, 136
Benegiamo, Giorgia, 110, 111
Benoit, Matthias, 14
Besenbacher, Søren, 113
Bhan, Irun, 116
Bick, Alexander, 95
Bischoff, Philip, 103
Bisognano, John, 29
Blair, Paul W., 117
Blanck, Maximilian, 160
Blankenberg, Daniel, 59
Blekhman, Ran, 53
Bou Sleiman, Maroun, 110, 111
Boucher, Christina, 10, 13, 148
Boyer, Jacob, 83
Bradley, Noah, 161
Bradshaw, Michael S., 60
Brajuka, Nadolina, 106
Braun, Terry, 136
Bresson, Xavier, 154
Brewer, Daniel S., 1
Brodie, Kent C., 169
Brown, C Titus, 50
Brown, Nathaniel, 16
Brown, Stuart, 79
Brudno, Michael, 20
Bryant, Asher, 36, 61
Buchholz, karin, 84
Buen Abad Najar, Carlos F., 27
Bulyk, Martha L., 114
Burg, Jonathon, 91
Bzikadze, Andrey, 42

Camerini-Otero, R. Daniel, 71
Campo, Elias, 54
Cao, Lei, 62
Carey, Vincent J., 118
Carpenter, Anne E., 22
Casavant, Tom, 136
Casper, Jonathan, 69
Cavalier, Sheridan H., 63
Cechova, Monika, 42, 131
Chandra, Ghanshyam, 12

< Back

Chao, Chris, 108
Chao, Kuan-Hao, 64, 149
Chapal, Noa, 98
Chen, Frances, 95
Chen, Kelvin, 19
Chen, Nae-Chyun, 112
Chen, Owen, 29
Chen, Yuhang, 28
Cheng, Haoyu, 65
Chesi, Alessandra, 31
Chhangawala, Sagar, 108
Chiba, Kenichi, 141, 150, 151
Chikhi, Rayan, 43, 70
Chilton, John M., 46
Chin, Hang Geong, 91
Choi, Eunwoo, 66
Choi, Jinmyung, 67
Chougule, Kapeel, 68
Chowdhury, Murad, 39, 41
Christensen, Erik, 20
Chu, Ci, 29
Chu, Timothy, 127
Chung, Dickson, 127
Churchman, Michael, 79
Clark, David J., 173
Clarke, Declan, 21
Clawson, Hiram, 69
Coley, Avril K., 116
Conrad, Christian, 103
Cooper, Colin S., 1
Coraor, Nate, 46
Corbett-Detig, Russ, 74
Cowles, Martis, 91
Coyle, Maxwell C., 161
Craig, Brian, 109
Cumbo, Fabio, 59
Curry, Kristen D., 70

Dahiya, Daisy, 71
Dai Yang, Karren, 19
Damaraju, Nikhita, 37, 45
Daniels, Craig, 54
Darbro, Ben, 136
Das, Arun, 72
Davis, John, 73
Davis, Sean, 118
Davuluri, Ramana V., 32, 163
Day, Chi-Ping, 24, 85
De Masi, Alessia, 111
De Pons, Jeffrey L., 169
Dean, Michael, 36, 61
Degrandi, Daniel, 84
Del Debbio, Luigi, 176
Demerath, Ellen W., 53
Demos, Wendy M., 169
Deng, Dori Z., 74
Dennis, Allison F., 173
Deshpande, Vikram, 116
Dhillon, Harsharan, 171
Di Venanzio, Gisela, 143
Dias, Paulo J., 11
Diaz Ortiz, Gerardo R., 75, 76
Diaz-Mejia, Javier, 20
Diaz-Navarro, Ander, 54
Diekhans, Mark, 131
Dilthey, Alexander, 84
DiRuggiero, Jocelyne, 134
Doheny, Kimberly, 109
Dong, Xiao, 162
Donmez, Ataberk, 3, 36, 61
Dooley, Helen M., 74
Dougherty, Brian, 155
Duarte, Sofia O., 11
Dubinkina, Veronika, 77
Dubocanin, Danilo, 137
Dutta, Pratik, 32, 163
Dwinell, Melinda R., 169

Eagles, William, 156
Eberle, Michael A., 171
Ebler, Jana, 15
Eddelbuettel, Dirk, 118
Eddy, Sean R., 17
Eichler, Evan E., 37, 42, 45, 90, 131
Einarsson, Hjörleifur, 137
Eizenga, Jordan M., 42
Emani, Prashant, 21
Emmett, Matthew J., 116
Engreitz, Jesse, 137
Eraslan, Basak, 19
Erdogdu, Beril, 78
Esiri-Bloom, Edward, 79
Esteve, Pierre, 91
Ettwiller, Laurence, 175
Ewer, Lexie, 29

< Back

Ewing, Ailith, 79, 155
Ezell, Ryan, 91

Fan, Yu, 62
Farhat, Maha, 126
Farooqi, Midhat, 36, 61
Feder, Andries, 2, 125
Feldman, Mario, 143
Feng, Xiaowen, 80, 88
Ferreira, Carlos R., 158
Ferretti, Pamela, 53
Ferro, Eddie, 13
Ferrone, Cristina R., 116
Fields, David A., 53
Figueiredo, Jane C., 167
Fitzgerald, Blaine, 14
Fleishman, Suzanne M., 50
Formenti, Giulio, 106
Frana, Daniel, 121
Franses, Joseph W., 116
Frary, Amy, 14
Frost, H. Robert, 23
Fulton, Robert S., 42

Gagie, Travis, 10, 16
Gagnon, Edeline, 14
Gale, Cheryl A., 53
Galey, Miranda, 37, 45
Gallardo-Alba, Cristóbal, 106
Ganesan, Shridar, 55
Gao, Arwen W., 111
Gao, Teng, 81
Gardner, Joshua, 36, 61
Garrido Martin, Diego, 21
Ge, Peter, 82
Ge, Yuchen, 1
Gearhart, Jessica, 109
Gecz, Jozef, 128
Geiger-Schuller, Katie, 19
Geistlinger, Ludwig, 118
Genereux, Diane, 95
Gentile, Iacopo, 14
Geraghty, Sara E., 83
Gerstein, Mark, 21, 115
Gerton, Jennifer L., 42
Gibson, Sophia B., 37, 45
Gignoux, Chris, 41
Gihawi, Abraham, 1
Gillis, Jesse, 14
Gillmor, Aaron, 171
Gitman, Melissa, 62
Givmanesh, Atieh, 29
Goffena, Joy, 37, 45
Gohar, Yomna, 84
Gonzalez, Jairo N., 69
Goodwin, Sara, 14
Goonasekera, Nuwan A., 46
Goretsky, Anton, 85
Gourley, Charlie, 79, 155
Govada, Pravallika, 86
Grant, Struan F., 31
Green, Richard E., 74
Grimes, Graeme, 79
Grüning, Bjorn, 106
Gschwind, Andreas, 137
Gulhan, A. Burak, 87
Guo, Xiaoyun, 108
Guo, Yujie, 88
Gustafson, J. (Gus), 37, 45

Hacohen, Nir, 19
Haeussler, Maximilian, 69, 87
Hall, Courtney, 124
Halldorsson, Bjarni, 41
Hankenson, Kurt D., 31
Harazono, Yoritaka, 89
Hardarson, Marteinn, 41
Hardy, Marie-Pierre, 54
Harris, Adam, 95
Harris, Andrew J., 90
Harris, Robert, 131
Hassett, Rebecca, 177
Haussler, David, 69
Hayman, Thomas G., 169
Heath, Lenwood, 58
Hebbar, Prajna, 131
Heidari, Elyas, 25
Heinzel, Andreas, 81
Heisel, Timothy, 53
Hendelman, Anat, 14
Herranz, Daniel, 56, 157, 168
Herrington, C Simon, 79
Hickman, Allison, 91
Hicks, Parker, 92
Hicks, Stephanie C., 78
Hildebrand, Mark, 41

< Back

Hiller, Michael, 69
Hinrichs, Angie S., 69
Hittinger, Chris T., 161
Hobolth, Asger, 113
Hoekzema, Kendra, 37, 45, 90
Hoffman, Michael M., 166
Holcik, Laurenz, 4
Hong, Jiyoon, 66
Hook, Paul, 48, 63
Hosea, Jessica, 109, 124
Hoshida, Yujin, 116
Howe, James, 136
Hsiao, William, 130
Hu, Richard, 149
Huang, Neng, 93
Huang, Rongting, 94
Huang, Xianjie, 94
Huang, Yina, 23
Huang, Yuanhua, 94
Huganir, Richard, 63
Hülse, Lisanna, 84
Husic, Mia, 20
Husted, Christopher, 95
Hwang, Shinwon, 101
Hwang, Stephen, 14, 96

Iida, Naoko, 150
Irber, Luiz, 50
Isaev, Keren, 97
Iyer, Sheila, 112

Jagoda, Evelyn, 137
Jain, Chirag, 12
Jain, Miten, 37, 45
Jansma, Abel, 176
Jarvis, Erich D., 106
Javidi, Alex, 165
Jenike, Katharine, 14, 48
Jensen, Matthew, 21
Jeong, Han, 132
Ji, Hyun Joo, 149
Ji, Yanrong, 32
Johnson, Bay, 29
Johnson, Kelsey, 53
Joshi, Jayadev, 59

Kadam, Aditee, 98
Kadash-Edmondson, Kathryn E., 51
Kakooza, Francis, 117
Kaldunski, Mary L., 169
Kandathil, Abraham J., 117
Kandemir, Mahmut, 87
Kang, Hyun Goo, 132
Kantipudi, Jayendra, 163
Karayel Eren, Ozge, 19
Karlsson, Elinor, 95
Kasahara, Masahiro, 89
Kastriti, Maria E., 81
Kataoka, Keisuke, 141, 151
Katharina, Raba, 84
Katsoula, Georgia, 40
Kaushansky, Nathali, 98
Kawaguchi, Kenji, 154
Kaykas, Ajamete, 29
Kazantseva, Ekatarina, 3
Kelso, Janet, 47
Kent, W. James, 69
Keogh, Michael-Christopher, 91
Kern, Lori S., 118
Keskus, Ayse, 36, 61, 85
Khamseh, Ava, 176
Khan, Jawad, 41
Kharchenko, Peter V., 81
Khiabanian, Hossein, 55, 56, 157, 168
Kim, Daehwan, 7
Kim, Daniel, 99
Kim, Juhyun, 100
Kim, Min, 132
Kim, Sangwoo, 66, 101, 132
Kim, Yongjun, 101
King, Nicole, 161
Kinney, Justin, 146
Kircher, Martin, 129
Kirilenko, Bogdan M., 69
Klein, Alison, 109
Knights, Dan, 53
Knowles, David A., 97
Knuth, Jordan, 45
Ko, Byung June, 106
Kobba, Kenneth, 117
Kogure, Yasunori, 141
Kokosinski, Michelle, 109

< Back

Kolmogorov, Mikhail, 3, 36, 52, 61, 85
Kolora, Rohit, 102
Kong, Yimeng, 62
Koo, Peter, 146
Koo, Tonny, 62
Koren, Sergey, 42, 44, 65, 131, 158
Korlach, Jonas, 171
Kota, Sankara, 163
Kovaka, Sam, 14, 48, 152
Koya, Junji, 141, 151
Kramer, Melissa, 14
Krasheninnikova, Ksenia, 49
Krieger, Teresa G., 103
Krishnan, Arjun, 92, 119, 142
Kronenberg, Zev, 171
Ksiezarek, Magdalena, 62
Kucher, Natalie, 104
Kuhn, Robert M., 69
Kuksenko, Olena, 19
Kumar, Pankaj, 105
Kumar, Sachin, 54
Kundaje, Anshul, 137
Kutschera, Eric, 51
Kwitek, Anne E., 169

Lal, Arnav, 125
Lalou, Amélia, 111
Lamorde, Mohammed, 117
Langmead, Ben, 7, 10, 13, 14, 16, 35, 96, 112, 152
Lariviere, Delphine, 106
Laulederkind, Stan J., 169
Laurent, Thomas, 154
Lautwein, Tobias, 84
Lawrence, John E G, 40
Layer, Jacob V., 108
Layer, Ryan, 39, 41, 57, 178
Lazareva, Olga, 25
le Sage, Carlos, 160
Le, Denise, 107
Lee, Alicia, 29
Lee, Brian T., 69
Lee, Christopher M., 69
Lee, HeeSeung, 66
Lee, Junwon, 132
Lehmitz, Jonas, 103
Lehner, Ben, 30
Leitmann, Bobby, 29
Li, Feng, 108
Li, Heng, 65, 80, 88, 93
Li, Jingyi Jessica, 174
Li, Qiuhui, 109
Li, Weijian, 32
Li, Xiaoxu, 110, 111
Li, Yang I., 27
Liebig, Johannes, 103
Lin, Mao-Jan, 112
Lin, Yixin, 113
Lippman, Zachary B., 14
Liss, Andrew, 140
Liu, Ann, 116
Liu, Han, 32
Liu, Lingjie, 177
Liu, Remy, 119
Liu, Xiao, 114
Liu, Yasmine, 111
Liu, Yuelin, 85
Ljungström, Viktor, 81
Lloyd, David, 29
Loh, Po-Ru, 81
Loose, Matt, 37, 45
Lopez, Romain, 19
Lou, Hong, 36, 61
Lou, Shaoke, 115
Lu, Chenyue, 116
Lu, Jennifer, 1, 82, 117
Lu, Zhenyuan, 68
Lucas, Julian, 42, 65
Luo, Ping, 20
Lust, Hannah, 23
Lyons, Leslie A., 90

Ma, Rosa, 137
Maass, Thorben, 129
Mahalanabis, Alaina, 20
Mahmoud, Alexandru, 118
Makinen, Veli, 9
Makova, Kateryna D., 131
Malikic, Salem, 24, 85
Malloy, Lynn, 169
Manabe, Yukari C., 117
Mancuso, Christopher, 119
Mano, Hiroyuki, 141
Manpearl, Keenan, 119

< Back

Mao, Alan, 64
Mariani, Luca, 114
Marie Sá Figueiredo, Agnes, 2
Marie, Kerrie L., 24
Marin, Maximillian, 126
Marini, Simone, 148
Markovic, Christopher, 42
Marosy, Beth, 109
Marschall, Tobias, 15, 42
Marti Gutierrez, Nuria, 121
Maslov, Alexander, 162
Mawhinney, Michelle, 109
Mazloom, Reza, 58
McBride, Matthew, 83
McCarthy, Shane A., 49, 120, 156
McCombie, W. Richard, 14
McDaniel, Jennifer, 140
McGonigle, John, 160
McKean, David, 109
McNulty, Brandy, 36, 42, 61
Mead, Edward A., 62
Meek, Isaac, 91
Megquier, Kate, 95
Meireles de Sousa, Ana, 19
Merlino, Glenn, 24, 85
Meynert, Alison, 79, 155
Miga, Karen, 36, 42, 61
Mikhalchenko, Aleksei, 121
Milano, Rose, 36, 61
Miller, Danny E., 37, 45
Minden, Mark, 98
Minkin, Ilia, 122
Mitalipov, Shoukhrat, 121
Mitchell, Rebecca N., 153
Mohr, David, 109
Molik, David C., 123, 172
Molloy, Erin, 85
Moltke, Ida, 38
Montano, Carolina, 109
Monteiro, Gabriel A., 11
Monument, Michael, 171
Moreira, Leonilde M., 11
Morel, Jean-David, 110, 111
Morina, Luke, 48, 109, 124
Morrissy, Sorana, 171
Moses, Alan M., 107
Moustafa, Ahmed M., 2, 125
Mulaudzi, Shandukani, 126
Mun, Taher, 35
Murphy, William J., 90
Musunuri, Rajeeva Lochan, 127

Naidas, Alaine, 20
Nakamura, Wataru, 141
Naor, Hadas, 98
Narechania, Apurva, 125
Narzisi, Giuseppe, 127
Nassar, Luis R., 69
Nawaz, Urwah, 128
Nawrocki, Cole, 116
Nazaretyan, Lusine, 129
Ndawula, Edgar C., 117
Negi, Shloka, 42
Nekrutenko, Anton, 46, 87, 106
Ng, Wing-On, 108
Nguyen, Eric, 21
Nguyen, Matthew, 130
Ni, Bohan, 109
Ni, Mi, 62
Ni, Zhanmo, 109
Nirmal, Ajit J., 94
Nitzan, Mor, 33
Novak, Adam, 127
Noyes, Noelle R., 75, 76, 148
Nurk, Sergey, 131

Oberbauer, Rainer, 81
Ochi, Yotaro, 141
Ogawa, Seishi, 141
Oguztuzun, Cerag, 165
Ojukwu, Christopher E., 39
Okada, Ai, 141, 150, 151
Oliva, Marco, 13
Olson, Andrew, 68
Oroperv, Carmen, 113
Ortiz-Marquez, Juan C., 143
Ostrovsky, Alex, 106
Ou, Shujun, 14

Pahl, Matthew C., 31
Pal, Karol, 131
Palmada-Flores, Marc, 106
Pandelakis, Matthew, 108
Pani, Samarendra, 15
Pankaj, Amaya, 116

< Back

Park, Dongwoo, 132
Park, Jimin, 36, 61
Park, Peter J., 81
Paschall, Justin, 109
Passalacqua, Michael, 14
Patel, Bidish K., 116
Patel, Chirayush, 163
Paten, Benedict, 36, 42, 61, 131
Patton, Elizabeth, 145
Pearce, Owen, 108
Perdomo, Jonathan E., 133
Perez Fernandez, Cesar A., 134
Perez, Gerardo, 69
Perez-Guijarro, Eva, 24
Perez-Lopez, Andres, 130
Perino, Alessia, 110
Perkett, Matthew, 160
Perreault, Claude, 54
Pertea, Mihaela, 1, 64, 78, 82, 149
Pevzner, Pavel A., 42
Pfeffer, Klaus, 84
Pflug, Florian, 4
Pflughaupt, Patrick, 135
Philippakis, Anthony, 104
Phillippy, Adam M., 42, 44, 100, 106, 131, 158, 164
Phillips, Zack, 29
Pickett, Brandon D., 106, 131
Pierce-Ward, N Tessa, 50
Pietan, Lucas, 136
Pinto, Brendan J., 131
Pippin, James A., 31
Planet, Paul J., 2, 125
Pointon, Damon, 156
Poisson, Johanne, 110, 111
Ponnaluri, Chaithanya, 91
Ponting, Chris, 176
Pop, Mihai, 3
Porubsky, David, 42
Potapova, Tamara, 42
Pradhan, Sriharsa, 91
Prajapati, Malyaj R., 105
Pratto, Florencia, 71
Prohens, Jaime, 14
Prosperi, Mattia, 148
Puente, Xose S., 54
Pugh, Trevor, 20
Puiu, Daniela, 1, 82

Qian, Elizabeth, 2
Qin, Wenning, 108
Qiu, Wei-Lin, 137

Raabe, Michael, 116
Raba, Veronica, 84
Rabbani, Keon, 106
Rabinowitz, Joshua, 83
Rajagopalan, Anugraha, 19
Rajput, Jyotshna, 12
Ramakrishnan, Srividya, 14
Ramalingam, Rajasekaran, 86
Ramani, Arun, 20
Ramdass, Amanda C., 138
Ramesh, K.T., 134
Ramos, Marcel, 118
Raney, Brian J., 69
Ranu, Navpreet, 29
Rapin, Alexis, 110
Rashidi Mehrabadi, Farid, 24
Raubenolt, Bryan, 59
Rautiainen, Mikko, 158
Ravi, Janani, 142
Razaghi, Roham, 124
Refahi, Mohammadsaleh, 8
Regev, Aviv, 19
Revsine, Mahler, 82
Rhie, Arang, 100, 131
Risso, Davide, 118
Rivera, Alberto M., 139
Rivers, Adam R., 50
Roberts, Nicholas, 109
Robine, Nicolas, 127
Robinson, Matthew L., 117
Robitaille, Gina, 14
Rocha, Eduardo, 70
Rodriguez, Isabel, 36, 61
Rokas, Antonis, 161
Röner, Sebastian, 129
Rosconi, Federico, 143
Rosen, Gail, 8, 140
Rossi, Massimiliano, 10, 13
Rossi, Nicole M., 36, 61
Roxburgh, Patricia, 79, 155
Rozenblatt-Rosen, Orit, 19
Rychkova, Anna, 102

< Back

Sahakyan, Aleksandr B., 135
Sahinalp, S. Cenk, 24, 85
Saito, Yuki, 151
Sakaguchi, Shimon, 19
Sakamoto, Yoshitaka, 141
Salick, Max R., 29
Salzberg, Steven L., 1, 64, 78, 82, 117, 122
Samart, Kewalin, 142
Sanchez, Sydney, 143
Santos, José M., 11
Santos, Ryan, 14
Särkinen, Tiina E., 14
Saterlee, James, 14
Savage, Michelle, 144
Schatz, Michael C., 14, 48, 72, 104, 106, 109, 144, 164
Schiebout, Courtney, 23
Schierup, Mikkel H., 113
Schilling, Rebecca, 169
Schmit, Stephanie L., 167
Schmitz, Martin, 154
Schneider, Kristen E., 41
Schoonjans, Kristina, 110
Schubach, Max, 129
Scott, Alan, 109
Scott, Lucy C., 145
Scrace, Simon, 160
Sedlazeck, Fritz, 37, 45, 178
Segal, Eran, 5
Seitz, Evan, 146
Seligmann, Will, 74
Semple, Colin A., 79, 155
Sen, Shurjo K., 147
Senff-Ribeiro, Andrea, 54
Senger, Donna, 171
Serajian, Mohammadali, 148
Serra Mari, Rebecca, 42
Sharon, Eilon, 29
Shaw, Jim, 6
Shelton, Jennifer, 127
Sheth, Maya, 137
Shibata, Tatsuhiro, 151
Shilo, Shay, 98
Shinder, Ida, 149
Shiraishi, Yuichi, 141, 150, 151
Shivakumar, Vikram S., 152
Shlush, Liran, 98
Shooshtari, Parisa, 20
Shuai, Shimin, 54
Shumate, Alaina G., 153
Siepel, Adam, 177
Šikic, Mile, 154
Silk, Ryan, 79, 155
Simpson, Jared T., 34
Sims, Ying, 156
Singh, Amartya, 55, 56, 157
Singh, Mona, 83
Sinha, Maisha, 37, 45
Sivanandan, Srinivasan, 29
Skadron, Kevin, 57
Slizovskiy, Ilya B., 76
Slotfeld Viana, Alice, 2
Smith, Anja, 160
Smith, Jennifer R., 169
Soga, Keisuke, 89
Sokhansanj, Bahrad, 8
Solar, Steven J., 158
Son, Hyeonju, 101
Song, Dongyuan, 174
Song, Li, 7
Song, Xiaoyu, 67
Southam, Lorraine, 40
Speir, Matthew L., 69
Spengler, Jennifer, 91
Stahlke, Amanda, 123
Standage, Daniel, 153, 159
Starita, Lea, 45
Stastny, Tiana, 160
Steenwyk, Jacob L., 161
Stegle, Oliver, 25
Stein, Lincoln, 54
Steinegger, Martin, 117
Steinmetz, Lars, 137
Stephens, Matthew, 27
Stirba, Josh, 108
Stitziel, Nathan O., 42
Sudy, Alexander, 103
Sugawa, Masahiro, 141
Sulc, Jonathan, 110, 111
Sun, Shixiang, 162
Sun, Zu-wen, 91
Sundar, Shreya, 127
Sunitha Kumary, Vishnu U., 91
Surana, Pallavi, 163
Suresh, Hamsini, 14

< Back

Sutton, Gavin, 128
Sweeten, Alexander P., 164
Swift, Diane, 40
Sztanka-Toth, Tamas Ryszard, 165

Takizawa, Akiko, 169
Tan, Anthony, 137
Tan, Luomeng, 166
Tanaka, Yosuke, 141
Tanay, Amos, 18
Tang, Patrick, 130
Taylor, Martin S., 145, 170
Taylor, Michael, 54
Teer, Jamie K., 167
Teichmann, Sarah, 40
Tepper, Mariano, 41
Thai, Christopher V., 168
Thakore, Pratiksha, 19
Theiller, Erin, 2, 125
Thomas, David L., 117
Timp, Winston, 48, 63, 109, 124
Ting, David T., 116
Tischler, Arthur S., 81
Tonielli, Amy, 142
Torremorell, Montserrat, 76
Tourancheau, Alan, 62
Trang, Khanh B., 31
Treangen, Todd J., 70
Trinh, Quang, 54
Tseng, Elizabeth, 100, 171
Tsui, Clement, 130
Tu, Shengjiang, 29
Turinsky, Andrei, 20
Tutaj, Marek A., 169
Tutaj, Monika, 169
Tutino, Mauro, 40

Uhler, Caroline, 19
Underwood, Jason, 171

Vaddadi, Naga Sai Kavya, 35
Vaidya, Anup, 91
Van Eck, Joyce, 14
Varabyou, Ales, 78
Varki, Rahul, 13
Vaughn, Hayley, 136
Vedi, Mahima, 169
Venkatachalapathy, Saradha, 29
Venters, Bryan, 91
Verburg, Jan C., 170
Verhey, Ted, 171
Vijg, Jan, 162
Vinatzer, Boris A., 58
Viner, Coby, 166
Voineagu, Irina, 128
Volden, Roger, 171
Vollmers, Christopher, 74
von Haeseler, Arndt, 4
Vrcek, Lovro, 154

Wagley, Yadav, 31
Wagner, Justin, 140
Waldron, Levi, 118
Walsh, Kieran, 128
Wang, Kai, 133
Wang, Pei, 67
Wang, Shur-Jen, 169
Wang, Ting, 42
Wang, Xingang, 14
Wang, Yuanyuan, 51
Wang, Yujue, 162
Ware, Doreen, 68
Warren, Wesley C., 90
Wei, Sharon, 68
Wells, Andrew D., 31
Wick, Brittney, 69
Wilkinson, J Mark, 40
Williams, Evan, 110
Wilson, Melissa A., 131
Wind, Daniel, 84
Wokaty, Jen, 118
Wolfe, Jareth, 176
Wood, Jonathan, 49, 156
Wright, Adam J., 172
Wyatt, Cameron, 145

Xie, Dongyue, 27
Xie, Yi, 36, 61
Xie, Zhijie, 51
Xing, Yi, 51
Xu, Amanda, 1, 82
Xu, Nova, 116
Xu, Shuangyong, 175
Xu, Zhuwei, 173

< Back

Yadollahpour, Payman, 19
Yan, Guan'ao, 174
Yang, Weiwei, 175
Yao, Yuelin, 176
Yedla, Roshan R., 163
Yin, Shiyi, 108
Yoo, Byunggil, 36, 61
Yu, Yun William, 6
Yuan, Guo-Cheng, 67
Yuan, Hao, 92
Yung, Andrea, 19

Zakeri, Mohsen, 16, 152
Zeggini, Eleftheria, 26, 40
Zemel, Babette S., 31
Zeng, Huiqing, 131
Zhang, Xue-Song, 62
Zhao, Lily, 134
Zhao, Weihao, 115
Zhao, Yixin, 177
Zheng, Xinchang, 178
Zhou, Zhihan, 32
Zhu, Bryan, 127
Zimin, Aleksey V., 179
Zook, Justin, 140

< Back

MAJOR DATA ANALYSIS ERRORS INVALIDATE CANCER MICROBIOME FINDINGS

Abraham Gihawi*[1], Yuchen Ge*[2,3], Jennifer Lu*[2,3], Daniela Puiu[2,3], Amanda Xu[2], Colin S Cooper[1], Daniel S Brewer[1,4], Mihaela Pertea[2,3,5], Steven L Salzberg[2,3,5,6]

[1]Norwich Medical School, University of East Anglia, Norwich, United Kingdom, [2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [4]Earlham Institute, Norwich Research Park, Norwich, United Kingdom, [5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, [6]Department of Biostatistics, Johns Hopkins University, Baltimore, MD

We re-analyzed the data from a recent large-scale study [1] that reported strong correlations between DNA signatures of microbial organisms and 33 different cancer types, and that created machine learning predictors with near-perfect accuracy at distinguishing among cancers. We found at least two fundamental flaws in the reported data and in the methods: (1) errors in the genome database and the associated computational methods led to millions of false positive findings of bacterial reads across all samples, largely because most of the sequences identified as bacteria were instead human; and (2) errors in transformation of the raw data created an artificial signature, even for microbes with no reads detected, tagging each tumor type with a distinct signal that the machine learning programs then used to create an apparently accurate classifier.

One consequence of the first error was that the vast majority of microbial read counts reported across more than 17,000 samples, both tumor and normal, were severely over-estimated. For example, read counts were *at least 10 times too high* for 98% of the microbes found in bladder cancer, 93% of the microbes in head and neck cancer, and 98% of the microbes in breast cancer. The likely cause of these over-estimates was that the metagenomics database used for the original study included thousands of draft genomes, which are known to be contaminated with human sequences.

A consequence of the second error was that some machine learning models included species that had never been reported in humans, and that were associated only with extreme environments, ocean-dwelling species, plants, or other non-human environments.

Each of these major errors invalidates the results, leading to the conclusion that the microbiome-based classifiers for identifying cancer presented in the study are entirely wrong. These flaws have subsequently affected more than a dozen additional published studies that used the same data and whose results are likely invalid as well.

References

[1] Poore GD, Kopylova E, et al. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579:567-574.

EMPOWERING GLOBAL DISEASE SURVEILLANCE: THE CURED TOOL FOR RAPID IDENTIFICATION OF UNIQUE CLONAL BIOMARKERS

Erin Theiller[1], Elizabeth Qian[2], Andries Feder[3], Alice Slotfeld Viana[4], Agnes Marie Sá Figueiredo[4,5], Paul J Planet[3,6,7], Ahmed M Moustafa[1,6,8]

[1]Children's Hospital of Philadelphia, Division of Gastroenterology, Hepatology and Nutrition, Philadelphia, PA, [2]University of Pennsylvania, School of Arts and Sciences, Philadelphia, PA, [3]Children's Hospital of Philadelphia, Division of Infectious Diseases, Philadelphia, PA, [4]Universidade Federal do Rio de Janeiro, Departamento de Microbiologia Médica, Rio de Janeiro, Brazil, [5]Universidade Federal Fluminense, Faculdade de Medicina, Programa de Pós-graduação em Patologia, Niterói, Brazil, [6]University of Pennsylvania, Department of Pediatrics, Perelman School of Medicine, Philadelphia, PA, [7]American Museum of Natural History, Sackler Genomic Institute, New York, NY, [8]Children's Hospital of Philadelphia, Department of Biomedical and Health Informatics, Philadelphia, PA

Amid the escalating imperative for swift tracking of emerging hypervirulent and adapted pathogenic clones, the diminishing financial burden of whole-genome sequencing (WGS) presents a promising prospect. However, numerous low-to-middle income regions persistently lack access to such resources, thereby highlighting the urgent need for developing new bioinformatics tools that can leverage WGS data to quickly find diagnostic, apomorphic sequence mutations that create a new and unique restriction site in the clone of interest. These identified biomarkers can then be used to create a diagnostic screen. Addressing this challenge, we introduce the "Classification Using Restriction Enzyme Diagnostics" (CURED) tool that quickly identifies k-mers with restriction enzyme recognition sites that are exclusive to a specific clonal lineage. The culmination of this pipeline is the development of primers and PCR test that can amplify the region where the unique K-mer that is highly specific to the clone of interest is. This is then followed by endonuclease digestion showing a unique gel electrophoresis pattern in the clone of the interest. Using the CURED tool, we efficiently analyzed more than 71,000 *Staphylococcus aureus* genomes in just 30 minutes, requiring less than 32GB memory usage. This rapid analysis enabled us to identify a unique diagnostic K-mer with a distinct restriction enzyme site for a new emerging clone of *S. aureus* in Rio de Janeiro, Brazil. This K-mer then served as a biomarker for a rapid PCR test for surveillance of this clone. This novel pipeline promises to be the catalyst for equitable disease surveillance worldwide, revolutionizing the way developing public health risks are confronted and contained.

STRAINY: ASSEMBLY-BASED METAGENOMIC STRAIN PHASING USING LONG READS

Ekaterina Kazantseva*[1], Ataberk Donmez*[2,3], Mihai Pop[3], Mikhail Kolmogorov[2]

[1]ITMO University, Computer Science, St. Petersburg, Russia, [2]National Institutes of Health, National Cancer Institute, Bethesda, MD, [3]University of Maryland, Department of Computer Science, College Park, MD
* Equal contribution

INTRODUCTION: Metagenomic sequencing of human microbiome and other complex microbial communities reveals extensive heterogeneity on sub-species levels. Reconstruction of heterogeneous bacterial species (represented by multiple strains) is particularly challenging for long-read metagenomic assemblers. Here we present an algorithm for phasing and assembly of closely related strains from long reads called Strainy.

METHODS: The goal of the Strainy is to cluster long reads based on their strain of origin. Given a set of all reads aligned to a unitig, Strainy builds a connection graph, where nodes correspond to aligned reads, and edges connect reads that share the same SNP genotypes. Then, Strainy iteratively uses a community detection algorithm to partition the connection graph into densely connected components. Reads from each connected component are assembled into strain unitigs. Strainy then uses the overlap graph approach to extend strain unitigs into longer contigs. The final strain-specific contigs are integrated back to the original de novo assembly graph.

RESULTS: We benchmark Strainy against hifiasm, Strainberry, and metaFlye using the following ONT and PacBio HiFi datasets:

1. Simulated datasets for different bacterial species (E. coli, P. aeruginosa, L. monocytogenes, S. aureus) with 2-5 strains in different proportions using PacBio HiFi and ONT reads
2. Zymo mock community dataset sequenced with both PacBio HiFi and ONT reads
3. ONT sequencing of an activated sludge from an anaerobic digester (Sereika et al., 2022).

On simulated ONT datasets, Strainy reconstructed a substantially higher portion of unique strain sequence (88.7% mean reference coverage), improving over Strainberry (71.4%) and metaFlye (35.5%) while having fewer misassemblies. Similarly, for PacBio HiFi reads, Strainy (90.5%) and hifiasm (89.3%) both produce higher mean genome coverage than Strainberry (71.9%) and metaFlye (48.7%), while having low number of misassemblies. Analysis of real metagenomic sequencing of mock communities resulted in similar conclusions, with Strainy accurately recovering a high fraction of unique strain haplotypes.

We also applied Strainy to untangle strains in metagenomic sequencing of activated sludge from an anaerobic digester bacterial community. From the de novo assembly graph, Strainy selected 13 high-quality species bins containing heterozygous SNVs for further phasing. Overall, we reconstructed an additional 40% of strain-specific sequence, with some species showing 3 or more co-existing lineages. We reconstructed strain SNVs and SVs from assembled haplotypes using dipcall, revealing specific evolutionary dynamics for different species.

Strainy is available at: https://github.com/katerinakazantseva/Strainy

GC-BIAS AWARE SPECIES ABUNDANCE ESTIMATION FROM METAGENOMIC DATA WITH GUACAMOLE INCREASES ACCURACY AND COMPARABILITY

Laurenz Holcik[1], Florian Pflug[2], Arndt von Haeseler[1]

[1]University of Vienna, Center for Integrative Bioinformatics, Vienna, Austria,
[2]Okinawa Institute of Science and Technology, Okinawa, Japan

The importance of the microbiome in medical research, ecology, and other fields has become apparent with the rapid advances in sequencing technologies. Through sequencing technologies, it is not only possible to determine the presence of, for example, specific bacteria, but one can also estimate the abundance of individual species based on the sequenced reads.
However, it is known that read counts as generated by next generation sequencing protocols are influenced by factors like GC content or library preparation. For the abundance estimation of microbial organisms the genomic GC content is one major cause of systematic errors. Although this has been described in the literature, there is currently no method to overcome this issue.
To account for GC content, we present GuaCAMOLE (Guanine Cytosine Aware Metagenomic Opulence Least squares Estimation), a fast computational method to estimate relative taxon abundance from shotgun metagenomic data. GuaCAMOLE builds on the assumption that sequencing reads with a similar GC content will be equally affected by any GC bias present in the data. It bins the reads into discrete bins by organism and GC content. Then the method compares read counts of the different organisms per GC bin. Thereby reads with similar GC contents are compared which are equally affected by GC bias. This model-free approach is implemented as a least squares problem which is computationally fast to solve and results in unbiased abundance estimates.
We show on data from a microbial mock community that GuaCAMOLE reduces the relative error of abundance estimates by up to 50 % compared to commonly used methods like Bracken and MetaPhlAn4.
For many sequencing protocols the sequencing efficiency of organisms with a low genomic GC-content is particularly bad. For example, the abundance of Fusobacterium nucleatum (GC-content 27%), which is associated with colorectal cancer amongst other diseases, is systematically underestimated with current methods. On metagenomic data from stool samples of coloractal cancer patients we show that GuaCAMOLE estimates abundances of F. nucleatum roughly 50% higher than established methods.
We show that for identical metagenomic samples, the library preparation protocol has a substantial impact on the GC bias. By its model-free approach, GuaCAMOLE automatically adjusts for protocol specific biases and will thereby contribute to comparability of results in microbiome research.
Additionally, GuaCAMOLE can identify mircobial species more reliably by comparing the GC distributions of the observed reads in the sample to the expected distributions from the reference genomes. We show on a metagenomic mock community that GuaCAMOLE reduces the number of falsely detected species by over 75% while still always detecting the truly present species.

PERSONALIZED MEDICINE BASED ON DEEP HUMAN PHENOTYPING

Eran Segal

Weizmann Institute of Science, Department of Computer Science, Rehovot, Israel

Recent technological advances allow large cohorts of human individuals to be profiled, presenting many challenges and opportunities. I will present The Human Phenotype Project, a large-scale (>10,000 participants) deep-phenotype prospective longitudinal cohort and biobank that we established, aimed at identifying novel molecular markers with diagnostic, prognostic and therapeutic value, and at developing prediction models for disease onset and progression. Our deep profiling includes medical history, lifestyle and nutritional habits, vital signs, anthropometrics, blood tests, continuous glucose and sleep monitoring, and molecular profiling of the transcriptome, genetics, gut and oral microbiome, metabolome and immune system. Our analyses of this data provide novel insights into potential drivers of obesity, diabetes, and heart disease, and identify hundreds of novel markers at the microbiome, metabolite, and immune system level. Overall, our predictive models can be translated into personalized disease prevention and treatment plans, and to the development of new therapeutic modalities based on metabolites and the microbiome.

ULTRAFAST, COVERAGE-CORRECTED GENOME SIMILARITY QUERIES FOR METAGENOMIC SHOTGUN SAMPLES WITH SYLPH

Jim Shaw[1], Yun William Yu[1,2]

[1]University of Toronto, Department of Mathematics, Toronto, Canada,
[2]Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA

In shotgun metagenomics, the sequencing of all DNA in an environmental sample, the first computational step often involves identifying the genomes present in the sample. k-mer sketching methods, where samples are broken down into “bags of k-mers” called a sketch, allow for efficient analysis of metagenomic samples without assembly. These methods can provide an estimate of average nucleotide identity (ANI) of a genome directly against a metagenome, thus providing a continuous notion of presence.

We introduce sylph (https://github.com/bluenote-1577/sylph), a method for nearest neighbour ANI queries of genomes directly against shotgun metagenomes. sylph uses a k-mer sketching strategy with a zero-inflated Poisson model to impute low-coverage k-mers, allowing for accuracy down to 0.1x coverage. sylph takes seconds per sample and is 45x faster than MetaPhlAn for screening databases.

sylph operates directly on the genome-level, enabling strain-level analyses. On a Parkinson’s disease cohort, we applied sylph to obtain ANIs from 289,232 genomes against over 5.5 terabases of gut metagenome samples in under a day. We conducted a metagenome-wide association study with sylph’s ANIs and found 26 species with significant strain-level associations. The resulting bacterial phenotypes highlight the previously known inverse association between pro-butyrate bacteria and Parkinson’s disease, showing the power of sylph to quickly generate large-scale biological hypotheses.

CENTRIFUGER: LOSSLESS COMPRESSION OF MICROBIOME GENOMES FOR EFFICIENT AND ACCURATE METAGENOMIC SEQUENCE CLASSIFICATION

Li Song[1], Daehwan Kim[2], Ben Langmead[3]

[1]Dartmouth College, Department of Biomedical Data Science, Hanover, NH, [2]University of Texas Southwestern Medical Center, Lyda Hill Department of Bioinformatics, Dallas, TX, [3]Johns Hopkins University, Department of Computer Science, Baltimore, MD

Centrifuger is an efficient taxonomic classification method that compares the sequencing read against the microbiome genome database. Due to the boom of the available microbiome genomes, many methods store the database approximately to ensure that the memory footprint is within an affordable range. Centrifuger losslessly compresses the Burrows-Wheeler transformed (BWT) sequence from microbiome genomes with a novel compression algorithm called run-block compression. We prove that the run-block compression achieves sublinear space complexity, $O(n/\sqrt{L})$, where n is the sequence length and L is the average run length. This space complexity falls between the no-compression wavelet tree representation using $O(n)$ space and the run-length compression representation using $O(n/L)$ space. Run-block compression has low memory overhead and is effective in compressing microbiome genome databases, such as RefSeq, where the average run length of the BWT sequence is low, e.g., about 6.8. Built upon the run-block compressed BWT sequence and compact representations of other components in the Ferragina-Manzini (FM) index, Centrifuger reduces the memory cost by half compared to its predecessor, Centrifuge. The lossless compression helps Centrifuger obtain better accuracy than other methods at low taxonomy levels. Additionally, run-block compression supports rapid rank queries, and the time complexity for the rank query is of the same order as the rank query in the wavelet tree. Therefore, even though Centrifuger runs with a compressed data structure, it is as efficient as Centrifuge in terms of processing speed. Centrifuger is free and open source at https://github.com/mourisl/centrifuger.

LEVERAGING LARGE LANGUAGE MODELS FOR METAGENOMIC ANALYSIS

Mohammadsaleh Refahi, Bahrad Sokhansanj, Gail Rosen

Drexel University, Electrical and Computer Engineering, Philadelphia, PA

Analyzing metagenomic data is challenging due to the high diversity of taxa and genes within a microbiome. We propose to use triplet networks with contrastive learning of RoBERTa-based embeddings to discern taxonomic and functional information. Pre-training RoBERTa, a large language model, on full genomes equips it to capture intrinsic nucleotide patterns not attainable through conventional supervised training. We assess the effectiveness of our method on taxa and gene datasets, showcasing its superiority in taxonomic classification, even in challenging genomic regions. Extending the triplet network to a new task -- identifying genes -- enables additional insight from the same data. Using embeddings for multitask classification holds significant promise for health and environmental diagnostics. It advances our capacity to process and interpret complex microbiome data, offering potential benefits in biomarker identification and the monitoring of disease and environmental health.

We conduct downstream analysis on three distinct datasets:16s, 28s, and ITS. By utilizing datasets containing 16S rRNA exclusive to bacteria and eukaryotic mitochondria, as well as datasets containing 28S rRNA and ITS specific to eukaryotes (such as fungi), we were able to assess the performance of RoBERTa embeddings across diverse genomic regions. We show that we achieve better taxonomic classification performance than other pre-trained models and are similar to state-of-the-art taxonomic prediction models. However, state-of-the-art classifiers do not learn an embedded representation of the sequences.

< Back

PANGENOMICS WITH FOUNDER SEQUENCES AND GRAPHS

Veli Makinen

University of Helsinki, Department of Computer Science, Helsinki, Finland

Consider a multiple sequence alignment (MSA) representing a sample of haplotypes from a species. Such representation is redundant as the sequences share large parts in common. These identical parts are typically merged to derive a variation graph, where paths represent the haplotypes. Construction of such graphs can be formalized as finding a segmentation of the MSA into haplotype blocks, using some optimization criteria to characterize the resulting graph. For example, one can minimize the maximum number of founder segments (nodes) in a block, with the goal of minimizing the number of paths required to cover the graph. Such path cover induces set F of founder sequences: each haplotype (row in the input MSA) can be represented as a recombination of F. The talk reviews some recent linear time segmentation algorithms to produce these kind of founder graphs and founder sequences. Also, the talk considers how to use the technique for enhancing tasks such as genotyping and long read alignment with a pangenomics component.

COMPRESSED LINEAR PANGENOME INDEXES FOR MORE ROBUST READ CLASSIFICATION

Omar Ahmed[1], Massimiliano Rossi[2], Travis Gagie[3], Christina Boucher[2], Ben Langmead[1]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]University of Florida, Computer Science, Gainesville, FL, [3]Dalhousie University, Computer Science, Halifax, Canada

Recent years have seen drastic improvements in compressed indexing for pangenomics. Pangenome indexes are increasingly used for read alignment [1] and classification [2]. However, these indexes have so far not been usable for the taxonomic classification problem, chiefly because the indexes grow too quickly with the number of genomes in the taxonomic database. In recent work we proposed ways of (a) performing binary and multi-class read classification using a compressed pangenome index, and (b) performing full-fidelity "document listing," i.e. determining which genomes a query string occurs in. But no pangenome method has been able to combine these features while maintaining a small and efficient index.

We present the first compressed pangenome index capable of performing taxonomic classification at the same scale as k-mer based tools like Kraken. Building on our earlier SPUMONI work that applied the r-index to read classification problems, we propose a new taxonomy-aware compression scheme called LCP cliffs. This allows us to store a small fraction of the information that would be needed for full document listing. Our experiments on the ribosomal RNA sequences from the SILVA database show that our compression scheme can reduce the index size by over 200x, going from O(rd) space complexity to O(r) for this special classification problem. Additionally, we present a general compression scheme called Top-K LCPs which can be used in any read classification setting where there is no taxonomy. These different compression techniques allow us to take full advantage of the r-index in the context of these large-scale classification problems.

The r-index, a pangenome FM-index, is a full-text index allowing us to store genomes in their true linear form, which in turn, allows us to avoid multiple sequence alignments. Furthermore, this approach allows us to fully respect structural variation and linkage disequilibrium, which can be difficult for graphs in small windows with many variants [3]. Our earlier work, SPUMONI, utilized the r-index to rapidly classify sequencing reads using matching statistics or a newly developed value called pseudo-matching lengths. Our results show that our index scales sublinearly as we add in more genomes due to the repetitiveness of genomic databases. This method gives us the flexibility to find exact matches of variable length, instead of using a single value of k. In summary, combining the strengths of the r-index based approach and our new compression techniques for “document listing” will allow us to begin applying compressed pangenome indexes for taxonomic classification.

[1] 10.1038/nbt.4227
[2] 10.1186/s13059-023-02958-1
[3] 10.1186/s13059-018-1595-x

GRAPH-BASED AND GENE-BASED PANGENOME OF *LACTOCOCCUS LACTIS* AND *LACTOCOCCUS CREMORIS*

Paulo J Dias[1,2], José M Santos[1,2], Diogo L Antunes[1,2], Sofia O Duarte[1,2], Leonilde M Moreira[1,2,3], Gabriel A Monteiro[1,2,3]

[1]iBB - Institute for Bioengineering and Biosciences, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, [2]Associate Laboratory i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, [3]Bioengineering Department, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

In this study, we have constructed the Gene-based and Graph-based Pangenomes of *L. lactis* and *L. cremoris* with the aim of characterizing the genetic variation and genome dynamics of strains belonging to these Lactic Acid Bacteria (LAB) species.

Two databases were constructed focusing on the LAB species. GenomeDB compiles biological and genome information for a total of 1,225 million genes encoded in 256 different LAB species. BlastDB compiles all possible pairwise comparisons between the amino acid sequences of the genes included in GenomeDB. Clusters of amino acid sequence similarity were determined using a blastp network traversal approach. Using the numeric codes assigned to these clusters, we employed the Pagoo pipeline to construct the Gene-based Pangenomes of *L. lactis* and *L. cremoris*, utilizing the complete genome sequences of 33 and 16 strains, respectively. This pipeline allowed determining the Core, Shell, and Cloud Genes and to identify genes unique of each LAB species. The Graph-based Pangenome was constructed using the following steps. The Anchorwave software was used to create pairwise genome alignments between all possible combinations of the *L. lactis* and *L. cremoris* strains selected for analysis. Since this software suite uses a synteny-guided algorithm, the resulting PAF alignment files also define blocks of synteny. Subsequently, the Seqwish software was used to induce the Graph-based Pangenomes of these two LAB species in the variation graph model. To obtain a VCF file as output within a reasonable time frame, we used the odgi and vg toolkits to detect complex regions in these objects and simplify the graphs and to construct XG and GBWT indexes. Finally, we used the vg deconstruct command to genotype the two LAB species at the (low) sequence level, allowing us to determine several model-free Population Genomics metrics.

The genome dynamics of *L. lactis* and *L. cremoris* at the gene region and chromosomal level were characterized using the following approaches. Scripting was used to query the genome coordinates of each synteny block and retrieve the numeric codes of the genes encoded in orthologous regions. This data was passed as input to the Smith-Waterman algorithm to construct pairwise genome alignments using the numeric cluster codes as words. The resulting tab-format files allowed identifying indels occurring at a gene regional level in a simple way. On the other hand, scripting was used to manipulate the PAF alignment files to identify homologous chromosomes and construct adequate input files for Syri. This software suite allowed enumerating all chromosomal-level structural variations occurring in the two LAB species.

CO-LINEAR CHAINING ON PANGENOME GRAPHS

Jyotshna Rajput, Ghanshyam Chandra, Chirag Jain

Indian Institute of Science, Computational and Data Sciences, Bengaluru, India

Pangenome reference graphs are essential to compactly represent the genetic diversity of a species, a capability that linear references lack. Pangenome graphs can incorporate complex variations like inversions, duplications, and copy number variations by representing them using cycles. However, efficiently aligning reads or assembly contigs to pangenome graphs with complex topology and cycles still needs to be improved. Co-linear chaining is an established technique for accelerating short and long-read mapping. After computing a set of short exact matches during the seeding stage, co-linear chaining identifies a coherent subset of the matches that can be combined to produce an alignment. This technique has been well-studied theoretically for sequence-for-sequence alignment and implemented in several aligners, including minimap2 and BWA. Recent works show how to solve the co-linear chaining problem for acyclic pangenome graphs by exploiting their small width [Makinen et al., TALG'19] and how incorporating gap cost in the scoring function improves alignment accuracy [Chandra and Jain, RECOMB'23]. However, these algorithms are restricted to directed acyclic graphs (DAGs) and are not easily generalizable to cyclic graphs.

We present the first practical formulation and an exact algorithm for co-linear chaining on cyclic pangenome graphs. Our algorithm builds upon the known chaining algorithms for DAGs. We propose a novel iterative algorithm to handle cycles and provide a rigorous proof of its correctness. Our algorithm also takes advantage of the small-width property of pangenome graphs to optimize its runtime. However, computing the width of a directed graph is a known NP-Hard problem. To overcome this challenge, we use a heuristic instead. The runtime complexity of our chaining algorithm is a function of the count of input matches, and other parameters that we show are small in practice.

We implemented our algorithm as an open-source tool, PanAligner (https://github.com/at-cg/PanAligner). PanAligner is an end-to-end long-read aligner for cyclic pangenome graphs. We evaluated its speed and accuracy by aligning simulated long reads to a cyclic pangenome graph constructed from 94 high-quality haplotype-resolved assemblies and CHM13 human genome assembly. We achieved the highest long-read mapping accuracy, 98.7%, using PanAligner compared to existing methods Minigraph (98.1%) and GraphAligner (97.0%).

MONI-ALIGN: AN R-INDEX BASED PANGENOMICS ALIGNER

Rahul Varki*[1], Eddie Ferro*[1], Massmiliano Rossi*[1], Marco Oliva[1], Ben Langmead[2], Christina Boucher[1]

[1]University of Florida, Department of Computer and Information Science and Engineering, Gainesville, FL, [2]John Hopkins University, Department of Computer Science, Baltimore, MD

* Authors contributed equally

The use of a single human reference genome for alignment has caused downstream analysis to be biased towards European populations due to the human reference genome being primarily composed of European DNA. In order to better understand and represent the genetic diversity across populations, pangenomic aligners have been created but this area is still in a nascent stage as pangenomes are inherently large in size and pose a significant challenge. Recently, the development of the r-index, a variation of the FM-index, has allowed for index construction of pangenomes in sub-linear space. Until now, there has been no read aligner developed for the r-index.

MONI-align is the first read aligner built specifically for the r-index. MONI-align extends the work of MONI by Rossi et al. which showed a novel algorithm that uses Prefix-Free-Parsing (PFP) to build the r-index and the auxiliary threshold data structure simultaneously in order to support finding maximal exact matches (MEMs) with the r-index. PFP preprocesses the data in a way that takes advantage of the inherent repetitiveness of the pangenome. MONI-align maps reads to pangenome references using MEMs found between the read and reference as seeds. These seeds are then extended to find approximate matches for the entire read. MONI-align is a full read aligner in that it calculates the MAPQ value of the alignments and outputs an accurate SAM file that can be used as input to a variant caller.

We will present multiple results on both simulated and real data from human and *Arabidopsis thaliana* that demonstrate the precision and recall of MONI-align. First, we aligned two million short reads simulated from HG002 and HG00111 to GRCh38 and an increasingly larger number of haplotypes taken from the 1000 Genomes Project, which illustrates that the precision and recall rivals that of both VG and Giraffe. Next, we show how MONI-align can be combined with GATK and DeepVariant to find variants using a pangenome. Hence, we present genetic variants identified from aligning 30x coverage HG002 reads to a human pangenome using data from the Human Pangenome Consortium and using those alignments as input to DeepVariant and GATK. We present both the precision and recall of the identified variants. Lastly, we build the first ever pangenome of *Arabidopsis thaliana* using the 1001 Genomes Project and identify variants in the *Arabidopsis thaliana* accession Shahdara (Sha) by aligning to this pangenome and using GATK and DeepVariant to perform the variant calling.

PANAGRAM: ALIGNMENT-FREE AND INTERACTIVE PAN-GENOME VISUALIZATION

Katharine Jenike[1], Sam Kovaka[1], Matthias Benoit[2,3], Srividya Ramakrishnan[1], Shujun Ou[1,4], James Saterlee[3], Stephan Hwang[1], Iacopo Gentile[3], Anat Hendelman[3], Michael Passalacqua[3], Xingang Wang[3], Michael Alonge[1], Hamsini Suresh[3], Ryan Santos[3], Blaine Fitzgerald[3], Gina Robitaille[3], Edeline Gagnon[5], Melissa Kramer[3], Sara Goodwin[3], W. Richard McCombie[3], Jaime Prohens[6], Tiina E Särkinen[7], Amy Frary[8], Jesse Gillis[3,9], Joyce Van Eck[10,11], Ben Langmead[1], Zachary B Lippman[2,3], Michael C Schatz[1]

[1]JHU, Computer Science, Baltimore, MD, [2]HHMI, Cold Spring Harbor, NY, [3]CSHL, Biological Sciences, Cold Spring Harbor, NY, [4]Ohio State University, Molecular Genetics, Columbus, OH, [5]Technical University of Munich, Phytopathology, Munich, Germany, [6]Universitat Politècnica de València, València, Spain, [7]Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom, [8]Mount Holyoke College, Biological Sciences, South Hadley, MA, [9]University of Toronto, Physiology, Toronto, Canada, [10]Boyce Thompson Institute, Ithaca, NY, [11]Cornell University, Plant Breeding & Genetics, Ithaca, NY

Pan-genomes are collections of genomes from individuals within a population, species, or clade. The sequence variation revealed by pan-genomes is critical in many diverse contexts, such as crop engineering and human health, where uncharacterized variation may impact phenotypes in unexpected ways. Pan-genomic sequence variation can be analyzed using methods like multiple-sequence alignment (MSA) or de Bruijn graphs, however these are computationally intensive and potentially inaccurate when comparing highly divergent genomes.

We developed Panagram, an alignment-free interactive pan-genome visualization and analysis tool for large-scale comparative genomics studies. Importantly, Panagram efficiently encodes variation as the presence or absence of k-mers at genomic loci, avoiding MSA or complex graph traversals. We applied Panagram to a genus-level pan-genome composed of over 20 ecologically and agriculturally diverse species in the crop-rich Solanum. Within the Solanum pan-genome, we identified the most conserved 200kbp block, which contained 68 genes, 37 of these genes were annotated and associated with key biological processes while the remaining 31 were uncharacterized. Detailed inspection revealed stretches of sample-specific k-mers within the highly conserved block, which overlapped annotated transposable elements, possibly indicating regions that are tolerant of transposable element invasions. Additionally, we found that the nucleotide level Panagram scores correlate with multiple sequence alignment based measures of conservation such as PhyloP and PhastCons. Panagram provides a fast and informative way to visualize the inherently large datasets and complex comparisons present in this, and other, pan-genomes.

GIGGLES: PANGENOME-BASED GENOME INFERENCE USING LONG READS

Samarendra Pani[1,2], Jana Ebler[1,2], Tobias Marschall[1,2]

[1]Heinrich Heine University, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany, [2]Heinrich Heine University, Center for Digital Medicine, Düsseldorf, Germany

Whole genome sequencing ultimately aims at reconstructing the full (diploid) genome of a study sample. To approximate this genome, traditionally a variant-centric viewpoint is taken where variant discovery and genotyping workflows map short-read sequences to linear reference genomes (e.g. GRCh38). However, the linear reference fails to capture the allelic diversity of the human population, which leads to reference biases, and the short-read data have profound limitations in regions with structural variants (SVs). In contrast, our previous tool PanGenie uses a pangenome reference, and specifically the linkage information in its haplotype paths, with short-read data to perform genome inference, i.e. to determine the genotypes of all variants simultaneously in a haplotype-aware manner. This amplifies our ability to detect structural variants by a factor of two. At the same time, long-read sequencing is becoming more widespread and corresponding tools for SV discovery and genotyping such as Delly2, Sniffles, and CuteSV have been proposed. However, tools leveraging the information provided by a pangenome in connection with long-read sequencing data have been mainly missing. Here, we propose Giggles, a tool for pangenome-based genome inference from long reads.

Giggles is an HMM-based method which uses long read alignments to pangenome graphs to perform genome inference. Giggles focuses on the panel of haplotype paths provided with the pangenome reference to leverage the linkage information inherent to the long reads. We integrate ideas from the Li-Stephens model and read-based phasing to jointly model read partitioning by haplotypes and express the sample haplotypes as a pair of mosaics resulting from recombinations between reference haplotypes, resulting in maximum likelihood genotype at each site. Heuristic approaches have been developed to make the core HMM constant-time with respect to coverage.

We performed leave-one-out experiments to evaluate the method. We used a pangenome reference (CHM13-Minigraph released by the Human Pangenome Reference Consortium) and excluded the sample HG01258 from the graph. Using ONT data of the same sample, Giggles genotyped 99% of biallelic variants with a concordance of 81% to 96%. For multiallelic variants, Giggles genotyped 49% to 73% of variants at a concordance of 84% to 86%. Downsampling of reads, even to 5x, showed very little drop in performance, which testifies to its ability to exploit the linkage information inherent to both long reads and the reference haplotypes. Giggles features heuristic speed-ups and requires 5 milliseconds per variant position for its core algorithm (peak memory usage of 6GB), making it extremely scalable. The scalability enables its use in projects producing population scale long read data sets, such as the “All of Us” program.

REAL-TIME NANOPORE ADAPTIVE SAMPLING WITH MOVI

Mohsen Zakeri[1], Nathaniel Brown[1], Omar Ahmed[1], Travis Gagie[2], Ben Langmead[1]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Dalhousie University, Computer Science, Halifax, Canada

Pangenome indexes enable efficient analysis of sequencing reads while avoiding reference bias. There is an increasing need for efficient algorithms for indexing and querying pangenomes. Recently, the "move structure" was proposed as an alternative to the BWT-based indexes like FM index and r-index that underlie popular alignment tools. The move structure grows with O(r), where r measures the amount of distinct sequence in the input. When indexing a pangenome, r grows slowly compared to n, where n is the total input length. The move structure is capable of executing key queries (e.g. the LF mapping) in constant time. Unlike other BWT-based indexes, the move structure is represented as a table with remarkably simple query access patterns, giving strong locality of reference and few cache misses. These benefits come at a price: the constant factor in the O(r) growth is high relative to other approaches like the r-index.

We describe a new move-structure-based pan-genome indexing and querying tool called Movi. When speed is paramount, Movi can greatly outperform other methods using its constant-time queries and locality of reference. We illustrate that Movi can compute the pseudo-matching lengths for long reads (e.g., ONT) and is about 10 times faster than the fastest available methods. Movi is able to keep up in real time with the sequencing throughput of the fastest available nanopore devices (PromothION) while other approaches fail to do so. We show that Movi achieves this speed by minimizing cache misses, incurring close to the minimum possible number of cache misses on average for large pangenomes.

Movi operates in three main modes: the “fully constant” mode, the “one-bit constant” mode, and the “approximately constant” mode. The constant time queries in the move structure require extra pre-processing steps as well as increasing the constant factor of O(r) in the index size. The third mode requires less pre-processing compared to the other two and smaller storage space compared to the fully constant mode. However, in practice, we observe that the amortized cost of queries with the approximately constant mode is highly similar to the cost of the queries with the constant modes.

COMPUTATIONAL ANALYSIS OF RNA STRUCTURE AND FUNCTION

Sean R Eddy

HHMI & Harvard University, Department of Molecular & Cellular Biology, Cambridge, MA

Many functional RNAs have evolutionarily conserved secondary structures composed of conserved base pairing interactions. Conserved base pairing induces strong pairwise sequence correlations between the interacting positions in RNA multiple alignments. This powerful statistical signal can be exploited in many kinds of comparative genome sequence analysis of structural RNAs, including structure-based homology searches, RNA sequence alignment, consensus RNA structure prediction, and computational genomic screens for new RNAs. I will give an overview of both the history and the state of the art of these computational methods, with a focus on two main issues in RNA biology. The first will be on how these methods help clarify the controversial question of whether many of the long noncoding RNAs (lncRNAs) discovered by transcriptomics do or do not have conserved secondary structures. The second will be on RNA homology search and alignment using probabilistic RNA structure/sequence models, which has (at last) become computationally tractable for even the largest RNAs.

< Back

SPATIO-TEMPORAL QUANTITATIVE MODELS FOR EMBRYONIC DEVELOPMENT

Amos Tanay[1,2]

[1]Weizmann Institut, Department of Computer Science and Applied Mathematics, Rehovot, Israel, [2]Weizmann Institute, Department of Molecular Cell Biology, Rehovot, Israel

The development of mammalian embryos involves a highly coordinated and interactive parallel processes driving cell proliferation and differentiation in space and time. We are developing new models for this process based on analysis of single cells in single embryos. Partial spatial profiling of embryos is added using a novel staining and sequencing technology. Our spatio-temporal models are then inferring differentiation flows over time bins and spatial bins, thereby balancing proliferation, migration and differentiation to harmoniously represent single cell data from over 300 gastrulating mouse embryos. The outcome is a canonical and quantitative model for mouse gastrulation, which we use as the basis for understanding cell intrinsic and extrinsic regulation of germ layer differentiation, axis formation and specification of key lineages. Generalization of our modelling approach to other species and its integration with epigenomic profiling will be discussed

PERTURBDECODE, A PROBABILISTIC ANALYSIS FRAMEWORK TO RECOVER REGULATORY CIRCUITS AND PREDICT GENETIC INTERACTIONS FROM LARGE-SCALE PERTURBATION SCREENS

Basak Eraslan[1,2,3], Katie Geiger-Schuller[2], Kelvin Chen[4], Romain Lopez[1,2], Payman Yadollahpour [2,3], Olena Kuksenko[3], Pratiksha Thakore[2], Ozge Karayel Eren[2], Andrea Yung[2], Anugraha Rajagopalan[2], Ana Meireles de Sousa[2], Karren Dai Yang[5], Nir Hacohen[3], Caroline Uhler[5], Orit Rozenblatt-Rosen[2], Shimon Sakaguchi [4], Aviv Regev[2]

[1]Stanford University, Genetics Department, Stanford, CA, [2]Genentech Research and Early Development, GRED, South San Francisco, CA, [3]Broad Institute of MIT and Harvard, Klarman Cell Observatory, Cambridge, MA, [4]Osaka University, Immunology Frontier Research Center, Osaka, Japan, [5]Massachusetts Institute of Technology, Department of Informatics, Cambridge, MA

Pooled perturbation screens with uni- and multimodal single cell readouts open up new avenues in dissecting the function of genes and deciphering the gene regulatory networks at multiple layers of regulation. However, it is challenging to analyze large-scale screens with thousands of perturbations and millions of profiled cells due to noise from varying degrees of efficiency of the CRISPR-based perturbations, complex interrelation between omics layers and the need to predict effects of combinations of perturbations that have not been observed experimentally. Here, we present PerturbDecode, a framework for the automated analysis of such screens, including ComBVAE, a novel probabilistic deep generative model to identify effective CRISPR guides and significantly perturbed cells, and for predicting the outcome of the unseen combinations of perturbations. To test PerturbDecode we analyzed two new large-scale screens, i) a Perturb-seq screen spanning 3,390 perturbations of 1,130 E3 ligase and family members across 838,201 primary immune dendritic cells, including 660,330 single and 177,871 combinatorial perturbations ii) a Dogma-seq screen of 290 epigenetic inhibitors at 4 different dosages, spanning 2,400 conditions across 314,000 primary human CD4+ T cells with surface protein, transcriptome and epigenome readouts. Applying PerturbDecode to these data, we demonstrate that it is powerful in increasing the signal to noise ratio, identifying the mechanisms of regulation of the upstream and downstream genes of pathways that are induced by perturbations, and predicting the genetic interactions determined with combinatorial perturbations. PerturbDecode grouped the E3 ligase family members into co-functional modules that were enriched for physical interactions and impacted specific programs through substrate transcription factors. With multimodal readouts, it identified the dosage specific effects of the epigenetic modifiers, revealing the transcription factors and their target genes responding to the chromatin structure changes. PerturbDecode provides an efficient statistical analysis framework to recover causal regulatory circuits from large-scale perturbation screens.

ASSESSING PERFORMANCE OF SUPERVISED AND UNSUPERVISED CELL TYPE LABELING ALGORITHMS FOR CANCER scRNA-SEQ DATA

Erik Christensen[1], Ping Luo[2], Andrei Turinsky[3], Mia Husić[3], Alaina Mahalanabis[3], Alaine Naidas[1], Javier Diaz-Mejia[2], Michael Brudno[2], Trevor Pugh[2], Arun Ramani[3], Parisa Shooshtari[1]

[1]Western University, London, Canada, [2]University Health Network, Toronto, Canada, [3]The Hospital for Sick Children, Toronto, Canada

Single-cell RNA sequencing (scRNA-seq) clustering and labeling methods are used to determine the precise cellular composition of tissue samples. Automated labeling methods rely on either unsupervised, cluster-based approaches or supervised, cell-based approaches to identify cell types. The high complexity of cancer poses a unique challenge, as tumor microenvironments are often composed of diverse cell types with unique functional effects that may lead to disease progression, metastasis, and treatment resistance. Here, we assessed 17 cell-based and 9 cluster-based scRNA-seq labeling algorithms using 8 cancer datasets, providing a comprehensive large-scale assessment of such methods in a cancer-specific context.

We first built a cancer scRNA-seq database and search tool called TMExplorer (https://github.com/shooshtarilab/TMExplorer). TMExplorer provides an interface to easily access and fast query of cancer scRNA-seq datasets and their metadata. We selected 8 datasets with cell type labels from TMExplorer. This provided a collection of datasets from different cancer types, sequencing technologies, and a wide range of cells and genes numbers. Then we evaluated the performance of 26 scRNA-seq labeling algorithms on these datasets, taking into account several aspects and conditions that are of particular interest when working with cancer data. This includes assessing the performance of the algorithms in detecting under-represented or specific categories of cell types, the effects of imbalanced datasets on cell type predictions, as well as patient-based analyses.

We showed that supervised cell-based methods generally achieved higher performance and were faster compared to cluster-based methods. Cluster-based methods more successfully labeled non-malignant cell types, likely because of a lack of gene signatures for relevant malignant cell subpopulations. Larger cell numbers present in some cell types in training data positively impacted prediction scores for cell-based methods. Finally, we examined which methods performed favorably when trained and tested on separate patient cohorts in scenarios similar to clinical applications, and which were able to accurately label particularly small or under-represented cell types in the given datasets. We conclude that scPred and SVM show the best overall performances with cancer-specific data, and provide further suggestions for algorithm selection.

Our study exemplifies the specific challenges associated with labeling cancer scRNA-seq data, and highlights which algorithms perform well in a cancer context, helping cancer researchers and clinicians select the preferred tools for their analyses.

ISOFORM QUANTITATIVE TRAIT LOCI ANALYSIS OF NEUROPSYCHIATRIC DISORDERS IN ADULT BRAINS AT SINGLE-CELL RESOLUTION

Eric Nguyen[1], Matthew Jensen[1], Diego Garrido Martin[2], Declan Clarke[1], Prashant Emani[1], Mark Gerstein[1]

[1]Yale University, Yale Computational Biology & Bioinforamtics, New Haven, CT, [2]University of Barcelona, Department of Genetics, Microbiology and Statistics, Barcelona, Spain

Gene expression is intricately controlled by various elements, such as transcription factors, regulatory RNA, and other proteins. These collectively contribute to specific phenotypes, traditionally studied through quantitative trait loci (QTLs). Alternative splicing introduces another layer of complexity, necessitating the investigation of isoform quantitative trait loci (isoQTLs). Regulation of alternative splicing has been shown to play an especially prominent role in neuropsychiatric disorders, potentially due to the large diversity of cell types in brain tissues. Our study represents the first of its kind to directly identify isoQTLs from single-cell resolution datasets across an integrated cohort of 388 human adult prefrontal cortex (PFC) samples from the PsychENCODE Consortium, including healthy individuals and those with neuropsychiatric disorders. After building a novel analysis pipeline that integrates several isoform analysis tools, we generated transcript-level expression matrices for ~2.8 million nuclei with short-read snRNA-Seq data, correcting for the 3’ bias inherent in 10X sequencing technologies, and identified isoQTLs from expression and genotype datasets. Our analysis spanned 20 neuronal and non-neuronal cell types and identified over 2000 significant cell type-specific isoQTLs and 90 significant spliced genes (sGenes) after permutation-based multiple testing. High-resolution cell typing allowed us to profile changes in isoQTLs effect sizes and functional enrichment across specific cell type classes in the PFC. Moreover, the identified sGenes include several genes linked to neuropsychiatric disorders, such as LYPD6 and GABARAPL1, in Pax6 and L5/6 NP cell types respectively, which are candidate genes for schizophrenia and autism spectrum disorder. We validated significant results through visualization and comparison to orthogonal regulatory datasets. such as snATAC-Seq for select samples. Furthermore, our isoQTL pipeline was rigorously designed to validate results between short-read datasets and long-read snRNA-Seq approaches, such as SMRT-Seq. Overall, our work fills a significant gap given the paucity of single-cell resolution studies in isoQTL analysis and addresses the need to uncover intricate relationships between isoQTLs and neuropsychiatric disorders. As our optimized computational pipeline becomes increasingly robust, we anticipate its broader application in elucidating isoQTLs in neuropsychiatric disorders and other clinical phenotypes, thereby contributing valuable insights to complex gene expression mechanisms.

FUNCTIONAL GENOMICS USING IMAGE-BASED PROFILING: FROM VARIANT IMPACT TO DRUG SCREENING

Anne E Carpenter

Broad Institute of Harvard and MIT, Cambridge, MA

Cell images contain a vast amount of quantifiable information about the status of the cell: for example, whether it is diseased, whether it is responding to a drug treatment, or whether a pathway has been disrupted by a genetic mutation. We aim to go beyond measuring individual cell phenotypes that biologists already know are relevant to a particular disease. Instead, in a strategy called image-based profiling, often using the Cell Painting assay, we extract hundreds of features of cells (or other biological samples, such as tissues or whole organisms) from images. Just like transcriptional profiling, the similarities and differences in the patterns of extracted features reveal connections among diseases, drugs, and genes.

We are harvesting similarities in image-based profiles to identify, at a single-cell level, how diseases, drugs, and genes affect cells, which can uncover small molecules’ mechanism of action, discover gene functions, predict assay outcomes, discover disease-associated phenotypes, identify the functional impact of disease-associated alleles, and find novel therapeutic candidates.

< Back

CELL TYPE-SPECIFIC INTERACTION ANALYSIS USING DOUBLETS IN scRNA-SEQ (CIcADA)

Courtney Schiebout[1], Hannah Lust[2], Yina Huang[3], H. Robert Frost[1]

[1]Dartmouth College, Biomedical Data Science, Hanover, NH, [2]MDI Biological Laboratory, STEM, Undergraduate & Graduate Training, Bar Harbor, ME, [3]Dartmouth College, Microbiology and Immunology, Hanover, NH

A doublet in single-cell RNA-sequencing (scRNA-seq) data occurs when two cells fail to dissociate during sample processing. This is typically assumed to be the result of accidental adherence of cells, whereby two cells that have no biological association become stuck together during sample preparation/processing. As a result of this assumption, cells that appear to be doublets in scRNA-seq data are usually excluded from downstream analysis. While this approach avoids the impact of artifactual doublets, it fails to consider the potential for doublets to be the result of a biologically meaningful association. Specifically, cells undergoing a juxtacrine interaction in situ may maintain adherence to one another throughout the scRNA-seq sample preparation pipeline, resulting in doublet data that actually carries important information for the sample being studied. This is especially true in models of the tumor microenvironment (TME), where the interactions among immune cells can have notable implications for recurrence and prognosis. To utilize this information, however, interacting doublets have to be separated from artifactual doublets, so that only the biologically relevant data is being considered. To meet this need, we developed Cell type-specific Interaction Analysis using Doublets in scRNA-seq (CIcADA), a tool for discerning and evaluating biologically relevant doublets in scRNA-seq data, with the option to integrate Cellular Indexing of Trancriptomes and Epitopes by Sequencing (CITE-seq) information. CIcADA operates by identifying potentially interacting doublets in a dataset based on a high overlap of scores for two separate cell types. These potential doublets are then compared to synthetic doublets constructed from the same cell type composition to distinguish between doublets with expression profiles that are canonical for those cell types and doublets whose expression profiles result from true interactions. We evaluated CIcADA on three TME scRNA-seq datasets: a partially-dissociated B16F10 mouse model of melanoma dataset, a mucosa-associated lymphoid tissue (MALT) dataset, and a lymphoma dataset. In all three cases, we found that doublets identified as truly interacting by CIcADA consistently had upregulated genes involved in immune response. We further validated our findings from the B16F10 dataset with 10x Visium spatial transcriptomics data generated on tumors from the same experimental setup. We found that genes identified as upregulated in doublets in the scRNA-seq data were also found to be upregulated in Visium spots where cell co-occurrence was detected.

ROBUST AND SCALABLE INTRATUMOR HETEROGENEITY AND TUMOR PROGRESSION TREE INFERENCE AND ASSESSMENT THROUGH SINGLE-CELL RNA SEQUENCING DATA

Farid Rashidi Mehrabadi[1], Salem Malikic[1], Eva Perez-Guijarro[2], Kerrie L. Marie[2], Glenn Merlino[2], Chi-Ping Day[2], S. Cenk Sahinalp[1]

[1]Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health, Bethesda, MD, [2]Laboratory of Cancer Biology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD

Lineage tracing at the single cell level is a key approach for understanding intratumor heterogeneity (ITH) and studying subclonal evolution in cancer progression, as well as investigating metastatic seeding and therapy resistance. For such purposes, single-cell tumor progression trees inferred by the use of genetic alterations provide detailed cellular lineages and an overall evolutionary history of a tumor. Currently, the vast majority of the available single-cell sequencing data sets are generated by single-cell RNA sequencing (scRNAseq). However, the existing tools for tumor progression tree inference, which are typically primarily designed for DNA sequencing data, are either not scalable or can not handle the expression variation, technical and biological dropouts, and sequencing error characteristics of scRNAseq data to accurately perform lineage tracing based on expressed genomic alterations. Furthermore they produce a single tree for modeling tumor progression and, even though such a tree may feature several highly reliable clades (subtrees), it may also include other clades that are less reliable. None of the available methods offers the ability to assess the reliability of a given clade in a principled manner.

To overcome these limitations, we have developed Trisicell (https://trisicell.rtfd.io), a set of computational methods for inferring tumor progression history from full-length scRNAseq data using expressed somatic mutations, assessing its reliability, and investigating the implied genetic and non-genetic ITH. To evaluate the accuracy of Trisicell, we generated sample-matched bulk exome, bulk transcriptome and scRNAseq datasets from single-cell derived clonal sublines from a well known mouse melanoma cell line and demonstrated that Trisicell is more scalable and/or accurate than all available alternatives.

When applied to tumor scRNAseq data from preclinical study conducted in this work, as well as data from previously conducted clinical studies, Trisicell enabled us to identify essential drivers of tumor subclonal evolution and explore crucial questions related to the relationship between adaptation and selection, the connection between genotype and phenotype and the impact of immune checkpoint blockade on tumor progression. In summary, the features offered by Trisicell bring new and essential capabilities for analyzing scRNAseq data and advancing our understanding of ITH and evolution.

SPACETREE: DECIPHERING TUMOR MICROENVIRONMENTS BY JOINT MODELING OF CELL STATES AND GENOTYPE-PHENOTYPE RELATIONSHIPS IN SPATIAL OMICS DATA

Olga Lazareva[1,2,3], Elyas Heidari[1,2,4], Omer Bayraktar[5], Oliver Stegle[1,2,6]

[1]German Cancer Research Center, Division of Computational Genomics and Systems Genetics, Heidelberg, Germany, [2]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany, [3]German Cancer Research Cente, Junior Clinical Cooperation Unit, Multiparametric Methods for Early Detection of Prostate Cancer, Heidelberg, Germany, [4]German Cancer Research Center, Division of AI in Oncology, Heidelberg, Germany, [5]Wellcome Sanger Institute, Wellcome Genome Campus Hinxton, Cambridge, United Kingdom, [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Spatial omics technologies offer unique opportunities to study tumor evolution in the spatial context. However, the clonal composition of tumors remains a major challenge, especially because genetic subclones are intrinsically coupled with cellular composition and the spatial makeup of a tumor. Additionally, tumor heterogeneity manifests in distinct spatial niches, which govern cell type, and clonal composition within the tumor microenvironment. Existing analysis methods fail to capture this complexity, either assuming known discrete clones matched to spatial profiles or they employ label transfer to map scRNA-seq to the corresponding spatial omics but ignoring clonal relationships. Consequently, these tools provide a fragmented view, capturing certain aspects of the integration task but falling short in jointly modeling cell types, clonal profiles, and their intricate spatial relationships.

To address this, we here propose SpaceTree, an end-to-end framework that jointly conducts copy number inference (leveraging external tools), spot decomposition (if needed), mapping of cell types and clones, multi-sample integration, and niche detection. SpaceTree builds on a multi-task graph neural network architecture coupled with label propagation. SpaceTree jointly models spatially smooth cell type- and clonal state composition, employing a hierarchical tree loss, thereby avoiding discretizing clonal profiles or cell types. SpaceTree employs Graph Attention mechanisms, capturing information from spatially close regions when reference mapping falls short, enhancing both interpretation and quantitative accuracy.

A significant merit of SpaceTree is its technology-agnostic nature, allowing clone-mapping in sequencing- and imaging-based assays. SpaceTree allows to borrow evidence strength across samples or different tissue blocks and technologies, as long as the same reference data is provided. The model outputs can be used to characterize spatial niches that have consistent cell type and clone composition.

We applied SpaceTree to published scRNA-seq, Visium, and Xenium data from human breast cancers, as well as data from an unpublished GBM tumor atlas. SpaceTree yielded results consistent across platforms, accurately identifying clonal profiles and niches concurrent with annotated, morphologically distinct ducts. The results show the potential of SpaceTree in elucidating the genetic heterogeneity of intricate tissues. Additionally, we benchmarked SpaceTree's cell-type deconvolution against leading cell-type deconvolution algorithms for Visium, further emphasizing its superior potential across platforms.

TRANSLATIONAL GENOMICS OF COMPLEX DISEASE

Eleftheria Zeggini[1,2]

[1]Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Translational Genomics, Neuherberg, Germany, [2]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine and Health, Munich, Germany

In this talk, I will give an overview of how we have used translational genomics approaches to enhance our understanding of complex diseases like osteoarthritis and type 2 diabetes, shed novel biological insights, and provide a stepping stone for bridging the gap between basic discovery and translation.

REFERENCE-FREE DIFFERENTIAL ISOFORM ANALYSIS USING SHORT-READ RNA-SEQ DATA

Carlos F Buen Abad Najar[1], Dongyue Xie[2], Matthew Stephens[2], Yang I Li[1]

[1]University of Chicago, Department of Medicine, Chicago, IL, [2]University of Chicago, Department of Statistics, Chicago, IL

RNA alternative splicing, alternative 3’ UTR ends, and intron retention are ubiquitous forms of isoform variation that greatly increase transcriptome diversity. Short-read RNA sequencing (RNA-seq) is a powerful tool for quantifying transcriptome expression. However, short reads usually cover at most one splice junction. Moreover, characterization and quantification of intron retention and alternative UTR events is limited to the availability of transcriptome annotations. This makes the study of entire transcripts and regulatory connections between splicing events difficult to study. To our knowledge, there is no reference-free approach for differential isoform usage of whole transcripts.

Empirical Bayes Poisson matrix factorization (EBPMF) is a novel method developed for variational inference in non-Gaussian data. We apply EBPMF, incorporating smoothing factor constraints, to thousands of short-read RNA-seq samples from multiple tissues from the GTEx consortium. EBPMF factors capture major isoform variation events, including alternative splicing, intron retention, and extended 3’ UTRs. As an annotation-free approach, EBPMF uncovers variational relationships between these events, potentially capturing hidden regulatory connections. For example, we find that nonsense-mediated decay isoforms in the gene SRSF3 are associated with shorter 3’UTRs in the terminal exon.

We confirm the robustness of our approach using long-read RNA-seq data from the GTEx consortium to corroborate structural connections between events. Coupling EBPMF with splice-junction based approaches such as Leafcutter increases the power to detect differential isoform usage of whole transcripts across tissues in a reference-free manner. Our approach outperforms established reference-dependent methods such as rMATS.

TransferQTL EXPANDS EXISTING eQTL CATALOGS ACROSS HUMAN TISSUES.

Yuhang Chen

Yale University, CBB, New Haven, CT

eQTL catalogs constitute a valuable resource to study the effects of genetic variants on gene expression regulation, yet they can be obtained only by sampling tissues from sufficiently large cohorts of individuals. In fact, high-power eQTL catalogs so far have been obtained mostly on a few readily available tissues, such as blood or skin, while they are less available for harder-to-profile tissues, such as heart. By leveraging the ENTEx resource, we have found that eQTL SNVs have stronger chromatin signals in the tissues in which they are active than in the tissues in which they are not. This suggests that the chromatin around an SNV may influence its chance of being an eQTL in a tissue. Thus, by combining the EN-TEx chromatin data and the GTEx eQTL catalog, we developed transferQTL, a classifier that transfers the activity of an eQTL from a given donor tissue (e.g., skin) to another target tissue (e.g., heart). SNVs with chromatin activity, especially high H3K36me3 signal, were more likely to be transferred. We observed the opposite for SNVs associated with genes that are tissue-specific or have distant transcription start sites. Overall, when compared with known GTEx eQTLs, our predictions are highly accurate, independent of which donor or target tissues are employed. To showcase the value of our approach to enhance existing eQTL catalogs, we applied it to a set of 1.5M blood eQTLs from a recent large cohort study. We were able to transfer up to 60% of these blood eQTLs, enhancing the current GTEx catalog with ~500K new potential eQTLs per tissue. We further expanded this framework by combining predictions from multiple models, and obtained a “consensus” tissue-agnostic model that can help expand eQTL catalogs beyond the tissues profiled by GTEx.

UNCOVERING NOVEL TARGETS IN HUMAN TUBEROUS SCLEROSIS AND NON-ALCOHOLIC FATTY LIVER DISEASE MODELS BY INTEGRATING MULTIOMICS AND AUTOMATION WITH POOLED OPTICAL SCREENING

Max R Salick, Srinivasan Sivanandan, Bobby Leitmann, Saradha Venkatachalapathy, Atieh Givmanesh, Rahul Atmaramani, Shengjiang Tu, Owen Chen, Zack Phillips, John Bisognano, Alicia Lee, Lexie Ewer, Bay Johnson, Navpreet Ranu, Eilon Sharon, David Lloyd, Ajamete Kaykas, Ci Chu

insitro, South San Francisco, CA

Optical pooled screening (OPS), high content imaging, and spatial transcriptomics have emerged as powerful, cost-effective methods to connect genetic perturbations to cellular phenotypes. When combined, these techniques have been shown to unearth mechanisms of antiviral responses, essential gene effects, and cytoskeletal organization, and when processed using modern computer vision approaches, they allow researchers to create maps of biology. While these approaches have mostly been applied to general questions of pathway biology so far, they are primed for utilization in human disease modeling and target discovery. Here we describe the Pooled Optical Screening in Human Cells (POSH) platform, which includes several industrialized components such as fully closed-loop automation of the OPS in situ sequencing process. We also describe the data pipelines and productionized machine learning models that are critical for properly disentangling the complex biology underlying these data-rich screens. We next describe examples of how the platform has been deployed to identify novel disease-reverting targets in both environmentally stressed or genetically modified induced pluripotent stem cell models of disease. Machine learning methods were applied to the high-content screen data, enabling accurate prediction of healthy vs. disease states of individual cells. When the cells are processed with OPS-derived gRNA readouts, several known modulators of the mTOR pathway were identified in the tuberous sclerosis (TSC) model, along with known lipid processing components in a fatty liver disease model. We performed parallel perturb-seq experiments and found these orthogonal datasets to be complementary and effective at finding targets that were validated in follow-up functional assays. Lastly, we demonstrate a unified screening platform that combines OPS with fluorescence in situ hybridization (FISH) via dual-probe intramolecular ligation and amplification, to allow for paired morphological and transcriptional profiling of genetic perturbations at scale. To summarize, we have built an industrialized platform for conducting large-scale optical pooled screens, and demonstrate the deployment of multiple target discovery screens on human disease models to identify novel modulators of disease.

FINDING ALL OUR SWITCHES MUTATING EVERYTHING TO UNDERSTAND ALLOSTERY

Ben Lehner[1,2]

[1]Wellcome Sanger Institute, Generative and Synthetic Genomics Programme, Cambridge, United Kingdom, [2]CRG, Systems and Synthetic Biology, Barcelona, Spain

Thousands of proteins have now been genetically-validated as therapeutic targets in hundreds of human diseases. However, very few have actually been successfully targeted and many are considered ‘undruggable’. This is particularly true for proteins that function via protein-protein interactions: direct inhibition of binding interfaces is difficult, requiring the identification of allosteric sites. However, most proteins have no known allosteric sites and a comprehensive allosteric map does not exist for any protein. We have addressed this shortcoming by charting multiple global atlases of inhibitory allosteric communication in KRAS, a protein mutated in 1 in 10 human cancers. We quantified the impact of >26,000 mutations on the folding of KRAS and its binding to six interaction partners. Genetic interactions in double mutants allowed us to perform biophysical measurements at scale, inferring >22,000 causal free energy changes, a similar number of measurements as the total made for proteins to date. These energy landscapes quantify how mutations tune the binding specificity of a signalling protein and map the inhibitory allosteric sites for an important therapeutic target. Allosteric propagation is particularly effective across the central beta sheet of KRAS and multiple surface pockets are genetically-validated as allosterically active, including a distal pocket in the C-terminal lobe of the protein. Allosteric mutations typically inhibit binding to all tested effectors but they can also change the binding specificity, revealing the regulatory, evolutionary and therapeutic potential to tune pathway activation. Using the approach described here it should be possible to rapidly and comprehensively identify allosteric target sites in many important proteins.

3D GENOMIC FEATURES ACROSS >50 DIVERSE CELL TYPES REVEAL INSIGHTS INTO THE DIFFERING GENOMIC ARCHITECTURES OF BMD DETERMINATION AND OSTEOPOROTIC FRACTURE PATHOGENESIS

Khanh B Trang[1], Matthew C Pahl[1], James A Pippin[1], Alessandra Chesi[2], Yadav Wagley[3], Babette S Zemel[4], Kurt D Hankenson[3], Andrew D Wells[1,2], Struan F Grant[1,4]
[1]The Children's Hospital of Philadelphia, Human Genetics, Philadelphia, PA, [2]University of Pennsylvania, Pathology, Philadelphia, PA, [3]University of Michigan, Orthopedic Surgery, Ann Arbor, MI, [4]University of Pennsylvania, Pediatrics, Philadelphia, PA

The incomplete correlation between determining bone mineral density (BMD) and understanding the development of osteoporotic fractures suggests differences in the underlying biological processes. Genome-wide association studies (GWAS) have uncovered various genetic variants linked to either BMD or osteoporotic fractures, with only partial genetic overlap between them. Much remains to be learned about these genetic loci, including their cellular context, the specific causal variants involved, and the corresponding genes that exert their effects. To gain insights into these distinct traits, we harnessed our extensive 3D genomic datasets encompassing over 50 diverse human cell types. These datasets include high-resolution promoter-focused Capture-C/Hi-C, ATAC-seq, and RNA-seq data. Among these cell types are those closely related to bone biology, such as human mesenchymal stem cell (hMSC)-derived osteoblasts, as well as cell types spanning a broader range, including metabolic, neuronal, and immune cell types. Utilizing stratified LD regression, we quantified the proportion of genome-wide SNP heritability attributed to our cell type-specific features, integrating the latest summary statistics from pediatric and adult BMD and osteoporosis fracture GWAS. Our analyses unveiled a statistically significant enrichment ($P<0.05$) of genetic associations with osteoporotic fractures in multiple immune cell types, such as plasmacytoid and classical CD1c+ dendritic cells, various T-cell classes, pancreatic α-cells, and adipocytes. Conversely, features related to BMD showed greater enrichment in bone-related cell types, including hMSC-derived osteoblasts and the human fetal osteoblast hFOB cell line, as well as several neuron and neural progenitor cell lines. Moreover, longitudinal BMD data in pediatric patients exhibited statistically significant enrichment in multiple naïve T-cell types. The enrichments observed were particularly pronounced in the extended 500bp regions around our defined cis-regulatory elements (cREs) for the immune cell types, suggesting that the genetic signals originate from the vicinity rather than being tightly confined within the cREs. Subsequent application of our chromatin contact-based 'variant-to-gene' mapping revealed key "hub" genes, functional pathways, and biological processes involved, including the regulation of extracellular matrix remodeling and immune response. These findings further support the prevailing hypothesis that variants implicated in GWAS exert their effects through complex polygenic regulatory mechanisms involving critical signaling pathways. In summary, our comprehensive examination of 3D genomic datasets across diverse cell types enhances our genomic understanding of both the commonalities and distinctions between BMD determination and the pathogenesis of osteoporotic fractures.

< Back

DNABERT-2: EFFICIENT FOUNDATION MODEL FOR MULTI-SPECIES GENOMES

Zhihan Zhou[1], Yanrong Ji[2], Weijian Li[1], Pratik Dutta[3], Han Liu[1], Ramana V Davuluri[3]

[1]Northwestern University, Computer Science, Evanston, IL, [2]Northwestern University, Biomedical Informatics, Chicago, IL, [3]Stony Brook University, Biomedical Informatics, Stony Brook, NY

Decoding the linguistic intricacies of the genome is a crucial problem in biology, and pre-trained foundational models such as DNABERT and Nucleotide Transformer have made significant strides in this area. Existing works have largely hinged on k-mer, fixed-length permutations of A, T, C, and G, as the token of the genome language due to its simplicity. However, we argue that the computation and sample inefficiencies introduced by k-mer tokenization are primary obstacles in developing large genome foundational models. We provide conceptual and empirical insights into genome tokenization, building on which we propose to replace k-mer tokenization with Byte Pair Encoding (BPE), a statistics-based data compression algorithm that constructs tokens by iteratively merging the most frequent co-occurring genome segment in the corpus. We demonstrate that BPE not only overcomes the limitations of k-mer tokenization but also benefits from the computational efficiency of non-overlapping tokenization. Based on these insights, we introduce DNABERT-2, a refined genome foundation model that adapts an efficient tokenizer and employs multiple strategies to overcome input length constraints, reduce time and memory expenditure, and enhance model capability. Furthermore, we identify the absence of a comprehensive and standardized benchmark for genome understanding as another significant impediment to fair comparative analysis. In response, we propose the Genome Understanding Evaluation (GUE), a comprehensive multi-species genome classification dataset that amalgamates 28 distinct datasets across 7 tasks, with input lengths ranging from 70 to 1000. Through comprehensive experiments on the GUE benchmark, we demonstrate that DNABERT-2 achieves comparable performance to the state-of-the-art model with 21× fewer parameters and approximately 56× less GPU time 1 in pre-training. Compared to DNABERT, while being 3× more efficient, DNABERT-2 outperforms it on 23 out of 28 datasets, with an average improvement of 6 absolute scores on GUE. The code, data, and pre-trained model are publicly available at https://github.com/Zhihan1996/DNABERT_2.

RECOVERING HIDDEN LAYERS OF INFORMATION IN SINGLE-CELL DATA

Mor Nitzan

Hebrew University of Jerusalem, Jerusalem, Israel

Gene expression profiles of a cellular population, generated by single-cell RNA sequencing, contain rich, 'hidden' information about biological state and collective multicellular behavior that is lost during the experiment or not directly accessible, including cell type, cell cycle phase, gene regulatory patterns, cell-cell communication, and location within the tissue-of-origin. In this talk I will discuss several methods, based on a combination of spectral, machine learning, and dynamical systems approaches, to disentangle and enhance particular spatiotemporal signals that cellular populations encode and interpret their manifestation across space and time in tissues.

CALLING SOMATIC MUTATIONS FROM LONG READ TUMORS WITHOUT MATCHED NORMAL SAMPLES

Jared T Simpson

University of Toronto, Ontario Institute for Cancer Research, Dept. of Molecular Genetics, Dept. of Computer Science, Toronto, Canada

The detection of somatic mutations acquired during the initiation and progression of a tumor underpins much of cancer genomics and is increasingly used to inform treatment decisions. The gold standard method of detecting mutations is to sequence DNA from the tumor along with a matched normal sample to distinguish somatic mutations from the far more abundant polymorphisms inherited from the individual's parents. Despite the dominance of the tumor-normal approach many groups have tried to sequence only the tumor, either due to the lack of available normals or to simplify clinical workflows. These methods typically have limitations however and may require extensive filtering with population variant databases. Prior work by Darby et al. proposed that haplotype phasing patterns can inform whether a given variant is somatic or polymorphic. In this talk I will describe how recent advances in long read sequencing, coupled with a novel mutation calling framework that assesses the evidence for a mutation within phased haplotypes, can improve the accuracy of tumor-only somatic mutation calling. In addition I will discuss the critical experimental parameters (sequencing depth and error rate along with tumor cellularity) that make this approach feasible.

MINIMIZING REFERENCE BIAS WITH AN IMPUTE-FIRST APPROACH

Naga Sai Kavya Vaddadi, Taher Mun, Ben Langmead

Johns Hopkins University, Computer Science, Baltimore, MD

Pangenome indexes have the potential to reduce reference bias in genomic research. However, it has also been shown repeatedly that personalized references, such as a diploid human reference constructed to match the alleles in the individual being sequenced, achieve the best possible reduction in reference bias. We present a novel *Impute-first* alignment framework that combines elements of genotype imputation with pangenome alignment. This framework begins by performing genotyping and genotype imputation using only a subset of the input data. Using a massive reference panel and modern genotype imputation software (e.g., GLIMPSE), the framework then imputes a personalized diploid reference genome. Finally, it applies a linear or graph-based aligner (e.g., VG-Giraffe) to align reads with respect to the personalized reference, avoiding reference bias and improving downstream results such as variant calls.

Our *Impute-first* framework achieves improved accuracy of downstream results while also using a simpler and more efficient (imputed) pangenome representation compared to a typical pangenome. To evaluate, we used the HG001 sample with an Illumina read set and benchmarked it against the v4.2 GIAB HG001 high-confidence call set.

Our *Impute-first* method outperformed graph-based pangenome references in multiple variant calling measures. It exhibited higher sensitivity (99.54% vs 99.37%), increased precision (99.36% vs 99.18%), and improved overall accuracy (99.45% vs 99.28%). By integrating the advantages of traditional reference-based methods and pangenome approaches, *Impute-first* can minimize reference bias at low computational cost.

SEVERUS: A TOOL FOR AUTOMATIC CHARACTERIZATION OF COMPLEX GERMLINE AND SOMATIC REARRANGEMENTS IN CANCER USING LONG-READ SEQUENCING.

Ayse Keskus[1], Tanveer Ahmad[1], Ataberk Donmez[1], Asher Bryant[1], Isabel Rodriguez[2], Nicole M Rossi[2], Yi Xie[2], Byunggil Yoo[5], Rose Milano[2], Hong Lou[3], Jimin Park[4], Joshua Gardner[4], Brandy McNulty[4], Karen Miga[4], Midhat Farooqi[5], Benedict Paten[4], Michael Dean[2], Mikhail Kolmogorov[1]

[1]CCR, NCI, Bethesda, MD, [2]DCEG, NCI, Rockville, MD, [3]Leidos Biomedical Research, Inc., Frederick, MD, [4]UCSC Genomics Institute, Santa Cruz, CA, [5]Children's Mercy Hospital, Kansas City, MO

Structural variants (SVs) are one of the hallmarks of cancer, with recent pan-cancer studies showing that 55% of driver mutations are explained by SVs. Long-read sequencing substantially improved SV detection compared to short reads, yet most current long-read SV-calling tools were designed for "healthy" genomes and fail to capture somatic variants, complex rearrangements, and heterogeneity in cancer genomes.

Here, we present Severus to detect and annotate a wide range of somatic SVs, from simpler indels to complex rearrangements with multiple breakpoints, using long-read sequencing. Severus is taking advantage of phased long-read alignments and producing haplotype-specific calls. The algorithm automatically detects mismapped reads from the collapsed duplication regions, which is a major source of false positive calls. Once the breakpoint signatures are detected from split alignments, Severus builds a haplotype-specific breakpoint graph to cluster multi-break rearrangements and represent the derived chromosomal structure.

Severus outperforms existing tools in SV recall measured against the multiplatform validated somatic SV set for the COLO829 cell line. Further, we sequenced five cancer cell lines and their matching normals and showed that Severus outperformed other tools in terms of recall and precision measured against the ensemble method. We show that Severus can accurately detect known and novel complex rearrangements, chromothripsis, chromoplexy, breakage-fusion-bridge (BFB), inversions with amplifications/deletions, and intrachromosomal long insertions.

We also sequenced HPV-infected cervical cancer cell lines (CaSki, SNU1000, SCC152, HT3), revealing mosaic amplifications near the HPV integration sites. Our analysis revealed enrichment of BFB events with YAP1 amplification, associated with worse prognosis, in all HPV-infected cell lines, whereas chromothripsis was exclusive to low-risk HPVs.

Finally, we applied Severus to the matching tumor/normal sequencing of two unresolved clinical cases of pediatric AML with Pacbio HiFi. We found an in-frame deletion in the TTN gene affecting multiple exons and a complex translocation between chr10 and chr11 with inversion-deletion, leading to KMT2A/MLLT10 fusion, a known prognostic marker.

< Back

LONG-READ SEQUENCING OF 1000 GENOMES SAMPLES TO BUILD A COMPREHENSIVE CATALOG OF HUMAN GENETIC VARIATION

Nikhita Damaraju[1], J. (Gus) Gustafson[1], Sophia B Gibson[1], Miranda Galey[1], Kendra Hoekzema[1], Joy Goffena[1], Maisha Sinha[1], The 1000 Genomes ONT Sequencing Consortium[2], Fritz Sedlazeck[3], Matt Loose[4], Miten Jain[5], Evan E Eichler[1], Danny E Miller[1,6]

[1]University of Washington, NA, Seattle, WA, [2]The 1000 Genomes ONT Sequencing Consortium, NA, NA, WA, [3]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [4]University of Nottingham, School of Life Sciences, Nottingham, United Kingdom, [5]Northeastern University, Department of Bioengineering, Department of Physics, Boston, MA, [6]Brotman Baty Institute for Precision Medicine, NA, Seattle, WA

Growing evidence demonstrates that structural variants (SVs) significantly contribute to genetic diversity and disease susceptibility. Identification and complete characterization of SVs has been challenging using short read sequencing (srGS) approaches which are not capable of fully resolving the location or structure. This difficulty is likely part of the reason more than half of individuals with a suspected Mendelian condition remain undiagnosed after a comprehensive clinical evaluation, limiting their ability to benefit from precision or N-of-1 therapies. Thus, there is broad interest in using new technologies, such as long-read sequencing (LRS), to accurately evaluate variation in regions of the genome difficult to study using srGS and resolve the structure of complex SVs. However, use of LRS is hindered by the lack of a comprehensive dataset of SVs from healthy controls, which is crucial for filtering, categorizing, and prioritizing SVs.

To address these limitations, the 1000 Genomes ONT Sequencing Consortium is generating LRS data from 800 reportedly healthy individuals (~40% African, 40% Asian, 10% European, and 10% American ancestry) from the 1000 Genomes Project. Our goal is to characterize normal genome-wide SV patterns as well as to identify variants in areas of the genome that are difficult to evaluate using srGS (e.g., highly repetitive regions).

We present preliminary data from the first 100 genomes (average coverage, 30x; read N50, 50 kbp). Our analysis includes both alignment and de novo assembly-based approaches (contig N50 values > 30 Mb), haplotype-resolved variant calling with DeepVariant, and SV calling with Sniffles2 and cuteSV. Using joint SV callsets from this dataset and positive controls for pathogenic SVs, we optimized a pipeline for filtering and prioritizing SVs in unsolved clinical cases. The dataset generated from this project is publicly available and analysis efforts are open to all. This will be of broad utility to the human genetics community leading to discovery of disease-causing variants among individuals with suspected Mendelian conditions sequenced on LRS platforms.

INSIGHTS INTO THE GENETIC ARCHITECTURE OF DIABETES AND OTHER COMPLEX TRAITS IN THE GREENLANDIC POPULATION

Ida Moltke

University of Copenhagen, Department of Biology, Copenhagen, Denmark

So far most medical genetics studies have been performed in Europeans, which means that most of our knowledge about the role of genetics and the genetic architecture of common diseases is based mainly on genetic studies of Europeans.

I will present genetic data from a non-European population, namely the Greenlandic Inuit population, which has historically been small and isolated. I will show how this demographic history has led to a very different allele frequency distribution and genetic architecture of common diseases compared to that observed in Europeans. Briefly, in Greenlanders a much larger proportion of genetic variants are common, exactly as would be expected based on population genetic theory, and unlike in Europeans, we have identified several high-impact variants that affect common diseases. For instance, more than a hundred variants have been found to be associated with type 2 diabetes in Europeans, but they are all either rare or have moderate effects. In contrast, in our studies of diabetes in the Greenlandic population, we have identified two Inuit specific variants that both have relatively high allele frequencies (0.23 and 0.02 in Greenlandic Inuit) and high effect sizes (ORRECESSIVE=10.3 and ORADDITIVE 4.4, respectively). These two variants alone affect almost 1/5 of the Greenlanders with diabetes.

I will end by briefly discussing some of the implications of these findings both for the potential of improving the prevention and treatment of diseases in Greenland and for future medical genetics studies.

STIX: A NOVEL APPROACH FOR COMPREHENSIVE SOMATIC STRUCTURAL VARIATION DETECTION AND GENE FUSION IDENTIFICATION

Christopher E Ojukwu[1], Murad Chowdhury[1,2], Ryan Layer[1,2]

[1]University of Colorado Boulder, Computer Science, Boulder, CO,
[2]University of Colorado Boulder, BioFrontiers Institute, Boulder, CO

Structural variants (SVs) play a crucial role in the development of cancer and Mendelian disorders. However, integrating SVs into disease analysis poses persistent challenges. These challenges include limitations in short-read SV calling, biases related to reference genomes, and heuristics specific to patient cohorts. These issues hinder our ability to comprehensively assess the frequencies and clinical implications of SVs in populations.

In cancer research, distinguishing between germline and somatic SVs relies on comparing tumor tissue to control tissue, which can be sensitive to normal sample issues. An alternative method uses unrelated samples instead of matched-normal tissue but comes with significant time and computational costs. Furthermore, large-scale DNA sequencing projects rely on SV catalogs for identifying inherited variants. However, practical constraints, such as short-read limitations and false-positive filtering, complicate variant detection and assessment, leaving the true nature of observed SVs in patients unresolved.

To address these challenges and achieve precise SV detection and accurate allele frequency assignment, we introduce a novel approach called STIX. STIX involves a systematic search of raw alignments across numerous samples, providing a count for each alignment supporting a specific deletion, duplication, inversion, or translocation variant. We operate under the assumption that deleterious variants are rare. Thus, an SV supported by evidence in many healthy samples is likely a common germline variant or a result of systematic noise, such as alignment artifacts, making it less likely to be disease-causing. STIX's reliance on raw alignments helps overcome previous issues with false negatives and eliminates potentially erroneous associations between variants and diseases.

For example, when analyzing PCAWG's prostate cancer samples, Manta identified 1,335 unique fusions. We can refine this list efficiently with STIX, narrowing it down to fusions common in cancer (PCAWG) but rare in germline (1KG) populations. This refined list includes both well-known fusions (e.g., ERG-TMPRSS2) and novel ones. We are expanding our method to consider all possible gene fusions and much larger datasets, including PCAWG's 2000+ samples across 20+ cancer types and the Cancer Cell Line Encyclopedia's 1000 cell lines. Since querying all gene pairs at this scale is impractical, we've added a new gene burden query to STIX, enabling us to investigate potential fusions efficiently. These new capabilities will support a gene fusion validation platform where clinicians and researchers can enhance their call sets using population-scale data.

THE MOLECULAR ENVIRONMENT OF OSTEOARTHRITIS RISK GENES IN PRIMARY CARTILAGE

Georgia Katsoula[1,2,3], John E G Lawrence[4,5], Mauro Tutino[2], Ana Luiza Arruda[1,2], Petra Balogh[6], Lorraine Southam[2], Diane Swift[7], Sam Behjati[5,8], Sarah Teichmann[5], J Mark Wilkinson[7,3], Eleftheria Zeggini[2]

[1]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine and Health, Munich, Germany, [2]Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Munich, Germany, [3]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine and Health, Munich, Germany, [4]Department of Trauma and Orthopaedics, Cambridge University Hospitals NHS Foundation Trust, Addenbrooke's Hospital, Cambridge, United Kingdom, [5]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [6]Department of Cellular and Molecular Pathology, Royal National Orthopaedic Hospital, Brockley Hill, Stanmore, United Kingdom, [7]School of Medicine and Population Health, University of Sheffield, Sheffield, United Kingdom, [8]Department of Paediatrics, University of Cambridge, Cambridge, United Kingdom

Osteoarthritis is a complex joint disease with no curative therapy. It is mainly characterized by loss of cartilage integrity and its replacement with bone, a process including phenotypic alterations in the main cell type of cartilage, the chondrocyte. Despite the identification of numerous effector genes, the promise of translation of GWAS signals remains largely unrealized. To address this issue, we performed RNA-seq of macroscopically intact and degraded *ex vivo* knee osteoarthritis cartilage from 300 osteoarthritis patients, the largest expression study in osteoarthritis to date. We: (1) examine gene expression differences between intact and degraded cartilage, (2) use a consensus co-expression network approach combined with GWAS summary statistics to identify coordinated deregulation in knee osteoarthritis, and (3) perform scRNA-seq on human developing embryonic limbs to place these transcriptional networks in a cellular context. We detect broad transcriptional changes between intact and degraded cartilage (3159 up- and 2990 down-regulated genes in degraded cartilage at 5% FDR, 1253 novel). We identify six gene modules enriched for genetic risk among three osteoarthritis phenotypes (knee osteoarthritis, total knee and total joint replacement) ($P_{adj}<0.05$). Through module scoring at the single-cell level, we find that these modules represent distinct expression patterns during endochondral ossification, a process entailing gradual replacement of cartilage by bone. A module linked to embryonic morphogenesis is associated with early knee osteoarthritis (LMM: $\beta=-0.071$, $P_{adj}<0.001$) and enriched for knee osteoarthritis risk ($P_{adj}<0.001$). This module reflects the differentiation trajectory from chondroprogenitor to the pre-hypertrophic chondrocytes. This suggests that alterations in chondrocyte homeostasis in the early phases of disease may mirror the processes of chondrocyte differentiation seen during limb development. We provide evidence of convergence of osteoarthritis genetic risk in distinct modules reflecting both biological processes and cell types in the fetal limb providing anchors for putatively causal gene networks in knee osteoarthritis.

< Back

QUERYING BIOBANK-SCALE GENOMES TO RAPIDLY IDENTIFY GENETICALLY MATCHED COHORTS.

Kristen E Schneider[1], Murad Chowdhury[1], Mariano Tepper[6], Mark Hildebrand[6], Jawad Khan[6], Marteinn Hardarson[4], Bjarni Halldorsson[4,5], Chris Gignoux[2,3], Ryan Layer[1]

[1]University of Colorado Boulder, Computer Science, Boulder, CO, [2]University of Colorado Denver, Anschutz Medical Campus, Biomedical Informatics, Aurora, CO, [3]Colorado Center for Personalized Medicine, Biomedical Informatics, Aurora, CO, [4]deCODE genetics, Sequence Analysis, Reykjavik, Iceland, [5]Reykjavik University, Biomedical Engineering, Reykjavik, Iceland, [6]Intel Labs, Machine Learning, Hillsboro, OR

Genomic data is disproportionately dominated by European samples. This translates to a Eurocentric bias in genome wide association studies (GWAS) and polygenic risk scores (PRSs). Importantly, PRSs are remarkably less accurate for non-European samples, playing a significant role in health disparities experienced by underrepresented populations. Efforts to add non-European reference samples from a selection of global populations has been an exciting first step to correct for these biases; but there remains notable variation within populations that are not captured by these reference samples. Furthermore, as the field continues to deepen its knowledge of the genetic diversity of human health, there will be no finite number of reference genomes which can properly represent an increasingly diverse global human population. To mitigate these challenges, we propose a method to match patients to “people-like-me” cohorts. With this, we can deliver evidence-based precision treatments to clinical providers in real-time.

Our approach focuses on a machine learning infrastructure which enables automatic patient-centered cohort creation. The primary objective is to generate patient genotype embeddings whose distance serves as a proxy to genetic similarity. Representing patient genomic information as an embedding enables AI similarity search databases to efficiently and accurately assemble patient-matched cohorts. This approach is advantageous over traditional exact-match search strategies (e.g., an SQL database) because these traditional methods do not scale for large and complex biobanks. Our genetic embedding results demonstrate that we can preserve genetic relationships at both population and familial levels. When run on the 1000 Genomes data, our method creates cohorts for individuals who share the same subpopulation. For simulated and real pedigree data, our embedding vectors show that an individual’s parents’ embeddings are closer than their grandparents, which are closer than their cousins. Exploring the landscape of population-wide genotype data to determine genetic similarities and differences between individuals at increasingly granular levels provides a necessary foundation for future work in precision medicine and population genetics alike.

COMPLETE GENOMES TO EXPAND STUDIES OF GENETIC AND EPIGENETIC INHERITANCE OF CENTROMERES

Monika Cechova[1], Sergey Koren[2], Julian K. Lucas[1], Rebecca Serra Mari[3], Mobin Asri[1], David Porubsky[4], Jordan M. Eizenga[1], Brandy McNulty[1], Andrey Bzikadze[5], Shloka Negi[1], Christopher Markovic[6], Tamara Potapova[7], Jennifer L. Gerton[7], Pavel A. Pevzner[8], Evan E. Eichler[4], Benedict Paten[1], Adam M. Phillippy[2], Ting Wang[9], Nathan O. Stitziel[10,] Robert S. Fulton[6], Tobias Marschall[3], Karen H Miga

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, [2]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [3]Institute for Medical Biometry and Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [4]Dept. of Genome Sciences, University of Washington School of Medicine, Seattle, WA, [5]Graduate Program in Bioinformatics and Systems Biology, University of California San Diego, San Diego, CA, [6]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, Stowers Institute for Medical Research, Kansas City, MO, [8]Dept. of Computer Science and Engineering, University of California San Diego, San Diego, CA, [9]Dept. of Genetics, Washington University School of Medicine, St. Louis, MO, [10]Dept. of Medicine, Washington University School of Medicine, St. Louis, MO

The complete genomes provide an opportunity for a new biological discovery in repetitive parts of the genomes that were frequently missing, incomplete, or misrepresented in the past, such as centromeres. Here, we present a study aiming to release complete, T2T diploid assemblies of four individuals, representing a three-generational pedigree of African-American ancestry. To this end, we have utilized a combination of high-coverage, long-read (51x HiFi, PacBio, and 71x Nanopore Ultra-long data 100kb+, Oxford Nanopore Technologies) and short-read paired technologies for haplotype phasing (including 76x coverage Omni-C, Cantata Genomics, 39x paired read parental data for trio-based phasing, as well as Strandseq and Pore-C data). Using two iterative, graph-based methods with ultra-long nanopore integration (verkko (Rautiainen et al. 2023) and hifiasm-UL (Cheng et al. 2022)), we achieved a high number (e.g. 28/46) of automated telomere-to-telomere (T2T) chromosome assemblies, with additional improvements with the two assembly methods combined. This allowed us to study transgenerational inheritance in biologically critical regions that are highly repetitive and present copy number variation within the population. We provide evidence for genetic and epigenetic inheritance across large tandem repeats that define human centromeres. For example, we found that the centromeric sequence for chromosome 12 was 100% identical across the three generations, spanning the length of 3,321,121 basepairs, allowing us to study the biological variation in methylation patterns in this region, such as the variation in CDR (Centromere Dip Region), marked by large dips in methylation that underlie the binding of the centromeric protein CENP-A. Moreover, we found preliminary evidence of the enrichment of another modification, 5-Hydroxymethylcytosine (5hmC), in the peri(centromeric) region, and especially in the flanking HSat3 satellite arrays. Additionally, our assemblies captured several rDNA arrays, such as on chromosomes 21 and 22. In summary, these results provide a high-quality multi-generational pedigree that serves as a community resource for tracing of transgenerational inheritance, as well as genetic and epigenetic variation of centromeres, satellite DNA, and rDNA arrays.

CURRENT TOOLS FOR PETA-SCALE SEQUENCE EXPLORATION

Rayan Chikhi

Institut Pasteur, Department of Bioinformatics, Paris, France

Petabytes of valuable sequencing data reside in public repositories, doubling in size every two years. They contain a wealth of genetic information, in particular about viruses, which can help us monitor spillovers and anticipate future pandemics. In this talk, I will present some of the analysis routes and computational tools that are available, or in development, to explore such data. We recently developed a cloud infrastructure, Serratus (Edgar et al, Nature, 2022), to perform petabase-scale sequence alignment. It enabled the discovery of 10x more RNA viruses than previously known, including a new family of coronaviruses. Serratus pioneered peta-scale biological data analysis, yet there is much more to be accomplished in this field. In particular, the development of k-mer methods is of special interest given their simplicity and efficiency.

VERKKO’S APPROACHES FOR ONE-BUTTON T2T RESOLUTION FOR DIPLOID HUMAN-SIZED GENOME

Dmitry Antipov, Adam Phillippy, Sergey Koren

NHGRI, Genome informatics section, Bethesda, MD

Recent advances in sequencing technologies, especially the development of long-read sequencing, dramatically improve the contiguity and correctness of genome assemblies. The increasing accuracy of long reads simultaneously allows haplotype-resolved assembly, faithfully reconstructing the full diploid genome. The recently published verkko assembler relies on a combination of long accurate (typically PacBio HiFi) and ultra-long (typically Oxford Nanopore) reads to automatically generate chromosome-scale contigs. However, due to homozygous regions, long-read sequencing only usually does not allow to assemble complete chromosomes. Usual approach to improve long read assemblies is the usage of the trio data. However, parents are not always available for sequencing. Moreover, even with trio data, current assemblers usually do not allow to get at least 1⁄3 chromosomes t2t-assembled for human samples (including all acrocentric chromosomes) without manual processing. Here we present different methods based on usage of hi-c data for both phasing and scaffolding, sequence similarity and special processing of rDNA regions that are integrated in verkko assembler that allows to get majority (if not all) chromosomes assembled from telomere to telomere, show the reasons why remaining chromosomes are not t2t and discuss directions of further improvements.

HAPLOTYPE-RESOLVED CHARACTERIZATION OF REPEAT EXPANSIONS AND PATTERNS OF METHYLATION FROM 1000 GENOMES ONT CONSORTIUM DATA

Sophia B Gibson[1], J. (Gus) Gustafson[1], Nikhita Damaraju[1], Miranda Galey[1], Kendra Hoekzema[1], Joy Goffena[1], Jordan Knuth[1], Maisha Sinha[1], 1000 Genomes ONT Sequencing Consortium[2], GREGoR Consortium[3], Fritz Sedlazeck[4], Matt Loose[5], Miten Jain[6], Lea Starita[1,7], Evan Eichler[1], Danny E Miller[1,7]

[1]University of Washington, NA, Seattle, WA, [2]1000 Genomes ONT Sequencing Consortium, NA, NA, WA, [3]GREGoR Consortium, NA, NA, WA, [4]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [5]University of Nottingham, School of Life Sciences, Nottingham, United Kingdom, [6]Northeastern University, Department of Bioengineering, Department of Physics, Boston, MA, [7]Brotman Baty Institute for Precision Medicine, NA, Seattle, WA

The 1000 Genomes ONT Consortium is an effort to generate whole-genome long-read sequencing (LRS) data on 800 of the approximately 3,200 individuals available in the 1000 Genomes Project collection. This dataset is unique in that it will allow us to catalog normal patterns of human structural variation (SVs), identify variants in difficult-to-study regions of the genome, and evaluate haplotype-resolved genome-wide patterns of methylation from a large cohort of presumably healthy individuals. For SVs, this includes more accurate characterization of repeat expansions, which can be used to evaluate individuals for candidate disease-causing expansions. This level of accuracy is possible because LRS data is able to span the entire length of most repeat expansions, providing resolution previously difficult to achieve with short-read sequencing. Using the 1000G ONT data, we have identified known and novel repeat expansions across samples as well as further quantified the variance in expansion length in known repeat regions. These data will allow for both novel repeat expansion sites in individuals undergoing clinical evaluation and also allow us to compare known expansion sizes to benchmarks when assessing the ability for the expansion to be tolerated. Separately, LRS data allows us to analyze haplotype-resolved patterns of methylation in this cohort. We have quantified CpG methylation genome wide and compared methylation between haplotypes to both confirm known differential methylation regions (DMRs) on autosomes as well as identify novel DMRs. We have also quantified the variation in methylation patterns between X chromosome haplotypes on 46,XX individuals and used it to create computational methods for determining the X chromosome inactivation (XCI) status of these individuals. We have built a database for common DMRs found in 1000G samples that is used in automated pipelines to rapidly identify differences in methylation that could be attributed to imprinting disorders, X-linked disorders, or in relation to large structural variants.

ADAPTING ANALYSIS TOOLS TO A WORKFLOW-CENTRIC WORLD

Nate Coraor[1], John M Chilton[1], Nuwan A Goonasekera[2], Anton Nekrutenko[1], The Galaxy Team[1]

[1]Pennylvania State University, Biochemistry and Molecular Biology, University Park, PA, [2]University of Melbourne, Melbourne Bioinformatics, Melbourne, Australia

Modern genomic analysis tools often follow the tried-and-true UNIX philosophy of “do one thing and do it well.” This makes such tools incredibly powerful, but also dependent on a myriad of underlying dependencies. Importantly, no analysis is performed with just one tool. Assembling these tools into a cohesive pipeline is the foundation of performing end-to-end analyses.

In the Galaxy platform, such pipelines are known as *Galaxy Workflows*. Galaxy facilitates running workflows on disparate computational resources transparent to the users who run them. Adequately sizing resource requests (cores, memory, etc.) for each tool is the job of the Galaxy system administrator, from simple text manipulation tools to massive genome assemblers.

Traditionally, Galaxy submits jobs to a local cluster and the administrator configures tool resource requests according to their cluster and needs. But with over 4,000 tools available in Galaxy and the explosion in computational analysis needs, the local cluster and manual configuration does not scale to meet demand.

Galaxy has long had a system to send jobs to remote clusters (Pulsar) but has lacked a “meta-scheduler” to efficiently distribute the load when multiple clusters are available. The introduction of the Total Perspective Vortex (TPV) has provided this critical missing piece, and much more. TPV matches Galaxy tools and their resource requirements to available resources by matching requirements and tags, and ranking them based on configurable rules. In addition, the administrators of large Galaxy servers, particularly Galaxy Australia and Galaxy Europe, have developed a shared database of TPV tool definitions reusable by any Galaxy server, to alleviate the time consuming administrative task of configuring tool resource requirements.

Because tools are only useful in the context of an end-to-end workflow, and because Galaxy is not running each tool by hand with highly tuned resource requirements as would be done by a savvy command line user submitting to a single cluster of known composition, tool authors can contribute greatly to the usability of their tools running well in Galaxy or any other workflow system such as NextFlow, CWL, WDL, etc.

In this presentation we use our multi-year experience with managing a public resource populated with thousands of analysis utilities to provide recommendations to tool developers. Adopting these recommendations will greatly expand the usability of analysis tools. Importantly, they will directly benefit tool developers by making it easier to expose their software to a large audience of users.

UNDERSTANDING GENETIC VARIATION IN MODERN AND ARCHAIC HUMAN GENOMES

Janet Kelso

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

The genomes of archaic and ancient modern humans offer a unique way to learn about their histories and to gain insights into their unique physiologies. However, the sequencing and analysis of DNA from ancient humans is complicated by DNA degradation, chemical modifications and contamination. Recent technological advances have made it possible retrieve and sequence DNA from bones and other remains found at archaeological excavations, and we have been able to reconstruct the genomes of several Neandertals, as well as the genome of am individual from previously unknown extinct Asian hominin group who we call Denisovans. The genomes of our extinct Neandertal and Denisovan relatives offer a unique opportunity to learn about the similarities and differences between us. We have used these archaic genome sequences to identify genetic changes that are unique to the modern and archaic human lineages. Further, we have also shown that the ancestors of some of present-day people interbred with both Neandertals and Denisovans such that all present-day people outside of Africa carry approximately 2% Neandertal DNA, and that some populations, largely in Oceania, also carry DNA inherited from Denisovans. This introgressed DNA has been shown to have both positive and negative outcomes for present-day carriers: underlying apparently adaptive phenotypes such as high altitude adaptation, as well as influencing immunity and disease risk. In recent work we have identified Neandertal haplotypes that are likely of archaic origin and determined the likely functional consequences of these haplotypes using public genome, gene expression, and phenotype datasets.

DNA AND RNA MODIFICATION DETECTION VIA RAPID NANOPORE SIGNAL ALIGNMENT, ANALYSIS, AND VISUALIZATION WITH UNCALLED4

Sam Kovaka[1], Paul W Hook[2], Katharine Jenike[3], Luke Morina[2], Winston Timp[2,3], Michael C Schatz[1,3]

[1]Johns Hopkins University, Computer Science, Balitmore, MD, [2]Johns Hopkins University, Biomedical Engineering, Balitmore, MD, [3]Johns Hopkins University, Genetic Medicine, Balitmore, MD

Epigenetic and epitranscriptomic profiling is a unique strength of nanopore sequencing, as modifications can be detected from native sequencing data without specialized library preparation. Nanopore sequencing operates by measuring ionic current as DNA or RNA molecules pass through a pore, where different bases and modifications produce different signal characteristics. While modern basecallers can accurately detect CpG methylation, other modifications (including over 150 in RNA) require specialized algorithms, usually beginning with alignment of raw signal to a nucleotide reference. Generalized nanopore signal alignment is challenging since signal characteristics vary with molecule type and pore chemistry, and few methods exist to optimize signal alignment in different contexts.

Here we present Uncalled4: a toolkit for nanopore signal alignment, analysis, and visualization. Uncalled4 uses basecaller metadata to guide signal alignment, improving both speed and accuracy compared to Nanopolish and Tombo. Alignments are stored within BAM tags, enabling both compression and reference-based indexing. We apply Uncalled4 to two major applications: pore model training and modification detection. Pore models map k-mers to expected signal characteristics (i.e. current mean, standard deviation, and dwell time) and are specific to pore chemistry and molecule type. Uncalled4 can train new pore models by repeatedly aggregating signal characteristics across many alignments. We use this process to generate a model for ONT's latest r10.4.1 DNA pore, which we use to detect CpG methylation in Drosophila melanogaster. This reveals current-to-sequence patterns which reflect the biophysical structure of the r10.4.1 pore, as well as putative errors in ONT's official r10.4.1 pore model. Uncalled4 alignments also yield higher-accuracy RNA modification calls both using builtin comparison statistics and when input to third-party tools like xPore and m6Anet. We apply m6Anet using Uncalled4 and Nanopolish alignments to several human cell lines, with Uncalled4 showing consistently higher sensitivity with similar precision, especially at lower coverage levels. Finally, Uncalled4 features a variety of signal alignment visualizations, which can be viewed and manipulated interactively using a browser-based application. Uncalled4 is implemented in C++ and Python, and is available at skovaka.github.io/UNCALLED.

MULTI-HAPLOTYPE CURATION FOR HIGH PLOIDY LEVEL PLANT GENOMES – THE URTICA DIOICA STORY

Dominic Absolon, Ksenia Krasheninnikova, Shane A McCarthy, Jonathan M Wood

Wellcome Sanger Institute, Tree of Life, Cambridge, United Kingdom

Genome curation is an integral part of the Tree of Life program at the Wellcome Sanger Institute.
It assists in producing the highest possible quality assembly from the available data that is produced from any given species. Curation tools and methodology have been developed and established with a primary-alt system where by a single haplotype of a diploid genome is curated to submit a polished primary representation of the genome.
This methodology works extremely well for the majority of diploid genomes. However, as the EBP expands, a greater number of species with a higher ploidy level than 2 are beginning to be sequenced and will require curation. One of the first to come through our pipeline of this nature is Urtica dioica (great nettle - drUrtDioi1), a tetraploid plant that was initially processed through our standard assembly pipeline. However, the resulting single haplotype was of insufficient quality to curate to an adequate representation of the genome.
Instead we have explored and developed methodology for scaffolding and subsequently curating all four haplotypes at once to ensure a quality output.

ENABLING PETABASE-SCALE SEARCH OF MILLIONS OF METAGENOMES WITH BRANCHWATER

Luiz Irber[1], N Tessa Pierce-Ward[1], Suzanne M Fleishman[2], Adam R Rivers[2], C Titus Brown[1]

[1]UC Davis, Population Health and Reproduction, Davis, CA, [2]USDA-ARS, Genomics and Bioinformatics Research Unit, Gainesville, FL

Substantial growth in publicly available nucleotide sequencing data (DNA and RNA) has occurred over the last decade, driven by decreases in sequencing costs. In particular the Sequence Read Archive now has over 15 million entries containing 26 PB of data.

Shotgun metagenomes, generated by random sequencing of mixtures of microbes sampled from a microbiome are a particularly interesting resource stored in the SRA. In the past decade, hundreds of thousands of new bacterial and archaeal genomes have been isolated from public metagenomes, and several entirely new branches of life have been discovered purely through analysis of public data.

Solutions for searching this trove of genomic data are limited to metadata or queries over a small subset of all the data available. Another issue is how often the index is updated, due to the challenges of keeping up with all the new data arriving and how to add new entries without needing to recalculate the full index.

We present branchwater, a new disk-based inverted index over sourmash FracMinHash signatures that allows querying more than a million metagenomes in seconds. Implemented in Rust, it has low resource usage and efficient utilization of multiple CPUs for parallelization for the indexing and searching processes, as well as dynamically updating the index with new sequencing datasets. This index is also available for use in sourmash, allowing scaling and speeding up many of its operations, especially taxonomic profiling with gather.

The new search capability enabled by branchwater was already used to study the global biogeography of antarctic cyanobacteria, and large scale analysis of a *Klebsiella pneumoniae* outbreak. As a demonstration of the features of this method and to allow new use cases and discovery we built an index for 1M+ SRA metagenomes and made it publicly available for use at https://branchwater.sourmash.bio, exposing content-based search and result visualization. The server for this index can be run with frugal resources (4 cores, 16GB of RAM, 2TB of SSD storage) and can scale to cover more entries or increased resolution if more resources are available.

rMATS-TURBO: AN EFFICIENT AND FLEXIBLE COMPUTATIONAL TOOL FOR ALTERNATIVE SPLICING ANALYSIS OF LARGE-SCALE RNA-SEQ DATA

Jenea I Adams[1,2], Yuanyuan Wang[1,3], Zhijie Xie[1], Eric Kutschera[1], Kathryn E Kadash-Edmondson[1], Yi Xing[1,2]

[1]The Children's Hospital of Philadelphia, Center for Computational and Genomic Medicine, Philadelphia, PA, [2]University of Pennsylvania, Pathology and Laboratory Medicine, Philadelphia, PA, [3]University of California, Microbiology, Immunology & Molecular Genetics, Los Angeles, CA

Pre-mRNA alternative splicing is a prevalent mechanism for diversifying eukaryotic transcriptomes and proteomes. Regulated alternative splicing plays a role in many biological processes, and dysregulated alternative splicing is a feature of many human diseases. Short-read RNA sequencing (RNA-seq) is now the standard approach for transcriptome-wide analysis of alternative splicing. Since 2011, our lab has developed and maintained rMATS, a computational tool for discovering and quantifying alternative splicing events from RNA-seq data. The rMATS software has been widely used by the research community. Here we describe the contemporary version of rMATS – called rMATS-turbo – a fast and scalable re-implementation that maintains the statistical framework and user interface of the original rMATS software, while incorporating a revamped computational workflow with substantial improvements in speed and data storage efficiency. The rMATS-turbo software scales up to massive RNA-seq datasets with tens of thousands of samples. To illustrate the utility of rMATS-turbo, we describe two representative application scenarios. Firstly, we describe a broadly applicable two-group comparison to identify differential alternative splicing events between two sample groups, including both annotated and novel alternative splicing events. Secondly, we describe a quantitative analysis of alternative splicing in a large-scale RNA-seq dataset (~1,000 samples), including the discovery of alternative splicing events associated with distinct cell states. We detail the workflow and features of rMATS-turbo that enable efficient parallel processing and analysis of large-scale RNA-seq datasets on a compute cluster. We anticipate that rMATS-turbo will be useful for studying alternative splicing in diverse biological systems.

HAPLOTYPE-SPECIFIC KARYOTYPES RECONSTRUCTION AND COPY NUMBER ABERRATIONS/PROFILING FROM LONG READS SEQUENCING DATA

Tanveer Ahmad, Mikhail Kolmogorov

Center for Cancer Research, NCI, Bethesda, MD

Introduction. Copy number alterations/variations (CNA/CNV) is a phenomenon during cancer progression where some sections of the affected genome are duplicated or deleted. This results in heterogeneous collections of cancer cells or clones, profiling and classification of these distinct classes play a vital role in understanding the cancer heterogeneity and progression to better inform diagnosis and treatment. Here we present Wakhan, a tool to analyze haplotype-specific chromosome-scale somatic copy number aberrations and aneuploidy using long reads (Oxford Nanopore, PacBio).
Methods. Wakhan extracts Single Nucleotide Polymorphisms (SNPs) frequencies from tumor data using pileup statistics from BAM files based on phased heterozygous SNPs present in normal samples. Haplotype-specific SNPs coverage is calculated as mean in 50k bins. Due to loss-of-heterozygosity or lower coverage in some genome regions, phasing tools break the phase-sets contiguity resulting in phase-switch errors in diploid genomes. Wakhan detects those phase switch regions and corrects them by taking into consideration the changes in haplotype-specific coverage.
To detect contiguous segments with distinct (haplotype-specific) copy numbers, GaussianHMM is used to precisely cluster the SNPs bins. For integer copy-numbers assignment and karyotype reconstruction, Wakhan employs HMM again to select longer contiguous segments and assign copy numbers. Wakhan generates interactive plots for overall genome coverage with customized bin size, phase blocks, SNPs frequencies and copy number segmentation, as well as integer copy numbers.
Wakhan is available at: https://github.com/KolmogorovLab/Wakhan
Results. We applied Wahkan to detect copy numbers changes resulting from breakage-fusion-bridge (BFB) events and copy-neutral loss-of-heterozygosity (LOH) events in multiple HPV-infected cell lines (CaSki, SNU1000, SCC152, HT3). In child acute myeloid leukemia (AML) samples, normal karyotypes have been observed. Additionally, we analyzed several tumor/normal cell lines (COLO829, H2009, H1937, H1437 and H1954) which exhibit copy number and LOH events. In COLO829, considerable aneuploidy was observed across the genome. We also compared Wakhan against short-read based copy-number profiling methods, such as HATCHet on these cancer cell lines. Overall, through long-reads we got consistent haplotype specific copy-number results as compared to Illumina. In the H1437 cell line we also observed sub-clonal events in chromosomes 6 and 21. Similarly, H1937 and H1954 cell lines clearly indicate these are tetraploid (two copies) genomes. In some cell lines, we also noticed LOH in one or both arms of the chromosome even in normal samples.

COMPOSITION, FUNCTION AND STRAIN-SHARING BETWEEN MATERNAL BREAST MILK AND THE INFANT GUT MICROBIOME

Mattea Allert*[1], Pamela Ferretti*[2], Kelsey Johnson[1], Timothy Heisel[3], Cheryl A Gale[3], Ellen W Demerath[1], Dan Knights[4,5], David A Fields[6], Frank W Albert[1], Ran Blekhman[2]

[1] University of Minnesota, Department of Genetics, Cell Biology, and Development, Minneapolis, MN, [2]University of Chicago, Section of Genetic Medicine, Division of Biological Sciences, Chicago, IL, [3] University of Minnesota, Department of Pediatrics, Minneapolis, MN, [4]University of Minnesota, Department of Computer Science and Engineering, Minneapolis, MN, [5]University of Minnesota, BioTechnology Institute, College of Biological Sciences, Minneapolis, MN, [6]the University of Oklahoma Health Sciences Center, Department of Pediatrics, Oklahoma City, OK

The infant gut microbiome plays an important role in immune system modulation, nutrient adsorption and resistance to pathogen colonization. The maternal microbial contribution to the infant's gut microbiome assembly has been extensively studied, with emphasis on the role of the mother's oral, intestinal and vaginal microbiomes. However, the human breast milk microbiome remains understudied, with milk representing less than 0.3% of all the human metagenomic samples currently deposited on public repositories. To investigate the maternal milk microbiome taxonomic composition and functional potential in relation to the infant's gut microbiome, we collected breast milk and infant fecal samples in the first semester postpartum, as part of the MILk cohort (Mothers and Infants Linked for Healthy Growth). A total of 507 samples (n=173 milk, n=334 stool) were collected from 195 mother-infant pairs. The taxa most frequently present in both maternal milk and infant gut microbiome were Bifidobacterium longum, B. breve and B. bifidum. We identified strain sharing between the milk and the infant's gut for commensal as well as for potentially pathogenic species. In addition, we found that unrelated infants born at the same time in the same hospital shared a higher number of strains than unrelated infants born at the same time in different hospitals. Functional profiling analysis showed that the early infant gut microbiome (1 month) was characterized by higher biosynthetic potential compared to 6 months of age, suggesting that early microbes are more versatile and metabolically independent compared to later colonizers. The infant gut microbiome at 1 month was also characterized by an increased carriage rate of antimicrobial resistance genes (ARGs) compared to 6 months of age, suggesting the potential role of hospital exposure during birth in antimicrobial resistance carriage.

LONG-READ SEQUENCING-BASED PIPELINE FOR NEO-EPITOPE CANDIDATES FOR IMMUNOTHERAPEOTIC TARGETING OF U1 snRNA MUTATION IN CANCER

Fatemeh Almodaresi[1], Ander Diaz-Navarro[1], Andrea Senff-Ribeiro[1], Marie-Pierre Hardy[4], Quang Trinh[1], Sachin Kumar[2,5], Shimin Shuai[9], Craig Daniels[2], Xose S Puente[6,7], Elias Campo[6,8], Michael Taylor[2,3], Claude Perreault[4], Lincoln Stein[1]

[1]Ontario Institute for Cancer Research, department of Oncology, Toronto, Canada, [2]The Hospital for Sick Children, Division of Neurology, Brain Tumour Research Centre, Toronto, Canada, [3]Texas Children Hospital, Department of Pediatrics, Houston, TX, [4]University of Montreal, Immunobiology, Montreal, Canada, [5]Boston Children’s Hospital, Boston, MA, [6]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain, [7]University of Oviedo, Asturias, Spain, [8]Hospital Clínic de Barcelona, -, Barcelona, Spain, [9]Southern University of Science and Technology (SUSTech), Department of Human Cell Biology and Genetics, Shenzhen, China

Recent breakthroughs have illuminated a novel class of somatic driver mutations prevalent in challenging-to-treat cancers, including chronic lymphocytic leukemia (CLL), liver hepatocarcinoma (Liver-HCC), and pediatric medulloblastoma (MB). These mutations target the U1 small nuclear RNA (snRNA), a pivotal component of the spliceosome, inducing consistent mis-splicing patterns across thousands of genes and potentially generating aberrant proteins. We are looking to computationally identify and experimentally confirm the expression of these aberrant proteins in order to develop an off-the-shelf immunotherapy strategy targeting recurrent neo-antigens in U1 mutant tumors. Our primary objective is to identify MHC-associated peptides (MAPs) originating from mis-spliced transcripts in U1-MUT tumors, with a focus on highly expressed and transcriptionally robust candidates for adoptive T-cell therapy.

We have developed a pipeline that comprises the following steps:

* **Read Quality Control and Transcript Assembly**: We remove low-quality long-read data, and align reads with Pacbio (isoseq3) and Nanopore (Pychopper + Minimap2) -optimized protocols. We then perform quality assessment using NanoPack and LongQC. Reads are assembled using StringTie2 and combined with GffCompare.
* **Transcript Validation**: We eliminate lowly expressed simple genes while retaining complex genes with different mutant and canonical isoforms.
* **Finding Peptides**: We identify the potential ORFs for each transcript and generate ORF peptides (9-11 amino acids).
* **Filtering and Quality Control**: We exclude known and human peptides by referencing the Human proteome peptide database, to select a set of neo-peptide candidates.

After the compilation of the pipeline, we perform pathway-enrichment analysis, focusing on (U1)Mut-specific transcripts in overexpressed pathways like translation and immune signaling. Our next steps involve employing computational techniques and mass spectrometry to select strong MHC class I binding subset of peptides and quantify the abundance of discovered Neo-MAPs in cell line samples as a validation step. we plan to employ. Our pipeline enhances our understanding of splicing dysregulation in cancer and offers a potential immunotherapy avenue for U1 mutant tumors, providing hope for improved treatment options.

INTEGRATED ANALYSIS OF IMBALANCED ALLELIC EXPRESSION TO INFER GENE REGULATORY PATTERNS IN CANCER

Mona Arabzadeh, Amartya Singh, Hossein Khiabanian, Shridar Ganesan

Rutgers Cancer Institute of New Jersey, Rutgers Biomedical and Health Sciences, New Brunswick, NJ

In a diploid genome, the relative abundance of transcripts that express each allele can be dysregulated by germline single nucleotide polymorphisms (SNP) or somatic mutations in regulatory regions such as promoters and enhancers of the genes, resulting in allelic imbalance (AI). Accurate measures of AI, reflected in mutant allele frequency (AF) and relative expression count measurements in RNA, may therefore identify regulatory loci that, when genetically altered, drive transcriptomic changes in cancer. Previous efforts to identify molecular processes and pathways that are altered in tumors have primarily focused on gene-level transcriptomic analyses and do not consider allele-specific expression (ASE). In the context of tumors, ASE/AI assumes immense significance since alterations in the alleles that are preferentially transcribed or induce allele-specific expression are more likely to impact the molecular processes that involve the aberrant expression of corresponding genes.

We tested the hypothesis that altered regulation in cancer is associated with the extent of co-exhibited, imbalanced allelic expression. We developed a quantitative framework for accurately measuring AI reflected in the transcription of alleles, corrected by DNA copy-number and specimen tumor content. We integrated DNA and RNA data to investigate the underlying mechanisms of AI and analyzed allelic expressions from ~900 TCGA breast cancer data and ~300 GTEx samples from normal breast tissue. We showed that bulk sequencing eliminates enrichment of allelic expression across normal tissue and that tumor aberrations are associated with enrichment and co-exhibition of imbalanced allelic expression in genes involved in known and novel oncogenic processes.

LONGITUDINAL SINGLE-CELL GENOMIC AND TRANSCRIPTOMIC ANALYSIS OF RELAPSED PEDIATRIC AML

Byron Avihai, Amartya Singh, Hossein Khiabanian, Daniel Herranz

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

Background and hypotheses: Little is known about the mutations driving therapeutic resistance in pediatric AML, and diagnostic and therapeutic options for relapsing patients, only 30% of whom survive, are lacking. We hypothesize that relapsed pediatric AML is associated with distinct molecular signatures that are detectable at the single-cell resolution at diagnosis or at remission, and that resistance-conferring mutations correlate with changes in transcriptional profiles.

Methods: Using novel bioinformatics pipelines, we integrated single-cell DNA (MissionBio) and RNA (10x Genomics) data from 78 samples from 40 patients collected at diagnosis, remission, and/or relapse (including 9 trios), and identified genomic alterations and transcriptomic signatures associated with relapse.

Preliminary results: Genomic data from one case revealed strong associations between a pathogenic KRAS mutation (KRAS^{G13D}), an intronic BCRO mutation, and leukemic subclones. The main diagnosis leukemic clone was KRAS heterozygous and was eclipsed at relapse by a homozygous KRAS clone. This suggests that either the homozygous KRAS mutation conferred resistance to chemotherapy or that resistance-conferring mutations arose in a homozygous KRAS subclone. Meanwhile, clones carrying BCRO mutations at diagnosis never reappeared at relapse, indicating that these diagnosis clones carried a molecular signature associated with chemotherapy sensitivity. In another case, transcriptomic data revealed an expression profile of a minor diagnosis subclone (0.9% cells) that mapped to an expanded relapse population (20.3%). Associated marker genes were enriched for erythrocytic cells and bone marrow pathways, indicating that relapsed cells correspond to a prominent leukemic clone originating from a minor diagnosis leukemic subclone.

Conclusions: Integrated clonal (DNA) and biological pathway (RNA) analyses of serial samples for each patient and across the cohort highlight the underlying mechanisms of therapy resistance in relapsing pediatric AML.

Acknowledgments: New Jersey Commission of Cancer Research (NJCCR) grant (COCR23PRG006).

< Back

ENHANCING PANDEMIC PREPAREDNESS: A NOVEL PARTIAL MATCHING APPROACH FOR IDENTIFYING SIMILAR GENETIC MATERIAL FROM DIVERSE SOURCES FOR PATHOGEN SURVEILLANCE

Morteza Baradaran[1], Ryan Layer[2], Kevin Skadron[1]

[1]University of Virginia, Computer Science, Charlottesville, VA, [2]University of Colorado, Computer Science, Boulder, CO, [3]University of Virginia, Computer Science, Charlottesville, VA

Metagenomics sampling will play a pivotal role in preparing for future pandemics by enabling early identification of potential pathogens and facilitating timely intervention strategies. Multiple projects are underway to perform such environmental sampling, but data collection is scaling faster than our current analytic capacity, which primarily involves extracting and assembling viral genomes. K-mer search methods like Kraken (Wood 2014) can identify known pathogens but may miss novel pathogens. Pseudo-alignment methods (Fan 2023) rapidly estimate the mapping reads to a reference genome but do not provide the precise information needed for epidemiological or virological analysis. To ensure that the data we are collecting helps us prepare for the next pandemic, we need a search tool that can identify and extract all potentially relevant reads across thousands of unassembled and unaligned metagenomic read files that have sufficient overlap with a query sequence. Such a tool would enable epidemiologists to source an outbreak, for example, by connecting pathogens observed in a patient to environmental samples, and allow virologists to monitor viral trends such as binding and immune evasion or connecting relevant reads from different surveillance locations.

To facilitate such a search capability, we propose a novel partial matching (PM) method. PM identifies alignments with sufficient overlap in prefix/suffix regions between two sequences, characterized by notable consecutive exact matches and a few edit distances. Prior large-scale sequence indexes identify and remove sequencing errors to control the index size. With unstructured and unclassified data, identifying sequencing errors is impossible, making us vulnerable to performance issues due to the inability to shrink the index size by removing errors. To mitigate this, we categorized errors into k-mer classes based on their frequency: cheap, medium, and expensive, akin to cheap k-mer selection in FastHash (Xin 2013). By concentrating on the cheap k-mers, we crafted a storage-efficient index that not only has a fast probing time but also retrieves similar reads to the query from a vast database of millions of reads. To simulate a pandemic surveillance scenario, we queried a database of 16 million reads from a bat feces metagenomics sample, using reads simulated from three COVID-related genomes. As compared with BWA (Li 2009), PM not only finds the best alignment (no false negatives) but also uncovers all the alignments that meet the user's criteria. PM retrieved similar reads about twice as fast as BWA, peaking at nearly 1,500 reads per second. At this rate, we expect that PM will scale to the thousands of metagenomic environmental samples the community will generate in the coming years, allowing public health officials and researchers to fully exploit this critical resource's capabilities.

ENHANCED GENOME-BASED TAXONOMY FOR PRECISE IDENTIFICATION OF FUSARIUM PATHOGENS

Kassaye H Belay[1,2], Reza Mazloom[3], Lenwood Heath[3], Boris A Vinatzer[2]

[1]Virginia Tech, Genetics, Bioinformatics, and Computational Biology, Blacksburg, VA, [2]Virginia Tech, School of Plant and Environmental Sciences, Blacksburg, VA, [3]Virginia Tech, Computer Science, Blacksburg, VA, [4]Virginia Tech, Computer Science, Blacksburg, VA

The genus Fusarium contains species detrimental to plant, animal, and human health due to their pathogenicity and mycotoxin production. Despite substantial research, the taxonomy of the genus is still contentious. Genome sequencing and analysis can address the limitations of more traditional taxonomic methods. In addition, accurate whole genome-based classification can improve the resolution of pathogen identification, which in turn can help to contain disease outbreaks. Therefore, as a first step towards a whole-genome-based taxonomy of Fusarium, we examined poorly resolved taxonomic relationships involving 276 fungal species. We combined BUSCO to detect single-copy genes, MUSCLE for alignment, TrimAI for sequence refinement, and IQ-Tree for a core genome tree construction. Also, we used sourmash and Life Identification Numbers (LINs) for genome classification. Results from sourmash and LINs were consistent with the core genome tree constructed from 391 single copy genes showing that these computationally efficient methods can be used to complement, or even replace, phylogenetic methods, accelerating both classification and identification of Fusarium and other fungi, crucial for maintaining plant health and food security.

QUANTUM COMPUTING COMES TO THE GALAXY

Bryan Raubenolt[1], Fabio Cumbo[1], Jayadev Joshi[1], Daniel Blankenberg[1,2,3,4]

[1]Cleveland Clinic, Center for Computational Life Sciences, Cleveland, OH, [2]Cleveland Clinic, Genomic Medicine Institute, Cleveland, OH, [3]Cleveland Clinic, Lerner Research Instituted, Cleveland, OH, [4]CCLCM of CWRU CoM, Molecular Medicine, Cleveland, OH

The era of quantum computing (QC) is upon us.

As QC continues to evolve from being a technology in its infancy, to something that is poised to revolutionize many aspects of daily life, it continues to capture the interest of academia, government, and industry alike. I, myself, am peaked. It is thus of great importance for members of our scientific community to familiarize themselves with the theory and applications of this new computing method, and most importantly, how to select and develop problems and potential algorithms for these devices.

Here, we introduce the incorporation of Qiskit, IBM's Python-based software stack for quantum computing, into the genomics-centric Galaxy ecosystem. This combined Open Source set of packages is fully available and licensed under Apache 2.0. (I am at least as surprised as you, perhaps more.)

The basic theories of Quantum Computing is re-visited in this presentation, along with the many important components of currently existing coding frameworks and syntaxes. Furthermore, we introduce an apdapted series of previously developed Qiskit tutorials and Jupyter notebooks, now integrated, into the Galaxy Platform, tuned towards HealthCare and Life Sciences, presenting new ways of solving some of our world's most pressing problems.

In our second half, we explore the dissemination of Quantum Computing Education while leveraging the collaborative Galaxy Collective. Our possibilities are endless, but the useful surface area is much-less so.

Collectively, the Clinic invites formally interested participants to engage with our Discovery Accelerator

WHAT ABOUT B.O.B.? BIO-ONTOLOGY-BIASES: THE EFFECTS OF STUDY BIAS ON LESS STUDIED GROUPS AND CLASSES IN KNOWLEDGE GRAPH EMBEDDING METHODS

Michael S Bradshaw[1,2]

[1]University of Colorado Boulder, Computer Science, Boulder, CO, [2]University of Colorado Boulder, Biofrontiers Institute, Boulder, CO

Knowledge graph (KG) embedding (KGE) algorithms have emerged as a promising method for numerous biomedical tasks such as drug target prediction, disease module identification, and variant prioritization. Unlike traditional graph-based tools, KGEs use higher order patterns and edge type information. Like their predecessors, KGEs are hindered by the biases of their underlying graphs.

Edges in biological networks are typically touted as ground truth, but they are in fact neither complete nor perfectly accurate. If they were, KG maintainers would not regularly release updates. Recently, Lucchetta et al. 2023 showed that study bias and unequal use of proteins as bait or prey in binding experiments and the aggregation of many small networks resulted in topologically imbalanced PPI networks that exhibited power-law (PL) degree distributions. The hub nodes of these PL networks were enriched for various diseases, notably cancers. When bait-prey imbalance was controlled for, the hub nodes changed, and no enrichment was found.

We found similar results using a large aggregate PPI network in which the top 10 highest degree nodes were enriched for over 400 biological processes including tumor suppression. In contrast, when we used a bait-prey balanced network the top 10 hub nodes were not enriched for any processes.

When KGE models are then fit to imbalanced networks the bias is passed on to their predictions. Bonner et al. 2022 demonstrated this well and showed that node degree and predictive scores for that node being connected to any other node were very strongly correlated. This stood true for numerous biological interactions. Even if a lower degree node had an edge to a specific target node included in the training set, the low degree node would be outranked by unrelated higher degree nodes. Based on this finding, if certain classes of genes, drugs, and diseases rank disproportionately well due to study bias and subsequent topological imbalance we must also ask what the same system deprioritized.

We evaluate several groups of genes, phenotypes, and diseases in a link prediction task. We trained and evaluated our KGE link predictor in a realistic way, dividing our training, validation, and test sets of edges based on the year they were added to the KG. Among our results, we found that there were 240% more disease-phenotype annotations specific to males than females but that sex-specific differentially expressed genes as a whole were significantly ($p<0.0001$) deprioritized in link prediction tasks. Additionally, we found that there were 1,116 ancestry specific variant phenotype annotations for European populations but only 249 and 317 for African and Latino populations respectively. Despite this large discrepancy in the number of annotations, the difference in link prediction results was not quite significant with $p=0.15$ for African and $p=0.25$ Latino.

BENCHMARKING LONG-READ SOMATIC STRUCTURAL VARIANT CALLERS ON A COLLECTION OF TUMOR/NORMAL CELL LINES

Asher Bryant[1], Ayse Keskus[1], Tanveer Ahmad[1], Ataberk Donmez[1], Isabel Rodriguez[2], Nicole M Rossi[2], Yi Xie[2], Byunggil Yoo[5], Rose Milano[2], Hong Lou[3], Jimin Park[4], Joshua Gardner[4], Brandy McNulty[4], Karen Miga[4], Midhat Farooqi[5], Benedict Paten[4], Michael Dean[2], Mikhail Kolmogorov[1]

[1]Center for Cancer Research, National Cancer Institute, Bethesda, MD, [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, [3]Leidos Biomedical Research, Inc., National Laboratory for Cancer Research, Frederick, MD, [4]Genomics Institute, UC Santa Cruz, Santa Cruz, CA, [5]Children's Mercy Hospital, Children's Mercy Hospital, Kansas City, MO

Cancer genomes acquire a number of somatic aberrations, including structural variants (SV). These large-scale (> 50bp) and complex alterations are known drivers of tumorigenesis in ~50% of cancers. SV detection could be key in cancer diagnosis, treatment and prevention. Yet many cancer genomics studies have only reliably detected single nucleotide variants and small indels due to technological limitations of short-read sequencing.

While structural variants can be inferred from discordant short reads, as much as 70% of SVs remain undetected. This is especially true for SVs in regions where GC content is greater than 45% or regions which harbor tandem repeats and segmental duplications. Advances in sequencing technology through long reads have enabled direct observation of SVs across a larger portion of the genome. To harness this technology, several germline and somatic long-read SV callers have recently been developed.

SV callers output records in variant call format (VCF), originally introduced by the 1000 Genomes Project to allow comparison of variants across many samples. VCF is exceptionally customizable with minimal required fields to suit the broadest range of datasets. Unfortunately, this is a drawback of the format for benchmarking, as each of the long-read callers can represent the same SV in VCF in different ways. Another challenge of benchmarking somatic SV callers is the dearth of truth sets, with only one relatively small benchmark available to date.

In response, we developed a benchmarking and merging script compatible with VCF files of recently developed long-read SV callers: Severus, NanomonSV, Savana and Sniffles2. Second, in addition to benchmarking the callers against a COLO829 melanoma cell line truthset, we sequenced five other tumor/normal cancer cell lines using ONT, Illumina, Hi-C and PacBio HiFi. We show that our script offers easy-of-use and reliability, without the need for preprocessing VCF files. Further, we show that for the COLO829 truth set, Severus has the highest recall, while Savana has the highest precision. For validation using the newly sequenced cell lines, we merged the calls from multiple long-read methods, a common strategy for short-read SV calling. Compared to other methods, Severus has the fewest private calls and recovered the most calls shared by at least 2 other callers, indicative of both high recall and precision.

mENRICH-SEQ: METHYLATION-GUIDED ENRICHMENT SEQUENCING AND NOVEL INFORMATIC METHODS TO INVESTIGATE SPECIFIC BACTERIAL TAXA OF INTEREST DIRECTLY FROM MICROBIOME

Lei Cao[1], Yimeng Kong[1], Yu Fan[1], Mi Ni[1], Alan Tourancheau[1], Magdalena Ksiezarek[1], Edward A Mead[1], Tonny Koo[2], Melissa Gitman[2], Xue-Song Zhang[3]

[1]Icahn School of Medicine at Mount Sinai, Genetics and Genomic Sciences, New York, NY, [2]Icahn School of Medicine at Mount Sinai, Department of Pathology, New York, NY, [3]Rutgers University, Center for Advanced Biotechnology and Medicine, New Brunswick, NJ

Background: Although metagenomics has enabled the comprehensive study of microbiomes, there’s an increasing need to examine certain bacterial taxa of interest, rather than sequencing all the species in a microbiome sample. This remains challenging due to the presence of diverse bacterial species and uneven sequencing depth caused by skewed abundance distribution. Although culture is one way to isolate specific bacteria, it is time- and resource-consuming, and hard to scale. Some existing tools can deplete mammalian host gDNA from bacterial gDNA, but they cannot effectively differentiate among the diverse bacterial taxa.

Core idea: We developed a method that takes advantage of bacterial DNA methylation, which is the foundation of restriction modification systems, naturally differentiates self from non-self DNA, and has been exploited as natural epigenetic barcodes to group highly similar species and strains in previous metagenomic analyses. In this work, we rationally choose individual, or combinations of methylation-sensitive restriction enzymes (REs) that digest and eliminate host DNA and background microbial DNA at non-methylated sequence recognition sites while leaving the gDNA from targeted bacterial taxa intact. This core idea is integrated with library preparation procedures in a way that only non-digested DNA libraries are sequenced.

Methods: We developed novel informatics methods for analyzing mEnrich-seq reads, examine antibiotic resistance genes, and improve metagenomic assembly. We evaluated the mEnrich-seq using a diversity of mock microbiome samples and applied the method to enrich pathogenic or beneficial bacteria from human urine and fecal samples, as well as low-abundance bacteria. In addition, we developed novel methods for analyzing and in-depth characterization of mEnrich-seq reads to differentiate species enriched due to matched DNA methylation vs. carry-over because of sparse frequency of restriction sites on genome.

Results & conclusions: 1) We demonstrated that mEnrich-seq is able to enrich both well-characterized bacteria (pathogenic E. coli and commensal Akkermansia muciniphila) and genomes with non-conserved or unknown methylation motifs coupled with de novo methylation discovery, achieving up to 117-fold enrichment without systematic coverage bias. 2) With novel informatics methods, we found mEnrich-seq significantly reduces sequencing reads from background bacteria, therefore enhancing and simplifying the assembly of the enriched bacteria. 3) We also assessed 4601 bacterial strains for the broad applicability of mEnrich-seq and found that ~68% of bacterial genomes can be targeted by mEnrich-seq with at least one RE, representing 54.78% of the species examined. 4) mEnrich-seq is compatible with different long- and short- read sequencing platforms and is generally applicable to other microbiomes such as environmental samples.

SINGLE-CELL, LONG-READ SEQUENCING OF THE MOUSE HIPPOCAMPUS REVEALS LEARNING-INDUCED ALTERNATIVE SPLICING PATTERNS AND TRANSCRIPT ISOFORM EXPRESSION ACROSS CELL TYPES

Sheridan H Cavalier[1], Paul Hook[2], Winston Timp[2], Richard Huganir[1]

[1]Johns Hopkins SOM, Neuroscience, Baltimore, MD, [2]Johns Hopkins SOM, Biomedical Engineering, Baltimore, MD

Neurons respond to experience-induced changes in brain activity through differential gene expression and alternative splicing. Although this response is necessary for learning and memory formation, our understanding of the incredibly diverse activity-induced transcriptome has been limited to gene expression-level assays or long read isoform profiling of targeted, specific genes of interest. The inherent cellular heterogeneity of brain tissue and the wide possible spectrum of individual cellular responses to activity also pose challenges to developing the methodology and analysis needed to generate a dataset that profiles learning-induced alternative splicing with single-cell resolution. To circumvent these challenges and to profile the learning-induced transcriptome, we leveraged 10X Genomics and Oxford Nanopore Technologies to create the first single-cell long-read sequencing dataset of the mouse hippocampus. In addition, we have developed an analysis framework for flagging genomic loci where learning-induced alternative splicing events occur.

Following a contextual learning paradigm, full-length, single-cell barcoded cDNA was generated from mouse hippocampal tissue using 10X Genomics. cDNA was circularized and amplified using a modified R2C2 error-correction protocol (Volden et al. 2018) prior to long-read sequencing on the ONT PromethION. Using this strategy, we demultiplexed an average 92% of reads and generated the sequencing depth per cell needed to identify 22 hippocampal cell subtypes. Furthermore, our average length of mapped reads sits at ~1kb, ten times the length of the canonical 10X Genomics 3’ Expression library when sequenced with Illumina. Reads were assembled into a transcriptome using FLAIR and were mapped and quantified using Salmon. The resulting isoform-count-per-cell matrix was used to calculate isoform abundances at each locus for each cell within a cell type. For each gene we measured the Jensen Shannon Divergence of each cell from the average naive isoform distribution at that locus. We then used linear modeling to interrogate the relationship between learning and splicing at each gene, and flagged genes that have a learning-induced departure from the typical naive splicing pattern. In this way we were able to profile learning-induced changes in splicing across thousands of genes in tens of cell types for fewer than 1 billion total reads.

< Back

SPLAM: A DEEP-LEARNING-BASED SPLICE SITE PREDICTOR THAT IMPROVES SPLICED ALIGNMENTS

Kuan-Hao Chao[1,2], Alan Mao[1,2,3], Steven L Salzberg[1,2,3,4], Mihaela Pertea[2,3]

[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, [4]Johns Hopkins University, Department of Biostatistics, Baltimore, MD

The process of splicing messenger RNA to remove introns plays a central role in creating genes and gene variants. Here we describe Splam, a novel method for predicting splice junctions in DNA based on deep residual convolutional neural networks. Unlike some previous models, Splam looks at a relatively limited window of 400 base pairs flanking each splice site, motivated by the observation that the biological process of splicing relies primarily on signals within this window. Additionally, Splam introduces the idea of training the network on donor and acceptor pairs together, based on the principle that the splicing machinery recognizes both ends of each intron at once. We compare Splam's accuracy to recent state-of-the-art splice site prediction methods, particularly SpliceAI, another method that uses deep neural networks. Our results show that Splam is consistently more accurate than SpliceAI, with an overall accuracy of 96% at predicting human splice junctions. Splam generalizes even to non-human species, including distant ones like the flowering plant Arabidopsis thaliana. Finally, we demonstrate the use of Splam on a novel application: processing the spliced alignments of RNA-seq data to identify and eliminate errors. We show that when used in this manner, Splam yields substantial improvements in the accuracy of downstream transcriptome analysis of both poly(A) and ribo-depleted RNA-seq libraries. Overall, Splam offers a faster and more accurate approach to detecting splice junctions, while also providing a reliable and efficient solution for cleaning up erroneous spliced alignments.

Preprint: https://doi.org/10.1101/2023.07.27.550754

SCALABLE TELOMERE-TO-TELOMERE ASSEMBLY FOR DIPLOID, POLYPLOID AND CANCER GENOMES WITH DOUBLE GRAPH

Haoyu Cheng[1,2], Mobin Asri[3], Julian Lucas[3], Sergey Koren[4], Heng Li[1,2]

[1]Dana-Farber Cancer Institute, Department of Data Sciences, Boston, MA, [2]Harvard Medical School, Department of Biomedical Informatics, Boston, MA, [3]University of California, Santa Cruz, Genomics Institute, Santa Cruz, CA, [4]National Human Genome Research Institute, National Institutes of Health, Genome Informatics Section, Computational and Statistical Genomics Branch, Bethesda, MD

The emergence of accurate PacBio High-Fidelity (HiFi) long reads has revolutionized the assembly of large genomes, making high-quality haplotype-resolved assembly a routine procedure. However, HiFi reads are often not long enough to resolve long exact repeats, resulting in fragmented components around repeat-rich regions such as centromeres. Recent advances by Oxford Nanopore Technologies (ONT) have enabled the generation of ultra-long reads, which are approximately 5-10 times longer than HiFi reads though at relatively lower accuracy. Existing Verkko assembler has demonstrated that, combining HiFi and ultra-long reads could produce telomere-to-telomere assemblies of diploid samples. However, it does not fully phase a single diploid sample without parental data and thus results in incomplete assembly. It may produce relatively fragmented assembly at lower read coverage and is unable to produce haplotype-resolved assemblies of polyploid samples. Verkko is also compute intensive, making it costly to deploy Verkko to a large number of samples.

For the efficient near telomere-to-telomere assembly, we developed hifiasm (UL) that tightly integrates PacBio HiFi, ONT ultra-long, Hi-C reads and trio. Unlike Verkko that is based on the multiplex de Bruijn graph, hifiasm (UL) represents sequences with two string graphs. The first string graph is built from HiFi reads, while the second string graph is built from ultra-long reads in reduced representation. Hifiasm (UL) then merges the two graphs to produce the final assembly graph. The use of two assembly graphs at different scales separates hifiasm (UL) from other assemblers. By utilizing twenty-two human and two plant genomes, we demonstrate that our algorithm is around an order of magnitude cheaper than existing methods, while producing better diploid and haploid assemblies. Notably, our algorithm is the only feasible solution to the haplotype-resolved assembly of polyploid genomes. We also successfully applied our algorithm to produce near telomere-to-telomere assemblies for complex cancer cell lines.

SOMATIC DRIVER GENE ALTERATIONS ARE ASSOCIATED WITH PREDICTIVE ANTICANCER RESPONSE AND PROGNOSTIC ASSESSMENT IN PANCREATIC DUCTAL ADENOCARCINOMA

Eunwoo Choi[1,2], Jiyoon Hong[1,2], Seungmin Bang[3], HeeSeung Lee[3], Sangwoo Kim[1,2]

[1]Yonsei Univesrsity College of Medicine, Graduate School of Medical Science, Brain Korea 21 Project, Seoul, South Korea, [2]Yonsei Univesrsity College of Medicine, Department of Biomedical Systems Informatics, Seoul, South Korea, [3]Yonsei Univesrsity College of Medicine, Division of Gastroenterology, Department of Internal Medicine, Seoul, South Korea

The prognosis of Pancreatic ductal adenocarcinoma (PDAC) is extremely poor, and most patients with PDAC still receive palliative chemotherapy. There is an increased use of fluorouracil, leucovorin, irinotecan, and oxaliplatin (FOLFIRINOX) in the treatment of PDAC; however, it is noteworthy that there are few validated biomarkers of anticancer response, particularly in Asian populations. Using a cohort of 120 Korean patients, we identify genes associated with clinical features and drug sensitivities. SMAD4, frequently mutated in PDAC patients, was found as a gene associated with a favorable prognosis in response to FOLFIRINOX chemotherapy. In the responder group, 77 mutations (including SMAD4) with high variant allele frequency are associated with the DNA damage response (DDR) and the TGF-beta signaling pathway; Somatic mutation of platinum drug resistance pathway driver genes (including BRCA1, MLH1 and ATM) is mutually exclusive of loss of TGF-beta pathway driver genes (including SMAD4), suggesting distinct alternative causes of response to FOLFIRINOX. Considering the known interactions between DDR, TGF-beta pathways, and the hedgehog pathway, it's intriguing to suggest that their complex interplay, combined with FOLFIRINOX's unique hedgehog inhibition, may contribute to FOLFIRINOX responsiveness. Clinical features, including the primary tumor location, and metastatic progression demonstrate a significant correlation with SMAD4 mutations, leading to a notably improved prognosis with FOLFIRINOX compared to alternative chemotherapies such as Gemcitabine and Abraxane. Finally, several mutations (including BRCA1, BRCA2 and SMAD4) predict FOLFIRINOX sensitivity based on some clinical features. Together, these findings shed new light on the potential for development of anticancer therapy and prognostic prediction platform using the personalized model of pancreatic cancer.

QuadST: A POWERFUL AND ROBUST APPROACH FOR IDENTIFYING CELL-CELL INTERACTION CHANGED GENES ON SPATIALLY RESOLVED TRANSCRIPTOMICS

Jinmyung Choi[1,2], Pei Wang[3,4], Guo-Cheng Yuan[3,5], Xiaoyu Song[1,2]

[1]Icahn School of Medicine at Mount Sinai, Institute for Health Care Delivery Science, Department of Population Health Science and Policy, New York, NY, [2]Icahn School of Medicine at Mount Sinai, Tisch Cancer Institute, New York, NY, [3]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences, New York, NY, [4]Icahn School of Medicine at Mount Sinai, Icahn Institute for Data Science and Genomic Technology, New York, NY, [5]Icahn School of Medicine at Mount Sinai, New York, Charles Bronfman Institute for Personalized Medicine, New York, NY

Recent advances in spatially resolved transcriptomics enabled to profile spatial organization of cells and their transcriptome in native tissue. In conjunction, various statistical and computational methods have been developed to infer cell-cell interactions, and genes and regulatory processes involved in the interaction from spatial transcriptome data. As an effort to improve upon existing methods, we developed a novel statistical framework called QuadST that allows unbiased discovery of cell-cell interaction changed genes (ICGs) from spatial transcriptome data.

QuadST is motivated by an idea that genes involved in specific cell type pair's interaction can show spatially variable expression levels, which can depend on the cell type pairs' distance. This allows to infer ICGs in a specific cell type pair's interaction from modeling and testing the association between the cell type pair's distance and gene expression level. In order to test the presence of such distance-expression association, QuadST controls False Discovery Rate (FDR) empirically, by contrasting cumulative associations at symmetrically lower and higher distant quantiles, simultaneously across all genes.

To evaluate the performance of QuadST, we designed simulation studies with a set of spatial cell compositions and gene expression levels. We illustrated that QuadST 's modeling, and inference approach provide better or comparable power than other compared methods, and consistent FDR control, even in the presence of unadjusted confounder. To demonstrate the applicability of QuadST, we applied it to seqFISH+ and MERFISH spatial transcriptome datasets, profiled from mouse brains. We showed that QuadST can identify ICGs that are presumed to play a role in specific cell type pairs' interaction (e.g., synaptic pathway genes among excitatory neuron cell's interaction). These suggest that QuadST can be a useful tool to discover genes and regulatory processes involved in specific cell type pair's interaction unbiasedly.

SWIFT PAN-GENOMIC METHODS FOR COMPREHENSIVE GENOME ANNOTATION IN CROP GENOMES

Kapeel Chougule[1], Sharon Wei[1], Zhenyuan Lu[1], Andrew Olson[1], Doreen Ware[1,2]

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

While the production of high-quality genome assemblies from long reads has become a common practice thanks to advanced assembly algorithms, the accurate annotation of gene structures remains a significant challenge. This challenge arises due to the predictive nature of the algorithms and the inconsistency in available transcriptome evidence. A single reference genome annotation often falls short in representing the full coding potential of a species. De novo or ab-initio gene annotations also encounter issues, whether it be sensitivity or specificity problems stemming from the absence of accession-specific evidence or inadequately trained HMMs for gene prediction. As more accessions are sequenced and annotated within a species, there arises a need to establish pan-genes, which encompass all known syntenic orthologs for a gene model and can be traced back to their original sources. To tackle this challenge, we have developed a pan-genomic approach that leverages representative pan-gene models selected through a comparative analysis of gene family trees created using the Ensembl Compara pipeline. We have compared and benchmarked this approach against other methods that rely on phylogeny and alignment for clustering pan-genes.To propagate these pan-gene representatives onto the genome assemblies of other unannotated accessions, we employ Liftoff and subsequently enhance the gene structures using available transcriptome evidence through PASA. This approach has been benchmarked across multiple genome assemblies of maize, oryza, sorghum, and grapevine varieties.To assess the quality of gene structural annotations, we employ the Gramene gene tree curation tool, allowing us to visually identify inconsistent gene models and flag them for potential manual curation. Furthermore, we characterize pan-gene sets based on taxonomic age and their presence in each genome, classifying them as core, shell, or orphan genes.

GENARK: TOWARDS A MILLION UCSC GENOME BROWSERS.

Hiram Clawson[1], Brian T Lee[1], Brian J Raney[1], Bogdan M Kirilenko[2,3,4], Jonathan Casper[1], Michael Hiller[2,3,4], Robert M Kuhn[1], Jairo N Gonzalez[1], Angie S Hinrichs[1], Christopher M Lee[1], Luis R Nassar[1], Gerardo Perez[1], Brittney Wick[1], Joel Armstrong[1], Matthew L Speir[1], David Haussler[1], W. James Kent[1], Maximilian Haeussler[1]

[1]University of California, Genomics Institute, Santa Cruz, CA, [2]LOEWE Centre for Translational Biodiversity Genomics, Frankfurt, Germany, [3]Senckenberg Research Institute, Frankfurt, Germany, [4]Goethe University Frankfurt, Institute of Cell Biology and Neuroscience, Frankfurt, Germany

Interactive graphical genome browsers are essential tools in genomics, but they do not contain all the recent genome assemblies. We create Genome Archive (GenArk) collection of UCSC Genome Browsers from NCBI assemblies.
Built on our established track hub system, this enables fast visualization of annotations. Assemblies come with gene models, repeat masks, BLAT, and in silico PCR. Users can add annotations via track hubs and custom tracks. We can bulk-import third-party resources, demonstrated with TOGA and Ensembl gene models for hundreds of assemblies. Three thousand two hundred sixty-nine GenArk assemblies are listed at **https://hgdownload.soe.ucsc.edu/hubs/** and can be searched for on the Genome Browser gateway page.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-03057-x

RHEA: RECOVERING HORIZONTAL GENE TRANSFER EVENTS FROM METAGENOME ASSEMBLY GRAPHS

Kristen D Curry[1], Rayan Chikhi[2], Eduardo Rocha[3], Todd J Treangen[1]

[1]Rice University, Computer Science, Houston, TX, [2]Institut Pasteur, Computational Biology, Paris, France, [3]Institut Pasteur, Microbial Genomics, Paris, France

Along with the influx of publicly available metagenomic sequences and advances in long read sequences comes great potential for new computational methods capable of extracting novel biological discoveries directly from existing metagenomes[1]. Methods of reducing a metagenome to a profile described by either taxonomic classification relative abundance, gene function presence, or k-mer counts do not capture the gene sharing events or evolutionary activity occurring in the community. Microbiomes are populations of microbes rife in mobile genetic elements driving gene transfer within and across bacterial species[2]. To track gene flux within microbial populations we have developed RHEA, a novel method for reducing metagenomes based on the structure of its contig assembly graph. Contrary to the conventional binning based approaches, we partition the assembly graph into “integration clusters.” Each cluster is inclusive of variations manifested within genomes corresponding to a specified taxonomic rank. These clusters can be utilized to identify signatures of horizontal gene transfer (HGT), host environments of mobile genetic elements, and evolving adaptive defense mechanisms. By taxonomically classifying clusters on the contig graph, we can detect which genomes are undergoing active mutations, insertions, and deletions, and can additionally establish shared genes between distally related genomes. We demonstrate the effectiveness of RHEA in two comparative metagenomic analyses; one in relation to type 2 diabetes in elderly women (Illumina sequences) and the other regarding fermentation in cheese making (Oxford Nanopore Technologies sequences). Additionally, we show how contig graphs in RHEA can be used to establish hosts of inserted prophages, hosts and donors of CRISPRs, and detect novel mobile genetic elements in long read sequences. In conclusion, RHEA holds substantial promise for providing a deeper understanding of strain dynamics and interactions present within metagenomic data.

[1] Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., & O’Sullivan, C. (2022). The Sequence Read Archive: A decade more of explosive growth. Nucleic Acids Research, 50(D1), D387–D390. https://doi.org/10.1093/nar/gkab1053

[2] Brito, I. L. (2021). Examining horizontal gene transfer in microbial communities. Nature Reviews Microbiology, 19(7), Article 7. https://doi.org/10.1038/s41579-021-00534-7

DEVELOPMENT OF A HAPLOTYPE-AWARE ASSEMBLY PIPELINE FOR ANALYSIS OF REARRANGEMENTS AT THE HUMAN CYP2D6 LOCUS

Daisy Dahiya, Benjamin Alleva, Florencia Pratto, R. Daniel Camerini-Otero

National Institutes of Health, NIDDK, Bethesda, MD

Meiosis is a specialized cell division that leads to the formation of gametes. DNA Double Strand Breaks (DSBs) are formed and subsequently repaired during meiosis to allow the exchange of genetic material between homologous chromosomes. In humans, DSBs cluster at regions in the genome called hotspots which are dependent upon the DNA-binding protein PRDM9. DSB repair may also occur using non-allelic sequences which could result in gross genomic rearrangements. In the germline, this type of repair can lead to heritable diseases.

We are interested in examining whether DSB hotspots in repeat regions of the genome lead to a higher propensity for rearrangements at those loci in humans, which is not yet fully understood. We analyzed the repetitive CYP2D6 locus known to harbour multiple DSB hotspots defined by different alleles of PRDM9. CYP2D6 is highly variable and located within a 40kb region containing two pseudogenes (CYP2D7 and CYP2D8) with >90% homology to CYP2D6. We used a long-range, overlapping PCR assay to analyse rearrangements in 319 individuals including 57 parent-offspring trios (1K Genomes Project).

In order to effectively and efficiently analyze our CYP2D6 locus long read (nanopore sequencing) amplicon data for rearrangements, we sought to create a diploid assembly for each sample. Unfortunately, no existing assembly software for long reads works successfully with our data, therefore, we developed an in-house haplotype-aware assembly pipeline. We first used Clair3 and whatshap for variant calling, phasing and separating reads from each amplicon into two haplotypes. We then map sequences from the first amplicon to the 1kb region upstream of the 40kb CYP2D6 region using BLAST. Corresponding to the two haplotype sequence files, we get two sets of mapped reads, from which we create consensus sequences using Flye to construct two contigs per amplicon. Next, we check for shared variants between overlapping amplicon consensus sequences to determine the haplotype. Then we pick 1kb region at the end of these two consensus sequences and map the two haplotype files from the next amplicon to each of these 1kb sequences. We repeat these steps for all the amplicons and get two contigs corresponding to each amplicon and then stitch the contigs for each haplotype, resulting in a seamless haplotype-aware assembly of the CYP2D6 locus in diploid samples. Here, we present an assembly pipeline providing a new method for assembling difficult to analyze regions of the genome when using a long-range overlapping PCR assay as input.

DIVERSITY AND REPRESENTATION OF SOUTH ASIAN GENOMES

Arun Das, Michael C Schatz

Johns Hopkins University, Department of Computer Science, Baltimore, MD

The rapid growth in genomics has not been uniform across the full range of human diversity, leaving the majority of the world's populations poorly represented and resulting in systemic biases that can have serious impacts on clinical interpretations and other analysis. In this work, we aim to utilize recent advancements in genome sequencing and assembly to better understand the variation present in South Asian (SA) populations.

Using high quality short read data from 635 SA individuals in the 1000 Genomes Project and Simons Genome Diversity Project, we investigated the variation between these individuals relative to linear and pangenome references. To do this, we follow a similar pipeline used in the creation of the African Pan-Genome (Sherman et al. 2019) which assembles contigs from reads that were unaligned and/or poorly aligned relative to a chosen reference genome. We then attempt to place the larger contigs in the reference, allowing us to identify variants and novel sequences.

We evaluate GRCh38, T2T CHM13 and draft human pan-genome references from the Human Pangenome Reference Consortium (HPRC). When using CHM13, we observe improved alignment rates (+0.5%) relative to GRCh38, reflecting the newly resolved regions. Interestingly, we still assemble ~1 Mbp of sequence from unaligned reads and ~15 Mbp of sequence from poorly aligned reads per individual, highlighting that widespread population-specific sequence remains missing. Improvements in alignment rate (+0.3%) and reductions in the amount of sequence assembled continue when using the HPRC pangenome references. Still, we find ~1 Mbp of sequence from unaligned reads and ~10 Mbp of sequence per individual assembled from poorly aligned reads, including some contigs that are over 60 Kbp in size. We also validate these contigs using long read data from two individuals, and find that ~38% of their previously unaligned reads and ~85% of >1 Kbp contigs assembled from them align well to a long read assembly.

We attempt to place the unaligned read contigs against the reference, with placements increasing as we move from GRCh38 to CHM13 and the pangenome references (+10%). These placements are evenly across the chromosomes, and overlap a range of biologically significant regions. We also find the majority of these contigs to be private to the individual they are assembled from, with just 20% of all >5 Kbp contigs across the full SA set being found in more than one individual, comprising just 50 Mbp out of the 350 Mbp of total novel sequence contained in >1 Kbp unaligned read contigs.

We also aligned RNA-seq data from 140 SA individuals against the assembled contigs to investigate their transcriptional potential. Across this set, we find that on average 885 reads align to the unaligned read contigs, and 1.56M to the poorly aligned read contigs per individual. These alignments are concentrated in a small subset of the contigs, with a few contigs accumulating the vast majority of aligned RNA-seq reads.

DATA ACCESS ARCHITECTURE AT "GALACTIC" SCALE: LESSONS LEARNED (SO FAR)

John Davis

Johns Hopkins University, Biology, Baltimore, MD

Galaxy (galaxyproject.org) is a globally-distributed open source software platform that connects analysis tools, datasets, compute resources, a graphical user interface, and a programmatic API. It enables accessible, reproducible, and collaborative data science and is used by thousands of scientists. It has been in continuous development for more than 15 years and is among the most active open source projects in the world.

From an engineering standpoint, Galaxy is a large distributed data-driven system, with data taking on many forms - from configuration files and reference data to tool wrappers and physical datasets. At the core of the system is the data model: a definition of the data objects and their associations that facilitate the application's business logic. These data objects are persisted in a relational database. Data, or database access in the context of a distributed multi-process and multi-threaded application is a nontrivial set of challenges, from architectural decisions to the daily grind of global collaborative development and system maintenance. This talk aims to decompose this complexity using Galaxy as an example.

I will start with an overview of Galaxy's data model, which can be represented as a dense graph of 150+ entities and 400+ relationships. I will explain how a Python application interacts with a database using concepts and protocols defined in Python's DBAPI, and why such an approach is unrealistic in a system like Galaxy without an extra layer of abstraction. This extra layer is SQLAlchemy - a SQL automation tool kit and object-relational mapper, as well as the de facto tool of choice for large Python applications. The core of this talk will be a whirlwind tour of SQLAlchemy in the context of Galaxy, including our use of its data definition language, its SQL expression language, object-relational mapping, as well as its database transaction control and the unique challenges that we have faced as a consequence of relying too much on framework "magic".

Finally, given Galaxy's state of active ongoing development (70,000+ commits over the past decade), the use of a database migration tool (that automates the management of changes to a database schema) is an absolute requirement for us. We have recently moved our code base to Alembic, which is the recommended database migration tool for SQLAlchemy. Again, we faced unique challenges, both due to the "galactic" scale of our code base, as well as the multitude of Galaxy instances around the world, which had to be provided with a migration path accommodating a wide variety of configurations and upgrade scenarios.

CAPTURING AND FUNCTIONAL ANNOTATING THE IMMUNOGLOBULIN LOCI OF VARIOUS SPECIES WITH THIRD-GENERATION SEQUENCING

Dori Z Deng[1], Will Seligmann[2], Helen M Dooley[3], Richard E Green[2], Russ Corbett-Detig[2], Christopher Vollmers[2]

[1]University of California Santa Cruz, Department of Molecular, Cellular, and Developmental Biology, Santa Cruz, CA, [2]University of California Santa Cruz, Department of Biomolecular Engineering, Santa Cruz, CA, [3]University of Maryland, Department of Microbiology and Immunology, College Park, MD

All animal species, from sponges to mammals, have various mechanisms to protect themselves from infection, collectively called immunity. Over the course of evolution, immune mechanisms have evolved toward the same goal, yet, paradoxically, they have become very different. Vertebrates are the only groups of species that have developed one of the most complex immune mechanisms, the adaptive immune system. By encoding a vast repertoire of antibodies with repeated and highly homologous sets of immunoglobulin genes, the adaptive immune system can recognize an extensive array of non-self molecules with exquisite specificity. Evolution has shaped the highly repetitive immunoglobulin loci with many insertion, duplication, and deletion events. Our understanding of the immunoglobulin loci and the antibody diversification process is limited due to the inability of sequencing technologies to capture both genomic loci and the transcripts they ultimately encode accurately.

Taking advantage of the recent advancements in long-read sequencing technology, we have developed a primer-independent and, therefore, unbiased full-length immunoglobulin heavy chain transcripts sequencing method with 99.9% per-base accuracy. We used this method to generate the immunoglobulin heavy chain transcript repertoires for two individuals from eight species (Rat, Rabbit, Chicken, Rhesus Monkey, Pig, Cow, and Nurse Shark). To complement the immunoglobulin transcripts, we also assembled high-quality, chromosome-scale genomes from the same individuals using sequencing reads from Oxford Nanopore Technologies, Pacific Biosciences, and Illumina (in the form of Omni-C). We then combined transcript repertoires and genome assemblies to functionally annotate the immunoglobulin heavy chain loci of these species. With these species' comprehensive annotated immunoglobulin loci, we have conducted a comprehensive analysis to elucidate the evolutionary events that have shaped the adaptive immune system.

COMPARISON OF 16S rRNA GENE SEQUENCING AND SHOTGUN METAGENOMIC SEQUENCING FOR RUMEN MICROBIOME ANALYSIS

Gerardo R Diaz Ortiz, Noelle N Noyes

University of Minnesota, Department of Veterinary Population Medicine, Saint Paul, MN

Microbiome studies rely heavily on the sequencing approach used. Currently, 16S rRNA sequencing (16S) and shotgun metagenomic sequencing (SMS) are the most used approaches. Despite trade-offs between each of these sequencing approaches, they have not been formally compared for taxonomic profiling of samples obtained from cattle rumen. We aimed to compare 16S rRNA sequencing hypervariable region V4 (16S-V4) and shotgun metagenomic sequencing at the lowest (SMS cs0) and highest (SMS cs1) classification confidence for rumen microbiome analysis. We used beef cattle rumen DNA samples collected in a previous longitudinal study that identified changes in the rumen microbiome composition associated with collection day and weaning strategy. These samples were subjected simultaneously to both sequencing methods. 16S-V4 sequences were processed using dada2, and the SILVA nr99 v138.1 database, producing 1 count matrix. Shotgun metagenomic sequences were processed using the minor_kraken2.nf module of AMR++ v2 and NCBI RefSeq database, setting 2 confidence scores for taxonomic classification: cs0 (lowest confidence) and cs1 (highest confidence), producing 2 count matrices. The taxonomic profiling performance, classification agreement, and microbiome diversity inferences were compared between the 3 count matrices. From 403 genera identified by 16S-V4, 148 were also identified by SMS-cs0 and SMS-cs1, while 253 were exclusively identified by 16S-V4. Sub-setting the 148 genera identified by all 3 methods, we identified a high Pearson’s correlation between genera abundance identified by SMS-cs0 and SMS-cs1 (average R2= 0.94, $P < 0.05$), and a lower correlation of genera abundance between 16S-V4 and SMS cs1 (R2= 0.85, $P < 0.05$). Despite different estimates for alpha and beta diversity indices produced by the 3 methods, all of them inferred the same significant association of microbiome diversity changes with collection day and weaning strategy, consistently with the previous study. When the correlation between diversity indices was measured, we found that genus-level Shannon’s indices and dissimilarity matrices of SMS-cs0 and SMS-cs1 methods had the highest correlation (Pearson R2= 0.88, $P < 0.001$; Procrustes m2=0.17, $P<0.05$). Although we observed discrepancies in taxonomic detection, resolution, and classification, all methods led to similar ecological inferences. The inherent technical nature of each sequencing method and inconsistent databases explained the taxonomic discrepancies, while a high correlation between genus abundance and a moderate correlation between diversity indices may explain why ecological inferences were similar despite these technical discrepancies.

A METAGENOMIC INVESTIGATION OF THE WORK-RELATED MICROBIOME IN US SWINE WORKERS

Gerardo R Diaz Ortiz, Ilya B Slizovskiy, Montserrat Torremorell, Noelle R Noyes

University of Minnesota, Department of Veterinary Population Medicine, Saint Paul, MN

The human microbiome, influenced by genetics and environmental factors, plays a pivotal role in health and disease. Occupational environments, such as workplaces with animal exposure, significantly impact workers' microbiomes. In particular, individuals working with swine exhibit distinct oral and nasal microbiomes, but the reasons and consequences for these differences remain unclear. Moreover, the efficacy of interventions to mitigate biological hazards in swine workers may hinge on their microbiome profiles, necessitating a deeper understanding of them. This study will explore the unique microbiomes of U.S. swine workers as a base to develop targeted health interventions in the U.S. swine industry.

Firstly, we will quantify the contribution of on-farm microbiomes (swine, high-contact surfaces, aerosols) to swine workers' microbiomes assigned to different job tasks. To do this, we will perform 16S rRNA sequencing of the skin and nasal microbiomes of 80 swine workers, and related swine and environmental sources in 10 different farms in Minnesota. A source tracker analysis will be conducted to determine the contribution of different on-farm sources to the worker’s microbiome.

Secondly, we will assess the impact of swine contact and farm biosecurity practices, like use of personal protective equipment and showering, on skin and nasal microbiomes. Our preliminary data suggest direct swine contact significantly influences workers skin microbiome. We will recruit 40 swine workers from farms with varying biosecurity practices and monitor their daily swine contact and microbiome composition before and after swine contact over 7 days.

Thirdly, we will investigate metagenomic signatures associated with the swine workplace, enabling a comprehensive comparison of U.S. swine worker microbiomes with non-swine workers. We will perform shotgun metagenomic sequencing on skin and nasal samples collected from 80 swine workers and compare the data with healthy U.S. adults from the Human Microbiome Project. Bioinformatic analysis will involve identifying genes related to antimicrobial resistance, virulence factors, and strain-level dynamics.

Upon successful completion, this research will identify the distinctive features of U.S. swine worker microbiomes, how on-farm factors shape them, and whether these differences correlate with disease risks. This knowledge will support targeted workplace interventions to enhance the health and safety of U.S. swine workers and facilitate further investigations into the health implications of these microbiome differences.

PHYLOGENETIC DIVERSITY PATTERNS AMONG GASTROINTESTINAL BACTERIAL STRAINS.

Veronika Dubinkina*

The Gladstone Institute of Data Science and Biotechnology, San Francisco, CA, [2]University of California, San Francisco, Department of Epidemiology and Biostatistics, San Francisco, CA, [3]Chan-Zuckerberg Biohub, San Francisco, CA

Along with traditional isolation and sequencing, metagenomic assembly has dramatically increased the coverage of bacterial taxa in sequence databases, and the vast majority of prevalent species in the human gut microbiome now have representative genomes. However, many efforts are now focused on capturing intra-species diversity, revealing that even strains within the same species can exhibit up to a 25% difference in their gene content. These microvariations in species genomes can result in vastly different phenotypes, which is critical to consider in ecological and epidemiological studies. Moreover, metagenomic studies of environmental samples are now relying on these extended databases to capture this micro-diversity and its consequences.

Nevertheless, one big challenge we face while exploring strain diversity in complex communities is strain phasing, or our ability to differentiate which genes are present in each strain. We aim to quantify the extent to which Single Nucleotide Polymorphisms (SNPs) predict gene presence/absence across species and genes. To do that, we explore patterns of intra-specific genomic variability across a recent collection of gut species, the Universal Human Gut Genome (UHGG) database. On the intra-species level, high rates of homologous recombination and horizontal gene transfer will lead to variable degrees of linkage between SNP alleles and genes. We show that accessory genes exhibit different levels of correlation with species population structure inferred from core genome SNPs within individual species. Notably, we observe that uncorrelated genes are enriched in mobile genetic elements, rendering their presence more challenging to predict. We hypothesize that due to the mechanisms of intra-species genome evolution, their presence/absence should be tightly linked with core-genome SNPs in their vicinity. These findings will enable us to leverage SNP profiling to infer most of the gene content of specific strains. Our study provides a basis to aid in the strain phasing of species, paving the way for more precise strain-level analysis of the gut microbiome data.

DETECTING DIFFERENTIAL TRANSCRIPT USAGE IN COMPLEX DISEASES WITH SPIT

Beril Erdogdu[1,2], Ales Varabyou[1,3], Stephanie C Hicks[1,4,5], Steven L Salzberg[1,2,3,4], Mihaela Pertea[1,2,3]

[1]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [2]Johns Hopkins School of Medicine and Whiting School of Engineering, Department of Biomedical Engineering, Baltimore, MD, [3]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [4]Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD, [5]Johns Hopkins University, Malone Center for Engineering in Healthcare, Baltimore, MD

Differential transcript usage (DTU) plays a crucial role in determining how gene expression differs among cells, tissues, and different developmental stages, thereby contributing to the complexity and diversity of biological systems. In abnormal cells, it can also lead to deficiencies in protein function and underpin disease pathogenesis. Analyzing DTU via RNA-Seq data is vital, but the genetic heterogeneity in populations with complex diseases presents an intricate challenge due to diverse causal events and undetermined subtypes. Although the majority of common diseases in humans are categorized as complex, the state-of-the-art DTU analysis methods overlook this heterogeneity in their models. SPIT is the first statistical tool that identifies predominant subgroups in transcript usage within a population along with their distinctive sets of DTU events. This study provides comprehensive assessments of SPIT's methodology and applies it to analyze brain samples from individuals with Schizophrenia, revealing previously unreported DTU events in six candidate genes.

INVESTIGATING THE ORIGINS AND IMPACTS OF STRUCTURAL VARIATION AND DNA METHYLATION IN HIGH-GRADE SEROUS OVARIAN CANCER

Edward Esiri-Bloom[1], Stuart Aitken[1], Graeme Grimes[1], Ailith Ewing[1], Alison Meynert[1], Ryan Silk[1], Stuart Brown[1], Michael Churchman[2], C Simon Herrington[3], Patricia Roxburgh[4,5], Charlie Gourley[2], Colin A Semple[1]

[1]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, [2]Nicola Murray Centre for Ovarian Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, [3]Edinburgh Pathology, Cancer Research UK Edinburgh Centre, MRC IGMM, University of Edinburgh, Edinburgh, United Kingdom, [4]Institute of Cancer Sciences, Wolfson Wohl Cancer Research Centre, University of Glasgow, Glasgow, United Kingdom, [5]Beatson West of Scotland Cancer Centre, University of Glasgow, Glasgow, United Kingdom

With a five-year survival rate of less than 50%, high-grade serous ovarian cancer (HGSOC) is one of the deadliest gynaecological cancers and one of the most structurally diverse tumour types. This tumour type is characterised by high levels of genomic rearrangement involving large and complex structural variation (SV), including catastrophic events such as chromothripsis. Such variation may be key to understanding tumorigenesis and drug resistance in this tumour. To address this, we have generated high-coverage (~30x) ONT long-read whole-genome sequencing (LR-WGS) data from 13 locally collected HGSOC tumours. We are combining SV calls in these samples with variant calls from existing deep short-read WGS and optical mapping data for the same tumours, providing orthogonal validation and accurate characterisation of complex structural rearrangements. Preliminary results from Sniffles2 calls show that our long-reads recall ~70% of SVs in our high-confidence short-read consensus call-set with a 50% reciprocal overlap and, importantly, contain many well-supported novel SVs not detected by short-read WGS. Methylation calls and matched RNA-seq for these tumours provide the opportunity for integrative analyses of genomic, transcriptomic, and epigenomic data to determine the functional impacts of structural complexity. Analysis of methylation array data from another HGSOC cohort (n=79) has revealed distinct clusters of samples which are enriched for Homologous Recombination Repair Deficiency ($p \leq 0.01$), which will be interrogated in our own dataset. Overall, we aim to accurately describe the patterns of known and novel complex SVs in HGSOC, shedding new light on their origins, their functional consequences, and ultimately their roles in tumour evolution.

STRAIN-RESOLUTION LONG READ METAGENOME ASSEMBLY

Xiaowen Feng[1,2], Heng Li[1,2]

[1]Dana-Farber Cancer Institute, Data Science, Boston, MA, [2]Harvard Medical School, DBMI, Boston, MA

In real microbial communities, bacteria and archea species are often distinctive enough to be separated or even fully assembled de novo using highly accurate long reads. We saw two problems following this. First, abundant species are often believed to assemble well given their deeper coverage. This conjuncture is rarely tested or evaluated in practice. We often do not know how many abundant species we are missing and do not have an approach to recover them. Second, limited by current technology, long reads are not long enough to fully resolve strains. Medium and low abundance species tend to have complicated or entangled unitig graphs. Is it possible to reliably preserve sub-species information?

To address the first issue, we proposed k-mer based and 16S RNA based methods to measure the completeness of metagenome assembly. We showed that even with PacBio High-Fidelity (HiFi) reads, abundant species are often not assembled as high strain diversity may lead to fragmented contigs. Regarding the second issue, we further develop hifiasm-meta. We notice that the string graph is closely related to interval graphs when phasing information is not considered, but otherwise needs to be treated as partial-ordered set (poset), which can be obtained by converting the bidirectional string graph to an undirected graph and applying orientations. We identify optimal walks of unitigs based Dilworth's theorem using weighted maximum cardinality matching algorithms, which is followed by graph cleaning, generation of strains and purging of redundant contigs.

Our work stresses the importance of metagenome completeness which is often overlooked before, and moves towards improving strain resolution in string graph-based de novo assembly.

SENSITIVE AND SPECIFIC DETECTION OF MOSAIC CHROMOSOMAL ALTERATIONS FROM LARGE-SCALE RNA-SEQ DATASETS

Teng Gao[1], Maria E Kastriti[2,3], Viktor Ljungström[1], Andreas Heinzel[4], Arthur S Tischler[5], Rainer Oberbauer[4], Po-Ru Loh[6,7], Igor Adameyko[2,3], Peter J Park[1], Peter V Kharchenko[1,8]

[1]Harvard Medical School, Department of Biomedical Informatics, Boston, MA, [2]Medical University of Vienna, Department of Neuroimmunology, Center for Brain Research, Vienna, Austria, [3]Karolinska Institutet, Department of Physiology and Pharmacology, Solna, Sweden, [4]Medical University of Vienna, Department of Nephrology, Internal Medicine III, Vienna, Austria, [5]Tufts Medical Center, Department of Pathology and Laboratory Medicine, Boston, MA, [6]Brigham and Women's Hospital and Harvard Medical School, Division of Genetics, Department of Medicine, Boston, MA, [7]Broad Institute of MIT and Harvard, Cambridge, MA, [8]Altos Labs, San Diego Institute of Science, San Diego, CA

Characterizing the multitude of mosaic mutations inside the human body is a new frontier of human genetics. While recent studies have revealed the landscape of somatic single-nucleotide variants across normal tissues, the repertoire of large mosaic chromosomal alterations (mCAs) remains largely unknown. Here, we present a computational method, HaHMMR, that augments RNA-based mCA detection with population-based haplotype phasing. Validation using paired RNA-seq and WGS data shows that HaHMMR can detect subclonal mCAs (cell fraction as low as 10%) with high sensitivity and low false positive rate (<0.0001 per autosome). Applying HaHMMR to 16,672 RNA-seq samples in GTEx, we identified hundreds of mCAs in more than 20 tissue types of the human body. We found that over a quarter of the subjects carry a clonally-expanded mCA in at least one tissue, with incidence strongly correlated with age. The prevalence and genome-wide patterns of mCAs vary considerably across tissue types, suggesting tissue-specific mutagenic exposure and selection pressures. The mCA landscapes in normal adrenal and pituitary glands resemble those in tumors arising from these tissues, whereas the same is not true for the esophagus and skin. Together, our findings demonstrate a widespread, age-dependent emergence of mCAs across normal human tissues with intricate connections to tumorigenesis.

RE-ANALYSIS OF MICROBIAL CONTENT FOUND IN TUMORS SEQUENCED BY THE CANCER GENOME ATLAS PROJECT

Peter Ge[1,2], Jennifer Lu[1,2], Daniela Puiu[1,2], Mahler Revsine[1,3], Amanda Xu[1,3], Mihaela Pertea[1,2,3], Steven L Salzberg[1,2,3,4]

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

In recent years a growing number of publications have reported that microbiomes in human tumors are specific to different cancer types [1,2,3]. We have re-analyzed the results in one of those studies, Poore et al. 2020 [1], and discovered two fundamental flaws: (1) read counts identified as microbial were greatly inflated, primarily due to human reads false identified as microbial; and (2) artificial signatures linking read counts to cancer types were introduced during normalization of the raw data [4]. Our original work was based on re-analysis of 3 cancer types from the original study, which we have now expanded to include all 25 cancer types for which whole-genome sequencing (WGS) data is available from The Cancer Genome Atlas (TCGA) project. In total, we downloaded the unmapped reads from 5666 WGS and 79 RNA-seq samples that had been aligned to the GRCh37 or GRCh38 human reference genomes and then realigned them to the CHM13 human reference genome. We then took these two-pass filtered reads and used KrakenUniq [5] to match them against a carefully curated database that included all complete genomes of bacteria, archaea, and viruses from RefSeq, as well as the GRCh38 human genome. Furthermore, we extracted the unclassified reads from the previous step and re-ran KrakenUniq using a database built from all complete fungi genomes from GenBank. We are releasing this high-quality dataset as carefully screened set of microbial reads that were found in the original TCGA samples.

References:
[1] Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., ... & Knight, R. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature, 579(7800), 567-574.
[2] Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., ... & Straussman, R. (2020). The human tumor microbiome is composed of tumor type–specific intracellular bacteria. Science, 368(6494), 973-980.
[3] Narunsky-Haziza, L., Sepich-Poore, G. D., Livyatan, I., Asraf, O., Martino, C., Nejman, D., ... & Straussman, R. (2022). Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. Cell, 185(20), 3789-3806.
[4] Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C. S., ... & Salzberg, S. (2023). Major data analysis errors invalidate cancer microbiome findings. mBio, in press.
[5] Breitwieser, F. P., Baker, D. N., & Salzberg, S. L. (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome biology, 19(1), 1-10.

INTEGRATIVE COMPUTATIONAL FRAMEWORK, *DYSCOVR*, LINKS MUTATED DRIVER GENES TO METABOLIC DYSREGULATION ACROSS 22 CANCER TYPES

Sara E Geraghty[1], Jacob Boyer[1,3], Matthew McBride[1,3,4], Joshua Rabinowitz[1,3], Mona Singh[1,2]

[1]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [2]Princeton University, Computer Science, Princeton, NJ, [3]Princeton University, Chemistry, Princeton, NJ, [4]Rutgers Cancer Institute of New Jersey, Chemical Biology, New Brunswick, NJ

Though somatic mutations play a critical role in driving cancer initiation and progression, the functional impacts of these mutations—particularly, how they alter expression patterns across the genome and within hallmark pathways like cancer metabolism—are not yet well-understood, even for mutations in well-studied cancer driver genes like *KRAS* or *PIK3CA*. Given that cancer therapies are increasingly designed to target commonly mutated driver genes, a nuanced understanding of the molecular mechanisms at play is important in the development of truly personalized and effective cancer treatment plans. From this premise, we designed an integrative machine learning model that draws on high-throughput sequencing data from 22 cancer types in the Cancer Genome Atlas to uncover putative relationships between nonsynonymous mutations in key cancer driver genes and downstream transcriptional dysregulation across the genome. This framework, *Dyscovr*, is unique in that it combines patient mutation, copy number alteration, and methylation information with various other molecular and clinical features, such germline variation, tumor mutational burden, and immune cell infiltration, to disentangle the molecular mechanisms underlying transcriptional dysregulation. We applied *Dyscovr* pan-cancer and within individual cancer types, producing thousands of both broadly relevant and cancer type-specific links. These links are specific to known driver genes—as no links are uncovered for highly mutated passenger genes—and are enriched in cancer-relevant genes and pathways. For drivers with gold standard sets of known transcriptional targets, such as *TP53*, we find that *Dyscovr* effectively prioritizes these relationships, lending credibility to its hundreds of novel associations. As a case study, we used *Dyscovr* to uncover effects of nonsynonymous mutations in driver genes on metabolic targets, a class of targets with particular therapeutic relevance. We used a suite of techniques to validate our model's predictions that nonsynonymous mutations in drivers such as *KRAS* upregulate key genes involved in targetable pathways such as nucleotide and amino acid metabolism. Altogether, this work suggests that *Dyscovr* is a clinically-relevant tool that sheds light on how—on a granular, functional level—driver mutations hijack regulatory pathways to enable cancer hallmarks.

CHARACTERIZING HOST-PATHOGEN INTERACTION DYNAMICS FOR *TOXOPLASMA GONDII* WITH SINGLE-CELL RNA SEQUENCING: A PILOT STUDY

Yomna Gohar[1], Veronica Raba[1], Tobias Lautwein[2], Daniel Wind[1], Lisanna Hülse[1], karin Buchholz[1], Raba Katharina[3], Daniel Degrandi[1], klaus Pfeffer[1], Alexander Dilthey[1]

[1]Heinrich Heine University Düsseldorf, Institute of Medical Microbiology and Hospital Hygiene, Düsseldorf, Germany, [2]Heinrich Heine University Düsseldorf, Biologisch-Medizinisches-Forschungszentrum (BMFZ), Genomics & Transcriptomics Laboratory, Düsseldorf, Germany, [3]Heinrich Heine University Düsseldorf, Core Facility Flow Cytometry, Düsseldorf, Germany

Toxoplasma gondii infects most warm-blooded animals and nucleated cells. While often asymptomatic, it poses a risk for immunocompromised patients. The mechanisms by which *T. gondii* manipulates and evades host immune response are incompletely understood. Single-cell RNA sequencing (scRNA-seq) may enable improved characterization of host-parasite interactions and immune evasion strategies at the level of individual cells; however, it is not clear whether current scRNA-seq protocols enable the joint characterization of the host and parasite transcriptomes.

We thus carried out a pilot experiment to investigate the applicability of the Smart-seq3 technology to the characterization of host-pathogen interactions. Hs27 cells were exposed to *T. gondii* ME49 at a Multiplicity of Infection (MOI) of 10 and fixed in methanol using a protocol adapted from Chen et al. (2018) at 2 hours post-infection (pi); a control cell population was exposed to the same experimental conditions and similarly fixed 2 hours pi. scRNA-seq data were generated for 52 cells from the *T. gondii*-exposed population and 31 cells from the controls; read data was analyzed using Seurat and a combined host-parasite reference.

The average number of sequencing reads for the *T. gondii*-exposed was 1,273,775, enabling the detection of 7602 host and 406.6 parasite genes on average. The proportion of *T. gondii*-derived reads ranged from 0.02% to 7.5% per cell, suggesting heterogeneity in the number of infecting parasites, which was confirmed using immunofluorescent staining. Most SRS genes, which are crucial for the parasite's ability to invade host cells, were not detected in the *T. gondii* exposed cells, with the exception of SAG1, which was detected in most cells. Cluster analysis from the experiment highlighted diverse transcriptional responses of host cells to *T. gondii* infection. Control cells formed one distinct cluster, while infected cells were divided into two separate clusters. Visual inspection revealed that clustering did not correlate with the proportion of *T. gondii*-derived reads, suggesting that the host cells' transcriptional response to *T. gondii* might not be directly tied to infection load. The clustering of both highly infected and minimally infected cells hints at a potential cell signaling process.

In conclusion, Smart-seq3 proved effective in analyzing host and parasite transcriptomes. However, our study also identified important challenges, such as heterogeneity in the number of infecting parasites. Our future work will involve a time-course experiment conducted at 0.5, 2, 8, 24, and 48 hours pi to track transcriptional changes across various infection stages. Additionally, we will explore the role of IFN-γ in the immune response against *T. gondii*, given its importance in anti-parasitic mechanisms.

MULTI-SAMPLE NANOPORE SEQUENCING PROVIDES INSIGHTS INTO MELANOMA HETEROGENEITY AND EVOLUTION

Anton Goretsky[1,2], Yuelin Liu[1,2], Ayse Keskus[1], Salem Malikic[1], Glenn Merlino[3], Chi-Ping Day[3], Erin Molloy[2], S. Cenk Sahinalp[1], Mikhail Kolmogorov[1]

[1]National Cancer Institute, Center for Cancer Research, Cancer Data Science Laboratory, Bethesda, MD, [2]University of Maryland, Department of Computer Science, College Park, MD, [3]National Cancer Institute, Center for Cancer Research, Laboratory of Cancer Biology and Genetics, Bethesda, MD

Melanoma is the most serious form of skin cancer, developed by the malignant evolution of melanocytes. Malignant melanoma incidence is increasing faster than most other cancers. While stage zero melanoma is highly treatable, in its advanced stages survivability dramatically decreases. Melanoma has shown to be one of the most heterogeneous cancers from RNA and exome analyses by The Cancer Genome Atlas and other groups. A better understanding of the key genomic and epigenomic events that characterize the diverse subclonal populations in melanoma may reveal key insight into what drives its progression and therapeutic resistance.

In this study, we leverage Nanopore long read sequencing to study the evolution of the mouse B2905 melanoma cell line. Twenty-four distinct clonal sublines were derived in vitro from single cells of the cell line, and the genetically homogenous population from each subline was sequenced using PromethION R10 flowcells. Enabled by the possibility to perform haplotype phasing and directly call 5mC base modifications with such data, our goal is to to integrate small variants, structural variants and detected CpG methylation to better our understanding of melanoma evolution, and build upon prior analyses of short read sequenced sublines.

We first compare various SNV calling approaches, such as DeepVariant and Clair; and develop a tool for multi-sample somatic SV calling called Severus. Our examination of structural variation and small variants has revealed significant numbers of somatic mutations, from translocations to various indels and single nucleotide variants, and potential losses of heterozygosity. Several chromosomes showed significantly higher proportions of variants than others, such as chr13 was enriched with somatic SNVs and SV in neurotransmitter- and GPCR-related genes.

Our preliminary results show that the phylogeny constructed using the CpG methylation profiles derived from long read data corroborates prior phylogenies built using small variants found in bulk exome / transcriptome short read data. Phylogenetic trees constructed solely from genomic variants (SNVs and SVs) offer mixed corroborative results, likely due to extensive aneuploidy and heterozygosity. As such, we take an evaluative approach to phylogenetic tree construction and variant calling and evaluate various parsimony, distance, and likelihood models, such as Dollo parsimony. By incorporating structural variations and small variant data in the phylogenetic construction, our analysis offers a better characterization of the epigenetic mechanism of subclonal evolution in melanoma. These results inform an on-going study about subclone-specific responses to immune checkpoint blockade therapy on a preclinical melanoma model.

LANDSCAPE OF DIFFERENTIATION INDUCED ONCOGENESIS REGULATED BY PSEUDOGENES: A NEURAL NETWORK BASED STUDY OF GASTROINTESTINAL TRACT

Pravallika Govada, Rajasekaran Ramalingam

Vellore Institute of Technology, Department of Integrative Biology,
Vellore, India

Pseudogenes form a subset of non-coding RNAs that are well-studied for their role in regulating the development of organism and cellular differentiation. However, their context-dependent role as promoters of differentiation during cancer initiation, progression and metastasis remains largely elusive. Furthermore, while pseudogenes exhibit tissue-specific expression pattern, they are seldom explored as a viable option in drug discovery to potentially reduce non-cancerous tissue damage. Since aberrant cell differentiation due to prolonged tissue damage leads to metaplastic epithelium with diverse cell signatures; often a precursor of tumorigenesis within the gastro-intestinal tract (GI), we utilized GI tract tumors as a model to study the role of pseudogenes in regulating the events of differentiation to promote oncogenic transformation. In accordance with our hypothesis, convolutional neural networks (CNN) based analysis of esophageal, gastric, colon and rectal carcinomas indicate tissue-specific expression pattern with distinct stage-wise expression of pseudogenes as well as their interacting partners. Furthermore, tissue-specific homeostatic expression of pseudogenes prompts their effective use as valid diagnostic biomarkers to identify distinct cancer types and their respective stages. Previously, to quantitatively characterize the extent of differentiation independent of histopathological tumor grading, we defined three unique metrics which along with gene regulatory network and topology analysis allowed us to identify the relationship between the extent of differentiation and pseudogene expression pattern. In fact, we report combinatorial de-regulation of a subset of differentially expressed pseudogenes as a potential driving factor of aberrant differentiation giving rise to metaplastic epithelium, especially for patient stratified Stage II esophageal carcinoma. In addition to the combinatorial de-regulation of pseudogenes, transcription factor analysis helped us reveal SOX2 as a key regulatory factor upstream as well as potentially downstream of pseudogenes that aids in differentiation induced oncogenic transformation. We additionally identified FEV, PRRX1 and TFAP2A as differentiation-associated transcription factors promoting tumorigenesis. Finally, CNN-based survival analysis of the GI tract tumors implicates distinct pseudogenes as valuable prognostic markers, similar to our previous reports of ARSDP1 and GYG2P1 as promoters of poor prognosis across a distinct landscape of pseudogene expression associated with differentiation within esophageal carcinoma.

SPEEDING UP WHOLE GENOME ALIGNMENT BY INCREASING HARDWARE UTILIZATION

A. Burak Gulhan[1], Mahmut Kandemir[1], Maximilian Haeussler[2], Anton Nekrutrenko[3]

[1]Penn State, Dept. of Computer Science, State College, PA, [2]UC Santa Cruz, Genomics Institute, Santa Cruz, CA, [3]Penn State, Huck Institutes for Life Sciences, State College, PA

The number of fully sequenced genomes grows rapidly. However, a sequenced genome is just a big text file with little utility on its own. So why are so many genomes being sequenced? Because of the need to understand the functional underpinnings of global biodiversity and use this for basic and applied research. Acquiring such an understanding involves comparing genomes to identify functional and evolving regions, which requires alignments at a massive scale. Recent benchmarks show that mammalian whole-genome alignment (WGA) takes ~800 CPU-hours. In light of thousands of sequences in the public genome archives, current WGA software stacks need improvements, as millions of CPU hours will be needed otherwise. One major component and bottleneck of such pipelines is the WGA tool lastZ, which is based on a ‘Seed, Filter & Extend’ paradigm. LastZ is used in several packages, such as Cactus and the UCSC multiZ pipeline, which require at least n pairwise alignments for n genomes. GPUs can dramatically reduce alignment costs. A replacement for lastZ, called SegAlign, uses GPU to accelerate the ‘seed’ and ‘filter’ stages, improving performance by a factor of 5-10x.

We found that, in SegAlign, there exists a significant issue of hardware underutilization. This frequently leads to cases where all hardware resources, both CPU and GPU, wait for a few CPU threads to complete their work. In extreme cases, a single CPU thread can work for over 66% of the total runtime while all other hardware resources stay idle. This problem cannot be fixed with more hardware resources; in fact, doing so exacerbates the problem by resulting in more idling resources. We also observed that underutilization is input dependent; less divergent sequences have more idling, due to higher CPU workloads.

We address this problem using two methods: 1) Restricting the maximum amount of work sent to a single CPU thread, that is, to the ‘extend’ stage of SegAlign which uses lastZ. This increases utilization by splitting a previously large unit of work among several CPU threads. 2) Using Multi-Instance GPU (MIG) and Multi-Process Service (MPS) to allow multiple workloads to run, in parallel, within a single GPU. For this method, we split our genome inputs into blocks and run each pair with SegAlign in parallel using MIG+MPS, leading to higher CPU and GPU utilization in addition to hiding overheads – cases where GPU or CPU becomes idle. Using these two methods, we run tests on an A100 GPU with 32 cores and obtain up to 3x speedup, where in the best case we reduce runtime from over 36.5 hours to 11 hours when aligning human (hs1) and chimpanzee (panTro6) genomes.

EVALUATION OF HAPLOTYPE-AWARE LONG-READ ERROR CORRECTION WITH HIFIEVAL

Yujie Guo[1,2], Xiaowen Feng[1,2], Heng Li[1,2]

[1]Dana-Farber Cancer Institute, Data Science, Boston, MA, [2]Harvard Medical School, Biomedical Informatics, Boston, MA

The PacBio High-Fidelity (HiFi) sequencing technology produces long reads of 99% in accuracy. It has enabled the development of a new generation of *de novo* sequence assemblers, which all have sequencing error correction as the first step. As HiFi is a new data type, this critical step has not been evaluated before. Here, we introduced hifieval, a new command-line tool for measuring over- and under-corrections produced by error correction algorithms. We assessed the accuracy of the error correction components of existing HiFi assemblers on the CHM13 and the HG002 datasets and further investigated the performance of error correction methods in challenging regions such as homopolymer regions, centromeric regions, and segmental duplications. Hifieval will help HiFi assemblers to improve error correction and assembly quality in the long run.

FOR THE DEVELOPMENT OF A PCR PRIMER DESIGN PIPELINE FOR DETECTION OF CONTAMINATION IN FOODS

Yoritaka Harazono[1], Keisuke Soga[2], Masahiro Kasahara[1]

[1]the University of Tokyo, Department of Computational Biology and Medical Sciences, Kahiwa-shi, Japan, [2]National Institute of Health Sciences, Division of Biochemistry, Kawasaki-shi, Japan

Even a tiny amount of allergens in food can be life-threatening. However, identification of such allergens in processed foods may not be easy. While PCR is generally used for high-sensitivity detection of such allergens, we need to consider three issues: (1)DNA molecules of allergens can be fragmented during food processing, (2)allergens may be present in only minute proportions, and (3)a PCR primer pair that specifically amplifies the genome of target allergens may not be available. To these ends, we are developing a pipeline to design primers for the ultra-sensitive detection of allergens in processed foods. We set the following three criteria for the region amplified by the primer: (1)a primer pair amplifies as many regions as possible in a given target allergen, (2) the length of the amplified region is smaller than 200 bp, and (3)the primer pair does not amplify genomes of other known species. The issue with (1) and (2) is that possible amplicons may not be of the same length, and therefore we needed a method to find regions of variable lengths in the target genome such that both ends of all of the regions match the primer pair. To this end, we developed a new counting approach, *LR-tuple* counting, that counts the number of such regions of smaller than a given threshold (200 bp by default), in a given target genome. The LR-tuple counting method segments a sliding nucleotide sequence window of variable lengths into left, middle, and right parts, allowing sequence fluctuation in the middle part that corresponds to PCR amplicons. To meet (3), the primers must not bind to the genome sequences of any known species except for the genome of the target species. Considering the relatively small genome size and the availability of high-quality NGS reads, we began with designing highly sensitive primers to detect rice. We implemented *swordfish*, a tool to find frequently occurring sequences in a given genome. Swordfish enumerates LR-tuples that appear more than a specified threshold. We used Primer3 to design primers on the frequent LR-tuples. Then, we performed alignments against all known genome sequences in the NCBI nt database using BLAST. Primers which were aligned with other species were excluded from the final primer candidates. The designed primers could amplify rice with a high sensitivity, but some false positives (cross reaction to other species) were also amplified. Such false positives were found to be a part of LTR retrotransposons, namely, Gypsy and Copia. We discuss how to avoid the effects of such cross reactions.

TOWARD TELOMERE-TO-TELOMERE FELID GENOMES

Andrew J Harris[1,2], Leslie A Lyons[3], Wesley C Warren[4], Kendra Hoekzema[5], Evan E Eichler[5,6], William J Murphy[1,2]

[1]Texas A&M University, Veterinary Integrative Biosciences, College Station, TX, [2]Texas A&M University, Interdisciplinary Program in Genetics & Genomics, College Station, TX, [3]University of Missouri, Department of Veterinary Medicine & Surgery, Columbia, MO, [4]University of Missouri, Department of Animal Sciences, Columbia, MO, [5]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [6]University of Washington School of Medicine, Howard Hughes Medical Institute, Seattle, WA

Recent advancements in genome assembly algorithms and long-read sequencing technologies have ushered in the era of telomere-to-telomere (T2T) genomes. With Oxford Nanopore “ultra-long” (>100kb) sequenced reads, we can now assemble through previously inaccessible regions of the genome like large centromeres and macrosatellites. With the successful completion of the human genome, researchers can finally take the lessons learned from the human T2T project and implement them throughout the Tree of Life, expanding the catalog of T2T genome assemblies. Here, we report progress toward T2T genome assemblies for several cat species. We present results and analyses from near T2T assemblies for the domestic cat and Geoffroy’s cat using a Safari cat F1-interspecific hybrid of the two species. We generated PacBio HiFi (50x) and Oxford Nanopore “ultra-long” (47x) sequenced reads and obtained short-read Illumina libraries from representatives of each parental species for haplotype phasing. We assembled the two parental haplotypes using Verkko in trio mode, producing haplotype phased genome assemblies with N50s of 148Mb and 153Mb for the domestic cat and Geoffroy’s cat, respectfully. Verkko assembled 16/19 chromosomes from the domestic cat and 14/18 chromosomes from the Geoffroy’s cat into a single sequence, with 11 domestic cat and 4 Geoffroy’s cat chromosomes being T2T. Comparisons to the current PacBio continuous long-read (CLR) based assemblies show improvements in both haplotypes, especially within complex regions like centromeres and FA-SAT macrosatellites. We highlight similar results from a domestic cat x African serval F1 hybrid trio and our ongoing progress toward generating T2T genomes across the entire cat family.

TEM-SEQ: AN ULTRASENSITIVE MULTIOMIC PLATFORM FOR EPITOPE-TARGETED DNA METHYLATION MAPPING

Allison Hickman[1], Vishnu U Sunitha Kumary[1], Bryan Venters[1], Jennifer Spengler[1], Anup Vaidya[1], Ryan Ezell[1], Jonathon Burg[1], Zu-wen Sun[1], Martis Cowles[1], Hang Geong Chin[2], Pierre Esteve[2], Chaithanya Ponnaluri[2], Isaac Meek[2], Sriharsa Pradhan[2], Michael-Christopher Keogh[1]

[1]EpiCypher, Inc, Durham, NC, [2]New England BioLabs, Ipswich, MA

Gene expression is regulated by the complex molecular crosstalk between DNA methylation (DNAme) and other chromatin features, such as histone post-translational modifications (PTMs) and chromatin-associated proteins (CAPs). Changes in the chromatin landscape can have a profound impact on DNAme patterning (and vice versa) in both development and disease. However, an understanding of how DNAme co-occurs and co-ordinates with other chromatin features to control gene expression is limited by a lack of reliable genomic tools.

EpiCypher has partnered with New England Biolabs (NEB) to develop **Targeted Enzymatic Methylation-sequencing (TEM-seq)**, an ultrasensitive multiomic genomic mapping technology that delivers high-resolution DNAme profiles at epitope-defined chromatin features. This multiomic workflow integrates EpiCypher’s quantitative high-sensitivity CUTANA CUT&RUN assay for genomic mapping together with NEB’s enzymatic methyl-seq (EM-seq) for unbiased DNAme analysis.

CUT&RUN uses antibodies to locally tether protein A/G-micrococcal nuclease (pAG-MNase) to chromatin in intact cells or nuclei, followed by controlled activation of the MNase to cleave nearby DNA. EM-seq leverages enzymatic conversion of DNAme to generate high resolution, unbiased DNAme profiles with less sample.

Previously, we demonstrated assay specificity, sensitivity, and dynamic range of TEM-seq by mapping two histone PTMs with co-occurring methylation states in human K562 cells. Recently, we successfully performed TEM-seq to map six targets across three cell lines (K562, MCF7, and GM12878), demonstrating the robust nature of this platform. We monitored the TEM-seq reaction using a two-step spike-in system, in which EpiCypher's designer nucleosome spike-ins are used to monitor the specificity and efficiency of the CUT&RUN reaction, and a separate plasmid DNA spike-in control is used to monitor the EM-seq reaction.

We demonstrate that TEM-seq is highly reproducible, specific, and efficient (<10% off-target binding, >90% enzymatic conversion of DNAme, <0.5% conversion of unmethylated DNA). TEM-seq is also incredibly sensitive, requiring only 5-50M sequencing reads per assay, demonstrating the disruptive potential of the assay to offer a low-cost solution for targeted DNAme analysis. We are currently working to validate additional targets (focusing on CAPs) and continuing to develop our controls, including spike-in DNA for 5hmC DNA.

INTERPRETABLE TEXT-BASED MACHINE LEARNING FOR INFERRING SYSTEMATIC TISSUE AND DISEASE ANNOTATIONS OF PUBLIC TRANSCRIPTOME SAMPLES

Parker Hicks[1], Hao Yuan[2,3], Mansooreh Ahmadian[4], Arjun Krishnan[1]

[1]University of Colorado Anschutz Medical Campus, Department of Biomedical Informatics, Aurora, CO, [2]Michigan State University, Genetics and Genome Sciences Program, East Lansing, MI, [3]Michigan State University, Ecology, Evolution and Behavior Program, East Lansing, MI, [4]University of Colorado Anschutz Medical Campus, Department of Biostatistics and Informatics, Aurora, CO

The foremost barrier in reusing existing publicly-available omics data is that all the studies and samples are described using metadata that are in the form of unstructured plain text. The advent of computational language models has opened the door for new approaches to overcome this metadata barrier. Recently, we developed *txt2onto*, a method that uses language modeling and machine learning to classify samples to their tissue and cell type of origin based on their unstructured text descriptions. A key step in *txt2onto* is the conversion of sample descriptions into numerical representations, or embeddings, by averaging all embeddings of all the words in the description. Although this approach seems to preserve useful information in the metadata, the signal from informative and infrequent biomedical words could be dampened when averaging the embeddings of all the words in a sample's description. Further, the average embedding of the whole description cannot be used to interpret the ML classifiers trained on them. Here, we address these issues by constructing feature vectors that preserve word- and phrase-level information, which can then be used to identify biologically meaningful entities in sample metadata that are relevant to a given tissue or disease (mapped to corresponding ontology terms). Using this approach, we improve both the interpretability and performance of *txt2onto*. We report that our novel method outperforms *txt2onto v1* overall for all tissue, cell-type, and disease classification tasks. It also excels in scenarios with limited positive samples in the training dataset. Moreover, our approach enables the identification of words and phrases that are distinctive to specific tissues/diseases, thus providing insights into key snippets of text in sample descriptions that are either biologically meaningful or are just artifacts.

COMPLEASM: A FASTER AND MORE ACCURATE REIMPLEMENTATION OF BUSCO

Neng Huang[1,2], Heng Li[1,2]

[1]Dana-Farber Cancer Institute, Department of Data Sciences, Boston, MA, [2]Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Evaluating the gene completeness is critical to measuring the quality of a genome assembly. An incomplete assembly can lead to errors in gene predictions, annotation, and other downstream analyses. BUSCO is a widely used tool for assessing the completeness of genome assembly by testing the presence of a set of single-copy orthologs conserved across a wide range of taxa. However, BUSCO is slow particularly for large genome assemblies. It is cumbersome to apply BUSCO to a large number of assemblies. Here, we present compleasm, an efficient tool for assessing the completeness of genome assemblies. Compleasm utilizes the miniprot protein-to-genome aligner and the conserved orthologous genes from BUSCO. It is 14 times faster than BUSCO for human assemblies and reports a more accurate completeness of 99.6% than BUSCO's 95.7%, which is in close agreement with the annotation completeness of 99.5% for T2T-CHM13.

COMPUTATIONAL ANALYSIS OF COPY NUMBER VARIATIONS IN SPATIAL TRANSCRIPTOMICS DATA

Rongting Huang[1,3,4], Xianjie Huang[1,2], Ajit J Nirmal[3,4], Yuanhua Huang[1,2,5]

[1]The University of Hong Kong, School of Biomedical Sciences, LKS Faculty of Medicine, Hong Kong, Hong Kong, [2]Hong Kong Science and Technology Park, Center for Translational Stem Cell Biology, Hong Kong, Hong Kong, [3]Harvard Medical School, Laboratory of Systems Pharmacology, Boston, MA, [4]Brigham and women's hospital, Department of Dermatology, Boston, MA, [5]The University of Hong Kong, Department of Statistics and Actuarial Science, Hong Kong, Hong Kong

Somatic copy number variations (CNVs) are major mutations that contribute to the development and progression of various cancers. Spatial transcriptomics (ST) has emerged as a powerful tool for unraveling the intricacies of the tumor microenvironment and the progression of cancer by analyzing gene expression patterns in tissues while preserving their spatial context. Despite a few computational methods proposed to detect CNVs from single-cell transcriptomic data that can be theoretically extended to ST, the technical sparsity of such data makes it challenging to identify allele-specific CNVs, particularly in complex clonal structures. In this study, we leverage our recently developed method XClone (Huang et al., 2023.) to analyze CNV in ST data. Uniquely, this method strengthens the signals of read depth and allelic imbalance by effective smoothing on cell neighborhood and gene coordinate graphs to detect haplotype-aware CNVs from scRNA-seq data. Here, we further extended XClone to embrace spatial information provided by ST aimed at detecting allele-specific copy number variations (CNV) in diverse cancer samples using ST data. We have applied XClone on prostate cancer samples (probed by 10x Visium ST platform) to observe different CNV profiles in different tumor microenvironments. Currently, we are examining the power of integrating spatial transcriptomics and corresponding histology images to achieve higher-resolution CNV analysis.

Our methodology not only enhances the precision of CNV detection but also provides a spatial context for the genetic aberrations within the tumor microenvironment. Furthermore, our findings have the potential to inform therapeutic strategies by identifying key genetic drivers and their spatial associations within the tumor.

Reference

Huang, R., Huang, X., Tong, Y., Yan, H. Y. N., Leung, Y., Stegle, O., & Huang, Y. (2023.). XClone: detection of allele-specific subclonal copy number variations from single-cell transcriptomic data.https://doi.org/10.1101/2023.04.03.535352

ENVIRONMENTAL AND GENETIC INSIGHTS INTO CARCINOGENESIS: AN APPROACH USING PASSIVE SAMPLING AND CHIP ANALYSIS IN THE COMPANION DOG

Christopher Husted[1,2], Kate Megquier[2], Adam Harris[3], Diane Genereux[1,2], Kim Anderson[4], Alexander Bick[5], Frances Chen[1,2], Elinor Karlsson[1,2]

[1]University of Massachusetts Chan Medical School, Bioinformatics and Integrative Biology, Worcester, MA, [2]Broad Institute of MIT and Harvard, Vertebrate Genome Biology, Cambridge, MA, [3]College of Veterinary Medicine and Biomedical Sciences, Colorado State University, Department of Microbiology, Immunology and Pathology, Fort Collins, CO, [4]Oregon State University, Department of Environmental & Molecular Toxicology, Corvallis, OR, [5]Vanderbilt University Medical Center, Division of Genetic Medicine, Department of Medicine, Nashville, TN

Studying the impact of environmental exposures on cancer risk in humans is difficult due to our long lifespans and the long latency between exposure and cancer diagnosis. Companion dogs present a unique opportunity to identify carcinogenic exposures due to their shorter lifespans and shared environment with humans. In this study, we implemented approaches to identify environmental exposures and clonal hematopoiesis of indeterminate potential (CHIP), a genetic biomarker of increased cancer risk, in pet dogs. Passive sampling quantifies environmental exposures, whereas CHIP signals genetic vulnerabilities.

We identified environmental exposures and their association with cancer in 101 dogs utilizing Mann-Whitney U tests and regression analysis. Owners reported their dogs' health status, including the presence of >20 cancer types via an online survey. Silicone tags worn by the dogs were analyzed for >1500 analytes using GC-MS to identify environmental exposures. Preliminary results highlighted two chemicals correlated with cancer status, which will be further investigated in expanded cohorts.

In an independent cohort of 641 dogs, we examined CHIP mutations as markers for cancer vulnerability. Targeted sequencing of genes connected to CHIP and canine malignancies, with a central focus on canine lymphoma, identified mutations in *ASXL1*, *JAK2*, and *TET2*, reflecting findings in human CHIP research and underscoring genetic parallels in cancer predispositions between humans and dogs.

These two methodologies reveal environmental factors and genetic changes related to cancer. We plan to integrate these approaches using machine learning methods, enabling dogs as environmental sentinels for comparative studies. Our findings demonstrate the potential of passive sampling of dogs as models to guide future preventive strategies, benefiting interspecies health by offering a deeper understanding of environmental and genetic interaction in cancer development.

COMPRESSED INDEXING FOR PANGENOME SUBSTRING QUERIES

Stephen Hwang[1], Omar Y Ahmed[2], Ben Langmead[2]

[1]Johns Hopkins School of Medicine, XDBio Program, Baltimore, MD,
[2]Johns Hopkins University, Department of Computer Science, Baltimore, MD

There is a growing availability of collections of genomes, including pangenomes and taxonomic databases. However, tools to query substrings from these collections are limited to fixed k-mer size, rely on computationally intensive multiple sequence alignment, or scale poorly in space to larger databases. Here we present Maximal Exact Match Ordered (MEMO), a tool to efficiently index a collection of genomes and query substrings of any length. We compare the MEMO index to k-mer-based indexes for pangenome substring queries including global and local count, presence/absence, and sequence conservation. The MEMO index is created starting from the full document array profile (P_{DA}) which can be efficiently constructed alongside the Burrows-Wheeler Transform. The P_{DA} stores similarity statistics in the form of longest common prefixes (LCPs) between positions of a query genome and all other genomes in a pangenome. Order-specific-MEMs of the sorted per-position LCPs yield match intervals to at least an Nth-order number of other documents in the collection. Our index is formed by compressing these consecutive overlapping order-MEM intervals whereby a substring fully spanning an interval is absent in that Nth-order number of documents. MEMO allows for user-selected k-flexible visualization of pangenome sequence homology in $O(rd)$ space index, where r is the number of Burrows-Wheeler runs and d is the number of documents. MEMO indexed relative to a single pivot genome is ~1.5-4x larger than a k-mer index with fixed k. MEMO is a compressed index that enables efficient k-flexible substring queries, allowing for visualizing pangenome gene homology and chromosome-wide sequence conservation without the need for re-indexing at different k-mer lengths or genome resolutions.

INVESTIGATING RNA SPLICING AS A SOURCE OF CELLULAR DIVERSITY USING A BINOMIAL MIXTURE MODEL

Keren Isaev[1,2], David A Knowles[1,2,3,4]

[1]Columbia University, Systems Biology, New York, NY, [2]New York Genome Center, New York Genome Center, New York, NY, [3]Columbia University, Computer Science, New York, NY, [4]Columbia University, Data Science Institute, New York, NY

Alternative splicing (AS) contributes significantly to RNA and protein variability as well as gene expression regulation, yet its role in cellular diversity is not fully delineated. One challenge to studying AS is that it can be difficult to quantify in single cells. 10X-based methods, which are commonly used for single-cell RNA sequencing (scRNA-seq), have limited and biased transcript coverage. In contrast, high-coverage techniques like Smart-seq2 offer clearer RNA splicing insights, albeit at lower throughput. Another challenge is that most computational tools for scRNA-seq differential splicing analysis focus on simple binary events such as exon skipping, and rely on predefined cell type labels or low-dimensional gene expression representations. This limits their ability to detect more complex AS events, while relying on prior knowledge of cell classifications. Here, we present Leaflet, a splice junction centric and annotation-free approach inspired by Leafcutter, our tool for quantifying RNA splicing variation with bulk and short read RNA-seq data. Leaflet is a probabilistic mixture model designed to infer AS-driven cell states without the need for cell type labels. We detail Leaflet's generative model, inference methodology, and its efficiency in detecting differentially spliced junctions. By applying Leaflet to the Tabula Muris Smart-seq2 brain cell dataset, we highlight cell type specific splicing patterns, offering a deeper insight into cellular diversity beyond that captured by gene expression alone.

DETECTION, CHARACTERIZATION, AND PREVENTION OF MMEJ DELETIONS

Aditee Kadam[1], Shay Shilo[1], Hadas Naor[1], Mark Minden[2], Nathali Kaushansky[1], Noa Chapal[1], Liran Shlush[1]

[1]Weizmann Institute of Science, Department of Molecular Cell Biology, Rehovot, Israel, [2]Princess Margaret Cancer Centre, Department of Medical Oncology & Hematology, Toronto, Canada

Background: Detecting medium-sized deletions in short-reads is highly challenging due to reference biases and mapping issues. Although recent advances have been made in long-read sequencing, most of the available data still comes from inadequately analyzed short-read sequencing. We developed an algorithm that enables de novo detection of medium-sized deletions: Del-Read. The algorithm focuses on a specific type of deterministic deletion with a well-defined genetic mechanism - Micro-Homology mediated End Joining deletions (MMEJ-del). Using prior knowledge of the MMEJ mechanism, our algorithm compiles a complete set of potential deletions in the exome. Subsequently, it maps these deletions to sequencing reads, thereby reducing reliance on mapping differences to a reference genome.

Aims: To explore the somatic and germline MMEJ-del landscape using Del-Read, and provide insights into preventing somatic deletions through genome editing.

Methods: The Del-Read algorithm was applied to two datasets - Beat AML and TCGA-breast - which comprised of tumor-control paired exomes (N=359 and 225, respectively). A subset of these mutations underwent deep targeted sequencing in a cohort of 500 healthy individuals. The somatic MMEJ-del *ASXL1* was then edited across the homology in the K562 cell line using CRISPR-Cas9.

Results: The Del-Read algorithm identified reported (N=82), novel germline (N=486), and somatic (N=20) MMEJ-del in the datasets. A subset of these mutations (N=37) was validated with comparable population frequencies using targeted sequencing of healthy individuals (N=500) in ethnicity-matched controls.

The magnitude of novel MMEJ-del discovered allowed us to associate them with genomic features of replication stress such as G-quadruplexes and minisatellites. Interestingly, we also observed a new class of MMEJ-del characterized by mismatches in the homologies, although not all mismatches were equally tolerated. Further, we demonstrated that a specific single-base substitution can restrict the occurrence of pre-leukemic MMEJ-del in the *ASXL1* gene.

Summary: Our Del-Read algorithm is a promising new route in detecting medium-sized deletions and provides insights into their association with genomic features of replication stress. Additionally, our findings suggest the potential for preventing somatic MMEJ-del through genome editing of homologies. Collectively, our results underlie the potential of our algorithm in revealing previously undetected deletions and provide insights into the prevention of somatic deletions through genome editing.

CCSA: CONCURRENT FORCE-BASED POSITION SOLVING AND QUANTIZATION FOR MULTIPLE SEQUENCE ALIGNMENT

Daniel Kim

Horace Greeley High School, Applied Programming, Chappaqua, NY

We present a novel multiple sequence alignment (MSA) scheme from an original approach in which alignments of all the sequences are adjusted concurrently so that the alignment quality is maximized spontaneously in a single convergence process, avoiding an exhaustive, time-consuming search across the combinatorial space of possible alignments. The proposed scheme, CCSA (Concurrent Continuous Sequence Alignment), models a set of sequences as dynamically interacting parallel chains of sliding bases.

The procedure begins by mapping each input sequence into a set of 1-dimensional coordinates representing the sequential positions of the bases. The system defines attractive and repulsive interactions among bases within or across sequences. Steady-state coordinates that balance interacting forces can be obtained by updating coordinates using incremental small step sizes, as in transient circuit simulations or training neural networks. The solved steady-state coordinates provide not discrete positions but instead continuous values. A quantization process is performed to translate the final coordinates into integer values to obtain the final alignment results.

Many different cross-interaction relations among bases and sequences have been studied to understand the convergence and dynamics of the system. Algorithms and modeling have been optimized to maximize accuracy while maintaining a reasonable alignment time. Tests have been conducted using both simulation datasets and biological datasets, including 16S, and we show that the scheme can achieve comparable or even better accuracy in many datasets compared to prevalent MSA algorithms.

REVEALING HIDDEN TRANSCRIPTS WITH A COMPLETE REFERENCE AND PERSONALIZED TRANSCRIPTOME GRAPH

Juhyun Kim[1,2,3], Elizabeth Tseng[4], Adam M Phillippy[1], Arang Rhie[1]

[1]NIH, NHGRI, Bethesda, MD, [2]Seoul National University College of Medicine, Department of Biomedical Sciences, Seoul, South Korea, [3]Seoul National University, Genomic Medicine Institute, Seoul, South Korea, [4]Pacific Biosciences, Menlo Park, CA

A complete human reference genome (hereby CHM13) was recently released, encompassing an additional 200 Mbp of sequence. It corrects thousands of structural errors and unveils the most intricate regions of the human genome. By incorporating genes absent in GRCh38, it offers a chance to uncover novel transcripts previously excluded from analysis. Additionally, a personalized diploid reference can be employed for analyzing haplotype-aware transcripts, which is relevant for allele-specific expression. However, this task is challenging due to the high similarity between maternal and paternal sequences.

In this study, we utilized two long-read PacBio MAS-Seq read sets derived from GM26105/HG002 and asked which reference discovers the most full-length isoforms. We compared mapping statistics between GRCh38, CHM13, and a personalized, complete, diploid assembly of HG002 as the reference (HG002-dip). We found 4,900 transcriptome reads that mapped to CHM13 but not GRCh38. These reads mapped to 62 regions on CHM13, among which 19 loci (30.6%) were found to be in non-syntenic regions with respect to GRCh38. Furthermore, the number of discovered transcripts in CHM13 was greater than that in GRCh38 by around 6,000. Out of 124,762 total transcripts, 40,794 featured novel combinations of known splicing junctions, while there were 1,326 fewer transcripts with novel exons or splice junctions compared to GRCh38. This indicates that the gene annotation on CHM13 is more comprehensive than GRCh38. When utilizing HG002-dip as the reference, an additional hundreds of transcripts aligned compared to CHM13, however 70.52% of the total reads were aligned ambiguously, hindering accurate interpretation of allele-specific expression.

To overcome this problem, we used HG002-dip to construct a diploid genome graph of HG002 and projected the annotation of CHM13. Next, using the projected annotation, we generated a haplotype-aware complete transcriptome graph. This approach improved the mapping quality compared to using CHM13 or HG002-dip, due to the collapse of perfectly homozygous sequence combined with preservation of haplotype-specific variants in the graph.

In conclusion, we demonstrate that the more comprehensive CHM13 reference enhances overall transcript mappability and that a personalized, diploid transcriptome graph enables the comprehensive assignment of allele-specific transcripts. This enhancement is crucial for accurate isoform identification and the discovery of novel transcripts, enabling the reference to be applicable across various fields, including not only genomics but also transcriptomics

MOSCAL: DETECTION OF MOSAIC VARIANTS USING LINKED-READ SEQUENCING

Yongjun Kim[1,2], Shinwon Hwang[2,3], Hyeonju Son[1,2], Sangwoo Kim[1,2]

[1]Yonsei University College of Medicine, Graduate School of Medical Science, Brain Korea 21 Project, Seoul, South Korea, [2]Yonsei University College of Medicine, Department of Biomedical Systems Informatics, Seoul, South Korea, [3]Yonsei University College of Medicine, Department of Medicine, Physician-Scientist Program, Seoul, South Korea

Genomic mosaicism describes the presence of multiple cell lineages derived from distinct fertilized eggs. Detection of mosaic variants has unraveled the genomic pathogenicity of many diseases including early developmental disorders and cancers. However, accurate identification of mosaic variants has been frequently confounded owing to the low variant allele frequency and the absence of a clear matched control for germline variant filtration. While several strategies to achieve higher precision, such as read-backed phasing with nearby heterozygous germline SNPs have been employed, the lack of such SNPs within a short read hindered wider application. To secure a sufficient number of phased heterozygous SNPs, we applied the Linked-read sequencing technology (10x Genomics), which leverages the barcode DNA to generate data type provides contextual information about the genome from short-reads. Using this technology, we developed a novel variant pipeline MOSCAL that utilizes distant heterozygous germline SNPs that are phased into the regions of interest. In benchmarks on datasets used for Single-sample Mosaic SNV calling with linked read (Samovar) training, our pipeline achieved improved accuracy (F1-score of 0.729 and 0.604 in 60x and 30x sequencing) than previous methods (MuTect2; 0.542 and 0.318, and Samovar; 0.619 and 0.477, in 60x and 30x sequencing, respectively). We expect that the use of linked-read sequencing would provide new options for identifying mosaic variants.

LEVERAGING PUBLIC DATASETS TO UNDERSTAND PARKINSON'S DISEASE PROGRESSION

Rohit Kolora, Anna Rychkova

Alector Therapeutics, Target Discovery and Genomics, South San Francisco, CA

Parkinson's disease (PD) is one of the most prevalent age-related neurodegenerative diseases affecting over 10 million people worldwide. Studies have shown ageing to be a primary risk factor of most neurodegenerative disorders, however disease associated changes in the brain occur years before the onset of symptoms making pre-clinical diagnosis crucial. Central Nervous System (especially the brain) is extremely vulnerable to the detrimental effects of ageing. Multi-Omics datasets from longitudinal studies of diverse genetic cohorts can help detect signals associated with early onset of the disease as well as its progression. This enables us to identify biomarkers and prioritize drug targets, which help in developing therapies. Parkinson's Progression Markers Initiative (PPMI) harnesses such PD-related ~Omics datasets offering the opportunity to better identify, understand and treat PD. We analyzed transcriptomic, proteomic and methylation data from PD-affected and healthy controls from PPMI. Accounting for confounding factors such as sex, age of onset and demography helped us identify variably expressed genes and methylated loci with age-specific effects implicated in various stages of PD-related etiology. Proteomics data from plasma and cerebrospinal fluid is useful in identifying biomarkers in addition to the well-known ones such as Synuclein alpha and Neurofilament light. Furthermore, using a machine learning based classifier and guilt-by-association approaches, we observed enrichments in PD-related mechanisms such as lysosomal pathways, mTOR signaling and immune-responses, ultimately helping us propose a candidate list of potential drug targets against PD.

GENE EXPRESSION PREDICTION FROM HISTOPATHOLOGY IMAGES OF COLORECTAL CANCER

Jonas Lehmitz[1,2], Philip Bischoff[1], Alexander Sudy[3], Johannes Liebig[3], Christian Conrad[3], Teresa G Krieger[1,3]

[1]Charité, Institute of Pathology, Berlin, Germany, [2]RWTH Aachen, Institute of Medical Informatics, Aachen, Germany, [3]Berlin Institute of Health, Digital Health Center, Berlin, Germany

Recent spatial transcriptomics technologies enable the study of tumour cells and their microenvironment in spatial context, but their applicability is limited by time and financial constraints. In contrast, haematoxylin & eosin (H&E) stained whole-slide images are cheap and routinely acquired for histopathological assessments. Predicting gene expression based on H&E images would thus enable us to extend results from small spatial transcriptomics studies to larger datasets and validate hypotheses in clinical cohorts.

Here, we used a direct in-situ hybridisation approach to detect transcription of 192 target genes in ten colorectal cancer (CRC) tissue sections. A transfer learning approach was implemented to develop deep learning based models for (i) the automated detection of functional tissue area types and (ii) the prediction of CRC subtype and stromal marker gene expression from H&E images. While the small dataset size as well as low tissue and image quality were identified as major constraints for model performances, spatial transcriptional domains could be matched with relevant histological regions.

Our approach may enable cost-efficient screenings for spatial molecular biomarkers to aid in cancer subtype diagnosis, therapy selection, and patient stratification for clinical trials in the future.

THE NHGRI GENOMIC DATA SCIENCE ANALYSIS, VISUALIZATION, AND INFORMATICS LAB-SPACE (AnVIL)

Natalie Kucher[1], Michael C Schatz[1,2], Anthony Philippakis[3]

[1]Johns Hopkins University, Department of Biology, Baltimore, MD, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [3]Broad Institute, Cambridge, MA

* The full list of contributors is available at: https://anvilproject.org/about/team.

The traditional model of genomic data sharing – centralized data warehouses such as dbGaP from which researchers download data to analyze locally – is increasingly unsustainable. Not only are transfer/ download costs prohibitive, but this approach also leads to redundant siloed compute infrastructure and makes ensuring security and compliance of protected data highly problematic.

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space, or AnVIL, inverts this model, providing a cloud environment for the analysis of genomic and related datasets. By providing a unified environment for data management and compute, AnVIL eliminates the need for data movement, allows for active threat detection and monitoring, and provides elastic, shared computing resources as needed. AnVIL currently provides harmonized access to >600,000 genomes, with many more on the horizon, as well as thousands of software tools and several options for interactive and batch analysis.

The platform is built on a set of established components that have been used in a number of flagship scientific projects. First, we discuss how AnVIL supports the Telomere-to-Telomere (T2T) consortium through a large-scale reanalysis of the 1000 Genomes cohort orchestrated through the Workflow Description Language (WDL) on Terra. This enabled us to identify over one million variants in the newly resolved regions of the human genome, including within the recently completed chromosome Y sequence. Next, we present new results detecting single- and multi-tissue SV-eQTLs by genotyping SVs discovered with long-reads within the GTEx short-read sequencing data. This work was quickly and securely executed through WDLs, Jupyter notebooks, and R/Bioconductor, and lead to the discovery of 5,580 SV-eQTLs where the SV has the highest CAVIAR score (a metric of causality) over other nearby SNVs. Finally, we discuss how to analyze pangenomes and haplotype diversity using Galaxy within Terra. This is critically important to diversity and disease studies, especially to capture and analyze variation not found in any single reference genome.

Long-term, AnVIL will provide a unified platform for ingestion and organization for a multitude of current and future genomic and genome-related datasets; ease the process of acquiring access to protected datasets for investigators, and reduce the burden of performing analyses across many datasets to fully realize the potential of ongoing data production efforts.

TRANSCRIPTOME ANALYSIS OF *HELICOVERPA ARMIGERA* (HUBNER) (LEPIDOPTERA: NOCTUIDAE) INFECTED WITH *METARHIZIUM ANISOPLIAE*

Pankaj Kumar, Malyaj R Prajapati

Sardar Vallabhbhai Patel University of Agriculture and Technology, Bioinformatics, Meerut, IN

Helicoverpa armigera (Hubner), stands out as a serious agricultural menace, inflicting substantial damage to a wide range of crops, including cotton, corn, tomato, and soybean. One promising approach involves the use of entomopathogenic fungi, which are naturally occurring pathogens that infect and kill insects. Among them, the fungus *Metarhizium anisopliae* has shown great potential as a biocontrol agent against various insect pests. Aim of this study, high-throughput sequencing and assembly techniques to investigate the transcriptome of *Helicoverpa armigera* under *Metarhizium anisopliae* treated and non-treated conditions. Illumina NovaSeq6000 sequencing generated a substantial number of raw reads for both conditions. After filtering and quality control steps, the processed reads were subjected to de novo assembly, resulting in contigs. Functional annotation of the H. armigera transcriptome was performed by comparing the contigs to sequences in the NCBI-NR database and utilizing GO, COG database and KEGG database. The annotation revealed significant similarity with sequences from various insect species. Additionally, the analysis identified immune-related differentially expressed genes in *H. armigera*, including those involved in Toll and IMD signaling pathways. Upregulation of genes associated with Toll and IMD pathway was observed in response to treatment, suggesting their role in the immune response against fungal pathogens. This study provides valuable insights into the transcriptomic profile and immune-related gene expression of *H. armigera* under treated conditions. The findings provide valuable information regarding the functional categories, metabolic pathways, and immune-related genes in this insect pest. This knowledge contributes to our understanding of *H. armigera* biology and opens avenues for further research on pest management strategies and immune modulation.

AUTOMATED REFERENCE GENOME ASSEMBLY ON PUBLIC INFRASTRUCTURE WITH GALAXY

Delphine Lariviere[1], Linelle Abueg[2], Nadolina Brajuka[2], Cristóbal Gallardo-Alba[3], Bjorn Grüning[3], Byung June Ko[4], Alex Ostrovsky[5], Marc Palmada-Flores[6], Brandon D Pickett[7], Keon Rabbani[8], Erich D Jarvis[2], Adam M Phillippy[7], Anton Nekrutenko[1], Michael Schatz[5], Giulio Formenti[2]

[1]Pennsylvania State University, Eberly College of Science, University Park, PA, [2]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, [3]Albert-Ludwigs-University Freiburg, Bioinformatics Group, Department of Computer Science, Freiburg, Germany, [4]Seoul National University, Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul, South Korea, [5]Johns Hopkins University, Department of Biology, Baltimore, MD, [6]Universitat Pompeu Fabra-CSIC, Department of Medicine and Life Sciences (MELIS), Institut de Biologia Evolutiva, Barcelona, Spain, [7]National Human Genome Research Institute, National Institutes of Health, Genome Informatics Section, Computational and Statistical Genomics Branch, Bethesda, MD, [8]University of Southern California, Department of Quantitative and Computational Biology, Los Angeles, CA

Recent improvements in genome sequencing and assembly promise to generate high-quality reference genomes for many species. Yet the genome assembly process is still laborious and costly, requiring substantial expertise, and is generally not scalable to the goals of multispecies scientific efforts. To democratize the training and assembly process, we implemented the latest version of the Vertebrate Genomes Project assembly pipeline in Galaxy (https://galaxyproject.org/projects/vgp/).

The automated pipeline performs de novo assembly based on PacBio HiFi reads, with optional extended graph phasing using HiC or parental data and scaffolding using Bionano optical mapping data and HiC via modular workflows. The workflows include quality control throughout the assembly process using GenomeScope, gfastats, Merqury, BUSCO, and Pretext. We will show how to use these workflows via the Galaxy interface or via the command line to rapidly generate dozens of assemblies using the free public resources available. We will also discuss the quality of generated assemblies and how it is impacted by the technology used.

Within the Vertebrate Genome Project, these workflows have already been applied to de novo assemble the genomes of over a hundred species. The VGP long-term goal is to use these workflows to generate high-quality, complete reference genomes for all of the roughly 70,000 extant vertebrate species and to help enable a new era of discovery across the life sciences.

This partnership with the Vertebrate Genomes Project has led to several enhancements to Galaxy's ability to be utilized on large-scale projects. In addition to the new Galaxy assembly workflows, we will highlight Galaxy's (1) new features improving the development of large modular workflows and (2) current efforts on the globalization and integration of public computational infrastructure from the EU, US, and Australia.

Complete author list : https://doi.org/10.5281/zenodo.8370925

EXPLORING FUNCTIONAL DIVERGENCE IN PARALOGS USING EMBEDDINGS FROM PROTEIN LANGUAGE MODELS

Denise Le[1], Alan M Moses[1,2,3]

[1]University of Toronto, Department of Cell and Systems Biology, Toronto, Canada, [2]University of Toronto, Department of Computer Science, Toronto, Canada, [3]University of Toronto, Department of Ecology and Evolutionary Biology, Toronto, Canada

Gene duplication is an important evolutionary mechanism that drives the evolution of new genes with novel or altered protein functions. While a majority of gene duplicates are lost, a significant number are retained, leading to paralogs (gene duplicates) with diverged functions. Protein embeddings are numerical representations of protein sequences obtained from deep language models and capture important features related to protein structure and function. Embedding spaces generated from deep learning methods have been used to predict evolutionary relationships, evolutionary dynamics, and function of proteins. Here, we investigate the use of the embedding space to detect and visualize functional divergence of protein paralogs after gene duplication. Preliminary results in yeast proteins suggest a correlation between the separation of paralogs in the embedding space and divergence in Gene Ontology (GO) term annotations. Additionally, we observe examples where visualizations of the reduced dimensionality embeddings align with functional divergence patterns in yeast subcellular localization and mirror shifts in their functional traits. We aim to leverage the spatial distribution of embeddings of protein families to predict divergence of protein function and visualize changes in evolution. This work explores the ability of protein embeddings to enhance our understanding of patterns in protein evolution and functional divergence after gene duplication.

A FULLY ASSEMBLED, PHASED YUCATAN REFERENCE GENOME ENABLES ACCURATE ON-TARGET AND OFF-TARGET ANALYSIS

Feng Li, Xiaoyun Guo, Shiyi Yin, Owen Pearce, Wing-On Ng, Chris Chao, Josh Stirba, Matthew Pandelakis, Jacob V Layer, Wenning Qin, Ranjith P Anand, Sagar Chhangawala

eGenesis, Inc, Cambridge, MA

The persistent organ shortage, such as kidney, heart, and liver, poses a significant challenge in medical care. An innovative approach to this challenge is xenotransplantation, which involves the transplantation of genetically engineered organs from large mammals, such as pigs, to humans. Ensuring the safety and efficacy of these gene edited organs requires the careful analysis of the edited genomic loci and the thorough characterization of unintended genomic alterations. At eGenesis, we employed CRISPR-Cas9 technology to eliminate xenoantigens, inactivate porcine endogenous retroviruses (PERVs), and express human transgenes to mitigate the risk of organ rejection and immune responses. We focused on the Yucatan miniature pig breed.

Here, we created the first high-quality haplotype-resolved genome assembly for Yucatan pig breed using PacBio HiFi, Bionano Optical Genome Mapping and Hi-C technologies, with hifiasm, Bionano hybrid scaffolding and YaHS. The contig N50 is 42 Mb and 47.6 Mb for haplotype 1 and 2, respectively. The scaffold N50 is 140.8 Mb and the BUSCO score for both haplotypes are 96.4% and 96.1 % for haplotype-phased chromosomes. This indicates a highly contiguous and complete genome assembly, which is used as reference for Yucatan-specific analyses in our work.

One of our primary concerns in xenotransplantation is the potential risk of zoonotic transmission posed by PERVs. To address this concern, we developed a comprehensive strategy based on BLAST analysis to identify all full-length and truncated PERV copies, along with their genomic loci, within the porcine genomes. Our findings, based on query coverage and identity to PERV *pol*, *gag*, and *env* domains, reveal that a representative Yucatan haplotype assembly carries 28 full-length PERV copies, including 20 copies of PERV-A, 5 copies of PERV-B, and 3 copies of PERV-C, along with suggestive evidence of 8 to 9 truncated PERV copies. Furthermore, we have employed a phylogenetic approach utilizing MAFFT and PhyML to classify PERV copies into sub-families, shedding light on the genetic diversity among these closely related PERVs. Notably, PERV-B and PERV-C form distinct clades, while PERV-A branches into two separate clades, underscoring the genetic variations within the PERV-A group. This information informs our primer design for PERV regions and enhances the accuracy of our PERV knockout characterization.

To our knowledge, this is the first fully assembled, phased Yucatan reference genome. This genomic resource and PERV detection tool will serve the xenotransplantation community, facilitating accurate characterization of genomic on-target, as well as off-target edits.

ANALYSIS OF PANCREATIC CANCER RISK VARIANTS USING LONG READ SEQUENCING

Qiuhui Li[1], Carolina Montano[2], Jessica Hosea[2], Luke Morina[2], Bohan Ni[1], Justin Paschall[3], Beth Marosy[3], Michelle Kokosinski[3], Jessica Gearhart[3], Brian Craig[3], Alan Scott[3], David Mohr[3], Michelle Mawhinney[3], David McKean[4,5], Nicholas Roberts[4,5], Zhanmo Ni[4,5], Alexis Battle[1,2], Kimberly Doheny[3], Winston Timp[2], Michael Schatz[1], Alison Klein[4,5]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [3]Johns Hopkins University, Department of Genetic Medicine, Baltimore, MD, [4] The Johns Hopkins Medical Institutions, Department of Oncology, The Sol Goldman Pancreatic Cancer Research Center, Baltimore, MD,
[5]The Johns Hopkins Medical Institutions, Department of Pathology, The Sol Goldman Pancreatic Cancer Research Center, Baltimore, MD

Pancreatic cancers are the third leading cause of cancer-related deaths in the United States, and the annual number of diagnoses worldwide has more than doubled over the past 25 years. Despite recent improvements in detection and treatment, the 5-year survival rate in the US remains a dismal 12%. Family history is one of the most important known risk factors and increases risk by 2.3 to 32-fold. Association studies leveraging short-read sequencing have identified germline genetic risk factors in BRCA2, PALB2, ATM, and a few other genes. However, only a small fraction (~10%) of familial cancers have recognizable mutations, leaving many patients unaware of their true risk. Comprehensive variation profiling is critical to advancing cancer susceptibility research. However, investigating complex structural variations and certain SNVs - particularly those in highly repetitive regions - remains challenging due to short-read sequencing constraints.

Addressing this need, we are using long-read sequencing to comprehensively assess variants in high-risk familial pancreatic cancer patients. Long-read sequencing produces reads tens of kilobases long that can span difficult-to-assemble regions and improve variation identification. To date, we have performed long-read whole genome sequencing on germline samples from >30 familial cancer patients on the Oxford Nanopore PromethION platform (R10.4 flowcells and LSK114 library prep) to an average mean depth of 58X and n50 of 27.5 kb. We produced highly accurate variant calls using Sniffles2, PAV, and Clair3. Notably, using long reads, we detect SVs and other complex variants in these patient genomes that are not detectable using short reads in genes known to increase risks for hereditary pancreatic cancer and other hereditary cancers. Using SVs cataloged within the families and those identified in other control samples, we are designing an integrative algorithm to detect likely pathogenic germline alterations in patients' genomes. Our ultimate goal is to build a comprehensive variant dataset comprising over 400 familial pancreatic patient samples and unravel the importance of structural variants in cancer. We expect our efforts to substantially improve patient genetic profiling, especially in high-risk families, thereby improving pancreatic cancer prevention and therapy and ultimately saving lives.

GENETIC AND DIETARY MODULATORS OF THE INFLAMMATORY RESPONSE IN THE GASTRO-INTESTINAL TRACT OF THE BXD MOUSE GENETIC REFERENCE POPULATION

Xiaoxu Li[1], Jean-David Morel[1], Giorgia Benegiamo[1], Johanne Poisson[1], Alexis Bachmann[1], Alexis Rapin[1], Jonathan Sulc[1], Evan Williams[3], Alessia Perino[2], Kristina Schoonjans[2], Maroun Bou Sleiman*[1], Johan Auwerx*[1]

[1]École Polytechnique Fédérale de Lausanne, Laboratory of Integrative Systems Physiology, Institute of Bioengineering, Lausanne, Switzerland, [2]École Polytechnique Fédérale de Lausanne, Laboratory of Metabolic Signaling, Institute of Bioengineering, Lausanne, Switzerland, [3]University of Luxembourg, Luxembourg Centre for Systems Biomedicine, Luxembourg, Luxembourg

Inflammatory gut disorders, including inflammatory bowel disease (IBD), can be impacted by dietary, environmental and genetic factors. While the incidence of IBD is increasing worldwide, we still lack a complete understanding of the gene-by-environment interactions underlying inflammation and IBD. Here, we profiled the colon transcriptome of 52 BXD mouse strains fed with a chow or high-fat diet (HFD) and identified a subset of BXD strains that exhibit an IBD-like transcriptome signature on HFD, indicating that an interplay of genetics and diet can significantly affect intestinal inflammation. Using gene co-expression analyses, we identified modules that are enriched for IBD-dysregulated genes and found that these IBD-related modules share cis-regulatory elements that are responsive to the STAT2, SMAD3, and REL transcription factors. We used module quantitative trait locus (ModQTL) analyses to identify genetic loci associated with the expression of these modules. Through a prioritization scheme involving systems genetics in the mouse and integration with external human datasets, we identified Muc4 and Epha6 as the top candidates mediating differences in HFD-driven intestinal inflammation. This work provides insights into the contribution of genetics and diet to IBD risk and identifies two candidate genes, MUC4 and EPHA6, that may mediate IBD susceptibility in humans.

SYSTEMS GENETICS OF METABOLIC HEALTH IN THE BXD MOUSE GENETIC REFERENCE POPULATION

Xiaoxu Li[1], Jean-David Morel[1], Alessia De Masi[1], Amélia Lalou[1], Jonathan Sulc[1], Giorgia Benegiamo[1], Johanne Poisson[1], Yasmine Liu[1], Arwen W Gao[2], Maroun Bou Sleiman[1], Johan Auwerx[1]

[1]École Polytechnique Fédérale de Lausanne, Laboratory of Integrative Systems Physiology, Institute of Bioengineering, Lausanne, Switzerland, [2]University of Amsterdam, Laboratory Genetic Metabolic Diseases, Amsterdam Gastroenterology, Endocrinology and Metabolism Institute, Amsterdam, Netherlands

Susceptibility to the metabolic syndrome (MetS) is dependent on genetics, the environment and gene-by-environment interactions (GxE), rendering the study of underlying mechanisms challenging. The majority of mouse experiments do not incorporate genetic variation and lack specific evaluation criteria to define and monitor the MetS. Here, we addressed these two shortcomings using multi-omic data from genetically diverse BXD mouse strains in which we develop a continuous metabolic health score (MHS) based on standard clinical parameters. The metric was designed to capture general metabolic health and, when derived in human subjects of the UK Biobank, predicts the metabolic syndrome and future disease incidence. Using quantitative trait locus analyses in mice, three MHS-associated genetic loci were identified and validated in different, unrelated, mouse populations. Through a prioritization scheme involving systems genetic analysis in BXD mice and integration with human databases such as the UK Biobank, we identified TNKS, MCPH1, and PCSK5 as candidates mediating differences in the MHS. Our findings provide insights into the understanding of sustaining metabolic health across species and the identification of candidates that might regulate metabolic health.

MEASURING, VISUALIZING AND DIAGNOSING REFERENCE BIAS WITH BIASTOOLS

Mao-Jan Lin, Sheila Iyer, Nae-Chyun Chen, Ben Langmead

Johns Hopkins, Computer Science, Baltimore, MD

A goal of recent alignment methods is to reduce reference bias, which occurs when reads containing non-reference alleles fail to align to their true point of origin. However, there is a lack of methods for systematically measuring, categorizing, and diagnosing reference bias. We present biastools, which analyzes and categorizes instances of reference bias.

Biastools offers different sets of functionality tailored to different scenarios, i.e. (a) when the donor genome is well-characterized and input reads are simulated, (b) when the donor is well-characterized and reads are real, and (c) when the donor is not well-characterized and reads are real. In scenario (a), biastools categorizes instances of reference bias based on their causes: bias due to loss, flux, or local misalignment. Loss events refer to the mapping of alternative allele reads to locations other than their true origin. Flux events indicate bias rise from inclusion of mismapped reads originated from elsewhere. Local bias events stem from local repeat content and sequencing errors that create ambiguity in gap placement. In scenario (b), biastools predicts the reference biases based on the allelic balance and mapping quality of the reads. In scenario (c), biastools operates in scan mode to detect large-scale mapping artifacts due to structural variation and flaws in the reference representation.

We assess reference bias with Bowtie 2, BWA MEM, and VG Giraffe with different numbers of variants in the graph genome. Our findings confirm that including more variants in the graph genome alignment method results in fewer reference biases. Additionally, we observe that end-to-end alignment modes are effective in reducing bias at insertions and deletions, compared to local aligners that allow soft clipping. Finally, we use biastools to characterize the ways in which using the new telomere-to-telomere human reference can improve bias at a large scale. In short, biastools is a tool uniquely focused on reference bias, making it a valuable resource as the field continues to develop new aligners and pangenome representations to reduce bias.

USING BETTER BASE QUALITY(BBQ) TO DETECT LOW-FREQUENCY SOMATIC MUTATIONS ACCURATELY

Yixin Lin[1], Carmen Oroperv[1], Claus L Andersen[1], Mikkel H Schierup[2], Asger Hobolth[3], Thomas Bataillon[2], Kristian Almstrup[4], Søren Besenbacher[1]

[1]Aarhus University, Department of Molecular Medicine, Aarhus, Denmark,
[2]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark,
[3]Aarhus University, Department of Mathematics, Aarhus, Denmark,
[4]Rigshospitalet, Department of Growth and Reproduction, Copenhagen, Denmark

Next-generation sequencing (NGS) has revolutionized genomics studies by providing high-throughput and cost-effective DNA sequencing. The extensive exploration of genetic variants from whole-genome sequencing(WGS) data has catalyzed the development of diverse variant calling tools. However, detecting low-frequency somatic mutations from WGS data remains problematic, particularly in scenarios where sequencing error rates at some sites might be similar to the variant allele frequency. Therefore, it is important to estimate the error rate of a specific base in a specific read accurately.

In paired-end sequencing, when DNA fragments are shorter than twice the read length, the overlapping region emerges as an ideal resource for training models to discern specific base error rates, as the mismatches inside the overlaps must be an error generated either by sequencing procedures or alignment. In this context, we introduce, ‘Better Base Quality’(BBQ), a method that leverages the mismatches within read overlaps to precisely estimate the error rate associated with different sequencing contexts. Specifically, we use the 7-mer encompassing the mismatch to predict error rates for all combinations of FASTQ base quality(BQ) and mutation type using kmerPaPa. In addition to the sequencing context, we hypothesized that multiple alternatives at a site within the read overlap suggest a higher likelihood of error; thus, we updated our model with the position-specific information. We applied a beta-binomial distribution to allow the posterior error rate to differ between sites based on the number of observed mismatches.

The model has been tested on both cell-free DNA cancer plasma samples and testis samples and can demonstrate significantly improved precision in calling low-frequency single nucleotide variations compared to Mutect2, Strelka2 and Lofreq as shown by the precision-recall curve. This work can elevate the accuracy of detecting circulating tumor DNA mutations and ultimately benefit early cancer detection and timely relapse monitoring.

< Back

DNA BENDABILITY REGULATES TRANSCRIPTION FACTOR PIONEER BINDING TO NUCLEOSOMES

Xiao Liu[1,2], Luca Mariani[2], Martha L Bulyk[2,3]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, [3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

The nucleosome is a complex formed by 147 DNA bases wrapped around a histone octamer for 1.65 turns. Being the fundamental unit of chromatin packaging, nucleosomes regulate gene expression by posing a barrier to transcription factors (TFs) binding to DNA. Recent studies have investigated how TFs can access nucleosomal DNA ("TF pioneer binding"), unveiling different interaction modes including binding at nucleosome ends, dyad and spanning gyres. Furthermore, single molecule looping assays revealed that the nucleosomal DNA ends unwrap asymmetrically with the stiffer half being more prone to opening. We hypothesized that this asymmetric unwrapping facilitates TF pioneer binding by exposing the less bendable half.

To test this hypothesis, we analyzed the DNA bendability of in vitro nucleosomal libraries bound by TFs and histone octamers from published NCAP-SELEX datasets for 154 TFs. To quantitatively evaluate the bendability of DNA sequences assembled into nucleosomes, we employed DNAcycP, an existing deep learning model based on loop-seq datasets. Loop-seq experiments measure the ability of 100-bp sequences to form loops ("DNA cyclizability"), creating a proxy for nucleosomal DNA bendability. Consistent with our hypothesis, we discovered that most TFs prefer binding at nucleosome ends adjacent to the less cyclizable ends, a trend that we did not observe in DNA libraries selected exclusively by either TFs or histone octamers. This behavior is also less pronounced for TFs which bind more strongly at the nucleosome dyad, suggesting that TFs capable of overcoming the histone octamer barriers do not rely on the nucleosome unwrapping for pioneer binding. We further analyzed the nucleosomal sequences selected by the gyre-spanning binding factor T(Brachyury). We noticed that sequences bound by T on a single gyre exhibit cyclizability asymmetry, while sequences with T spanning across two gyres do not. This agrees with our hypothesis that DNA bendability differences underlie the exposure of TF binding sites at the ends of nucleosomes. Integrated analyses of MNase-seq data with ChIP-seq data for C/EBP TFs support our model for how bendability regulates pioneer binding of TFs.

In summary, we found that TF binding sites at nucleosome ends adjacent to sequences with low bendability are preferentially exposed, and propose a model for how genomic sequence composition ~100-bp away from TF DNA binding sites on nucleosomes can influence TF binding through nucleosome dynamics.

PATHWAYGAT: A METHOD TO TRACE BACK BIOLOGICAL INTERACTIONS FROM MICROBES TO HOST PHENOTYPES

Shaoke Lou[1,2], Weihao Zhao[1,2], Mark Gerstein[1,2]

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

Diseases often arise from or are influenced by pathogen infections, leading to diverse clinical manifestations. While past research has shed light on the basic mechanisms of pathogen infection, a holistic understanding of the infection "interactome" remains elusive. This interactome encompasses not just direct protein-protein interactions but also factors like human miRNAs, additional protein-coding genes, and external microbes. We've pioneered a deep-learning approach to delve into multi-layered host-pathogen interactions, revealing how pathogens influence host phenotypes. Our model integrates data from microbiome abundance, host genetics, and inferred gene interactions, leveraging our MLCrosstalk framework. Using transfer learning, we first train a graph attention model to identify biological signatures linked to cancer phenotypes. We then fine-tune this model using a smaller dataset from Honduras, enriched with details on environment, diet, social interactions, and diseases. Our approach offers insights into the intricate relationship between pathogens and the broader host environment.

SPATIAL TRANSCRIPTOMICS DATA ANALYSES REVEALED CANCER-ENDOTHELIAL CELL COMMUNICATION IN HEPATOCELLULAR CARCINOMA

Chenyue Lu[1,2,3,4], Amaya Pankaj[1], Michael Raabe[1], Cole Nawrocki[1], Ann Liu[1], Nova Xu[1], Bidish K Patel[1], Matthew J Emmett[1], Avril K Coley[1,5], Cristina R Ferrone[5], Vikram Deshpande[6], Irun Bhan[1], Yujin Hoshida[7], David T Ting[1], Martin A Aryee[3,4,8], Joseph W Franses[1]

[1]Massachusetts General Hospital, Cancer Center, Boston, MA, [2]Harvard-MIT, HST, Cambridge, MA, [3]Dana-Farber Cancer Institute, Data Science, Boston, MA, [4]Broad Institute, Epigenomics, Cambridge, MA, [5]MGH, Surgery, Boston, MA, [6]MGH, Pathology, Boston, MA, [7]UT Southwestern, Medicine, Dallas, TX, [8]Harvard T.H. Chan School of Public Health, Biostatistics, Boston, MA

Hepatocellular carcinoma (HCC) is the most prevalent primary liver cancer and a leading cause of cancer-related deaths in the world, with only a 30% response rate to the current therapy for advanced disease. HCC tumors are complex ecosystems consisting of cellular components such as cancer cells, endothelial cells, fibroblasts, immune cells as well as non-cellular components such as cytokines and chemokines. Recent breakthroughs have targeted the vascular components, yet the role of endothelial cells is not well understood in HCC biology.

To dissect the heterogeneity in 41 HCC patients' tumor FFPE specimens, we used the Nanostring 1800-plex GeoMx Digital Spatial Profiling platform to profile 65 regions, each 400 µm x 400 µm in size. For each region, we separately quantified the transcriptome in adjacent cancer and endothelial spatial compartments. Hierarchical clustering revealed that the dominant HCC subtyping schema (Hoshida 2009) is largely driven by the proportion of endothelial cells in the tumor. We validated this finding in two orthogonal spatial and single-cell datasets.

Next we characterized the cancer cell compartments across all samples and found three distinct clusters, which we term T1, T2, and T3 subtypes, with different enriched pathways, highlighting transcriptomic heterogeneity in tumor cells from different patients. We then used canonical correlation analysis (CCA) to identify the most highly correlated features between cancer cells and their adjacent endothelial cells. Our CCA analysis allows for hypothesis generation regarding possible EC-cancer interactions in intact tissues. For example, we found that the expression of key chemokines (CXCL2, CXCL12) in the T2 tumor niche were correlated with multiple different chemokines (CCL19, CCL21) in the associated vessels. These findings could imply significant co-modulation of the microenvironment and cancer cells.

Our work uncovers molecular heterogeneity previously missed in tumor bulk transcriptome-based subtyping and demonstrates the power and promise of spatial technologies in cancer precision medicine. The results motivate the development of novel biomarkers and drug targets in HCC.

THE KRAKEN PROTOCOL IN ACTION: IDENTIFYING POTENTIAL PATHOGENS IN UGANDAN INDIVIDUALS WITH UNEXPLAINED ACUTE FEBRILE ILLNESS.

Jennifer Lu[1,2,3], Abraham J Kandathil[4], Raghavendran Anantharam[4], Kenneth Kobba[4], Paul W Blair[4,6], Matthew L Robinson[4], Edgar C Ndawula[5], Francis Kakooza[5], Mohammed Lamorde[5], David L Thomas[4], Martin Steinegger[7], Yukari C Manabe[4], Steven L Salzberg[1,2,8]

[1]Center for Computational Biology, Baltimore, MD, [2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, [3]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD, [4]Department of Infectious Diseases, Johns Hopkins School of Medicine, Baltimore, MD, [5]Infectious Diseases Institute, Makerere University, Kampala, Uganda, [6]Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc, Bethesda, MD, [7]School of Biology, Seoul National University, Seoul, South Korea, [8]Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD

Metagenomic experiments utilize high-throughput sequencing to reveal the microbial composition of any given environment. Accurate characterization of this microbial environment relies on the computational analysis of the sequencing data. For efficient and reproducible metagenomic analysis, we developed a Kraken protocol as an end-to-end pipeline for classification, quantification, and visualization of metagenomic datasets.

To illustrate the clinical application of this protocol, we used it to identify potential microbial infections in 44 hospital patients from Uganda who were suffering from unexplained, acute febrile illness. For each patient sample, we collected plasma and performed RNA sequencing using a NovaSeq 6000, generating an average of 25.7 million reads per sample. We analyzed each sample using the Kraken protocol, identifying strong signals of microbial infection in 8 of the study participants. The microbial infections identified included *Pegivirus hominis* (n=2), *Human herpes virus 8*, *Plasmodium falciparum*, *Helicobacter pylori*, *Rickettsia conorii*, *Staphylococcus haemolyticus*, and *Candida parapsilosis*.

TOWARDS A CLOUD-AGNOSTIC SCALABLE ECOSYSTEM FOR OPEN GENOMIC DATA SCIENCE WITH BIOCONDUCTOR AND GALAXY

Alexandru Mahmoud[1], Enis Afgan[2], Dirk Eddelbuettel[3], Marcel Ramos[4], Ludwig Geistlinger[5], Jen Wokaty[4], Lori S Kern[6], Davide Risso[7], Sean Davis[8], Levi Waldron[4], Vincent J Carey[1]

[1]Harvard Medical School/Mass General Brigham, Channing Division of Network Medicine, Boston, MA, [2]Johns Hopkins University, Biology, Baltimore, MD, [3]University of Illinois Urbana-Champaign, Statistics, Urbana-Champaign, IL, [4]City University of New York School of Public Health, Epidemiology and Biostatistics, New York, NY, [5]Harvard Medical School, Center for Computational Biomedicine, Boston, MA, [6]Roswell Park Comprehensive Cancer Center, Bioinformatics, Buffalo, NY, [7]University of Padova, Statistics, Padova, Italy, [8]University of Colorado, Biomedical Informatics, Aurora, CO

New single cell and image-oriented assays induce requirements for large storage volumes and distributed analytic computing, leading naturally to consideration of cloud technologies. There is also hope that improved uptake of cloud computing methodology will accelerate sound collaborative solutions to problems of data sharing and workflow benchmarking and reuse throughout genome biology. We describe approaches in Bioconductor and Galaxy to cloud scale infrastructure enhancement for genomic data science. **Storage:** We established egress-free access to large immunoflourescence imaging experiments, scalable representations of genomes called against the telomere-to-telomere reference, and resources for new assessments of variant pathogenicity based on AlphaFold. **Analytic computation support:** Allocations from NSF Jetstream2 and Microsoft Genomics enable flexible approaches to workflow design, instrumentation, and deployment. The Jetstream2 allocation includes various GPU setups, facilitating exploration of new AI/ML strategies arising in all aspects of genome biology. **Infrastructure solutions:** A significant barrier to innovative computations for genomic data science arises from the challenge of configuring software tools for effective use in available and familiar but performant platforms. We describe how focusing on package management technologies for specific operating systems (our example is Ubuntu Linux) allows investigators to configure and use readily available "bare metal" systems in cloud or elsewhere with verified reliability simply by specifying the names of desired packages. **Training:** workshop.bioconductor.org is a Kubernetes-based system supporting automated submission of teaching materials, production of workshop-specific containers with all analytic modules and data, and a Galaxy-backed deployment framework for teaching or self-guided learning. The system provides a growing library of conference workshops that have been conducted all over the world.

LEVERAGING REPRESENTATIONS OF MULTI-SPECIES NETWORKS AND ONTOLOGIES TO IMPROVE GENE CLASSIFICATION

Keenan Manpearl[1], Remy Liu[2], Christopher Mancuso[3], Arjun Krishnan[1]

[1]University of Colorado, Anschutz, Department of Biomedical Informatics, Aurora, CO, [2]Michigan State University, Department of Computational Mathematics, East Lansing, MI, [3]University of Colorado, Anschutz, Department of Biostatistics and Informatics, Aurora, CO

Computationally predicting the role of genes in functions, phenotypes, and diseases (i.e. gene classification) is a key problem in biomedical research. Large-scale molecular interaction networks are useful for addressing this problem because a gene's role depends on the other genes it is working with. Thus, these networks lend themselves to the application of the 'guilt-by-association' principle to predict the role of a gene based on the roles of other genes in the same network neighborhood. Previous network-based gene classification approaches have successfully built machine learning models to predict genes related to a biological term (i.e. a function or disease), but building one model per term requires sufficient training examples of previously known associated genes and fails to account for relationships between terms. These predictions can be improved by using the hierarchical relationships between terms that are captured in biomedical ontologies such as the Gene Ontology (GO). Nevertheless, as ontologies are abstract and likely incomplete, directly mapping between genes and terms still limits our discoveries. More recent approaches have leveraged this relational information by learning a mapping between genes (in a network) and an embedded representation of an ontology derived from the textual descriptions of terms or the structure of the ontology graph. These approaches also allow for classifying genes to terms that have no or very few known gene associations. However, these approaches have not been comprehensively tested for diverse networks and ontologies. Further, they all use species-specific networks, which are noisy and incomplete. Using networks from multiple species can leverage high evolutionary conservation across species to improve gene classification within a species and knowledge transfer across species. We aim to build a unified neural network that can learn a mapping between a joint multi-species network representation and an embedding of a biomedical ontology graph to improve prediction of gene -function, -phenotype, or -disease associations. Our work provides a quantitative evaluation of the existing GO embedding methods clusDCA and HiG2Vec for their ability to accurately represent graph structure and an evaluation of these methods applied to the MONDO disease ontology and uPheno phenotype ontology. We also compare our model's ability to predict gene function against the STRING2GO method, which makes a direct prediction for each gene-term pair and the joint embedding methods clusDCA and HiG2Vec, and expand these predictions to include disease and phenotype associations.

TOOLS AND WORKFLOWS ENABLING SCALING OF GENOME ASSEMBLY ACROSS THE TREE OF LIFE

Shane A McCarthy[1,2], on behalf of Tree of Life Programme[1], and Darwin Tree of Life Project[1]

[1]Wellcome Sanger Institute, Tree of Life Programme, Hinxton, United Kingdom, [2]University of Cambridge, Department of Genetics, Cambridge, United Kingdom

As part of the Tree of Life Programme at the Wellcome Sanger Institute we have been developing a “Genome Engine” able to take samples from any Eukaryotic specimen and produce a high quality chromosome level reference genome. As of August 2023 we have produced and openly released assemblies for over 1,000 diverse species from taxa ranging across over 400 families, including the largest genome assembled to date - Viscum album (mistletoe) at over 90G. The majority of genomes have been generated in collaboration with the Darwin Tree of Life project aimed at producing genomes for all Eukaryotic species in Britain and Ireland, with additional challenging genomes coming from the Aquatic Symbiosis Project aiming to sequence 500 aquatic symbiotic systems. We also collaborate with the Vertebrate Genomes Project and European Reference Genome Atlas under the coordination of the Earth BioGenome Project. To enable the scaling of the assembly part of the Genome Engine, we have developed tools and workflows to tackle the range of data and genome characteristics that we encounter. Purge-dups is an efficient tool to identify and remove excess haplotypic duplication from a primary assembly. YaHS is a state of the art Hi-C scaffolder that, in combination with the PreText suite for visualising and manipulating Hi-C contact maps, has helped reduce the burden of manual curation after the automated assembly generation. MitoHifi consistently produces animalia mitochondrial assemblies from PacBio HiFi data, and we have an in-development tool that addresses organellar assemblies for plants and other taxa. We are also developing strategies to handle polyploid genomes and to assemble cobionts and metagenomic communities found sequenced with our samples. To empower the community to use these tools with their own data, we have been developing standardised workflows written in nf-core compatible Nextflow to cover the whole informatics pipeline, including QC, genome assembly, cobiont/contamination checking, curation, and post-genome analysis. These are all publicly available under the sanger-tol GitHub organisation (https://github.com/sanger-tol).

UNDERSTANDING AND MITIGATING AMPLIFICATION BIASES IN SINGLE-CELL DNA SEQUENCING FOR ACCURATE GENOTYPE CALLING

Aleksei Mikhalchenko[1], Nuria Marti Gutierrez[1], Daniel Frana[1], Paula Amato[1,2], Shoukhrat Mitalipov[1]

[1]Oregon Health & Science University, Center for Embryonic Cell and Gene Therapy, Portland, OR, [2]Oregon Health & Science University, Division of Reproductive Endocrinology, Department of Obstetrics and Gynaecology, Portland, OR

OBJECTIVE: Single-cell DNA sequencing, a pivotal tool in reproductive genomics, relies on whole genome amplification (WGA) to obtain sufficient DNA input, yet WGA introduces amplification bias in heterozygous sites, risking loss of heterozygosity. Our research addresses this issue and its implications, emphasizing the role of expertise at the intersection of library preparation and genome informatics.

METHODOLOGY: We conducted whole genome sequencing on skin fibroblasts from a female proband to identify heterozygous sites, selecting 608 loci distributed across all chromosomes for a custom sequencing panel. To generate embryos genetically identical to the donor fibroblasts, we employed somatic cell nuclear transfer, yielding 11 cloned cleavage-stage embryos. DNA from individual blastomeres and donor fibroblasts underwent WGA, followed by sequencing (143X coverage) and variant analysis for assessing allelic dropout (ADO) among captured loci. We explored mitigation strategies, including a transition from traditional probabilistic germline variant calling to custom thresholds for heterozygous call assignment. We also explored pooling DNA from multiple blastomeres.

RESULTS: Our custom panel targeting 608 loci achieved 100% accuracy with heterozygous genotype calls in bulk DNA. However, when applied to single blastomeres of cloned embryos and donor fibroblasts subjected to WGA, germline variant calling based on Bayesian methods revealed significant rates of false-positive homozygous calls (27% in cleaving embryos, 37% in G1-fibroblasts), indicating ADOs. Reassigning heterozygous genotypes for sites with a variant allele frequency ≥5% reduced false-positive homozygous calls to 18% and 25% in embryos and G1-fibroblasts, respectively. While pooling single-cell samples (groups of 2, 3, and 6 cells) reduced false-positive homozygosity, it may not be suitable for scenarios expecting embryo mosaicism.

CONCLUSION: Our findings emphasize the significant distortion in genotype calling accuracy introduced by WGA biases in single-cell DNA sequencing. We caution against the use of end-to-end germline variant calling pipelines due to frequent distortion of original 50/50 allele frequency.

IMPACT STATEMENT: Understanding and mitigating WGA biases will empower bioinformaticians, researchers, and clinicians to make more informed decisions when analyzing single-cell DNA sequencing data, ultimately impacting reproductive genomics and precision medicine.

QUALITY ASSESSMENT OF HUMAN SPLICE SITE ANNOTATION BASED ON CONSERVATION IN 470 SPECIES

Ilia Minkin[1,3], Steven L Salzberg[1,2,3,4]

[1]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [3]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [4]Johns Hopkins School of Medicine, Department of Genetic Medicine, Baltimore, MD

The annotation of the human genome is one of the most fundamental resources in genomics research. However, years after assembling the genome itself, its annotation remains inaccurate due to the noisy nature of technologies used to sequence and assemble RNA, and to noise inherent in the transcription itself. Existing genome annotations of the same human reference genome often contain contradictory information and do not agree on basic statistics like the number of genes. Hence, refinement of the human genome annotation is an important challenge.

A promising approach to address this problem is using evolutionary conservation information. In theory, functional elements should be conserved in many species while mis-annotated ones would not be. However, finding errors based on conservation alone is challenging. For example, a supervised machine learning approach would require a “gold standard” training dataset which is difficult to obtain. This problem was recently addressed by creating MANE, a manually curated consensus dataset between two popular gene catalogs, RefSeq and GENCODE. In addition, the USCS Genome Browser team published a whole-genome alignment of 470 species, which gives an opportunity to study conservation at unprecedented scale.

Given these resources, we studied patterns of conservation of splice sites in human annotation. We found that both donor and acceptor splice sites from protein-coding genes in MANE are consistently conserved across more than 400 species. Furthermore, we studied splice sites from RefSeq, GENCODE and CHESS 3 that are not present in MANE, from both protein-coding genes and lncRNAs. To do so, we trained a logistic regression model that distinguishes between a conservation pattern exhibited by a splice site from MANE and a randomly generated one from a sequence not under selection. We found that up to 30% of splice sites from protein-coding and 40% of the “extra” splice sites outside of MANE exhibit conservation patterns closer to random sequence as opposed to highly-conserved splice sites from MANE.

Our study highlights potentially erroneous splice sites that might require further scrutiny. In addition, the striking pattern of conservation in many species exhibited by splice sites from MANE is of interest in itself. With the advent of large-scale whole-genome alignments, these patterns can be used as powerful aids for annotating genomes of other species that are still in nascent stages.

OTB (ONLY THE BEST GENOME ASSEMBLY TOOLS): A PHASED GENOME ASSEMBLY NEXTFLOW PIPELINE

David C Molik[1], Amanda Stahlke[2]

[1]USDA Agricultural Research Service, Arthropod-borne Animal Diseases Research Unit, Manhattan, KS, [2]Colorado Mesa University, Department of Biological Sciences, Grand Junction, CO

Phased genomes, which determine the DNA sequence of both chromosomes of a diploid organism, serve as valuable tools for analyzing genetic variation in populations [Snyder 2015]. In agricultural research, where the evolution and spread of resistance genes can impact outbreaks of insect pests, understanding genetic variation is of particular importance [Leftwich 2015]. There are numerous applications for a substantial quantity of phased arthropod genomes [Tewhey 2011]. However, the continuous generation of new genomes poses several data management challenges, especially in creating and storing genome assemblies. To address this, we have introduced a new HiC/HiFi phased genomics assembly pipeline, named 'Only The Best' (otb). Our goal is to reduce the time spent on data organization, bioinformatic tool installation and calibration, and analysis. Additionally, we have developed scripts to facilitate data movement and labeling for archiving.

The implementation of this environment has significantly reduced the time required to produce a usable genome. By carefully managing data and standardizing processes, we have further reduced team effort in genome creation. otb is built using Nextflow [Di Tommaso 2017], accessible through Bash scripting with a Singularity environment [Kurtzer 2017]. It conducts software and environment checks to debug problems and ensure smooth operation. Utilizing Nextflow offers parallel task execution and efficient compute resource management. Leveraging Singularity ensures a consistent compute environment for all otb users. An additional benefit of Singularity is the ability for users within the same environment to share containers, reducing software duplication across high-performance computing clusters (HPC).

otb's process begins with environment checks, ensuring container availability. It executes analyses using Nextflow, filtering sequencing data, assembling with HiFiASM and hicstuff, and optionally running Busco. Post-assembly, Shhquis.jl clusters and orients contigs. otb offers three genome assembly polishing methods: Merfin, DeepVariant, and 'Simple.' 'Simple' employs error-corrected reads for scaffolding. Users can optionally run KMC/Yahs and Busco. otb automates multiple assembly steps, simplifying the process of genome assembly.

https://github.com/molikd/otb

[Di Tommaso 2017] doi: 10.1038/nbt.3820
[Kurtzer 2017] doi: 0.1371/journal.pone.0177459
[Leftwich 2015] doi: 10.1111/eva.12280
[Tewhey 2011] doi: 10.1038/nrg2950
[Snyder 2015] doi: 10.1038/nrg3903

CELL-TYPE DECONVOLUTION WITH LONG-READ, SINGLE MOLECULE METHYLATION

Luke B Morina[1], Courtney Hall[1], Jessica Hosea[1], Roham Razaghi[1], Winston Timp[1,2]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, Molecular Biology and Genetics, Baltimore, MD

Single-molecule, long-read sequencing technologies have enormous potential in epigenetic applications; unlike traditional sequencing-by-synthesis technologies, single-molecule platforms (PacBio/ONT) distinguish covalently modified nucleotides directly. This base modification data is stored in the SAM/BAM file as standardized tags (MM and ML). We can take advantage of the long read lengths – tens to hundreds of kilobases – to phase methylation across >85% of the genome, enabling allele-specific differential methylation analysis. However, analysis is limited by the heterogeneous nature of tissue, with a complex distribution of cell types within the sample. This can lead to either different cell mixtures incorrectly identified as sample specific differentially methylated regions (DMRs), or actual differential methylation lost as it only occurs within a subset of cells. Profiling the cell-type distribution in a tissue sample will allow for improved DMR identification among other downstream epigenetic analysis, especially in known regions of cell-type-specific methylation as identified by established short-read methods. Existing single cell sequencing methods attempt to profile this through oligo based cell barcoding, but these methods often result in sparse data with many gaps per cell. Using long, single molecules we can interrogate the heterogeneous distribution of cellular states within tissue samples.

We are developing a toolset to better leverage the long single-molecule nature of these datasets. We are currently validating this approach by applying it to an in silico mix of well-studied human blood and colon cells. We calculate Euclidean distance between reads using either the canonical discrete (unmodified/uncertain/modified) status or the modification probabilities from the ML BAM tag of individual CpG sites; low-confidence and missing modification calls are taken into account in this processing. We are also able to smooth CpGs within a genomic window size on a per-read basis, removing singleton CpGs and applying Gaussian filtering over the CpG modification probabilities to remove technical and biological noise. By applying hierarchical clustering of individual DNA molecules using their modification state we can test the efficiency of molecular deconvolution, using the known origin of the reads (lymphoblastoid HG002 versus colon epithelial HCT116) as a ground truth. We then applied these same strategies to brain data (from the CARD consortium) and human whole blood, finding that at known differentially methylated regions (e.g. TREM2 for glia vs. neurons) we can parse out different cell types Ultimately, we hope to apply these preprocessing strategies to improve the viability of single-molecule methylation clustering and its applications in epigenetic analysis.

REDCARPET: A TOOL FOR RAPID RECOMBINATION DETECTION AMIDST EXPANDING GENOMIC DATABASES

Ahmed M Moustafa[1,2,3], Erin Theiller[1], Arnav Lal[4], Andries Feder[5], Apurva Narechania[6], Paul J Planet[2,5,6]

[1]Division of Gastroenterology, Hepatology and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA, [2]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, [3]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Philadelphia, PA, [4]School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, [5]Division of Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA, [6]Sackler Genomic Institute, American Museum of Natural History, New York, NY

Genomic recombination plays a pivotal role in enhancing biological diversity among microbial populations, facilitating their adaptation to various environments, hosts, and niches. Traditional recombination detection methods are computationally intensive, primarily relying on alignment of genomic sequences, phylogenetic analyses, and comparative techniques. This approach becomes especially challenging given the exponential increase in available whole-genome sequences. Addressing this challenge, we introduce Redcarpet, an innovative, alignment-free, database-driven tool designed for recombination detection. This technique leverages the distribution of exact protein matches within a genomic database, building upon the WhatsGNU algorithm—a tool focused on exact proteomic compression. Redcarpet inputs a single query genome and, for each encoded protein, identifies genomes in the database with exact protein sequence matches. It then computes the Jaccard similarity coefficient between these genome sets based on pairwise protein comparisons. Such operations operate under the assumption that genes with identical sequences are more likely to be present in similar genome sets due to shared evolutionary histories. Redcarpet's results are visualized as a 2-D heatmap, highlighting recombination regions and enabling the identification of recombination tracts. Furthermore, probabilistic changepoint analysis can pinpoint likely recombination breakpoints. When applied to known recombination events in *Staphylococcus aureus* and *Klebsiella pneumoniae*, Redcarpet's efficiency was evident. Beyond recombination detection, it can also deduce the probable origins of genomic segments and define a genomic "core" for subsequent phylogenetic analyses. Overall, Redcarpet can be used to rapidly identify recombination tracts in any species that has a large database of genomic sequences.

IMPROVING THE ACCURACY OF MIRO-VARIANT DETECTION IN WHOLE GENOME SEQUENCING DATA

Shandukani Mulaudzi, Maximillian Marin, Maha Farhat

Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Background: Unfixed or micro-variants present within a population at lower mapped read frequencies (typically <75%) present an opportunity for early diagnosis of drug resistance (DR) in *Mycobacterium tuberculosis* (*Mtb*) and other pathogens. Micro-variants are also applicable in somatic mutation detection in human cancers and understanding genetic diversity in microbiome analysis. In addition to true micro-variation, another common reason for observing unfixed variants in short-read sequencing data is read mapping error due to sequence homology. Existing tools for micro-variant detection have not been specifically benchmarked in genomic regions of low mappability.

Methods: Here we benchmark the accuracy of three methods including LoFreq, VarScan and BinoSNP against MixedVar, a new method we developed that balances the accuracy of site-specific micro-variant calling with computational time, and controls for regions of extreme coverage as an approach to reducing false positives in regions of homology. Two Illumina simulation tools were used to simulate a total of 3,800 paired-end *Mtb* sequencing runs each with 20 mutations in DR genes and 20 mutations in low mappability regions. We studied the effect of the following parameters: (1) Mutational position: 10 different sets of 40 mutations, (2) Sequencing depth ranging between 50x and 700x, and (3) Mutant allele frequency ranging between 1% and 50%. Each pair of Mtb reads underwent read trimming (Prinseq), alignment to the *Mtb* H37Rv reference genome (BWA-MEM) and duplicate read removal (Picard). MixedVar, LoFreq and VarScan were then run on the resulting alignments.

Results: BinoSNP was computationally prohibitive as it was estimated to take ~300 hours to be run on the whole *Mtb* genome and was dropped from the benchmarking exercise. Across simulated data sets, MixedVar and LoFreq achieved an average precision of 0.99 and 1.00 and average recall of 0.72 and 0.76 respectively. VarScan was inferior in performance, achieving precision and recall of 0.37, and 0.07 respectively. MixedVar demonstrated higher recall than LoFreq in low mappability regions, especially at depths of 200x and above (0.67 versus 0.62 respectively, t-test for difference in number of true positives detected with p-value$<10^{-3}$) at the expense of a slightly higher false positive rate (<1 per genome for MixedVar versus 0.0 for LoFreq, t-test p-value$<10^{-10}$).

Conclusion: MixedVar shows comparable to improved micro-variant calling accuracy in *Mtb* over existing methods. Future directions include the use of local genome assemblies generated with long-read sequencing data to improve micro-variant recall in regions of low mappability.

LANCET2: IMPROVED PERFORMANCE AND GENOTYPING OF SOMATIC VARIANTS USING LOCALIZED GENOME GRAPHS

Rajeeva Lochan Musunuri[1], Bryan Zhu[1], Dickson Chung[2], Shreya Sundar[2], Adam Novak[2], Timothy Chu[1], Jennifer Shelton[1], Nicolas Robine[1], Giuseppe Narzisi[1]

[1]New York Genome Center, Computational Biology, New York, NY,
[2]University of California, Santa Cruz, Genomics Institute, Santa Cruz, CA

One of the central challenges in cancer genomics is the ability to accurately detect somatic mutations in heterogeneous tumors, and precisely determine their clonal origin and evolution. This fundamental knowledge is central to the discovery of new cancer therapies. In recent years, reductions in the cost of whole-genome sequencing have enabled researchers to address these questions in unprecedented detail. However, indels and genomic rearrangement can create complex events that defy the traditional linear reference representation and become more and more challenging to detect and inspect with traditional read alignment tools. Graphical structures are becoming increasingly more popular to encode variants of genomic sequences from multiple related samples, but these developments have been limited so far to the analysis of germline variants.

Towards addressing these shortcomings, we have developed Lancet, and its successor Lancet2 (https://nygenome.github.io/Lancet2/), a somatic variant caller which leverages local assembly and joint analysis of tumor-normal paired data using region-focused colored de Bruijn graphs. The assembly graphs built by Lancet are small-scale sequence graphs that represent the local genome structures of the tumor and normal samples. Lancet2 is a complete redesign with focus on improved performance, software maintainability, and genotyping accuracy compared to the original tool. We will present our recent efforts, including: (1) extensive benchmarking of Lancet2 against state-of-the-art methods using multiple matched tumor/normal pairs from high-depth sequencing studies; (2) improved genotyping approach using multiple sequence alignment of graph haplotype paths and re-alignment of reads to count support; (3) GFA support to export and visualize the Lancet2 graph using specialized visualization software; and (4) enhancement of the Sequence Tube Map visualizer to allow the inspection of somatic variants within the Lancet2 graphs using elegant tube-map visualizations.

EXPLORING GENE EXPRESSION PROPERTIES OF THE HUMAN BRAIN WITH BITHUB

Urwah Nawaz[1], Kieran Walsh[2], Gavin Sutton[2], Jozef Gecz[1], Irina Voineagu[2]

[1]The University of Adelaide, Adelaide Medical School, Robinson Research Institute, Adelaide, Australia, [2]University of New South Wales, School of Biotechnology and Biological Sciences, Sydney, Australia

Large-scale transcriptomic consortia such as GTEx and BrainSpan have been instrumental in characterizing gene expression in the human brain across developmental stages, brain regions and cell-types. Despite this wealth of data, it is currently difficult to extract and compare gene expression information across these datasets. The usability of publicly available gene expression data is often limited by the availability of high-quality, standardized biological phenotype and experimental condition information. Additionally, while some datasets are accompanied by data access portals (GTEx), others require significant efforts to retrieve and explore the data. Consequently, comparison across datasets, which is crucial for determining whether observations are replicable, is cumbersome and thus rarely carried out.

We introduce Brain Integrative Transcriptome Hub (BITHub), a web tool which allows integrative exploration of gene expression across large-scale transcriptome consortia of the post-mortem human brain (https://voineagulabunsw.github.io/BITHub/ - currently in development). Curated datasets include RNA-seq from GTEx, PsychENCODE, BrainSeq, BrainSpan, the Human Developmental Biology Resource (HDBR) and FANTOM5; and several large-scale single-nucleus RNA-seq, such as the Human Cell Atlas and recently sequenced tissue samples from HDBR.

Our resource includes 6,933 samples from 2,933 donors spanning over 21 developmental stages,; and contains the expression profiles of 49 different brain regions. We have performed rigorous curation and harmonization of metadata associated with each dataset, providing consistent nomenclature of brain regions and developmental stages; performed cellular deconvolution on bulk RNA-seq datasets; and used applied a mixed-linear model framework to prioritize drivers of variation across all datasets. Users are able to explore the gene expression properties of their gene of interest through multiple interactive plots that can be populated according to user-selected brain regions, cell-types, age intervals and other technical or biological covariates of interest. BITHub also allows users to compare the gene expression profiles across datasets through scatterplots by providing mean expression values that have been log2 transformed and scaled as z-scores. Overall, BITHub facilitates the exploration of gene expression analyses of several large-scale databases and improves their reproducibility as individual researchers do not have to manually curate the sample metadata.

CADD V1.7: USING PROTEIN LANGUAGE MODELS, REGULATORY CNNs AND OTHER NUCLEOTIDE-LEVEL SCORES TO IMPROVE GENOME-WIDE VARIANT PREDICTIONS

Max Schubach[1], Thorben Maass[2], Lusine Nazaretyan[1], Sebastian Röner[1], Martin Kircher[1,2]

[1]Exploratory Diagnostic Sciences, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany, [2]Institute of Human Genetics, University Hospital Schleswig-Holstein, University of Lübeck, Lübeck, Germany

Machine Learning-based scoring and classification of genetic variants aids the assessment of clinical findings and is employed to prioritize variants in diverse genetic studies and analyses. Combined Annotation-Dependent Depletion (CADD) is one of the first methods for the genome-wide prioritization of variants across different molecular functions and has been continuously developed and improved since its original publication. Here, we present our most recent release, CADD v1.7. We explored and integrated new annotation features, among them state-of-the-art protein language model scores (Meta ESM-1v), regulatory variant effect predictions (from sequenced-based convolutional neural networks) and sequence conservation scores (Zoonomia). We evaluated the new version on data sets derived from ClinVar, ExAC/gnomAD and 1000 Genomes variants. For coding effects, we tested CADD on 31 Deep Mutational Scanning (DMS) data sets from ProteinGym and, for regulatory effect prediction, we used saturation mutagenesis reporter assay data of promoter and enhancer sequences. The inclusion of new features further improved the overall performance of CADD. As with previous releases, all data sets, genome-wide CADD v1.7 scores, scripts for on-site scoring and an easy-to-use webserver are readily provided to the community.

GENOMIC CHARACTERIZATION AND GLOBAL CONTEXTUALIZATION OF ESBL-PRODUCING *E. COLI* FROM PEDIATRIC PATIENTS IN QATAR

Matthew Nguyen[1,2,3], Clement Tsui[4,5,6], Patrick Tang[7,8], Andres Perez-Lopez[7,8], William Hsiao[3]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]University of British Columbia, Bioinformatics Graduate Program, Vancouver, Canada, [3]Simon Fraser University, Faculty of Health Sciences, Vancouver, Canada, [4]National Centre for Infectious Diseases, Infectious Diseases Research Laboratory, Singapore, Singapore, [5]Nanyang Technological University, Lee Kong Chian School of Medicine, Singapore, [6]University of British Columbia, Vancouver, Faculty of Medicine, Vancouver, Qatar, [7]Sidra Medicine, Department of Pathology, Doha, Qatar, [8]Weill Cornell Medicine-Qatar, Department of Pathology and Laboratory Medicine, Doha, Qatar

Extended spectrum beta-lactamase (ESBL) producing *Enterobacteriaceae* such as *E. coli* and *K. pneumoniae* are found globally in hospital and community settings and represent a major healthcare challenge due to their limited effective treatment options. Several lineages are considered high-risk clones with an enhanced ability to spread and persist and increased pathogenicity and antimicrobial resistance (AMR). The prevalence of ESBL-producing *Enterobacteriaceae* in Qatar has steadily increased in the last two decades. Whole-genome sequencing can be used for the genomic characterization of these Qatari organisms; however, previous studies has focused only on core genes. Moreover, the genetic relationship of Qatari *E. coli* lineages to globally circulating high-risk clones has not been studied. This work presents an in-depth genomic characterization of ESBL-producing *E. coli* from a pediatric hospital in Qatar, with a focus on characterizing mobile genetic elements and plasmids. Through a pangenome approach, both the core and accessory genome can be used to better understand the genetic diversity of the different strains present in Qatar. This work is also the first study to globally contextualize Qatari isolates against globally circulating lineages of ESBL-producing *E. coli*. By performing large-scale clustering of Qatari isolates and globally circulating lineages, this study reveals the presence of a putative emerging high-risk clone that may be endemic to Qatar, ST8881. Although the lineage was sampled multiple times in Qatar, it had not been previously studied nor sampled at a significant prevalence. Through phylogenetic reconstruction from different genomic features, ST8881 is genetically distinguishable from other Qatari and global lineages by its mobile genetic element, virulence factor, and AMR marker profiles. ST8881 is likely a descendant of the globally circulating high-risk clone ST648. ST8881 is also characterized by factors affecting the stability of AMR genes: chromosomal integration of beta-lactamase genes, as well as the presence of a multiple-replicon plasmid encoding for five AMR genes. With Qatar being a major international travel hub, a high-risk clone could be easily exported; therefore, it is crucial to understand the genetic landscape of ESBL-producing *E. coli* in Qatar. The discovery and characterization of ST8881 provides a strong basis for future studies of AMR in Qatar.

EVOLUTIONARY INSIGHTS INTO PRIMATE SEX CHROMOSOMES: SHARING AND GENE CONTENT OF PALINDROMES

Karol Pal[1], Robert Harris[1], Monika Cechova[2], Sergey Koren[3], Sergey Nurk[4], Huiqing Zeng[1], Arang Rhie[3], Melissa A Wilson[5], Brandon D Pickett[3], Brendan J Pinto[5], Prajna Hebbar[6], Mark Diekhans[6], Benedict Paten[6], Evan Eichler[7], Adam M Phillippy[3], Kateryna D Makova[1]

[1]Pennsylvania State University, Department of Biology, University Park, PA, [2]University of California, Department of Biomolecular Engineering, Santa Cruz, CA, [3]NIH, NHGRI, Bethesda, MD, [4]Oxford Nanopore Technologies Inc., Oxford, United Kingdom, [5]Arizona State University, Center for Evolution and Medicine, Tempe, AZ, [6]University of California, Genomics Institute, Santa Cruz, CA, [7]University of Washington, Department of Genome Sciences, Seattle, WA

Using the methods employed for generating the gapless assemblies of the human female genome and of the human Y chromosome, the Telomere-To-Telomere (T2T) Primates Consortium is building complete genome assemblies for a number of primate species. As part of this effort, we have recently assembled the T2T sequence of the X and Y chromosomes for five great apes (bonobo, chimpanzee, gorilla, Bornean and Sumatran orangutans) and one lesser ape, the siamang gibbon—all are endangered species. The distinct mating patterns of all studied species could be associated with the architecture of the quickly evolving Y chromosome. These new assemblies have allowed us to elucidate the evolution of complex chromosomal structures such as palindromes—long inverted repeats known to undergo intrachromosomal gene conversion between their arms. For the first time we were able to identify palindromes on the sex chromosomes and annotate protein-coding genes which they carry. We show the full extent of divergence and sharing of palindromes on the X and Y chromosomes in the studied species. The X chromosome palindromes displayed a high degree of homology across species. Gene density in X palindromes shared among species is higher than in any other region on the X chromosome (29-47 genes/Mb). Most of the genes located in such palindromes were housekeeping. In contrast, the Y chromosome palindromes displayed a lower degree of homology, with only closely related species sharing a substantial number of palindromes. Gene density at Y palindromes was only 3-6 genes/Mb. We found an expansion of particular gene families in Y palindromes of some ape genera. For instance, we detected an expansion of the CDY family in Pongo, and of the RBMY family in Pan—both gene families have functions in spermatogenesis. Additionally, we detected movement and copy number expansion of several autosomal genes to palindromes on the Y chromosome (e.g., *KRT8* in Pan, and *PTPN13* in gorilla). The potential role of these genes in determining mating patterns and levels of sperm competition should be evaluated in further functional studies.

DIAGNOSIS OF OCULAR INFECTION USING NANOPORE METAGENOMIC SEQUENCING

Dongwoo Park*[1,2], Junwon Lee*[3], Hyun Goo Kang*[4], Han Jeong*[1,4], Sangwoo Kim[1,2], Min Kim[3]

[1]Yonsei University College of Medicine, Graduate School of Medical Science, Brain Korea 21 FOUR Project for Medical Science, Seoul, South Korea, [2]Yonsei University College of Medicine, Department of Biomedical Systems Informatics, Seoul, South Korea, [3]Institute of Vision Research, Department of Ophthalmology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, South Korea, [4]Institute of Vision Research, Department of Ophthalmology, Severance Eye Hospital, Yonsei University College of Medicine, Seoul, South Korea

Early identification of infection and pathogen identification is essential to the visual recovery of patients with intraocular inflammation, however, there remain challenges in the accurate classification via clinical presentation, securing sufficient amount of ocular samples, and the single species-level verification of pathogens that requires a high cost and labor. Here, we developed a novel bioinformatics method for ocular infection using Oxford Nanopore metagenomic sequencing that addresses previous hurdles by detecting DNAs from all potential pathogens in a single test. Our algorithm distinguishes infectious patients from noninfectious, and identify the pathogens in a species-level with a confidence score. In a test on the intraocular aqueous and vitreous samples from 28 patients (18 infected, and 10 uninfected), our algorithm marked overall 89% accuracy (sensitivity 94%, specificity 80%). Moreover, we could identify the potential pathogens, including cytomegalovirus, *toxoplasma gondii*, *mycobacterium tuberculosis*, *streptococcus sanguinis* and *clostridium septicum*, which were further validated by clinicians, and could be utilized in the treatment procedure. We expect that sequencing-based diagnosis of ocular infection would overcome the shortage of conventional diagnostic tests and make transition between bedside to bench with higher accuracy and lower cost and time.

CONTEXTSV: A NOVEL COMPUTATIONAL METHOD FOR CALLING STRUCTURAL VARIANTS AND INTEGRATING INFORMATION ACROSS SEQUENCING PLATFORMS

Jonathan E Perdomo[1,2], Kai Wang[2,3]

[1]Drexel University, School of Biomedical Engineering, Science and Health Systems, Philadelphia, PA, [2]Children's Hospital of Philadelphia, Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Philadelphia, PA, [3]University of Pennsylvania, Department of Pathology and Laboratory Medicine, Philadelphia, PA

Structural variants (SVs) are genomic alterations >50 bp which form the largest source of human genome variation. SV identification and characterization are important for gaining insight into the genetic basis of diseases: Identifying SVs associated with specific clinical phenotypes empowers clinical diagnoses and enables researchers to identify potential molecular mechanisms for these genomic alterations. Emerging long-read sequencing technologies such as the Oxford Nanopore (ONT) and Pacific Biosciences (PacBio) platforms provide the resolution required to resolve larger and more complex SVs. Nevertheless, variable error rates in these technologies possibly result in a high false positive rate and low robustness for SVs detected using only long-read sequencing data. The rich repertoire of available technologies, such as Illumina short-read sequencing and Bionano optical mapping, can be leveraged to resolve these limitations. Here we introduce ContextSV, a novel structural variant calling method that uses a hybrid approach to achieve high-accuracy, robust SV calling: Long read sequencing data is used to identify SV candidates, while short read sequencing provides high-accuracy sites for resolving breakpoints in complex SVs, and optical maps are used as long-range scaffolds for high-quality read assembly prior to running SV detection algorithms. To improve accuracy, we train a binary classification model which is used to score candidate SVs based on coverage and genomic context, which are key features for SV validation. We then filter likely false positives by thresholding SVs based on score. Finally, we plan to incorporate support for pangenome graph reference formats in ContextSV: Pangenome graph structures can better represent common haplotypes in the human population relative to a single linear reference genome, and thus they would form a more comprehensive reference for identifying SVs. Large efforts led by projects such as the Human Pangenome Reference Consortium (HPRC) aim to release a pangenome graph reference representing a large, diverse set of human genome sequences, and thus there is a growing importance for future SV callers to provide support for these formats. In summary, our ContextSV method enables capturing large, complex SVs with high accuracy and robustness by leveraging information across multiple genome analysis technologies and using a machine learning model to compute confidence scores, while providing support for future developments in pangenome graph reference formats.

THE EFFECT OF DYNAMIC PRESSURE ON THE GENE EXPRESSION OF *DEINOCOCCUS RADIODURANS* R1

Cesar A Perez Fernandez[1,2], Lily Zhao[3,4], K.T. Ramesh[2,3,5], Jocelyne DiRuggiero[2,5]

[1]Universidad Privada Boliviana, Faculty of Engineering and Architecture, Santa Cruz, Bolivia, [2]Johns Hopkins University, Department of Biology, Baltimore, MD, [3]Johns Hopkins University, Department of Mechanical Engineering, Baltimore, MD, [4]Johns Hopkins University, Hopkins Extreme Materials Institute, Baltimore, MD, [5]Johns Hopkins University, Department of Earth and Planetary Sciences, Baltimore, MD

Impact events are common in space in the form of meteoroid and asteroid impacts. These events generate extreme pressure for a brief period in the colliding bodies. Dynamic pressure is important in the context of studying extraterrestrial life, as organisms need to withstand such impacts. Numerous experiments have been conducted to assess microbial survival under these conditions, but have not generated consistent conditions. We developed an experiment to evaluate survival and molecular response to impact events using *Deinococcus radiodurans* R1 as a model. Approximately,10^9 cells were filtered, covered with aluminum foil, and placed between two steel plates in a sandwich-like configuration. The specimen was impacted at 1.4 GPa and 2.4 GPa using a hypervelocity gun. Cells were recovered from the filter by vortexing. Survival rates were measured using plate count, and RNA was also extracted through mechanical lysis in liquid nitrogen and purification with Trizol. *D. radiodurans* had a survival rate of ~99% at 1.4 GPa, and ~70% for 2.4 GPa. The experiment seemed to affect RNA quality; however, yield and amount were enough to sequence the transcriptome suggesting our experiment is ideal for studying the dynamic pressure. Changes in differential expression indicated major changes in response to 2.4 GPa. We found that replication, repair, and defense mechanisms were up-regulated in genes involved in stress response such as *recA, ddrA, MazG, MazE,* and *uvrB*. Metabolic pathways such as transcription, translation, energy production, and transporters displayed a bimodal curve of differential expression indicating groups of genes that are up and down-regulated inside these categories. One of the most over-expressed transcripts was in the iron uptake via heme or siderophores after 2.4 GPa impact. It is important to remark on the high expression of the mobilome, which may be indicative of a DNA double-strand break. Manganese is an important cofactor in the adaptation to oxidative stress. We only detect differential expression in *MazG* and *MazE* as manganese-related enzymes, but not others like catalase and superoxide dismutases that are essential in cell response to irradiation. Our results point out that dynamic pressure can hinder metabolism, but it remains unclear whether it leads to oxidative stress

SEQUENCE-ASSOCIATED MECHANISTIC INSIGHTS OF DNA FRAGILITY

Patrick Pflughaupt, Aleksandr B Sahakyan

University of Oxford, MRC WIMM Centre for Computational Biology,
Radcliffe Department of Medicine, Oxford, United Kingdom

Genomic insertion and deletion alterations, which occur through the formation of DNA strand breaks, are the second most significant DNA modifications after point mutations. However, unlike point mutations, the various sequence-context dependencies of DNA strand breakpoints and the detailed regional variation of breakage propensities in the genome have not been thoroughly explored through cutting-edge computational means. This computational project aims to understand the genomic sequence dependencies in tissues experiencing different physiological, spontaneous and pathological processes, revealing the sequence-driven commonalities and differences across these processes leading to DNA strand breaks.

Our work identified how the DNA sequence context influences the propensity of a breakpoint appearing under different conditions, showing that the sequence context can be separated into three distinct short-, medium-, and long-range effects. Focusing on the short-range effect, we quantified the k-meric breakage propensities and revealed the relationship between the DNA sequence and various types of induced breakages. We also applied sequence-based feature engineering to datasets of chromatin, epigenetic, and structural features of the genome. By combining these sequence determinants, we built a robust machine learning engine that understands different mechanical, biological and physicochemical aspects of DNA fragility. Our high-quality sequence-based features demonstrated significant performance enhancement in machine learning, as compared to the usage of basic triplet counts in a certain range surrounding a breakpoint location. This work is currently being developed and applied towards a generalised DNA fragility model for cancer applications.

PRIORITIZATION OF FLUORESCENCE IN SITU HYBRIDIZATION (FISH) PROBES FOR DIFFERENTIATING PRIMARY SITES OF NEUROENDOCRINE TUMORS WITH MACHINE LEARNING

Lucas Pietan[1,2], Hayley Vaughn[1,3], James Howe[4], Andrew Bellizzi[5], Ben Darbro[1,3], Terry Braun[1,2,6], Tom Casavant[1,2,6,7]

[1]University of Iowa, Interdisciplinary Graduate Program in Genetics, Iowa City, IA, [2]University of Iowa, Department of Biomedical Engineering, Iowa City, IA, [3]University of Iowa, Stead Family Department of Pediatrics, Iowa City, IA, [4]University of Iowa, Healthcare Department of Surgery, Iowa City, IA, [5]University of Iowa, Department of Pathology, Iowa City, IA, [6]University of Iowa, Center for Bioinformatics and Computational Biology, Iowa City, IA, [7]University of Iowa, Department of Electrical and Computer Engineering, Iowa City, IA

Determining neuroendocrine tumor (NET) primary sites is pivotal for patient care. Among the approximate 170,000 NET cases in the United States, pancreatic NETs (pNETs) account for 17-20%, while small bowel NETs (sbNETs) constitute 50-55%, each demanding distinct treatment approaches. Numerous clinical tests can be performed to assist in identifying the primary sites of NETs, including blood tests, imaging, and histological assessments with immunohistochemistry and fluorescence in situ hybridization (FISH) tests. The diagnostic power and prioritization of FISH assay biomarkers for establishing the primary site has not been thoroughly investigated using machine learning (ML) techniques. We trained several ML models on FISH assay metrics from 85 sbNET and 59 pNET samples for primary site prediction. Exploring multiple methods for imputing missing data, the impute-by-median dataset coupled with a support vector machine using a radial basis function (SVM-RBF) model achieved the highest classification accuracy of 93.1% on a held-out test set and 85.4% on the initial training set, with the top importance variables originating from the ERBB2 FISH probe. Due to the greater interpretability of decision tree (DT) models, we fit DT models to ten dataset splits, achieving optimal performance with k-nearest neighbors (KNN) imputed data and a transformation to single categorical biomarker probe variables, with mean accuracy of 81.4% on the held-out test sets and 75.9% on the training sets. ERBB2 and MET variables ranked as top performing features in 9 of 10 DT models and in the full dataset model. The full dataset DT model utilizes the ERBB2, CDKN2A, and SMAD4 variables as splits with 3 of the 4 terminal nodes having predictive probabilities greater than 0.796. These findings offer probabilistic guidance for FISH testing, emphasizing the prioritization of the ERBB2 FISH probe in diagnosing NET primary sites.

< Back

INTEGRATIVE MODELING OF ACTIVITY, RESPONSIVENESS AND CONTACT (ARC) REVEALS ENHANCER-GENE CONNECTIONS USING SINGLE-CELL DATA

Wei-Lin Qiu*[1,2], Maya Sheth*[2,3,4], Rosa Ma[3,4], Andreas Gschwind[3,4], Anthony Tan[3,4], Hjörleifur Einarsson[1,2], Danilo Dubocanin[3], Evelyn Jagoda[2], Lars Steinmetz[3], Anshul Kundaje[3], Jesse Engreitz$[2,3,4], Robin Andersson$[1,2]

[1]University of Copenhagen, Department of Biology, Copenhagen, Denmark, [2]Broad Institute of MIT and Harvard, The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Cambridge, MA, [3]Stanford University, Department of Genetics, Stanford, CA, [4]Lucile Packard Children’s Hospital, BASE Initiative, Stanford, CA

A major challenge for understanding gene regulation and the genetic mechanisms of complex disease is connecting enhancers to their target genes. While several methods based on single-cell data have been proposed to this end, a thorough evaluation of their performances has been lacking. Based on single-cell multiome (RNA-seq + ATAC-seq) data, we here benchmark the predictions of 10 methods against more than 10,000 CRISPRi-tested enhancer-gene (E-G) pairs in human K562 cells. Our benchmarking results demonstrate that methods based solely on association between enhancer accessibility and gene expression (or promoter accessibility) in single-cell data generally fail to reach the performance of the scATAC-seq activity-by-contact (ABC) model (AUPRC: 0.471). To fully utilize the information provided in the single-cell multiome data, we derive an effective estimate of E-G responsiveness based on the Kendall correlation, a non-parametric method to examine the association between enhancer accessibility and target gene expression. We construct an integrated score, ARC-E2G, which combines the pseudo-bulk ABC score with the E-G responsiveness estimate and surpasses the performances of all other methods (AUPRC: 0.550). Finally, we extend the ARC-E2G score with additional scATAC-derived measurements in a classification framework. The final model demonstrates superior performance (AUPRC: 0.619). Altogether, we provide an effective method to predict E-G connections using single-cell data that will help to build genome-wide maps of human gene regulation and facilitate the interpretation of regulatory genetic variants associated with diseases.

*: equal contribution; $: corresponding authors

GENOMIC ANALYSIS OF NOVEL HYDROCARBONOCLASTIC *CHRYSEOBACTERIUM ORANIMENSE* STRAIN COTT, A PUTATIVE BIOREMEDIATION AGENT WITH MULTI-DRUG RESISTANCE AND ENZYMES FOR INDUSTRY

Amanda C Ramdass, Sephra N. Rampersad

The University of the West Indies, Life Sciences, St. Augustine, Trinidad And Tobago

Petrogenic hydrocarbons, especially oil-induced polycyclic aromatic hydrocarbons (PAHs), are priority pollutants based on toxicity, carcinogenicity, potential for human exposure and frequency of occurrence in the environment. The discovery and characterization of new microbes with the ability to degrade these hydrocarbons are important for efficient biodegradation. Bioprospecting of hydrocarbonoclastic microbes is ongoing since their enzymes are important biocatalysts used by numerous industries and they have biotechnological applications. The multibillion-dollar microbial enzyme market continues to grow due to the discovery of new enzymes, in particular those from exotic environments. Few bacteria in the genus *Chryseobacterium* have been implicated in petrogenic hydrocarbon utilization. While *C. hungaricum*, *C. nepalense* and *C. indoltheticum* have been reported as petroleum degraders, no genomic or bioinformatic studies have been conducted. In this study, a novel hydrocarbonoclastic *Chryseobacterium oranimense* strain COTT was isolated from chronically polluted crude oil soil with over a century of petrogenic hydrocarbon exposure. Comparative genomic analysis revealed that COTT has extraordinary genome plasticity suggesting that COTT has independently evolved its own repertoire of degradative enzymes and genetic features to survive in oil such as its unique repair mechanisms, social behavior, and stress tolerance. Its genome encoded several proteins of putative value including major classes of enzymes that are used in industrial bioprocesses; >200 oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases were identified. The COTT strain was multidrug resistant and its resistome consisted of 21 β-lactam resistance genes, 13 cationic antimicrobial peptide (CAMP) resistance genes, 5 vancomycin resistance genes, and 2 tetracycline resistance genes. The analysis revealed that the COTT strain is a potential candidate for various bioremediation and industrial applications and may be of clinical value. The genome sequence provides valuable genetic information for further study of *C. oranimense*.

DEDUCING THE EVOLUTION OF ALLORECOGNITION AND PRIMORDIAL IMMUNITY IN CNIDARIANS

Alberto M Rivera, Andy Baxevanis

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

The allorecognition gene complex (ARC) in cnidarians provides a model for understanding the evolution of the primordial immune system in bilaterians. The allorecognition system of the hydroid Hydractinia symbiolongicarpus plays an essential role in mediating self vs. non-self recognition in cnidarian colonies. This ability to recognize genetic identity is analogous to the immune system of bilaterian animals. Similarities have been noted between the allorecognition complex in Hydractinia and the major histocompatibility complex (MHC) in higher vertebrates, including humans. At the genomic level, we find that both complexes have contiguous stretches of similar, highly duplicated genes, strongly suggesting a homologous relationship between the ARC and MHC. Protein structural prediction methods, such as AlphaFold2, also indicate structural similarity between the proteins encoded within the Hydractinia ARC and the MHC. We are currently elucidating the structure of the ARC in a related hydrozoan species (Podocoryna carnea) to determine the degree of conservation of the allorecognition complex between Hydractinia and Podocoryna. Through the use of genome annotation software (BRAKER) and protein structural prediction (AlphaFold2), we identified a number of candidate Podocoryna allorecognition proteins. Phylogenetic approaches suggest strong homology between cnidarian allorecognition genes, as well as the conservation of synteny in this gene complex across cnidarians. Future comparative genomics studies will use these findings as a foundation for determining whether there is macrosynteny between the ARC and the bilaterian MHC.

ON-GOING SEQUENCING AND ANALYSIS OF A NEW TUMOR CELL LINE FOR DEVELOPMENT OF A GENOME IN A BOTTLE TUMOR/NORMAL BENCHMARK

Gail Rosen[1], Justin Wagner[2], Jennifer McDaniel[2], Andrew Liss[3], Justin Zook[1]

[1]Drexel University, EESI Lab, Philadelphia, PA, [2]National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD, [3]Massachusetts General Hospital, Surgery, Boston, MA

We describe results from initial efforts towards the first tumor/normal benchmark from the Genome in a Bottle (GIAB) consortium and detail on-going sequencing from a variety of technologies. We performed and analyzed PCR-free Illumina WGS and Hi-C sequencing on a new cell from a broadly-consented pancreatic ductal adenocarcinoma patient with matched normal tissue. We are now performing an array of PacBio HiFi, ONT ultralong and duplex, Bionano, Bioskyrb single cell WGS, along with other technologies from a large batch of the tumor cell line. We plan to make this data public as soon as possible after it is generated and are forming an open working group to develop a GIAB tumor/normal benchmark.

For initial characterization of the tumor cell line, we performed preliminary analysis using HiC for copy number and structural variant analysis and identified substantial aneuploidy common in pancreatic tumors. We found ~17 large inversions and translocations and 16 chromosomes with extensive loss of heterozygosity due to missing >30% of one copy, in addition to smaller duplications. We also analyzed CNVs in single cell sequencing and found that many of the large deletions are in most or all cells, while others are in only some cells. The tumor contains the common G12V mutation in KRAS, and interestingly the ~2 Mbp region containing this mutation is likely triplicated.

While GIAB has developed expertise on accounting for errors and biases in germline variant calling, we have started by assessing sources of error somatic variant calling. We started by aligning the Illumina PCR-Free data at coverages ~150x on normal and ~180x on the tumor to GRCh38 then ran Lancet, Mutect2, Strelka2, DRAGEN, and NeuSomatic. We used hap.py to compare Mutect2 calls to Strelka2 with default filtering and found that after excluding difficult regions using the GIAB stratifications more than 90% of SNVs and small indels agree. For Strelka2, the somatic SNVs fell in 4 primary classes: ~4000 in most cells in diploid regions, ~2000 in most cells in haploid (loss of heterozygosity) regions, ~400 in only some cells in haploid regions, and ~400 in only some cells in diploid regions. We categorized disagreements to find that many occur because of systematic sequencing errors, alignment errors, low frequency, and proximity to germline variants. Examining discordant SNVs in difficult regions, most are erroneous somatic variants in regions with loss of heterozygosity that also have mapping errors due to segmental duplications, sequences missing from GRCh38, and/or tandem repeats.

A NOVEL STRUCTURAL VARIANT DETECTION PIPELINE IN CANCER GENOMES: THE PERSONALIZED MATCHED-CONTROL REFERENCE-BASED APPROACH

Yoshitaka Sakamoto[1], Masahiro Sugawa[1], Ai Okada[1], Yotaro Ochi[2], Yosuke Tanaka[3], Yasunori Kogure[4], Kenichi Chiba[1], Wataru Nakamura[1], Junji Koya[4], Hiroyuki Mano[3], Seishi Ogawa[2], Keisuke Kataoka[3,5], Yuichi Shiraishi[1]

[1]National Cancer Center Research Institute, Division of Genome Analysis Platform Development, Tokyo, Japan, [2]Graduate School of Medicine, Kyoto University, Department of Pathology and Tumor Biology, Kyoto, Japan, [3]National Cancer Center Research Institute, Division of Cellular Signaling, Tokyo, Japan, [4]National Cancer Center Research Institute, Division of Molecular Oncology, Tokyo, Japan, [5]Keio University School of Medicine, Division of Hematology, Department of Medicine, Tokyo, Japan

Accurate detection of structural variants (SVs) in cancer genomes is crucial for understanding the mechanisms of cancer progression. SVs in human cancer genomes are typically identified by aligning long-read sequencing data to the established human reference genomes. However, reference genomes contain regions with ambiguous sequences, such as telomeres and centromeres, where the analysis of somatic variants was challenging. Recently, the combination of long-read sequencing technologies, such as PacBio HiFi (HiFi) and Oxford Nanopore Technologies Ultra-long read sequencing (ONT UL), along with trio or Hi-C phase information has enabled comprehensive analysis of the human genome spanning from telomere to telomere.

In this study, we aimed to detect various types of somatic variants involving centromeres and telomeres using HiFi, ONT UL, and Hi-C sequencing data obtained from tumor samples and matched-control samples. We developed a novel bioinformatical pipeline, named “personalized matched-control reference-based approach”. In this approach, we first constructed a patient-unique diploid reference genome through de novo genome assembly using HiFi, ONT UL and Hi-C sequencing data from a matched-control sample, employing either Verkko or Hifiasm. Subsequently, sequencing data from the tumor sample was aligned to the constructed patient-specific reference genome using a haplotype-aware approach. Finally, copy-number analysis and SV calling were performed.

We applied our method to two cancer cell lines (H2009 and HCC1954) and their respective matched-control cell lines (BL2009 and HCC1954BL). Copy-number analyses of cancer cell lines based on this approach enabled the reconstruction of predicted copy numbers for these cancer cell lines by haplotypes, as determined from karyotyping data. SV calling results obtained using our approach covered more than 85% of the SVs based on GRCh38 including the SVs validated by PCR in previous research. Furthermore, our approach identified SVs that were not detected by GRCh38-based approach, suggesting high sensitivity of our approach. Additionally, our approach successfully identified some centromere-involved SVs that were challenging to detect in previous analyses using either long-read or short-read sequencing data. In summary, this approach can provide a comprehensive view of the cancer genome structures and provide new insights into cancer genome studies.

INTEGRATING MULTIPLE TRANSCRIPTOME-BASED METHODS TO REPURPOSE DRUGS FOR INFECTIOUS DISEASES

Kewalin Samart[1,3], Amy Tonielli[2], Arjun Krishnan[3], Janani Ravi[3]

[1]University of Colorado School of Medicine, Computational Bioscience Program, Aurora, CO, [2]Michigan State University, Biomedical Laboratory Science, East Lansing, MI, [3]University of Colorado Anschutz Medical Campus, Department of Biomedical Informatics, Center for Health Artificial Intelligence, Aurora, CO

Infectious diseases (InfD) like tuberculosis (TB) are a leading cause of fatalities worldwide. Further, the accelerating rise of antibiotic resistance warrants new avenues for treatment. To address this critical need, there is growing interest in developing host-directed therapeutics. Given the immense time and cost of developing new drugs for humans, drug repurposing is a very appealing approach. We have developed a computational workflow to integrate transcriptome-based methods to repurpose FDA-approved drugs for InfD, starting with TB. Transcriptome-based drug repurposing works by finding drugs that reverse the differential gene expression pattern in the disease (i.e., showing negative disease-drug 'connectivity'). Over the past decade, many methods have been developed to improve the accuracy and robustness of quantifying such disease-drug reversal relationships (DOI: 10.1093/bib/bbab161; Samart, Tuyishime, et al., 2021). While they have been applied to complex diseases such as cancer, they are not widely used for InfD. We construct InfD expression signatures from public microarray and RNA-seq datasets, and drug candidates are prioritized by integrating multiple methods to find drug-disease reversals. Since gene expression data for each infectious disease have been generated as part of multiple studies, it is critical to account for heterogeneity that exists across these data when using them for drug repurposing. Therefore, our approach includes methods to address heterogeneous datasets: experimental platforms, conditions, infection stages, and cell/tissue types using (i) individual and aggregated disease signatures, (ii) baseline comparisons to identify appropriate cell lines, and iii) gene/pathway-level comparisons. We are currently applying these methods to Mycobacterium tuberculosis infection data to identify (i) top-ranked drug candidates/families and (ii) genes/pathways and modes of action underlying drug-InfD pairs for experimental validation. By leveraging multiple public datasets and methodologies, our drug repurposing approach will prioritize a robust list of drug candidates and mechanisms for validation, thus expediting and minimizing the cost of drug development against infectious diseases.

IDENTIFYING NICHE-SPECIFIC GENETIC ADAPTATIONS IN *ACINETOBACTER BAUMANNII*

Sydney Sanchez[1], Gisela Di Venanzio[2], Mario Feldman[2], Federico Rosconi[1], Juan C Ortiz-Marquez[1]

[1]Boston College, Department of Biology, Boston, MA, [2]Washington University School of Medicine, Department of Molecular Microbiology, St. Louis, MO

The alarming increase in infections caused by the nosocomial bacterial pathogen *Acinetobacter baumannii* (*Ab*), especially the multidrug-resistant strains (MDR), presents a pressing healthcare challenge. Despite its significant global impact, relatively little is known about the pathogenesis of MDR *Ab* compared to other major multidrug-resistant pathogens. For instance, *Ab* infects various niches, including the respiratory tract, blood, urinary tract, and soft tissues. However, its strains are typically seen as a homogenous group of opportunistic pathogens with indiscriminate virulence in critically ill hosts. As a result, research efforts often extrapolate findings from one *Ab* strain in a single model of infection to draw conclusions about *Ab* as a whole. Here we challenge this conventional perception of niche-indiscriminate virulence and delve into the genetic determinants underpinning niche specificity. Comparative genomics was employed to determine the genetic elements responsible for niche-specific adaptations and virulence. A total of 39 *Ab* strains were isolated and sorted into separate virulence groups: hypervirulent, low-virulent, respiratory, and uropathogenic. Genomes were prepped and sequenced (WGSeq) using both short-length (Illumina) and long-length (PacBio) sequencing technologies. High-quality genome assemblies were obtained by polishing WGSeq reads with Pilon. To identify functional elements and potential virulence factors, comprehensive pan-genome analyses are implemented after genome annotation. Niche-specific virulence determinants are recognized in our set of *Ab* strains from separate virulence groups. These genetic adaptations may be novel targets for innovative antibiotic-independent therapies and could help combat or prevent future *Ab* outbreaks. Additionally, our findings would help to identify and characterize genetically modern uropathogenic and respiratory *Ab* strains that can be broadly adopted by the research community to better investigate two leading manifestations of *Ab* disease.

BIOMEDICAL AND BIOLOGICAL APPLICATIONS OF MACHINE LEARNING USING GALAXY PROJECT

Michelle Savage[1], Michael Schatz[1], Galaxy Project[2]

[1]Johns Hopkins University, Biology, Baltimore, MD, [2]UseGalaxy.org, Steering Committee, Baltimore, MD

The Galaxy Project has become a leader in bioinformatics research, renowned for its user-friendly interface, reproducibility, and collaborative capabilities. In recent years, the integration of machine learning techniques into the Galaxy platform has ushered in a new era of data-intensive science highlighting the strengths of the Galaxy platform. Galaxy has a rich collection of tools and libraries for designing, training, and evaluating models using neural networks, decision trees, and many other ML techniques. This presentation provides an overview of some of the most compelling biomedical and biological research using machine learning techniques currently available or theoretically feasible on the Galaxy Project.

Drug Discovery and Pharmacogenomics: The Galaxy Project is being used to democratize drug discovery and pharmacogenomics research. The potential integration of machine learning approaches further enables the identification of potential drug candidates and personalized treatment strategies, ultimately accelerating drug development.
Predictive Modeling for Genomic Data: Machine learning models are increasingly used for predicting genomic features such as gene expression, protein binding, and variant effects. The integration of these models into Galaxy allows for seamless and user-friendly access to predictive tools, making genomics research more accessible across subject-mater experts.
Multi-omics Integration: Researchers can integrate multi-omics data using machine learning-driven techniques within Galaxy. This would enable holistic analyses of complex biological systems, revealing intricate relationships between genomics, transcriptomics, proteomics, and metabolomics data.
Image Analysis and Single-Cell Sequencing: Machine learning image analysis tools in the Galaxy ecosystem provide insights for the analysis of single-cell sequencing. This facilitates the identification of rare cell types, characterization of cellular heterogeneity, and identifies discoveries in areas like immunology and developmental biology.
Real-time Data Analysis and Quality Control: Machine learning-based real-time data analysis and quality control modules are being developed within the Galaxy Project. These modules can provide immediate feedback to researchers during experiments, ensuring data accuracy and reducing experimental errors.

CELL LINEAGE INFERENCE USING PATTERNS OF DNA DAMAGE

Lucy C Scott[1], Cameron Wyatt[1], Elizabeth Patton [1,2], Martin S Taylor [1]

[1]University of Edinburgh, MRC Human Genetics Unit, Edinburgh, United Kingdom, [2]University of Edinburgh, Cancer Research UK Edinburgh Centre, Edinburgh, United Kingdom

Lineage tracing technologies are powerful tools that offer insight into embryonic development by tracking the inheritance of a cell 'marker' across cell divisions. Existing lineage tracing tools have successfully constructed broad phylogenies of cellular lineages providing insight into important developmental events. However, these phylogenies often feature large gaps due to information loss, caused either by their marker not being inherited or through the occurrence of cell death.

We have developed a novel lineage tracing technique that overcomes this limitation by using inherited DNA damage as a cell 'marker'. Lesion segregation describes the process through which DNA damage can be inherited across multiple cell divisions (Aitken et al. 2020). When cells experience a discrete mutagenic event, lesions form on the DNA strands present and these lesions can evade the cell's repair machinery, instead being replicated across as the cell divides, introducing base pair mismatches to the newly synthesised DNA strand. Since replication necessarily occurs with every cell division, the marker must be inherited by subsequent daughter cells. Additionally, both DNA strands are independently damaged before segregating into separate daughter cells, generating complementary patterns of mutations. This complementarity creates an expectation that can be used to infer the loss of a lineage as for every pattern of mutations we observe, we expect to find another, complementary pattern. In contrast to the broad, inexact phylogenies generated by current tools, our technique provides more precise resolution over a smaller window of development.

To develop this technique we have been introducing DNA damage to cleavage phase zebrafish embryos, a phase of development during which DNA repair is minimal. We have trialled both UV irradiation and ENU as sources of DNA damage, titrating exposures to identify optimal dosage for introducing mutations without perturbing developmental progress. We have then generated both single cell RNA sequencing and whole genome sequencing data from individual zebrafish which have developed in the presence of this DNA damage and combined this data to identify resulting mutations in cell populations while simultaneously assigning cell identities. Drawing correlations between cell identity and inherited mutation patterns then enables relationships between cells of the cleavage phase embryo explored.

A SURROGATE MODELING FRAMEWORK FOR INTERPRETING DEEP NEURAL NETWORKS IN FUNCTIONAL GENOMICS

Evan Seitz, Justin Kinney, Peter Koo

Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY

Understanding the cis-regulatory grammars that coordinate how proteins interact to regulate transcription is a major goal in genomics. Deep neural networks (DNNs) applied to this task have greatly enhanced our ability to accurately predict experiments in regulatory genomics. Despite their impressive performance compared to traditional methods in computational genomics, it remains difficult to determine how these networks form their decisions. To address this gap, attribution methods are being increasingly used to gain mechanistic insights underlying DNN predictions. Attribution methods probe the trained DNN to assess the importance of each nucleotide in a sequence to produce an attribution map, which has been shown to visualize known functional motifs and their locations. However, current attribution methods are sensitive to the local function properties learned by the DNN, making identification of functional motifs difficult. Due to the high expressivity of DNNs, there is no guarantee this issue can be resolved by altering the DNN to learn smoother functions amenable to attribution maps.

Instead, we surmise that attribution-based explanations can be made more robust by approximating a larger region of function space anchored at a given sequence of interest with an interpretable surrogate model, for which the parameters provide direct interpretations of variant importance similar to attribution maps. Here we introduce this new surrogate modeling approach into genomics, where it has not yet been explored. Our approach, called SQUID for Surrogate QUantitative Interpretability of Deepnets, is a general framework that leverages interpretable surrogates to quantitatively model the sequence-function relationship learned by any black-box genomic model. We demonstrate our framework across several existing DNNs designed to perform a variety of regulatory genomics prediction tasks. For each of these DNNs, we show that SQUID outperforms existing attribution methods in studies spanning ensembles of high-functioning motifs and genomic sequences. From this comparison, we find that SQUID is able to more robustly characterize the direct effect of motifs and their higher-order interactions on predictions, consistently model larger sequence contexts, identify weaker binding sites that enable opportunities for better annotation, and provide better approximations to variant effect predictions. SQUID provides a leap forward in our ability to decipher the quantitative effects of cis-regulatory elements throughout the genome.

DEMOCRATIZING ACCESS TO GENOMICS AND DATA SCIENCE EDUCATION THROUGH THE CLOUD – THE GDSCN SUCCESS STORY

Shurjo K Sen, Genomic Data Science Community Network

NHGRI, OGDS, Bethesda, MD

The genome informatics workforce must reflect the diversity of our nation's population. This requires greatly increasing the number of individuals from underrepresented backgrounds who have the necessary training to pursue careers in computational genomics and data science. Currently, educational opportunities in genomics and genomic data science are primarily limited to resource-rich institutions with access to high-performance computing clusters. This results in an extreme lack of diversity in the genomic data science workforce. In 2020, NHGRI established the Genomic Data Science Community Network (GDSCN) contract to work with students and faculty at limited-resource institutions to develop and facilitate access to instructional materials for genomic data science, cloud computing and related topics. By supporting the development of training materials by faculty at institutions with less resources, and increasing dissemination and outreach activities, NHGRI seeks to enable a broader spectrum of undergraduate institutions to have educational and research access to genomic data science.

The GDSCN organizers at Johns Hopkins University, the Fred Hutchinson Cancer Center, and Carnegie Institution, together with NHGRI, have worked for the past two and a half years to diversify the genomic data science research community by working with 25 faculty from Historically Black Colleges and Universities (HBCUs), Hispanic Serving Institutions (HSIs), Tribal Colleges and Universities (TCUs), and Community Colleges (CCs) geographically distributed across the United States. In spite of the global pandemic, this network has built a strong community through virtual symposia, working groups, and in-person symposiums; tailored curricula using cloud-based resources that engage students in genomic analysis; and formed scientific collaborations that include manuscripts and grant proposals. Moving forward, continuity of effort is crucial to sustain momentum and build upon the hard-earned trust among the network. Towards this goal, the community has identified a collaborative urban soil microbiome research project involving network faculty and students as the logical next step to further engage these keystone institutions. This project will allow us to scale up our current pilot project to reach students from diverse GDSCN institutions to participate in all aspects of research from sample collection through data analysis. Success here will result in multiple manuscripts that will involve GDSCN faculty and students as co-authors and deepen connections amongst the network. In parallel, we are developing complementary educational materials to teach genomic data science and microbiome analysis to GDSCN students. Collectively, these efforts will provide a capstone research & education experience for GDSCN faculty and students, and will have major impacts on their future educational experiences and careers.

CLASSIFICATION OF ANTIBIOTIC RESISTANCE OF MYCOBACTERIUM TUBERCULOSIS VIA LINEAR AND NON-LINEAR MACHINE LEARNING

Mohammadali Serajian[1], Simone Marini[2], Jarno N Alanko[3], Noelle R Noyes[4], Mattia Prosperi[2], Christina Boucher*[1]

[1]University of Florida, Computer and Information Science, Gainesville, FL, [2]University of Florida, Epidemiology, Gainesville, FL, [3]University of Helsinki, Computer Science, Helsinki, Finland, [4]University of Minnesota, Veterinary Population Medicine, St. Paul, MN

Background. World Health Organization estimates that there were over 10 million cases of tuberculosis (TB) worldwide in 2019, resulting in over 1.4 million deaths, with a worrisome increasing trend yearly. The disease is caused by *Mycobacterium tuberculosis* (MTB) through airborne transmission. Treatment of TB is estimated to be 85% successful, however, this drops to 57% if MTB exhibits multiple antimicrobial resistance (AMR), for which fewer treatment options are available.

Methods. We develop a robust machine learning classifier using both linear and nonlinear models (i.e., LASSO logistic regression and random forests) to predict the resistance of MTB for a broad range of AMR classes. We use publicly available data from the CRyPTIC consortium to train our classifier. The data consists of both whole genome sequencing and phenotypic testing data for 16 different antimicrobial resistance classes over 6,500 different MTB isolates. To train our model, we assemble the sequence data into contigs, identify all unique 31-mers in the set of contigs, and build a feature matrix M, where $M[i, j]$ is equal to the number of times the i-th 31-mer occurs in the j-th genome. Due to the size of this feature matrix (over 350 million unique 31-mers), we first build a sparse representation using succinct data structures, and then extract a subset of rows to train the model. Hence, our method, which we refer to as MTB++, leverages succinct data structures and iterative methods to allow the screening of all the 31-mers in the development of both LASSO logistic regression and random forest.

Results. We use 5-fold cross-validation to compare our methods with other well-established AMR prediction algorithms: ResFinder, TBProfiler, Mykrobe, and KVarQ. MTB++ is able to achieve high discrimination, achieving a F-1 greater than 90% for the top four antibiotics, between 90-80% for the next six, and between 80-75% for the other two. MTB++ had the highest F-1 score in all but 3 classes, and was the most comprehensive since it had a F-1 score greater than 75% in all but four (rare) AMR classes. Lastly, we provide some biological insights into the more rare AMR classes based on our predictions; and an estimate of the amount of data that is needed in order to provide accurate predictions.

Availability. The models and source code are publicly available on Github at https://github.com/M-Serajian/MTB-plus-plus/.

[*] Corresponding author

EASTR: IDENTIFYING AND ELIMINATING SYSTEMATIC ALIGNMENT ERRORS IN MULTI-EXON GENES

Ida Shinder[1,2], Richard Hu[2,3], Hyun Joo Ji[2,3], Kuan-Hao Chao[2,3], Mihaela Pertea[2,3,4,5]

[1]Cross Disciplinary Graduate Program in Biomedical Sciences, Johns Hopkins School of Medicine, Baltimore, MD, [2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, [3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, [4]Department of Biomedical Engineering, Johns Hopkins School of Medicine and Whiting School of Engineering, Baltimore, MD, [5]Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD

Accurate RNA-seq alignment to reference genomes is a cornerstone of transcript assembly, annotation, and gene expression studies, essential in both biomedical research and basic sciences. Our study reveals that widely-used splice-aware aligners, such as STAR and HISAT2, introduce systemic alignment errors by incorrectly treating repeated sequences as if they were spliced, thus generating spurious spliced alignments. These errors highlight challenges that may call for adjustments in the computational logic used for spliced alignment, including aligners designed for long reads. Such systemic aligner-induced errors have led to the inclusion of "phantom" introns in widely-used genome annotation databases. Our multi-species RNA-seq analysis reveals that these erroneous alignments can make up to 20% of the spliced alignments in some datasets. In all gene catalogs we examined, we found hundreds of likely cases of mis-annotation. To address these issues, we developed EASTR (Emending Alignments of Spliced Transcript Reads), a software tool designed to detect and eliminate false spliced alignments from both alignment and annotation files. EASTR enhances the accuracy of spliced alignments in diverse species, such as human, maize, and *Arabidopsis thaliana*, by evaluating sequence similarities between the upstream and downstream regions directly flanking each intron, as well as their occurrence frequency. Pre-processing alignment files with EASTR prior to transcript assembly significantly reduces false positives in introns, exons, and transcripts, thereby improving overall transcript assembly accuracy. Moreover, EASTR also serves as a tool for identifying mis-annotations in existing annotation databases and transcript assemblies.

SYSTEMATIC DISCOVERY OF SPLICE-SITE CREATING VARIANTS FROM MASSIVE PUBLICLY AVAILABLE TRANSCRIPTOME SEQUENCING DATA

Yuichi Shiraishi, Naoko Iida, Ai Okada, Kenichi Chiba

National Cancer Center Research Institute, Division of Genome Analysis Platform Development, Tokyo, Japan

Many loss-of-function variants are associated with abnormal splicing. Among the most prevalent classes of splicing-associated variants (SAVs) are those that disrupt existing splice sites, leading to exon-skipping, alternative 5’ (or 3’) splice-site usage, and/or intron retention. In a prior study, we developed a statistical approach to identify these SAVs and applied it to matched genome and transcriptome sequencing data (Shiraishi et al., Genome Research, 2018). While this method is sensitive and accurate, a significant limitation is the requirement for both genome and matched transcriptome sequencing data—a combination that's uncommon in many current cohort studies. However, with the rapid accumulation of transcriptome sequencing data in public repositories like the Sequencing Read Archive, we designed a new approach, IRAVNet, that detects intron retention associated variants (IRAVs) using only transcriptome sequencing data, ensuring high precision (Shiraishi et al., Nature Communications, 2021).

In this presentation, we introduce "juncmut," a tool to detect another type of splicing-associated variant: splice-site creating variants (SSCVs) using only transcriptome sequencing data. These SSCVs can lead to the formation of cryptic splice sites deep within exons and introns, often overlooked in standard analyses. We applied juncmut to datasets from 10,724 TCGA samples and 312,314 SRA samples, identifying 33,368 splice-site creating variants. An analysis of the positional relationships of SSCVs within deep introns and Alu sequences revealed hotspots of novel splice sites at specific Alu sequence positions. Many of these SSCVs, which impact disease-causing variants, were found to generate premature termination codons, subsequently degraded by nonsense-mediated decay. Conversely, certain genes, such as TP53 and PIK3R1, exhibited gain-of-function SSCVs. In summary, our comprehensive SSCV catalog can shed light on new biological mechanisms related to splicing and genetic variation, serving as a valuable reference for potential drug discovery targets.

PRECISE CHARACTERIZATION OF SOMATIC COMPLEX STRUCTURAL VARIATIONS FROM TUMOR/CONTROL PAIRED LONG-READ SEQUENCING DATA WITH NANOMONSV

Yuichi Shiraishi[1], Junji Koya[2], Kenichi Chiba[1], Ai Okada[1], Yasuhito Arai[3], Yuki Saito[1], Tatsuhiro Shibata[3], Keisuke Kataoka[2]

[1]National Cancer Center Research Institute, Division of Genome Analysis Platform Development, Tokyo, Japan, [2]National Cancer Center Research Institute, Division of Molecular Oncology, Tokyo, Japan, [3]National Cancer Center Research Institute, Division of Cancer Genomics, Tokyo, Japan

Long-read sequencing technologies have garnered significant attention in the hope of enhancing the detection of structural variations (SVs). However, there remains a scarcity of software designed for identifying somatic SVs from tumor and matched control sequencing data. In this presentation, we introduce our novel software, nanomonsv, tailored for somatic SV detection from paired tumor/matched control data.

Nanomonsv is equipped with two detection modules: Canonical SV module and Single Breakend SV module. Using tumor/control paired long-read sequencing data from three cancer and their matched lymphoblastoid lines, we show that Canonical SV module can identify somatic SVs, which are also detectable by short-read technologies, but with a higher precision and recall than other existing methods. Additionally, we've developed a workflow for classifying mobile element insertions, shedding light on their intricate properties. These include 5′ truncations, internal inversions, and the identification of source sites for 3′ transductions. Moreover, Single Breakend SV module specializes in detecting complex SVs, ones that can only be discerned by long-reads. Examples include SVs related to highly repetitive centromeric sequences and those influenced by LINE1 and virus-mediated rearrangements.

In conclusion, when applied to cancer long-read sequencing data, our methodologies can uncover a broad spectrum of features of somatic SVs. This will pave the way for a more profound understanding of mutational processes and the functional consequences of these SVs.

SIGMONI: CLASSIFICATION OF NANOPORE SIGNAL WITH A COMPRESSED PANGENOME INDEX

Vikram S Shivakumar, Omar Y Ahmed, Sam Kovaka, Mohsen Zakeri, Ben Langmead

Johns Hopkins University, Computer Science, Baltimore, MD

Improvements in nanopore sequencing necessitate more efficient methods for long read classification. Particularly crucial are pre-filtering and adaptive sampling algorithms, which can enrich or deplete reads of interest with minimal computational effort. Previous methods rely on either aligning the raw signal or basecalled reads to a genomic reference index. The neural network basecaller presents a computational bottleneck in this pipeline, requiring hardware acceleration and significantly limiting the potential savings of read filtering in practice. Previous signal-based approaches circumvent basecalling by identifying similar signal “seeds” in a reference index. However, current methods are unable to scale efficiently with large, repetitive references such as a pangenome, limiting classification to partial references or individual genomes and potentially misclassifying reads due to reference bias.

We introduce Sigmoni, a rapid nanopore signal-based multiclass classification method based on the r-index, scaling to references of hundreds of Gbps and efficiently identifying reads based on exact matching. Sigmoni uses an ultra-fast signal quantization procedure in lieu of a basecaller to project the electrical signal into a discrete alphabet of bins representing picoamp ranges. The discretized signal is queried against an r-index, a compressed index which scales with the number of runs in the Burrows-Wheeler transform, to compute matching statistics. In addition, Sigmoni utilizes an augmented document array data-structure to retrieve approximate co-linearity statistics.

Sigmoni is 10-300X faster than previous signal-based methods in classifying short signal “chunks”, crucial for real-time read identification in adaptive sampling. Additionally, Sigmoni computes matching statistics in linear time with respect to read length, unlike previous methods which use seed-chain-extend approaches. It can index a human pangenome (44 diploid assemblies) in 12.6 GB, ~282X smaller than RawHash2. Sigmoni outperforms nanopore signal-based tools in both binary and multiclass classification from full-length signal in a mock community, crucial for read filtering tasks in metagenomics applications, as well as in classifying read chromosomal-origin in human nanopore reads.

Sigmoni can accurately classify nanopore reads directly from electrical signal, significantly improving classification speed while scaling efficiently to query against pangenomic indexes. It outperforms previous tools in speed and scalability, while maintaining the necessary sensitivity for classification in metagenomic datasets. This approach highlights the potential for sublinear index-based methods for pangenomic classification, enabling rapid adaptive sampling and filtering on large pangenomic databases.

CLASPY: CELL LINE AUTHENTICATION WITH STRs IN PYTHON

Alaina G Shumate, Rebecca N Mitchell, Daniel S Standage

National Bioforensic Analysis Center, Bioinformatics, Frederick, MD

Mammalian cell lines are a widely used resource in biomedical research. Handling and sharing of cell line materials over time sometimes leads to contamination and misidentification, compromising the integrity of experimental findings. To mitigate these issues, profiling of short tandem repeat (STR) markers has emerged as the primary strategy for cell line verification. We introduce Claspy, a software package written in Python for cell line authentication using STRs. Claspy computes a pairwise similarity score between two cell lines based on shared versus disjoint alleles. This similarity score can be used to rank cell line profiles in a database such as Cellosaurus and report likely profile matches. The freely available Claspy software package has a light footprint and can be installed using popular package managers such as conda and pip.

This work was funded under Contract No. HSHQDC-15-C-00064, awarded by the DHS S&T to NBACC, a DHS federal laboratory operated by BNBI. Views and conclusions contained herein are those of the authors and should not be interpreted to represent policies, expressed or implied, of the DHS or S&T.

AI APPROACH FOR DE NOVO GENOME ASSEMBLY

Lovro Vrček[1,4], Xavier Bresson[2], Thomas Laurent[3], Martin Schmitz[1,2], Kenji Kawaguchi[2], Mile Šikić[1,4]
[1]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Laboratory of AI in Genomics, Singapore, Singapore, [2]National University of Singapore, Department of Computer Science, Singapore, Singapore, [3]Loyola Marymount University, Mathematics Department, Los Angeles, CA, [4]University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Signal Processing, Zagreb, Croatia

We propose a new de novo genome assembly paradigm utilising telomere-to-telomere resolved genomes and AI architecture. The results are better or on par with the state-of-the-art de novo assemblers based solely on Oxford Nanopore Technologies or Pacific Biosciences long reads.

For four decades, de novo assembly has relied on an algorithmic method that organizes sequenced DNA reads into extended contiguous sequences. Traditional methods convert overlapping sequence fragments into a graph structure, which is then resolved to reconstitute the genome. Methods are categorized into Overlap-Layout-Consensus (OLC) and de Bruijn graph assemblers. Despite the theoretical potential for accurate reconstruction, sequencing errors and repetitive sequences often introduce complexities that these algorithms struggle to resolve. Additionally, optimal solutions remain computationally infeasible for longer genomes. Numerous heuristic methods have been crafted to simplify graphs. However, these methods largely rely on manually selected parameters obtained by testing performances on several well-known genomes and datasets.

The achievement of a complete T2T reconstruction of the CHM13 human genome has revolutionized genomic research, integrating multiple sequencing techniques with extensive manual curation. While promising, this multi-faceted approach is resource-intensive.

Our proposal harnesses the accumulated data from T2T projects, combined with AI, to streamline genome assembly. This method allows for the simulation of numerous sequencing scenarios to anticipate various assembly challenges.

We refine the layout phase of the OLC process using the cutting-edge hifiasm tool for graph generation. The starting and ending of simulated reads' positions on a reference genome enable us to correctly label all edges in a graph as positive and negative. We have developed a novel graph neural network methodology to predict edge labels, capitalizing on the inherent symmetries within assembly graphs. Subsequently, a greedy algorithm decodes the graph to reconstruct the genome, ensuring comprehensive coverage by initiating traversal from multiple positions.

We trained the network on several chromosomes taken from the HG002 Q100 de novo assembly project. Testing the approach on CHM13 genome and inbred M. musculus and A. thaliana genomes show that genome contiguity of resolved genomes measured in L90 and NGA50 is longer or equal to hifiasm results and better than Hi-Canu. In addition, we show the generalisation of the approach by inferring edge labels on previously unseen Raven graphs constructed from ONT data. Surprisingly, results show on CHM13 and inbred Arabidopsis datasets that our approach outperforms assemblies obtained with Raven and Flye.

MITOCHONDRIAL MUTATION AND DYSFUNCTION IN HIGH GRADE SEROUS OVARIAN CANCER

Ryan P Silk[1], Alison M Meynert[1], Ailith Ewing[1], Brian Dougherty[2], Patricia Roxburgh[3,4], Charlie Gourley[5], Colin A Semple[1]

[1]University of Edinburgh, MRC Human Genetics Unit, Edinburgh, United Kingdom, [2]AstraZeneca, Oncology R&D, Waltham, MA, [3]University of Glasgow, Wolfson Wohl Cancer Research Centre, Glasgow, United Kingdom, [4]University of Glasgow, Beatson West of Scotland Cancer Centre, Glasgow, United Kingdom, [5]University of Edinburgh, Nicola Murray Centre for Ovarian Cancer Research, Edinburgh, United Kingdom

In recent years, there has been a growing body of evidence linking somatically acquired mitochondrial dysfunction to cancer. However, robust estimates of the prevalence, patterns and impact of these events remain limited due to small sample sizes and incomplete analyses. Therefore, it is not yet understood how these alterations could provide a mechanism for tumour initiation and growth, and also how they may affect a patient's response to treatment. This lack of systematic analysis for mitochondrial dysfunction is most prevalent in ovarian cancer. Here, we have analysed 324 whole-genome sequenced High Grade Serous Ovarian Cancer (HGSOC) samples with blood matched normals. We find frequent somatic mutations in the mitochondrial DNA, the most deleterious of which are associated with reduced overall survival.

TREEVAL: DATA GENERATION FOR THE CURATION OF CHROMOSOME-SCALE GENOMES

Damon Pointon, Ying Sims, William Eagles, Jonathan Wood, Shane McCarthy

Wellcome Sanger Institute We, DTOL, Cambridge, United Kingdom

The Tree of Life Project is scaling up in order to, over the next decade, sequence the entirety of the complex life in Britain and Northern Ireland as well as produce high-quality genomic assemblies for the estimated 70,000 species that live in this region. The size of this task cannot be understated and requires a tectonic shift in how the project manages it’s data flow. Amongst the first steps to automate this process we present TreeVal, a bipartite project which aims to generate and display the plurality of data required for the manual curation of genomic assemblies into reference quality assemblies.
The pipeline is available at: https://github.com/sanger-tol/treeval.

SIMPLIFYING AND IMPROVING SINGLE-CELL GENE EXPRESSION ANALYSIS WITH PICCOLO

Amartya Singh[1], Hossein Khiabanian[2], Daniel Herranz[1]

[1]Rutgers Cancer Institute of New Jersey, Pathology, New Brunswick, NJ,
[2]Regeneron Pharmaceuticals, Regeneron Genetics Center, Tarrytown, NY,
[3]Rutgers Cancer Institute of New Jersey, Pathology, New Brunswick, NJ

Single-cell RNA sequencing (scRNA-seq) studies can uncover cellular heterogeneity and map dynamic changes during differentiation, treatment, and evolution. Yet, distinguishing between technical and biological sources of variation poses a significant challenge to appropriate and effective analysis and interpretation. Thus, preprocessing and normalization of scRNA-seq counts data plays a pivotal role in all scRNA-seq analyses. This is typically accomplished by re-scaling the observed counts to reduce the differences in total counts between the cells and then log transforming the scaled counts to stabilize the variances. Often, this is followed by feature selection to identify genes that capture most of the biologically meaningful variation across the cells. Recently, Ahlmann-Eltze and Huber provided a thorough comparison between some of the widely used transformation methods for scRNA-seq data (Comparison of transformations for single-cell RNA-seq data, Nature Methods, April 2023) and concluded that despite the conceptual appeal of residuals-based methods the simple log-transformation based normalization method performed just as well or better, particularly when evaluated based on consistency of nearest-neighbor cells. However, they themselves also highlighted that this normalization is unable to effectively reduce sampling depth differences between cells thus compromising on its principal objective.

We explored the fundamental nature of scRNA-seq counts data to propose a simplified feature selection method based on estimates of quasi-Poisson dispersion coefficients and show that we can perform feature selection before normalization. We also propose a novel residuals-based normalization method that includes a variance stabilization transformation to ensure effective stabilization of the residual variances in contrast to previously proposed residuals-based methods that relied on raw counts to compute residuals. Using this novel normalization approach, we first show that it reduces technical biases much more effectively compared to the log-transformation based method. We further demonstrate significant improvements in downstream clustering analyses through the application of our feature selection and normalization methods to truth-known biological as well as simulated counts data sets. Based on these results, we make the case for a revised scRNA-seq analysis workflow wherein feature selection precedes and in fact informs our residuals-based normalization. This novel workflow has been implemented in an R package called Piccolo which can even be used with the popular Seurat package for scRNA-seq analyses.

TRIOKALA: A TRIO CO-ASSEMBLY APPROACH FOR *DE NOVO* VARIANT DETECTION

Steven J Solar[1], Carlos R Ferreira[2], Sergey Koren[1], Dmitry Antipov[1], Mikko Rautiainen[3], Adam Phillippy[1]

[1]National Human Genome Research Institute, National Institutes of Health, Genome Informatics Section, Computational and Statistical Genomics Branch, Bethesda, MD, [2]National Human Genome Research Institute, National Institutes of Health, Medical Genomics and Metabolic Genetics Branch, Bethesda, MD, [3]Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, Helsinki, Finland

Variant calling is important for investigating genotype-phenotype relationships, improving our understanding of diseases, and informing care and future research. This is typically achieved by sequencing patient samples, aligning the reads to a reference genome, and identifying places where reads differ from the reference. This process is subject to reference bias, and can struggle to identify variants in repetitive regions, regions with significant deviation from the reference, and gaps when using incomplete references.

We investigated a trio co-assembly approach, treating a proband's parents as their “personal reference genome” to call *de novo* variants in the child. We used 30x HiFi sequencing for two trios with unaffected parents and a child with genochondromatosis: a skeletal dysplasia suspected to be caused by an unknown autosomal dominant mutation. Trios were co-assembled with Verkko, and assembly graph nodes colored by read origin. Nodes containing only child reads were marked as *de novo* nodes, and the supposed original parental sequence pre-mutation was identified. Small and structural variants were called, considering the parental sequence as the reference, and filtered to exclude extraneous haplotype variation. The remaining variants were lifted to T2T-CHM13, filtered with gnomAD, and annotated with VEP. This resulted in roughly 30 small variants and 10 structural variants per trio, including one frameshift mutation in *ITPRID2*. No variants were shared between probands in either the co-assembly or alignment (our “control”) approach. For one proband, one variant was shared between the two approaches, but no variants were readily identified as possible causal variants for genochondromatosis.

TrioKala provides a reference-free method for the identification of *de novo* variants. This enables the comprehensive identification of *de novo* variants in regions of the genome not well represented by, or substantially different from, available references. Current limitations of this approach include missed variants within homopolymers and a lack of variant quality scores. Future work will focus on the automatic detection of recombination events (and the identification of their breakpoints), and the determination of variant quality by the interrogation of reads underlying the child-specific haplotype paths.

MICROHAPDB: A COMPREHENSIVE CATALOG OF HUMAN MICROHAPLOTYPE VARIATION

Daniel Standage

National Bioforensic Analysis Center, Bioinformatics, Frederick, MD

Microhaplotypes (MHs) are a novel class of genetic marker defined as a set of SNPs occupying a short genomic region, such that a single NGS read is capable of genotyping and empirically phasing all variation at the locus. In contrast to single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs), MH markers can exhibit numerous allelic combinations and are unaffected by PCR stutter artifacts. Interest in the use of MH markers for anthropology, genetic genealogy, forensics, and clinical diagnostics has grown in recent years, with applications to inference of biogeographical ancestry, kinship analysis, human identification, and DNA mixture detection and deconvolution. A growing number of studies have published MH markers as proofs-of-concept, as proposed panels, and as whole-genome surveys, with the total number of published markers now exceeding 3000. MicroHapDB is a comprehensive catalog of published human MH markers aggregated from the relevant literature, supporting the efforts of a growing community to develop and standardize MH panels—along with ancillary kits and information resources—for various research and operational applications. In addition to information about each marker's SNPs and location, MicroHapDB contains estimates of allele frequency and allelic diversity (effective number of alleles) for 26 global population groups. MicroHapDB is available as an open-access community resource (https://github.com/bioforensics/MicroHapDB), installed locally as a software package, and accessed through a command line interface or Python API.

This work was funded under Contract No. HSHQDC-15-C-00064, awarded by the DHS S&T to NBACC, a DHS federal laboratory operated by BNBI. Views and conclusions contained herein are those of the authors and should not be interpreted to represent policies, expressed or implied, of the DHS or S&T.

CRISPRsc: POOLED CRISPR SCREENING WITH SINGLE-CELL TRANSCRIPTOME RESOLUTION

Tiana Stastny, Maximilian Blanck, Matthew Perkett, John McGonigle, Simon Scrace, Carlos le Sage, Anja Smith

Revvity, Research and Development, Lafayette, CO

Pooled knockout screening with CRISPR-Cas91 has revolutionized drug discovery and has evolved to become the backbone of many drug discovery pipelines. Pooled CRISPR knockout screens offer a fast, cost effective and precise toolbox to determine drug mechanism of action and to identify genetic resistance and sensitizing factors that enable patient stratification2,3,4. Though powerful in its own right, pooled CRISPR screening is often restricted to the measurement of a single phenotype such as proliferation or cell survival, or simple phenotypic changes, such as, measurement of changes in activity of a single gene through the generation of a reporter cell line. Recent publications have demonstrated the feasibility of coupling pooled CRISPR screening to highly complex readouts, such as single cell gene expression profiling5,6. It is widely appreciated that tumor heterogeneity drives metastasis7, invasion and chemotherapy resistance8. Consequently, high levels of cell heterogeneity are also observed in cancer cell lines9 and results in altered drug resistance10 as well as immune response to neoantigens that facilitate tumor rejection in vivo11. To better understand drug responses and the occurrence of treatment resistance it is paramount to investigate the underlying mechanisms at single cell resolution. By probing the transcriptome of each individual cell for its response to a treatment or a stimulus, we can link gene expression signature with drug response which enables the detection of biomarkers and the development of personalized medicines. CRISPRsc (CRISPR single cell) combines the flexibility of pooled screening with the power of transcriptomics at single cell resolution. By linking a specific guide RNA to single cell whole transcriptome readout, CRISPRsc offers the opportunity to interrogate the phenotype of any given genetic perturbation in unprecedented detail. This allows identification of nuanced cellular responses that are integral to solving complex biological questions. Revvity’s CRISPRsc screening platform offers a streamlined approach to elucidating valuable and intriguing biological information.

References:

1. Wang, T. et al., Science 350, 1096–1101 (2015)
2. Cross, B. C. S. et al., Sci. Rep. 6 (2016)
3. Le Sage, C. et al., Sci. Rep. 7 (2017)
4. Blanck, M. et al., CRISPR J. 3, 211–222 (2020)
5. Datlinger, P. et al., Nat. Met. 14, 297–301 (2017)
6. Hill, A. et al., Nat. Met. 15, 271–274 (2018)
7. Kim, N. et al., Nat. Commun. 11, 2285 (2020)
8. Liu, D. et al., Nat. Commun. 8, 2193 (2017)
9. Dexter, D.L. et al., Am. J. Med. 71, 949–956 (1981)
10. Ben-David, U. et al, Nature 560, 325–330 (2018)
11. Wolf, Y. et al., Cell 179, 219–235 (2019)
12. Mercer, T. et al., Nat. Protoc. 9, 989–1009 (2014)

INNOVATION, CONSTRAINT, AND THE EVOLUTION OF GENETIC NETWORKS IN MAJOR EUKARYOTIC LINEAGES

Jacob L Steenwyk[1], Maxwell C Coyle[1], Noah Bradley[2], Chris T Hittinger[3], Antonis Rokas[4], Nicole King[1]

[1]UC-Berkeley / HHMI, Molecular and Cell Biology, Berkeley, CA, [2]Northwestern University, Department of Molecular Biosciences, Evanston, IL, [3]University of Wisconsin - Madison, Department of Genetics, Madison, WI, [4]Vanderbilt University, Biological Sciences, Nashville, TN

Genetic networks depict the intricate relationships among genes, their pathways, and cellular functions. Here, we infer genetic networks using coevolutionary information across 26 major lineages of animals and fungi. Analysis of network features uncovers both analytical and biological factors influencing their structural properties, including evolutionary rate and signal-to-noise ratios. Ancestral reconstructions of complex gene-gene relationships uncover patterns of gain and loss, mirroring gene families, but substantially more dynamic. Guided by these findings, we construct individual genetic networks for Animals and Choanoflagellates and identify complex gene relationships shared in both lineages, thus likely predating animal origins. Shared hubs of genes encode ancient cellular functions, such as cell cycle regulation, DNA replication, and ciliary processes. The principle of 'guilt-by-association' emerges as a promising approach for uncovering genotype-to-phenotype relationships. These findings uncover innovation and constraint in genetic network evolution and suggest that gene-gene association changes are a dynamic and underexplored mode of genome evolution.

SOMAMUTDB: A DATABASE OF SOMATIC MUTATIONS IN NORMAL HUMAN TISSUES

Shixiang Sun[1], Yujue Wang[1], Alexander Maslov[1,2], Xiao Dong[3], Jan Vijg[1,4]

[1]Albert Einstein College of Medicine, Genetics, Bronx, NY, [2]Voronezh State University of Engineering Technology, Laboratory of Applied Genomic Technologies, Voronezh, Russia, [3]University of Minnesot, Genetics, Cell Biology and Development, Minneapolis, MN, [4]Shanghai Jiao Tong University School of Medicine, Center for Single-Cell Omics, Shanghai, China

De novo mutations, a consequence of errors in DNA repair or replication, have been reported to accumulate with age in normal tissues of humans and model organisms. This accumulation during development and aging has been implicated as a causal factor in aging and age-related pathology, including but not limited to cancer. Due to their generally very low abundance mutations have been difficult to detect in normal tissues. Only with recent advances in DNA sequencing of single-cells, clonal lineages or ultra-high-depth sequencing of small tissue biopsies, somatic mutation frequencies and spectra have been unveiled in several tissue types. The rapid accumulation of such data prompted us to develop a platform called SomaMutDB (https://vijglab.einsteinmed.org/SomaMutDB) to catalog the 5.77 million single nucleotide variations (SNVs) and small insertions and deletions (INDELs) thus far identified using these advanced methods in twenty-eight human tissues or cell types as a function of age or environmental stress conditions. SomaMutDB employs a user-friendly interface to display and query somatic mutations with their functional annotations. Moreover, the database provides six powerful tools for analyzing mutational signatures associated with the data. We believe such an integrated resource will prove valuable for understanding somatic mutations and their possible role in human aging and age-related diseases.

INVESTIGATION OF TISSUE-SPECIFIC TRANSCRIPTOME AT ISOFORM-LEVEL IN GTEX AND TCGA DATASETS

Pallavi Surana, Pratik Dutta, Chirayush Patel, Jayendra Kantipudi, Roshan R Yedla, Sankara Kota, Ramana V Davuluri

Stony Brook University, Biomedical Informatics, Stony Brook, NY

Although human tissues carry out common molecular processes, different tissues can be distinguished by gene expression patterns. While numerous informatics methods have addressed this problem, most studies have focused on gene-level analysis. It is now well known that majority of human genes produce multiple transcript-variants and protein isoforms, which could be involved in different functional pathways. Moreover, altered expression of transcript-variants and protein isoforms for numerous genes is linked with cancer and its prognosis. These cells manipulate regulatory mechanisms to express specific isoforms that confer drug resistance and survival advantages. We, therefore, hypothesize that gene expression signatures at isoform-level would allow us to identify novel tissue-specific genes at splice/transcript-variant level and generate a more robust gene-expression based classifier for tissue and cancer classification. We propose a novel tissue-specificity score (ΔP), combining subsampling-based p-values and fold-change estimates, to develop a pipeline identifying tissue-specific and enhanced transcripts using GTEx RNA-seq data. We identified 16,666 tissue-specific, 858 tissue-enhanced, 23,397 widely expressed, and 9,114 housekeeping transcripts. Testis has the most (13,004) tissue specific isoforms followed by brain (629), liver (603), muscle (389), and spleen (271). These groups of transcripts are made available in Transcript-level Tissue Expression Database (TransTExDb) - https://bmi.cewit.stonybrook.edu/transtexdb

We investigated tissue-specific transcript expression patterns by differential expression and clustering analyses (pan and organ-specific cancers). Also, we developed deep-learning based classification models to classify the promoters of brain, testis, and liver-specific transcripts from the promoters of housekeeping transcripts, using DNABERT (Ji et al. 2021, Bioinformatics), a large language model (LLM) our group developed for decoding the DNA language in the human genome. By applying DNABERT-snv, we identified candidate variants (germline and somatic) affecting core-promoters and splice-sites in TransTExDb from TCGA and dbSNP databases. Using DNABERT-snv, which leverages DNABERT-Prom and Splice models, we pinpointed 196,734 functional variants across core promoters and splice regions. While 50% of these variants impact the core promoter predictions, the rest impact the splice donor and acceptor regions. Of these variants, we found 3852 variants across 2219 functional sites as clinically relevant in ClinVar database.

These results emphasize how potentially important transcript variants could be missed by solely focusing on the canonical isoform. The predicted SNVs and somatic mutations that alter the core-promoter and splice-sites offer a rich set of candidates for further experimental validation.

MODDOTPLOT: A RAPID AND INTERACTIVE VISUALIZATION OF TANDEM REPEATS

Alexander P Sweeten[1,2], Michael C Schatz[1], Adam M Phillippy[2]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]National Human Genome Research Institute, Computational and Statistical Genomics Branch, Bethesda, MD

A common method for analyzing genomic repeats is to produce a sequence similarity matrix, which can be succinctly visualized via a dot plot. In order to adapt to the current “Telomere to Telomere” era of fully completed genomes, software such as StainedGlass have been fine-tuned to produce dot plots that can visualize the large multi-megabase tandem repeats these complete genomes contain. However, these tools face challenges from the high computational overhead of sequence alignment, as well as decreasing accuracy that alignment tends to cause in these repetitive regions.

In this work we introduce ModDotPlot, an interactive and alignment-free dot plot viewer. By approximating the Average Nucleotide Identity via the Containment Index, ModDotPlot can produce an accurate plot orders of magnitude faster than StainedGlass. This feat is accomplished through the use of a modimizer scheme, which sketches *k*-mers based on a modulo function. This allows users of ModDotPlot to control the sparsity of *k*-mer sets, selecting an appropriate amount based on the size of the input sequence.

Using Mod.Plot, we can visualize an entire 135 Mbp genome of *Arabadopsis thaliana* in under 2 minutes on a single laptop. This represents a significant speedup over StainedGlass, which takes over an hour to produce the same plot, while maintaining near-identical topology. We further show that by removing the dependence of sequence alignment, ModDotPlot more accurately reflects the similarity of the high order repeats present in human centromeres.

Finally, we bundle ModDotPlot as a Python package with an interactive graphical user interface using Plotly. This interface allows users to adjust genomic coordinates by panning and zooming, and receive an updated dot plot immediately. Mod.Plot is open source and available at github.com/marbl/ModDotPlot.

ESTIMATING BACKGROUND PROTEIN SIGNALS TO ENHANCE DATA NORMALIZATION IN CITE-SEQ

Cerag Oguztuzun[1], Tamas Ryszard Sztanka-Toth[2], Alex Javidi[1]

[1]Data Science and Digital Health, Janssen R&D, Johnson & Johnson, Spring House, PA, [2]Data Science and Digital Health, Janssen R&D, Johnson & Johnson, Neuss, Germany

Multimodal single-cell profiling approaches, such as CITE-seq [1], have emerged as powerful tool for characterizing cellular heterogeneity in almost all RNA biology study areas by profiling surface proteins and transcriptomes simultaneously. As every CITE-Seq experiment generates massive amount of data at single cell resolution, this approach provides insights into cell types and states, particularly in the immune system where surface protein profiles are associated with cellular subsets and functions. To accurately analyze single-cell data, effective normalization methods that consider technical variations are necessary. As an initial normalization technique Stoeckius et al. suggested using the centered log-ratio (CLR) transformation as it preserves the compositionality of the data [1], however, CLR does not remove the background signal arising from free floating antibodies within droplets. Removing background noise from actual signal proteins is crucial for normalization as background proteins introduce variability and could distort valid biological signals. In 2022 Mulè et al. demonstrated that protein counts in empty droplets can estimate the expected background protein signals associated with each antibody using ‘denoised and scaled by background’ (dsb) normalization [2]. Our project had two main aspects. Firstly, by using publicly available CITE-seq data from 10x Genomics, we sought to analyze dsb normalization and compared it with CLR based on how well they perform on background protein signal elimination. Secondly, we propose a variational-autoencoder (VAE) framework which tries predict the background protein signal mean and standard deviation based on the foreground protein distribution.

[1] Stoeckius, M., Hafemeister, C., Stephenson, W. et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods 14, 865–868 (2017). https://doi.org/10.1038/nmeth.4380
[2] M. P. Mulè, A. J. Martins, and J. S. Tsang, “Normalizing and denoising protein expression data from droplet-based single cell profiling,” Nature Communications, vol. 13, no. 1, 2022. doi:10.1038/s41467-022-29356-8

NEW COMPUTATIONAL METHODS FOR THE ANALYSIS OF TRANSCRIPTION FACTOR CUT&RUN DATA

Luomeng Tan, Coby Viner, Michael M Hoffman

University Health Network, Toronto, Canada

Introduction. Identifying sequence-specific transcription factor binding sites (TFBSs) is crucial for understanding the regulation of gene expression. Cleavage Under Targets and Release Using Nuclease sequencing (CUT&RUN-seq) offers a cost-effective improvement over Chromatin Immunoprecipitation sequencing (ChIP-seq), requiring significantly fewer cells while maintaining a higher signal-to-noise ratio and lower background. Existing computational tools for CUT&RUN-seq are often poorly adapted from ChIP-seq methods, leading to suboptimal TFBS elucidation.

Methods. To refine the elucidation of transcription factor motifs from CUT&RUN datasets, we conducted a comprehensive evaluation of existing computational tools tailored for CUT&RUN analyses. Utilizing both public data and some currently unpublished datasets from our collaborators, we scrutinized the impact of multiple variables on motif elucidation. These variables included fragment length, the use of spike-in controls, the selection of peak callers, and the handling of low-quality reads.

Results. We developed an automated pipeline that processes raw CUT&RUN-seq data and outputs quality control metrics and comprehensive motif analyses to optimize downstream analyses while improving reproducibility and ensuring ease-of-use. We used the pipeline to analyze CUT&RUN data from various labs, including samples from human and mouse cells. From our benchmarking results, we found that filtering for fragments with a length of less than or equal to 120 bp and spike-in calibration usually improved motif enrichment. Two peak callers, MACS2 and SEACR, showed advantages on data in different conditions, and while SEACR tended to work more poorly overall, it was better for samples with a smaller number of cells. Removing reads with poor mapping qualities was also essential to improve motif elucidation, with SEACR. We are actively developing easy-to-use software that optimizes transcription factor motif elucidation using CUT&RUN data, by refining our pipeline. We plan to extend it to call statistically significant peak regions, making judicious use of spike-in controls, as an integrated algorithmic refinement.

Conclusion. Our benchmarking compared existing computational methods for CUT&RUN data on TFBS elucidation. Using this information, we refined our integrated computational method for CUT&RUN data processing and improved TFBS elucidation. Our work allows for improved analyses of CUT&RUN datasets, allowing researchers to more rapidly and reliably utilize this new technology, towards bettering our understanding of downstream genome regulation.

< Back

LESS IS MORE: TRIMMING LONG MASSIVELY PARALLEL SEQUENCE READS CAN IMPROVE MAPPING RATES AND DEPTH OF COVERAGE

Jamie K Teer[1], Jane C Figueiredo*[2], Stephanie L Schmit*[3,4]

[1]H. Lee Moffitt Cancer Center, Department of Biostatistics and Bioinformatics, Tampa, FL, [2]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Department of Medicine, Los Angeles, CA, [3]Lerner Research Institute, Cleveland Clinic, Genomic Medicine Institute, Cleveland, OH, [4]Case Comprehensive Cancer Center, Population and Cancer Prevention Program, Cleveland, OH

Advances in massively-parallel sequencing technology continue to increase sequence generation rates and decrease costs. Improvements in chemistry, flowcell layout, and image analysis have increased the length of high quality reads. This enables greater average depth of coverage that is critical for variant detection experiments, including our ongoing work addressing cancer heath disparities through the Latino Colorectal Cancer Consortium (LC3). We have recently identified examples of lower-than-expected human genome mapping rates from 150 base pair paired-end whole exome sequencing results. This experiment included 38 FFPE-derived LC3 DNA samples across a range of age, quality, and quantity. Interestingly, trimming the reads to shorter lengths resulted in increased mapping rates as well as increased average depth of coverage in many samples. Trimming to 100 base pairs resulted in median depth of coverage fold change of 2.19 (min, max = 0.96, 5.15). Trimming further to 75bp resulted in an even higher median depth of coverage fold change of 2.38 (min, max = 0.77, 7.66). We did observe a decrease in average depth of coverage after trimming (fold change <1) in some samples, although these samples generally had higher overall coverage depths. We also observed examples where trimming did not improve mapping or depth of coverage but did result in fewer errors with certain downstream analysis tools. We hypothesize that longer reads may result in artificial sequences when read length exceeds fragment size, and these false sequences may result in poor or non-alignment. Therefore, there is an opportunity to potentially improve sequencing results quality from challenging FFPE samples by trimming reads.

To determine when trimming may be beneficial, we are currently developing a tool to test a fraction of reads across a variety of trim lengths to determine the optimal (if any) trimming to apply to a sample. This tool is being developed in the Workflow Descriptor Language using docker containers for portability and will be released publicly under an open source software license.

*JCF and SLS contributed equally.

GUIDING SINGLE-CELL RNA-SEQ CLUSTERING WITH RANK-BASED METRICS

Christopher V Thai, Hossein Khiabanian, Daniel Herranz

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

Single-cell RNA sequencing measures gene expression of individual cells in a tissue sample. Cells are first grouped into clusters based on how similarly they express different genes. Then, differentially expressed genes in each cluster relative to other clusters are used as marker genes to assign functional identities to cell populations. Parameters of common clustering algorithms, such as the resolution in Leiden clustering, determine the number and membership of inferred clusters and are applied globally across an entire dataset, despite the possible presence of distinct cell subpopulations with varying size. As a result, clustering algorithms can be confounded by measurement noise, especially in genes specifically expressed in these subpopulations, leading to over- or under-clustering and mischaracterization of novel cell populations and their marker genes.

Here, we demonstrate that rank-based metrics quantify the similarity of marker gene lists in a biologically meaningful way by considering both their overlap and the order in which genes are differentially expressed. We show that this quantification of marker gene list similarity can identify which cell clusters may have the same functional identity and propose a workflow that utilizes average overlap, a rank-based metric, to inform and guide identification of unique cell identities. We find that average overlap is normal-like when calculated over randomly shuffled ranked lists. This empirical distribution is used to calculate the probability that two clusters may arise from a single functional cell identity, which is then used as a criterion for subsequent unification of clusters. We then evaluate the approach using simulated and “truth-known” data in the presence of over-clustering due to stochastic noise in gene expression measurements and arbitrarily high-resolution clustering parameters.

VARIANT ANALYSIS IN AN INBRED RAT POPULATION – A LESSON FROM THE HYBRID RAT DIVERSITY PANEL

Monika Tutaj[1], Akiko Takizawa[1], Lynn Malloy[1], Rebecca Schilling[1], Kent C Brodie[2], Jeffrey L De Pons[1], Wendy M Demos[1], Thomas G Hayman[1], Mary L Kaldunski[1], Stan J Laulederkind[1], Jennifer R Smith[1], Marek A Tutaj[1], Mahima Vedi[1], Shur-Jen Wang[1], Anne E Kwitek[1], Melinda R Dwinell[1]

[1]Medical College of Wisconsin, Physiology, Milwaukee, WI, [2]Medical College of Wisconsin, Clinical and Translational Science Institute, Milwaukee, WI

Currently, many human genetic studies identify disease-associated loci and variants using whole exome sequencing (WES)/whole genome sequencing (WGS) methods. However, the associated variants are often of unknown significance and/or not even within gene coding regions, making identifying causal variants challenging. Therefore, researchers utilize multiple model organisms, different genetic backgrounds and environmental stressors to link the variants to orthologous genes, pathways, molecular networks and eventually disease phenotypes. The Rat Genome Database provides information on genomic variants across laboratory rat strains for all rat genome assemblies and integrates it with strain phenotype data to aid in interpretation of the variants. 96 strains from the Hybrid Rat Diversity Panel (HRDP) were sequenced and analyzed at MCW as part of the Hybrid Rat Diversity Program, a resource to rederive classic and recombinant inbred rat strains, sequence their genomes to identify variations, and make the data and the strains accessible to research community. In 2020, we used mRatBN7 and Rnor_6.0 to assess the detection rate of homozygous, heterozygous, genic, and intergenic single nucleotide variants (SNVs) and short indels from WGS of 47 HRDP strains. We also performed variant discovery using non-reference rat genomes assembled and released in 2022 (SHRSP/BbbUtx, WKY/Bbb). All datasets were analyzed with Genome Analysis Toolkit from the Broad Institute, designed and optimized for human data. There are no existing recommendations for variant discovery in non-human, inbred populations like rat with low effective sizes. In addition, the systematic comparisons of variant callers were not conducted in rat populations, so we lack the truth and training datasets to evaluate variant call accuracy and therefore to efficiently exclude false positive calls. We propose strategies for selecting and prioritizing candidate variants for disease model studies. We differentiate variants present in likely misassembled genomic regions in the BN reference, in repetitive regions, and in regions with accumulation of heterozygous variants. Thus, scientists can identify potential disease-causing mutations in the context of RGD integrated multi-species data, have comparative insight in data alignment and distribution (JBrowse2, VCMap) and prioritize variants for validation in the available HRDP strains. Finally, they can confirm if the elected SNVs and indels occur in QTLs and genes associated with disease phenotypes, and have impact on transcripts (SnpEff) and proteins (PolyPhen2) shared between strains representing the same disease model.

ACCURATE COMPARISON OF INSERTION AND DELETION MUTATION RATES USING SEQUENCE COMPOSITION CORRECTION WITH NOVEL SEQUENCE AMBIGUITY SCORING

Jan C Verburg, Martin S Taylor

Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom

Dysregulated cellular processes and exposure to exogenous agents are commonly attributed to cancer, and the patterns of mutations associated can be summarised as ‘signatures’. Constructing mutational signatures is becoming increasingly important in the clinical treatment of patients. Though insertions and deletions contribute to a significant proportion of the mutations in cancer, much of the research into signatures has been restricted to single nucleotide substitutions due to their consistent trinucleotide context. Indels are folded into 83 distinct categories when creating the indel mutational signature due to varying event size and sequence context length. Indel mutational signatures only allow fair comparison of regions with the same sequence composition. This impedes investigating underlying mechanisms of indel-causing processes. Without a null expectation or other genomic region to serve as comparison, it is impossible to make a meaningful deduction from observed mutations. This underlines a need for an approach to account for variability in sequence composition in order to make accurate comparison of indel mutation rates across the genome and better resolve aetiology of the events and processes leading to cancer. We propose a novel framework and metric that quantifies sequence ambiguity inherent in indel alignment to systematically score sequence context. In combination with sequence size, we retain sequence identity across each of the ID83 categories. This sequence ambiguity scoring framework provides means for sequence composition correction and to generate null expectations from any sequence. Moreover, it allows for compositionally adjusted indel rates to be compared between genomic regions including between species. This has previously been impossible when using indel mutational signatures. We show that compositionally adjusted indel rates in ID83 format allow for direct comparison between coding and non-coding regions and how insertion and deletion rates relate to each other on leading and lagging strands. Additionally, we show compositionally adjusted indel rates in early/late replicating regions in mismatch repair deficient tumours. In short, systematically scoring indels for sequence ambiguity opens up new avenues to explore context specific mechanisms of indel causing mutational processes.

PBFUSION: DETECTING GENE FUSIONS AND OTHER TRANSCRIPTIONAL ABNORMALITIES USING PACBIO HIFI DATA

Roger Volden[1], Daniel Baker[1], Zev Kronenberg[1], Aaron Gillmor[2], Ted Verhey[2], Michael Monument[2,3], Donna Senger[2,4], Harsharan Dhillon[1], Jason Underwood[1], Elizabeth Tseng[1], Primo Baybayan[1], Michael A Eberle[1], Jonas Korlach[1], Sorana Morrissy[2]

[1]PacBio, Menlo Park, CA, [2]University of Calgary, Charbonneau Cancer Institute, Calgary, Canada, [3]University of Calgary, Department of Surgery, Calgary, Canada, [4]McGill University and the Lady Davis Institute for Medical Research, Gerald Bronfman Department of Oncology, Montreal, Canada

Sarcomas are a broad group of soft tissue and bone cancers that can be difficult to treat. Sarcomas comprise two broad genomic classes: (1) simple karyotypes, where a single oncogenic structural variant clonally expands a subtype that is relevant to tumor burden tracking; and (2) complex karyotypes, genomic instability, where SVs continuously arise throughout tumor evolution resulting in heterogeneous cellular subtypes. Class two sarcomas are harder to characterize using genome sequencing because there may be multiple low-frequency mutations. Accurate and sensitive detection of fusion transcripts is needed to interpret functional consequences, to understand tumor biology and evolution, and potentially identify new targets for therapy. Many fusions have complex structures that cannot be uniquely resolved using short reads due to a lack of exon connectivity. PacBio full-length RNA isoform sequencing resolves complex fusions, providing more accurate breakpoints, and a complete sequence readout of the associated fusion transcript. Recent advances in bulk RNA library preparation due to concatenation with MAS enable us to find more fusions. Here we present a fusion detection tool, pbfusion, specifically designed for HiFi sequence data, and apply it to sarcoma samples from both classes. pbfusion converts mapped HiFi sequences into transcript objects that are annotated with reference gene models. Annotations determine whether transcripts are discordantly mapped, overlap differing genes, strand swap, transcriptional readthrough, or contain novel exons. The discordant exonic boundaries are treated as breakpoints between two genomic locations. All breakpoints are clustered with a multi-directional chaining algorithm and annotated with exonic information, gene names, and quality information.

To test our method, we applied pbfusion to 12 samples from 8 sarcoma patients. We discovered both known and novel fusions, including validated driver events in the fusion-driven samples (e.g., ASPSCR1-TFE3 in alveolar soft part sarcoma and SS18-SSX2/1 fusion in synovial sarcoma). We were also able to phase these fusion transcripts for allele-specific isoform calling and ORF prediction. This approach demonstrates the utility of HiFi sequence data for identification of fusion transcripts in samples, and the use of pbfusion in quantifying and annotating these events for future neoantigen detection. pbfusion provides a user-friendly interface, can process a sample in a few minutes, and is freely available to the research community on Bioconda.

FAIR BIOHEADERS: FAIR HEADER REFERENCE GENOME (FHR)

Adam J Wright[1], David C Molik[2]

[1]Ontario Institute for Cancer Research, Adaptive Oncology, Toronto, Canada, [2]USDA ARS, Abadru, Manhattan, KS

The increasing publication of omics data across diverse platforms increases the likelihood of data modification (intentional and unintentional) [Grossman 2019], posing a threat to interoperability and future relevance. This trend is linked to the decentralization of data publishing, which often leads to multiple incompatible systems and standards. Decentralization, while inevitable in scientific research, introduces technical and policy challenges and threatens the archivability of data. A key component to solving these problems is data standards that consider how easy it is to adopt the new technology, given current operating procedures. Data standards with simple and easy-to-use principles have a high return with respect to the investment of time in implementation and the amount of return in data quality. It should be noted that usability and returns are measurable metrics that can be used in metadata implementation.

To address these challenges in one single specific data type, in this case, the reference genome assembly, we introduced the Fair Header Reference genome (FHR), aiming to enhance interoperability and archivability in reference assemblies. FHR goes beyond providing a specification, offering multiple serializations, validation software, and conversion tools. It enables users to access core information and related resources within FASTA files easily, ensuring data stability and accuracy across platforms. FHR computationally verifies genomes, records data provenance, and supports multiple data formats, with a preference for FASTA headers that include metadata. This design aligns with the often overlooked FASTA comment feature, improving compatibility with existing tools. FHR adheres to the principles of FAIR and TRUST [Lin 2020, Wilinkson 2016], which guide the implementation of data-centric software. However, FHR faces adoption challenges, such as limited FASTA comment support in current libraries and a time cost for metadata formatting. Efforts are underway to address these challenges, including creating pull requests for library updates and developing tools to streamline FHR data formatting and transfer. Collaborating with early adopters is crucial to achieving broad adoption, as the biggest obstacle to adoption is that interoperability efforts only work if multiple parties agree to adopt the standard [Rocca Serra 2015].

https://github.com/FAIR-bioHeaders

[Grossman 2019] doi.org/10.1016/j.tig.2018.12.006
[Lin 2020] doi.org/10.1038/s41597-020-0486-7
[Wilinkson 2016] doi.org/10.1038/sdata.2016.18
[Rocca Serra 2015] doi.org/10.1007/s11306-015-0879-3

EXAMINING CHROMATIN HETEROGENEITY THROUGH PACBIO LONG-READ SEQUENCING OF M.ECOGII METHYLATED GENOMES: AN M^{6}A DETECTION EFFICIENCY AND CALLING BIAS CORRECTING PIPELINE

Zhuwei Xu, Allison F Dennis, David J Clark

NIH, National Institute of Child Health and Human Development,
Bethesda, MD

Recent studies have combined DNA methyltransferase footprinting of genomic DNA in nuclei with long read sequencing, resulting in detailed chromatin maps for multi-kilobase stretches of genomic DNA from one cell. Theoretically, nucleosome footprints and nucleosome-depleted regions can be identified using M.EcoGII, which methylates adenines in any sequence context, providing a high-resolution map of accessible regions in each DNA molecule. Here we report PacBio long-read sequence data for budding yeast nuclei treated with M.EcoGII and a bioinformatic pipeline which corrects for three key challenges undermining this promising method. First, detection of m^{6}A in individual DNA molecules by the PacBio software is inefficient, resulting in false footprints predicted by random gaps of seemingly unmethylated adenines. Second, there is a strong bias against m^{6}A base calling as AT content increases. Third, occasional methylation occurs within nucleosomes, breaking up their footprints. After correcting for these issues, our pipeline calculates a correlation coefficient-based score indicating the extent of chromatin heterogeneity within the cell population for every gene. Although the population average is consistent with that derived using other techniques, we observe a wide range of heterogeneity in nucleosome positions at the single-molecule level, probably reflecting cellular chromatin dynamics.

SCREADSIM: A SINGLE-CELL RNA-SEQ AND ATAC-SEQ READ SIMULATOR

Guan'ao Yan[1], Dongyuan Song[2], Jingyi Jessica Li[1,2]

[1]University of California, Los Angeles, Department of Statistics, Los Angeles, CA, [2]University of California, Los Angeles, Bioinformatics, Los Angeles, CA

Benchmarking single-cell RNA-seq (scRNA-seq) and single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) computational tools demands simulators to generate realistic sequencing reads. However, none of the few read simulators aim to mimic real data. To fill this gap, we introduce scReadSim, a single-cell RNA-seq and ATAC-seq read simulator that allows user-specified ground truths and generates synthetic sequencing reads (in FASTQ and BAM formats) by mimicking real data. At both read-sequence and read-count levels, scReadSim mimics real scRNA-seq and scATAC-seq data. Moreover, scReadSim provides ground truths, including unique molecular identifier counts for scRNA-seq and open chromatin regions for scATAC-seq. In particular, scReadSim allows users to design cell-type-specific ground-truth open chromatin regions for scATAC-seq data generation. In benchmark applications of scReadSim, we show that UMI-tools achieves better accuracy in scRNA-seq UMI deduplication, and HMMRATAC and MACS3 achieve top performance in scATAC-seq peak calling.

A METAGENOMICS GENOME-PHENOME ASSOCIATION (METAGPA) STUDY REVEALS 2-AMINOADENINE (DZ) BIOSYNTHETIC PATHWAY IN UNCULTURABLE PHAGE METAGENOME

Weiwei Yang, Shuangyong Xu, Laurence Ettwiller

New England Biolabs, Research Department, Ipswich, MA

In the evolutionary arm race between phage and host, phages develop a variety of strategies to combat the host-encoded antiphage defensive mechanisms. Full modification of the genome is one typical strategy used by phages to prevent cleavage by host restriction endonucleases. Understanding the different types of DNA modification carried by phage microbiomes and how these modifications are synthesized and maintained in the phage genomes inspire innovations and applications in nucleic acid products and biotechnology. Here, we applied our previously established metagenomics genome-phenome association (metaGPA) pipeline to study phage metagenomes containing 2-aminoadenine (diaminopurine or dZ) from environmental samples. We established a next generation sequencing method to selectively enrich dZ genome sequences and through this study, we gain access to thousands of unculturable, novel dZ containing phage genomes. We then applied MetaGPA association analysis to scrutinize protein domains associated with dZ genomes. These protein domains are potentially involved in production and maintenance of dZ genome. We were able to recapitulate the known phage biosynthetic pathway for the production of dZ genome demonstrating that our approach is selecting for dZ containing phages. Futhermore, using co-occurrence network, we identify a potentially novel component in the phage dZ pathway. In addition, we performed metaGPA differential residue association analysis on two conversed genes in related phage genomes: PurZ (homolog of adenylosuccinate synthetase) and DpoZ (DNA polymerase in dZ genome) sequences from de novo and public databases. Our findings implicate key residues which may play roles on the dZ vs dA specificity.

STATOR: HIGH ORDER EXPRESSION DEPENDENCIES FINELY RESOLVE CRYPTIC STATES AND SUBTYPES IN scRNA-SEQ DATA

Yuelin Yao[1,2], Abel Jansma[2,4], Jareth Wolfe[2], Luigi Del Debbio[4], Sjoerd Beentjes[2,3], Chris Ponting[2], Ava Khamseh[1,2,4]

[1]University of Edinburgh, School of Informatics, Edinburgh, United Kingdom, [2]University of Edinburgh, Institute of Genetics and Cancer, Edinburgh, United Kingdom, [3]University of Edinburgh, School of Mathematics, Edinburgh, United Kingdom, [4]University of Edinburgh, Higgs Centre for Theoretical Physics, School of Physics and Astronomy, Edinburgh, United Kingdom

Advances in scRNA-seq techniques are resolving cell (sub)types among complex cell populations. However, conventional approaches are less able to reveal the continuous spectrum of cell states, such as cell cycle phases, in large part due to a reliance on proximity in reduced dimensional gene expression space. We introduce *Stator*, a novel method that finely resolves cell types, subtypes and states among cells whose transcriptomes appear homogeneous upon clustering. *Stator* takes advantage of lowly-expressed as well as not-expressed genes, and can identify rare biological states (~0.2% of 10k single cells). The approach: (i) applies a model-free estimator of higher-order interactions to quantify expression dependencies amongst n-tuples of genes (beyond pair-wise), (ii) extracts significantly deviating gene combinations (*tuples*) driving these higher-order gene dependencies, and finally (iii) combines tuples into *Stator* states when they commonly co-occur in the same cell. Typically, *Stator* labels a cell not just by type and sub-type but also by biological state, for example an immature interneuron in G2/M cell cycle phase. *Stator* generates molecular and cellular hypotheses for subsequent experimental testing. To facilitate this, we provide a *Stator* Shiny App (https://shiny.igc.ed.ac.uk/MFIs/). This flexible app takes the output of *Stator*'s Nextflow pipeline and - through an interactive and user-friendly interface - performs downstream analyses such as (i) comparing *Stator* state labels with external annotations or experimental conditions, (ii) calculating differential gene expression per *Stator* state, (iii) automatically annotating states using a user-provided gene list, and (iv) analysing Gene Ontology and KEGG Pathway enrichment for tuples' genes and differentially expressed genes. We demonstrate *Stator*'s ability to extract biologically meaningful cell (sub)types and states at far greater resolution than hitherto using publicly available scRNA-seq datasets of mouse developmental astrocytes or neurons, and a hepatocellular carcinoma dataset.

MODEL-BASED CHARACTERIZATION OF THE EQUILIBRIUM DYNAMICS OF TRANSCRIPTION INITIATION AND PROMOTER-PROXIMAL PAUSING IN HUMAN CELLS

Yixin Zhao[1], Lingjie Liu[1,2], Rebecca Hassett[1], Adam Siepel[1,2]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Stony Brook University, Graduate Program in Genetics, Stony Brook, NY

In metazoans, both transcription initiation and the escape of RNA polymerase (RNAP) from promoter-proximal pausing are key rate-limiting steps in gene expression. These processes play out at physically proximal sites on the DNA template and appear to influence one another through steric interactions. Here, we examine the dynamics of these processes using a combination of statistical modeling, simulation, and analysis of real nascent RNA sequencing data. We develop a simple probabilistic model that jointly describes the kinetics of transcription initiation, pause-escape, and elongation, and the generation of nascent RNA sequencing read counts under steady-state conditions. We then extend this initial model to allow for variability across cells in promoter-proximal pause site locations and steric hindrance of transcription initiation from paused RNAPs. In an extensive series of simulations, we show that this model enables accurate estimation of initiation and pause-escape rates. Furthermore, we show by simulation and analysis of real data that pause-escape is often strongly rate-limiting and that steric hindrance can dramatically reduce initiation rates. Our modeling framework is applicable to a variety of inference problems, and our software for estimation and simulation is freely available.

INVESTIGATING MOSAIC STRUCTURAL VARIATIONS ACROSS THOUSANDS OF GENOMES WITH STIX

Xinchang Zheng[1], Ryan M Layer[2,3], Fritz J Sedlazeck[1,4]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]University of Colorado, BioFrontiers Institute, Boulder, CO, [3]University of Colorado, Department of Computer Science, Boulder, CO, [4]Rice University Computer Science Department, Department of Computer Science, Houston, TX

Structural variants (SVs) are strongly associated with various diseases, such as cancer and neurological disease. Importantly, somatic SVs play crucial roles in diseases' progression and recurrence. However, the prevalence and impact of somatic SVs have not been fully studied. Even though the remarkable explosion of high-throughput sequencing data made it possible to investigate mosaic SVs in large populations. The vast growth of long read sequencing enables us to utilize these technologies also for annotation of SV. This is required as long-read sequencing often better represents SV and variants in tandem repeats. To accomplish this we have extended STIX to work with long read sequencing and extended its concept to genotype insertions. Over the GIAB HG002 benchmark we reach 89.7 (Pacbio) and 88.52% (ONT) re-identification rates with only smaller rates of false positives of 2.8% when genotyping HG0733 in HG002 data.

Utilizing STIX we now further investigate how it can be leveraged to further not only annotate germline SV but also mosaic SV (ie. low frequency mutations) in the population. For this purpose we are investigating how and if mosaic SV are shared across the human population. These alleles could highlight common instabilities in the human genome but further will help to correctly annotate and identify pathogenic mosaic mutation. We identified multiple mosaic mutations that are indeed shared across multiple individuals, especially recombinant deletions. These seem to be frequency co-occurring with ALU insertions and thus repeat mediated recominations that we can now routinely annotate using STIX. We found these events across brain tissues, cell lines and other samples that we already indexed over STIX.

In our presentation we will also highlight our plans to utilize STIX across the SMaHT consortia to index and build a multiple tissue catalog of SV. This will be important and helpful to annotate not only germline and mosaic SV but also will help to prioritize cancer mediated mutations without matching normal tissues or blood. Overall STIX can now be applied across short and long read catalogs across all SV types.

DATA BEATS MACHINE LEARNING FOR GENOME ANNOTATION

Aleksey V Zimin

Johns Hopkins University, Biomedical Engineering, Baltimore, MD

For many years *de novo* gene finding approaches based on machine learning were dominant in gene annotation of eukaryotic genomes. These approaches are utilized in popular annotation packages such as BRAKER, MAKER, and GENEMARK. High cost and low availability of gene expression data were the main reasons why machine learning approaches were needed in the past. Several transcriptome sequencing technologies were introduced in the recent years, such as RNA-seq by Illumina, ISOseq by PacBio and RNA/CDNA sequencing by Oxford Nanopore. These technologies can yield abundant gene expression data at low cost. We found that current annotation packages are unable to utilize the data to its full potential, and thus we developed a novel data-driven eukaryotic gene annotation software called EviAnn. EviAnn outperforms current state-of-the-art genome annotation approaches such as MAKER2 and BRAKER3 in speed and quality. EviAnn is purely data-driven, it does not utilize any *de novo* gene finding techniques. It utilizes protein homology instead. We show that with the same input data EviAnn outperforms current state-of-the-art packages such as BRAKER3, while utilizing less computer time. EviAnn has few dependencies, and thus it is easy to configure and install. EviAnn is an open source software, available at https://github.com/alekseyzimin/EviAnn_release/releases .

NOTES

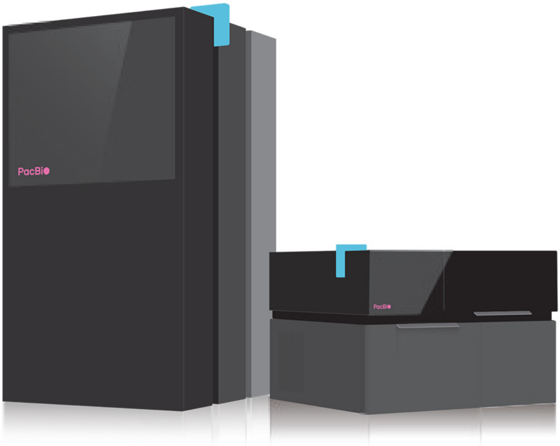
NOTES

NOTES

NOTES

Say hello to

Revio + Onso



Get ready to make discoveries that will change the world

Scan the QR code and explore the unparalleled accuracy of our new long-read and short-read sequencing systems.

Participant List

Mr. Dominic Absolon
Welcome Sanger Institute
da16@sanger.ac.uk

Dr. Eren Ada
Harvard Medical School
eren_ada@hms.harvard.edu

Jenea Adams
University of Pennsylvania
jiadams@pennmedicine.upenn.edu

Tanveer Ahmad
NIH/NCI
tahashmipk@gmail.com

Mr. Omar Ahmed
Johns Hopkins University
omaryfekry@gmail.com

Mattea Allert
University of Minnesota
alle0499@umn.edu

Dr. Fatemeh Almodaresi
Ontario Institute for Cancer Research (OICR)
falmodaresi@oicr.on.ca

Ms. Anusha Aman
University of Oxford
anusha.aman@oncology.ox.ac.uk

Dr. Dmitry Antipov
NHGRI
dmitrij.antipov@gmail.com

Dr. Mona Arabzadeh
Rutgers Biomedical and Health Science
mona.arabzadeh@rutgers.edu

Mr. Byron Avihai
Rutgers' Cancer Institute of New Jersey
ba365@rwjms.rutgers.edu

Morteza Baradaran
University of Virginia
rgq5aw@virginia.edu

Mr. Kassaye Belay
Virginia Tech
kassayeb@vt.edu

Dr. Daniel Blankenberg
Cleveland Clinic
dan.blankenberg@gmail.com

Mr. Michael Bradshaw
CU Boulder
michael.bradshawIii@colorado.edu

Dr. Michael Broe
The Ohio State University
broe.1@osu.edu

Dr. Amanda Brown
Texas Tech University
amanda.mv.brown@ttu.edu

Asher Bryant
National Cancer Institute
asher.bryant@nih.gov

Carlos Buen Abad Najar
University of Chicago
cnajar@uchicago.edu

Dr. Lei Cao
Icahn School of Medicine at Mount Sinai
lei.cao@mssm.edu

Dr. Jennifer Capaccio
Regeneron Pharmaceuticals
jennifer.capaccio@regeneron.com

Dr. Tiago Capote
New York University Abu Dhabi
tcc414@nyu.edu

Dr. Anne Carpenter
Broad Institute of MIT and Harvard
anne@broadinstitute.org

Dr. Javier Carpinteyro-Ponce
Carnegie Institution for Science
jcarpinteyro@carnegiescience.edu

Ms. Sheridan Cavalier
Johns Hopkins School of Medicine Dept of Neuroscie
scavali1@jh.edu

Dr. Ling Cen
Moffitt Cancer Center
ling.cen@moffitt.org

Mr. Kuan-Hao Chao
Johns Hopkins University
kuanhao.chao@gmail.com

Ms. Yuhang Chen
Yale University
yuhang.chen@yale.edu

Dr. Nae-Chyun Chen
Exai Bio
naechyunc@exai.bio

Dr. Haoyu Cheng
Dana-Farber Cancer Institute & HMS
hcheng@ds.dfci.harvard.edu

Dr. Chao Cheng
Baylor College of Medicine
chao.cheng@bcm.edu

Dr. Rayan Chikhi
Institut Pasteur
rayan.chikhi@pasteur.fr

Dr. Jinmyung Choi
Icahn School of Medicine at Mount Sinai
jinmyung.choi@mountsinai.org

Ms. Eunwoo Choi
Yonsei University College of Medicine
ehfkdpahd65@naver.com

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

Dr. Hiram Clawson
U.C. Santa Cruz Genomics Institute
hclawson@ucsc.edu

Mr. Nathan Coraor
Penn State University
nate@bx.psu.edu

Kristen Curry
Rice University
kdc10@rice.edu

Dr. Daisy Dahiya
National Institutes of Health
daisy.dahiya@nih.gov

Dr. Jieqiong Dai
Roche Sequencing & Life Science
jieqiong.dai@roche.com

Prof. Mehdi Damaghi
Stony Brook University
mehdi.damaghi@stonybrookmedicine.edu

Nikhita Damaraju
University of Washington
nikhita@uw.edu

Mr. Arun Das
Johns Hopkins University
arun.das@jhu.edu

Dr. Erwin Datema
Keygene N.V.
ellen.vaessen@keygene.com

Dr. John Davis
Johns Hopkins University
jdavcs@gmail.com

Prof. Ramana Davuluri
Stony Brook University
Ramana.Davuluri@stonybrookmedicine.edu

Zhiqian Deng
University of Californiam, Santa Cruz
zdeng7@ucsc.edu

Prof. Michael Deyholos
Univ British Columbia, Okanagan
michael.deyholos@ubc.ca

Dr. Paulo Dias
iBB - Institute for Bioengineering and Biosciences
pjdias@tecnico.ulisboa.pt

Mr. Gerardo Diaz Ortiz
University of Minnesota
diazo005@umn.edu

Zhiyi Dong
Stony Brook University
zhiyi.dong@stonybrook.edu

Mr. Ataberk Donmez
National Institutes of Health
ataberk@umd.edu

Dr. Veronika Dubinkina
Gladstone Institutes
veronika.dubinkina@gladstone.ucsf.edu

Carolyn Dunlap
Metagenomi
carolyn.dunlap@metagenomi.co

Philip Ebert
Eli Lilly and Company
ebert_philip_j@lilly.com

Dr. Sean Eddy
HHMI/Harvard University
seaneddy@fas.harvard.edu

Ms. Melise Edwards
University of Massachusetts Amherst
medwards@umass.edu

Dan Ehninger
DZNE
dan.ehninger@dzne.de

Basak Eraslan
Stanford University
eraslab1@gene.com

Beril Erdogdu
The Johns Hopkins University
berdogd1@jhu.edu

Mr. Edward Esiri-Bloom
University of Edinburgh
s1909930@ed.ac.uk

Xiao Fan
University of Florida
xiaofan@ufl.edu

Dr. Xiaowen Feng
Dana-Farber Cancer Institute
xfeng@ds.dfci.harvard.edu

Mr. Logan Fink
DGS Division of Consolidated Laboratory Services
logan.fink@dgs.virginia.gov

Mr. Vinicius Franceschini-Santos
The Netherlands Cancer Institute
v.franceschini@nki.nl

Hildreth Frost
Dartmouth College
hildreth.r.frost@dartmouth.edu

Dr. Minakshi Gandhi
Cold Spring Harbor Laboratory
mgandhi@cshl.edu

Mr. Teng Gao
Harvard Medical School
tgao@g.harvard.edu

Mr. Peter Ge
Johns Hopkins School of Medicine
yge15@jhmi.edu

Ms. Tatiana Gelaf Romer
BlueRock Therapeutics
tgelafromer@bluerocktx.com

Ms. Sara Geraghty
Princeton University
scamilli@princeton.edu

Dr. Edoardo Giacopuzzi
Human Technopole Foundation
edoardo.giacopuzzi@fht.org

Ms. Sophia Gibson
University of Washington
sophiabg@uw.edu

Ms. Yomna Gohar
Heinrich-Heine-Universitat Dusseldorf
gohar.yomna@gmail.com

Anton Goretsky
National Institutes of Health
anton.goretsky@nih.gov

Ms. Pravallika Govada
Vellore Institute of Technology
pravallika.govada2018@vitstudent.ac.in

Dr. Anna Green
University of Massachusetts Amherst
annagreen@umass.edu

A. Burak Gulhan
The Pennsylvania State University
gulhan@psu.edu

Yujie Guo
Dana Farber Cancer Institutue
yguo@ds.dfci.harvard.edu

Mr. Yoritaka Harazono
the University of Tokyo
harazono_yoritaka_17@stu-cbms.k.u-tokyo.ac.jp

Mr. Andrew Harris
Texas A&M University
ajharris@cvm.tamu.edu

Dr. Michael Heskett
Stanford University
heskett@stanford.edu

Dr. Allison Hickman
EpiCypher
ahickman@epicypher.com

Mr. Parker Hicks
University of Colorado Anschutz Medical Campus
parker.hicks@cuanschutz.edu

Celine Hoh
Johns Hopkins University
celinehohzm@gmail.com

Laurenz Holcik
University of Vienna
laurenz.holcik@univie.ac.at

Dr. Neng Huang
Dana Farber Cancer Institute
neng@ds.dfci.harvard.edu

Ms. Rongting Huang
The University of Hong Kong
rthuang@connect.hku.hk

Mr. Win Hung
Bayside High School
winhung@aol.com

Sanjida Huseni Rangwala
NCBI/NLM
rangwala@nih.gov

Mr. Christopher Husted
UMASS Chan Medical School
christopher.husted@umassmed.edu

Stephen Hwang
Johns Hopkins University
shwang45@jh.edu

Dr. Parisa Imanirad
Bristil Myers Squibb
parisa.imanirad@bms.com

Mr. Luiz Carlos Irber Junior
University of California Davis
lcirberjr@ucdavis.edu

Ms. Karin Isaev
Columbia University
kisaev@nygenome.org

Ms. Komal Jain
Frederick National Laboratory for Cancer Research
komal.jain@nih.gov

Dr. Kan Jang
NIAMS/NIH
jiangk@mail.nih.gov

Ms. Katharine Jenike
Johns Hopkins University
kate.jenike@gmail.com

Ms. Aditee Kadam
Weizmann Institute of Science
aditee.kadam@weizmann.ac.il

Dr. Andre Kahles
ETH Zurich
andre.kahles@inf.ethz.ch

Dr. Masahiro Kasahara
The University of Tokyo
mkasa@edu.k.u-tokyo.ac.jp

Ms. Georgi Katsoula
Helmholtz Zentrum Munchen (GmbH)
georgia.katsoula@helmholtz-munich.de

Dr. Hideya Kawaji
Tokyo Metropolitan Institute of Medical Science
kawaji-hd@igakuken.or.jp

Dr. Joanna Kelley
University of California-Santa Cruz
jokelley@ucsc.edu

Dr. Janet Kelso
Max Planck Institute for Evolutionary Anthropology
kelso@eva.mpg.de

Dr. Ayse Keskus
NIH-NCI
keskusayse@gmail.com

Dr. Sam Khalouei
Personalis
sam.khalouei@personalis.com

Daniel Kim
Horace Greeley High School
dnk1492@gmail.com

Dr. Juhyun Kim
NIH
juhyun.kim.0203@gmail.com

Mr. Yongjun Kim
Yonsei University College of Medicine
dragonf97@naver.com

Dr. Mikhail Kolmogorov
National Institutes of Health
mikhail.kolmogorov@nih.gov

Dr. Rohit Kolora
Alector Tx
rohit.kolora@alector.com

Wenjun Kong
Calico Life Sciences
kwj@calicolabs.com

Dr. Sergey Koren
NIH
sergekoren@gmail.com

Dr. Sam Kovaka
Johns Hopkins University
skovaka@gmail.com

Dr. Teresa Krieger
Charite
teresa.krieger@charite.de

Dr. Annika Kroeger
The Univeristy of Birmingham
a.t.kroeger@bham.ac.uk

Rujuta Kshirsagar
Kymera Therapeutics
rujuta227@gmail.com

Nataliya Kucher
Johns Hopkins University
nkucher3@jhu.edu

Prof. Pankaj Kumar
Weill Cornell Medical College in Qatar
panks.svpuat@gmail.com

Mr. Ashwin Kumar
MIT
ashwinsk@mit.edu

Sunita Kumari
Cold Spring Harbor Lab
kumari@cshl.edu

Prof. Benjamin Langmead
Johns Hopkins University
langmea@cs.jhu.edu

Sarah Laperriere
Metagenomi
sarah.laperriere@metagenomi.co

Dr. Delphine Lariviere
The Pennsylvania State University
lariviere.delphine@gmail.com

Dr. Olga Lazareva
German Cancer Research Center (DKFZ)
olga.lazareva@dkfz-heidelberg.de

Megan Le
Dana Farber Cancer Institute
meganle@mit.edu

Ms. Denise Le
University of Toronto
denise.le@mail.utoronto.ca

Dr. Soo Ching Lee
National Institutes of Health
sooching.lee@nih.gov

Dr. Ben Lehner
Centre for Genomic Regulation
ben.lehner@crg.es

Steven Lewis
Cold Spring Harbor Lab
slewis@cshl.edu

Dr. Nancy Li
Ontario Institute for Cancer Research
nancy.li@oicr.on.ca

Dr. Feng Li
eGenesis Inc.
feng.li@egenesisbio.com

Dr. Qiuhui Li
Johns Hopkins University
liqiuhui09@gmail.com

Dr Hua Li
Stowers Institute for Medical Research
hul@stowers.org

Ms. Xiaoxu Li
EPFL
xiaoxu.li@epfl.ch

Shumin Liang
Hong Kong Baptist University
20483120@life.hkbu.edu.hk

Ms. Yixin Lin
Aarhus University
yixinlin@clin.au.dk

Mr. Mao-Jan Lin
Johns Hopkins University
mlin77@jhu.edu

Mr. Connor Littlefield
University of Utah
Connor.littlefield@utah.edu

Ms. Xiao Liu
Harvard Medical School
xkliu0424@gmail.com

Dr. Jia Liu
Leidos Biomedical Research Inc
liuj18@mail.nih.gov

Dr. Shaoke Lou
Yale University
loushaoke@gmail.com

Chenyue Lu
Dana-Farber Cancer Institute
chenyue_lu@dfci.harvard.edu

Dr. Jennifer Lu
Johns Hopkins University
jlu26@jhu.edu

Ms. Wen Luo
Leidos Biomedical Research, Inc.
wen.luo@nih.gov

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Alexandru Mahmoud
Harvard Medical School/Mass General Brigham
mahmoudalexandru@gmail.com

Prof. Veli Makinen
University of Helsinki
veli.makinen@helsinki.fi

Dr. Salem Malikic
National Institutes of Health
salem.malikic@nih.gov

Sharvari Mankame
Translational Genomics Research Institute
smankame@tgen.org

Keenan Manpearl
University of Colorado Anschutz
keenan.manpearl@cuanschutz.edu

Dr. Shane McCarthy
Wellcome Sanger Institute
sm15@sanger.ac.uk

Prof. William McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Prof. Pall Melsted
University of Iceland
pmelsted@hi.is

Dr. Karen Miga
UCSC Genomics Institute
khmiga@ucsc.edu

Dr. Aleksei Mikhalchenko
Oregon Health & Science University
mikhalch@ohsu.edu

Dr. Pamela Milani
Bristol Myers Squibb
pamela.milani@bms.com

Dr. Ilia Minkin
Johns Hopkins University
ivminkin@gmail.com

Ms. Soheila Moeini
University of Montreal
soheila.moeini@umontreal.ca

Dr. David Molik
USDA Agricultural Research Service
david.molik@usda.gov

Ms. Ida Moltke
University of Copenhagen
ida@binf.ku.dk

Mr. Luke Morina
Johns Hopkins University
lmorina2@jhmi.edu

Dr. Ahmed Moustafa
Children's Hospital of Philadelphia
moustafaam@chop.edu

Shandukani Mulaudzi
Harvard Medical School
smulaudzi@g.harvard.edu

Dr. Nicola Mulder
University of Cape Town
nicola.mulder@uct.ac.za

Mr. Rajeeva Lochan Musunuri
New York Genome Center
rmusunuri@nygenome.org

Mr. Michael Nagy
Praxis Genomics
fidel.nagy@gmail.com

Dr. RK Narayanan
Cold Spring Harbor Laboratory
narayan@cshl.edu

Ms. Urwah Nawaz
University of Adelaide
urwah.nawaz@adelaide.edu.au

Lusine Nazaretyan
Berlin Institute of Health at Charité
lusine.nazaretyan@bih-charite.de

Dr. Anton Nekrutenko
Penn State
anton@nekrut.org

Matthew Nguyen
Johns Hopkins University
matthewnguyen.667@gmail.com

Mr. Eric Nguyen
Yale University
eric.nguyen@yale.edu

Dr. Mor Nitzan
Hebrew University of Jerusalem
mor.nitzan@mail.huji.ac.il

Deborah Nusskern
Luminex Corporation
dnusskern@luminexcorp.com

Christopher Ojukwu
University of Colorado, Boulder
christopher.ojukwu@colorado.edu

Karol Pal
Pennsylvania State University
kxp5629@psu.edu

Mr. Samarendra Pani
Heinrich Heine University
samarendra.pani@hhu.de

Mr. Dongwoo Park
Yonsei University College of Medicine
woooh324@yuhs.ac.kr

Chloe Parker
Northwestern University
Chloe.Parker@northwestern.edu

Dr. Pei-Hua Peng
Chang Gung Memorial Hospital at Linkou
jinger0908@gmail.com

Ms. Jialin Peng
Hong Kong Baptist University
20483236@life.hkbu.edu.hk

Mr. Jonathan Perdomo
Drexel University
perdomojonathan5@gmail.com

Dr. Cesar Perez Ferandez
Universidad Privada Boliviana
cesarperez@upb.edu

Prof. Mihaela Pertea
Johns Hopkins University
mpertea@jhu.edu

Dr. Marcus Pezzolesi
University of Utah
marcus.pezzolesi@hsc.utah.edu

Mr. Patrick Pflughaupt
University of Oxford
patrick.pflughaupt@imm.ox.ac.uk

Mr. Lucas Pietan
University of Iowa
lpietan@uiowa.edu

Dr. Rishvanth Prabakar
Cold Spring Harbor Laboratory
kaliapp@cshl.edu

Mr. Malyaj Prajapati
Sardar Vallabhbhai Patel University of Agriculture
malyajrprajapati@gmail.com

Dr. Weilin Qiu
University of Copenhagen
weilin.qiu@bio.ku.dk

Ms. Jyotshna Rajput
Indian Institute of Science
jyotshnar@iisc.ac.in

Mr. Ajay Ramakrishnan Varadarajan
Illumina Inc
avaradarajan@illumina.com

Ms. Amanda Ramdass
The University of the West Indies
amanda.ramdass@my.uwi.edu

Mr. Mohammadsaleh Refahi
Drexel University
sr3622@drexel.edu

Dr. Giulietta Riboldi
New York University
giulietta.riboldi@nyulangone.org

Dr. Alberto Rivera
National Institutes of Health - NIH
albertomarcosrivera@gmail.com

Dr. Gail Rosen
Drexel University
glr26@drexel.edu

Dr. Yoshitaka Sakamoto
National Cancer Center Research Institute
yosakam2@ncc.go.jp

Dr. Max Salick
insitro
max@insitro.com

Prof. Steven Salzberg
Johns Hopkins University
steven.salzberg@gmail.com

Ms. Kewalin Samart
University of Colorado Anschutz Medical Campus
kewalin.samart@cuanschutz.edu

Sydney Sanchez
Boston College
sanchesl@bc.edu

Dr. Fatemeh Sanjar
Pivot Bio
fsanjar@pivotbio.com

Michelle Savage
Johns Hopkins University
Michelle.Savage@jhmi.edu

Dr. Michael Schatz
Johns Hopkins University
mschatz@cs.jhu.edu

Dr. Courtney Schiebout
Dartmouth College
courtney.t.schiebout@dartmouth.edu

Dr. Zina Schlachetzki
Illumina Inc
zschlachetzki@illumina.com

Kristen Schneider
University of Colorado, Boulder
krsc0813@colorado.edu

Ms. Lucy Scott
The University of Edinburgh
l.scott-19@sms.ed.ac.uk

Dr. Eran Segal
Weizmann Institute of Science
eran@weizmann.ac.il

Evan Seitz
Cold Spring Harbor Laboratory
seitz@cshl.edu

Prof. Colin Semple
MRC Human Genetics Unit
colin.semple@ed.ac.uk

Dr. Shurjo Sen
NHGRI
sensh@mail.nih.gov

Mr. Mohammadali Serajian
Univertsity of Florida
m.serajian@ufl.edu

Jim Shaw
University of Toronto
jshaw@math.toronto.edu

James Shaw
Kymera Therapeutics
jshaw@kymeratx.com

Dr. Heejung Shim
University of Melbourne
heejung.shim@unimelb.edu.au

Ida Shinder
Johns Hopkins School of Medicine
ishinde1@jhmi.edu

Dr. Yuichi Shiraishi
National Cancer Center Research Institute
yuishira@ncc.go.jp

Mr. Vikram Shivakumar
Johns Hopkins University
vshivak1@jhu.edu

Dr. Parisa Shooshtari
Western University
pshoosh@uwo.ca

Alaina Shumate
National Bioforensic Analysis Cetner
alaina.shumate@st.dhs.gov

Mile Sikic
Genome Institute of Singapore
mile_sikic@gis.a-star.edu.sg

Mr. Ryan Silk
University of Edinburgh
r.p.silk@sms.ed.ac.uk

Dr. Jared Simpson
Ontario Institute for Cancer Research
Jared.Simpson@oicr.on.ca

Dr. Ying Sims
Welcome sanger Institute
yy5@sanger.ac.uk

Dr. Amartya Singh
Rutgers Cancer Institute of New Jersey
as2197@scarletmail.rutgers.edu

Mr. Steven Solar
NHGRI
steven.j.solar@gmail.com

Dr. Li Song
Dartmouth College
Li.Song@dartmouth.edu

Dr. Bao-Hua Song
University of North Carolina at Charlotte
bsong5@uncc.edu

Mr. Joon-Hyun Song
Stony Brook University
joon-hyun.song@stonybrook.edu

Mr. Stephen Staklinski
Cold Spring Harbor Laboratory
staklins@cshl.edu

Dr. Daniel Standage
National Bioforensic Analysis Center
daniel.standage@st.dhs.gov

Ms. Tiana Stastny
Revvity
Tiana.stastny@horizondiscovery.com

Dr. Jacob Steenwyk
HHMI & UC-Berkeley
jlsteenwyk@berkeley.edu

Dr. Oliver Stegle
German Cancer Research Center
o.stegle@dkfz-heidelberg.de

Bo Sun
UCSF
bo.sun2@ucsf.edu

Dr. Shixiang Sun
Albert Einstein College of Medicine
shixiang.sun@einsteinmed.org

Mengyi Sun
Cold Spring Harbor Laboratory
msun@cshl.edu

Ms. Pallavi Surana
SUNY-Stony Brook University
pallavi.surana@stonybrook.edu

Dr. Hillary Sussman
Genome Research, Executive Editor
hsussman@cshl.edu

Mr. Alexander Sweeten
Johns Hopkins University
alex.sweeten@gmail.com

Tamas Ryszard Sztanka-Toth
Janssen-Cilag GmbH, Johnson and Johnson
tsztanka@its.jnj.com

Dr. Shaoyuan Tan
St. Jude Children's Research Hospital
shaoyuan.tan@stjude.org

Ms. Luomeng Tan
University of Toronto
luomeng.tan@mail.utoronto.ca

Dr. Amos Tanay
Weizmann Institute of Science
amos.tanay@weizmann.ac.il

Ms. Alison Tang
Freenome
alison.tang@freenome.com

Dr. Yaoliang Tang
Augusta University
yaotang@augusta.edu

Prof. Martin Taylor
University of Edinburgh
martin.taylor@ed.ac.uk

Dr. Jamie Teer
H. Lee Moffitt Cancer Center & Research Institute
Jamie.Teer@moffitt.org

Christopher Thai
Rutgers University
ct647@iqb.rutgers.edu

Ms. Erin Theiller
Children's Hospital of Philadelphia
theillere@chop.edu

Dr. Winston Timp
Johns Hopkins University
wtimp@jhu.edu

Dr. Khanh Trang
Children's Hospital of Philadelphia
trangk@chop.edu

Dr. Tim Triche
VAI
tim.triche@vai.org

Dr. Frances Turner
University of Edinburgh
fturner@ed.ac.uk

Dr. Monika Tutaj
The Medical College of Wisconsin, Inc
motutaj@mcw.edu

Courtney Vaccaro
Illumina
cvaccaro@illumina.com

Naga Sai Kavya Vaddadi
Johns Hopkins University
kvaddad1@jh.edu

Mr. Rahul Varki
University of Florida
rvarki@ufl.edu

Meghana Vemulapalli
NIH
meghana.vemulapalli@nih.gov

Mr. Jan Verburg
Institute of Genetics and Cancer
s2010688@ed.ac.uk

Dr. Roger Volden
PacBio
rvolden@pacificbiosciences.com

Justin Wagner
National Institute of Standards and Technology
justin.wagner@nist.gov

Yinan Wan
Pacific Biosciences
ywan@pacificbiosciences.com

Ms. Jiahui Wang
Cancer Genomics Research Laboratory
jiahui.wang2@nih.gov

Minqian Wang
PacBio
miwang@pacb.com

Dr. Difei Wang
Leidos Biomedical Research, Inc.
wangdi@mail.nih.gov

Dr. Doreen Ware
Cold Spring Harbor Lab/USDA-ARS
ware@cshl.edu

Dr. Sharon Wei
Cold Spring Harbor Laboratory
weix@cshl.edu

Dr. Sherman Weissman
Yale Univeristy
sherman.weissman@yale.edu

John Wilson
Personalis
johnd.wilson9@gmail.com

Dr. Jan Witkowski
Cold Spring Harbor Laboratory
witkowsk@cshl.edu

Dr. Adam Wright
Ontario Institute for Cancer Research
adam.wright@oicr.on.ca

Dr. Siyang Xia
Intellia Therapeutics
siyang.xia@intelliatx.com

Prof. Yan Xu
Cincinnati Children's Hospital Medical Center
yan.xu@cchmc.org

Dr. Zhuwei Xu
NIH
xuz5@nih.gov

Guanao Yan
Univeristy of California, Los Angeles
gayan@g.ucla.edu

Dr. In Seok Yang
Yonsei Univesrsity College of Medicine
channeli@yuhs.ac

Dagyeong YANG
NIH/NIDDK
yangd12@nih.gov

Dr. Weiwei Yang
New England Biolabs
wyang@neb.com

Dr. Yuelin Yao
The University of Edinburgh
s1914230@ed.ac.uk

Mr. Mohsen Zakeri
Johns Hopkins University
mohs.zakeri@gmail.com

Dr. Eleftheria Zeggini
Helmholtz, Munich
eleftheria.zeggini@helmholtz-munich.de

Mr. Gao Zhanyu
Beijing Institute of Genomics
gaozhy@big.an.cn

Dr. Yixin Zhao
Cold Spring Harbor Laboratory
yizhao@cshl.edu

Haizi Zheng
Regeneron Pharmaceuticals
haizi.zheng@regeneron.com

Dr. Xinchang Zheng
Baylor College of Medicine
Xinchang.Zheng@bcm.edu

Dr. Man Zhou
University of Maryland College Park
mzhou888@umd.edu

Dr. Alexey Zimin
Johns Hopkins University
alekseyz@jhu.edu

CODE OF CONDUCT FOR ALL PARTICIPANTS IN CSHL MEETINGS

Cold Spring Harbor Laboratory (CSHL or the Laboratory) is dedicated to pursuing its twin missions of research and education in the biological sciences. The Laboratory is committed to fostering a working environment that encourages and supports unfettered scientific inquiry and the free and open exchange of ideas that are the hallmarks of academic freedom. To this end, the Laboratory aims to maintain a safe and respectful environment that is free from harassment and discrimination for all attendees of our meetings and courses as well as associated support staff, in accordance with federal, state and local laws.

Consistent with the Laboratory's missions, commitments and policies, the purpose of this Code is to set forth expectations for the professional conduct of all individuals participating in the Laboratory's meetings program, both in person and virtually, including organizers, session chairs, invited speakers, presenters, attendees and sponsors. This Code's prohibition against discrimination and harassment is consistent with the Laboratory's internal policies governing conduct by its own faculty, trainees, students and employees.

By registering for and attending a CSHL meeting, either in person or virtually, participants agree to:

1. Treat fellow meeting participants and CSHL staff with respect, civility and fairness, without bias based on sex, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, nationality or national origin, citizenship status, disability status, veteran status, marital or partnership status, age, genetic information, or any other criteria prohibited under applicable federal, state or local law.

2. Use all CSHL facilities, equipment, computers, supplies and resources responsibly and appropriately if attending in person, as you would at your home institution.

3. Abide by the CSHL Meeting Alcohol Policy (*see below)*.

Similarly, meeting participants agree to refrain from:

1. Harassment and discrimination, either in person or online, in violation of Laboratory policy based on actual or perceived sex, pregnancy status, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, creed, nationality or national origin, immigration or citizenship status, mental or physical disability status, veteran status, military status, marital or partnership status, marital or partnership status, familial status, caregiver status, age, genetic information, status as a victim of domestic violence, sexual violence, or stalking, sexual reproductive health decisions, or any other criteria prohibited under applicable federal, state or local law.

2. Sexual harassment or misconduct.

3. Disrespectful, uncivil and/or unprofessional interpersonal behavior, either in person or online, that interferes with the working and learning environment.

4. Misappropriation of Laboratory property or excessive personal use of resources, if attending in person.

BREACHES OR VIOLATIONS OF THE CODE OF CONDUCT

Cold Spring Harbor Laboratory aims to maintain in-person and virtual conference environments that accord with the principles and expectations outlined in this Code of Conduct. Meeting organizers are tasked with providing leadership during each meeting, and may be approached informally about any breach or violation. Breaches or violations should also be reported to program leadership in person or by email:

- Dr. David Stewart, Grace Auditorium Room 204, 516-367-8801 or x8801 from a campus phone, stewart@cshl.edu
- Dr. Charla Lambert, Hershey Laboratory Room 214, 516-367-5058 or x5058 from a campus phone, clambert@cshl.edu

Reports may be submitted by those who experience harassment or discrimination as well as by those who witness violations of the behavior laid out in this Code.

The Laboratory will act as needed to resolve the matter, up to and including immediate expulsion of the offending participant(s) from the meeting, dismissal from the Laboratory, and exclusion from future academic events offered by CSHL.

If you have questions or concerns, you can contact the meeting organizers, CSHL staff.

For meetings and courses funded by NIH awards:

Participants may contact the Health & Human Services Office for Civil Rights (OCR). See this page for information on filing a civil rights complaint with the OCR; filing a complaint with CSHL is not required before filing a complaint with OCR, and seeking assistance from CSHL in no way prohibits filing complaints with OCR. You may also notify NIH directly about sexual harassment, discrimination, and other forms of inappropriate conduct at NIH-supported events.

For meetings and courses funded by NSF awards:

Participants may file a complaint with the NSF. See this page for information on how to file a complaint with the NSF.

Law Enforcement Reporting:

- For on-campus incidents, reports to law enforcement can be made to the Security Department at 516-367-5555 or x5555 from a campus phone.
- For off-campus incidents, report to the local department where the incident occurred.

In an emergency, dial 911.

DEFINITIONS AND EXAMPLES

Uncivil/disrespectful behavior is not limited to but may take the following forms:

- Shouting, personal attacks or insults, throwing objects, and/or sustained disruption of talks or other meeting-related events

Harassment is any unwelcome verbal, visual, written, or physical conduct that occurs with the purpose or effect of creating an intimidating, hostile, degrading, humiliating, or offensive environment or unreasonably interferes with an individual's work performance. Harassment is not limited to but may take the following forms:

- Threatening, stalking, bullying, demeaning, coercive, or hostile acts that may have real or implied threats of physical, professional, or financial harm
- Signs, graphics, photographs, videos, gestures, jokes, pranks, epithets, slurs, or stereotypes that comment on a person's sex, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, nationality or national origin, citizenship status, disability status, veteran status, marital or partnership status, age, genetic information, or physical appearance

Sexual Harassment includes harassment on the basis of sex, sexual orientation, self-identified or perceived sex, gender expression, gender identity, and the status of being transgender. Sexual harassment is not limited to sexual contact, touching, or expressions of a sexually suggestive nature. Sexual harassment includes all forms of gender discrimination including gender role stereotyping and treating employees differently because of their gender. *Sexual misconduct* is not limited to but may take the following forms:

- Unwelcome and uninvited attention, physical contact, or inappropriate touching
- Groping or sexual assault
- Use of sexual imagery, objects, gestures, or jokes in public spaces or presentations
- Any other verbal or physical contact of a sexual nature when such conduct creates a hostile environment, prevents an individual from fulfilling their professional responsibilities at the meeting, or is made a condition of employment or compensation either implicitly or explicitly

MEETING ALCOHOL POLICY

Consumption of alcoholic beverages is not permitted in CSHL's public areas other than at designated social events (wine and cheese reception, picnic, banquet, etc.), in the Blackford Bar, or under the supervision of a licensed CSHL bartender.

No provision of alcohol by meeting sponsors is permitted unless arranged through CSHL.

Meeting participants consuming alcohol are expected to drink only in moderation at all times during the meeting.

Excessive promotion of a drinking culture at any meeting is not acceptable or tolerated by the Laboratory. No meeting participant should feel pressured or obliged to consume alcohol at any meeting-related event or activity.

VISITOR INFORMATION

EMERGENCY (to dial outside line, press 3+1+number)	
CSHL Security	**516-367-8870 (x8870 from house phone)**
CSHL Emergency	**516-367-5555 (x5555 from house phone)**
Local Police / Fire	**911**
Poison Control	**(3) 911**

CSHL SightMD Center for Health and Wellness *(call for appointment)* Dolan Hall, East Wing, Room 111 cshlwellness@northwell.edu	**516-422-4422** **x4422 from house phone**
Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2000**
Dentists Dr. William Berg Dr. Robert Zeman	**631-271-2310** **631-271-8090**
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400**

GENERAL INFORMATION

Meetings & Courses Main Office

Hours during meetings: M-F 9am – 9pm, Sat 8:30am – 1pm
After hours – See information on front desk counter
For assistance, call Security at 516-367-8870
(x8870 from house phone)

Dining, Bar

Blackford Dining Hall (main level):
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Blackford Bar (lower level): 5:00 p.m. until late

House Phones

Grace Auditorium, upper / lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

Books, Gifts, Snacks, Clothing

CSHL Bookstore and Gift Shop
516-367-8837 (hours posted on door)
Grace Auditorium, lower level.

Computers, E-mail, Internet access

Grace Auditorium
Upper level: E-mail and printing in the business center area
WiFi Access: GUEST (no password)

Announcements, Message Board Mail, ATM, Travel info

Grace Auditorium, lower level

Russell Fitness Center

Dolan Hall, east wing, lower level

PIN#: **(On your registration envelope)**

Laundry Machines

Dolan Hall, lower level

Photocopiers, Journals, Periodicals, Books

CSHL Main Library
Open 24 hours (with PIN# or CSHL ID)
Staff Hours: 9:00 am – 9:00 pm
Use PIN# (On your registration envelope) to enter Library
See Library staff for photocopier code.
Library room reservations (hourly) available on request between 9:00 am – 9:00 pm

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m.
Two tennis courts open daily.

Local Interest

Fish Hatchery	631-692-6758
Sagamore Hill	516-922-4788
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -
Take CSHL Shuttle OR Uber/Lyft/Taxi to Syosset Train Station
Long Island Railroad to Penn Station
Train ride about one hour.

TRANSPORTATION

Limo, Taxi

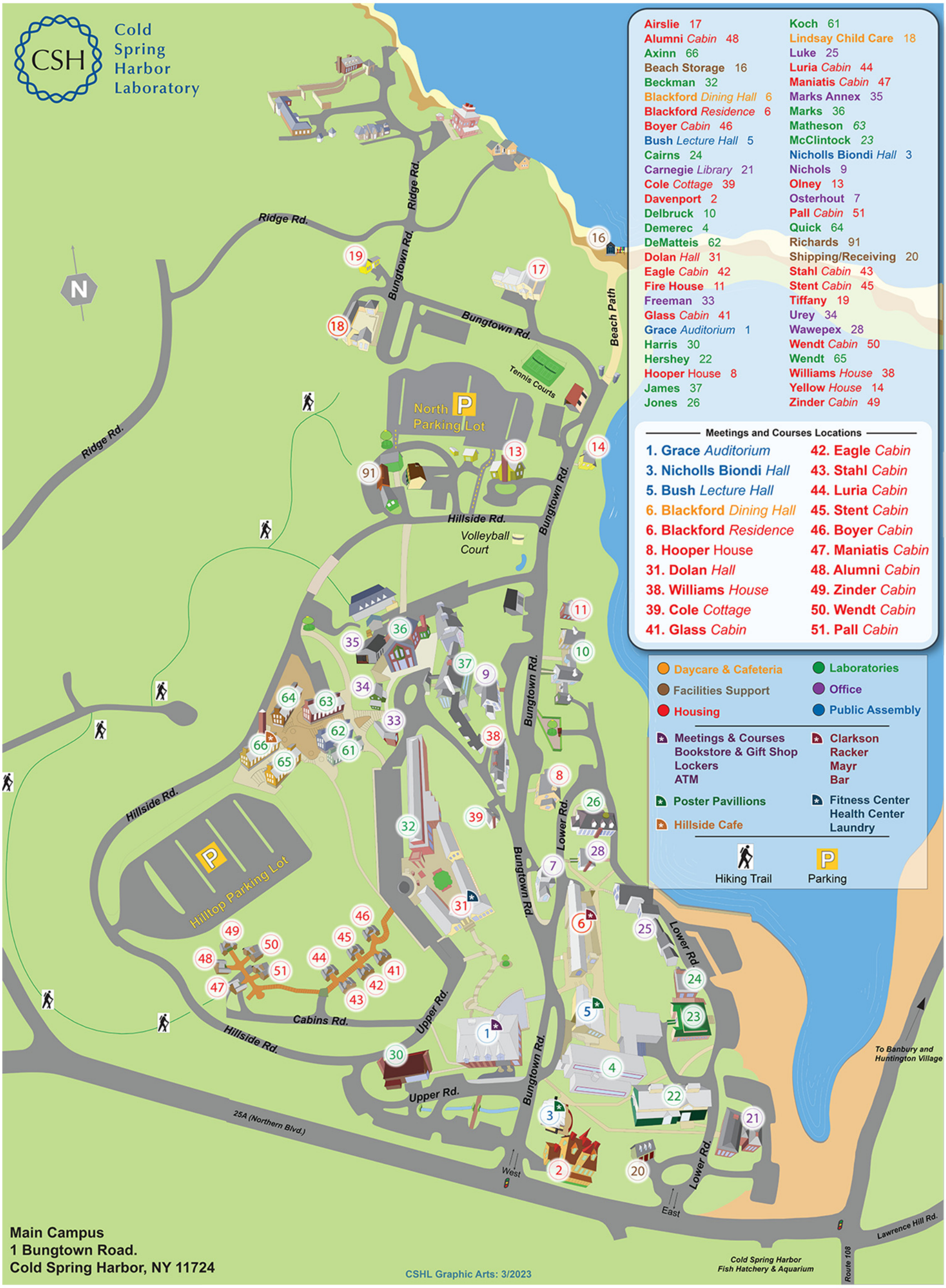
Syosset Limousine	516-364-9681
Executive Limo Service	516-826-8172
Limos Long Island	516-400-3364
Syosset Taxi	516-921-2141
Orange & White Taxi	631-271-3600
Uber / Lyft	

Trains

Long Island Rail Road	718-217-LIRR (5477)
Amtrak	800-872-7245
MetroNorth	877-690-5114
New Jersey Transit	973-275-5555

CSHL Campus Map

CSHL Map


Cold Spring Harbor Laboratory
N
Ridge Rd.
Ridge Rd.
Ridge Rd.
Bungtown Rd.
Bungtown Rd.
Bungtown Rd.
Bungtown Rd.
Bungtown Rd.
Beach Path
Tennis Courts
North Parking Lot
Hillside Rd.
Volleyball Court
Hillside Rd.
Hillside Rd.
Hilltop Parking Lot
Cabins Rd.
Upper Rd.
Upper Rd.
Lower Rd.
Lower Rd.
Lower Rd.
25A (Northern Blvd.)
West
East
To Banbury and Huntington Village
Lawrence Hill Rd.
Route 108
Cold Spring Harbor Fish Hatchery & Aquarium

Airslie	17	Koch	61
Alumni *Cabin*	48	Lindsay Child Care	18
Axinn	66	Luke	25
Beach Storage	16	Luria *Cabin*	44
Beckman	32	Maniatis *Cabin*	47
Blackford *Dining Hall*	6	Marks Annex	35
Blackford *Residence*	6	Marks	36
Boyer *Cabin*	46	Matheson	63
Bush *Lecture Hall*	5	McClintock	23
Cairns	24	Nicholls Biondi *Hall*	3
Carnegie *Library*	21	Nichols	9
Cole *Cottage*	39	Olney	13
Davenport	2	Osterhout	7
Delbruck	10	Pall *Cabin*	51
Demerec	4	Quick	64
DeMatteis	62	Richards	91
Dolan *Hall*	31	Shipping/Receiving	20
Eagle *Cabin*	42	Stahl *Cabin*	43
Fire House	11	Stent *Cabin*	45
Freeman	33	Tiffany	19
Glass *Cabin*	41	Urey	34
Grace *Auditorium*	1	Wawepex	28
Harris	30	Wendt *Cabin*	50
Hershey	22	Wendt	65
Hooper House	8	Williams *House*	38
James	37	Yellow *House*	14
Jones	26	Zinder *Cabin*	49

Meetings and Courses Locations

1. Grace *Auditorium*
3. Nicholls Biondi *Hall*
5. Bush *Lecture Hall*
6. Blackford *Dining Hall*
6. Blackford *Residence*
8. Hooper House
31. Dolan *Hall*
38. Williams *House*
39. Cole *Cottage*
41. Glass *Cabin*
42. Eagle *Cabin*
43. Stahl *Cabin*
44. Luria *Cabin*
45. Stent *Cabin*
46. Boyer *Cabin*
47. Maniatis *Cabin*
48. Alumni *Cabin*
49. Zinder *Cabin*
50. Wendt *Cabin*
51. Pall *Cabin*

● Daycare & Cafeteria
● Facilities Support
● Housing
● Laboratories
● Office
● Public Assembly

Meetings & Courses
Bookstore & Gift Shop
Lockers
ATM

Poster Pavillions

Hillside Cafe

Clarkson
Racker
Mayr
Bar

Fitness Center
Health Center
Laundry

Hiking Trail

Parking

Main Campus
1 Bungtown Road.
Cold Spring Harbor, NY 11724

CSHL Graphic Arts: 3/2023