2017 COMPUTATIONAL GENOMICS SUMMER INSTITUTE RETREAT

BIG BEAR LAKE, CALIFORNIA JULY 6-8

Program Contents

WELCOME

The CGSI retreat will be a unique opportunity for faculty and Long Course participants to become acquainted and learn about each other's research in an informal setting. The 2017 CGSI retreat schedule includes research talks, flash talks, journal clubs, and social events.

RESEARCH TALKS are 30-minute explorations of current problems in computational genomics by participating post-doctoral scholars and faculty.

FLASH TALKS are brief presentations by participating graduate students.

JOURNAL CLUBS are 50-minute sessions that give participants an opportunity to critically evaluate recent articles and stay up-to-date on relevant new research.

In addition, we will hold a special extended JOURNAL CLUB BRUNCH on Saturday, July 8, in The Olympian Room of the Best Western Big Bear Chateau.

ACCOMMODATIONS

The 2017 CGSI retreat will be held at the Best Western hotel in Big Bear Lake, California. Hotel rooms have already been assigned to guests; please check in with your name. Complimentary breakfast served from 6:00 a.m. to 9:30 a.m.

Best Western Big Bear Chateau
42200 Moonridge Rd, Big Bear Lake, CA 92315
(909) 866-6666

ACTIVITIES

Arrangements for optional social events include dinners at local restaurants, hiking in the San Bernardino Mountains, soccer competitions at a nearby park, and free time at the hotel. The Best Western Big Bear Chateau features an outdoor heated swimming pool and hot tub, as well as a game room with a pool table.

2017 CGSI PROGRAMS

Long Course Retreat: July 6 – 8
Short Course: July 10 – 14
Long Course: July 10 – 26

CGSI ORGANIZING COMMITTEE

Eleazar Eskin, UCLA, CGSI Director
Russel Caflisch, UCLA, IPAM Director
Francesca Chiaromonte, Pennsylvania State University
Eran Halperin, UCLA
David Koslicki, Oregon State University
John Novembre, University of Chicago
Ben Raphael, Princeton University

Visit our website for more about CGSI:
computationalgenomics.bioinformatics.ucla.edu

SCHEDULE

Wednesday July 5

16:00 Check into Best Western Big Bear Chateau
42200 Moonridge Rd, Big Bear Lake, CA 92315 goo.gl/maps/ZUN8Hmz8hpt

18:00 Dinner @ Nottinghams Tavern
40797 Big Bear Blvd, Big Bear Lake, CA 92315 goo.gl/maps/S2yizdBgB7B2

Thursday July 6

06:00 - 09:00 Breakfast in the Best Western Breakfast Room
07:00 - 09:00 Registration
Pick up your name tag in the hotel breakfast room!

09:00 - 15:10 Retreat Day One
The Olympian Room, Best Western Big Bear Chateau

- 09:00 Introduction
- 09:15 Flash Talks
  Mohammed Alser, Eliran Avni, Brian Hill, Lisa Gai, and Colin Farrell
- 09:25 Vasilis Ntranos
  Clustering a million cells: Large-scale scRNA-Seq data analysis
- 09:55 Coffee Break
- 10:30 Flash Talks
  Christopher Robles, Nolan Donoghue, Kaiyuan Zhu, Judy Du, and Joel Mefford
- 10:40 Serghei Mangul
  Squeezing the last drop out of next generation sequencing data
- 11:10 Journal Club
- 12:00 Lunch
- 13:00 Flash Talks
  Joseph Marcus, Arun Durvasula, Jeffrey West, and Catharine Krebs
- 13:10 Andy Dahl
  Adjusting for principal components of molecular phenotypes induces replicating false positives
- 13:40 Coffee Break
- 14:00 Flash Talks
  Jennifer Zou, Nadav Brandes, Stephen Rong, Ana Kenney, and Shahar Shohat
- 14:10 Ilan Gronau
  Population phylogenomics: A genealogical perspective
- 14:40 Na Cai
  Heterogeneity in depression

16:00 Hiking in the San Bernardino Mountains @ Big Bear Discovery Center
40971 North Shore Drive/ Hwy 38, Fawnskin, CA 92333 goo.gl/maps/ZKkd7SLv48q

19:00 Dinner @ Azteca Grill
40199 Big Bear Blvd, Big Bear Lake, CA 92315 goo.gl/maps/C6MNnZQvvYx

Friday July 7

06:00 - 09:00 Breakfast in the Best Western Breakfast Room

09:00 - 15:30 Retreat Day Two
The Olympian Room, Best Western Big Bear Chateau

- 09:00 Flash Talks
  Vivian Link, Lina Zheng, Debmalya Nandy, and Kelsey Johnson
- 09:10 Marzia Cremona
  Functional data analysis testing and linear modeling for high-resolution "omics" data
- 09:40 David Koslicki
  Improving Min Hash for Metagenomic Classification
- 10:10 Coffee Break
- 10:30 Flash Talks
  Brian Jo, Ariel Gewirtz, Julia Matsieva, and Chloe Robins
- 10:40 Pejman Mohammadi
  Using ASE data to facilitate diagnosis for unresolved rare diseases
- 11:10 Journal Club
- 12:00 Lunch
- 13:00 Flash Talks
  Igor Mandric, James Boocock, Nemanja Marjanovic, and Fatma Kahveci
- 13:10 Anil Ori
  Integration of longitudinal gene expression with polygenic disease risk establishes human neuronal differentiation as a model to study schizophrenia
- 13:40 Coffee Break
- 14:00 YoSon Park
  Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci
- 14:30 Nikita Alexeev
  Estimation of the rate of transpositions and the true evolutionary distance
- 15:00 Loes Olde Loohuis
  Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia

15:30 Sports @ Meadow Park
41220 Park Ave, Big Bear Lake, CA 92315 goo.gl/maps/XLYxLyhcE192

18:00 Dinner @ Paoli's Italian Country Kitchen
40821 Village Dr, Big Bear Lake, CA 92315 goo.gl/maps/mY6C4YtzzPH2

20:00 Free Time in the Best Western Game Room and Outdoor Pool

Saturday July 8

09:00 - 11:00 Journal Club Brunch
The Olympian Room, Best Western Big Bear Chateau

12:00 Check out of hotel.

TALK TITLES AND PAPERS

This year's retreat features twelve 30-minute research talks presented by emerging scholars in the field of computational genomics. Each presenter compiled a list of relevant papers that provide a more in-depth exploration of their research talk material. Click to open a paper in your internet browser.

01 Vasilis Ntranos: Clustering a million cells: Large-scale scRNA-Seq data analysis

1. Ntranos, V., Kamath, G.M., Zhang, J.M., Pachter, L. and David, N.T., 2016. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome biology, 17(1), p.112.

02 Serghei Mangul: Squeezing the last drop out of next generation sequencing data

No papers assigned.

03 Andy Dahl: Adjusting for principal components of molecular phenotypes induces replicating false positives

1. Dahl, A., Guillemot, V., Mefford, J., Aschard, H. and Zaitlen, N., 2017. Adjusting For Principal Components Of Molecular Phenotypes Induces Replicating False Positives. bioRxiv, p.120899.

2. Leek, J.T. and Storey, J.D., 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet, 3(9), p.e161.

3. Leek, J.T. and Storey, J.D., 2008. A general framework for multiple testing dependence. Proceedings of the National Academy of Sciences, 105(48), pp.18718-18723.

04 Ilan Gronau: Population phylogenomics: A genealogical perspective

1. Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A., 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature genetics, 43(10), pp.1031-1034.

2. Rasmussen, M.D., Hubisz, M.J., Gronau, I. and Siepel, A., 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet, 10(5), p.e1004342.

05 Na Cai: Heterogeneity in depression

1. Cai, N., Bigdeli, T.B., Kretzschmar, W.W., Li, Y., Liang, J., et al., 2017. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Scientific data, 4.

2. Peterson, R.E., Cai, N., Bigdeli, T.B., Li, Y., Reimers, M., et al., 2017. The genetic architecture of major depressive disorder in Han Chinese women. JAMA psychiatry, 74(2), pp.162-168.

3. Converge Consortium, 2015. Sparse whole genome sequencing identifies two loci for major depressive disorder. Nature, 523(7562), p.588.

4. Cai, N., Chang, S., Li, Y., Li, Q., Hu, J., Liang, J., Song, L., Kretzschmar, W., Gan, X., Nicod, J. and Rivera, M., 2015. Molecular signatures of major depression. Current Biology, 25(9), pp.1146-1156.

06 Marzia Cremona: Functional data analysis testing and linear modeling for high-resolution "omics" data

1. Campos-Sánchez, R., Cremona, M.A., et al, 2016. Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. PLoS Comput Biol, 12(6), p.e1004956.

2. Cremona, M.A., Campos-Sánchez, R., Pini, A., Vantini, S., Makova, K.D. and Chiaromonte, F., 2017. Functional data analysis of "Omics" data: how does the genomic landscape influence integration and fixation of endogenous retroviruses?. In Functional Statistics and Related Fields (pp. 87-93). Springer, Cham.

3. Cremona, Pini, Chiaromonte, Vantini (2017). IWTomics: Interval-Wise Testing for Omics Data. R package version 1.0.0.

07 David Koslicki: Improving Min Hash for Metagenomic Classification

1. Broder, A.Z., 1997, June. On the resemblance and containment of documents. In Compression and Complexity of Sequences 1997. Proceedings (pp. 21-29). IEEE.

2. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., et al., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology, 17(1), p.132.

08 Pejman Mohammadi: Using ASE data to facilitate diagnosis for unresolved rare diseases

1. Cummings, B.B., Marshall, J.L., Tukiainen, T., et. al., 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science translational medicine, 9(386), p.eaal5209.

2. Mohammadi, P., Castel, S.E., Brown, A.A. and Lappalainen, T., 2016. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. bioRxiv, p.078717.

09 Anil Ori: Integration of longitudinal gene expression with polygenic disease risk establishes human neuronal differentiation as a model to study schizophrenia

1. Tai, Y.C. and Speed, T.P., 2006. A multivariate empirical Bayes statistic for replicated microarray time course data. The Annals of Statistics, 34(5), pp.2387-2412.

2. Aryee, M.J., Gutiérrez-Pabello, J.A., Kramnik, I., Maiti, T. and Quackenbush, J., 2009. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). BMC bioinformatics, 10(1), p.409.

3. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K. and Ripke, S., 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics, 47(11), pp.1228-1235.

4. de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D., 2015. MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology, 11(4), p.e1004219.

10 YoSon Park: Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci

1. Pashos, E.E., Park, et al., 2017. Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. Cell Stem Cell, 20(4), pp.558-570.

11 Nikita Alexeev: Estimation of the rate of transpositions and the true evolutionary distance

1. Alexeev, N. and Alekseyev, M.A., 2017. Estimation of the true evolutionary distance under the fragile breakage model. BMC Genomics, 18(4), p.356.

2. Alexeev, N., Aidagulov, R. and Alekseyev, M.A., 2015. A computational method for the rate estimation of evolutionary transpositions. arXiv preprint arXiv:1501.07546.

3. Yancopoulos, S., Attie, O. and Friedberg, R., 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics, 21(16), pp.3340-3346.

4. Biller, P., Guéguen, L., Knibbe, C. and Tannier, E., 2016. Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation. Genome biology and evolution, 8(5), pp.1427-1439.

5. Lin, Y. and Moret, B.M., 2008. Estimating true evolutionary distances under the DCJ model. Bioinformatics, 24(13), pp.i114-i122.

6. Erdos, P. and Rényi, A., 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1), pp.17-60.

12 Loes Olde Loohuis: Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia

1. Mangul, S., Loohuis, L.M.O., Ori, A., Jospin, G., Koslicki, D., Yang, H.T., et al., 2016. Total RNA Sequencing reveals microbial communities in human blood and disease specific effects. bioRxiv, p.057570.

JOURNAL CLUBS

As part of our first annual retreat, we will be breaking into five small groups for daily journal clubs on July 6 and 7. Before departing on Saturday, July 8, we will have a special extended journal club with catered brunch. Journal clubs provide an excellent opportunity to discuss current work and doing so with colleagues with shared interests can be insightful, productive, and fun. Note required reading for each journal club.

01 Microbiome analysis: Computational techniques and challenges

Hosted by Serghei Mangul
serghei@cs.ucla.edu

Technological advances and the decreasing costs of 'next-generation' sequencing (NGS) make it the technology of choice for many applications, including studying the human microbiome composed of bacterial, viral, fungi and other eukaryotic communities. Recently, high-throughput sequencing has revolutionized microbiome research by enabling the study of thousands of microbial genomes directly in their host environments. This approach, which forms the field of metagenomics, avoids the biases incurred with traditional culture-dependent analysis. The metagenomics approach also allows the comparison of microbial communities' composition in their natural habitats across different human tissues and environmental settings. Specifically, metagenomic profiling is proven useful for analyzing microbes such as eukaryotic and viral pathogens, which were previously impossible to study in an unbiased way with target 16S ribosomal RNA gene.

Tentative list of topics: We will be discussing recent methods to study microbial communities. We will be discussing the challenges in metagenomics analysis and limitation of the current methods. The goal will be to identify the best strategy to analyze metagenomics data.

We will start with discussion the most popular 'marker genes' methods, which are suggested to have poor sensitivity and also may result in false positives, detecting dangerous pathogens which are not present in the metagenomics sample. We also will discuss methods aimed to study microbiome at strain level and methods to study non-bacterial organisms, including viruses and fungi.

Outcomes: One possible outcome can be a joint effort to write an educational paper introducing metagenomics for researchers with no background in computational genomics or bioinformatics.

Difficulty: Intro/Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):

1. Escobar-Zepeda, A., de León, A.V.P. and Sanchez-Flores, A., 2015. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in Genetics, 6.

2. Simon, C. and Daniel, R., 2011. Metagenomic analyses: past and future trends. Applied and Environmental Microbiology, 77(4), pp.1153-1161.

3. Schmidt, C., 2017. Living in a microbial world. Nature Biotechnology, 35(5), p.401.

Papers to discuss (read before the meeting when they are scheduled to be discussed):

4. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E. and Bremges, A., 2017. Critical Assessment of Metagenome Interpretation– a benchmark of computational metagenomics software. Biorxiv, p.099127.

5. Nayfach, S., Rodriguez-Mueller, B., Garud, N. and Pollard, K.S., 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research, 26(11), pp.1612-1625.

6. Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S. and Ahsanuddin, S., 2015. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Systems, 1(1), pp.72-87.

7. Huffnagle, G.B. and Noverr, M.C., 2013. The emerging world of the fungal microbiome. Trends in Microbiology, 21(7), pp.334-341.

● ● ●

02 Statistical methods to refine and redefine phenotypes

Hosted by Andy Dahl
andywdahl@gmail.com

With many multitrait datasets--like EHRs--the observed traits do not parsimoniously or precisely represent the underlying biology. For downstream analysis, the observed traits would ideally be summarized by a small number of latent and unknown traits that describe distinct and clear biological mechanisms. I hope to read papers on related dimensionality reduction problems, including both relevant methodological stats/ML papers and genetics papers using rigorous multitrait methods.

Difficulty: Advanced

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Van Der Maaten, L., Postma, E. and Van den Herik, J., 2009. Dimensionality reduction: a comparative. J Mach Learn Res, 10, pp.66-71.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
1. Lawrence, N., 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of machine learning research, 6(Nov), pp.1783-1816.

2. Cortes, A., Dendrou, C., Motyer, A., Jostins, L., Vukcevic, D., Dilthey, A., Donnelly, P., Leslie, S., Fugger, L. and McVean, G., 2017. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. bioRxiv, p.105122. PLUS SUPPLEMENT.

3. Joshi, S., Gunasekar, S., Sontag, D. and Joydeep, G., 2016, December. Identifiable Phenotyping using Constrained Non-Negative Matrix Factorization. In Machine Learning for Healthcare Conference (pp. 17-41).

● ● ●

03 Modern statistical methods with application to genomics

Hosted by Marzia Cremona
mac78@psu.edu

The goal of this journal club is to review and discuss modern statistical approaches that have been used in genomics research, or that can potentially be applied to analyze genomics data. We will discuss both theoretical aspects and genomics applications. Topics can include functional data analysis, variable selection and sufficient dimension reduction, inference methods, methods for big data, and will be chosen on the basis of participants' interest.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Wang, J.L., Chiou, J.M. and Müller, H.G., 2016. Functional data analysis. Annual Review of Statistics and Its Application, 3, pp.257-295.

Additional papers to potentially discuss:
2. Reimherr, M. and Nicolae, D., 2014. A functional data analysis approach for genetic association studies. The Annals of Applied Statistics, 8(1), pp.406-429.

3. Matsui, H. and Konishi, S., 2011. Variable selection for functional regression models via the L1 regularization. Computational Statistics & Data Analysis, 55(12), pp.3304-3310.

4. Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S. and Miyano, S., 2016. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. Biostatistics, 17(2), pp.235-248.

5. Taylor, S. and Pollard, K., 2009. Hypothesis tests for point-mass mixture data with application to 'omics data with many zero values. Statistical Applications in Genetics and Molecular Biology, 8(8), pp. 1-43.

6. Nye, T.M., 2011. Principal components analysis in the space of phylogenetic trees. The Annals of Statistics, pp.2716-2739.

• • •

04 Outbreak detectives in the genomics era: Computational methods in molecular epidemiology
Hosted by Pavel Skums
pskums@gsu.edu

Molecular epidemiology is a new computationally-intensive discipline, which seek to allow to investigate disease outbreaks and track pathogen transmissions using viral genomic data sampled from infected individuals. In the recent years, computational genomics methods were successfully used for emerging diseases outbreaks (such as Ebola and Zika), as well as for the long-standing epidemics (such as HIV and HCV). The ultimate goal of computational molecular epidemiology is to develop methods allowing to reconstruct transmission histories and answer the question, who infected whom. This task is complicated by incomplete and noisy sequencing and epidemiological data, as well as by the extreme genetic heterogeneity of many viruses, which rapidly evolve within their hosts. We plan to discuss recent computational advances in the area, as well as pose and discuss open computational problems.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Read introductions to papers 3 and 4 (and references therein). There are no review papers in this field yet.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
1. Campo, D.S., Xia, G.L., Dimitrova, Z., Lin, Y., Forbi, J.C., Ganova-Raeva, L., Punkova, L., Ramachandran, S., Thai, H., Skums, P. and Sims, S., 2015. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. The Journal of infectious diseases, 213(6), pp.957-965.

2. Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. and Ferguson, N., 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology, 10(1), p.e1003457.

3. De Maio, N., Wu, C.H. and Wilson, D.J., 2016. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS computational biology, 12(9), p.e1005130.

4. Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D. and Bunimovich, L., 2017. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics.

• • •

05 Introduction to single-cell genomics and new research directions towards a human cell atlas

Hosted by Vasilis Ntranos
ntranos@berkeley.edu

Our main goal in this journal club will be to familiarize ourselves with some of the key problem formulations in single-cell genomics and get exposed to the computational challenges emerging from the new types of data that are becoming available in this field [R1, R2]. After the initial overview [R3], participants will be free to choose specific papers/methods that best align with their interests and have them discussed in more detail by the group. Suggested topics include spatial reconstruction [P1], single-cell entropy in differentiation [P2] and lineage tracing by genome editing [P3]. Participants with diverse backgrounds are welcome, as we would like to engage in broad discussions about new research directions and potentially draw connections to existing methods and ideas from related fields such as phylogenetics and metagenomics.

Difficulty: Introductory/Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
R1. Yuan, G.C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S. and Quackenbush, J., 2017. Challenges and emerging directions in single-cell analysis. Genome Biology, 18(1), p.84.

R2. Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. and Clevers, H., 2017. The Human Cell Atlas. bioRxiv, p.121202.

R3. Wagner, A., Regev, A. and Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
P1. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nature biotechnology, 33(5), pp.495-502.

P2. Teschendorff, A.E. and Enver, T., 2017. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nature Communications, 8, p.15599.

P3. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F. and Shendure, J., 2016. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science, 353(6298), p.aaf7907.

COMPUTATIONAL GENOMICS SUMMER INSTITUTE

SHORT COURSE JULY 10-14

Program Contents

WELCOME

The short course provides didactic training in computational and statistical genomic methods development. This part of the program is relevant to a wide range of trainees who are interested in developing or enhancing the methodology development aspect of their research program. The short course is taught by program faculty who are world leading experts in computational and statistical genomics methodology development. The 2017 CGSI retreat schedule includes research talks, tutorials, journal clubs, and social events.

RESEARCH TALKS are 45-minute explorations of current problems and research in computational genomics by participating faculty.

TUTORIALS are interactive, guided 45-minute workshops that aim to help participants apply new techniques to research problems.

JOURNAL CLUBS are 45-minute sessions that give participants an opportunity to critically evaluate recent articles and stay up-to-date on relevant new research.

LOCATION
The 2017 Computational Genomics Summer Institute Short Course will be held in the California Room at the UCLA Faculty Center. The Faculty Center is situated in the heart of the UCLA campus, a short walk away from Mathematical Sciences, Boyer Hall, and IPAM. An attractive and comfortable private club that opened in 1959, the UCLA Faculty Center offers many well-furnished meeting rooms, lounges, patios, and dining facilities are available for a wide range of social and professional activities.

480 Charles Young Drive, Los Angeles, CA 90095-1617
goo.gl/maps/BY23jYAUmAM2

PARKING
The closest lot to the Faculty Center is Parking Structure 2: maps.ucla.edu/campus/?locid=194

2017 CGSI PROGRAMS
Long Course Retreat: July 6 – 8
Short Course: July 10 – 14
Long Course: July 10 – 26

CGSI ORGANIZING COMMITTEE
Eleazar Eskin, UCLA, CGSI Director
Russel Caflisch, UCLA, IPAM Director
Francesca Chiaromonte, Pennsylvania State University
Eran Halperin, UCLA
David Koslicki, Oregon State University
John Novembre, University of Chicago
Ben Raphael, Princeton University

Visit our website for more about CGSI:
computationalgenomics.bioinformatics.ucla.edu

SCHEDULE

Monday July 10

08:00 - 16:45 Short Course Day One
California Room, UCLA Faculty Center

- 08:00 Registration
  Check in and pick up your name tag during breakfast!
- 09:00 Introduction
  CGSI Organizing Committee
- 09:15 Lior Pachter
  Research Talk: An introduction to pseudoalignment
- 10:00 Ilan Gronau
  Tutorial: Demography inference: From parameter estimation to model selection
- 10:45 Coffee Break
- 11:15 Journal Club
- 12:00 Lunch Break (Lunch is on your own – we recommend the Faculty Center, Wolfgang Puck, the Greenhouse salad and soup bar, and Bombshelter Bistro)
- 13:30 Jennifer Listgarten
  Research Talk: Machine learning for CRISPR gene editing
- 14:15 Coffee Break
- 14:45 Dana Pe'er
  Research Talk: Imputation in Single Cell RNA-seq Data
- 15:30 Coffee Break
- 16:00 Michael Schatz
  Research Talk: Advances in genome sequencing and assembly

17:00 - 18:00 Happy Hour @ Wolfgang Puck Express
Ackerman Union, Level 1, UCLA Campus goo.gl/maps/n4BKLAaynuu

Tuesday July 11

07:30 - 08:30 Morning run and stretching with Dr. Sagi Snir
Meet in front of the UCLA Faculty Center

08:30 - 12:45 Short Course Day Two
California Room, UCLA Faculty Center

- 08:30 Breakfast
- 09:15 James Zou
  Research Talk: Deep learning in genomics: Introduction and examples
- 10:00 Alexander Schönhuth
  Tutorial: Assembling polyploid genomes
- 10:45 Coffee Break

11:15 William (Xiaoquan) Wen
Research Talk: Bayesian approaches for integrative genetic association analysis: Data integration and scalable computation

12:00 Can Alkan
Next-generation sequence characterization of complex genome structural variation

13:00 - 14:30 Teaching Bioinformatics Lunch
UCLA Faculty Center

Grab food and snacks on your way over to the Hollywood Bowl Picnic Area!

13:00 - 18:00 Self-Guided Tour of Hollywood
We encourage CGSI participants to Lyft or Uber to Hollywood for a self-guided tour of the neighborhood.

18:00 - 19:30 Picnic @ Hollywood Bowl Picnic Area 10
Highland Ave, north of Odin St goo.gl/maps/y19NUQ11wc82

20:00 - 22:00 Dudamel & Stars of Ballet featuring Misty Copeland @ Hollywood Bowl
2301 Highland Ave, Los Angeles, CA 90068 goo.gl/maps/ZDsfbcMrbTq

Wednesday July 12

08:30 - 16:45 Short Course Day Three
California Room, UCLA Faculty Center

08:30 Breakfast

09:15 David Koslicki
Tutorial: Comparative metagenomic analysis

10:00 David Tse
Tutorial: How to solve NP-hard assembly problems in linear time

10:45 Coffee Break

11:15 Journal Club

12:00 Lunch Break (Lunch is on your own – we recommend the Faculty Center, Wolfgang Puck, the Greenhouse salad and soup bar, and Bombshelter Bistro)

12:00 - 13:30 UCLA Career Opportunities Lunch
UCLA Faculty Center

13:30 Ran Blekhman
Research Talk: Human genomic control of the microbiome

14:15 Coffee Break

14:45 Daniel Wegmann
Tracing the spread of farming using ancient DNA: Bioinformatic challenges and population genetic insights

15:30 Coffee Break

16:00 Brian Browning
Research Talk: Genotype phasing with large sample sizes

17:00 - 20:00 Picnic, Volleyball, & Soccer @ UCLA Sunset Canyon Recreation Center, Amphitheater Lawn
111 Easton Drive, Los Angeles, CA 90024 goo.gl/maps/Ni8yy6Adcfx

Thursday July 13

07:30 - 08:30 Morning run and stretching with Dr. Sagi Snir
Meet in front of the UCLA Faculty Center

08:30 - 16:30 Short Course Day Four
California Room, UCLA Faculty Center

08:30 Breakfast

09:00 Or Zuk
Tutorial: Co-evolution analysis: Methods and applications

09:45 Itsik Pe'er
Tutorial: Identity by descent in medical and population genomics

10:30 Coffee Break

11:00 Journal Club

11:45 Lunch Break (Lunch is on your own – we recommend the Faculty Center, Wolfgang Puck, the Greenhouse salad and soup bar, and Bombshelter Bistro)

13:00 Lior Pachter
Tutorial: Differential analysis of count data in genomics

13:45 Melissa Gymrek
Analyzing complex repeat variation from high throughput sequencing data

14:30 Coffee Break

15:00 Jonathan Flint
Research Talk: Using low pass sequence data to analyse complex traits

15:45 Elhanan Borenstein
Research Talk: Multi-omic and model-based analysis of the human microbiome

Grab food and snacks on your way over to Santa Monica!

17:30 - 19:00 Bike Ride - Meet at Spokes N' Stuff to bike along Marvin Braude Bike Trail
1715 Ocean Ave, Santa Monica, CA 90401 goo.gl/maps/kx2si2YBkxK2

19:00 - 21:00 Twilight Concert Series @ Santa Monica Pier; meet at Spokes N' Stuff
1715 Ocean Ave, Santa Monica, CA 90401 goo.gl/maps/kx2si2YBkxK2

Friday July 14

08:30 - 16:30 Short Course Day Five
California Room, UCLA Faculty Center

- 08:30 Breakfast
- 09:15 Saharon Rosset
  Tutorial: Quality preserving databases for statistically sound "big data" analysis on public databases
- 10:00 Jason Ernst
  Research Talk: Computational approaches for deciphering the non-coding human genome
- 10:45 Coffee Break
- 11:15 Journal Club
- 12:00 Lunch Break (Lunch is on your own – we recommend the Faculty Center, Wolfgang Puck, the Greenhouse salad and soup bar, and Bombshelter Bistro)
- 13:30 Alex Zelikovsky
  Research Talk: Inference of metabolic pathway activity from metatranscriptomic reads
- 14:15 Coffee Break
- 14:45 Cenk Sahinalp
  Research Talk: Computational methods for intra-tumor heterogeneity detection and modeling clonal evolution of cancer through the use of bulk and single cell sequencing data
- 15:30 Coffee Break
- 16:00 Sagi Snir
  Research Talk: A universal PaceMaker as a better explanation of evolution and aging

17:00 - 18:00 Happy Hour @ Wolfgang Puck Express
Ackerman Union, Level 1, UCLA Campus goo.gl/maps/n4BKLAaynuu

Saturday July 15

09:00 - 12:00 Volleyball & Soccer @ Santa Monica Beach; meet at Perry's Café and Beach Rentals
2600 Ocean Front Walk, Santa Monica, CA 90405 goo.gl/maps/CqTnMH4YBaJ2

LOOKING FOR A FANTASTIC RESTAURANT RECOMMENDATION FOR DINNER ANYWHERE IN LOS ANGELES?

ASK THE LOCAL CGSI ORGANIZERS FOR ADVICE!

PRESENTATION TITLES AND PAPERS

This year's Short Course features twenty-five 45-minute research talks and tutorials presented by prominent scholars in the field of computational genomics. Each speaker compiled a list of relevant papers that provide a more in-depth exploration of their presentation material. Click to open a paper in your internet browser.

01 Lior Pachter: Research Talk: An introduction to pseudoalignment

1. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology, 34(5), pp.525-527.

02 Ilan Gronau: Tutorial: Demography inference: From parameter estimation to model selection

1. Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A., 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature genetics, 43(10), pp.1031-1034.

2. Hey, J. and Nielsen, R., 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. PNAS, 104(8), pp.2785-2790.

3. Rannala, B. and Yang, Z., 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Systematic Biology, p.syw119.

4. Chung, Y. and Hey, J., 2016. Bayesian Analysis of Evolutionary Divergence with Genomic Data Under Diverse Demographic Models. bioRxiv, p.080606.

03 Jennifer Listgarten: Research Talk: Machine learning for CRISPR gene editing

1. Fusi, N., Smith, I., Doench, J. and Listgarten, J., 2015. In silico predictive modeling of CRISPR/Cas9 guide efficiency. bioRxiv, p.021568.

2. Listgarten, J., Weinstein, M., Elibol, M., Hoang, L., Doench, J. and Fusi, N., 2016. Predicting off-target effects for end-to-end CRISPR guide design. bioRxiv, p.078253.

3. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R. and Virgin, H.W., 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology.

04 Dana Pe'er: Research Talk: Imputation in Single Cell RNA-seq Data

1. van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A.J., et al., 2017. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv, p.111591.

2. Prabhakaran, S., et al., 2016. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In Proc. of The 33rd International Conference on Machine Learning (pp. 1070-1079).

05 Michael Schatz: Research Talk: Advances in genome sequencing and assembly

1. Schatz, M.C., Delcher, A.L. and Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. Genome research, 20(9), pp.1165-1173.

2. Berlin, K., Koren, S., Chin, C.S., Drake, J.P., et al., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nature biotechnology, 33(6), pp.623-630.

3. Loman, N.J., Quick, J. and Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature methods, 12(8), pp.733-735.

4. Weisenfeld, N.I., Kumar, V., Shah, P., Church, D. and Jaffe, D.B., 2016. Direct determination of diploid genome sequences. bioRxiv, p.070425.

5. Zook, Justin M., et al. “Extensive sequencing of seven human genomes to characterize benchmark reference materials.” Scientific data 3 (2016).

06 Alexander Schönhuth: Tutorial: Assembling polyploid genomes

1. Baaijens, J.A., El Aabidine, A.Z., Rivals, E. and Schönhuth, A., 2017. De novo assembly of viral quasispecies using overlap graphs. Genome Research, 27(5), pp.835-848.

2. Välimäki, N., Ladra, S. and Mäkinen, V., 2012. Approximate all-pairs suffix/prefix overlaps. Information and Computation, 213, pp.49-58.

3. Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, p.btw152.

07 William (Xiaoquan) Wen: Research Talk: Bayesian approaches for integrative genetic association analysis: Data integration and scalable computation

1. Wen, X., Lee, Y., Luca, F. and Pique-Regi, R., 2016. Efficient integrative multi-SNP association analysis via Deterministic Approximation of Posteriors. The American Journal of Human Genetics, 98(6), pp.1114-1129.

2. Wen, X., et al, 2016. Integrating Molecular QTL Data into Genome-wide Genetic Association Analysis: Probabilistic Assessment of Enrichment and Colocalization. PLOS Genetics. 2017 Mar 13(3): e1006646.

08 James Zou: Research Talk: Deep learning in genomics: Introduction and examples

No papers assigned.

09 Can Alkan: Research Talk: Next-generation sequence characterization of complex genome structural variation

1. Alkan, C., Coe, B.P. and Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nature Reviews Genetics, 12(5), pp.363-376.

2. Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O. and Sahinalp, S.C., 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics, 41(10), pp.1061-1067.

3. Hormozdiari, F., Alkan, C., Eichler, E.E. and Sahinalp, S.C., 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome research, 19(7), pp.1270-1278.

4. Chiatante, G., Miroballo, M., Tang, J., Ventura, M., Amemiya, C.T., Eichler, E.E., Antonacci, F. and Alkan, C., 2017. Discovery of large genomic inversions using long range information. BMC genomics, 18(1), pp.65-65.

5. Layer, R.M., Chiang, C., Quinlan, A.R. and Hall, I.M., 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome biology, 15(6), p.R84.

10 David Koslicki: Tutorial: The CAMI Project: Assessment of computational techniques in metagenomics

1. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E. and Bremges, A., 2017. Critical Assessment of Metagenome Interpretation– a benchmark of computational metagenomics software. Biorxiv, p.099127.
2. CAMI: Critical Assessment of Metagenomic Interpretation. Website. Accessed July 5, 2017.

3. CAMI: Critical Assessment of Metagenomic Interpretation. Data. Website. Accessed July 5, 2017.

11 David Tse: Tutorial: How to solve NP-hard assembly problems in linear time

1. Bresler, G., Bresler, M.A. and Tse, D., 2013. Optimal assembly for high throughput shotgun sequencing. BMC bioinformatics, 14(5), p.S18.

2. Shomorony, I., Kim, S.H., Courtade, T.A. and David, N.C., 2016. Information-optimal genome assembly via sparse read-overlap graphs. Bioinformatics, 32(17), pp.i494-i502.

3. Kamath, G.M., Shomorony, I., Xia, F., Courtade, T. and David, N.T., 2017. Hinge: Long-read assembly achieves optimal repeat resolution. Genome Research, pp.gr-216465.

4. Kannan, S., Hui, J., Mazooji, K., Pachter, L. and Tse, D., 2016. Shannon: An Information-Optimal de Novo RNA-Seq Assembler. bioRxiv, p.039230.

12 Ran Blekhman: Research Talk: Human genomic control of the microbiome

1. Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T. and Spector, T.D., 2014. Human genetics shape the gut microbiome. Cell, 159(4), pp.789-799.

2. Blekhman, R., Goodrich, J.K., Huang, K., Sun, Q., Bukowski, R., Bell, J.T., Spector, T.D., Keinan, A., Ley, R.E., Gevers, D. and Clark, A.G., 2015. Host genetic variation impacts microbiome composition across human body sites. Genome biology, 16(1), p.191.

3. Burns, M.B., Lynch, J., Starr, T.K., Knights, D. and Blekhman, R., 2015. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome medicine, 7(1), p.55.

4. Lynch, J., Tang, K., Sands, J., Sands, M., Tang, E., Mukherjee, S., Knights, D. and Blekhman, R., 2016. HOMINID: A framework for identifying associations between host genetic variation and microbiome composition. bioRxiv, p.081323.

5. Burns, M.B., Montassier, E., Abrahante, J., Starr, T.K., Knights, D. and Blekhman, R., 2016. Discrete mutations in colorectal cancer correlate with defined microbial communities in the tumor microenvironment. bioRxiv, p.090795.

13 Daniel Wegmann: Research Talk: Tracing the spread of farming using ancient DNA: Bioinformatic challenges and population genetic insights

1. Kousathanas, A., Leuenberger, C., Link, V., Sell, C., Burger, J. and Wegmann, D., 2017. Inferring heterozygosity from ancient and low coverage genomes. Genetics, 205(1), pp.317-332.

2. Broushaki, F., Thomas, M.G., Link, V., López, S., van Dorp, L., Kirsanow, K., Hofmanová, Z., Diekmann, Y., Cassidy, L.M., Díez-del-Molino, D. and Kousathanas, A., 2016. Early Neolithic genomes from the eastern Fertile Crescent. Science, 353(6298), pp.499-503.

3. Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-del-Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V. and Kirsanow, K., 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. Proceedings of the National Academy of Sciences, p.201523951.

14 Brian Browning: Research Talk: Genotype phasing with large sample sizes

1. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T. and Sulem, P., 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nature genetics, 40(9), pp.1068-1075.

2. Browning, S.R. and Browning, B.L., 2011. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics, 12(10), pp.703-714.

3. O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.F., Delaneau, O. and Marchini, J., 2016. Haplotype estimation for biobank-scale data sets. Nature Publishing Group.

4. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R. and Durbin, R., 2016. Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics, 48(11), pp.1443-1448.

15 Or Zuk: Tutorial: Co-evolution analysis: Methods and applications

1. Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist, 125(1), pp.1-15.

2. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences, 96(8), pp.4285-4288.

3. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T. and Weigt, M., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences, 108(49), pp.E1293-E1301.

16 Itsik Pe'er: Tutorial: Identity by descent in medical and population genomics

1. Palamara, P.F., Francioli, L.C., et al, 2015. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. The American Journal of Human Genetics, 97(6), pp.775-789.

2. Genome of the Netherlands Consortium, 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature Genetics, 46(8), pp.818-825.

3. Palamara, P.F. and Pe'er, I., 2013. Inference of historical migration rates via haplotype sharing. Bioinformatics, 29(13), pp.i180-i188.

4. Palamara, P.F., Lencz, T., Darvasi, A. and Pe'er, I., 2012. Length distributions of identity by descent reveal fine-scale demographic history. The American Journal of Human Genetics, 91(5), pp.809-822.

5. Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., et al., 2012. The architecture of long-range haplotypes shared within and across populations. Molecular biology and evolution, 29(2), pp.473-486.

6. Gusev, A., Kenny, E.E., Lowe, et al., 2011. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. The American Journal of Human Genetics, 88(6), pp.706-717.

7. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M. and Pe'er, I., 2009. Whole population, genome-wide mapping of hidden relatedness. Genome research, 19(2), pp.318-326.

17 Lior Pachter: Tutorial: Differential analysis of count data in genomics

1. Anders, S. and Huber, W., 2010. Differential expression analysis for sequence count data. Genome biology, 11(10), p.R106.

2. Soneson, C., Love, M.I. and Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research, 4.

3. Pimentel, H., Bray, N.L., Puente, S., Melsted, P. and Pachter, L., 2017. Differential analysis of RNA-Seq incorporating quantification uncertainty. Nature Methods.

18 Melissa Gymrek: Research Talk: Analyzing complex repeat variation from high throughput sequencing data

1. Gymrek, M., Golan, D., Rosset, S. and Erlich, Y., 2012. lobSTR: a short tandem repeat profiler for personal genomes. Genome research, 22(6), pp.1154-1162.

2. Willems, T., Zielinski, D., Gordon, A., Gymrek, M. and Erlich, Y., 2016. Genome-wide profiling of heritable and de novo STR variations. bioRxiv, p.077727.

3. Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., et al., 2015. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature genetics.

4. Gymrek, M., Willems, T., Reich, D.E. and Erlich, Y., 2016. A framework to interpret short tandem repeat variation in humans. bioRxiv, p.092734.

19 Leonid Kruglyak: Research Talk: Complex Traits and Simple Systems

No papers assigned.

20 Elhanan Borenstein: Research Talk: Multi-omic and model-based analysis of the human microbiome

1. Manor, O. and Borenstein, E., 2017. Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome. Cell Host & Microbe, 21(2), pp.254-267.

2. Noecker, C., Eng, A., Srinivasan, S., Theriot, C.M., et al, 2016. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. mSystems, 1(1), pp.e00013-15.

3. Greenblum, S., Carr, R. and Borenstein, E., 2015. Extensive strain-level copy-number variation across human gut microbiome species. Cell, 160(4), pp.583-594.

21 Saharon Rosset: Tutorial: Stochastic process models for mutations, their estimation from data, and their uses

1. Huelsenbeck, J.P. and Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics, 28(1), pp.437-466.

2. Tamura, K. and Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular biology and evolution, 10(3), pp.512-526.

3. Nielsen, R., 2005. Statistical methods in molecular evolution (Vol. 6). New York: Springer.

4. Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G. and Sibly, R.M., 2003. Likelihood-based estimation of microsatellite mutation rates. Genetics, 164(2), pp.781-787.

22 Jason Ernst: Research Talk: Computational approaches for deciphering the non-coding human genome

1. Ernst, J. and Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology, 28(8), pp.817-825.

2. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., et al., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), pp.43-49.

3. Ernst, J. and Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. Nature methods, 9(3), pp.215-216.

4. Ernst, J. and Kellis, M., 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature biotechnology, 33(4), pp.364-376.

5. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., et al, 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nature Biotechnology, 34(11), pp.1180-1190.

23 Alex Zelikovsky: Research Talk: Inference of metabolic pathway activity from metatranscriptomic reads

1. Temate-Tiagueu, Y., Al Seesi, S., Mathew, M., et al, 2016. Inferring metabolic pathway activity levels from RNA-Seq data. BMC genomics, 17(5), p.542.

2. Mathew, M., et al, 2016. Influence of symbiont-produced bioactive natural products on holobiont fitness in the marine bryozoan, Bugula neritina via protein kinase C (PKC). Marine biology, 163(2), pp.1-17.

3. Glebova, O., Temate-Tiagueu, Y., et al, 2016. Transcriptome Quantification and Differential Expression from NGS Data. Computational Methods for Next Generation Sequencing Data Analysis, pp.301-327.

4. Nicolae, M., Mangul, S., Măndoiu, I.I. and Zelikovsky, A., 2011. Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms for molecular biology, 6(1), p.9.

24 Cenk Sahinalp: Research Talk: HIT'nDRIVE: Patient-Specific Multi-Driver Gene Prioritization for Precision Oncology

1. El-Kebir, M., Satas, G., Oesper, L. and Raphael, B.J., 2016. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Systems, 3(1), pp.43-53.

2. El-Kebir, M., Oesper, L., Acheson-Field, H. and Raphael, B.J., 2015. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics, 31(12), pp.i62-i70.

3. Oesper, L., Satas, G. and Raphael, B.J., 2014. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics, 30(24), pp.3532-3540.

4. Salehi, S., Steif, A., Roth, A., Aparicio, S., et al., 2017. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. Genome Biology, 18(1), p.44.

5. Roth, A., McPherson, A., Laks, E., Biele, J., et al., 2016. Clonal genotype and population structure inference from single-cell tumor sequencing. Nature methods, 13(7), pp.573-576.

6. Ha, G., Roth, A., Khattra, J., et al., 2014. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome research, 24(11), pp.1881-1893.

7. Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., et al, 2014. PyClone: statistical inference of clonal population structure in cancer. Nature methods, 11(4), pp.396-398.

8. Jahn, K., Kuipers, J. and Beerenwinkel, N., 2016. Tree inference for single-cell data. Genome biology, 17(1), p.86.

25 Sagi Snir: Research Talk: A universal PaceMaker as a better explanation of evolution and aging

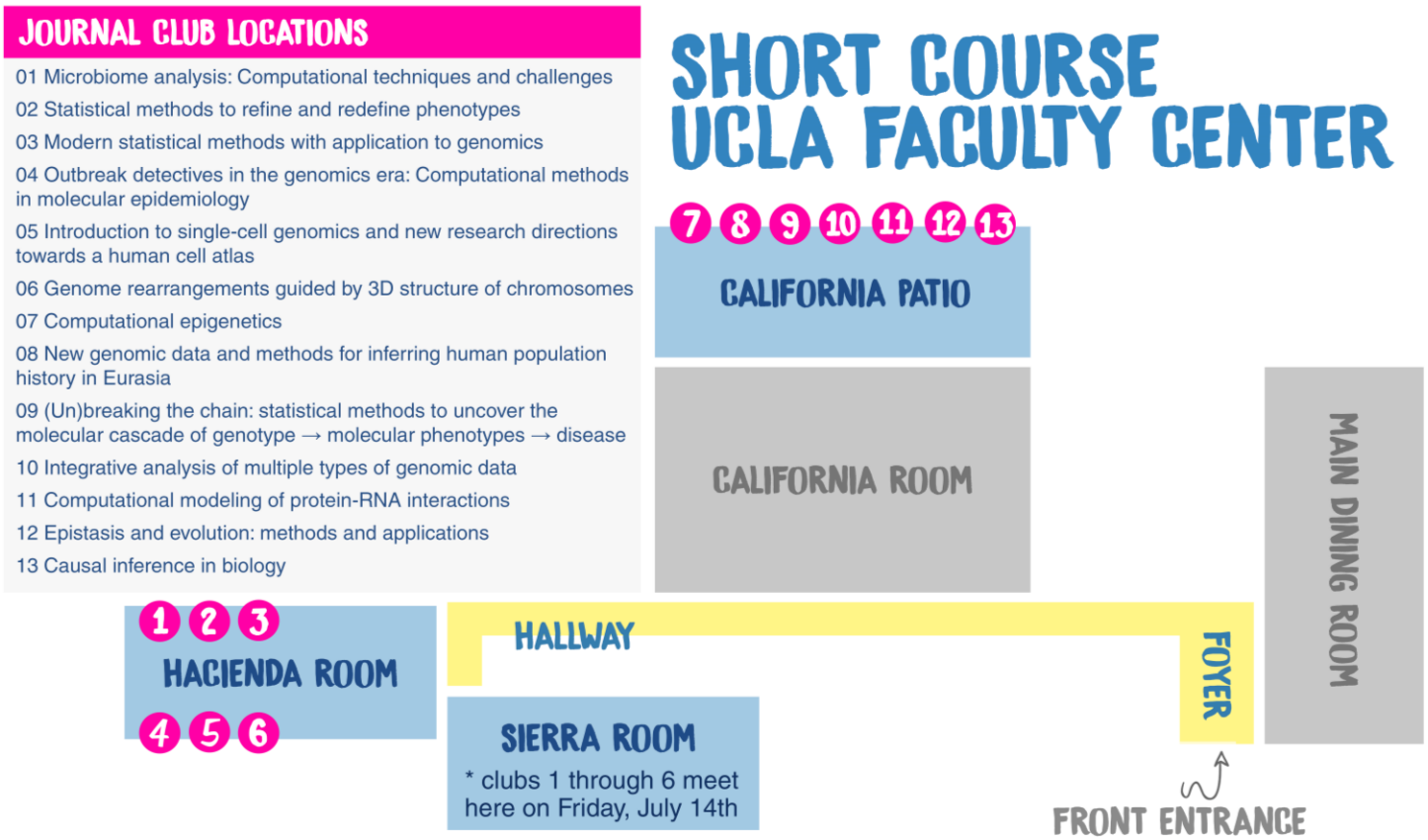
1. Snir, S., Wolf, Y.I. and Koonin, E.V., 2012. Universal pacemaker of genome evolution. PLoS Comput Biol, 8(11), p.e1002785.

2. Snir, S. and Pellegrini, M., 2016. A Statistical Framework to Identify Deviation from Time Linearity in Epigenetic Aging. PLoS Comput Biol, 12(11), p.e1005183.

JOURNAL CLUBS

As part of our Short Course, we will be breaking into small groups for daily journal clubs. Journal clubs provide an excellent opportunity to discuss current work and doing so with colleagues with shared interests can be insightful, productive, and fun.

Browse our list of thirteen journal clubs and use this form to sign up for a journal club of your choice: goo.gl/forms/yPeP92qoDlUNQMYN2. Note required reading for each.



01 Microbiome analysis: Computational techniques and challenges
Hosted by Serghei Mangul
serghei@cs.ucla.edu
July 6 – 26

Technological advances and the decreasing costs of 'next-generation' sequencing (NGS) make it the technology of choice for many applications, including studying the human microbiome composed of bacterial, viral, fungi and other eukaryotic communities. Recently, high-throughput sequencing has revolutionized microbiome research by enabling the study of thousands of microbial genomes directly in their host environments. This approach, which forms the field of metagenomics, avoids the biases incurred with traditional culture-dependent analysis. The metagenomics approach also allows the comparison of microbial communities' composition in their natural habitats across different human tissues and environmental settings. Specifically, metagenomic profiling is proven useful for analyzing microbes such as eukaryotic and viral pathogens, which were previously impossible to study in an unbiased way with target 16S ribosomal RNA gene.

Tentative list of topics: We will be discussing recent methods to study microbial communities. We will be discussing the challenges in metagenomics analysis and limitation of the current methods. The goal will be to identify the best strategy to analyze metagenomics data.

We will start with discussion the most popular 'marker genes' methods, which are suggested to have poor sensitivity and also may result in false positives, detecting dangerous pathogens which are not present in the metagenomics sample. We also will discuss methods aimed to study microbiome at strain level and methods to study non-bacterial organisms, including viruses and fungi.

Outcomes: One possible outcome can be a joint effort to write an educational paper introducing metagenomics for researchers with no background in computational genomics or bioinformatics.

Difficulty: Intro/Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):

1. Escobar-Zepeda, A., de León, A.V.P. and Sanchez-Flores, A., 2015. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in Genetics, 6.

2. Simon, C. and Daniel, R., 2011. Metagenomic analyses: past and future trends. Applied and Environmental Microbiology, 77(4), pp.1153-1161.

3. Schmidt, C., 2017. Living in a microbial world. Nature Biotechnology, 35(5), p.401.

Papers to discuss (read before the meeting when they are scheduled to be discussed):

4. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E. and Bremges, A., 2017. Critical Assessment of Metagenome Interpretation– a benchmark of computational metagenomics software. Biorxiv, p.099127.

5. Nayfach, S., Rodriguez-Mueller, B., Garud, N. and Pollard, K.S., 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research, 26(11), pp.1612-1625.

6. Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S. and Ahsanuddin, S., 2015. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Systems, 1(1), pp.72-87.

7. Huffnagle, G.B. and Noverr, M.C., 2013. The emerging world of the fungal microbiome. Trends in Microbiology, 21(7), pp.334-341.

02 Statistical methods to refine and redefine phenotypes

Hosted by Andy Dahl
andywdahl@gmail.com
July 6 – 26

With many multitrait datasets--like EHRs--the observed traits do not parsimoniously or precisely represent the underlying biology. For downstream analysis, the observed traits would ideally be summarized by a small number of latent and unknown traits that describe distinct and clear biological mechanisms. I hope to read papers on related dimensionality reduction problems, including both relevant methodological stats/ML papers and genetics papers using rigorous multitrait methods.

Difficulty: Advanced

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Van Der Maaten, L., Postma, E. and Van den Herik, J., 2009. Dimensionality reduction: a comparative. J Mach Learn Res, 10, pp.66-71.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
1. Lawrence, N., 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of machine learning research, 6(Nov), pp.1783-1816.

2. Cortes, A., Dendrou, C., Motyer, A., Jostins, L., Vukcevic, D., Dilthey, A., Donnelly, P., Leslie, S., Fugger, L. and McVean, G., 2017. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. bioRxiv, p.105122. PLUS SUPPLEMENT.

3. Joshi, S., Gunasekar, S., Sontag, D. and Joydeep, G., 2016, December. Identifiable Phenotyping using Constrained Non-Negative Matrix Factorization. In Machine Learning for Healthcare Conference (pp. 17-41).

03 Modern statistical methods with application to genomics
Hosted by Marzia Cremona
mac78@psu.edu
July 6 – 26

The goal of this journal club is to review and discuss modern statistical approaches that have been used in genomics research, or that can potentially be applied to analyze genomics data. We will discuss both theoretical aspects and genomics applications. Topics can include functional data analysis, variable selection and sufficient dimension reduction, inference methods, methods for big data, and will be chosen on the basis of participants' interest.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Wang, J.L., Chiou, J.M. and Müller, H.G., 2016. Functional data analysis. Annual Review of Statistics and Its Application, 3, pp.257-295.

Additional papers to potentially discuss:
2. Reimherr, M. and Nicolae, D., 2014. A functional data analysis approach for genetic association studies. The Annals of Applied Statistics, 8(1), pp.406-429.

3. Matsui, H. and Konishi, S., 2011. Variable selection for functional regression models via the L1 regularization. Computational Statistics & Data Analysis, 55(12), pp.3304-3310.

4. Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S. and Miyano, S., 2016. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. Biostatistics, 17(2), pp.235-248.

5. Taylor, S. and Pollard, K., 2009. Hypothesis tests for point-mass mixture data with application to 'omics data with many zero values. Statistical Applications in Genetics and Molecular Biology, 8(8), pp. 1-43.

6. Nye, T.M., 2011. Principal components analysis in the space of phylogenetic trees. The Annals of Statistics, pp.2716-2739.

04 Outbreak detectives in the genomics era: Computational methods in molecular epidemiology
Hosted by Pavel Skums
pskums@gsu.edu
July 6 – 26

Molecular epidemiology is a new computationally-intensive discipline, which seek to allow to investigate disease outbreaks and track pathogen transmissions using viral genomic data sampled from infected individuals. In the recent years, computational genomics methods were successfully used for emerging diseases outbreaks (such as Ebola and Zika), as well as for the long-standing epidemics (such as HIV and HCV). The ultimate goal of computational molecular epidemiology is to develop methods allowing to reconstruct transmission histories and answer the question, who infected whom. This task is complicated by incomplete and noisy sequencing and epidemiological data, as well as by the extreme genetic heterogeneity of many viruses, which rapidly evolve within their hosts. We plan to discuss recent computational advances in the area, as well as pose and discuss open computational problems.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Read introductions to papers 3 and 4 (and references therein). There are no review papers in this field yet.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
1. Campo, D.S., Xia, G.L., Dimitrova, Z., Lin, Y., Forbi, J.C., Ganova-Raeva, L., Punkova, L., Ramachandran, S., Thai, H., Skums, P. and Sims, S., 2015. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. The Journal of infectious diseases, 213(6), pp.957-965.

2. Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. and Ferguson, N., 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology, 10(1), p.e1003457.

3. De Maio, N., Wu, C.H. and Wilson, D.J., 2016. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS computational biology, 12(9), p.e1005130.

4. Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D. and Bunimovich, L., 2017. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics.

05 Introduction to single-cell genomics and new research directions towards a human cell atlas
Hosted by Vasilis Ntranos
ntranos@berkeley.edu
July 6 – 26

Our main goal in this journal club will be to familiarize ourselves with some of the key problem formulations in single-cell genomics and get exposed to the computational challenges emerging from the new types of data that are becoming available in this field [R1, R2]. After the initial overview [R3], participants will be free to choose specific papers/methods that best align with their interests and have them discussed in more detail by the group. Suggested topics include spatial reconstruction [P1], single-cell entropy in differentiation [P2] and lineage tracing by genome editing [P3]. Participants with diverse backgrounds are welcome, as we would like to engage in broad discussions about new research directions and potentially draw connections to existing methods and ideas from related fields such as phylogenetics and metagenomics.

Difficulty: Introductory/Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
R1. Yuan, G.C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S. and Quackenbush, J., 2017. Challenges and emerging directions in single-cell analysis. Genome Biology, 18(1), p.84.

R2. Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. and Clevers, H., 2017. The Human Cell Atlas. bioRxiv, p.121202.

R3. Wagner, A., Regev, A. and Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
P1. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nature biotechnology, 33(5), pp.495-502.

P2. Teschendorff, A.E. and Enver, T., 2017. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nature Communications, 8, p.15599.

P3. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F. and Shendure, J., 2016. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science, 353(6298), p.aaf7907.

06 Genome rearrangements guided by 3D structure of chromosomes
Hosted by Nikita Alexeev
nikita.v.alexeev@gmail.com
July 10 – 14

Genome rearrangements are evolutionary events that shuffle genomic architectures, which break a genome at several positions and glue the resulting fragments in a new order. They were widely studied with graph theory methods. The common belief is that rearrangements are possible between the fragile genome regions which are close in 3D. In this journal club we are going to discuss how knowledge about 3D structure of chromosomes obtained with Hi-C protocol allows us to understand the nature of genome rearrangements and discover the transformations that happened between different genomes.

Difficulty: Intermediate (requires a basic understanding of graph theory)

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. The following manuscript: Guillaume Fertin, Anthony Labarre, Irena Rusu, Eric Tannier and Stéphane Vialette. Combinatorics of Genome Rearrangements. MIT press, 2009.
Available at: https://mitpress.mit.edu/sites/default/files/titles/content/9780262062824_sch_0001.pdf

2. Chapters 5.1-5.2 of: Jones, N.C. and Pevzner, P., 2004. An introduction to bioinformatics algorithms. MIT press. Accessible at: https://www.dropbox.com/s/z97nf7cvk6j5cwb/Chaters5.2to5.2.pdf?dl=0

Papers to discuss (read before the meeting when they are scheduled to be discussed):
3. Swenson, K.M., Simonaitis, P. and Blanchette, M., 2016. Models and algorithms for genome rearrangement with positional constraints. Algorithms for Molecular Biology, 11(1), p.13.

4. Véron, A.S., Lemaitre, C., Gautier, C., Lacroix, V. and Sagot, M.F., 2011. Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. BMC Genomics, 12(1), p.303.

5. Pulicani, S., Simonaitis, P. and Swenson, K.M., 2017. Rearrangement Scenarios Guided By Chromatin Structure. bioRxiv, p.137323.

07 Computational epigenetics
Hosted by Chloe Robins
crobi27@emory.edu
July 10 – 14

A survey of computational methods for the analysis of epigenetic data, from DNA methylation to chromatin-level modifications. We will start with DNA methylation and move from there.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics, 13(7), pp.484-492.

2. Bock, C., 2012. Analysing and interpreting DNA methylation data. Nature Reviews Genetics, 13(10), pp.705-719.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
3. Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P. and Conneely, K.N., 2015. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Research, 43(21), p.e141.

4. Hansen, K.D., Langmead, B. and Irizarry, R.A., 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biology, 13(10), p.R83.

5. Slieker, R.C., van Iterson, M., Luijk, R., Beekman, M., Zhernakova, D.V., Moed, M.H., Mei, H., Van Galen, M., Deelen, P., Bonder, M.J. and Zhernakova, A., 2016. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biology, 17(1), p.191.

6. Bell, C.G., Xia, Y., Yuan, W., Gao, F., Ward, K., Roos, L., Mangino, M., Hysi, P.G., Bell, J., Wang, J. and Spector, T.D., 2016. Novel regional age-associated DNA methylation changes within human common disease-associated loci. Genome Biology, 17(1), p.193.

7. Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T. and Marsit, C.J., 2016. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics, 17(1), p.259.

8. Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G., Eskin, E., Zou, J. and Halperin, E., 2016. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nature Methods, 13(5), pp.443-445.

08 New genomic data and methods for inferring human population history in Eurasia
Hosted by Ilan Gronau
ilan.gronau@idc.ac.il
July 10 – 14

The past couple of years have produced an extreme wealth of genome sequence data that can be used to retell the story of human population dispersal out of Africa and into Eurasia. We will review some of the main sources of data that emerged from these studies (present-day and ancient DNA), as well as the statistical

methods used to produce demographic models from these data. Some of the interesting questions addressed in these studies: how many waves of migration out of Africa can we trace in present-day Eurasian populations? How was Europe populated? How do present-day populations relate to early farmers of the Middle East?

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. and Willerslev, E., 2017. Tracing the peopling of the world through genomics. Nature, 541(7637), pp.302-310.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
2. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. and Skoglund, P., 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature, 538(7624), pp.201-206.

3. Pagani, L., Lawson, D.J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L. and Wall, J.D., 2016. Genomic analyses inform on migration events during the peopling of Eurasia. Nature, 538(7624), pp.238-242.

4. Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A. and Nickel, B., 2016. The genetic history of Ice Age Europe. Nature, 534(7606), pp.200-205.

5. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K. and Connell, S., 2016. Genomic insights into the origin of farming in the ancient Near East. Nature, 536(7617), pp.419-424.

6. Malaspinas, A.S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergstrom, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E. and Heupink, T.H., 2016. A genomic history of Aboriginal Australia. Nature, 538(7624), pp.207-207.

7. Lipson, M. and Reich, D., 2017. working model of the deep relationships of diverse modern human genetic lineages outside of Africa. Molecular Biology and Evolution, p.msw293.

09 (Un)breaking the chain: statistical methods to uncover the molecular cascade of genotype → molecular phenotypes → disease
Hosted by Nick Mancuso
nmancuso@mednet.ucla.edu
July 10 – 14

Genome-wide association studies have been wildly successful in identifying genomic regions associated with disease risk. To date, thousands of loci have been reproducibly identified, yet most fail to provide mechanistic insight. Multiple lines of evidence have demonstrated significant enrichment of GWAS risk loci in functional regions of the genome, which suggests regulatory control of intermediate phenotypes (e.g., gene expression, splice variation, chromatin state). Taken together, this paints a broad landscape where genetic variation influences intermediate molecular phenotypes and ultimately disease risk. Recently, several nascent computational approaches have been proposed that link genetic variation to intermediate phenotype and disease risk. This journal club will review the state-of-the-art in this area of research and prepare attendees for independent investigation.

Difficulty: Intermediate/Advanced

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):

1. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., De Geus, E.J., Boomsma, D.I., Wright, F.A. and Sullivan, P.F., 2016. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics, 48(3), pp.245-252.

2. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A. and Pasaniuc, B., 2017. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics, 100(3), pp.473-487.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
3. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., McCaroll, S., Neale, B., Ophoff, R. and O'Donovan, M.C., 2016. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. bioRxiv, p.067355.

4. Park, Y., Sarkar, A.K., Bhutani, K. and Kellis, M., 2017. Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv, p.107623.

10 Integrative analysis of multiple types of genomic data
Hosted by William Wen
xwen@umich.edu
July 10 – 14

The goal here is to survey the current literature of integrative analysis of multiple types of genomic data to (1) perform fine-mapping of genetic association signals; (2) understand molecular mechanism of complex traits; and (3) variant effect prediction.

Difficulty: Beginner to intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. and Kim, D., 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics, 16(2), pp.85-97.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
2. Pickrell, J.K., 2014. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics, 94(4), pp.559-573.

3. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E. and Kähler, A.K., 2014. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics, 95(5), pp.535-552.

4. Ionita-Laza, I., McCallum, K., Xu, B. and Buxbaum, J.D., 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature genetics, 48(2), pp.214-220.

11 Computational modeling of protein-RNA interactions
Hosted by Yaron Orenstein
yaronore@mit.edu
July 10 – 14

Protein-RNA interactions, mediated through both RNA sequence and structure, play vital role in all cellular processes. In recent years, technologies have been developed to measure these interactions in high-throughput manner. In this journal club, we will discuss computational solutions in modeling protein-RNA binding from these data, and focus on how RNA structure is incorporate in these models.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Cook, K.B., Hughes, T.R. and Morris, Q.D., 2014. High-throughput characterization of protein–RNA interactions. Briefings in functional genomics, 14(1), pp.74-89.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
2. Maticzka, D., Lange, S.J., Costa, F. and Backofen, R., 2014. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biology, 15(1), p.R17.

3. Orenstein, Y., Wang, Y. and Berger, B., 2016. RCK: accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data. Bioinformatics, 32(12), pp.i351-i359.

12 Epistasis and evolution: methods and applications
Hosted by Or Zuk
or.zuk@mail.huji.ac.il
July 10 – 14

The effect of many genetic variants is mediated by other variants, leading to genetic interactions, also termed epistasis. Such interactions affect the selection forces acting on individual variants and can thus leave signals of co-evolution when looking at these variants in genomes of related species or of individuals from the same species. Consequently, these signals can be used to detect pairs of interacting variants, and to computationally infer the role of individual variants, including for example contacts of amino-acids in a protein, and compensatory cis-regulatory mutations. We will review recent papers which study the evolution of both regulatory and coding interacting variants and discuss their models and computational approaches, with the goal being proposing modifications and extensions of the new methods to large-scale genomic datasets.

Difficulty: Advanced

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Phillips, P.C., 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics, 9(11), pp.855-867.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
2. Jordan, D.M., Frangakis, S.G., Golzio, C., Cassa, C.A., Kurtzberg, J., Davis, E.E., Sunyaev, S.R. and Katsanis, N., 2015. Identification of cis-suppression of human disease mutations by comparative genomics. Nature, 524(7564), pp.225-229.

3. Sohail, M., Vakhrusheva, O.A., Sul, J.H., Pulit, S.L., Francioli, L.C., van den Berg, L.H., Veldink, J.H., de Bakker, P.I., Bazykin, G.A., Kondrashov, A.S. and Sunyaev, S.R., 2017. Negative selection in humans and fruit flies involves synergistic epistasis. Science, 356(6337), pp.539-542.

4. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P., Springer, M., Sander, C. and Marks, D.S., 2017. Mutation effects predicted from sequence co-variation. Nature Biotechnology, 35(2), pp.128-135.

13 Causal inference in biology
Hosted by Michael Bilow
bilow@cs.ucla.edu
July 10 – 14

In this journal club, we'll cover some of the basics of causal inference as it relates to molecular biology, with a particular focus on inference on biological networks. We'll also be covering computational tools to solve biologically-focused problems, especially distributed graph processing using GraphX.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):

1. Kleinberg, S. and Hripcsak, G., 2011. A review of causal inference for biomedical informatics. Journal of biomedical informatics, 44(6), pp.1102-1112.

2. Pearl, J., 2009. Causal inference in statistics: An overview. Statistics surveys, 3, pp.96-146.

Papers to discuss (read before the meeting when they are scheduled to be discussed):

3. Şenbabaoğlu, Y., Sümer, S.O., Sánchez-Vega, F., Bemis, D., Ciriello, G., Schultz, N. and Sander, C., 2016. A multi-method approach for proteomic network inference in 11 human cancers. PLoS computational biology, 12(2), p.e1004765.

4. Djordjevic, D., Yang, A., Zadoorian, A., Rungrugeecharoen, K. and Ho, J.W., 2014. How difficult is inference of mammalian causal gene regulatory networks?. PloS one, 9(11), p.e111661.

5. Siahpirani, A.F. and Roy, S., 2017. A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic acids research, 45(4), pp.e21-e21.

COMPUTATIONAL GENOMICS SUMMER INSTITUTE

JULY 6 - 26
UCLA CAMPUS

SOCIAL PROGRAM

Welcome to Los Angeles! As part of the CGSI Social Program, we would like to share with you several amazing events that are part of the "summer in LA" experience.

01 Tuesday, July 11th – Hollywood Excursion and Bowl Concert

Join us for an afternoon in Hollywood; show starts at 8:00pm. We will go to the iconic Hollywood Bowl to attend a live performance of the Stars of Ballet featuring Misty Copeland and conductor Gustavo Dudamel playing classical music. In addition, you can leave early in the afternoon for lunch and a self-guided tour in Hollywood!

Getting to Hollywood.
The CGSI Short Course ends early on Tuesday, July 11—at 12:45pm or, if attending the *Teaching Bioinformatics Lunch*, at 2:30pm. There are two ways to get to Hollywood:

1. Early in the afternoon, split into smaller groups and take Uber or Lyft to the historic Hollywood district. The *average* cost of an Uber or Lyft ride from Westwood to Hollywood is $18 (please note that we cannot control the rate of Uber or Lyft fares). We provide on the following page recommendations for lunch, sightseeing, and purchasing food and drinks for the picnic and show at Hollywood Bowl.

2. Late in the afternoon, take the Hollywood Bowl shuttle bus from Westwood (Veteran Ave at Wilshire Blvd) to Hollywood. This shuttle is $7 both ways, and the earliest departure is 5:30pm. You may request one shuttle ticket when RSVPing for the event. Participants who take the shuttle will arrive at the Bowl just in time for the show and will not arrive in time for the picnic. If you wish to take the bus, please let a volunteer know—we will walk you from campus to the bus stop. We recommend purchasing food and drinks in Westwood, before leaving, and bringing your personal picnic to the Hollywood Bowl as your dinner during the show.

Places to eat lunch.
We recommend visiting the following restaurants in smaller groups.

Musso & Frank Grill
6667 Hollywood Blvd, Los Angeles, CA 90028
(323) 467-7788
goo.gl/maps/Pi3Hxk3CLkp
Classic American fare at Hollywood's oldest eatery.

25 Degrees - Hollywood Roosevelt Hotel
7000 Hollywood Blvd, Los Angeles, CA 90028
(323) 785-7244
goo.gl/maps/fFvXWySCrDU2
Vintage pub with burgers, shakes, & glam decor.

Loteria Grill – Hollywood
6627 Hollywood Blvd, Los Angeles, CA 90028
(323) 465-2500
goo.gl/maps/5KAcuNKJ8V12
Creative takes on classic Mexican eats & a full bar.

Stout Burgers and Beer
1544 N Cahuenga Blvd, Los Angeles, CA 90028
(323) 469-3801
goo.gl/maps/xFTVNhbNae12
Casual gastropub with inventive burgers & craft beer.

Kino Sushi
6721 Hollywood Blvd, Los Angeles, CA 90028
(323) 465-4567
goo.gl/maps/UQHdRrJvjso
Popular Japanese, Korean & Chinese fare.

Pig 'N Whistle
6714 Hollywood Blvd, Los Angeles, CA 90028
(323) 463-0000
goo.gl/maps/FmWpPv3DUWH2
1920s pub serves up drinks & Continental fare.

Luv2eat Thai Bistro
6660 Sunset Blvd P, Los Angeles, CA 90028
(323) 498-5835
goo.gl/maps/KVzP2rLwLks
Classic Thai eatery for familiar favorites.

JINYA Ramen Express
6801 Hollywood Blvd. #317, Los Angeles, CA 90028
(323) 391-1916
goo.gl/maps/ZBjTbPzzjPT2
Counter service with ramen bowls & Japanese sides.

Umami Burger Hollywood
1520 N Cahuenga Blvd, Los Angeles, CA 90028
(323) 469-3100
goo.gl/maps/az53F9KLJtr
Lively chain serving elevated burgers & craft beer.

Urban Masala
6554 Hollywood Blvd, Los Angeles, CA 90028
(323) 957-9999
goo.gl/maps/ScbpHiZjhTw
Modern spot with quick-serve Indian eats.

Things to see.
Not sure where to go after lunch? Check out self-guided walking tours that are available on the internet:

Angeles Walk LA: Self-Guided Historic Trails
angelswalkla.org/walks_hollywood.html

Hollywood Cemetery Self-Guided Tour
tour.hollywoodcemetery.org

FieldTripper App
fieldtripper.com

The Real Los Angeles Self-Guided Tours
thereallosangelestours.com/tours/self-guided-tours

TimeOut: 22 must-see Hollywood attractions
timeout.com/los-angeles/things-to-do/what-to-see-in-hollywood

Fodor's Travel: Hollywood Sights
fodors.com/world/north-america/usa/california/los-angeles/things-to-do/sights/hollywood

Where to buy food for the picnic and concert.
Join us for a picnic at 6:00pm—bring your own food and drink—and walk with us to the bowl at 7:30pm. In addition, the evening's Bowl performance allows guests to bring their own food and drink. We suggest stopping by the following stores to grab food, snacks, and drinks. Bring your leftovers to the show!

Trader Joe's
1600 Vine St, Los Angeles, CA 90028 goo.gl/maps/XHDg7sAs1gu

Yucca Supermarket
1815 N Cahuenga Blvd, Los Angeles, CA 90028 goo.gl/maps/ad4R3ayMbKo

How to get to the Hollywood Bowl.
We reserved a picnic area for CGSI participants to meet and hang out before heading up to the venue. Picnic Area 10 is reserved from 6:00pm until 7:30pm. From the picnic area, the bowl is a 0.5-mile uphill walk and takes approximately 15 minutes.

THE HOLLYWOOD BOWL CONCERT BEGINS AT 8:00PM

PICNIC AREA 10 RESERVED FROM 6:00PM TO 7:30PM

WALK THIS WAY UP HIGHLWAND AVE TO PICNIC AREA 10

02 Wednesday, July 12th – Picnic, Volleyball, & Soccer at Sunset Canyon

Starting at 5:00pm, we will have a picnic and volleyball and soccer tournaments at Sunset Canyon Recreation Center, Amphitheater Lawn, on the northwest corner of UCLA campus. We look forward to having some 'not so competitive' fun. Food—BBQ and vegetarian options—will be provided.

03 Thursday, July 13th – Pacific Coastal Path Bike Ride

Starting at 5:30pm, we will meet at a bike rental in Santa Monica and ride bicycles along the beach path. At 7:00pm we will attend a free concert next to the Santa Monica Pier. The concert is Marcia Griffiths who is a Jamaican singer known particularly for her smooth and captivating live performances! The concert is free, but you will have to pay for the bike rental and for transportation to the beach (which is inexpensive when sharing Uber or Lyft). You are free to bring guests.

04 Tuesday, July 11th and Thursday, July 13th – Morning Exercise

At 7:30am there will be a morning run, and at 8:00am we will hold a cool down session—led by world-renown scholar Dr. Sagi Snir. We invite you to join us for a leisurely and scenic jog around campus. Wellness and sport are key aspects of the LA experience! Meet in front of the UCLA Faculty Center.

05 Monday, July 10th and Friday, July 14th – Happy Hours

Starting at 5:00pm, we will hold happy hour at Wolfgang Puck, a bar on campus located just a few minutes' walk away from the Faculty Center. We strongly encourage all to attend these happy hours and to get to know your fellow course participants.

COMPUTATIONAL GENOMICS SUMMER INSTITUTE

LONG COURSE
JULY 16-26

Program Contents

WELCOME

The long course provides additional instruction to more senior trainees, including advanced graduate students and post-docs. This part of the program provides a research residence program at UCLA where trainees will interact with world-leading researchers in the field of computational genomics. The program combines structured training programs with flexible time in order to encourage interaction and collaboration with other participants and program faculty.

RESEARCH TALKS are 45-minute explorations of current problems and research in computational genomics by participating faculty.

TUTORIALS are interactive, guided 45-minute workshops that aim to help participants apply new techniques to research problems.

JOURNAL CLUBS are 45-minute sessions that give participants an opportunity to critically evaluate recent articles and stay up-to-date on relevant new research.

LOCATION
The 2017 Computational Genomics Summer Institute Long Course will be held in the First Floor Conference Room in the Gonda Building. The Gonda (Goldschmied) Neuroscience and Genetics Research Center, is a state-of-the-art facility that helps researchers explore the depths of science.

200 Medical Plaza Driveway, Los Angeles, CA 90095
goo.gl/maps/xNdoP8jK5Eq

PARKING
The closest lots to the Gonda Building are Parking Structure 8 and Parking Structure 9.

2017 CGSI PROGRAMS
Long Course Retreat: July 6 – 8
Short Course: July 10 – 14
Long Course: July 10 – 26

CGSI ORGANIZING COMMITTEE
Eleazar Eskin, UCLA, CGSI Director
Russel Caflisch, UCLA, IPAM Director
Francesca Chiaromonte, Pennsylvania State University
Eran Halperin, UCLA
David Koslicki, Oregon State University
John Novembre, University of Chicago
Ben Raphael, Princeton University

Visit our website for more about CGSI:
computationalgenomics.bioinformatics.ucla.edu

SCHEDULE

Sunday July 16

18:00 Long Course Kick-off Party in Santa Monica
Keep an eye out for an email announcing location.

Monday July 17

08:30 - 14:45 Long Course Day One
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 David Tse
Research Talk: Maximally correlation and principal component analysis

10:00 Coffee Break

10:30 Saharon Rosset
Stochastic process models for mutations, their estimation from data, and their uses

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

13:30 Kin Fai Au
Research Talk: Transcriptome analysis at the gene isoform level using hybrid sequencing

14:15 - 14:45 Coffee Break

Tuesday July 18

08:30 - 15:30 Long Course Day Two
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 Jo Hardin
Research Talk: Prediction intervals for random forests with applications to high throughput data

10:00 Coffee Break

10:30 Ilan Gronau
Inferring a complex network of interbreeding between modern and archaic humans

12:00 - 13:30 Lunch Break

13:30 Speaker TBA
Title TBA

14:15 - 14:45 Coffee Break

Wednesday July 19

08:30 - 14:45 Long Course Day Three
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 John Novembre
Tutorial: Computational tools for understanding population structure in genetic variation data

10:00 Coffee Break

10:30 Ben Raphael
Tutorial: Inferring tumor evolution

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

12:00 - 13:30 Joint Lunch with BIG Summer
Gonda First Floor Seminar Room

13:30 Or Zuk
Research Talk: Estimating gene-specific selection parameters from human variation data: new methods and applications

14:15 - 14:45 Coffee Break

18:00 Dinner 1
Kay 'n Dave's
9341 Culver Blvd, Culver City, CA 90232 goo.gl/maps/zGGxeo9NtK32

Thursday July 20

08:30 - 14:45 Long Course Day Four
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 Jae-Hoon Sul
Research Talk: Large-scale genetic studies of human complex traits

10:00 Coffee Break

10:30 Fabio Vandin
Tutorial: Computational discovery of significantly mutated genes and pathways in cancer

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

13:30 David Koslicki
Research Talk: Using the Earth-mover's Distance to compare microbial communities

14:15 - 14:45 Coffee Break

16:30 - 20:00 Picnic, Volleyball, & Soccer @ Will Rogers State Historical Park, Picnic Area
1501 Will Rogers State Park Rd, Pacific Palisades, CA 90272 goo.gl/forms/sS9Jym61QFFDYvAF3

Friday, July 21

08:30 - 14:45 Long Course Day Five
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 Vineet Bafna
Research Talk: Detecting the favored allele in an ongoing selective sweep

10:00 Coffee Break

10:30 Francesca Chiaromonte
Research Talk: Statistics for large, complex data and its role in "Omics" research

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

13:30 Sagiv Shifman
Research Talk: Genomics approaches to study neurodevelopmental disorders

14:15 - 14:45 Coffee Break

Saturday, July 22

13:00 Hiking in the Santa Monica Mountains
Los Leones Canyon Trailhead goo.gl/maps/Ci5pTypfJuw

Monday, July 24

08:30 - 15:15 Long Course Day Six
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 Bogdan Pasaniuc
Tutorial: Genetic correlations to gain insights into relations between traits

10:00 Coffee Break

10:30 Jessica (Jingyi) Li
Research Talk: Neyman-Pearson (NP) classification algorithms and NP Receiver Operating Characteristic (NP-ROC) curves

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

13:30 Fabio Vandin
Research Talk: Computational methods for survival analysis in genome-wide cancer studies

14:15 - 15:15 Special Coffee Break

18:00 Dinner 2
Location TBA

Tuesday, July 25

08:30 - 14:45 Long Course Day Seven
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 Sriram Sankararaman
Research Talk: Probabilistic PCA for large-scale genetic data

10:00 Coffee Break

10:30 Kirk Lohmueller
Research Talk: Variation in positive and negative selection across the Tree of Life

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

13:30 Fereydoun Hormozdiari
Research Talk: Discovery of genetic variants and modules in neurodevelopmental disorders

14:15 - 14:45 Coffee Break

Wednesday, July 26

08:30 - 14:45 Long Course Day Eight
Gonda First Floor Seminar Room

08:30 Breakfast

09:15 Noah Zaitlen
Tutorial: Covariate adjustment in genetics and genomics

10:00 Coffee Break

10:30 Barbara Engelhardt
Research Talk: Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology

11:15 Journal Club
Franklin D. Murphy Sculpture Garden, UCLA Campus goo.gl/maps/CHzL2fccnFL2

12:00 - 13:30 Lunch Break

13:30 Speaker TBA
Title TBA

14:15 - 14:45 Coffee Break

LOOKING FOR A FANTASTIC RESTAURANT RECOMMENDATION FOR DINNER ANYWHERE IN LOS ANGELES?

ASK THE LOCAL CGSI ORGANIZERS FOR ADVICE!

PRESENTATION TITLES AND PAPERS

This year's Long Course features twenty-two 45-minute research talks and tutorials presented by prominent scholars in the field of computational genomics. Each speaker compiled a list of relevant papers that provide a more in-depth exploration of their presentation material. Click to open a paper in your internet browser.

26 David Tse: Research Talk: Maximally correlation and principal component analysis

1. Rényi, A., 1959. On measures of dependence. Acta mathematica hungarica, 10(3-4), pp.441-451.
2. Feizi, S. and Tse, D., 2017. Maximally Correlated Principle Component Analysis. arXiv preprint arXiv:1702.05471.

27 Saharon Rosset: Research Talk: Quality preserving databases for statistically sound "big data" analysis on public databases

1. Rosset, S., Aharoni, E. and Neuvirth, H., 2014. Novel Statistical Tools for Management of Public Databases Facilitate Community-Wide Replicability and Control of False Discovery. Genetic epidemiology, 38(5), pp.477-481.

2. Aharoni, E. and Rosset, S., 2014. Generalized αinvesting: definitions, optimality results and application to public databases. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(4), pp.771-794.

3. Aharoni, E., Neuvirth, H. and Rosset, S., 2011. The quality preserving database: A computational framework for encouraging collaboration, enhancing power and controlling false discovery. IEEE/ACM transactions on computational biology and bioinformatics, 8(5), pp.1431-1437.

28 Kin Fai Au: Research Talk: Transcriptome analysis at the gene isoform level using hybrid sequencing

1. Au, K.F., Sebastiano, V., Afshar, P.T., Durruthy, J.D., Lee, L., Williams, B.A., van Bakel, H., Schadt, E.E., Reijo-Pera, R.A., Underwood, J.G. and Wong, W.H., 2013. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences, 110(50), pp.E4821-E4830.
2. Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.J., Buck, D. and Au, K.F., 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research, 6.

29 Jo Hardin: Research Talk: Prediction intervals for random forests with applications to high throughput data

1. Chen, X. and Ishwaran, H., 2012. Random forests for genomic data analysis. Genomics, 99(6), pp.323-329.
2. Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E. and Zhao, H., 2006. Pathway analysis using random forests classification and regression. Bioinformatics, 22(16), pp.2028-2036.
3. Zhang, J., Hadj-Moussa, H. and Storey, K.B., 2016. Current progress of high-throughput microRNA differential expression analysis and random forest gene selection for model and non-model systems: an R implementation. Journal of Integrative Bioinformatics, 13(5), p.306.

30 Ilan Gronau: Research Talk: Inferring a complex network of interbreeding between modern and archaic humans

1. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y. and Hansen, N.F., 2010. A draft sequence of the Neandertal genome. science, 328(5979), pp.710-722.
2. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L. and Maricic, T., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 468(7327), pp.1053-1060.
3. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K.,

De Filippo, C. and Sudmant, P.H., 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science, 338(6104), pp.222-226.
4. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., De Filippo, C. and Li, H., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature, 505(7481), pp.43-49.
5. Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de La Rasilla, M. and Rosas, A., 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature, 530(7591), pp.429-433.

31 John Novembre: Tutorial: Computational tools for understanding population structure in genetic variation data

1. Novembre, J. and Stephens, M., 2008. Interpreting principal component analyses of spatial population genetic variation. Nature genetics, 40(5), pp.646-649.
2. Alexander, D.H., Novembre, J. and Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), pp.1655-1664.
3. Yang, W.Y., Novembre, J., Eskin, E. and Halperin, E., 2012. A model-based approach for analysis of spatial structure in genetic data. Nature genetics, 44(6), pp.725-731.
4. Petkova, D., Novembre, J. and Stephens, M., 2015. Visualizing spatial population structure with estimated effective migration surfaces. Nature Publishing Group.
5. Novembre, J. and Peter, B.M., 2016. Recent advances in the study of fine-scale population structure in humans. Current Opinion in Genetics & Development, 41, pp.98-105.

32 Ben Raphael: Tutorial: Inferring tumor evolution

No papers assigned.

33 Or Zuk: Research Talk: Estimating gene-specific selection parameters from human variation data: New methods and applications

1. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., et al, 2014. Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences, 111(4), pp.E455-E464.

2. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E.,et al., 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), pp.285-291.

3. Zuk O. et al. “Estimating and testing for gene-specific and population-specific selection in humans”; in prep.

34 Jae-Hoon Sul: Research Talk: Large-scale genetic studies of human complex traits

1. Sul, J.H., Cade, B.E., Cho, M.H., Qiao, D., Silverman, E.K., Redline, S. and Sunyaev, S., 2016. Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees. The American Journal of Human Genetics, 99(4), pp.846-859.
2. Zhu, Y. and Xiong, M., 2012. Family-based association studies for next-generation sequencing. The American Journal of Human Genetics, 90(6), pp.1028-1045.
3. Schaid, D.J., McDonnell, S.K., Sinnwell, J.P. and Thibodeau, S.N., 2013. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genetic epidemiology, 37(5), pp.409-418.
4. Sul, J.H., Raj, T., de Jong, S., de Bakker, P.I., Raychaudhuri, S., Ophoff, R.A., Stranger, B.E., Eskin, E. and Han, B., 2015. Accurate and fast multiple-testing correction in eQTL studies. The American Journal of Human Genetics, 96(6), pp.857-868.

35 Fabio Vandin: Tutorial: Computational discovery of significantly mutated genes and pathways in cancer

1. Vandin, F., Upfal, E. and Raphael, B.J., 2011. Algorithms for detecting significantly mutated pathways in cancer. Journal of Computational Biology, 18(3), pp.507-522.
2. Raphael, B.J., Dobson, J.R., Oesper, L. and Vandin, F., 2014. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome medicine, 6(1), p.5.
3. Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M. and Lawrence, M.S., 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics, 47(2), pp.106-114.

4. Vandin, F., Papoutsaki, A., Raphael, B.J. and Upfal, E., 2015. Accurate computation of survival statistics in genome-wide studies. PLoS Comput Biol, 11(5), p.e1004071.
5. Hansen, T. and Vandin, F., 2016. Finding Mutated Subnetworks Associated with Survival in Cancer. arXiv preprint arXiv:1604.02467.

36 David Koslicki: Research Talk: Using the Earth-mover's Distance to compare microbial communities

1. McClelland, J. and Koslicki, D., 2016. EMDUnifrac: Exact linear time computation of the Unifrac metric and identification of differentially abundant organisms. arXiv preprint arXiv:1611.04634.
2. Mangul, S. and Koslicki, D., 2016. Reference-free comparison of microbial communities via de Bruijn graphs. bioRxiv, p.055020.
3. Evans, S.N. and Matsen, F.A., 2012. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(3), pp.569-592.

37 Vineet Bafna: Research Talk: Detecting the favored allele in an ongoing selective sweep

1. Ronen, R., Tesler, G., Akbari, A., Zakov, S., Rosenberg, N.A. and Bafna, V., 2015. Predicting carriers of ongoing selective sweeps without knowledge of the favored allele. PLoS Genet, 11(9), p.e1005527.

38 Francesca Chiaromonte: Research Talk: Statistics for large, complex data and its role in "Omics" research

1. Liu, Y., Chiaromonte, F. and Li, B., 2016. Structured Ordinary Least Squares: A Sufficient Dimension Reduction approach for regressions with partitioned predictors and heterogeneous units. Biometrics.
2. Guo, Z., Li, L., Lu, W. and Li, B., 2015. Groupwise dimension reduction via envelope method. Journal of the American Statistical Association, 110(512), pp.1515-1527.
3. Ma, Y. and Zhu, L., 2013. A review on dimension reduction. International Statistical Review, 81(1), pp.134-150.
4. Bertsimas, D., King, A. and Mazumder, R., 2016. Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2), pp.813-852.
5. Campos-Sánchez, R., Cremona, M.A., Pini, A., Chiaromonte, F. and Makova, K.D., 2016. Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. PLoS Comput Biol, 12(6), p.e1004956.
6. Cremona, M.A., Sangalli, L.M., Vantini, S., Dellino, G.I., Pelicci, P.G., Secchi, P. and Riva, L., 2015. Peak shape clustering reveals biological insights. BMC bioinformatics, 16(1), p.349.
7. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pp.267-288.
8. Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), pp.301-320.

39 Sagiv Shifman: Research Talk: Genomics approaches to study neurodevelopmental disorders

1. Ben-David, E. and Shifman, S., 2012. Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. PLoS Genet, 8(3), p.e1002556.
2. Ben-David, E. and Shifman, S., 2013. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. Molecular psychiatry, 18(10), p.1054.
3. Shohat, S., Ben-David, E. and Shifman, S., 2016. Varying intolerance of gene pathways to mutational classes explain genetic convergence across neuropsychiatric disorders. bioRxiv, p.054460.
4. Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M. and Carrera, N., 2014. De novo mutations in schizophrenia implicate synaptic networks. Nature, 506(7487), pp.179-184.
5. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S. and Geschwind, D.H., 2013. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell, 155(5), pp.1008-1021.
6. Zhang, B. and Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology, 4(1), p.1128.
7. Xu, X., Wells, A.B., O'Brien, D.R., Nehorai, A. and Dougherty, J.D., 2014. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. Journal of Neuroscience, 34(4), pp.1420-1431.

8. Study, D.D.D., 2017. Prevalence and architecture of de novo mutations in developmental disorders. Nature, 542(7642), pp.433-438.

40 Bogdan Pasaniuc: Tutorial: Genetic correlations to gain insights into relations between traits

1. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A. and Pasaniuc, B., 2017. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics, 100(3), pp.473-487.
2. Pasaniuc, B. and Price, A.L., 2016. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics.
3. Shi, H., Kichaev, G. and Pasaniuc, B., 2016. Contrasting the genetic architecture of 30 complex traits from summary association data. The American Journal of Human Genetics, 99(1), pp.139-153.
4. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., De Geus, E.J., Boomsma, D.I., Wright, F.A. and Sullivan, P.F., 2016. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics.

41 Jessica (Jingyi) Li: Research Talk: Neyman-Pearson (NP) classification algorithms and NP Receiver Operating Characteristic (NP-ROC) curves

1. Tong, X., Feng, Y. and Li, J.J., 2016. Neyman-Pearson (NP) classification algorithms and NP receiver operating characteristic (NP-ROC) curves. arXiv preprint arXiv:1608.03109.
2. Li, J.J. and Tong, X., 2016. Genomic Applications of the Neyman–Pearson Classification Paradigm. In Big Data Analytics in Genomics (pp. 145-167). Springer International Publishing.

42 Fabio Vandin: Research Talk: Computational methods for survival analysis in genome-wide cancer studies

1. Vandin, F., Upfal, E. and Raphael, B.J., 2011. Algorithms for detecting significantly mutated pathways in cancer. Journal of Computational Biology, 18(3), pp.507-522.
2. Raphael, B.J., Dobson, J.R., Oesper, L. and Vandin, F., 2014. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome medicine, 6(1), p.5.
3. Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M. and Lawrence, M.S., 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics, 47(2), pp.106-114.
4. Vandin, F., Papoutsaki, A., Raphael, B.J. and Upfal, E., 2015. Accurate computation of survival statistics in genome-wide studies. PLoS Comput Biol, 11(5), p.e1004071.
5. Hansen, T. and Vandin, F., 2016. Finding Mutated Subnetworks Associated with Survival in Cancer. arXiv preprint arXiv:1604.02467.

43 Sriram Sankararaman: Research Talk: Probabilistic PCA for large-scale genetic data

No papers assigned.

44 Kirk Lohmueller: Research Talk: Variation in positive and negative selection across the Tree of Life

1. Galtier, N., 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLoS Genet, 12(1), p.e1005774.
2. Kim, B.Y., Huber, C.D. and Lohmueller, K.E., 2016. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. bioRxiv, p.071431.

45 Fereydoun Hormozdiari: Research Talk: Discovery of genetic variants and modules in neurodevelopmental disorders

1. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D. and Turner, E.H., 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature, 485(7397), pp.246-250.
2. Parikshak, N.N., Gandal, M.J. and Geschwind, D.H., 2015. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nature Reviews Genetics, 16(8), pp.441-458.
3. Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A. and Zody, M.C., 2016. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. The American Journal of Human Genetics, 98(1), pp.58-74.
4. Hormozdiari, F., Penn, O., Borenstein, E. and Eichler, E.E., 2015. The discovery of integrated gene networks for autism and related disorders. Genome research, 25(1), pp.142-154.

5. Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M. and Vitkup, D., 2011. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron, 70(5), pp.898-907.
6. Linh and Hormozdiari Recomb 2017

46 Noah Zaitlen: Tutorial: Covariate adjustment in genetics and genomics

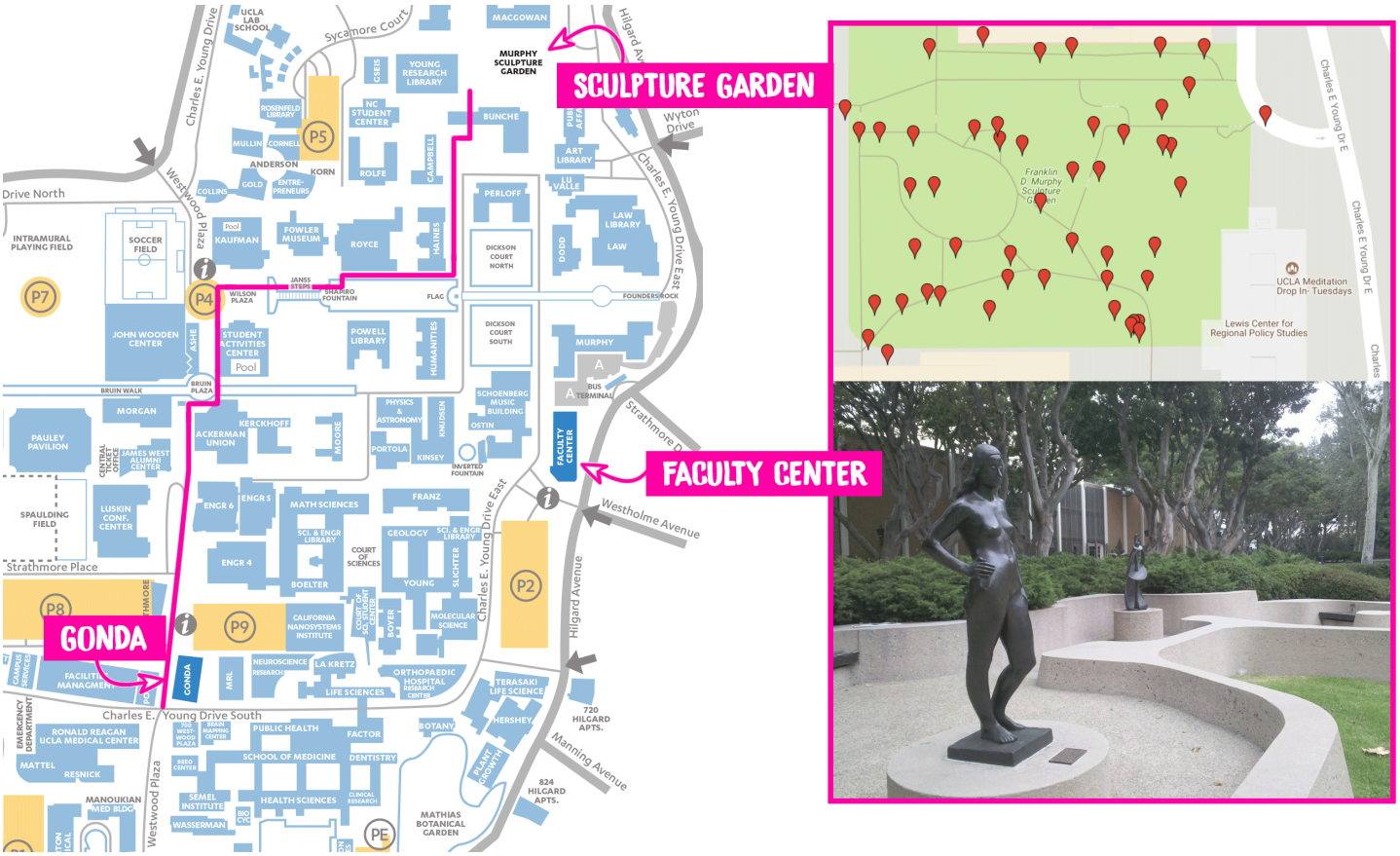
1. Zaitlen, N., Lindström, S., Pasaniuc, B., Cornelis, M., Genovese, G., Pollack, S., Barton, A., Bickeböller, H., Bowden, D.W., Eyre, S. and Freedman, B.I., 2012. Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet, 8(11), p.e1003032.
2. Aschard, H., Vilhjalmsson, B., Patel, C., Skurnik, D., Yu, J., Wolpin, B., Kraft, P. and Zaitlen, N., 2016. Playing Musical Chairs in Big Data to Reveal Variables Associations. bioRxiv, p.057190.
3. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L. and Kraft, P., 2015. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. The American Journal of Human Genetics, 96(2), pp.329-339.

47 Barbara Engelhardt: Research Talk: Transcriptional time series responses: Challenges, approaches, and opportunities

1. Heard, N.A., Holmes, C.C. and Stephens, D.A., 2006. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. Journal of the American Statistical Association, 101(473), pp.18-29.
2. Qin, Z.S., 2006. Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics, 22(16), pp.1988-1997.
3. McDowell, I.C., Manandhar, D., Vockley, C.M., Schmid, A., Reddy, T.E. and Engelhardt, B., 2017. Clustering gene expression time series data using an infinite Gaussian process mixture model. bioRxiv, p.131151.

JOURNAL CLUBS

As part of our Long Course, we will be breaking into five small groups for daily journal clubs. Journal clubs provide an excellent opportunity to discuss current work and doing so with colleagues with shared interests can be insightful, productive, and fun. Long Course journal clubs will take place **outdoors** in the Franklin D. Murphy Sculpture Garden, on UCLA Campus. **This area has plenty of shade trees, but you may wish to bring a hat and sunscreen.**



01 Microbiome analysis: Computational techniques and challenges
Hosted by Serghei Mangul
serghei@cs.ucla.edu
July 6 – 26

Technological advances and the decreasing costs of ‘next-generation’ sequencing (NGS) make it the technology of choice for many applications, including studying the human microbiome composed of bacterial, viral, fungi and other eukaryotic communities. Recently, high-throughput sequencing has revolutionized microbiome research by enabling the study of thousands of microbial genomes directly in their host environments. This approach, which forms the field of metagenomics, avoids the biases incurred with traditional culture-dependent analysis. The metagenomics approach also allows the comparison of microbial communities’ composition in their natural habitats across different human tissues and environmental settings. Specifically, metagenomic profiling is proven useful for analyzing microbes such as eukaryotic and viral

pathogens, which were previously impossible to study in an unbiased way with target 16S ribosomal RNA gene.

Tentative list of topics: We will be discussing recent methods to study microbial communities. We will be discussing the challenges in metagenomics analysis and limitation of the current methods. The goal will be to identify the best strategy to analyze metagenomics data.

We will start with discussion the most popular 'marker genes' methods, which are suggested to have poor sensitivity and also may result in false positives, detecting dangerous pathogens which are not present in the metagenomics sample. We also will discuss methods aimed to study microbiome at strain level and methods to study non-bacterial organisms, including viruses and fungi.

Outcomes: One possible outcome can be a joint effort to write an educational paper introducing metagenomics for researchers with no background in computational genomics or bioinformatics.

Difficulty: Intro/Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Escobar-Zepeda, A., de León, A.V.P. and Sanchez-Flores, A., 2015. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in Genetics, 6.

2. Simon, C. and Daniel, R., 2011. Metagenomic analyses: past and future trends. Applied and Environmental Microbiology, 77(4), pp.1153-1161.

3. Schmidt, C., 2017. Living in a microbial world. Nature Biotechnology, 35(5), p.401.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
4. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E. and Bremges, A., 2017. Critical Assessment of Metagenome Interpretation– a benchmark of computational metagenomics software. Biorxiv, p.099127.

5. Nayfach, S., Rodriguez-Mueller, B., Garud, N. and Pollard, K.S., 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research, 26(11), pp.1612-1625.

6. Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S. and Ahsanuddin, S., 2015. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Systems, 1(1), pp.72-87.

7. Huffnagle, G.B. and Noverr, M.C., 2013. The emerging world of the fungal microbiome. Trends in Microbiology, 21(7), pp.334-341.

02 Statistical methods to refine and redefine phenotypes
Hosted by Andy Dahl
andywdahl@gmail.com
July 6 – 26

With many multitrait datasets--like EHRs--the observed traits do not parsimoniously or precisely represent the underlying biology. For downstream analysis, the observed traits would ideally be summarized by a small number of latent and unknown traits that describe distinct and clear biological mechanisms. I hope to read papers on related dimensionality reduction problems, including both relevant methodological stats/ML papers and genetics papers using rigorous multitrait methods.

Difficulty: Advanced

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Van Der Maaten, L., Postma, E. and Van den Herik, J., 2009. Dimensionality reduction: a comparative. J Mach Learn Res, 10, pp.66-71.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
1. Lawrence, N., 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of machine learning research, 6(Nov), pp.1783-1816.

2. Cortes, A., Dendrou, C., Motyer, A., Jostins, L., Vukcevic, D., Dilthey, A., Donnelly, P., Leslie, S., Fugger, L. and McVean, G., 2017. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. bioRxiv, p.105122. PLUS SUPPLEMENT.

3. Joshi, S., Gunasekar, S., Sontag, D. and Joydeep, G., 2016, December. Identifiable Phenotyping using Constrained Non-Negative Matrix Factorization. In Machine Learning for Healthcare Conference (pp. 17-41).

03 Modern statistical methods with application to genomics
Hosted by Marzia Cremona
mac78@psu.edu
July 6 – 26

The goal of this journal club is to review and discuss modern statistical approaches that have been used in genomics research, or that can potentially be applied to analyze genomics data. We will discuss both theoretical aspects and genomics applications. Topics can include functional data analysis, variable selection and sufficient dimension reduction, inference methods, methods for big data, and will be chosen on the basis of participants' interest.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Wang, J.L., Chiou, J.M. and Müller, H.G., 2016. Functional data analysis. Annual Review of Statistics and Its Application, 3, pp.257-295.

Additional papers to potentially discuss:
2. Reimherr, M. and Nicolae, D., 2014. A functional data analysis approach for genetic association studies. The Annals of Applied Statistics, 8(1), pp.406-429.

3. Matsui, H. and Konishi, S., 2011. Variable selection for functional regression models via the L1 regularization. Computational Statistics & Data Analysis, 55(12), pp.3304-3310.

4. Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S. and Miyano, S., 2016. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. Biostatistics, 17(2), pp.235-248.

5. Taylor, S. and Pollard, K., 2009. Hypothesis tests for point-mass mixture data with application to ‘omics data with many zero values. Statistical Applications in Genetics and Molecular Biology, 8(8), pp. 1-43.

6. Nye, T.M., 2011. Principal components analysis in the space of phylogenetic trees. The Annals of Statistics, pp.2716-2739.

04 Outbreak detectives in the genomics era: Computational methods in molecular epidemiology
Hosted by Pavel Skums
pskums@gsu.edu
July 6 – 26

Molecular epidemiology is a new computationally-intensive discipline, which seek to allow to investigate disease outbreaks and track pathogen transmissions using viral genomic data sampled from infected individuals. In the recent years, computational genomics methods were successfully used for emerging diseases outbreaks (such as Ebola and Zika), as well as for the long-standing epidemics (such as HIV and HCV). The ultimate goal of computational molecular epidemiology is to develop methods allowing to reconstruct transmission histories and answer the question, who infected whom. This task is complicated by incomplete and noisy sequencing and epidemiological data, as well as by the extreme genetic heterogeneity of many viruses, which rapidly evolve within their hosts. We plan to discuss recent computational advances in the area, as well as pose and discuss open computational problems.

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Read introductions to papers 3 and 4 (and references therein). There are no review papers in this field yet.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
1. Campo, D.S., Xia, G.L., Dimitrova, Z., Lin, Y., Forbi, J.C., Ganova-Raeva, L., Punkova, L., Ramachandran, S., Thai, H., Skums, P. and Sims, S., 2015. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. The Journal of infectious diseases, 213(6), pp.957-965.

2. Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. and Ferguson, N., 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology, 10(1), p.e1003457.

3. De Maio, N., Wu, C.H. and Wilson, D.J., 2016. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS computational biology, 12(9), p.e1005130.

4. Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D. and Bunimovich, L., 2017. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics.

05 Introduction to single-cell genomics and new research directions towards a human cell atlas
Hosted by Vasilis Ntranos
ntranos@berkeley.edu
July 6 – 26

Our main goal in this journal club will be to familiarize ourselves with some of the key problem formulations in single-cell genomics and get exposed to the computational challenges emerging from the new types of data that are becoming available in this field [R1, R2]. After the initial overview [R3], participants will be free to choose specific papers/methods that best align with their interests and have them discussed in more detail by the group. Suggested topics include spatial reconstruction [P1], single-cell entropy in differentiation [P2] and lineage tracing by genome editing [P3]. Participants with diverse backgrounds are welcome, as we would like to engage in broad discussions about new research directions and potentially draw connections to existing methods and ideas from related fields such as phylogenetics and metagenomics.

Difficulty: Introductory/Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
R1. Yuan, G.C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S. and Quackenbush, J., 2017. Challenges and emerging directions in single-cell analysis. Genome Biology, 18(1), p.84.

R2. Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. and Clevers, H., 2017. The Human Cell Atlas. bioRxiv, p.121202.

R3. Wagner, A., Regev, A. and Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
P1. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nature biotechnology, 33(5), pp.495-502.

P2. Teschendorff, A.E. and Enver, T., 2017. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nature Communications, 8, p.15599.

P3. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F. and Shendure, J., 2016. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science, 353(6298), p.aaf7907.

• • •

08 New genomic data and methods for inferring human population history in Eurasia
Hosted by Chris Robles and Arun Durvasula
crrobles2grad@gmail.com and arun.durvasula@gmail.com
July 17 – 26

The past couple of years have produced an extreme wealth of genome sequence data that can be used to retell the story of human population dispersal out of Africa and into Eurasia. We will review some of the main sources of data that emerged from these studies (present-day and ancient DNA), as well as the statistical methods used to produce demographic models from these data. Some of the interesting questions addressed in these studies: how many waves of migration out of Africa can we trace in present-day Eurasian populations? How was Europe populated? How do present-day populations relate to early farmers of the Middle East?

Difficulty: Intermediate

Papers covering assumed knowledge (read or know in advance of the first journal club meeting):
1. Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. and Willerslev, E., 2017. Tracing the peopling of the world through genomics. Nature, 541(7637), pp.302-310.

Papers to discuss (read before the meeting when they are scheduled to be discussed):
2. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. and Skoglund, P., 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature, 538(7624), pp.201-206.

3. Pagani, L., Lawson, D.J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L. and Wall, J.D., 2016. Genomic analyses inform on migration events during the peopling of Eurasia. Nature, 538(7624), pp.238-242.

4. Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A. and Nickel, B., 2016. The genetic history of Ice Age Europe. Nature, 534(7606), pp.200-205.

5. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K. and Connell, S., 2016. Genomic insights into the origin of farming in the ancient Near East. Nature, 536(7617), pp.419-424.

6. Malaspinas, A.S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergstrom, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E. and Heupink, T.H., 2016. A genomic history of Aboriginal Australia. Nature, 538(7624), pp.207-207.

7. Lipson, M. and Reich, D., 2017. working model of the deep relationships of diverse modern human genetic lineages outside of Africa. Molecular Biology and Evolution, p.msw293.