<title> UCLA QCBio </title>
<body>

<h1> Bruins-in-Genomics </h1>
<p> COVID-19 Edition 2020 <p>

August 14th 1-3.30pm (Pacific)

B.I.G. Summer Research Symposium

1pm: Keynote: Victoria Sork
Dean of Life Sciences
"DNA sequences: from oak to the acorn genome"

1.15pm: 4 parallel Themes

Precision Medicine | **Computational Genetics** | **Epi-genomics** | **Systems Biology**

3pm: Awards Ceremony

UCLA Institute for Quantitative & Computational Biosciences

Precision Medicine – Bogdan Pasaniuc

1.15pm Session I – chaired by Jasmine Zhou
DEHOLLANDER, YANG (Zhou)
CONWAY (Fogel)
LEE, SHENOI (Boutros)
RANADE, CHEN (Pimentel)

1.45pm Session II – chaired by William Hsu
CHEN (Pasaniuc)
RAO (Pimentel)
CRISP, LADEROUTE (Boutros)
LAPINSKI (Pasaniuc)

2.15pm Session III – chaired by Bogdan Pasaniuc
DOCKSTADER, WU (Zhou)
LU, YANG (Tward)
SIU (Speier)
YUN (Hsu)

Computational Genetics – Sankararaman

1.15pm Session I – chaired by Janet Sinsheimer
CHEN (Garud)
SACHDEV (Lohmueller)
CHRISTIE (Sinsheimer/Papp)
HAN, RISSE-ADAMS (Sankararaman)

1.45pm Session II – chaired by S. Sankararaman
LIN (Sul)
DANIEL, GONZALEZ (Ophoff)
RODRIGUEZ (Sinsheimer/Papp)
GILLAM, GALLMEISTER (Flint/Eskin)

2.15pm Session III – chaired by Nandita Garud
CHENG/TANG (Sankararaman)
HWANG (Sul)
DELAO, SINGH (Sankararaman)
SMULLEN, ZHANG (Zaitlen)
DIEPPA, JACKSON (Garud)

Epigenomics – Matteo Pellegrini

1.15pm Session I – chaired by Chongyuan Luo
CHIDERAA (Luo)
ARDREN, YANG (Ernst)
CONCEPCION, PHILLIPS (Yang)
GALLASO, KALE (Ernst)

1.45pm Session II – chaired by Jason Ernst
HORSFALL, JACKSON, WALTERS (Arboleda)
KLEINSASSER, SUN (Pellegrini)
KOCH (Ernst)
LEE (Yang)

2.15pm Session III – chaired by Jingyi Jessica Li
MALEPATI (Li)
PFAHNL (Fan)
SINHA (Yang)
DERY (Pellegrini)

Systems Biology – Alexander Hoffmann

1.15pm: Session I – chaired by A. Hoffmann
KHAN (Butte)
FELT (Geschwind)
GUPTA (Geschwind)
SHEU (Graeber)
WAHLSTEN (Meyer)

1.45pm: Session II – chaired by Eric Deeds
JIANG (Hoffmann)
OKEKE, NGUYEN (Hoffmann)
TURNER (Alber)
KIM (Deeds)
SPIRO (Deeds)

2.15pm: Session III – chaired by Hilary Coller
EZE, PEREZ (Coller)
DICKSON (Park)
TYDINGS (Wollman)
LANDAU (Franco)
MATHUR (Savage)

1.15pm Session I – chaired by Jasmine Zhou

Tissue Phylogeny Reconstruction Based On DNA Methylation

ADAM DEHOLLANDER[1], EILEEN YANG[1], Ran Hu[2,3], Shuo Li[2,3], Xianghong Jasmine Zhou[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Pathology and Laboratory Medicine, UCLA
[3] Bioinformatics Interdepartmental Graduate Program, UCLA

DNA methylation is considered a key mechanism of tissue-specific transcriptional regulation. Although tissue-specific DNA methylation patterns exist in mammals, its role during tissue differentiation remains unknown. We examined DNA methylation data from thirteen tissue types to investigate methylation differences between tissues. We created phylogenetic trees to determine the relationships among tissues and identified differentially methylated regions (DMRs) unique to each tree branch. We discovered that tissues corresponding to the same germ layer clustered together in the phylogenetic tree. We then identified genes unique to the DMRs of each tree branch. By comparing heatmaps of methylation and corresponding gene expression in tissue-specific DMRs, we found that genes with differences in methylation patterns across tissues have corresponding differences in gene expression across tissues. Thus, DNA methylation-based tissue phylogeny and its associated DMRs can provide insight into the underlying mechanisms of tissue-specific gene expression and the role of DNA methylation in early development.

Using Transcriptional Profiling to Develop a Functional Assay for Amyotrophic Lateral Sclerosis, Type 4 (ALS4)

DANIEL CONWAY[1], Kathie Ngo[2,3], Brent Fogel[2,3,4]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics PhD Program, UCLA
Departments of [3] Neurology and [4] Human Genetics, David Geffen School of Medicine, UCLA

Amyotrophic Lateral Sclerosis, Type 4 (ALS4) is a rare dominant neurological disease due to gain-of-function mutations in the senataxin (*SETX*) gene and characterized by slow progressive motor neuron degeneration. Because rare private variants are often difficult to link to neurological diseases by sequence, we used transcriptional profiling to functionally identify patients with ALS4. Using weighted gene-correlation network analysis (WGCNA) on microarray data from two different ALS4 mouse models, we identified and characterized two disease-associated modules. Loss-of-function *SETX* mutations cause a distinct neurological disease, Ataxia with Oculomotor Apraxia, Type 2 (AOA2) but we observed that the ALS4 key modules did not overlap with the AOA2 key modules and were not associated with disease from AOA2 patient whole blood samples, confirming distinct disease-specific signatures. Whole blood RNA-sequencing data from ALS4 patients was compared with these key modules to test if this ALS4 transcriptional signature can be used to identify affected patients.

Down-sampling Effects on RNA Sequencing of Prostate Cancer

JOHN LEE[1,2,3,4], SAMUEL SHENOI[1,2,3,4], Julie Livingstone[2,3,4,5,6], Paul C. Boutros[2,3,4,5,6]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Human Genetics, University of California, Los Angeles
[3] Department of Urology, University of California, Los Angeles
[4] Institute for Precision Health, University of California, Los Angeles
[5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles
[6] Broad Stem Cell Research Center, University of California, Los Angeles

RNA-sequencing is used to help understand the state of a cancer. RNA is extracted from a population of cells and sequenced to identify transcripts and their abundances. Due to the tumoral heterogeneity of cancer, it is unclear how much sequencing must be performed to derive an accurate picture of the state of the transcriptome. We down-sampled a deeply sequenced set of prostate cancer tumors containing between 224.6 and 538.4 million reads/sample to four down-sampled percentages: 20%, 40%, 60% and 80%. This resulted in a minimum of 45.4 million reads/sample. The results of our analysis on the down-sampled dataset show that down-sampling maintains stable percentages of intragenic, intronic, and exonic reads across all down-sampled percentages. The results of this project will elucidate the relationship between sequencing depth and transcript detection, which can help in "forecasting" cancer progression using RNA-Seq and in optimizing studies to detect transcriptional products of subclonal mutations.

Exploring the Impact of Transcript Quantification on eQTL Analyses

ASHWIN RANADE[1], YIWEN CHEN[1], Harold Pimentel[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Departments of Computational Medicine and Human Genetics, UCLA
[3] David Geffen School of Medicine, UCLA

We aim to understand how transcript quantification and differential transcript usage affects expression quantitative trait loci (eQTL) analyses. It has been shown in small sample sizes that when there is differential transcript usage, differential gene expression estimates from naïve gene counts are very biased and expectation maximization-style transcript quantification techniques provide a gain in power. Since common eQTL pipelines use naïve gene counting when quantifying gene expression for eQTL, we aim to see if this bias is affecting eQTL analyses. In particular, we ran the two quantification methods (featureCounts and kallisto) on 87 Yoruba Lymphoblastoid cell lines. We then used QTLtools to discover eQTLs for each method, and observed how the results differed. We find overall much similarity, but a number of genes with very different effects resulting from inconsistencies in quantification. These results warrant further investigation on the differences between the two quantification techniques.

DISCUSSION

1.45pm Session II – chaired by William Hsu

ATLAS-hub: an R Shiny App for Phenome-wide Association Studies (PheWAS) results on the ATLAS BioBank

JESSIE CHEN [1], Ruth Johnson [2], Bogdan Pasaniuc [3,4,5,6]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computer Science, UCLA
[3] Department of Pathology and Laboratory Medicine, UCLA
[4] Department of Human Genetics, UCLA
[5] Department of Computational Medicine, UCLA
[6] Bioinformatics Interdepartmental Program, UCLA

Phenome-wide Association Studies (PheWAS) identify associations between a specific genetic variant and a wide range of phenotypes. However, most datasets with a wide variety of phenotypes currently lack representation of diverse populations. Due to the diversity of genetic ancestry in Los Angeles, UCLA's ATLAS Biobank has one of the largest proportions of non-European ancestry participants. With ATLAS-hub, we built a data visualization tool/web interface that displays PheWAS associations for 500K SNPs and approximately 1000 phenotypes. Phenotypes are structured into 'phecodes' (ICD-9/ICD-10 groupings of similar traits/diseases), providing associations for 4 major ancestry groups from the ATLAS Biobank: White/Caucasian, Black/African-American, Asian, Hispanic/Latino. The interface allows users to query associations on the SNP or gene level, particularly observing differences across populations for future implications in clinical assessment. ATLAS-hub can act as an additional resource to gain further insight into genetic variants for both researchers and physicians.

Quantifying Uncertainty in Heritability Estimation with Small Sample Sizes

JINGYOU RAO[1], Kathryn S. Burch[2], Harold Pimentel[3,4]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental Program, UCLA
[3] Department of Computational Medicine, David Geffen School of Medicine, UCLA
[4] Department of Human Genetics, David Geffen School of Medicine, UCLA

SNP-heritability is commonly used in genome-wide association studies (GWAS) to capture genetic architecture and quantifies the maximum possible accuracy of linear predictive models used in transcriptome-wide association studies. However, due to the small sample sizes of expression quantitative trait locus (eQTL) studies, GWAS heritability estimation tools suffer from lack of power resulting in large variance in the estimates. To understand the range of power and variance using GWAS heritability estimators in eQTL analyses, we built a gene expression model that simulates the isoform expression from real individual-level genetic data given the heritability and the isoform covariance matrix. Our simulations show that commonly used estimation methods have about 12.5% power for a gene with 10% heritability and 5% causal SNPs with 100 samples, thus indicating large opportunities for improvement with small sample sizes.

Cutpoint Optimization in Cox Proportional Hazards Modeling

ASHLYNN CRISP[1,2,3,4], MATTHEW LADEROUTE[1,2,3,4], Zhuyu Qiu[2,3,4], Paul Boutros[2,3,4,5,6,7,8]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Jonsson Comprehensive Cancer Center, UCLA
[3] Department of Human Genetics, UCLA
[4] Institute for Precision Health, UCLA
[5] Department of Urology, UCLA
[6] Broad Stem Cell Research Centre, UCLA
[7] Department of Medical Biophysics, University of Toronto
[8] Department of Pharmacology and Toxicology, University of Toronto

Cancer survival analyses commonly utilize Cox proportional hazards models with the parameters as exclusively continuous or discrete. Each of these approaches suggest a distinct biological mechanism through which the parameters impact the outcome for the patient. Using mRNA abundance data from 204 primary breast cancer tumor transcriptomes, we investigate how discretization methods affect gene significance in survival prediction. We found that over half the genes in our data set had differences in q-values greater than 0.1 when used as continuous vs. dichotomized parameters, indicating that discretization has a significant impact on survival prediction accuracy on a per gene basis. By finding how discretization methods affect gene significance, we can find characteristics of genes that are significant in all dichotomization approaches.

Uncertainty in Polygenic Risk Scores (PRS) and Its Implications for Clinical Use

SANDRA LAPINSKI, Yi Ding[2,3], Bogdan Pasaniuc[2,3,4,5]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental Program, UCLA
[3] Department of Pathology and Laboratory Medicine, UCLA
[4] Department of Human Genetics, UCLA
[5] Department of Computational Medicine, UCLA

Polygenic risk scores (PRS) predict an individual's genetic predisposition for disease by summing the effects of genetic variants across the human genome into a single score. When PRS is combined with lifestyle and clinical factors, it can help personalize preventative disease measures for patients. For example, it can stratify a population into high risk or low risk based on a certain threshold. However, current PRS methods report the point estimation of PRS without measures of uncertainty, which impacts its performance in clinical settings. Our approach for measuring uncertainty implements fine-mapping using a "Sum of Single Effects (SuSiE)" model to sample the posterior distribution of PRS, which will be used to construct 95% confidence intervals for PRS. By checking whether the PRS confidence interval overlaps with the diagnosis threshold, we can tell whether a patient has high uncertainty in diagnosis. The proportion of uncertain diagnosis varies with varying heritability. Based on our simulation, we found low patient proportions for patients in ambiguous low risk, ambiguous and unambiguous high risk categories where unambiguous refers to threshold overlap with confidence intervals. From these results, we can investigate the uncertainty of each patient and its implication for risk stratification.

2.15pm Session III – chaired by Bogdan Pasaniuc

Machine learning approach for cancer status prediction through fragment size analysis of tumor-derived cell-free DNA

JORDAN DOCKSTADER[1], JESSICA WU[1], Jim Liu[2], Mary Same[2], Jasmine Zhou[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA

Tumor-derived cell-free DNA (cfDNA) in human plasma opens up new avenues for non-invasive cancer diagnostics. cfDNA fragments are released into the bloodstream by apoptosis and generally have lengths consistent with the nucleosome-bound DNA released during this cellular process. However, past studies have reported aberrantly long and short lengths in cfDNA fragments derived from tumor tissues of cancer patients. Here, we expand this size analysis by exploring its cancer status prediction potential. Using a public dataset of cfDNA samples, we were able to perform numerous classification algorithms on cfDNA fragment length profiles to distinguish cancer and non-cancer samples. We also generated and utilized fragment length profiles from specific regions of the genome to uncover the relationship between fragment length and mapping position. Our study demonstrates how cfDNA size profiling shows promise in revolutionizing cancer diagnosis and monitoring through liquid biopsy.

Computational Algorithms for Revealing Microstructure in Brain Images with Deformable Registration and Deep Scattering Networks

JAQUELINE LU[1], ZI XI (OPHELIA) YANG[1], Daniel Tward[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computational Medicine, UCLA
[3] Brain Mapping Center, Department of Neurology, UCLA

We aim to quantify patterns of cell distribution in the brain, by building brain atlases from multiple neuroimages. Because the brain contains information at multiple spatial scales, atlases require alignment of high resolution data using deformable image registration. This calls for downsampling techniques that preserve information while decreasing image size for faster computations. Using novel methods based on the scattering transform, we extracted information from microstructures to produce low resolution images with high feature counts at each voxel. We examined how our downsampling method preserves information by predicting anatomical structures at each location using machine learning algorithms (LDA and random forests). Aligning these images requires a new approach to cross-modality image registration. We developed a method for working with this data, and also tested its performance on single-modality benchmark datasets. These techniques are being used to build better brain atlases, to study diseases and quantify variation in populations.

Estimating limbal stem cell densities in corneal tissue imaging in ImageJ

NATHAN SIU[1], William Speier[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Medical Informatics, Radiological Sciences, and Bioinformatics, UCLA

Limbal stem cell deficiency (LSCD) is a progressive corneal degenerative disease that renders the corneal epithelium unable to repair itself, which can lead to the eventual loss of vision. Although advances in technology have allowed for the growth of limbal stem cells ex-vivo for the purposes of transplantation, the current quantification methods used for quality control require ophthalmologists to manually count cells and calculate densities such that inter-observer error is unavoidable. In order to simplify the existing workflow, a plugin for the image processing software ImageJ was created. The plugin analyzes user-selected regions of interest, applies a color-thresholding method to predict cell centers, and provides a density calculation. Integrating these aspects into a user-friendly interface streamlines workflows, save time, and generates accurate, reproducible results.

Combining radiologist-interpreted and quantitative imaging features to classify pulmonary nodules as adenocarcinoma

MYOUNGJUN YUN[1], Anil Yadav[2,3], William Hsu[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Bioengineering, UCLA
[3] Medical & Imaging Informatics Group, Department of Radiological Sciences, UCLA

Lung cancer is the most common cause of cancer-related deaths in the United States. Lung cancer screening via computer tomography (CT) has been shown to reduce mortality, yet challenges remain including high false-positive rates, which result in costly biopsy procedures. Prior studies in this area have focused on the detection and classification of nodules using a limited number of clinical and imaging features. In this study, we attempt to fill a current gap in literature about the relationship between radiologist-interpreted semantic features and image-derived quantitative features in predicting adenocarcinoma. Our study examined 69 scans from patients (41 adenocarcinoma, 28 benign) seen at our institution. By interpreting both semantic features and feature extractions from the key slice of a patient's CT scan, we perform univariate and multivariable analysis to assess the relationship between individual and groups of features and adenocarcinoma. Our analysis can inform the design of future classification networks and, with further validation from external datasets, can help radiologists combine semantic and quantitative features to determine appropriate management of patients with indeterminate pulmonary nodules.

1.15pm Session I – chaired by Janet Sinsheimer

Species diversity begets genetic diversity in the human gut microbiome

DAISY CHEN[1], Naïma Madi[2], B. Jesse Shapiro[2], Nandita Garud[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Departement de sciences biologiques, Universite de Montreal, Canada
[3] Department of Ecology and Evolutionary Biology, UCLA

The effect of existing biodiversity on further diversification is a long-standing question of particular interest for complex, ubiquitous microbial communities. One hypothesis predicts that "diversity begets diversity" (DBD) via processes such as competition and niche construction. While previous work shows evidence of DBD in microbiomes through taxonomic ratios, it has yet to be tested using direct signatures of evolution, which can occur over short timescales in human gut microbiota. Here, we investigate the relationship between species diversity and evolutionary change in time-series metagenomic data from fecal samples of 249 healthy human adults. We observe that within-sample polymorphism positively correlates with species diversity, reflecting greater persistence of genetic variants. Inferring temporal changes in dominant lineages, we find higher numbers of SNP modifications in initially diverse communities, suggesting that DBD promotes faster adaptation rates across species. Our study poses new questions about mechanisms and health consequences of DBD in the human gut.

The Influence of Dominance of Deleterious Mutations in Detecting Archaic Introgressed Ancestry in Modern Humans

NINA SACHDEV[1], Xinjun Zhang[2], Kirk Lohmueller[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Ecology and Evolutionary Biology, UCLA
[3] Department of Human Genetics, David Geffen School of Medicine, UCLA

There is widespread evidence of archaic introgressed ancestry in modern human populations. Previous research suggests an influence of recessive deleterious mutations in detecting admixture levels in several regions of the human genome. However, given what is known about different genetic parameters in a realistic human demography, it is unclear how the dominance of deleterious mutations in an archaic population affects such levels of introgressed ancestry. Using the SLiM framework and Python, we created a pipeline that simulated admixture between Neanderthals and humans on different genomic regions as a function of dominance. We verified a previous study's observations, which showed elevated levels of introgressed ancestry in regions with high exon density and low recombination rate, as illustrated by the *HYAL2* gene. However, we did not observe this pattern in regions without similar genetic properties. Our work confirms deleterious variation as a variable that impacts observed levels of admixed ancestry in various regions of the genome.

Inheritance of Methylation in Grey Wolves from Yellowstone National Park

ROWAN CHRISTIE[1], Janet Sinsheimer[1,2,3,4], Jeanette C Papp[1,2], Bridgett M. vonHoldt[4], Chris German[2], Juyhun Kim[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[1] Department of Human Genetics, UCLA
[2] Department of Biostatistics, UCLA
[3] Department of Computational Medicine, UCLA
[4] Ecology & Evolutionary Biology, Princeton University, NJ

Our objective was to understand inheritance of methylation fraction in wolves by mapping their genes. To do this, we look at field observations, methylation data, and pedigree information from over 500 wolves in Yellowstone. We ran summary statistics on methylation data to determine variance at each site to determine which individuals had very little variation. We utilized OpenMendel software to perform GWAS and calculate theoretical kinship coefficients from pedigree data, which enabled us to find suspect pedigree structures. So far, our results show that there is little variance in methylation beta values across all individual wolves.

Assessment of power of principal component-based statistics to detect positive selection

ESTELLE HAN[1,2*], OONA RISSE-ADAMS[1,3*], Alec M. Chiu[4], Sriram Sankararaman[4,5,6,7]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Center for Computational Molecular Biology, Brown University
[3] Department of Mathematics, UC Berkeley
[4] Bioinformatics Interdepartmental Program, UCLA
[5] Department of Computer Science, UCLA
[6] Department of Human Genetics, UCLA
[7] Department of Computational Medicine, David Geffen School of Medicine, UCLA
* Denotes equal contribution

Discovering genetic variants with unusual differentiation between populations is a widely used approach for identifying putative signals of natural selection. Traditional set-based methods require discrete assignment of populations, neglecting phenomena such as admixture. Consequently, several statistics based on principal component analysis (PCA) have been developed as an alternative method to identify signals of natural selection that address such shortcomings. However, many PCA-based statistics are understudied in their sensitivity and power to detect variants under various models of natural selection. We assess three previously proposed PC-based selection statistics using data simulated under common models of selection designed to evaluate the qualities and characteristics of these statistics. We ultimately find that PCA-based statistics are generally underpowered, revealing a need for further developments in statistical methods to detect putative signals of selection.

DISCUSSION

1.45pm Session II – chaired by Sriram Sankararaman

Analysis of Risk Factor Genes for Congenital Heart Disease

DARREN LIN[1], Jae Hoon Sul[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Psychiatry and Biobehavioral Sciences, UCLA

Congenital heart disease is a disease characterized by abnormalities in the structure of the heart. It is one of the leading causes of infant mortality and occurs in around 1% of live births. Those who grow up with CHD tend to have other health complications in their adulthood, including heart failure and neurodevelopmental problems. We analyzed data from whole-genome sequencing of 711 trios (711 CHD children and 1422 parents) to better understand CHD risk factor genes. We focused on de novo single nucleotide mutations in coding regions of the genome found using the program Triodenovo and filtered by ABhet and ABhom values. We compared these mutations to interest areas suspected to be involved in CHD and its complications. This analysis reveals possible risk factor genes for CHD, such as NOTCH1 and CHD7, and supports preexisting research on the subject. Future CHD research can continue to focus on these gene areas.

Identification of Cerebrospinal Fluid Metabolites linked to Brain Disorders through Genetic Imputation

NOAH DANIEL[1], LIZBETH GONZALEZ[1], Toni Boltz[2], Roel Ophoff[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Human Genetics, UCLA
[3] Center for Neurobehavioral Genetics, UCLA

Brain disorders are obscured by a deficit in specific biomarkers to facilitate diagnosis and treatment. We hypothesized that metabolite data from cerebrospinal fluid (CSF) provides insight into brain health and disease. We previously gathered CSF from 500 healthy individuals and collected metabolomic and genotype data. We quantified levels of 11,000 metabolites of which 600 yielded significant genome-wide association (GWAS) results. For gene expression, it has been shown that imputation of its genetic regulation into disease GWAS results can be used to identify genes involved in these disorders. We extended this approach to include CSF metabolites. Using the TWAS FUSION software (Nature Genetics 48, pp.245), we imputed CSF metabolite levels into GWAS results of ten brain disorders, including Alzheimer's disease, Parkinson's disease, schizophrenia, and bipolar disorder. We identified hundreds of CSF metabolites with nominally-significant evidence of involvement in a brain disorder, which may imply that these metabolites can aid as biomarkers for disease.

SNP Selection and Characterization for Odontotaenius disjunctus

KARINA RODRIGUEZ[1], Benjamin Chu[2], Jeanette C. Papp[3], Janet S. Sinsheimer[2,3,4], Alexander M. Waldrop[5], Maria C. Rivera[6]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computational Medicine, UCLA
[3] Department of Human Genetics, UCLA
[4] Department of Biostatistics, UCLA
[5] The Research Computing Division, RTI International, Research Triangle Park, NC
[6] Department of Biology, Virginia Commonwealth University, Richmond, VA

The genotypes of over 200 patent leather beetles (*O. disjunctus*) were obtained from double digest RADseq experiments. First, we tested for Hardy-Weinberg Equilibrium (HWE) on each of the over 1300 loci using the VCFTools.jl package. We determined that the underlying Pearson's chi-square test statistic, which is based on large sample theory, is not appropriate for our dataset. Therefore, we implemented a number of exact probability models. Using Fisher's exact test statistic, we discovered there is a suppression of heterozygotes in the beetle genotypes, suggesting inbreeding or population substructure. Since populations of these beetles were historically isolated by past glaciation, distinguishing their origins from east or west of the Appalachian Mountains may restore HWE for most observed loci. Our results suggest all downstream data analysis involving beetle genetics must correct for population substructure induced by geography, which can be achieved using a simple heuristic.

Leveraging meta-analysis and fine mapping to facilitate causal gene identification

ANNIKA GILLAM[1], ELIZABETH GALLMEISTER[1], Nathan LaPierre[2], Jonathan Flint[3], Eleazar Eskin[4,5,6]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental PhD Program, UCLA
[3] CNG, Semel Institute for Neuroscience and Human Behavior, UCLA
[4] Department of Computer Science, UCLA
[5] Department of Human Genetics, UCLA
[6] Department of Computational Medicine, UCLA

Mouse models are useful for identifying causal genes for complex traits due to the ability to perform gene knockout experiments, which CRISPR has recently made cost-efficient. However, running functional tests for every gene is infeasible, and mouse GWAS studies tend to have low power due to small sample size. Here, we combine meta-analysis of mouse GWAS studies, fine mapping, and existing information on gene expression levels in relevant brain tissues to prioritize genes for knockout-based tests of causality for five anxiety-related behavioral traits. We found that, while some genes were well-supported by existing literature to have anxiety-related behavioral implications, others were novel candidates. The candidate genes will be analyzed further by quantitative complementation to confirm their causal role. These results will help identify the most effective methods for determining causal genes for future studies, which is critical for assessing these methods in human populations where knockout experimentation cannot be performed.

DISCUSSION

2.15pm Session III – chaired by Nandita Garud

Genetic Similarity Models Trained on Individual Level Data Outperform Conventional Models Trained on GWAS Summary Statistics in Phenotype Prediction

MICHAEL CHENG[1], DAVID TANG[1], Robert Brown[2], Sriram Sankararaman[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computer Science, UCLA
[3] Department of Human Genetics, UCLA

Polygenic risk scores (PRS) are used to predict an individual's phenotype based on their genotype. Because individual level phenotype and genotype data are publicly unobtainable, PRSs tend to rely on GWAS summary statistics for model training. This results in large prediction biases for individuals with ancestries dissimilar to the training population in linkage disequilibrium (LD) structure. However, with the recent growth of biobanks that include phenotype and genotype data, it is now feasible to construct PRSs with genetic similarity methods that do not rely so heavily on population matching assumptions. In this work, we compare the prediction accuracy of a PRS trained with GWAS marginal effects against a PRS trained with a model of genetic similarity. We show that using genetic similarity to inform PRSs leads to a 127 percent increase in prediction R2 when testing in admixed individuals with a quantitative phenotype simulated at a heritability of 0.3.

Accurate and fast detection of copy number variants from whole-genome sequencing with deep learning

STEPHEN HWANG[1], Albert Lee[2], Jae Hoon Sul[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Human Genetics, David Geffen School of Medicine, UCLA
[3] Department of Psychiatry and Biobehavioral Sciences, UCLA

Copy number variation (CNV) detection in whole-genome sequencing data provides valuable insights into human diseases and complex traits. Existing structural variant (SV) callers have poor performance with CNV detection due to the nature of short-read sequencing. Thus, researchers have developed ensemble methods combining results from several SV callers, but these methods still yield unsatisfactory results with high computational costs. Here, we propose SV-Net, a novel approach to CNV detection using a six-layer convolutional neural network (CNN) trained on reference mapped reads encoding base type, coverage, and read quality into RGB image color channels. SV-Net achieves an F1-score of 0.81 across insertions, deletions, and false-positives on the GIAB HG002 dataset, comparable to top ensemble methods featuring several SV callers. Future work involves further improvement of CNN accuracy and completing an efficient and streamlined pipeline from sequence alignment file to VCF file.

Predicting phenotype and identifying causal loci from simulated genotype data using machine learning models and SHAP

KEVIN DELAO[1], MAYA SINGH[1], Boyang Fu[3], Nandita Garud[2], Sriram Sankararaman[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Ecology and Evolutionary Biology, UCLA
[3] Department of Computer Science, UCLA

Comprehending the hidden structure between genotypes and phenotypic traits is a challenging problem in many fields. Part of the challenge is due to the number of biologically confounding factors in determining causal loci. We attempt to solve the problem of loci identification by using the SHAP (SHapley Additive exPlanations) feature interpreter on machine learning models run on simulated data with single causal loci, multiple causal loci, and multiple causal loci with interactions. The accuracy of SHAP in determining the causal loci is tested over multiple simulated trials with Linear Regression, Random Forest Regression, and Neural Network models. Varying biological factors in our simulations allows us to determine scenarios where SHAP is viable for causal loci identification. Applying feature interpretation with SHAP on machine learning models allows us to determine how the genetic information contained within genotypes can potentially be used to predict traits.

Revealing variation of predictive accuracy across quantiles and potential GxG or GxE interactions using quantile regression

MOLLY SMULLEN[1], FELIX ZHANG[1], Joel Mefford[3], Andrew Dahl[4], Nadav Rappoport[5], Noah Zaitlen[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computational Medicine, UCLA
[3] Department Neurology, UCLA
[4] Department of Medicine, U Chicago
[5] Department of Psychiatry and UCSF Weill Institute for Neurosciences, UCSF

Polygenic risk scores are an important method by which individuals can learn their risk of developing a disease, but questions persist about the accuracy PRS provides across quantiles of a phenotypic distribution. After simulating phenotypes generated by null and alternative genetic models, we use quantile regression to illustrate the variation in predictive accuracy PRS provides depending on the quantile of distribution and the presence or absence of gene-gene or gene-environment interactions. We develop a method using meta-regression to quantify instances of linear or non-linear variation across deciles due to GxG and GxE interactions, successful in detecting such variations while maintaining a low false positive rate. Our method illustrates that the covariance of quantile effect estimates must be considered when performing meta-regression for tests of homogeneity of effects, and that there are significant linear and quadratic variations on effect sizes for individual SNPs or PRS' due to GxG and GxE interactions.

Evaluating the predictive capability of gapped-kmers from microbiome data to improve phenotype prediction

ETAN DIEPPA[1] , SHAVONNA JACKSON[1] , Leah Briscoe[2] , Nandita Garud[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Ecology and Evolutionary Biology, UCLA

The human gut microbiome is a dynamic environment that plays important roles in an individual's well-being. Dysbiosis of the microbiome is associated with several diseases including Inflammatory Bowel Disease, Coronary Artery Disease, and Colorectal Cancer (CRC). Recently, studies have been able to predict disease from metagenomic data using k-mers, which are DNA substrings of length k. However, k-mers have inherent limitations, such as the lack of sequence coverage, which can be addressed by alternate forms of k-mers, called gapped kmers. In this study, we evaluate the accuracy of disease prediction with gapped k-mers compared to ungapped k-mers using CRC metagenomic data. To evaluate the effectiveness of our approach we implemented a machine learning classification algorithm (Random Forest). Our results reveal that certain gapped patterns are effective but fail to outperform ungapped k-mers. We conclude that the use of gapped k-mers is not as effective as ungapped k-mers for metagenomic analysis.

DISCUSSION

1.15pm Session I – chaired by Chongyuan Luo

Identifying cell-type specific chromatin interactions in human brain cell types

APAKAMA CHIDERAA[1], Chongyuan Luo[2]
[1] B.I.G Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Human Genetics, David Geffen School of Medicine, UCLA

Genomic function is regulated by an interplay between genome sequence and epigenomic modifications such as higher-order structure of chromatin in the nucleus. Characterizing patterns of epigenomic state has provided critical insights into the basic functional states of our genomes. An emerging challenge is characterizing features such as DNA methylation and 3D genome structure in complex mixtures of cells such as human tissue to study such features in their native in vivo setting. In this project, single-nucleus methyl-3C sequencing (snm3C-seq) was used to profile 3D genome structure and DNA methylation simultaneously in single cells obtained from postmortem prefrontal cortex tissue. By applying this method, we achieved loop level resolution of chromatin contacts, and identified cell type specific chromatin loops between adjacent brain regions. These results provide further insight into the 3D genome organization and the functional state of the human genome.

Extending the SHARPR-MPRA analysis pipeline with machine learning

MEGAN ARDREN[1], MUDI YANG[1], Tevfik Dincer[2,3], Jason Ernst[2,3,4]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental Program, UCLA
[3] Department of Biological Chemistry, David Geffen School of Medicine, UCLA
[4] Computer Science Department, UCLA

Analyzing regulatory regions of the noncoding genome has historically been low throughput and low resolution. SHARPR-MPRA employs a combined experimental and computational approach using Massively Parallel Reporter Assays (MPRAs) that allows for high throughput and high-resolution dissection of regulatory regions. Further work demonstrated the potential for machine learning models to uncover previously unknown DNA sequence patterns that have a marked effect on the transcriptional regulome. Here, we present a regression-based framework to extend the SHARPR model by allowing for incorporation of sequence features and MPRA tile features into the SHARPR probabilistic graphical analysis pipeline to improve identification of functional regulatory nucleotides. Our model will demonstrate if the incorporation of DNA sequence data and MPRA tile features into the SHARPR pipeline are able to obtain accurate, high-resolution information about activating and repressive nucleotides in a region.

Understanding the Interplay between RNA Binding Proteins and Repeat Elements

TORI CONCEPCION[1], BRIDGET PHILLIPS[1], Kofi Amoah[2], Xinshu (Grace) Xiao[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental Program, UCLA
[3] Department of Integrative Biology and Physiology, UCLA

Repetitive elements (RE) constitute a large fraction of the noncoding genome and interact with RNA binding proteins (RBPs). Many RBPs have been extensively studied using crosslinking and immunoprecipitation methods. Yet, the interplay between RBP and RE expression is not well-understood. Here, we investigate the effects of RBP levels on RE expression. Using RBP knockdown datasets from the K562 and HepG2 cell lines, we compared the expression levels of REs in knockdown samples versus controls. In K562, DDX47 regulated 972 REs, while in HepG2 TAF15 regulated 157 REs suggesting that DDX47 and TAF15 have the broadest impacts on RE expression. Also, the majority of the differentially expressed REs identified in K562 are from the MER and L1 families whereas those found in HepG2 are from the LTR and HERV families. Our analyses indicate that retrotransposons and endogenous retroviruses correlate with RBP expression and may explain how these factors regulate certain phenotypes.

Finding Phenotypic Similarities from GWAS data Across Species due to Biological Similarities of the Annotations Across Species

JOSEPH GALASSO[1], ANSHUL KALE[1], Jennifer Zou[2], Jason Ernst[2,3]
[1] Big Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computer Science, UCLA
[3] Department of Biological Chemistry, UCLA

GWAS studies using model organisms, such as mice and rats, are a useful way of studying the genetic basis of complex traits in a controlled environment, which often is not possible for humans. However, it is often unclear how well these results generalize to humans. The goal of this project is to compare GWAS studies in model organisms with GWAS studies in humans to assess how similar the GWAS variants are on a molecular level. To accomplish this, we utilized molecular data (ChIP-seq, chromatin states, DNase, RNA-seq, CAGE-seq) to compile 3113 mouse annotations and 8824 human annotations. We aligned these data sets and learned a shared embedding for these annotations using principal component analysis (PCA) followed by canonical correlation analysis (CCA). We then clustered the annotations in this embedding, computed enrichments of the GWAS variants in these clusters, and compared these enrichments across different mouse and human GWAS studies.

DISCUSSION

1.45pm Session II – chaired by Jason Ernst

Exploring the Presence of Genetic Compensation in KAT6A Syndrome

STEPHANIE HORSFALL[1], AMEENAH JACKSON[1], CYNNEY WALTERS[1], Leroy Bondhus[2], Angela Wei[3], Valerie Arboleda[2,3,4,5]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Human Genetics, UCLA
[3] Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA
[4] Department of Bioinformatics, UCLA
[5] Molecular Biology Institute, UCLA

Genetic compensation is a mechanism where genes of similar function to the mutated gene are expressed. *KAT6A* syndrome is a rare syndromic disorder characterized by intellectual disability, congenital heart defects, and distinctive facial features. The syndrome is caused by protein-truncating or missense mutations throughout the *KAT6A* gene. It has been observed that the location of the genetic mutation into the first half or second half of the gene is correlated with phenotypic severity. The underlying mechanism explaining the phenotypic variability in *KAT6A* syndrome remains unknown, which may be due to genetic compensation occurring. We processed RNA-seq data from *KAT6A* and *KAT6B* knockout cells and utilized differential expressed analyses to detect if genetic compensation was occurring. Furthermore, using BLAST we located genes homologous to *KAT6A* and tested for their enrichment in the differentially expressed genes. Our results showed genetic compensation was not observed in the *KAT6A* and *KAT6B* knockout HEK cells.

Identifying dependence of human cell-type composition on age and sex across human tissues

JANNA KLEINSASSER[1], ANTHONY SUN[1], Matteo Pellegrini[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Institute for Quantitative and Computational Biosciences, UCLA

GTEx is a large database that houses gene expression data from hundreds of individuals and from numerous tissues. Multiple tools are available that allow the decomposition of gene expression data into their constituent cell types. For example, the Gene Expression Deconvolution Tool, GEDIT, estimates cell type abundance from gene expression data. Using signature gene selection GEDIT takes input data and references a library of composition matrices to predict cell type abundances through row scaling and linear regression. Using this methodology, we asked whether the cell type abundance in human tissues depends on age and sex. We identified cell-type composition changes with age consistent with previous findings on immunosenescence, as well as novel correlations in monocytes, dendritic cells, and neutrophils with age. We also identified significant sex differences in abundance of neutrophils, dendritic cells, natural killer cells in blood, B cells, macrophages, monocytes, subcutaneous adipose, and CD8+ T cells.

A framework for identifying representative and differential chromatin state annotations within and across groups of samples

ZANE KOCH[1], Ha Vu[2,3], Petko Fiziev[9], Jason Ernst[2,3,4,5,6,7,8]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental Program, UCLA
[3] Department of Biological Chemistry, UCLA
[4] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA
[5] Computer Science Department, UCLA
[6] Jonsson Comprehensive Cancer Center, UCLA
[7] Molecular Biology Institute, UCLA
[8] Computational Medicine Department, UCLA
[9] Illumina Inc.

Sequencing technologies allowing for the examination of protein interactions with DNA have enabled the creation of genome-wide chromatin state maps. Given a group of biologically similar samples, it is often useful to have a chromatin-state annotation that is representative of the group. Here we introduce CSREP – a method that accepts a set of chromatin-state annotations from a group of samples and, using a logistic regression classifier, estimates the group's most representative chromatin-state annotation at the resolution of nucleosomes. Additionally, CSREP identifies differential chromatin regions between groups by comparing their representative chromatin-state maps. By applying CSREP to groups of reference genomes from the Roadmap Epigenomics Consortium, we demonstrate advantages of CSREP compared to a baseline method. Additionally, we identify biologically relevant epigenetic differences between male and female samples, as well as brain and embryonic stem cell samples, at a finer resolution than previous approaches.

Multi-omics Integration to Identify Network Perturbation of Glial Cells in Psychiatric Disorders

SANGWON (KARL) LEE[1], Yanning Zuo[2], Xia Yang[3]
[1] B.I.G. Summer Program, Institute of Quantitative and Computational Biosciences, UCLA
[2] Department of Biological Chemistry, UCLA
[3] Department of Integrative Biology and Physiology, UCLA

Millions of Americans suffer from mental illnesses, imposing a significant health burden costing 200 billion dollars annually. Despite the recent progress in psychiatric disorder genetics and transcriptomics, the pathogenesis mechanisms remain largely elusive. Here we elucidate disorder-related glial cell subtypes and key driver genes from frontal cortex and striatum by utilizing Mergeomics – a multi-omics pipeline integrating human genetics, functional genomics, and single cell transcriptomics – for common psychiatric disorders. We found that oligodendrocytes and their precursor cells are relevant for the pathogenesis of major depressive disorder, bipolar disorder, and autism spectrum disorder, suggesting myelination deficit as a potential pathogenesis factor. We predicted gene DHCR24 from frontal cortex oligodendrocytes to be a key driver for ASD, which was supported by an independent previous study that identified DHCR24 as a rare recessive mutation for ASD. Our study provides insights into the role of glial cells in psychiatric disorders and reveals potential therapeutic targets.

DISCUSSION

2.15pm Session III – chaired by Jingyi Jessica Li

A supervised ARI-based marker gene selection method for single-cell data

MANASVI MALEPATI[1], Ruochen Jiang[2], Jinfei Fang[3], Jingyi Jessica Li[2,4,5]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Statistics, University of California, Los Angeles, CA 90095-1554
[3] Department of Mathematics, University of California, Los Angeles, CA 90095-1555
[4] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[5] Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766

SCMarker is a gene-selection algorithm which uses the modality and expression levels of cells to identify and provide marker genes.Our main question was whether we could create an algorithm that utilizes the marker genes found by SCMarker and perform differential expression(DE) analysis using the Adjusted Rand Index(ARI) formula? Our first task was to reproduce the results of SCMarker by using R code to recreate the graphs and data analysis. The second task was to study the concepts behind the Adjusted Rand Index and formulate an algorithm which could cluster cells using SCMarker marker genes as part of our DE analysis. Through this algorithm, marker genes can be used for clustering in pilot studies with small data sets. In addition, our algorithm may be able to identify new, non-traditional marker genes which could be beneficial for cell clustering and gene-clustering in the future.

Determining Gene Expression Patterns between Human Retinal Cells and Mouse Retinal Cells

ARNOLD PFAHNL[1], Jing Wang[2], Guoping Fan[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Human Genetics, UCLA

The discovery of evolutionarily conserved and differentially expressed genes in various tissues has many important applications including the prediction of drug translation from animal to human models, and single-cell RNA sequencing (scRNA-seq) has been an instrumental tool in this process. Currently, there is very little understanding of the gene expression patterns between human retinal cells and mouse retinal cells. Here, we utilize scRNA-seq of mouse and human retinal cells as the basis for our analysis. We then perform statistical integration and clustering to find genes that are conserved and differentially expressed between the human and mouse retinal cells. We then examine the most statistically important genes and filter those that are most biologically significant.

Single-Cell Analysis of Astrocyte and Oligodendrocyte Subpopulations in Alzheimer's Disease

IRIKA SINHA [1,2], ALANNA STEWART [1,3], Jessica Ding [4], Xia Yang[1,4,5,6]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Biochemistry, University of Washington, Seattle
[3] Department of Biology, Spelman College
[4] Department of Integrative Biology and Physiology, University of California, Los Angeles
[5] Molecular Biology Institute, UCLA, Los Angeles, CA
[6] Brain Research Institute, UCLA, Los Angeles, CA

Alzheimer's Disease (AD) is a chronic neurodegenerative disease leading to cognitive decline and the leading cause of dementia. Glial cells have been increasingly recognized as important in AD pathogenesis, but astrocytes and oligodendrocytes are poorly investigated. In our study, we used single-cell RNA-sequencing on the hippocampus of the 5XFAD mouse model of AD to understand the roles of astrocytes and oligodendrocytes. We identified distinct astrocyte and oligodendrocyte subpopulations and found significant subtype specific transcriptional regulation induced by 5XFAD. Immune, complement, and cathepsin genes were found to be upregulated by 5XFAD in both astrocytes and oligodendrocytes. Lipid metabolism and oxidative phosphorylation were downregulated specifically in astrocytes, while hemostasis was downregulated in oligodendrocytes. Although the specific roles of these pathways require further confirmation through experimental testing, our findings provide insight into the roles of astrocytes and oligodendrocytes in AD.

DISCUSSION

Design of an Automated Program to Analyze Genomic Sequence Variants

ADAM DERY[1], Colin Farrell[2], Matteo Pellegrini[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Departement de sciences biologiques, Universite de Montreal, Canada
[3] Department of Ecology and Evolutionary Biology, UCLA

The development of methodologies to analyze annotated variants offers an approach to integrate varied data efficiently, using large-scale datasets. In this study, variant calls produced using high-throughput sequencers were converted to 23andMe formats so that we could obtain annotations for DNA markers in a genome. Freely available programs Clinvar, SNPedia, GEDmatch, and Python allowed automation of detailed reports. We further established a goal of using Excel to automate the transition process from the variant gene name to an individualized hyperlink function so that a variant annotation report could be generated without having to navigate through each variant one at a time. This function reduced the time to search through more than 600,000 variants to a smaller subset of 300 variants. Using these methodologies will help further advance our description of sequence variants.

1.15pm Session I – chaired by Alexander Hoffmann

Integrating Signaling and Polygenic Risk Scores to Predict Immune Dysregulation in Common Variable Immunodeficiency

HUMZA A. KHAN[1,2], Timothy J. Thauland[3], and Manish J. Butte[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Microbiology, Immunology, and Molecular Genetics, UCLA
[3] Division of Immunology, Allergy, and Rheumatology, Department of Pediatrics, UCLA

Common Variable Immunodeficiency (CVID) is a collection of monogenic disorders that are characterized by defective antibody production. Clinical presentation of these patients varies widely–from susceptibility to infection to autoimmunity to cancer. Patients with immune dysregulation require much more aggressive treatment, but we have no predictors of patient disease course. We employed mass cytometry (CyTOF), exome sequencing, and extensive clinical phenotyping to identify patients with immune dysregulation. By CyTOF, we found a number of aberrant signaling pathways in immune dysregulated patients, including a newly described defective T cell STAT3 signaling module. Other defects in the STAT3 and AKT signaling axes were found as well. We also utilized polygenic risk scores (PRS) to segregate CVID patient by clinical phenotype. A previously published PRS of absolute lymphocyte counts successfully distinguished autoimmune and non-autoimmune patients. Using this approach, we aim to combine genomics, phenomics, and phospho-CyTOF to classify patients and inform therapeutic interventions.

Accounting for genetic relationship in rare variant statistical testing

THOMAS FELT[1], Timothy Chang[2], Daniel Geschwind[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Neurology, David Geffen School of Medicine, UCLA

Progressive Supranuclear Palsy (PSP) is a rare neurodegenerative condition that has parkinsonian features and dementia. Association studies have identified common variants contributing to PSP risk, but common variation only accounts for part of disease heritability, indicating that rare variation likely contributes to the unexplained genetic heritability. However, traditional rare variant analyses have decreased power compared to common variant analyses due to the overall lower frequency of rare variants. Recent tests incorporate genetic relationship among samples to better estimate variance of rare variants. Here, we used whole genome sequencing data from 1668 PSP and 3272 control subjects, jointly-called, where we accounted for genetic relationships by removing 3rd degree or closer relatives. We then compared rare variant gene burden results in PSP accounting for (SKATO+GRM) and not accounting for (SKATO) the genetic relationship among samples. Although the correlations between SKATO+GRM and SKATO p-values were highly significant (Pearson cor=0.43, p<2.2E-16), they were lower than might be expected from the same cohort. This suggests that accounting for genetic relationship can have an impact on rare variant burden tests due to latent relatedness.

Cell type specific changes in transcriptional networks underlying ASD

SHRUTI GUPTA[1], Brie Wamsley[2], Daniel H Geschwind[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, UCLA
[3] Program in Neurobehavioral Genetics and Center for Autism Research and Treatment Semel Institute and Department of Human Genetics, David Geffen School of Medicine, UCLA

Advances in genomic technologies have played a major role in understanding the underlying cause of autism spectrum disorder (ASD). The cerebral cortex is built from a highly heterogenous group of cell types whose cooperative function underlies high-order cognitive functions commonly disrupted in ASD. Initial scRNAseq analysis has confirmed cell types that are most disrupted, but we lack an understanding of altered transcriptional networks across ASD cortical cells. We use SCENIC (single-cell regulatory network inference and clustering) on a large single-cell dataset (200,000 cells) composed of pre-frontal cortex from 10 unaffected individuals and 10 individuals diagnosed with ASD. SCENIC builds networks based on gene co-expression with transcription factors and their cis-regulatory elements found within each cell type. This analysis holds promise to extend our understanding of the molecular changes underlying ASD by unbiased linking of distinct transcriptional alterations to their genetic basis within specific cell types of the human cortex.

Genome-scale CRISPR-Cas9 Knockout Screen Identifies Genes Driving Chromosomal Instability in Cancer

KYLE M. SHEU[1,2], Kai Song[2], Nikolas G. Balanis[2], Daniel S. Yong[2], Xiangyu Ge[2], Thomas G. Graeber[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, UCLA

Chromosomal instability (CIN) is a hallmark of cancer and represents a state of high mutational frequency within the cell genome. This stochastic variance provides a Darwinian landscape through which cancer cell populations adopt characteristics favorable to drug-resistance, immune evasion, and metastasis; clinically, high CIN correlates with poor patient prognosis. However, the genes that drive CIN remain imprecisely determined. Here, we employ a forward genetic screen using genome-scale CRISPR-Cas9 knockout lentiviral barcoded libraries to identify these genetic determinants of CIN, with intent to identify potentially novel targets for molecular therapeutics. To assess gene contribution to CIN, we analyze sequencing data from CIN-high and CIN-low cell populations with STARS and correct false-positives from copy-number-amplified genomic regions with CERES. We then compare data from our screen to publicly available gene dependency screens to assess genetic perturbations and resultant CIN—as identified in our screen—and their potential mechanistic role in cancer cell fate.

Mapping a Melanoma Drug Resistance Program by Fitting a Data-Driven Dynamical Model

MADISON WAHLSTEN[1], Zhan Zhang[2], Farnaz Mohammadi[3], Aaron Meyer[3]
[1] BIG Summer Program, Institute of Quantitative and Computational Biosciences,
[2] Cross-disciplinary Scholars in Science and Technology, UCLA
[3] Department of Bioengineering, UCLA

Though effective therapies exist for melanoma, resistance to these drugs inevitably develops. Previous studies have shown that resistance arises from rare cancer cells that are reprogrammed from a pre-resistant state. Several genes, including EGFR, NGFR, and AXL, are disproportionately expressed in pre-resistant cells and have been comprehensively profiled through knockouts models and gene expression measurement. However, the broader regulatory events by which a cell enters this rare state are unclear. A unified model for how these components interact would help uncover drivers of this process. We built an ordinary differential equation model of the concentrations of mRNA corresponding to pre-resistant genes. We used this as a data-driven framework to identify gene-gene interactions by allowing all possible interactions, then comparing to gene expression measurements from each knockout using optimization implemented in Julia. The interaction parameters inferred by the model can be used to identify key regulators driving melanoma drug resistance development.

DISCUSSION

1.45pm Session II – chaired by Eric Deeds

Predicting LPA-induced gene expression dynamics in M1 and M2 macrophages with a multiple regression model of histone modifications

KEVIN JIANG[1], Katherine Sheu[2], Alexander Hoffmann[2]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Microbiology Immunology, and Molecular Genetics, UCLA

Macrophages encountering immune stimuli upregulate specific sets of genes, which may differ when macrophages are reprogrammed by polarizing cytokines into M1 and M2 states. Which genes are induced may be determined by histone modifications. However, the relationship between histone marks and stimulus-induced gene expression remains unclear. Here we aimed to study whether differences in histone modifications among macrophage states are predictive of peak fold-induction of LPA-induced genes. We processed ChIP-Seq data of four histone marks at baseline and RNA-Seq data for LPA-stimulated naive, M1, and M2 macrophages. We used a multiple regression model to correlate histone marks to either gene expression levels or peak fold-induction. We found that while histone modifications are predictive of baseline gene expression, they are poor predictors of peak gene induction upon stimulation. Our results emphasize the important role of signal-dependent transcription factors in the stimulus-response and suggest that their interaction with chromatin requires further study.

The transcription factor GAF induces gene expression in a cell-type-specific manner

CHIDERA OKEKE[1], MINH NGUYEN[1], Quen Cheng[2], Alexander Hoffmann[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Medicine, Infectious Diseases, UCLA
[3] Department of Microbiology, Immunology, and Molecular Genetics, UCLA

Interferon (IFN) cytokines are key signaling molecules of the immune system. IFNs activate two transcription factors (TFs), ISGF3 and GAF, in a coordinated manner to regulate interferon stimulated genes (ISGs). Previous work in epithelial cells demonstrated that GAF collaborates with ISGF3 to enhance ISG expression, but GAF binding alone is insufficient to induce expression of nearby genes. As GAF has been more extensively characterized in macrophages, we asked whether a similar phenomenon exists in this cell type. We examined ChIP-seq and RNA-seq data of IFN-stimulated macrophages and identified 2173 binding events, of which 1281 and 552 were classified as ISGF3 and GAF binding, respectively. ISGF3 and GAF binding events correlated with the induction of 19% and 22% of nearby genes, respectively. We conclude that GAF behaves differently in the two cell types. This specificity may be driven by collaborating TFs that are present in macrophages but not epithelial cells.

Prediction of Subnuclear Compartmentalization of Genomes Using 3D Structural Modeling and Machine Learning

SERAFINA TURNER[1], Asli Yildirim[2], Jitin Singla[2,3], Frank Alber[2]
[1] B.I.G Summer Program, Institute for Computational and Quantitative Biosciences, UCLA
[2] Department of Microbiology, Immunology, and Molecular Genetics, UCLA
[3] Department of Quantitative and Computational Biology, USC

The 3D structure of the genome plays an important role in various functions, including gene expression and replication. The genome is organized at different structural scales, with one layer of organization being its subnuclear compartmentalization: the composition of transcriptionally active A and inactive B subcompartments varies over cell types and governs chromatin co-segregation into functional microenvironments. Subcompartment detection requires high sequencing depths not available for all cell types. Here we combine 3D structural modeling and machine learning to identify subcompartments in different cell types. We use structural features extracted from 3D genome models generated from HiC data and compare the performance of unsupervised and supervised machine learning algorithms, such as k-means clustering, logistic regression and neural networks. We observed that logistic regression and neural networks achieved ~80% prediction accuracy. We aim to create a robust method to accurately predict subcompartments across cell types by including additional graph based structural features.

An accessible method for dynamical behavior analysis of large gene regulatory networks

SANDY KIM[1], Shamus Cooley[2,4], and Eric J. Deeds[3,4]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental Program, UCLA
[3] Department of Integrative Biology and Physiology, UCLA
[4] Institute for Quantitative and Computational Biosciences, UCLA

The temporal and spatial patterns of gene expression are fundamentally integral in all organisms. These patterns are governed by a set of genes and their interactions are known as a gene regulatory network (GRN) that underlie and influence many critical processes in the cell such as development, differentiation, and responses to environmental changes. Thus, dysregulation of GRNs is extremely detrimental and often leads cells to disease states or even senescence. The ubiquity and necessary role of GRNs across all organisms make them of great interest to study. However, due to their often large and highly complex nature, little is currently known about the dynamic properties of many GRNs. Although advances in high-throughput sequencing methods and their applications to temporal studies have produced time course data for gene expression, there remain many challenges in analyzing such high-dimensional data. Using principal component analysis, a dimensionality reduction technique, along with other mathematical methods, we developed an easily-accessible method to analyze the dynamics of high-dimensional time-course gene-expression data to infer the behavior and robustness of GRNs. We demonstrate its ability to uncover the dynamics underlying a wide variety of gene expression data by applying our tool to the analysis of simulations of very large GRNs.

Comparing the Levels of Distortion Introduced by Dimensionality Reduction Techniques

ANNA SPIRO[1], Shamus Cooley[2], Serena Hughes[2], Eric J Deeds[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Bioinformatics Interdepartmental PhD Program, UCLA
[3] Department of Integrative Biology and Physiology, UCLA

Existing techniques for reducing the dimensionality of high-dimensional datasets include linear approaches like Principal Component Analysis (PCA) and nonlinear approaches like t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). These techniques all introduce distortion, which we quantify using the Average Jaccard Distance (AJD), a measure of how much a lower-dimension embedding retains information about the local structure of the data. After developing the Deep Embedder (DE), a deep neural network approach to nonlinear dimensionality reduction, we compared the AJDs of the embeddings created by this technique to those created by PCA, t-SNE, and UMAP for nine machine learning datasets. We found that for sufficiently high embedding dimensions, the DE generally produces embeddings that are less distorted than t-SNE or UMAP embeddings and more distorted than PCA embeddings. We predict that adjustments to the DE algorithm will allow it to better approximate nonlinear manifolds than existing techniques.

DISCUSSION

2.15pm Session III – chaired by Hilary Coller

A Quantitative Approach to Study the Rates of Autophagosome Formation and Degradation for Fibroblasts in Cellular Quiescence

LAUREL EZE[1], IVAN A. PEREZ[1], Eric Deeds [2,3], Hilary Colller [2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Computational Biology, UCLA
[3] Department of Molecular, Cell and Developmental Biology, UCLA

The quiescent cell state is often overlooked due to its lack of genome replication and cell division. However, the transition to and from quiescence is a highly regulated process and is essential for cellular and pathophysiology. Previous studies suggest quiescence may be maintained by autophagy which recycles macromolecules into metabolites. Additionally, our lab has observed that transition from proliferation to quiescence is accompanied by an increase in autophagosomes. In order to understand the rates of autophagic flux in proliferating and quiescent cells, we used experimental data, statistical and dynamical modeling to solve an equation for proliferating, contact inhibited and serum starved fibroblasts. We determined that autophagosome formation is similar in both proliferative and quiescent cells but the rate of degradation decreases for cells in the quiescent state. These results suggest that decreased autophagosome degradation may play an important role in the viability or reversibility of quiescence.

Programmatic Identification of Information-Rich Tracers for Metabolic Flux Analysis

SEAN DICKSON[1], Jacob Prohroff[2], Keunseok Park[3], Aliya Lakhani[3], Junyoung Park[3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] CARE Program, Undergraduate Research Center (Sciences), UCLA
[3] Department of Chemical and Biomolecular Engineering, UCLA

Metabolic research is essential for understanding cell functions, identifying therapeutics for metabolic diseases, and engineering metabolism for biotechnological applications. Metabolic Flux Analysis (MFA) is currently the favored method for studying metabolism, relying on stable-isotope tracers and MS/NMR to gather information on metabolic networks. A major pitfall for MFA, however, is the exceedingly large number of tracers to choose from when designing MFA experiments. As a result, tracer selection to this point has been largely heuristic, making it nearly impossible to identify the most information rich tracers available for a given network. Here, we designed a tool that allows researchers to identify the most information rich tracers available for study of any metabolic network. We used the Elementary Metabolic Unit (EMU) model to simulate isotopologue distributions for all intermediates in a given network. Through these simulations, we determined information content of all possible EMU tracers and identified ones which gave us the most information about the network. Using this tool, metabolic researchers can avoid tedious trial-and-error tactics for finding information-rich tracers, effectively making MFA more efficient and fast-tracking metabolic research at large.

Deep Learning Predicts Early Apoptotic Commitment from Caspase 8 Activity

CLAIBORNE TYDINGS[1], Evan Maltz[2], Alon Oyler-Yaniv[2], Jennifer Oyler-Yaniv[2], Roy Wollman[2, 3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Institute for Quantitative and Computational Biosciences, UCLA
[3] Departments of Integrative Biology and Physiology and Chemistry and Biochemistry, UCLA

When cells are exposed to TNF, they build a death-inducing signaling complex, which includes caspase 8. Caspase 8 activation is a known commitment step for apoptosis. Is the apoptotic decision instantaneous or does it take into account accumulated information? A deep learning neural network, including an LSTM and an attention layer, was trained on caspase 8 activity collected from live-cell FRET reporter imaging, and used to predict cell death. Our neural network shows that early, pre-apoptotic caspase 8 information impacts the final decision of apoptotic commitment. As the point of apoptotic commitment is approached, the cumulative information in the caspase 8 activity better predicts apoptotic commitment. This work indicates that cellular decision making in the case of apoptosis is not an instantaneous decision, but depends upon accumulated cellular information.

Computational analysis and comparison of two recombinase-based oscillator designs with molecular sequestration

JUDITH LANDAU[1], Christian Cuba Samaniego[2], Elisa Franco[2,3]
[1] B.I.G. Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Mechanical and Aerospace Engineering, UCLA
[3] Department of Bioengineering, UCLA

Synthetic gene oscillators are canonical examples of dynamic biological circuits which allow autonomous cycling of cellular processes. Since serine recombinases can rearrange and then reverse-rearrange DNA when bound to their RDFs (recombination directionality factors), scientists have proposed their use in gene oscillators to invert promoters, an unconventionally dynamic application of recombinases. However, a recombinase-based oscillator has yet to be built so the optimal design is unknown. We used MATLAB to compare the dynamic models of two recombinase-based oscillator designs, each with an inverting promoter. Design 1 has one constitutive and one inverting promoter while design 2 is a novel oscillator with a single promoter. Our results unexpectedly showed that the more novel design 2 will likely perform better. This warrants thorough experimental testing of both designs. We are characterizing the conditions for oscillations and assessing tunability of period and amplitude to direct the selection of circuit components in experiments.

Allometric Scaling of Antibiotic Efficacy

SHAILI MATHUR[1], Portia M. Mira[2], Pamela J. Yeh[2,4], Christopher P. Kempes[4], Van M. Savage[2,3,4]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Ecology and Evolutionary Biology, UCLA
[3] Department of Biomathematics, David Geffen School of Medicine, UCLA
[4] Santa Fe Institute, Santa Fe, NM 87501, USA

How antibiotic efficacy varies with bacterial species is of basic and applied importance, including understanding of microbial dynamics in clinical and ecological contexts with possible consequences for the community structure of the microbiome. The scaling of cellular components in bacteria and their impact on metabolic, cellular, and evolutionary processes will help illuminate this question and possibly reveal an important role for cell size across bacterial species. Cellular components that antibiotics target—DNA, proteins, mRNA, tRNA, cellular envelope, and ribosomes—all scale non-linearly with cell volume. We model optimal strategies for cells to respond to antibiotics based on energetic constraints. We develop theory that shows how antibiotic efficacy may depend on cell size based on the specific cellular components targeted by the antibiotics and the nonlinearities between those components and cell size. Here, we present a general framework and detailed model for ribosome targeting antibiotics.

DISCUSSION


<title> UCLA QCBio </title>
<body>
<h1> Bruins-in-Genomics </h1>
<p> COVID-19 Edition 2020 <p>
UCLA Institute for Quantitative & Computational Biosciences