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