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 for 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.