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.