D2K Learning Lab Project – Baylor College of Medicine

Project Title: Inferring genomic signatures in age-related macular degeneration across different stages.

Project Pitch

Recent advancements in genomics (studying the structure and function of the genome) is transforming biomedical research into digitalized, data-intensive science that has invaded multiple aspects of biology and medicine. This project is aimed at utilizing data science and scientific computing to address challenges associated with big data in biomedical research. Specifically, we focus on uncovering the gene expression changes in common neurodegenerative disease, Age-related macular degeneration (AMD). AMD is a major cause of vision loss in the elderly that affects 11 million people in United States alone. It is a complex, multifactorial disease caused by the cumulative impact of multiple genetic variants, environmental stress, and advanced aging. Genome-wide association studies (GWAS) have identified 34 genetic loci (DNA variants) that account for ~65% of the total genetic contribution in AMD.

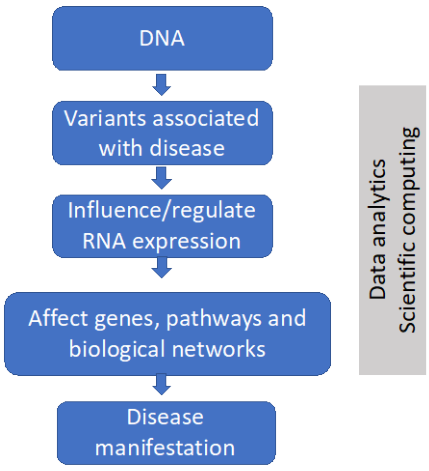

Figure 1: Transcriptome: Connecting the DNA variant to gene function and disease

Despite a strong genetic component underlying AMD, the progress in translating genetic findings into understanding disease mechanisms and harnessing new therapies has been slow. This is mainly attributed to the limited progress on translating DNA variants to functional understanding of the biology underlying disease risk. It has been hypothesized that disease-associated DNA variants are likely to influence the **disease risk through gene expression (RNA) regulation that in turn perturb biological pathways and networks (Figure 1)**. Thus, a comprehensive understanding of the global **transcriptome (collection of all RNA**) in disease-relevant cells and tissues represents the first logical step in functional understanding of AMD genetic findings. Towards this goal**,** we have generated large-scale transcriptome profiles of postmortem retinas (**the tissue that is affected in AMD**) from ~500 normal and cases at distinct stages of AMD**.** We built a comprehensive reference transcriptome of the human retina and expanded on understanding the role of genetic variants in regulating the gene expression. Here we propose to explore the transcriptome data for genes, pathways and biological networks in normal and disease contexts.

The broad, long-term goal of this project is to bridge the gap between genetic predisposition and biological mechanisms in AMD. Here we will implement advanced analytical methodologies to integrate large-scale, high-dimensional, genomic data. This line of investigation will generate analytical pipelines that will address the challenges of subtle gene expression changes associated

with complex diseases. These goals align with Baylor College of Medicine efforts of expediting the understanding of fundamental genetic and genomic principles through technical innovations.

Project Description

Our transcriptome data (17,389 genes) compromises of 105 normal, 175 early, 112 intermediate, and 61 advanced AMD donor retina as determined by the Minnesota Grading System (MGS) for their disease status. A preliminary search for genes that were altered in AMD yielded a few candidates owing to the inherently noisy and heterogeneous nature of RNA-seq data. However, gene expression measurements are highly correlated, which often reflects the activity of upstream biological pathways. Thus, we resorted to gene-set enrichment analysis (GSEA) that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. This analysis revealed multiple significantly enriched gene sets involved in metabolism, cell component organization, immune system, and stress response during disease progression. GSEA analysis also identified downregulated gene sets associated with synapse development and function that were predominant and largely exclusive to intermediate AMD. These findings suggest that intermediate AMD might have distinct molecular underpinning that does not represent a transitional stage between early and advanced AMD.

Based on these findings, we hypothesize that AMD progression from early to advanced stage may not be linear and involve both shared and stage-specific genes, pathways and cellular perturbations. Thus, we propose to apply computational methods such as soft clustering and unsupervised data structure deconvolution methods to elucidate the genes and pathways associated with distinct stages of the disease. We will then explore genes and pathways that are predictive of AMD stages. Perturbation in one gene or pathway can be propagated through the interactions, and affect other genes in the network. Thus, we will investigate the networks through which AMD pathways interact with each other and study the differences in networks associated with AMD stages. These approaches will contribute to a holistic understanding of complex diseases.

Project Objectives

- **Explore genes and genomic pathways associated with AMD stages.**
- **Discover genes and pathways predictive of distinct AMD stages.**
- **Explore genomic network signatures of AMD and specific changes in networks associated with the AMD stages.**

Data Confidentiality

1. Is your data confidential or does it have privacy and/or security requirements to be considered?
   No, the data is publicly available.

2. By what mechanism will you be sharing the data with Rice? (e.g. shared directly with the project group, kept on a sponsor-controlled server, or held on a preferred secure third-party platform)

   I will be sharing the data directly with project group.

Data Description

The data consist of post-mortem human donor eyes from 523 individuals that were procured by the Minnesota Lions Eye Bank. RNA was isolated from 50-100 mg of homogenized retina tissue. RNAseq libraries were prepared using TruSeq® Stranded mRNA Library Preparation Kit, and paired-end reads of 125 or 126 base pairs were obtained using the HiSeq 2500 platform (Illumina, San Diego, CA). The initial quality control process yielded 453 high-quality samples for gene expression analysis (105 MGS1, 175 MGS2, 112 MGS3, and 61 MGS4) that were included for all downstream analysis.

Primary analysis of the data has been already performed. Raw RNA-Seq reads were trimmed for Illumina adapters in Trimmomatic. Trimmed reads were aligned to the Ensembl release 85 (GRCh38.p7) human genome using STAR. Additional quality control metrics were calculated from Trimmomatic, FastQC and STAR using in-house Python and R scripts, including FASTQ and BAM file sizes, total number of reads, number of mapped and unmapped reads, and percentage of mapped reads. RNA-seq by expectation maximization (RSEM) was used to obtain estimated gene- and transcript expression levels. Normalization was performed using Trimmed Mean of M-values (TMM) in Counts per Million (CPM) using edgeR and then converted into $\log_2$ CPM with an offset of 1.

Student Teams

Experience with R and python. Previous experience on working with transcriptome or other genomic data is desirable but not necessary. This is also a research project and students interested in research and possibly working towards a publication are preferred.

Sponsor Information

I am an early-stage geneticist with a broad background and training in human genetics, genomics, and molecular biology. I am located at Baylor College of Medicine (BCM) in the Department of Ophthalmology at the Neurosensory Center. Advancement in next-generation sequencing technologies and their impact on human genetic research has largely shaped my research interests. My research efforts are aimed at harnessing the power of genomic technologies to address translational and clinical bioinformatics challenges pertaining to human health and diseases. I focus on retinal and macular degenerative diseases that display a spectrum of clinical phenotype ranging from rare monogenic to complex multifactorial form. I have attempted to understand the molecular genetic basis of both these extremes by studying DNA (whole exome sequencing) and transcriptome (RNA sequencing) to delineate the genes and pathways involved in disease causation. These approaches provide the path-forward for evidence-based practice to implement genomic findings in clinics as well as developing novel therapeutic strategies.

Sponsor Mentors

Rinki Ratnapriya
Instructor, Department of Ophthalmology
Baylor College of Medicine

Brief Bio

Identification of the causal genes/variants and understanding their roles in human diseases represent the central theme of my research. I have a broad background and training in human genetics, genomics, and molecular biology. During my Ph.D., I applied traditional linkage mapping in multigenerational families with epilepsies to identify genetic loci and causal genes. I did my postdoctoral training at the National Eye Institute (NEI), NIH, where I exploited next-generation sequencing (NGS) based, genome-wide methods for understanding the genetic basis of retinal and macular degenerative diseases. More recently, I have been focusing on age-related macular degeneration (AMD), which is a complex, multifactorial disease with 34 genetic loci identified through genome-wide association studies (GWAS). I utilized whole exome sequencing to expand the role of rare, coding variants in AMD. Additionally, I have aided in functional interpretation of AMD-associated loci by performing a large scale, integrative analyses of transcriptomes (RNA-seq) and their genetic variations from 453 individuals including both controls and cases at distinct stages of AMD. Integrative analysis of this data with AMD-GWAS has expanded the genetic landscape of AMD and created a useful resource for post-GWAS interpretation of multifactorial ocular traits.

I recently joined at Baylor College of Medicine where I am developing my independent research program that will focus on harnessing the power of genomic technologies to address translational and clinical bioinformatic challenges pertaining to AMD. These approaches are aimed at elucidating the disease circuitry by extracting valuable information from large genomic, transcriptomic and epigenetic data with an overreaching goal of identifying novel targets for therapeutic interventions. My training at NIH gave me ample opportunities to provide strategic, analytical and technical leadership across multiple projects. I developed strong teamwork skills through numerous national and international collaborators that culminated in publications of over three-dozen manuscripts. I am very excited to take on this new role as an independent investigator and have an opportunity to make significant contributions towards dissecting the mechanistic basis of AMD. Additional information regarding my research can be found at the link below:
https://www.bcm.edu/people/view/rinki-ratnapriya-ph-d/9b262067-17ec-11e9-b630-005056a012ee