Improved understanding of gene expression by mapping eQTLs to region-specific enhancers

Background

Psychiatric illnesses have been particularly difficult to understand because of the complexity of gene expression in the brain. However, new methods — particularly, analysis of tissue-specific gene expression — may provide novel insight into the etiology of psychiatric illnesses. My goal is to better understand how gene expression varies between different functional regions in the brain, and ultimately, to identify genetic factors contributing to these changes in expression levels.

One mechanism of interest is the potential variation in quantitative trait loci (QTLs) contained within transcription machinery binding sites, (e.g. enhancers and promoters). These locations are the sites of single nucleotide polymorphisms (SNPs), or mutations within the genome, and when they impede the binding of enzymes required to express the gene, they are labeled as expression quantitative trait loci (eQTLs). Probing the significance of these eQTLs has typically posed a challenge, but new methods for measurement and analysis of genomic expression provide mostly unexplored sources of multiple types of data. eQTLs have been previously examined in computational studies of numerous tissue types, ranging from blood cells to liver. This project will be part of a larger project aiming to integrate these diverse data to better understand gene expression in the brain. The overarching project includes the five following objectives:

1. Combining several large consortium data to analyze eQTLs in the brain.
2. Identifying genes with expression that is primarily confined to one brain region
3. Linking these genes to each other via common functional networks and pathways
4. Comparing the expression of these genes with the activity of neighboring promoters and enhancers throughout the genome to identify sites potentially involved in their expression
5. Mapping known eQTLs to enhancers or promoters associated with the brain region-specific genes
Last semester, I focused on objective #1 and an adjacent project to determine the optimal pipeline to carry out objective #5. This semester, my senior project will be focused on objective #5. I will draw from pre-existing methods for eQTL analysis to develop an optimal method to utilize the broad sets of available data for more statistically powerful identification of eQTLs involved in the expression of genes of interest.

**Data**

This project will be conducted as part of the work of the PsychENCODE consortium, a group of coordinated studies into the genetics of the brain, both healthy and unhealthy. [1] However, since this project requires tissue-specific data, we will be utilizing data from two other projects that focus on tissue-specific gene expression, namely GTEx and Brainspan. GTEx offers us genotype and expression data from different tissue types, including multiple specific regions of the brain, which will be used to accomplish objectives #1 and #2 (as listed above). [2] Brainspan generates RNA-seq and genotype of brain across various developmental stages (utilized in objectives #4 and #5). [3] Data points are labeled as fetal, infant, or adult and provide insight into the changes in gene expression at various time points in human development. This data will be used for objective #3.

**Previous work**

I completed the first half of this yearlong project last semester. I investigated some of the best practices for proceeding with the eQTL mapping portion of this project. First, I looked into the compatibility of different consortium data sets to determine if their formats were conducive to integration. RNA-seq data from GTEx and CommonMind have different formats; the former uses RPKMs (reads per kilobase of transcript per million mapped reads) while the latter uses read counts. However, even using read count-based software — like HTSeq — to quantify gene expression in GTEx data was not sufficient to fully reconcile the differences between the two approaches. Based on general opinion that read count and RPKM data should not be combined, as well as the general lack of consensus on the relative merits of the two systems, the conclusion was to analyze the data sets separately.
Then, I tested new genotype imputation methods to determine the optimal algorithm for making the most true positive eQTL identifications. In particular, I looked at the effectiveness of the recently released Haplotype Reference Consortium panel. [4] After imputing the genotype for one chromosome with the HRC panel, I compared the resulting SNP IDs with those obtained through imputation with the 1000 Genome reference panel. The HRC panel yielded more initial SNPs and fewer failures, but after rounds of stringent filtering, both panels offered the same number of SNPs. This matched what had been predicted by comparative studies. [4] However, since the 1000 Genomes panel had been used in the past and required much less CPU memory (0.13 GB vs 0.55 GB for the HRC panel), we determined that it would be appropriate to continue using the 1000 Genome panel.

Methods

Starting with the initial genotype and RNA-seq data, I will proceed through a proposed pipeline to impute the complete genotypes and link SNPs to changes in gene expression. The pipeline was developed through best practice analyses from last semester, detailed above.

Genotype imputation

The first step of eQTL analysis is genotype imputation. This is a common first step in many genome-wide analyses because it allows researchers to increase the statistical power of their studies. Genotype imputation is the process by which genetic data is inferred for sites that are not genotyped [4]. After assaying hundreds of thousands of single nucleotide variants (SNVs), combinations of SNVs shared between samples are used in conjunction with a reference genome to probabilistically determine whether another sample may have a given SNV, based on the rest of its haplotype. This project will utilize an algorithm that calculates the probabilities of potential genotypes by breaking the genome into smaller blocks to maximize the number of available reference haplotypes without needing to store and iterate through all the haplotypes in the entire genome. [5] The overall pipeline will have three main steps. First, the raw genotype calls will undergo quality control in PLINK to check for appropriate call rates, heterozygosity, relatedness, and proper gender reporting. Second, the genotype calls will be split by chromosome and each chromosome will undergo pre-phasing with Eagle2 in order to estimate haplotypes.
Finally, the actual imputation will be carried out using IMPUTE2, with imputed genotypes calculated against the Haplotype Reference Consortium panel (selected based on its performance on test data last semester). Code will be written in R and Python, and shell scripts will be run on the Yale High Performance Computing cluster. Statistical analysis will use models like Hidden Markov Model.

**Mapping eQTLs to enhancers**

Next, once genotype information is known across all SNV sites in the genome for all samples, I will calculate a set of known brain eQTLs and use the list of brain enhancers identified in objective #3 to form connections between eQTLs and the enhancers they disrupt or activate. eQTLs are defined for a given trait, and these eQTLs were identified by statistically testing the relationship between these genetic loci and psychiatric traits. [6] [7] Potential enhancers will be identified for the genes of interest, and eQTLs associated with transcription factor binding sites will be identified within this genomic region. The pipeline for this part of the project can be mostly completed with the Matrix eQTL package for R. It incorporates random sampling and ANOVA statistical tests will be used to determine whether the risk loci were significantly related to the up- or down-regulation of associated genes. The code for these tests will be written in R, Python and shell scripts, using models like Bayesian complexity control and joint modeling.

**Deliverables**

At the end of the semester, I will submit a paper outlining:

- The results of the experiment (mapping eQTLs to brain enhancers)
- A refined method/pipeline for integrating different brain genomic data (RNAseq expression, developmental data, known eQTLs)

I will also submit my code for the eQTL mapping.
References


3. [www.brainspan.org](http://www.brainspan.org)


