INVESTIGATING THE GENETIC BASIS OF GENE EXPRESSION USING EQTL TECHNIQUES

by

Andrew Quitadamo

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics

Charlotte

2018

Approved by:

Dr. Xinghua Shi

Dr. Cynthia Gibas

Dr. Jun-tao Guo

Dr. Bao-Hua Song

Dr. Way Sung

ABSTRACT

ANDREW QUITADAMO. Investigating the Genetic Basis of Gene Expression Using eQTL Techniques. (Under the direction of DR. XINGHUA SHI)

With advances in genome sequencing technology, datasets with large sample sizes can be generated relatively quickly and cheaply, especially compared to the past decade or so. We can utilize this data to analyze the associations between genetic variants and gene expression, and how that in turn relates to specific phenotypes. We will explore the impact of structural variants (SVs) on gene expression and microRNA expression in healthy individuals. This dissertation includes a description of a new eQTL analysis pipeline package, and an analysis of the impact of SVs on gene expression using eQTL techniques.

ACKNOWLEDGEMENTS

There are many people who deserve thanks for their contributions and help along the way. First my advisor Dr. Shi, who has given me guidance and support throughout graduate school. I'll be eternally grateful to her for allowing a first year PhD student to work on the 1000 Genomes project, for giving me room to grow and for listening to my ideas. Also, the other people in the Shi lab who have worked with me on many projects, and who have been part of many long and animated discussions.

I would also like to thank the members of my committee, Dr. Gibas, Dr. Guo, Dr. Sung and Dr. Song for the critiques, guidance, suggestions and of course their time. Their were professors at UNH, especially Dr. Kelly, who help start me on this path.

This work would not have been possible without funding from both the University of North Carolina at Charlotte, and a GAANN fellowship from the Department of Education.

On the more personal side, my parents provided lots of love, support and encouragement. Melissa, if anybody deserves more credit than me, it is you. You gave a tremendous support and understanding through this journey. You kept pushing me forward, and this would be half finished without you. Thank you for all your faith in me, even when I had none. And Emma, this is all for you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

While generating data for different aspects of cellular mechanics is relatively easy, the analysis and synthesis of this data is much more difficult. Many techniques and tools attempt to address these problems. In this dissertation we applied techniques from expression quantitative trait loci analysis to study how genetic variation impacts gene expression, and how that can translate into phenotypic differences. We used large, publicly available datasets and combined different data types (e.g genetic variation, gene expression, microRNA expression etc.) from the same individual to create a more holistic picture of genetic regulation.

This dissertation focuses on the impact of genetic variation and other cellular regulatory mechanisms. The first part of this dissertation is a description of an eQTL analysis software package built around MatrixEQTL. The second part uses the 1000 Genomes data to study the effects of SVs on gene expression and miRNA expression, as well as the effects of miRNA expression on gene expression.

1.1 Expression Quantitative Trait Loci

Expression quantitative trait loci (eQTL) analysis is a group of methods for analyzing the association between genotypes, or any quantitative trait, and gene expression [1]. eQTL analysis grew out of QTL analyses, which attempted to find links between sections of the genome (loci) and a trait, for example the number of bristles in

fruit flies [2]. eQTL studies could be compared to genome wide association (GWA) analysis, with gene expression values used as the factor of interest instead of a single trait. The introduction of high density genetic variation and gene expression microarrays allowed for the development and proliferation of eQTL analysis, and currently DNA and RNA sequencing are more commonly used. eQTL analysis was pioneered in yeast in 2001, has subsequently been applied to many other organisms, including humans starting in 2005 [3]. In humans the use of RNA-Seq data in eQTL analysis was introduced in 2010, and quickly replaced microarrays, as RNA-Seq can provide a wider range of data, such as allele specific expression [4]. eQTL analyses have been used to study both the function of normal and diseased cells, along with many other biological questions. There are an increasing number of studies that have generated eQTLs using a large sample size, including the Genome Tissue Expression Project (GTEx), and the Geuvadis project which used data from the 1000 Genomes Project. The GTEx group focuses on producing gene expression values, genotypes and eQTLs for multiple tissues from each individual in the study. Based on the most recent estimates a significant proportion (up to 80%) of genes are affected by an eQTL [5]. This estimate has increased since early human eQTL studies as methods and, more importantly, data have improved with the increasing sample size and the number of genotypes. Generally eQTLs are divided into two categories, *cis*- and *trans*-. In eQTL analyses the definition of *cis*- is usually based on a distance cutoff, i.e. within 1MB or 250kb, and *trans*- is anything that is not *cis*- [2]. This is slightly different than other definitions of *cis*- and *trans*- which can mean same vs. different chromosome. Most studies have focused mainly on *cis*-eQTLs due to the large multiple testing burden

for analyzing *trans*-eQTLs. It is easier to find eQTLs than to interpret the results and determine causality.

It has been shown that eQTLs overlap with GWAS SNPs, suggesting that eQTLs can help narrow down the genetic variant that cause the reproducible signals found in the GWA analysis [6]. Besides overlapping with GWAS SNPs, eQTLs are also enriched for regulatory features, including DNaseI hypersensitivity sites, open chromatin, and others [7]. Previous studies have repeatedly found that eQTLs cluster symmetrically around the transcription start site (TSS), and, to a lesser extent, asymmetrically around the transcription end site (TES) [8, 9]. These studies also indicate that most eQTLs occur within 100kb of the TSS, and this finding has been replicated many times. The GTEx analysis found that ∼80% of eQTLs are found in this area. eQTLs nearer the TSS appear to have stronger effects, and eQTLs shared among multiple populations occur closer to the TSS. There appears to be a high percentage of eQTLs that are shared among disparate populations, as well as a large percentage of eQTLs that are shared among different tissues [10, 9].

eQTL analysis has been extended to look at the impact of genetic variation on the protein expression. Protein QTLs (pQTLs) studies have also included eQTLs in the analyses to trace a genetic variant's effect from gene expression to protein expression. While there are genetic variants that impact both the gene and protein expression, many genes appear to have only an eQTL and not a pQTL [11, 12, 13, 14, 15]. This suggests that eQTL results do not translate directly to protein expression and then to a phenotype. Several studies have used samples from the 1000 Genomes Project to look at eQTLs and pQTLs, but these have focused on SNPs and indels and not larger

structural variants [16, 17]. Due to the realities of generating protein expression data the sample sizes tend to be smaller than projects that focus just on gene expression.

1.2 Structural Variants

Structural variants (SVs) are a group of genetic variation that includes large insertions, large deletions, mobile elements, duplications and others. There isn't an official size definition for a SV vs. an indel, but currently the most common definition is >50bp. This size cutoff is quite arbitrary, and has been reduced as the technology has improved (previously the consensus was >1000bp). While SVs occur less frequently than SNPs, they make up a larger proportion of the genetic changes in humans due to their size. Challenges in identifying SVs using short read technology means that they have been studied less than easier to identify types of genetic variations. An early analysis of the impact of copy number variation (CNVs), a type of SV, on gene expression showed that SNPs which tagged the CNVs were more likely to effect multiple genes compared to a null set [18]. Groups like the 1000 Genomes Project and GTEx have started to generate SV genotypes for a large number of individuals. With the data from these groups it is now possible to analyze the contribution SVs have on gene expression.

One of the first SV eQTL analysis was developed using CNV data from CGH arrays and gene expression arrays from four of the HapMap populations. They found 323-411 SNP eQTLs, and 44-96 CNV eQTLs depending on the population. In their analysis only about 20% of the CNV eQTLs could be captured using the SNP genotypes [19]. In 2011, Schlattl et al. conducted a SV eQTL analysis based on CNV

genotypes from both arrays and sequencing, and RNA sequencing in two of the 1000 Genomes populations. They were able to capture smaller CNVs than Stranger et al., as well as ascertain breakpoints for some of the CNVs. They found 50 and 73 CNV eQTLs for 110 unique genes, and a subset of these overlapped with the CNV eQTLs from Stranger et al., as well as prior studies that focused on a handful of deletions [20]. Using the breakpoints the authors were able to determine that in $\sim$20% of the eQTLs the CNV overlapped either part or the complete gene sequence. The 1000 Genome Project genotyped individuals from a number of populations from five continental groups [21]. The Structural Variation Analysis Group focused on genotyping structural variants in the 1000 Genomes Project data. The SVs include insertion-deletions, CNVs, nuclear mitochondrial insertions (NUMTs), and mobile element insertions (MEIs) which include Alu, L1 and SVA elements. The Geuvadis project performed RNA-Seq and miRNA-Seq on 465 lymphoblastoid cell lines (LCLs) from five 1000 Genomes populations. Four (CEPH/CEU, Finns/FIN, British/GBR, Toscani/TSI) are from the European super-population, and the last (Yoruba/YRI) is an African population. The Geuvadis paper included an eQTL analysis using SNPs and indels, and there was a separate SV eQTL analysis in "An integrated map of structural variation in 2,504 human genomes". While a SV eQTL analysis has already been done using this data, it was not the main focus of the paper. A further exploration of the results and an in depth look at the potential downstream consequences would add to the work. The GTEx consortium published an eQTL analysis of SVs which we used as a second dataset for validation and comparison. They analyzed SV eQTLs from 147 individuals across 12 tissue types [22].

1.2.1 SV Detection

There have been many algorithms and techniques developed to detect structural variants from next generation sequencing (NGS) data. Before the rise of NGS both array comparative genomic hybridization (array-CGH) and SNP based arrays were used to find SVs. In general there are only a few overall strategies to detect SVs from NGS reads, including using read depth, using paired end read, using split reads, and *de novo* assembly. Read depth algorithms are well suited to finding and genotyping CNVs, but may not have very good breakpoint resolution. The majority of the detection algorithms use the information from paired end reads. When a read maps normally one expects a certain length and orientation. If these expectations are broken, this is an indication of the presence of a SV. If the insert length differs from the expected length this could indicate an insertion or a deletion. Differences in the orientation of the paired ends could indicate a duplication or an inversion. A split-read strategy can be used to identify SVs when the reads contain the ends, or breakpoints, of the SV. While *de novo* assembly is both computationally expensive and can be error prone, it can find novel and larger variants. Many of the SV detection algorithms use a combination of these approaches to improve accuracy and precision [23, 24, 25, 26]. The 1000 Genomes Project Structural Variation Analysis Group used a combination of Breakdancer, Delly, VariationHunter, CNVnator, GenomeSTRIP, PINDEL, MELT, dinumt and read depth to detect and genotype SVs [27, 28, 29, 30, 31, 32, 33, 34]. The GTEx consortium relied on LUMPY for detection and genotyping [35].

1.2.2 SVs and Disease

The first instances of structural variation described almost 100 years ago were very large chromosomal changes in fruit flies. Karyotyping and fluorescent in situ hybridization (FISH) were used to find similar rearrangements in humans, all of which were associated with disease. These can range from chromosomal aneuploidy, like chromosome 21 triploidy and Down syndrome, to large chromosomal deletions, duplications and translocations, which were associated with diseases such as Potocki-Lupski syndrome and Smith-Magenis syndrome. Chromosomal translocations and gene fusions were found to be driving mutations in chronic myelogenous leukemia and Burkitt's lymphoma respectively. Because these large events were the only ones detectable, it was thought that SVs were always associated with disease. When array-CGH was developed multiple CNVs were found in healthy individuals in 2004 [36, 37]. With array-CGH, SNP arrays and then NGS data more SVs, and smaller SVs were discovered, both in healthy individuals and associated with disease. SVs are now implicated in neurodevelopmental disorders like autism and schizophrenia, as well as cancer, but they are also found in healthy individuals [38, 39, 40]. Duplications or deletions can affect the gene copy number, which can have phenotypic consequences, but SVs can also impact phenotype when they overlap functional elements like promoters.

1.2.3 1000 Genomes SV Paper

The SV Analysis Group for the 1000 Genomes Project used Illumina reads (100bp mean, 7.4x coverage) from the 1000 Genomes samples to discover and genotype SVs.

Because the 1000 Genomes SV analysis relied on low coverage WGS data, it is under powered rarer and complex SVs. Even with that caveat, the analysis was able to discover many more SVs, and SV consequences than previous studies. From the analysis it was estimated that there are 18.4 Mbp of SVs in an individual genome, mostly made of 11.3 Mbp of mCNVs, and 5.6 Mbp of biallelic deletions. The reads were mapped using BWA and MrFast, and nine different SV detection algorithms were used to create the call set. Mapping and genotyping SVs is difficult because they have a tendency to be found in repetitive regions. The final call set contained over 60,000 SVs, which included 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 multi-allelic copy-number variants, and 16,631 mobile element insertions. 71% of SVs were novel when compared to the previous 1000 Genomes release, and 60% were novel when compared to the Database of Genomic Variation. By using PacBio long reads, and combining them with the WGS reads, breakpoints were established for over half the SVs. Most SVs were found to be shared between population groups. 68% of SVs with a VAF $> 0.1\%$ are in LD with a nearby SNP ($r^2 > 0.6$), but this is highly dependent based on the SV type.

There were statistically fewer CDSs, UTRs, introns and TFBSs found in deletions than expected by a random model (P ¡ 0.001). There was a direct relationship between size and rarity of deletions, the larger deletions being rarer: but this relationship does not hold for duplications. 240 genes were found in homozygous knockouts, which indicates that they are not crucial for survival. Most of these genes were novel knockouts, but they tended not to be highly conserved, and quite tolerant to mutation. The genotyped SVs with breakpoints were analyzed for complex events.

6% of deletions intersected another deletion, and 16% had an insertion at the deletion breakpoints. Many SVs showed even more complex situations, often with deletions, insertions and duplications all occurring in the location.

An eQTL analysis was performed using the SV genotypes, and 462 individuals from the Geuvadis project. The joint analysis between SNPs, indels and the SVs found 54 eQTLs with a lead association that was a SV. That is compared to the 9,537 eQTLs with a lead SNP/indel association. For 166 of these eQTLs, there was a SV in LD ($r^2 > 0.5$), which is seven times more than expected when compared to random variants. SVs were enriched for eQTLs compared to SNPs and indels, up to 50 times depending on the SV class. Larger SVs tended to have larger effect sizes, and those that overlapped genes also had the largest effect sizes. 136 SVs were in strong LD ($r^2 > 0.8$) with SNPs that were found to be GWAS hits. All of this suggests that SVs have a bigger impact on gene expression than SNPs, relative to their frequency.

1.3 MicroRNAs

MicroRNAs (miRNAs) are small (18-22 base pair) nucleotides that are involved in gene regulation. There is a complex set of cellular mechanisms that take miRNAs and use them to knockdown gene expression. They were first discovered in 1993, and thousands of potential miRNAs have been identified in humans. The miRNA mechanisms appear to be highly conserved among organisms. Mature miRNAs are used by the RNA-induced silencing complex (RISC) to bind to complementary RNAs, which then reduces translation, or marks the transcript for degradation [41]. Less than 10 years after the discovery of miRNAs, their involvement in cancer was first

identified. miRNA expression changes in cancer have been extensively studied, and several groups have identified specific miRNA signatures for a variety of tumor types. Even with all the efforts to study miRNAs it is still unclear the extent of their involvement in disease, and how other genetic factors play a role [42]. Just as gene expression is regulated by a host of factors, the same holds true for miRNA expression.

The Geuvadis project conducted miRNA expression sequencing on over 450 samples, and used the data to analyze how SNPs impacted miRNA expression. They conducted an eQTL analysis that used miRNA expression in the place of gene expression. They also analyzed the potential downstream effects by looking at SNPs that influenced both miRNAs and genes, and miRNAs that repressed gene expression. The found cases where a SNP that was implicated in an eQTL was also found in the miRNA eQTL, and that miRNA was a regulator of the gene.

We published a paper that used miRNA expression as the X values, and gene expression as the Y values in a linear regression association analysis in ovarian cancer [43]. While we called it an "eQTL analysis" in the paper, what we did was more of a miRNA-to-gene association analysis. We used expression to analyze pairwise correlations between miRNA expression and gene expression. We applied a similar method in other contexts to explore the impact of miRNAs on gene expression.

CHAPTER 2: METHODS

2.1 eQTL Analysis

There are many different methods for finding eQTLs, but in general they can be divided into two categories, pairwise association methods or statistical/machine learning methods. The earliest eQTL analysis methods developed were pairwise association methods, and relied on correlation or regression (Figure 1). In these methods a statistical test, like a linear regression, is performed for each combination of genetic variant and gene. Because of the number of tests performed, the choice of multi-test correction method is important. While a family wise error rate method like Bonferroni correction can be used, other less stringent methods such as permutation and false discovery rate (FDR) tend to be used. A random permutation correction is based on running the eQTL analysis many (e.g. 10,000 times) while permuting the genotype and expression values. The permutation p-value is based on the number of times the SNP and gene combination appears in the permuted results.

There are many pairwise association eQTL methods including Merlin, R/qtl, and Plink. Merlin is a set of tools for association and linkage studies, which includes eQTL analysis. Merlin has a focus on pedigrees, and can include this information in the eQTL analyses [44]. R/qtl is an R based eQTL package that uses Hidden Markov Models, and Haley-Knott regression to perform a correlation analysis [45, 46]. PLINK

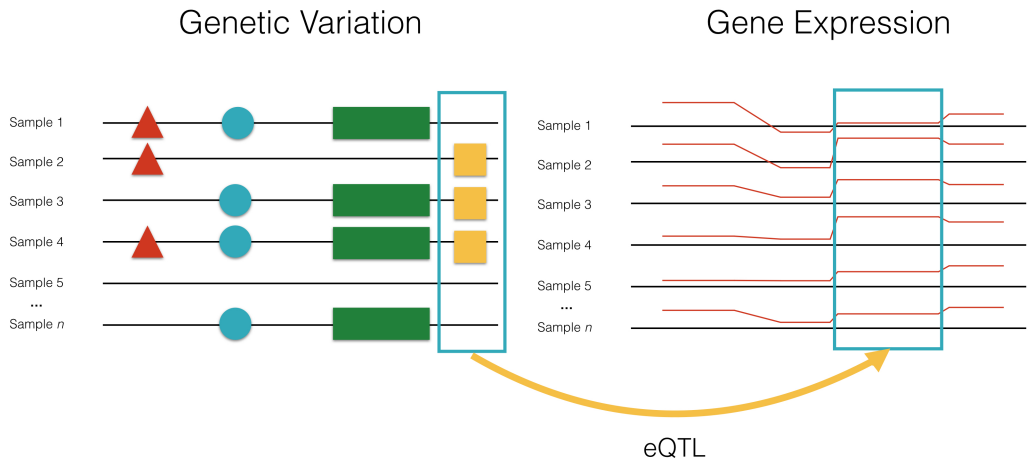


Figure 1: Pairwise Associations in an example eQTL

is a suite of tools for genome wide association studies, and can perform eQTL analyses. PLINK can calculate linkage disequilibrium, inbreeding coefficients as well as other statistics [47]. PLINK uses a binary file format that other eQTL packages have also used. Merlin, PLINK and R/qtl are older pairwise association methods, and tend to be computationally and time intensive when used on larger datasets, that are now standard. MatrixEQTL and FastQTL are among a newer generation of linear regression software packages [48, 49]. They implement a range of algorithmic methods in order to run eQTL analyses on a large number of individuals in a reasonable amount of time.

The second group of eQTL analysis methods are the statistical and machine learning methods. While linear regression treats each genetic variant/gene pair as independent this does not reflect biological reality and some of these methods attempt to address this. They include MTLasso2G, SCGGM, ICE, PANAMA and others [50, 51, 52, 53]. ICE is based on a linear regression analysis, but combines it with a statistical model to correct for complex correlation structures. It can correct for both

strong and moderate effects, and has been shown to reduce the number of false positive eQTLs in *trans*-, as well as increase the number of *cis*- associations. PANAMA uses a linear mixed model to jointly learn the confounding factors and the impact from the genetic variants. SCGGM is a probabilistic graphical model that models how SNPs perturb a gene network [54]. geQTL is a sparse regression method that can map single or groups of genetic variants to a single or a group of genes [55]. In the following sections we will discuss in depth one method from each type, MatrixEQTL from the pairwise association group, and MTLasso2G from the statistical learning group.

2.1.1 MatrixEQTL

We use MatrixEQTL to conduct our eQTL analyses. A naive implementation of linear regression for eQTL would be computational intensive. For example an eQTL analysis using the 1000 Genomes SNPs and genes has 224,857,198,181 potential pairwise associations. To address this MatrixEQTL works on the data in chunks of up to 10,000 lines. For each pair of 10,000 variable chunks MatrixEQTL calculates a correlation matrix, and then selects pairs that have a correlation above the defined threshold, and then MatrixEQTL calculates the linear regression only for those pairs. This approach drastically reduces the computational burden and time for an eQTL analysis even when using large datasets. We have written R code that extends MatrixEQTL and simplifies it use. MatrixEQTL requires a genotype matrix, a gene expression matrix and the positions of both the genes and genotypes. This can be generalized to a X and Y matrices. While not required by MatrixEQTL we also

include covariates to remove any potentially confounding effects. While we typically use a distance cutoff of 1 megabase and limit our analysis to just *cis*-, MatrixEQTL can perform both *cis*- and *trans*- eQTL analyses, and adjust the distance cutoff. MatrixEQTL outputs a file with eQTLs, and a qq-plot. See Figure 2 for an overview of the workflow.

2.1.1.1 Data Preprocessing

To obtain the best results the data needs to be preprocessed. This preprocessing could include a principal components analysis (PCA), quantile normalization and minor allele frequency filtering. The steps outlined below are what we use for MatrixEQTL, however similar steps are taken for most eQTL analyses. If the X matrix contains discrete values such as genotypes, then a minor allele frequency (MAF) filter should be used. This helps insure that there are enough individuals (e.g. 5% of the sample) with the minor allele so the associations aren't driven by a handful of outliers. We normalize the gene expression by using inverse quantile normalization. This maps the gene expression values onto the standard normal distribution based on their rank. Importantly this fulfills the assumption of normally distributed data required by linear regression, and it also helps limit the effect of outlier genes. It also removes any effects from differences due to the large variation in gene expression values. We also apply inverse quantile normalization to the X matrix if it contains continuous values, such as miRNA expression.

For covariates we include gender, population, and principle components on both the genotypes and the gene expression. We use a screeplot and the percentage variation

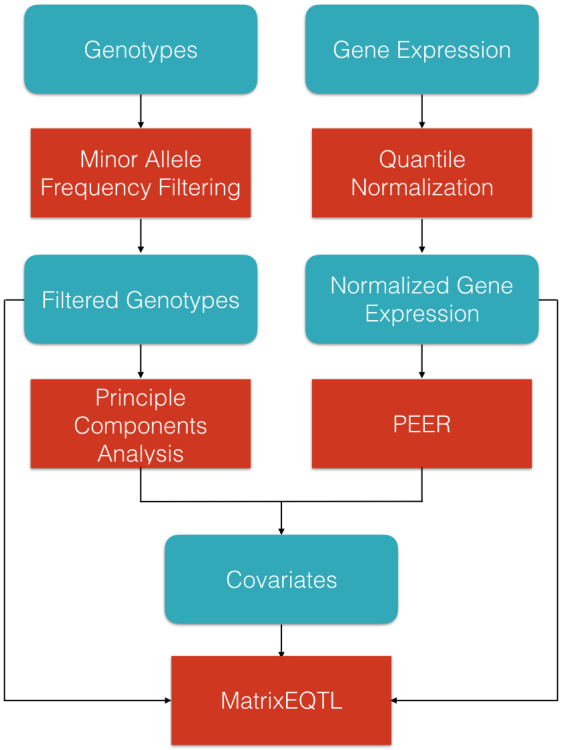


Figure 2: eQTL Workflow

explained to determine how many PCs to include. Instead of including gene expression PCs in the covariates, we can include PEER factors. This technique is commonly used in eQTL analyses and has been used in both the Geuvadis project and the GTEx project. Like PCs, PEER factors help account for some of the unknown covariance in the data. By removing covariates we hope that we are removing confounding elements in the data, and leaving only the biological signals. One method for evaluating the efficacy of the covariates, specifically the number of PCs, is to take a subset of the data and run multiple trials using it. The number of PCs and PEER factors that produces the greatest number of eQTLs is then used on the full dataset. Another approach which has been used by the GTEx project, is to simply set the number of PCs and PEER factors based on the number of samples.

2.1.2 MTLasso2G

Multitask Lasso 2G (MTLasso2G) is an orthogonal approach to conducting eQTL analyses based on a sparse modeling approach using Lasso. MTLasso2G utilizes the relationships between the genotypes and the relationships between the genes in the analysis instead of treating each pair as independent. The sparsity in MTLasso2G and other similar methods, is based on the assumption that there only exists a small number of associations between genetic variants and traits. MTLasso2G creates a graph or network based on the correlation matrix for both the genotypes and the gene expression. This also allows for the association between groups of related genetic variants and a trait. This reduces the dimensonality of the genotypes and gene expression and induces the sparsity. MTLasso2G uses these two graphs to inform the lasso portion of the analysis. The output includes the two graphs and the final association matrix. Any non-zero value represents an association between the respective genotype and gene. The input for MTLasso2G is similar to MatrixEQTL, with a few modifications. The genotypes are MAF filtered, and the gene expression is inverse quantile normalized. Instead of including PCs and PEER factors as covariates, they are regressed out of the respective matrices.

2.1.2.1 Hyperparameter Optimization

In the broadest possible terms machine learning is the process of optimizing a set of parameters given a set of inputs. Many machine learning methods depend on the selection of a set of hyperparameters. These hyperparameters include things like the learning rates in a neural network, and the regularization parameter in Lasso

type methods. Given the same dataset and different hyperparameters these methods will produce different results, which means the selection of the hyperparameters is a critical step in any machine learning analysis. For MTLasso2G there are three different hyperparameters (γ_1, γ_2, and λ) that need to be tuned, and SCGGM has two hyperparameters. We analyzed three different methods of hyperparameter selection and optimization. Hyperparamter optimization allows us to select a good set of hyperparameters without relying on guessing and checking, thus producing better eQTL results. This also allows us to have some degree of reproducibility in our hyperparameter selection process.

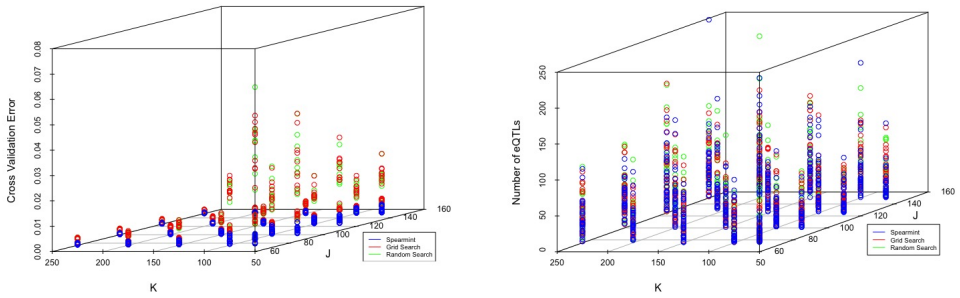
Hyperparameters can be manually tuned, that is the user can adjust the hyperparameters in an attempt to achieve better results. While there is no overhead for manual tuning the results are dependent on the user and not easily replicated. There are a few different methods for automatic hyperparameter optimization. The first, and simplest, is a grid search. Here a set of values for each hyperparameter is specified, and the analysis is run with every possible combination of the hyperparameter sets. A grid search is easy to implement, easily run in parallel, and given the same amount of time as manual tuning a grid search can usually find better hyperparameters. Another simple to implement technique is a random search. A range of hyperparameters is specified and an value from this range is randomly selected for each optimization run. Published results suggest that a random search can find better hyperparameters in a shorter amount of time, when compared to a grid search[56]. This is due to the fact that a grid search can easily miss sets of hyperparameters that aren't specified, where as a random search does not rely on specific values. Snoek et

al. applied Bayesian optimization to the hyperparameter search question, and found that it could out perform both grid and random search.

To test the differences between these methods we applied each one to eQTL analyses using MTLasso2G. The datasets were generated from the gene expression values and genotype data from the Yoruban population varying the number of samples, the number of SVs, and the number of genes. For each analysis run we compared the number of eQTLs, and the cross validation error. We wrote custom code (three nested loops) for the grid search, and used the Spearmint software package for the random search and Bayesian optimization. To create a fair comparison we allowed each method to run for 320 iterations. Figure 3 shows the cross validation error and number of eQTLs for grid search, random search and Bayesian optimization across 25 different combinations of number of SVs and number of genes. Overall we found that the Bayesian optimization worked the best, but the grid search was not far behind. A single eQTL analysis using MTLasso2G can take hours or even days, which makes hyperparameter tuning computationally intensive and time consuming. Since the results of each analysis is so highly dependent upon the hyperparameters, the time taken to optimize the hyperparameters is well spent. While the results from these tests only apply to MTLasso2G, the principle of using hyperparameter optimization for any other machine learning based bioinformatics methods is the same.

2.2 Network Analysis

In prior work we constructed integrated networks that combined multiple different data types and cellular interactions [57]. We start with the eQTL or miRNA-gene



(a) (b)

Figure 3: Results for grid search, random search, and Bayesian optimization on 10 randomly sampled datasets with the number of SVs J=[75,100,125,150] and the number of genes K=[75,100,150,200,250]. (a) CV errors; (b) The number of eQTL associations.

association results as the seed nodes for the network. We add protein-protein interactions and gene expression correlations using the genes from the results. If we are starting from miRNA-gene associations we also add miRNA targets from TarBase, or another similar miRNA target database, and miRNA expression correlations. Using this method we construct a network comprised of genetic variants, or miRNAs, and genes. We can apply network analysis methods, like node centrality and connectivity to analyze these networks and determine the importance of the genes or miRNAs. Network alignment can be used to compare networks in order to find the similarities and differences.

2.3 Downstream Analyses

After performing the eQTL analyses we can analyze the results a few different ways. Gene enrichment analysis can be used to evaluate the eQTL genes, and disease associations can be performed on both the eQTL genes, as well as the SVs and SNPs.

We can compare our eQTL results to those generated by other groups who used the same data. We can also use functional enrichment analysis to see if the genetic variants in our eQTLs overlap functional elements such as gene promoters. There are several methods for narrowing down which genetic variant is causal in a set of eQTLs including CAVIAR and DAP [58, 59].

CHAPTER 3: UTILIZING NETWORKS: OVARIAN CANCER AND DEEP LEARNING

3.1 Ovarian Cancer miRNA eQTL Analysis

Ovarian cancer is responsible for ∼5% of female cancer deaths. It is estimated that in 2018 there will be ∼22,240 new cases and 14,070 deaths from ovarian cancer [60]. Previous studies have suggested that miRNA dysregulation plays a role in ovarian cancer [61, 62]. The overall 5-year survival is below 50%, but it drastically improves when diagnosed in the earliest stages [63]. The TCGA found miRNA subtypes in their analysis of serous ovarian cancer. We published a paper in 2015 that analyzed the associations between miRNA expression and gene expression in ovarian cancer [57]. In the paper we constructed an integrated network using multiple levels of data including miRNA-gene associations, protein-protein interactions, and miRNA targets. In this analysis we conducted an eQTL analysis, a miRNA target search, a protein-protein interaction search, a miRNA-miRNA correlation and gene-gene correlation search and then performed a network integration (Figure 4).

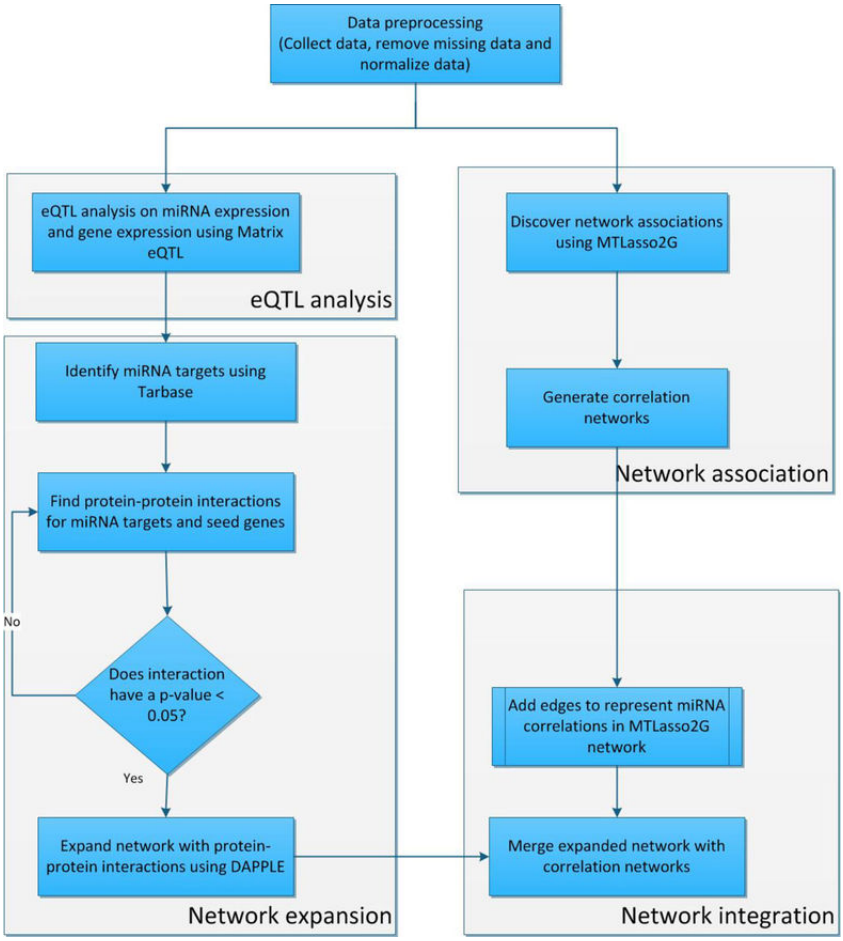


Figure 4: The workflow from "An Integrated Network of microRNA and Gene Expression in Ovarian Cancer".

For the analysis we used the miRNA isoform expression and mRNA expression from the TCGA, which contained 480 samples. Both expression datasets were generated on the Illumina HiSeq. We removed miRNAs and genes with missing values, giving us 183 miRNAs and 13,536 genes, and then applied between-sample quantile normalization to both. We used MatrixEQTL to perform the association analysis, with a *cis*-distance cutoff of 1 MB, and a FDR cutoff of 0.01. We used the association results as seed nodes to expand the network in a variety of directions. We used TarBase to add the genes that the miRNAs target, and DAPPLE to add protein-protein interactions

[64, 65]. For the DAPPLE expansion we only chose interactions with a corrected p-value of < 0.05. The initial integrated network contained the overlap of the miRNA-gene associations, miRNA targets and protein-protein interactions. To this network we added miRNA-miRNA correlations and gene-gene correlations from MTLasso2G. MTLasso2G was also used to generate a second set of eQTLs, which were added to the network. The resultant network demonstrates the potential complexities and interplay between gene and miRNA expression.

Using this technique we created a network that was composed of 167 nodes and 277 edges (Figure 5). The initial dataset was 44 miRNA-gene associations. We found 310 miRNA targets in TarBase. Of these only 244 could be used as inputs for DAPPLE. From DAPPLE we found 236 protein-protein interactions, with 145 having a corrected p-value of < 0.05. We created an initial network from the miRNA-gene associations, 145 protein-protein interactions, and 108 miRNA target interactions. We were able to add 9 miRNA correlations and 18 gene correlations from MTLasso2G, as well as 8 of 48 miRNA-gene associations. 26 of the 145 genes in the network were associated with cancer, 11 of which were specifically associated with ovarian cancer. The integrated network captured more cancer genes and ovarian cancer genes than the DAPPLE extended network, the miRNA targets or the MTLasso2G network.

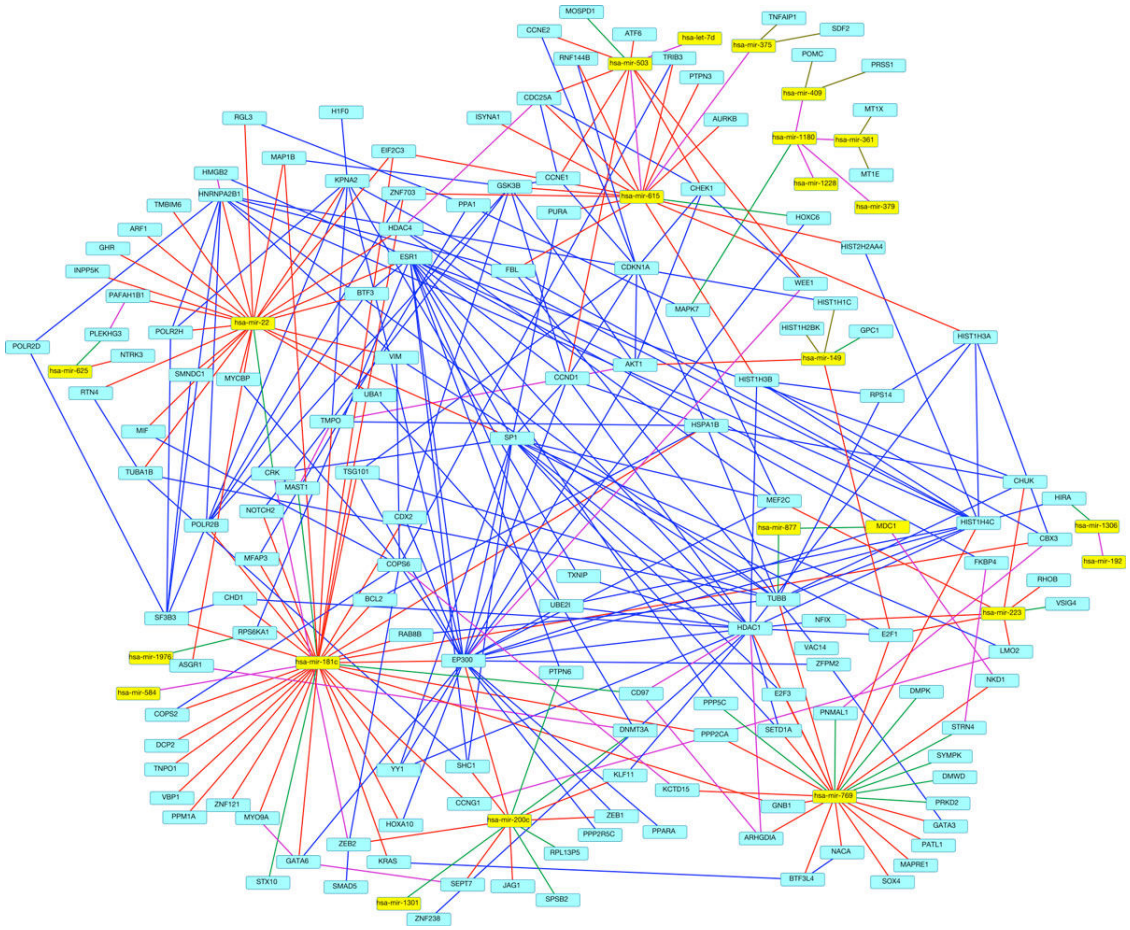


Figure 5: The finale integrated ovarian network containing 167 nodes and 277 edges.

In the integrated network we found $ATK1$, which has been implicated in tumerogenesis in ovarian cancer. $ATK1$ was found in each of the component networks as well. We also found a subnetwork containing $ATK1$ that had $ESR1$ in a central role. $ESR1$ has been associated with breast cancer, and its expression has been suggested to be a predictor of ovarian cancer survival. This subnetwork contained 5 miRNAs and 12 genes, connected by three miRNA-gene associations, six miRNA target associations and protein-protein interactions.

In our integrated network we found that $E2F1$ was targeted by three separate miRNAs. The down regulation of $E2F1$ has been shown to inhibit ovarian cancer

growth, and $E2F1$ has also been studied as a predictor of drug resistance in ovarian cancer. In both these cases, the $ATK1$ subnetwork, and the miRNA interactions with $E2F1$ the miRNA associations could provide insight into the roles and regulation of these genes. Studying these miRNA associations, and others like them, could shape our understanding of these genes role in ovarian cancer.

We used multiple databases to examine the miRNAs and their relation to cancer. Of the 22 miRNAs in our integrated network, 14 were associated with cancer in general, and 12 were associated with ovarian cancer specifically. Increased expression of hsa-miR-200c has been associated with better survival and increased responsiveness to paxlitacyl, a chemotheraputic. hsa-mir-22 has been associated with survival and recurrence of ovarian cancer, and in our network targets $ESR1$.

By combining these different types of data we are able to create a picture that better captures the interactions that occur in biology. This holistic approach allows us to analyze changes in miRNA and gene expression simultaneously and to take into account the impact that they have on each other. Examining the component parts individually does not have the same results, as each component only provides part of the picture. This same approach can not only be applied to other types of cancer, but can be expanded to include even more types of data.

3.2 Predicting Gene Expression from Genotypes Using Deep Learning

In "A deep auto-encoder model for gene expression prediction" we applied deep learning techniques to the genotype-to-expression problem [66]. We used a Multi-Layer Perceptron and Stacked Denoising Auto-encoder (MLP-SAE) to predict gene

expression of yeast from their genotypes. It has been long established that genetic variation can lead to phenotypic differences, and we know that some of the differences in gene expression levels are driven by genetic variation. These two facts are the rationale behind GWA and eQTL studies respectively. Studies have shown that GWAS SNPs are enriched for eQTLs, which suggests that both techniques are picking up a signal from the same mechanism [6]. This also means that eQTL studies can be used to examine the intermediate layer, i.e. gene expression, between genotype and phenotype.

One of the benefits of deep learning is that the hidden layers can capture previously unknown and complicated structures in the data. Another benefit is that, unlike linear regression based eQTL analyses, deep learning does not treat each individual genotype-gene expression pair as independent. Deep learning models have been previously applied to other areas of bioinformatics, including predicting splicing from RNAseq data, the functionality of non-coding variants and many areas in proteomics.

A multilayer perceptron maps input to output using a feedforward neural network, and each layer is fully connected to the next. The nodes in the hidden layer use non-linear activation functions. The model is learned by adjusting the weights of the connections between the nodes using the backpropagation algorithm. An auto-encoder is another neural network type which can be used for dimensionality reduction. An auto-encoder is also composed of input, output and hidden layers. Like MLPs, auto-encoders can be trained using backpropagation, where the errors between the expected and predicted results are used to adjust the weights between the nodes.

A denoising auto-encoder is an type of auto-encoder which can be used to separate signal and noise. The MLP-SAE model combines the MLP model and auto-encoders by using them as the hidden layers. The input layer takes SNP genotypes, and the output layer is predicted gene expression values.

The MLP-SAE model that we used was made up of four layers, one for each input and output, and two hidden layers in between. To prevent overfitting of the model, we employed dropout. To use dropout nodes, and any of its connections are temporarily removed from the network. This helps prevent overfitting by reducing the dependency of the network on any individual node. To evaluate how the deep learning model performs we compared it to other similar machine learning based methods, Lasso and Random Forests, using genomic data from yeast. We performed a series of tests varying the hyperparameter values (α for lasso, number of trees for random forests, and learning rate for MLP-SAE), and compared the cross validation error. The MLP-SAE model outperforms both Lasso and Random Forests when predicting gene expression from genotypes using this yeast dataset. The addition of dropout decreased the cross validation error further.

We used the yeast data generated by Brem et al. for our project [67]. There were 2,956 SNPs and 7,085 genes from 112 samples of a BY4716 and RM11-1a cross. After removing genes with missing expression values we were left with 6,611 genes. We used Scikit-Learn to impute and scale the SNP genotypes. When we compared the predictions from the MLP-SAE model with dropout to the actual gene expression values we found that they were well correlated. The model captured many of the features of the real data, as can be seen in Figures 6 and 7.

Figure 6: Predicted vs. observed gene expression. The deep learning model recapitulates many of the features that we find in the measured gene expression from the yeast.

Besides an increased accuracy in prediction from lasso or random forests, a benefit of a deep learning model is that it can incorporate other data like regulatory and epigenetic elements, biological pathways, and environmental conditions. The inclusion of these factors could lead to a better model, as would using a dataset with a larger number of samples. Other deep learning architectures, like Recurrent Neural Networks, could be applied to this and other similar problems. The MLP-SAE method could also be applied to other expression datasets beyond yeast, and problems like tissue specific expression.

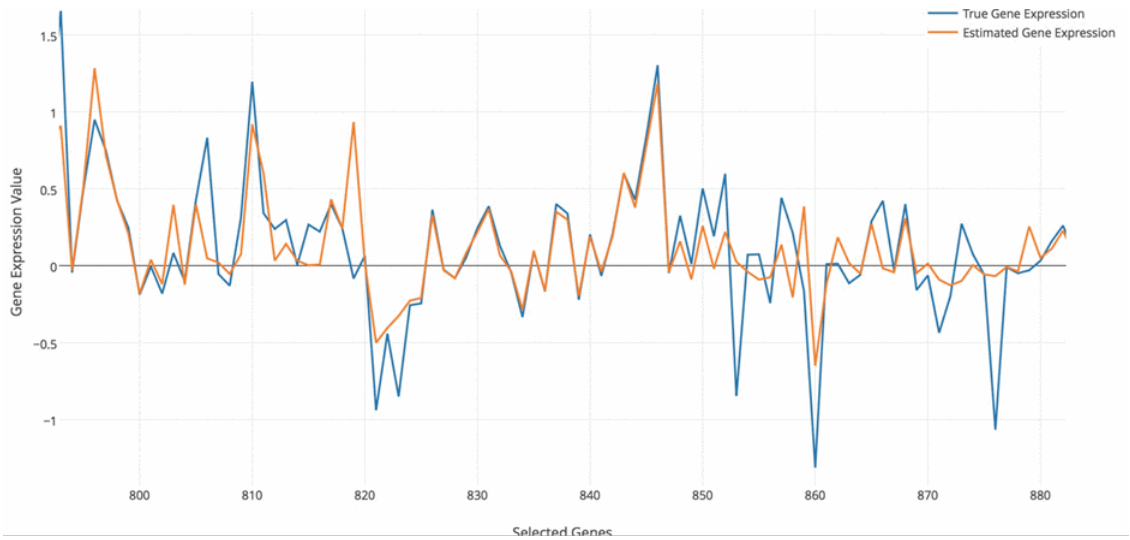


Figure 7: A subset of the predicted vs observed gene expression. Here it is clear that while the model does not perfectly represent the actual gene expression, it has many of the features seen in the data.

CHAPTER 4: EQTL ANALYSIS PIPELINE

There are many different eQTL methods available, but only some of the newer methods can easily handle the sample sizes that the standard today. One of these methods is MatrixEQTL, and while it is a good eQTL analysis tool, there are improvements that could be added, and additional tools that could be included for ease of use. We developed a pipeline based on MatrixEQTL, that adds functionality and tools to aid in the eQTL analysis process from beginning to end. We have added the ability to normalize the gene expression data, run PEER and perform a PCA to create covariates, filter the genotypes based on the minor allele frequency, and create exploratory plots of the results. The software is available at https://github.com/andrewquitadamo/matrix_eqtl_pipeline.

4.1 Pipeline Software Description

The pipeline is written in a combination of R and Python. In total it is over 1400 lines of code, 85% Python and 15% R. The Python code is used to wrap the R code, to parse the input files, and to provide the user interface. The R code is used to conduct the eQTL analysis and generates the plots and figures. We use rpy2 as a bridge between R and Python. The use of R was necessitated by MatrixEQTL, but as R is not the ideal environment for parsing files, Python was chosen to handle those tasks. The use of Python also allows us to utilize parts of Scikit-Learn.

There are a few things that need to be done to any raw data before MatrixEQTL can be used. These preprocessing steps ensure the most accurate results are obtained; see Methods for a discussion of these steps. The inputs for this pipeline are a VCF file, a gene expression matrix and gene positions. The overall workflow is documented in Figure 8

The simplest usage is to rely on the default settings. All the user has to do is supply the requisite files and specify the number of PEER factors to correct, as in Listing 1.

Listing 1: The simplest invocation of the eQTL pipeline

```
python3 matrix_eqtl_wrapper.py \
        -v data/snp_genotypes.vcf \
        -g data/gene_expression -p data/gene_positions -n 5
```

The first script in the pipeline, *remove_vcf_header.py*, removes the meta-informational lines in the header of the VCF file, i.e. those that begin with ##. This leaves only the header line with the sample IDs, and the genotype lines. The next script, *vcf_overlap.py*, overlaps the VCF file and the gene expression matrix file. The output files only contain the samples found in both files, and in the same order. MatrixEQTL will not accept input files that have a differing number of samples, or files that have samples in a different order. *filter_snps.py* filters the VCF file using a minor allele frequency (MAF) cutoff. The user can select their desired MAF cutoff, and SNPs with a minor allele frequency that falls below the cutoff are removed, and not used in the eQTL analysis. *parse.py* takes the VCF file and produces a

singular summed genotype. For example 0|0 becomes 0 and 0|1 or 1|0 becomes 1. *position.py* uses the VCF file and creates two different genotype position files, one that is in MatrixEQTL format, and one that contains the ID, chromosome, start position, stop position, and type for each genotype. *run_iqn.py* is a Python wrapper for *general_iqn_py.R*, which runs inverse quantile normalization on the gene expression matrix. *pc_covariates.py* calculates the principal components of the genotypes using the IncrementalPCA method from Scikit-Learn. The PCs are then used as covariates in the eQTL analysis. *run_peer.py* is a Python wrapper for *peer_function.R* which runs PEER on the gene expression matrix. The PEER factors can then used as covariates for the eQTL analysis. *combine_covariates.py* takes the genotype PCs, the PEER factors, and any additional covariates that the user supplies and combines them into one file to be used as input for MatrixEQTL. *run_matrix_eqtl.py* is the Python wrapper for *mxeqtl.R*, which is code that has the R functions to run MatrixEQTL. *modify_matrix_eqtl.py* is used as part of the installation process. It downloads the current MatrixEQTL distributions code, and makes some tweaks to the code that prints messages, so that it interacts with Python and rpy2 better. *CorrBoxPlot.py* is the wrapper for *CorrBoxPlotFile.R*, which produces both the correlation between genotype and gene expression for each eQTL and boxplots to represent the gene expression for each genotype in an eQTL. *manhattan.py* is the Python wrapper for *manhattan.R* which produces a Manhattan plot of the eQTL results. This can be used to find loci of interest, where there are peaks of significance.

matrix_eqtl_wrapper.py is the main entrypoint for the pipeline, and will run the analysis from start to finish. It provides the user control over the options for each

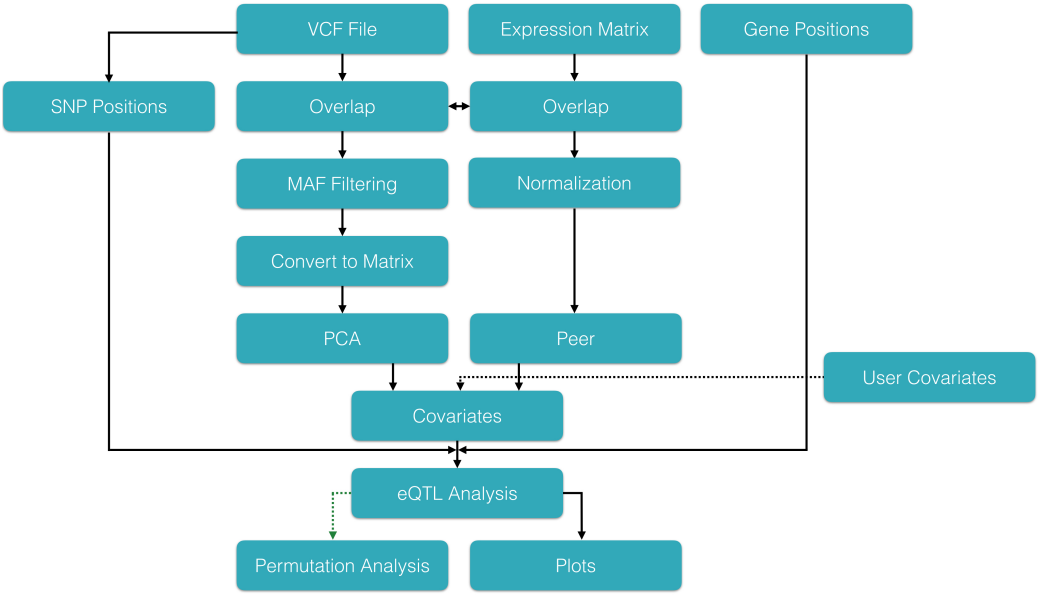


Figure 8: The workflow for the eQTL pipeline.

step of the process, such as specifying the number of genotype PCs to include, the *cis-* p-value, and the *cis-* distance cutoff. The complete list of options can be seen in Figure 9. This pipeline requires a Python3, R > 3.3, Numpy, Scipy, Pandas, Scikit-Learn, rpy2, and PEER. An example of how to install this pipeline on Ubuntu 18.04 can be found in the GitHub repository. Besides using the pipeline as a single tool, the individual component tools can be used by themselves as well.

4.2 Discussion

This software package provides useful extensions to MatrixEQTL, both pre- and post- analysis. There are functions to help preprocess the data, generate the required input files and functions to visualize the eQTL results. The ultimate purpose of this package is to help others use MatrixEQTL without any overhead or having to write custom code to preprocess data. This pipeline attempts to codify best practices for an eQTL analysis. Previously the functions around the periphery of the analysis in

```
usage: matrix_eqtl_wrapper.py [-h] -v -g -p -n
                              [--headless-vcf-filename]
                              [--overlap-extension] [--maf-cutoff]
                              [--filtered-filename] [--parsed-filename]
                              [--position-filename]
                              [--meqtl-position-filename] [--number-pcs]
                              [--pc-filename] [--normalized-filename]
                              [--peer-factor-filename]
                              [--combined-covariate-filename]
                              [--additional-covariates]
                              [--trans-output-file] [--trans-p-value]
                              [--model] [--cis-distance] [--cis-p-value]
                              [--no-header] [--no-rownames] [--missing]
                              [--sep] [--maf] [--qqplot]
                              [--eqtl-output-file] [--boxplot-pdf-file]
                              [--correlation-output-file]
                              [--manhattan-pdf-file]

optional arguments:
  -h, --help            show this help message and exit
  -v, --vcf-file    VCF file for MatrixEQTL
  -g, --gene-expression-file
                        Gene expression matrix file
  -p, --gene-position-file
                        Gene position file
  -n, --numfactors  Number of PEER factors to correct for
  --headless-vcf-filename
                        Name for the VCF file with ## removed
  --overlap-extension
                        File extension to use for overlapped files
  --maf-cutoff        Minor allele frequency cutoff for genotype filtering.
                        Default is 0.05
  --filtered-filename
                        Name for the MAF filtered VCF file
  --parsed-filename   Name for the parsed VCF file
  --position-filename
                        Name for the position and type file
  --meqtl-position-filename
                        Name for the MatrixEQTL formatted position file
  --number-pcs        Number of genotype principal components to correct
  --pc-filename       Name of the genotype PC file
  --normalized-filename
                        Name for the inverse quantile normalized gene
                        expression matrix
  --peer-factor-filename
                        Name for the PEER factors
  --combined-covariate-filename
                        Filename for the combined covariates
  --additional-covariates
                        Filename of any additional covariates to include in
                        the combined covariate file
  --trans-output-file
                        Filename for the trans- eQTL results
  --trans-p-value     Maximum p-value for eQTL results for trans- analysis
  --model             Model to use for eQTL analysis
  --cis-distance      eQTL search window size in base pairs
  --cis-p-value       Maximum p-value for eQTL results for cis- analysis
  --no-header
  --no-rownames
  --missing           String for missing data
  --sep               Character separating fields; Default is tab
  --maf               MAF cutoff for filtering by MatrixEQTL
  --qqplot            Name for the qq-plot PDF file
  --eqtl-output-file  Filename for the cis- eQTL results
  --boxplot-pdf-file  Name for the boxplots PDF file
  --correlation-output-file
                        Filename for Genotype-Gene correlations
  --manhattan-pdf-file
                        Filename for manhattan plot
```

Figure 9: The options for the eQTL pipeline.

MatrixEQTL, while essential, were somewhat neglected. This left users to develop their own scripts and processes to address this. Instead of having every MatrixEQTL user roll their own solution, a single, previously developed pipeline saves collective time and effort. For example the pipeline also allow the user to start directly from a VCF genotype file, which cannot be used as input for MatrixEQTL. The modular nature of the pipeline could allow for other eQTL analysis tools to be used instead of MatrixEQTL without retooling the entire workflow. The pipeline also allows novice users to get started with eQTL analysis using the default settings, and the downstream visualizations can help users explore the eQTL results. The pipeline has the option to produce boxplots for each eQTL, as well as a Manhattan plot. These visualizations bring together different aspects of the data, that the user would have to otherwise manually compile.

CHAPTER 5: STRUCTURAL VARIANT FUNCTIONAL ANALYSIS

The purpose of this chapter is to analyze the role that structural variants play in gene and miRNA expression. We produced SV eQTLs, joint SNP/SV eQTLs, SV miRNA-QTLs, joint SNP/SV miRNA-QTLs, and miRNA-gene associations. We studied the associations between genetic variation, gene expression and miRNA expression, and have applied some downstream analyses to the eQTL data.

5.1 SV and SNP eQTL Analyses

We performed an eQTL analysis using structural variants as the genotypes of interest, and a separate analysis which only included SNPs. We used the SV and SNP genotypes from the 1000 Genomes Project, and the gene expression from the Geuvadis project, and normalized the data as outlined in the methods. We used 37,172 SVs, and 9,411,447 SNPs, which we ran against 23,723 genes. At a FDR rate of 0.01, we found 974 SV-only eQTLs, which involved 488 SVs and 474 genes. We found 794,802 eQTLs which contained 499,212 unique SNPs and 8,010 unique genes. 465 (98.1%) of the genes in the SV-only analysis were also found in the SNP-only analysis.

5.2 Joint SNP-SV eQTL Analysis

We used both the SV and SNP genotypes to perform a joint eQTL analysis. This analysis is similar to the one in the 1000 Genomes paper [22].

After filtering and pre-processing this analysis used 9,448,610 genetic variants and

Table 1: SNP and SV FDR and Beta Distributions

	Minimum	First Quartile	Median	Mean	Third Quartile	Max
SNP FDR	2	2.855	4.442	7.964	8.390	143.8
SV FDR	2.001	2.901	4.555	7.331	8.277	128.8
SNP Beta	-2.2	-0.436	0.2947	0.04801	0.4766	2.801
SV Beta	-2.528	-.4967	0.3226	0.09828	0.522	3.776

23,723 genes, and found 11,269,037 eQTLs in total. We found 793,654 eQTLs with a FDR of 0.01 in the joint SV/SNP eQTL analysis, which involved 8,010 unique genes and 499,555 unique genetic variants. The genetic variants were involved in an average of 1.59 eQTLs (min. of 1, median of 1, max of 16). Most genes had multiple eQTLs, with an average of 99 (min. of 1, median of 15 and max of 11,490). The eQTLs could be broken down into 1,117 SV eQTLs and 793,685 SNP eQTLs. The SV distribution was similar to the overall genetic variant distribution, with each SV involved in an average of 1.55 eQTLs (min. of 1, median of 1, max of 13), however they impacted fewer genes (mean of 5.47, min. of 1, median of 1, max of 536). The FDR distribution for the SNP and SV eQTLs appears to be similar, but the SVs appear to have slightly larger effect sizes. (Table 1, Figure 10) In the joint analysis results there were 474/474 (100%) of the genes from the SV-only eQTL analysis, and 7,991/8010 (99.76%) of the genes from the SNP-only eQTL analysis.

When we look at the Manhattan plot we can see that at most loci the SNPs dominate, but at several loci there is a SV (or SVs) that are the lead association (Figure 11). The Manhattan plots for the SNP-only (Figure S1) and SV-only (Figure S2) are similar to that from the joint analysis.

We can create a similar plot that uses the beta values in place of the FDR. This

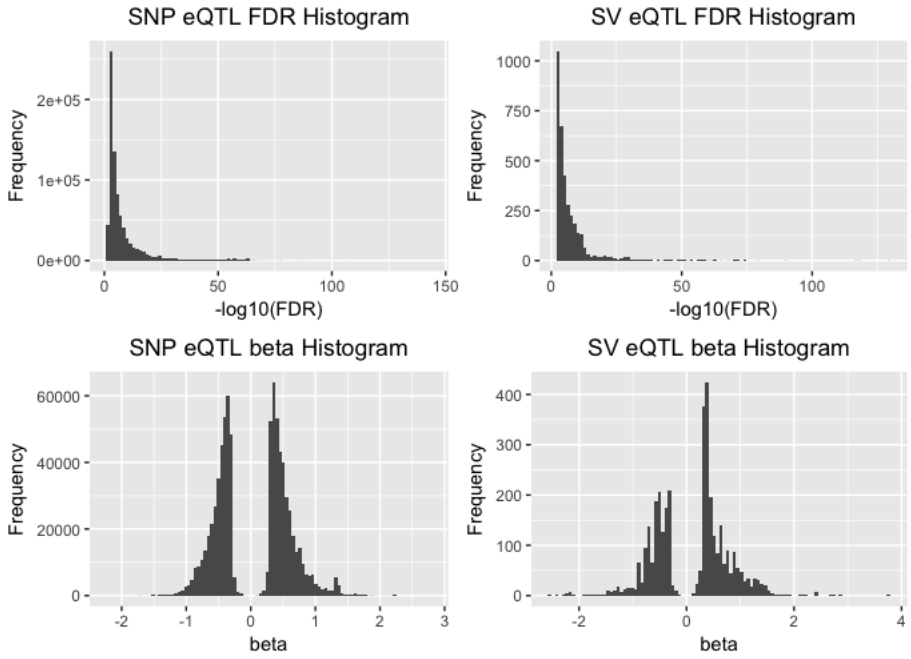


Figure 10: The SNP and SV eQTL FDR and Beta distribution histograms. Both the SNP and SV FDR distributions are similar, but the beta distribution for the SV eQTLs is not as symmetric as the SNP distribution, and is also shifted towards larger effect sizes.

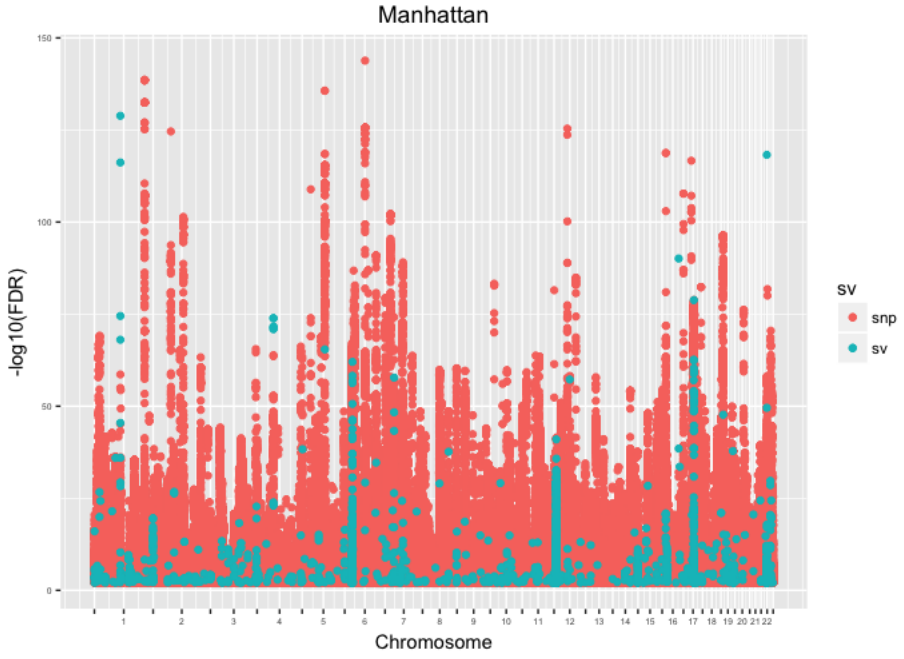


Figure 11: Whole genome FDR Manhattan plot. A Manhattan plot shows areas of the genome that have significant eQTLs. For most loci there are many genetic variants due to linkage disequilibrium. Here we can see loci where SVs contribute significantly.

plot differentiates between positive and negative effects, that is it separates out the genetic variants that are associated with an increase in gene expression and those that are associated with a decrease in expression. In this plot we can see that there are more loci where the SVs are the lead association when compared to the FDR plot(Figure 12). Figures S3 and S4 show the beta Manhattan plots for the SNP-only and SV-only analyses.

Figure 12: Whole genome beta "Manhattan" plot. This plot shows the effect size of each genetic variant. Like the traditional Manhattan plot we can see loci of interest, but in this plot we can also distinguish how the variant effects the gene expression. Using this we find more locations where an SV has the largest effect.

When the FDR is plotted against the distance to the transcription start site, we see a similar pattern to the one that was previously described in other eQTL analyses, where the more significant eQTLs cluster around the TSS (Figure 13).

Table 2: SVs Shared between Joint and SV-only analyses

	Joint	SV-only	Both
SVs	205	61	913

Table 3: SNPs Shared between Joint and SNP-only analyses

	Joint	SNP-only	Both
SNPs	0	323	793,684

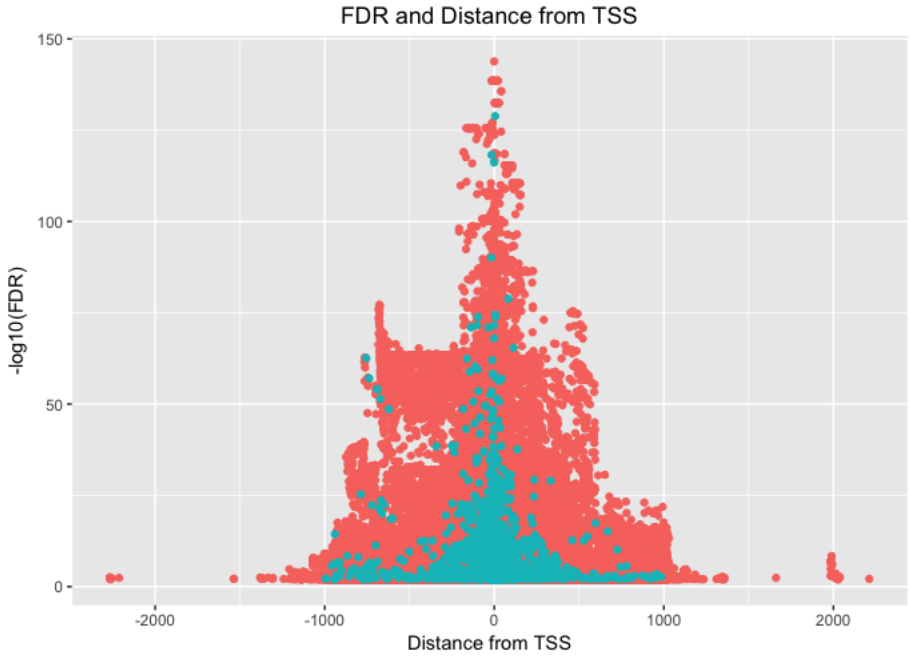


Figure 13: The eQTL FDR plotted against the distance to the TSS. As found by other groups in prior analyses the more significant eQTLs tend to cluster around the TSS.

When we compare the joint eQTLs to the SV-only and SNP-only eQTLs, the vast majority overlap. 913 (77.4%) of the SVs are shared between the Joint and SV-only analyses, and 793,684 (99.95%) SNPs are shared between the Joint and SNP-only analyses (Tables 2 and 3). For the highest ranking eQTLs the ordering is either the same, or very similar between the joint eQTLs and the SV/SNP-only eQTLs.

We found 32 eQTLs that are shared between our results and the GTEx SV eQTLs.

Because of the different approaches between the studies, there were only 412 potential eQTLs that shared genes from the GTEx eQTLs.

We found 133 eQTLs with a lead SV based on FDR. When we look for the genotype with the largest positive, or largest negative effect size we find that 139 eQTLs have a lead SV. When comparing the largest overall effect size, there are 185 eQTLs with a lead SV, 130 of which are not found in the FDR set.

By plotting the position of the gene, the SVs and the SNPs with their respective FDRs we can visualize the eQTLs around each gene. This can provide insights into how the genetic variants are related and how they impact the gene. We can combine this information with the linkage disequilibrium values. This is important as the most significant eQTL may not be causal, but be in LD with another genetic variant that is. Figures 14 and 15 show an example from ENSG000000184674 (GSTT1), which has a SV as the lead association. When we plot the SNP-only eQTLs for the same gene with LD information we can see that there are three other SNPs in very high LD with the lead SNP, along with other SNPs in high LD as well (Figure 16). However none of these SNPs are in high LD with the lead SV in the joint analysis.

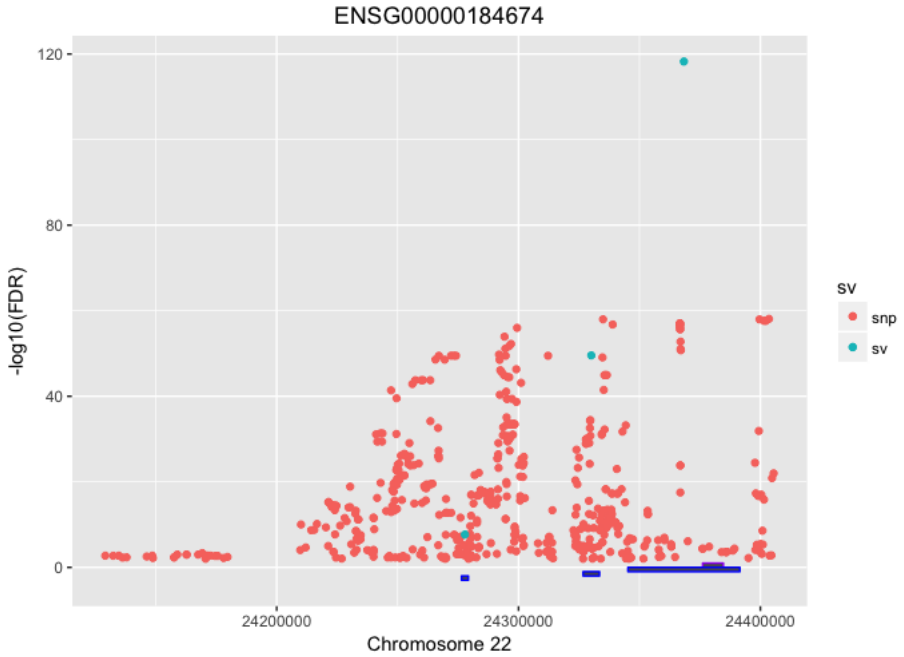


Figure 14: The eQTLs for ENSG00000184674 (GSTT1). The location of the gene is represented by a purple rectangle, and the SVs are blue rectangles. The lead association is a SV which overlaps the gene.

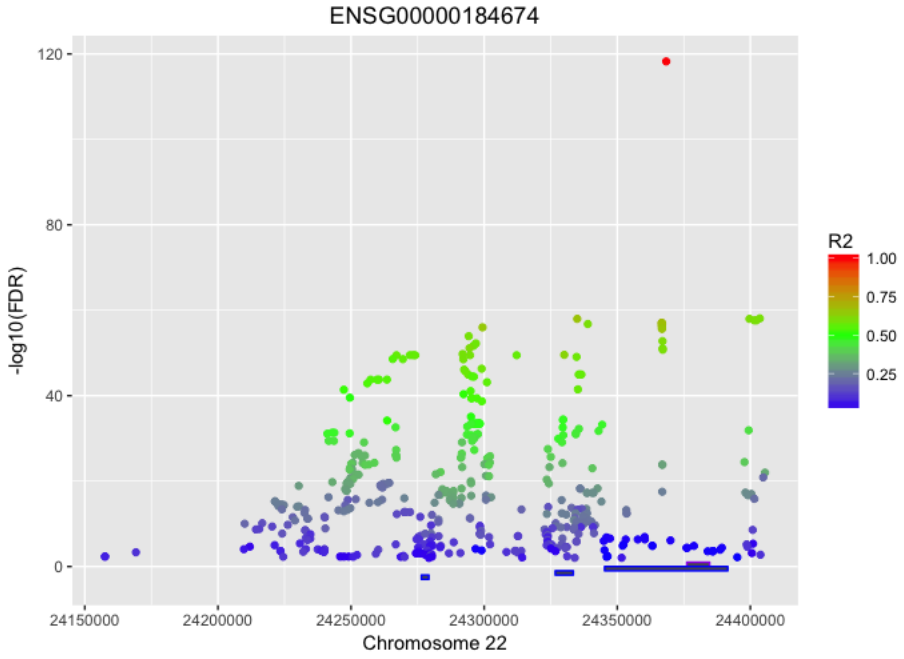


Figure 15: The eQTLs for ENSG00000184674 with linkage disequilibrium values. In this example there aren't other genetic variants that are highly linked to the SV.

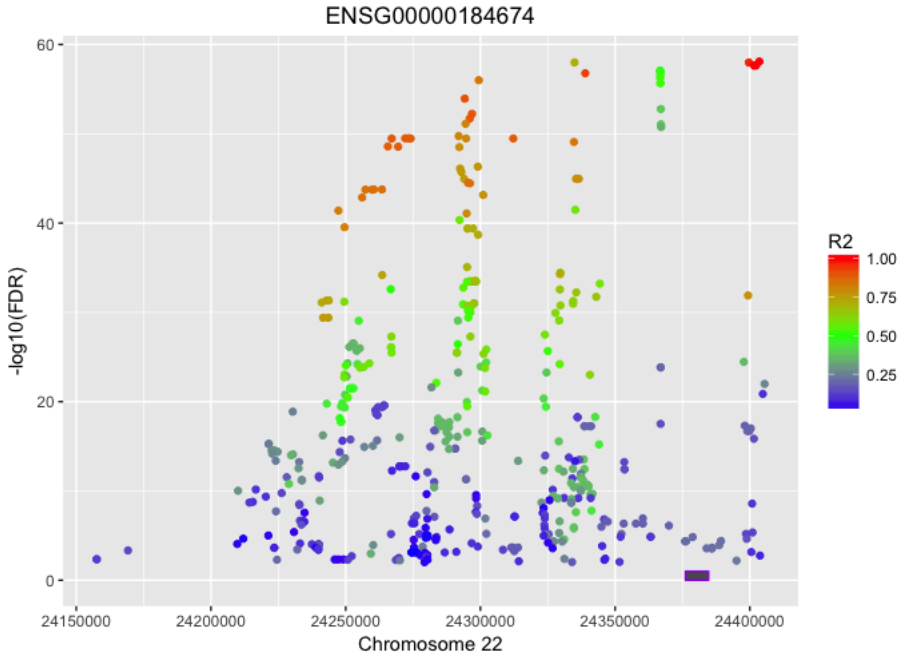


Figure 16: The SNP-only eQTLs for ENSG00000184674 with linkage disequilibrium values. There are several SNPs in high LD with the lead association.

5.3 miRNA eQTL Analysis

We performed SV-only, SNP-only and joint SV/SNP miRNA eQTL analyses. For all analyses we started with 716 miRNAs and 23,723 genes. At a FDR of 0.01 there were 14 miRNA-eQTLs in the SV-only analysis, which included 8 SVs and 8 miRNAs. In the SNP-only analysis there were 8,733 miRNA-eQTLs comprised of 6,309 unique SNPs and 116 unique miRNAs. For the joint SV/SNP analysis there were only 7,059 miRNA-eQTLs with 4,846 genetic variants and 92 miRNAs. We also applied eQTL analysis methods to study the impact of miRNA expression on gene expression. The miRNA expression acts as the X values, and the gene expression are the Y values.

At a FDR cutoff of 0.01 there are 250 miRNA-gene associations from 146 miRNAs and 163 genes, in the *cis*- miRNA-gene association analysis We were able to

find 4,108 triangular associations, like those in Figure 18. Since miRNAs can impact the expression levels of genes, we looked for instances where a SNP or SV is associated with a miRNA and a gene, and the miRNA targets that gene (Figure 18). This could indicate that the eQTL is a representation of miRNA regulation, where a genetic marker influences a miRNA, which in turn influences a gene. While we found 4,108 triangular associations, they only represented 42 distinct miRNA-gene pairs, from 22 unique miRNAs, and 25 unique genes. Figure 17 shows an example of several of these associations for hsa-miR-335-3p. There are 13 SNPs that were involved in an eQTL with the miRNA and also involved in an eQTL with a gene. Some SNPs were associated with multiple genes in the network, and the three genes in the eQTLs were also associated with the miRNA. According to GeneHancer predictions these three genes ($MEST$/ENSG00000106484, $CPA2$/ENSG00000158516 , $CPA4$/ENSG00000128510) share two enhancers, GH07J130362 and GH07J130354 [68].

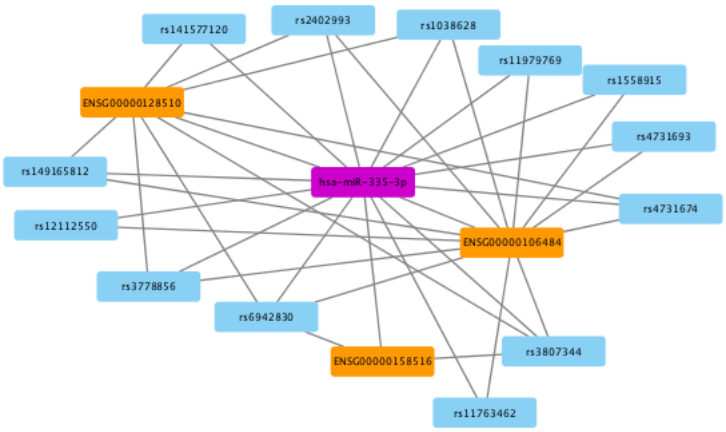


Figure 17: An example of the associations between hsa-miR-335-3p, 13 SNPs and three genes. Each SNP in the network not only is an eQTL for a gene, but also for the miRNA and each gene has an association with the miRNA.

These triangular associations also only represent local influence by the miRNA. Because we were using a *cis-* analysis, by definition the genes and miRNA have to be within a 1MB window of the genetic variant. In order for a triangular association to occur the miRNA expression and gene expression need to be associated with the same genetic variant, which means the maximum distance a miRNA could be from a gene in this analysis is 2MB. miRNA do not simply act in a *cis-* manner, but can interact with genes from all over the genome once they are expressed. To explore this dimension we analyzed instances where a genetic variant impacted a miRNA, and the miRNA impacted a gene. We conducted a *trans-* eQTL analysis between the genetic variants and the genes, as well as a *trans-* type miRNA-gene association analysis. We found 96,532 of the genetic variant-miRNA-gene associations from 2,551 miRNA-gene pairs. However we did not find any triangle associations in the *trans-* analysis. This could suggest that we are being too stringent with our FDR cutoff, or the effects of

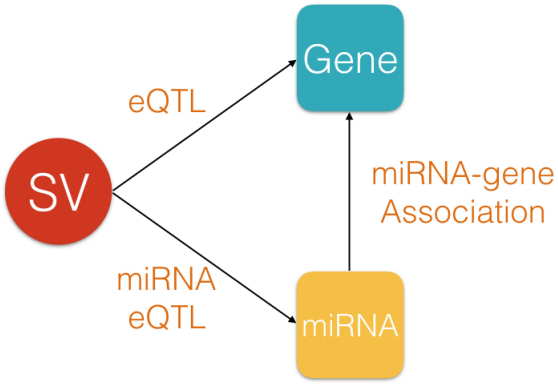


Figure 18: A set of associations between a SV, gene and miRNA. The SV impacts the gene and miRNA directly, and the miRNA impacts the gene.

these associations are too small to be picked out by this analysis, or that they don't exist at all.

Because the number of miRNAs is much smaller than the number of genes, the Manhattan plots for the joint analysis are sparser, but there are still identifiable loci with eQTL peaks (Figures 19 and 20). Unlike the gene eQTLs there aren't any loci that have a SV as the lead association.

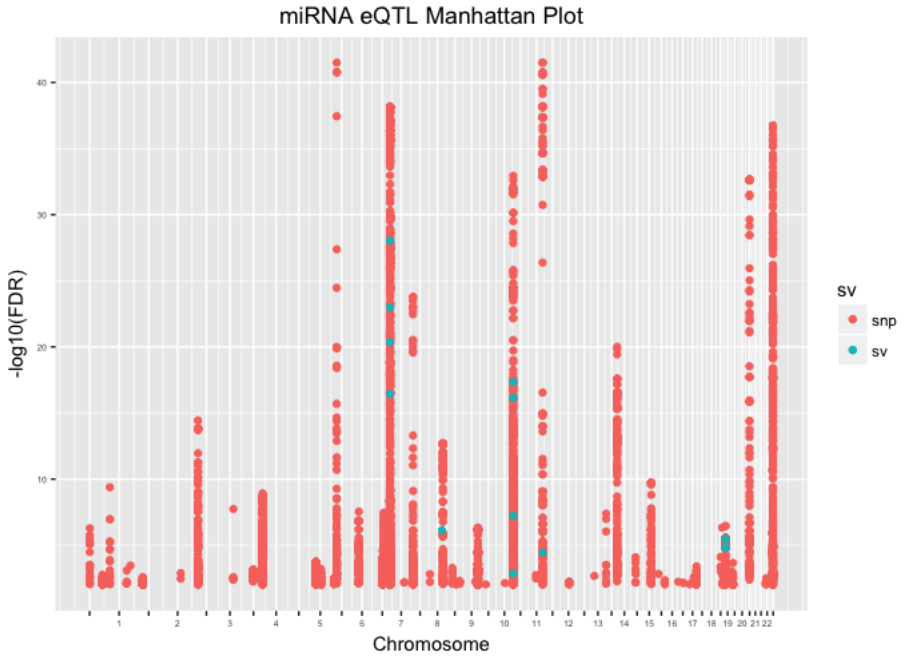


Figure 19: The Manhattan plot for the joint SV/SNP miRNA eQTLs. There are several loci with significant peaks.

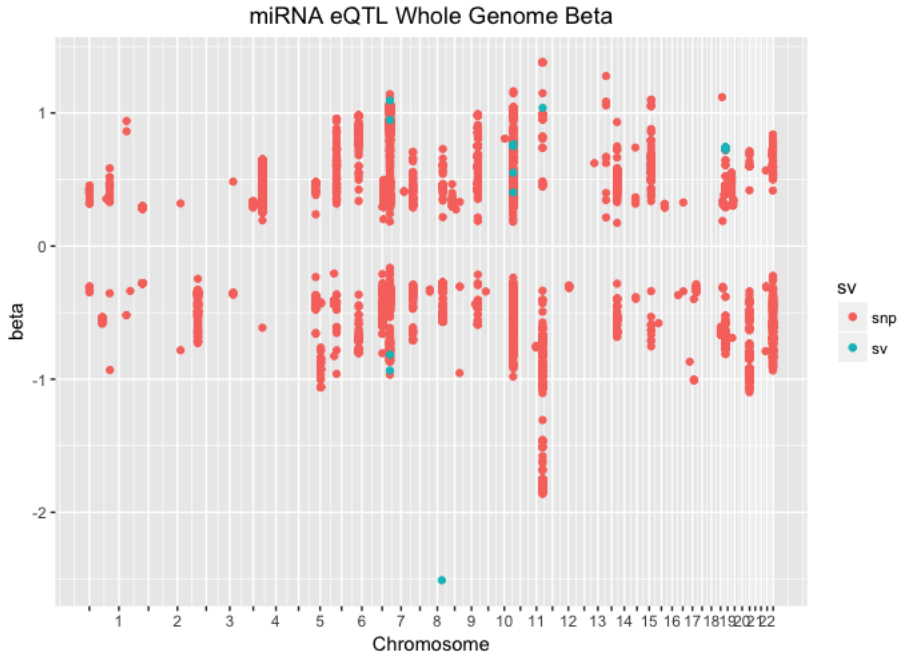


Figure 20: The Manhattan plot for the effect sizes from the SV/SNP miRNA eQTLs.

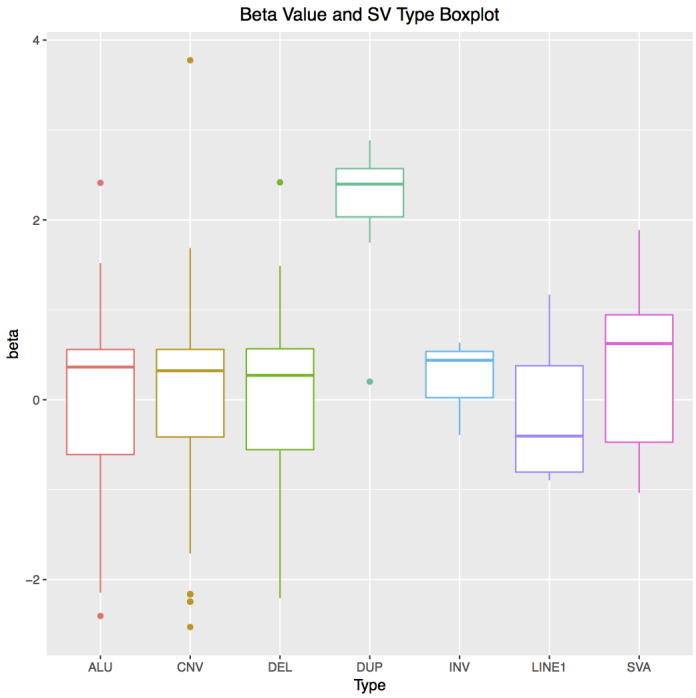


Figure 21: Boxplots of eQTL beta values broken down by SV type. While most of the SV types have both negative and positive beta values, duplications have exclusively positive effect sizes.

5.4 Downstream and Functional Analysis

To study the results from the eQTL analyses we used a variety of downstream analyses.

We compared the impact of a SV on eQTLs based on its type. There are not large differences in the distributions of the FDR by SV type (Figures S6, S7, S8, and S9), but when we look at the effect size distribution a few things become obvious. Duplication eQTLs all have a positive effect, and while ALUs, and deletions are relatively evenly split between negative and positive effect sizes, CNVs have a slight bias towards positive effect sizes (Figures 21 and 22).

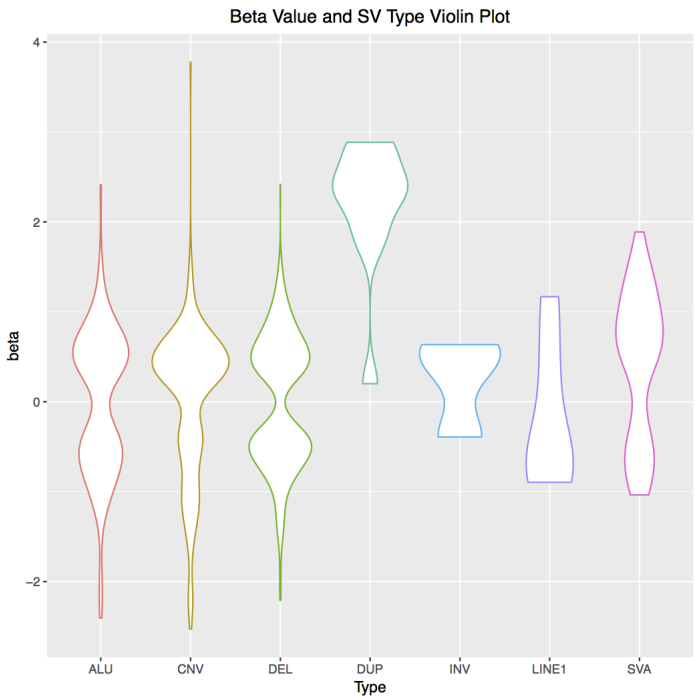


Figure 22: Violin plots of eQTL beta values broken down by SV type. Here we can see that CNV eQTLs have a slight bias towards positive effect sizes.

5.5 SV Examples

While a full examination of each of the SV-eQTLs is too extensive to be included, there are several examples that are of interest and might merit further study. *APOBEC3B* (ENSG00000179750) has been implicated in breast, cervical, lung, and bladder cancer among others. It is believed to be a DNA mutator in cancer, and the over expression of *APOBEC3B* has been consistently found across cancer types [69]. In the joint eQTL analysis, four of the five most significant eQTLs for *APOBEC3B* are due to SVs, and the SNP from the fifth, is in high LD with two of the SVs. Two of the SVs directly impact the gene, YL_CN_STU_4456 overlaps almost the entire gene, and DUP_gs_CNV_22_39359355_39379392 only overlaps the first exon. The other two SVs fall into an intergenic region. BI_GS_DEL1_B1_P2885_301 does not appear to

overlap any regulatory features and is the least significant of the SV eQTLs, but BI_GS_DEL1_B5_P2885_346 overlaps some transcription factor binding sites, as well as some DNAse hypersensativity sites. Aside from the single SNP in LD with two SVs, no other SNP is able to act as a tag SNP, which highlights the importance of including SVs in this sort of analysis. This is of extra importance with $APOBEC3B$, as changes in its expression have been linked with many cancers.

There are four duplications that were found to be SV-eQTLs for $DHTKD1$. While these eQTLs were not the most significant associations, they had the largest effect sizes. All four SVs directly impact the gene. DUP_uwash_chr10_12092804_12171370 is a complete gene duplication, but the others are partial gene duplications (Figure 26). Mutations in $DHTKD1$ have been linked to 2-Aminoadipic and 2-Oxoadipic Aciduria, as well as Charcot-Marie-Tooth Disease Type 2Q. CMT is a progressive disease which leads to muscle loss and loss of touch sensation. $Dhtkd1^{-/-}$ mice had symptoms of CMT. Interestingly mice given feed with 2-Aminoadeipic acid also showed symptoms similar to CMT [70, 71, 72].

$ZFN92$ was associated with seven SVs, all of which are deletions, and all were among the 25 largest negative effect sizes (Figure 27). While only two of the SVs directly overlapped the gene, they were in LD with each other.

DUP_uwash_chr7_64695322_64838588 had the largest negative effect size of all the eQTLs in the joint analysis. It does not overlap $ZFN92$ directly, but lands directly upstream. And while it is in LD with the other nearby SVs, it isn't in high LD. This could indicate that this deletion is impacting the upstream regulatory region, which in turn impacts the gene expression, and that the deletion is not just an association

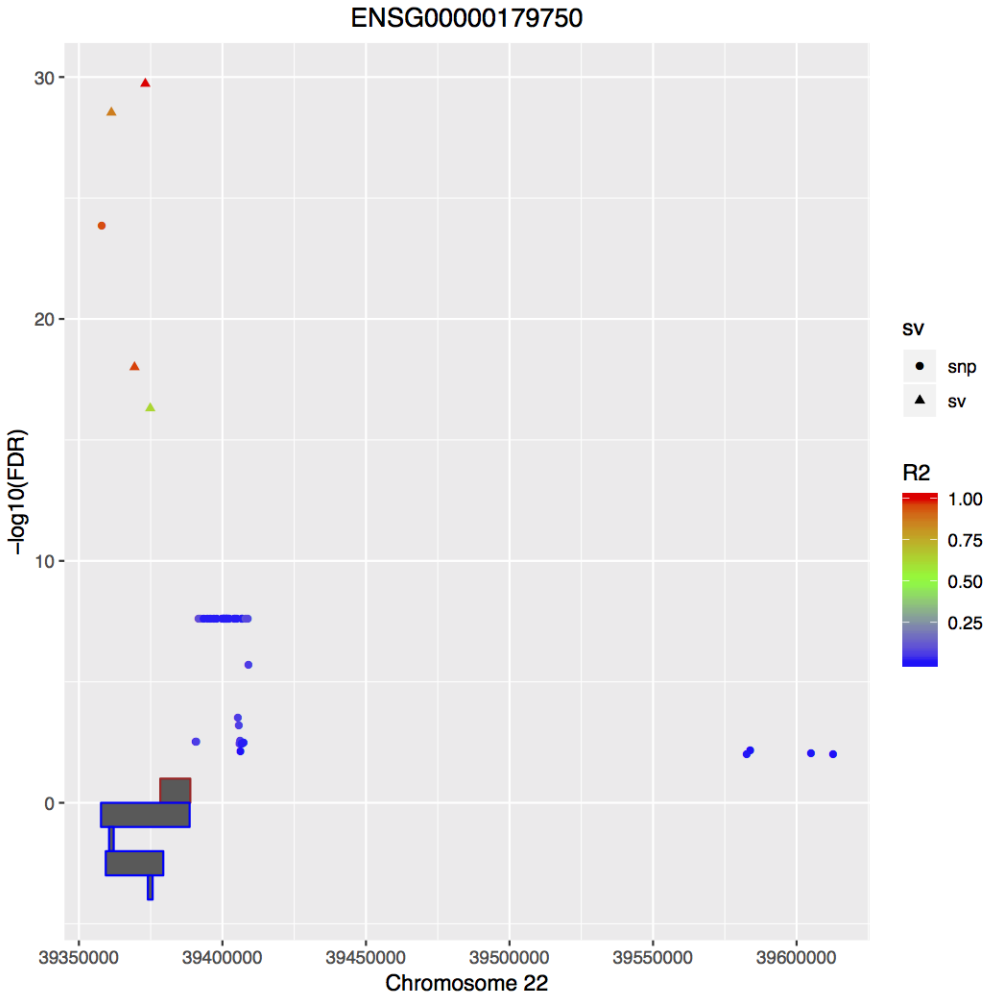


Figure 23: The gene plot for all eQTLs associated with *APOBEC3B*. The four SVs are involved in four of the five most significant eQTLs, and the SNP is in high LD with two of those SVs. No other SNPs are in LD with these SVs, which means that there is no way to use them as tag SNPs. The only way to find these relationships between the SVs and *APOBEC3B* expression is to interrogate them specifically.

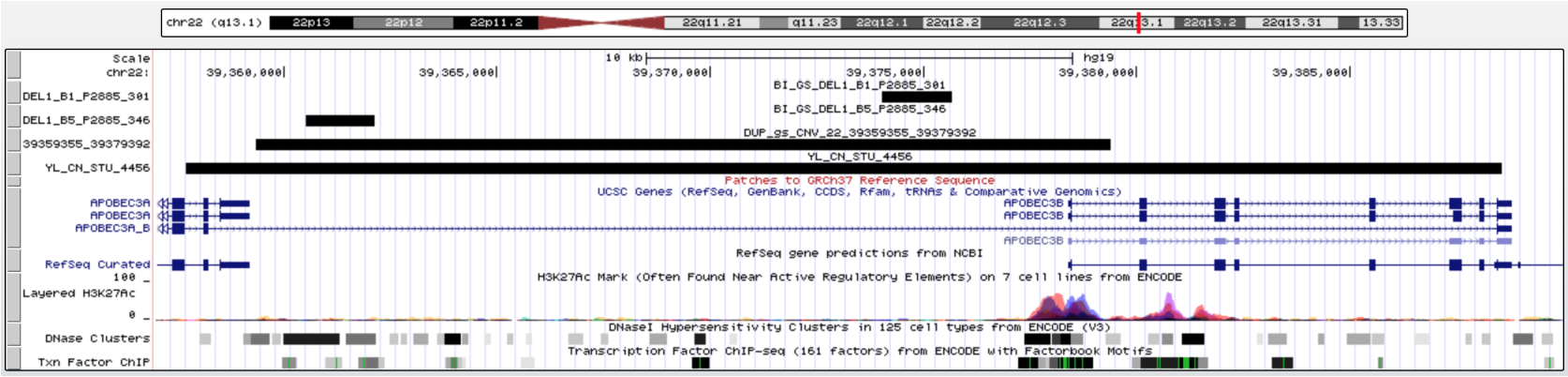


Figure 24: The positions of the *APOBEC3B*, nearby genes, the SVs involved in *APOBEC3B* eQTLs, as well as nearby regulatory features. YL_CN_STU_4456 and DUP_gs_CNV_22_39359355_39379392 overlap the gene, while BI_GS_DEL1_B1_P2885_301 and BI_GS_DEL1_B5_P2885_346 do not.

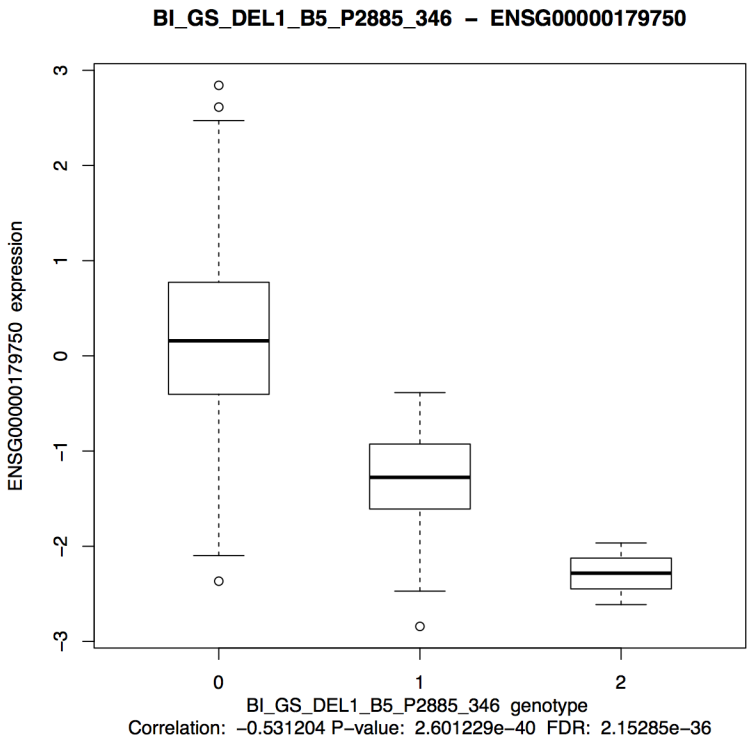


Figure 25: The boxplot for the relationship between BI_GS_DEL1_B5_P2885_346 genotype and *APOBEC3B* expression. One deletion reduces gene expression, but a homozygous deletion of the regulatory region reduces expression even more.

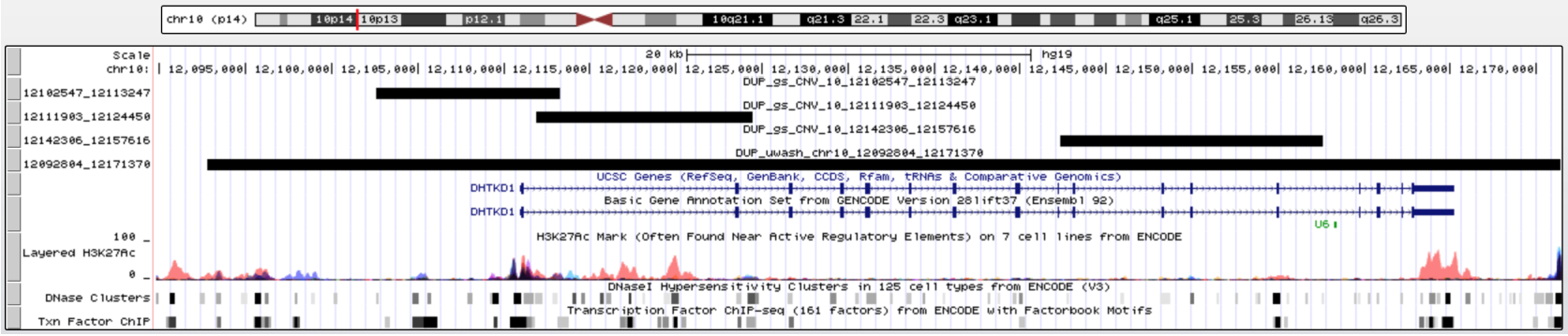


Figure 26: *DHTKD*1 has four SV associated with it in our results. These SV-eQTLs have some of the largest positive effect sizes of all the eQTLs. All four SVs, which are duplications, overlap the gene body in someway. Each of duplications is associated with an increase in gene expression.

because it is in LD with a deletion that impacts the coding region.

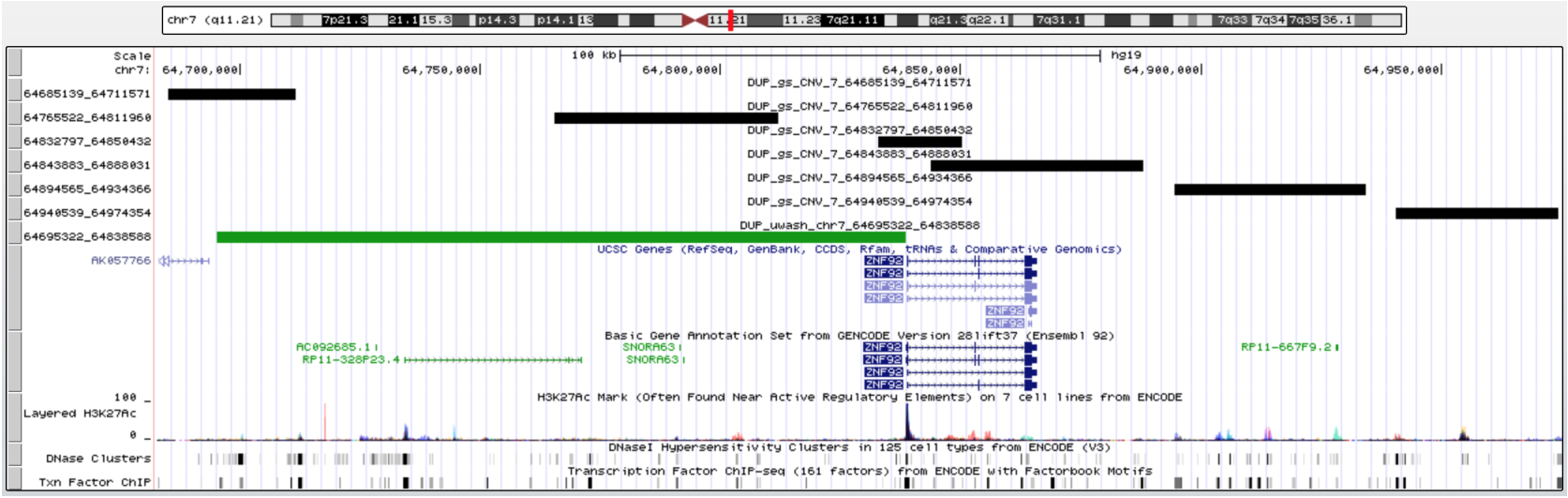


Figure 27: *ZFN92* has seven SVs associated with it in our results. Each SV is a deletion, and the related eQTLs have some of the most negative effect sizes of all the eQTLs. While only two of the SVs overlap the gene body, four of the non-overlapping SVs are in high LD with them.

CHAPTER 6: CONCLUSION

This proposal is broken down into two main projects, the eQTL pipeline and the 1000 Genomes eQTL analyses. The 1000 Genomes project is composed of sub-projects including a joint SV/SNP eQTL analysis, miRNA eQTL analysis. By combining multiple data types we can analyze the impact of genetic variants on gene expression and also disease phenotypes.

6.1 Discussion

The eQTL pipeline software provides not only an easy way to start performing eQTL analysis, but a set of tools to thoroughly customize them. The user can go from start to finish in one command, or they can use the component parts as individual tools. The pipeline not only provides eQTL results, but also produces visualizations that can be used to explore the results, or in reports and publications. Often eQTL analysis tools, like MatrixEQTL, only perform the eQTL analysis itself, and don't provide comprehensive start to finish support. This leads to users having to develop their own supporting software, which in turn leads to a replication of effort. The eQTL pipeline described in this dissertation attempts to provide a solution for these problems.

The impact of large alterations to the genome in the form of structural variants can be seen using eQTL analysis. SNPs are much more numerous than SVs, and while

the SVs aren't often the lead association they tend to have larger effect size. Because of the underlying genetic structure these SV eQTLs often aren't able to be discovered using SNPs alone. Many of the SVs involved in eQTLs don't directly overlap with the eGene, but are downstream or upstream of the gene. This indicates that the effects we find from these SV eQTLs are due to changes in the regulatory regions, or by disruption of another gene.

6.2 Future Work

The deep learning gene expression predictions should be explored further. Another analysis can be done with an expanded sample size, and using human data. While this will be more computationally intensive, it is probably closer to an actual use case. The neural network could then be expanded further to include epigentic markers and other regulatory features.

The eQTL pipeline can also be expanded further. A permutation analysis function could be added, as well as more visualizations, like the gene based plots in this dissertation. Another eQTL analysis engine, like FastQTL could be added, to complement MatrixEQTL. Instead of having the user specify the PEER factors, a function could be added to set the number of factors based on the sample size according to the GTEx example.

As more data is generated using the 1000 Genomes samples it could be incorporated into the eQTL analysis. The eQTL analysis could be expanded to include a pQTL analysis, to study how the changes in gene expression impact protein expression levels. Currently we use genotypes, gene expression, and microRNA expression, but could

add methylation, Hi-C data and others.

REFERENCES

[1] Lu Tian, Andrew Quitadamo, Frederick Lin, and Xinghua Shi. Methods for population-based eqtl analysis in human genetics. *Tsinghua Science and Technology*, 19(6):624–634, 2014.

[2] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics*, 24(8):408–415, 2008.

[3] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature reviews. Genetics*, 16(4):197, 2015.

[4] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768, 2010.

[5] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, et al. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*, 24(1):14–24, 2014.

[6] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[7] Daniel J Gaffney, Jean-Baptiste Veyrieras, Jacob F Degner, Roger Pique-Regi, Athma A Pai, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Dissecting the regulatory architecture of gene expression qtls. *Genome biology*, 13(1):R7, 2012.

[8] Jean-Baptiste Veyrieras, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K Pritchard. High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS genetics*, 4(10):e1000214, 2008.

[9] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS genetics*, 8(4):e1002639, 2012.

[10] Jun Ding, Johann E Gudjonsson, Liming Liang, Philip E Stuart, Yun Li, Wei Chen, Michael Weichenthal, Eva Ellinghaus, Andre Franke, William Cookson, et al. Gene expression in skin and lymphoblastoid cells: Refined statistical

method reveals extensive overlap in cis-eqtl signals. *The American Journal of Human Genetics*, 87(6):779–789, 2010.

[11] Amy L Stark, Ronald J Hause Jr, Lidija K Gorsic, Nirav N Antao, Shan S Wong, Sophie H Chung, Daniel F Gill, Hae K Im, Jamie L Myers, Kevin P White, et al. Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy. *PLoS genetics*, 10(4):e1004192, 2014.

[12] Ronald J Hause, Amy L Stark, Nirav N Antao, Lidija K Gorsic, Sophie H Chung, Christopher D Brown, Shan S Wong, Daniel F Gill, Jamie L Myers, Lida Anita To, et al. Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. *The American Journal of Human Genetics*, 95(2):194–208, 2014.

[13] Linfeng Wu, Sophie I Candille, Yoonha Choi, Dan Xie, Jennifer Li-Pook-Than, Hua Tang, and Michael Snyder. Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79, 2013.

[14] Zia Khan, Michael J Ford, Darren A Cusanovich, Amy Mitrano, Jonathan K Pritchard, and Yoav Gilad. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, 342(6162):1100–1104, 2013.

[15] Tamar Geiger, Anja Wehner, Christoph Schaab, Juergen Cox, and Matthias Mann. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & Cellular Proteomics*, 11(3):M111–014050, 2012.

[16] Alexis Battle, Zia Khan, Sidney H Wang, Amy Mitrano, Michael J Ford, Jonathan K Pritchard, and Yoav Gilad. Impact of regulatory variation from rna to protein. *Science*, 347(6222):664–667, 2015.

[17] Can Cenik, Elif Sarinay Cenik, Gun W Byeon, Fabian Grubert, Sophie I Candille, Damek Spacek, Bilal Alsallakh, Hagen Tilgner, Carlos L Araya, Hua Tang, et al. Integrative analysis of rna, translation, and protein levels reveals distinct regulatory variation across humans. *Genome research*, 25(11):1610–1621, 2015.

[18] Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. A study of cnvs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS genetics*, 7(2):e1001292, 2011.

[19] Barbara E Stranger, Matthew S Forrest, Mark Dunning, Catherine E Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P Bird, Anna De Grassi, Charles Lee, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[20] Steven A McCarroll, Tracy N Hadnott, George H Perry, Pardis C Sabeti, Michael C Zody, Jeffrey C Barrett, Stephanie Dallaire, Stacey B Gabriel, Charles

Lee, Mark J Daly, David M Altshuler, and The International HapMap Consortium. Common deletion polymorphisms in the human genome. *Nature Genetics*, 38:86 EP –, 12 2005.

[21] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[22] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75, 2015.

[23] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376, 2011.

[24] Haley J Abel and Eric J Duncavage. Detection of structural dna variation from next generation sequencing data: a review of informatic approaches. *Cancer genetics*, 206(12):432–440, 2013.

[25] Peiyong Guan and Wing-Kin Sung. Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods*, 102:36–49, 2016.

[26] Geòrgia Escaramís, Elisa Docampo, and Raquel Rabionet. A decade of structural variants: description, history and methods to detect structural variation. *Briefings in functional genomics*, 14(5):305–314, 2015.

[27] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, et al. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681, 2009.

[28] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[29] Alexej Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein. Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, 21(6):974–984, 2011.

[30] Fereydoun Hormozdiari, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, 19(7):1270–1278, 2009.

[31] Robert E Handsaker, Joshua M Korn, James Nemesh, and Steven A McCarroll. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, 43(3):269–276, 2011.

[32] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

[33] Eugene J Gardner, Vincent K Lam, Daniel N Harris, Nelson T Chuang, Emma C Scott, W Stephen Pittard, Ryan E Mills, Scott E Devine, 1000 Genomes Project Consortium, et al. The mobile element locator tool (melt): population-scale mobile element discovery and biology. *Genome research*, 27(11):1916–1929, 2017.

[34] Gargi Dayama, Sarah B Emery, Jeffrey M Kidd, and Ryan E Mills. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic acids research*, 42(20):12640–12649, 2014.

[35] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. Lumpy: a probabilistic framework for structural variant discovery. *Genome biology*, 15(6):R84, 2014.

[36] A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949, 2004.

[37] Jonathan Sebat, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, Hillary Massa, Megan Walker, Maoyen Chi, et al. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, 2004.

[38] Matthew E Hurles, Emmanouil T Dermitzakis, and Chris Tyler-Smith. The functional impact of structural variation in humans. *Trends in Genetics*, 24(5):238–245, 2008.

[39] Benjamin J Raphael. Structural variation and medical genomics. *PLoS computational biology*, 8(12):e1002821, 2012.

[40] Joachim Weischenfeldt, Orsolya Symmons, Francois Spitz, and Jan O Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138, 2013.

[41] Josie Hayes, Pier Paolo Peruzzi, and Sean Lawler. Micrornas in cancer: biomarkers, functions and therapy. *Trends in molecular medicine*, 20(8):460–469, 2014.

[42] Lyudmila F Gulyaeva and Nicolay E Kushlinskiy. Regulatory mechanisms of microrna expression. *Journal of translational medicine*, 14(1):143, 2016.

[43] Benika Hall, Andrew Quitadamo, and Xinghua Shi. Identifying microrna and gene expression networks using graph communities. *Tsinghua Science and Technology*, 21(2):176–195, 2016.

[44] Gonçalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlinrapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97, 2002.

[45] Karl W Broman, Hao Wu, Śaunak Sen, and Gary A Churchill. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.

[46] Danny Arends, Pjotr Prins, Ritsert C Jansen, and Karl W Broman. R/qtl: high-throughput multiple qtl mapping. *Bioinformatics*, 26(23):2990–2992, 2010.

[47] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[48] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[49] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2015.

[50] Xiaohui Chen, Xinghua Shi, Xing Xu, Zhiyong Wang, Ryan Mills, Charles Lee, and Jinbo Xu. A two-graph guided multi-task lasso approach for eqtl mapping. In *International Conference on Artificial Intelligence and Statistics*, pages 208–217, 2012.

[51] Lingxue Zhang and Seyoung Kim. Learning gene networks under snp perturbations using eqtl datasets. *PLoS computational biology*, 10(2):e1003420, 2014.

[52] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 2008.

[53] Nicoló Fusi, Oliver Stegle, and Neil D Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS computational biology*, 8(1):e1002330, 2012.

[54] Lingxue Zhang and Seyoung Kim. Learning Gene Networks under SNP Perturbations Using eQTL Datasets. *PLOS Computational Biology*, 10(2):e1003420, February 2014.

[55] Wei Cheng, Yu Shi, Xiang Zhang, and Wei Wang. Sparse regression models for unraveling group and individual associations in eQTL mapping. *BMC Bioinformatics*, 17, March 2016.

[56] James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[57] Andrew Quitadamo, Lu Tian, Benika Hall, and Xinghua Shi. An integrated network of microrna and gene expression in ovarian cancer. *BMC bioinformatics*, 16(5):S5, 2015.

[58] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[59] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[60] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.

[61] Gregory D Miles, Michael Seiler, Lorna Rodriguez, Gunaretnam Rajagopal, and Gyan Bhanot. Identifying microrna/mrna dysregulations in ovarian cancer. *BMC research notes*, 5(1):164, 2012.

[62] Lin Zhang, Stefano Volinia, Tomas Bonome, George Adrian Calin, Joel Greshock, Nuo Yang, Chang-Gong Liu, Antonis Giannakakis, Pangiotis Alexiou, Kosei Hasegawa, et al. Genomic and epigenetic alterations deregulate microrna expression in human epithelial ovarian cancer. *Proceedings of the National Academy of Sciences*, 105(19):7004–7009, 2008.

[63] N Howlader, AM Noone, M Krapcho, D Miller, K Bishop, CL Kosary, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, and KA Cronin. Seer cancer statistics review, 1975–2014. 2017. *National Cancer Institute: Bethesda, MD*, 2014.

[64] Dimitra Karagkouni, Maria D Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, Giorgos Skoufos, et al. Diana-tarbase v8: a decadelong collection of experimentally supported mirna–gene interactions. *Nucleic acids research*, 46(D1):D239–D245, 2017.

[65] Elizabeth J Rossin, Kasper Lage, Soumya Raychaudhuri, Ramnik J Xavier, Diana Tatar, Yair Benita, Chris Cotsapas, Mark J Daly, International Inflammatory Bowel Disease Genetics Constortium, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS genetics*, 7(1):e1001273, 2011.

[66] Rui Xie, Jia Wen, Andrew Quitadamo, Jianlin Cheng, and Xinghua Shi. A deep auto-encoder model for gene expression prediction. *BMC genomics*, 18(9):845, 2017.

[67] Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5):1572–1577, 2005.

[68] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database*, 2017(1), 2017.

[69] Jun Zou, Chen Wang, Xiangyi Ma, Edward Wang, and Guang Peng. Apobec3b, a molecular driver of mutagenesis in human cancers. *Cell & bioscience*, 7(1):29, 2017.

[70] Katharina Danhauser, Sven W Sauer, Tobias B Haack, Thomas Wieland, Christian Staufner, Elisabeth Graf, Johannes Zschocke, Tim M Strom, Thorsten Traub, Jürgen G Okun, et al. Dhtkd1 mutations cause 2-aminoadipic and 2-oxoadipic aciduria. *The American Journal of Human Genetics*, 91(6):1082–1087, 2012.

[71] Wang-yang Xu, Ming-min Gu, Lian-hua Sun, Wen-ting Guo, Hou-bao Zhu, Jianfang Ma, Wen-tao Yuan, Ying Kuang, Bao-jun Ji, Xiao-lin Wu, et al. A nonsense mutation in dhtkd1 causes charcot-marie-tooth disease type 2 in a large chinese pedigree. *The American Journal of Human Genetics*, 91(6):1088–1094, 2012.

[72] Wang-Yang Xu, Houbao Zhu, Yan Shen, Ying-Han Wan, Xiao-Die Tu, Wen-Ting Wu, Lingyun Tang, Hong-Xin Zhang, Shun-Yuan Lu, Xiao-Long Jin, et al. Dhtkd1 deficiency causes charcot-marie-tooth disease in mice. *Molecular and cellular biology*, pages MCB–00085, 2018.

Appendix A: Supplementary Figures

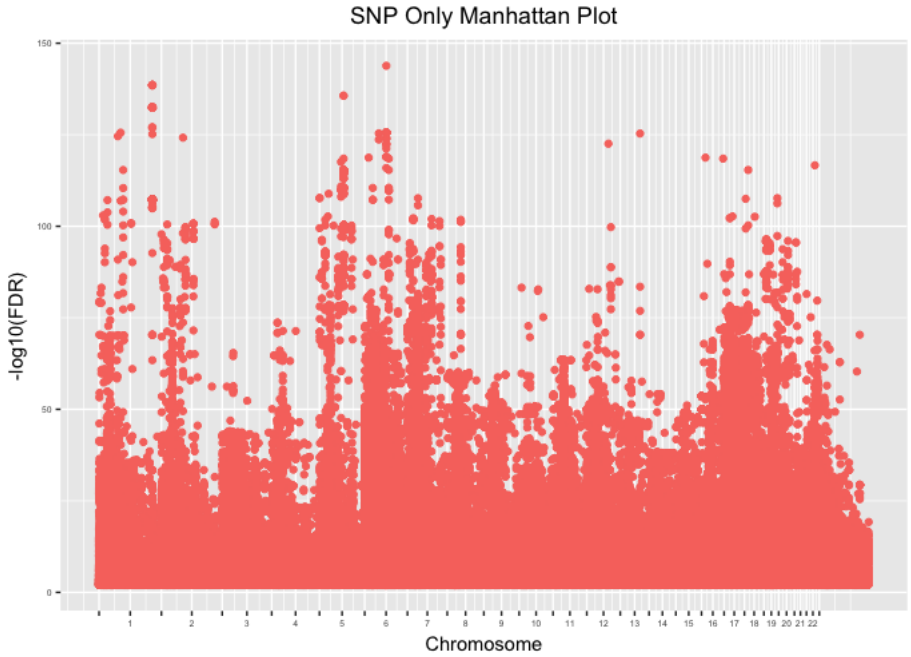


Figure S1: The manhattan plot for the SNP-only analysis. Overall it is similar to that from the joint analysis.

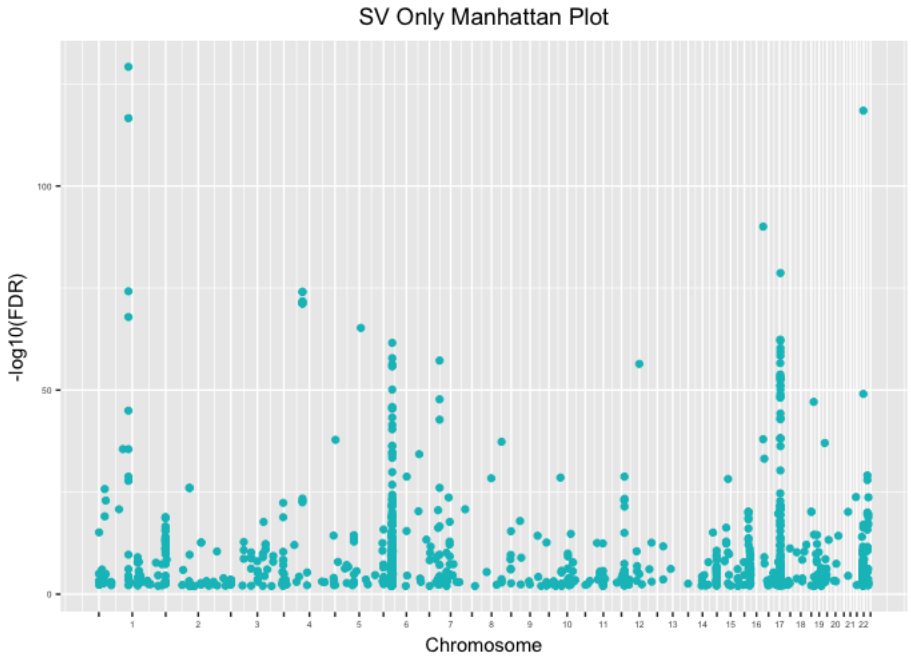


Figure S2: The manhattan plot for the SV-only analysis. While it is much sparser than that from the SNP-only analysis, there are still discernible loci where SVs are very significant.

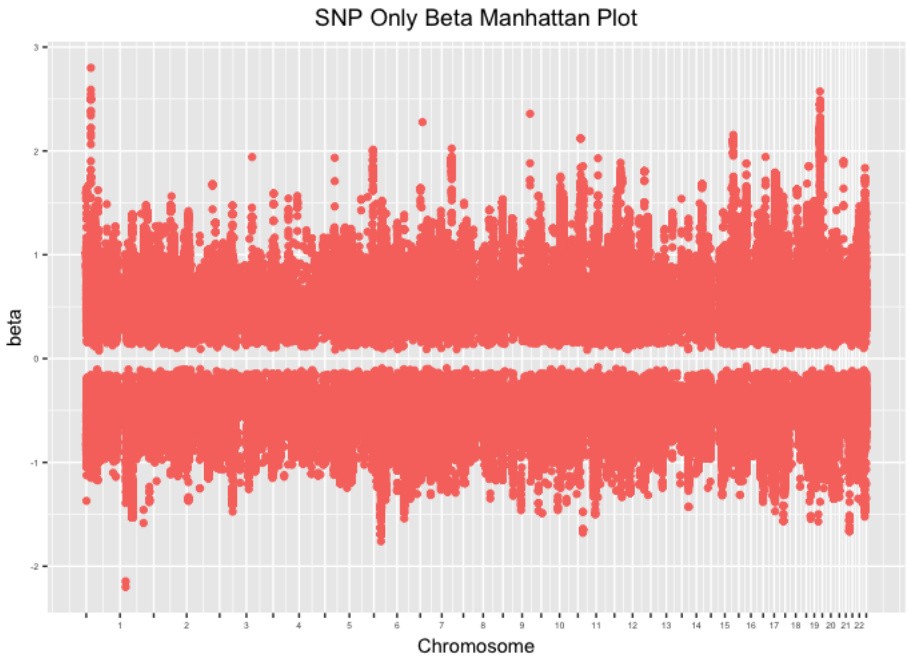


Figure S3: The manhattan plot for the effect sizes from the SNP-only analysis. Like the FDR manhattan plot it is similar to that from the joint analysis.

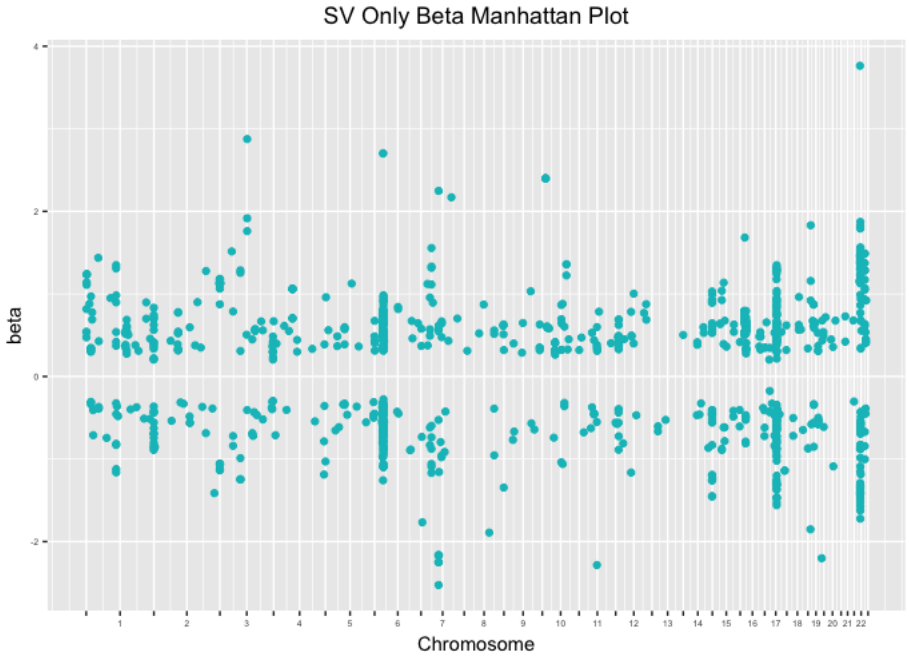


Figure S4: The manhattan plot for the effect sizes from the SV-only analysis.

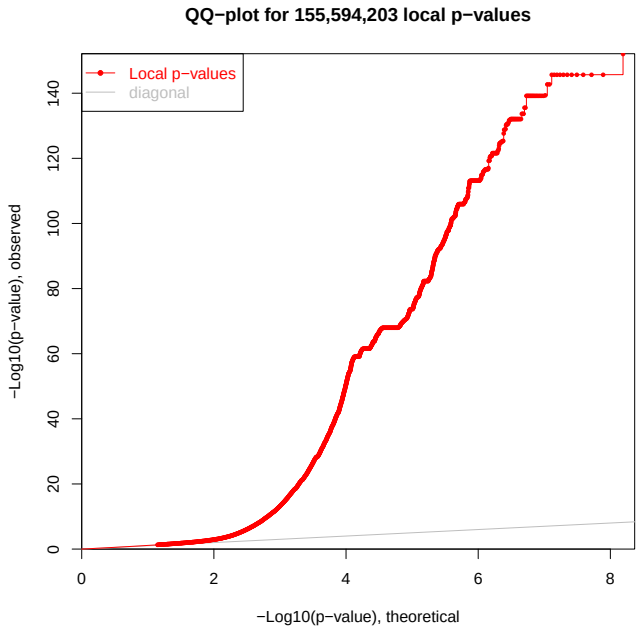


Figure S5: The qq-plot for the joint SV/SNP analysis. The horizontal runs are due to SNPs in perfect LD that share the same significance.

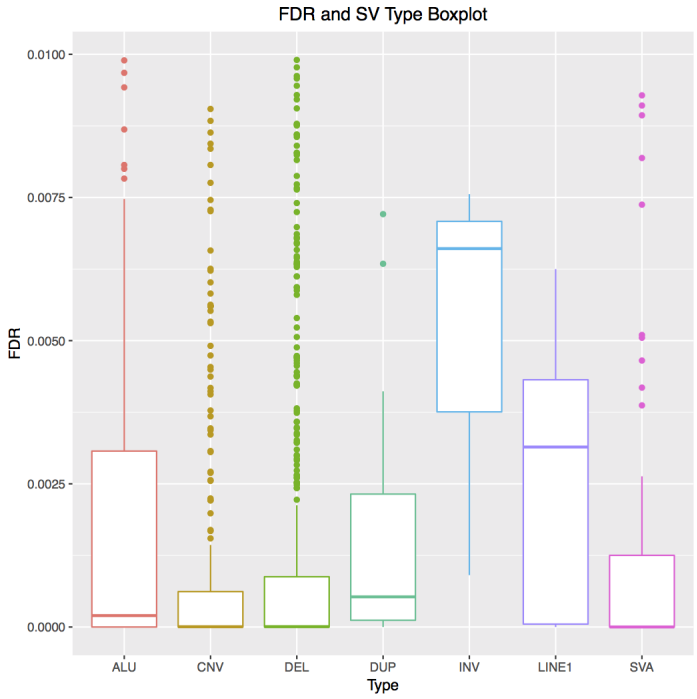


Figure S6: Boxplots of eQTL FDR values broken down by SV type.

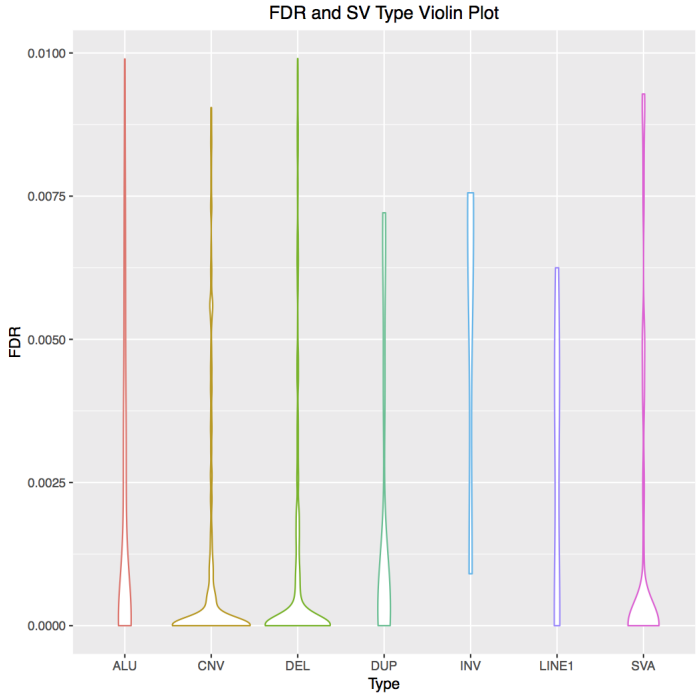


Figure S7: Violin plots of eQTL FDR values by SV type.

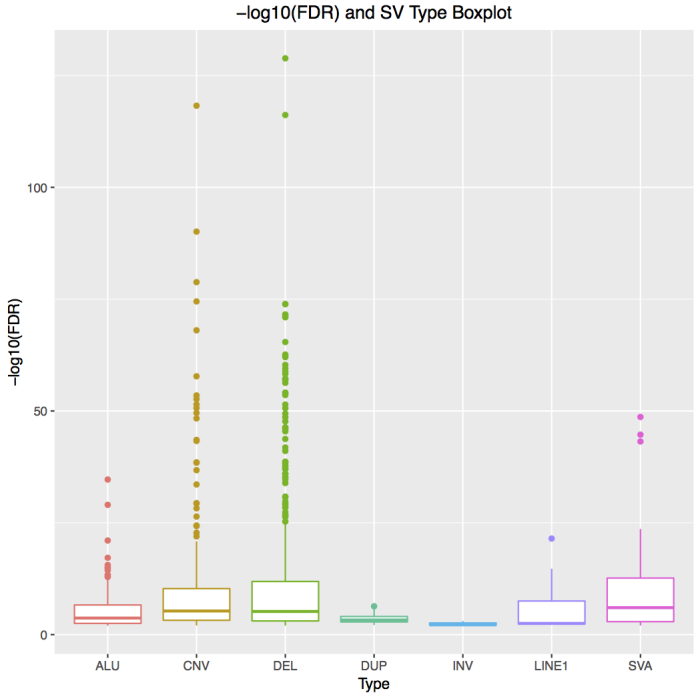


Figure S8: Boxplots of -log10(FDR) by SV type. Deletions and CNVs have more highly significant eQTLs than other types.

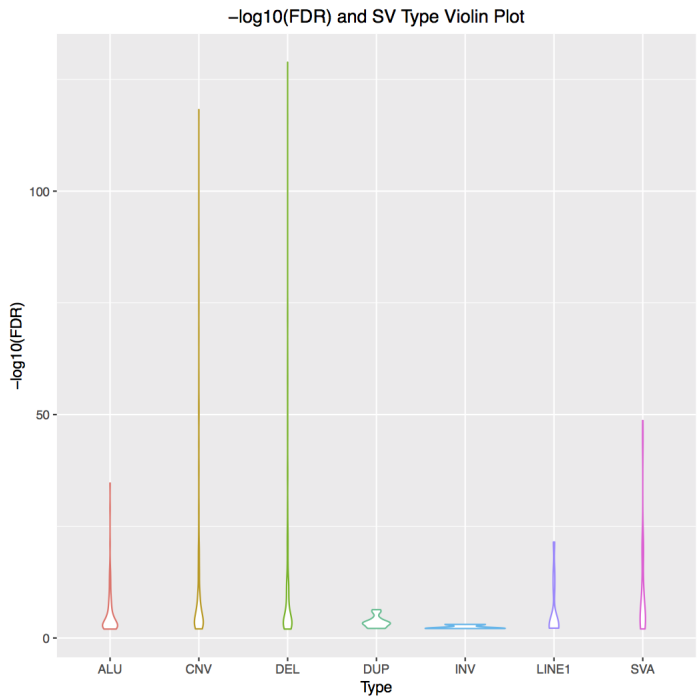


Figure S9: Violin plots of -log10(FDR) by SV type.

Appendix B: Published Work

Title	Authors	Publication	Year
A microRNA-Gene Network in Ovarian Cancer from Genome-Wide QTL Analysis	Quitadamo, Andrew; Lin, Frederick; Tian, Lu; Shi, Xinghua;	Bioinformatics Research and Applications: 10th International Symposium, ISBRA 2014, Zhangjiajie, China, June 28-30, 2014, Proceedings	2014
Methods for Population-Based eQTL Analysis in Human Genetics	Tian, Lu; Quitadamo, Andrew; Lin, Frederick; Shi, Xinghua;	Tsinghua Science and Technology	2014
An Integrated Network of microRNA and Gene Expression in Ovarian Cancer	Quitadamo, Andrew; Tian, Lu; Hall, Benika; Shi, Xinghua;	BMC Bioinformatics	2015
A Graph Community Approach for Constructing microRNA Networks	Hall, Benika; Quitadamo, Andrew; Shi, Xinghua;	International Conference on Big Data Computing and Communications	2015
A Phylogeny of Sand Flies (Diptera: Psychodidae: Phlebotominae), using recent Ethiopian collections and a broad selection of publicly available DNA sequence data	GraceLema, Danielle M; Yared, Solomon; Quitadamo, Andrew; Janies, Daniel A; Wheeler, Ward C; Balkew, Meshesha; Hailu, Asrat; Warburg, Alon; Clouse, Ronald M;	Systematic Entomology	2015
An Integrated Map of Structural Variation in 2,504 Human Genomes	Sudmant, Peter H; Rausch, Tobias; Gardner, Eugene J; Handsaker, Robert E; Abyzov, Alexej; Huddleston, John; Zhang, Yan; Ye, Kai; Jun, Goo; Fritz, Markus Hsi-Yang;	Nature	2015
A Global Reference For Human Genetic Variation	1000 Genomes Project Consortium;	Nature	2015
Identifying microRNA and Gene Expression Networks Using Graph Communities	Hall, Benika; Quitadamo, Andrew; Shi, Xinghua;	Tsinghua Science and Technology	2016
Analyzing microRNA Epistasis in Colon Cancer Using Empirical Bayesian Elastic Nets	Wen, Jia; Quitadamo, Andrew; Hall, Benika;	Bioinformatics Research and Applications: 12th International Symposium, ISBRA 2016, Minsk, Belarus, June 5-8, 2016, Proceedings	2016
A Predictive Model of Gene Expression Using a Deep Learning Framework	Xie, Rui; Quitadamo, Andrew; Cheng, Jianlin; Shi, Xinghua;	Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on	2016
Epistasis Analysis of microRNAs on Pathological Stages in Colon Cancer Based on an Empirical Bayesian Elastic Net Method	Wen, Jia; Quitadamo, Andrew; Hall, Benika; Shi, Xinghua;	BMC Genomics	2017