

# Functional annotation of imputed eQTL loci using different molecular data sources

---

Groningen, December 2013

The Netherlands

Matthieu Beukers (BFV-4)

[matthieu.beukers@gmx.com](mailto:matthieu.beukers@gmx.com)

Student number: 000332373

Hanze University of Applied Sciences, Groningen, the Netherlands

December 4, 2013

Project supervisor: Patrick Deelen

Second project supervisor: Morris Swertz, Harm-Jan Westra, Lude Franke

Supervisor Hanze: Arne Poortinga

## Acknowledgements

Patrick Deelen

Morris Swertz

Genomics Coordination Center

Lude Franke

Harm-Jan Westra

Cisca Wijmenga

## Abstract

Expression Quantitative Trait Loci (eQTLs) are genetic variants that influence gene expression levels in various ways. There are two types of eQTLs, *cis*-acting and *trans*-acting eQTLs. *Cis*-eQTL SNPs are located close to the genes they influence and are easier to identify than *trans*-acting eQTLs that regulate genes from more distant locations, sometimes even from other chromosomes. eQTLs are identified by correlating genotypes, generally SNPs, with gene expression levels measured with gene expression probes.

During this project, a *cis*-eQTL mapping has been performed on non-imputed and imputed blood genotype data containing 1240 samples. The effect of imputation on the eQTL mapping process was determined by comparing the most significant eQTLs per shared gene expression probe and performing annotations using RegulomeDB. These analyses showed that imputation has a positive effect on the eQTL mapping process because more eQTLs effect were detected after imputation, their significance increased, and they showed a better annotation using RegulomeDB, compared to a not imputed dataset.

With the results of the imputed dataset, enrichments on various data sources were performed. First, an enrichment analysis was performed on transcription factor binding sites using RegulomeDB. In total, binding sites for six transcription factors were significantly enriched for eQTL loci. The binding sites of three transcription factors were significantly depleted for eQTL loci. The fact that transcription factor binding sites are enriched is not surprising but does not fully explain entirely how eQTLs influence gene expression. An enrichment analysis was also performed on GENCODE annotation data which showed that eQTLs were significantly enriched for protein coding and retained intron transcripts whereas pseudogenes and antisense transcripts were significantly depleted. eQTLs are known to be enriched for protein coding transcripts. The observation that eQTLs were depleted for antisense transcripts was somewhat surprising, as they regulate the gene they originate from. Enrichments on repeat regions and retrotransposons yield no results indicating that eQTLs do not regulate gene expression by means of retrotransposons. No enrichments were found for histone marks. After performing the enrichment analyses, independent eQTL effects were identified by iteratively regressing out top effects from previous mapping. This allowed more eQTL effects to be identified.

Overall, imputation improves the identification and functional annotation of eQTLs. *Cis*-eQTLs are strongly enriched for transcription factor binding sites and protein coding and retained intron transcripts. *Cis*-eQTLs are not found to be enriched in retrotransposons or histone marks.

## List of Abbreviations

dbRIP	Database of Retrotransposon Insertion Polymorphisms
DNA	Deoxyribonucleic Acid
eQTL	Expression Quantative Trait Locus
FDR	False Discovery Rate
GoNL	Genome of the Netherlands
GWAS	Genome Wide Association Study
HLA	Human Leukocyte Antigen
LD	Linkage Disequilibrium
lincRNA	Long intergenic noncoding Ribonucleic Acid
LINE	Long Interspersed Element
miRNA	Micro Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
PC	Principal Component
PCA	Principal Component Analysis
P-Value	Probability Value
QTL	Quantative Trait Locus
SINE	Short Interspersed Element
siRNA	Small interference Ribonucleic Acid
SNP	Single Nucleotide Polymorphisms
SSR	Simple Sequence Repeat
SVA	SINE VNTR Alus
tRNA	Transfer Ribonucleic Acid
TSS	Transcription Start Site
UCSC	University of California, Santa Cruz

## Table of Contents

Acknowledgements .....	2
Abstract .....	3
List of Abbreviations .....	4
Introduction .....	7
Project goals .....	7
Theory .....	9
Regulation of gene expression .....	9
Single Nucleotide Polymorphisms .....	9
Linkage disequilibrium .....	10
Imputation .....	10
Materials & Methods .....	11
Materials .....	11
Gene expression data .....	11
Genotype data .....	11
Genome of the Netherlands imputed genotype data .....	11
Permutation data .....	11
RegulomeDB .....	11
Gencode Annotation data .....	12
Retrotransposon data .....	12
DBRPIP data .....	12
ENCODE Histone mark data .....	12
Methods .....	12
Genome build conversion .....	12
eQTL Mapping .....	13
Effect of imputation .....	13
Enrichment analyses .....	13
Transcription factor binding site enrichment .....	14
Transcript enrichment .....	14
Retrotransposon enrichment .....	14
Histone mark enrichment .....	14
Identifying independent eQTL effects .....	15
Second transcription factor binding site enrichment .....	15
Results .....	16

eQTL Mapping .....	16
Effect of imputation .....	16
Transcription factor binding site enrichment.....	19
Transcript enrichment.....	23
Retrotransposon enrichment .....	23
Histone mark enrichment .....	23
Identifying independent eQTL effects.....	24
Second transcription factor binding site enrichment.....	24
Conclusion & Discussion .....	26
Effect of Imputation .....	26
Transcription factor binding site enrichment.....	26
Transcript enrichment .....	26
Retrotransposon enrichment .....	27
Histone mark enrichment .....	27
Identifying independent eQTL effects.....	27
Second transcription factor binding site enrichment.....	27
References.....	29

## Introduction

Expression Quantitative Trait Loci (eQTLs) are loci on the genome that affect the expression of genes. Although the underlying mechanisms of eQTLs are understood more and more it is still not entirely clear how they affect gene expression. eQTLs can be categorized into two types, *cis*-acting and *trans*-acting. *Cis*-eQTLs are located close to the gene they regulate. *Trans*-eQTLs regulate genes from a large distance, sometimes even from a different chromosome<sup>[1, 2, 6, 8]</sup>. Most of the identified eQTLs in eQTL mapping studies are *cis*-acting and only a small proportion is *trans*-acting<sup>[1]</sup>. Because *cis*-eQTLs are close to the gene they influence, these types of eQTLs are easier to identify than *trans*-eQTLs<sup>[1]</sup>. *Cis*-eQTLs are identified by correlating the expression of genes measured with expression probes and the genotypes within 250kb from the expression probe. *Cis*-eQTLs have been identified for many different human cell lines and tissues: approximately 30% of all *cis*-eQTLs are tissue specific<sup>[2]</sup>.

*Cis*-eQTLs are often found near Transcription Start Sites (TSS) and within gene bodies. The *cis*-eQTLs near TSS are not clearly identified or mapped to specific TSS regions<sup>[1, 5]</sup>. Additionally, *cis*-acting eQTLs are often found at transcription factor binding sites, and histone modification marks that are associated with active promoters and enhancers (histone mark H3k4me3 for example)<sup>[5, 7]</sup>.

SNPs (Single Nucleotide Polymorphisms) associated with complex traits are more likely to be eQTLs compared to SNPs that are not associated with complex traits<sup>[3, 4]</sup>. Identifying the underlying mechanisms of eQTLs can give more insight about how SNPs regulate gene expression and perhaps how they are involved in complex traits.

## Project goals

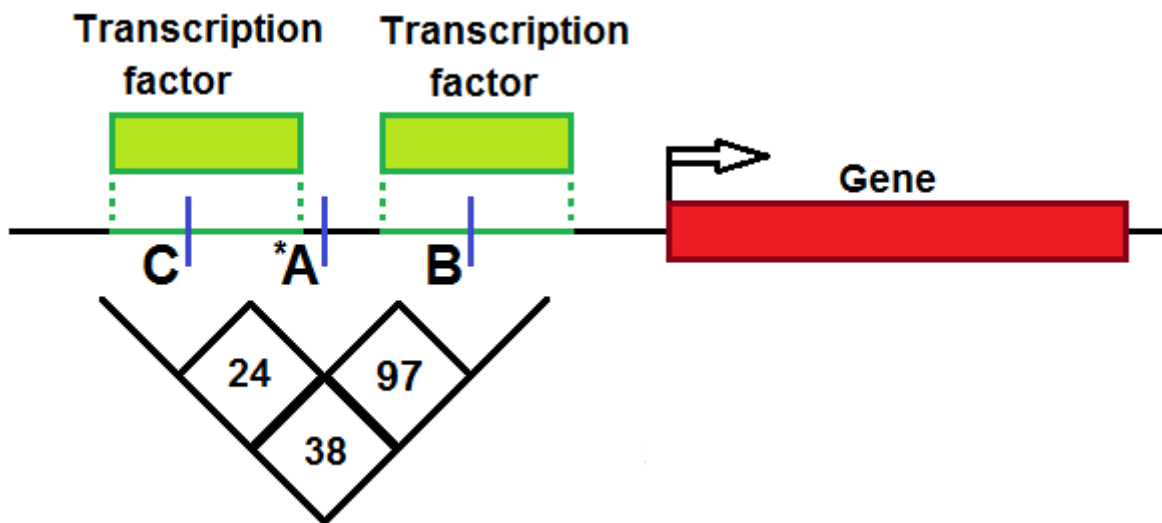


Figure 1: A genomic region with a gene, two transcription factors and three SNPs, each with an eQTL effect. SNP A has the most significant eQTL effect. SNP B which is strongly linked to SNP A and is located in the binding site of a transcription factor. By incorporating SNP B in the analysis of SNP A, a better understanding can be obtained as to how the same eQTL effect in SNP A and B might influence gene expression.

Figure 1 displays a region on the genome with a gene, two transcription factors and their binding sites and three SNPs represented by the vertical lines and labeled A, B and C. All three SNPs have a *cis*-eQTL effect. SNP A has the most significant effect and is therefore labeled with an asterisk. Below the SNPs is a small matrix which indicates the Linkage Disequilibrium (LD) between the SNPs. SNPs A and B have a strong linkage and have the same eQTL effect. SNP B is located in a binding site of a transcription factor and might be able to provide a better understanding how an eQTL influences the expression of the gene. By including strongly linked eQTLs in the analysis of eQTLs with a strong effect (in this case including SNP B in the analysis of SNP A), a better understanding as to how eQTLs influence gene expression can be obtained.

The process of imputation enables the identification of more SNPs with an eQTLs [15]. This project will try to use imputation to improve the functional annotation of eQTLs. Therefore, this project focuses on determining whether imputation improves the number of identified eQTLs, and determines whether imputation enhances the functional annotation of eQTLs. This can show that imputation has more added value in eQTL mapping studies than just identifying more SNPs. Additionally, this project aims to determine which regulatory factors and genomic sites are enriched for eQTLs. Gaining which regions and transcripts are enriched and depleted for eQTLs can give more insight as to how eQTLs might regulate gene expression.

To do this, this project performs two *cis*-eQTL mappings on a non-imputed and imputed blood genotype dataset. To determine the effect of imputation on the identification and annotation of eQTLs, the results of the two mappings are compared. Comparing the two mappings is done by looking at the difference between the most significant eQTLs effects per expression probe. Additionally, the identified eQTLs from both mappings will be annotated using RegulomeDB. To determine the enrichment of eQTLs for regulatory regions, enrichment analyses will be performed on the results of one of the two mappings using various data sources, after comparison of the two mappings.

In this study, the term enrichment describes that more eQTL effects are found than expected by chance. An enrichment for protein coding transcripts for instance, means that eQTLs are found more often in these transcript types than is expected by chance. The term depletion in this study describes the opposite of enrichment, which means that eQTLs are found less often than expected by chance. Enrichment analysis will be performed for transcription factor binding sites, and genomic transcripts such as micro RNAs, protein coding transcripts and retotransposons. Known histone marks will also be explored to see if these sites are enriched for eQTLs as well. Finally, more eQTL effects will be identified by iteratively regressing out the most significant eQTLs from previous mappings to see if more effects can be found after correcting for previous effects. It is expected that enrichments will be found for transcription factor binding sites, protein coding transcripts and histone marks as has been found in other studies<sup>[1,5]</sup>.



## Theory

### Regulation of gene expression

The expression of genes is regulated by many different factors such as transcription factors, micro RNAs and histones. Each of these factors regulates the expression of genes in different ways.

**Transcription factors** are proteins that regulate gene expression by binding to the DNA sequences in regulatory regions of genes, known as transcription factor binding sites. At these regulatory sites, transcription factors interact with other proteins at promoters near the Transcription Start Sites (TSS) of genes<sup>[16]</sup>. Binding of one or more transcription factors and RNA Polymerase II causes the transcription of genes. Although the majority of transcription factors regulate gene expression in a positive way (more transcription of a gene) there are some transcription factors that regulate gene expression in a negative way (less transcription of a gene)<sup>[18]</sup>.

**Micro RNAs (MiRNAs)** are a class of small single stranded non coding RNA's (ncRNA) ranging from 19 to 25 nucleotides in length and are transcribed by RNA Polymerase II. These small RNA molecules can pair with a target messenger RNA (mRNA), causing cleavage of the mRNA or translational repression. The majority of miRNAs are located either in intergenic regions or in antisense with respect to the genes they regulate<sup>[17]</sup>.

**Histones** are proteins that package the genome into ordered structures. This is done by packaging four types of histones, H2A, H2B, H3 and H4, in a protein octamer. When histones are methylated, a methyl group is added to the tail of the histone. The opposite process, in which a methyl group is removed from a histone tail, is called demethylation. Depending on which histone type is methylated or demethylated and the amount of methylation or demethylation, portions of the genome might become more or less accessible for transcription factors, which can result in a change in gene expression. Methylation mainly takes place at lysine residues. During trimethylation, three methyl molecules are added, or bound to the histone tail. Trimethylation of lysine residue 4 of histone H3 (H3K4me3) is associated with transcription activation, which means that transcription is increased. Dimethylation is the process of adding two methyl molecules to the histone tail. Dimethylation of lysine residue 9 on histone H3 (H3K9me2) is associated with transcriptional repression and a decrease of transcription.

**Retrotransposons** are DNA sequences that are able to move or copy themselves to other parts of the genome and constitute about a third of all human DNA. Although most of these elements are believed to be inactive, some elements are still active such as LINE1, Alu and SVA elements. Alu elements are of the Short Interspersed Elements (SINEs) class and LINE1 elements of Long Interspersed Elements (LINEs). SINE retrotransposons are short retrotransposons and more numerous than LINE retrotransposons but constitute a smaller proportion of the entire genome in combined length. LINE1 elements are the only autonomous retrotransposons, which means that they can move or copy themselves. Alu and SVA elements cannot move or copy themselves and depend on other factors to accomplish this<sup>[14]</sup>. Retrotransposons can also regulate gene expression<sup>[14, 21]</sup>, although the mechanism of regulation by these elements is still unclear. To give one example, LINE1 has been found to negatively regulate the gene expression of the *HLA-G* gene<sup>[22]</sup>.

### Single Nucleotide Polymorphisms

Single Nucleotide Polymorphisms are single bases in the genome that can differ between individuals. SNPs can be used for various analyses. For instance, SNPs can be used to determine the variation in genomic regions but can also be used as markers for eQTL studies.

## Linkage disequilibrium

Linkage Disequilibrium (LD) refers to the non-random association of two SNPs<sup>[19]</sup>. This means that two SNPs are inherited together more often than expected by chance. The LD is a scale that runs from 0 to 1. An LD value of 1 indicates that two SNPs are completely linked. This means that two SNPs are always inherited together. An LD of 0 indicates that two SNPs are not linked at all.

## Imputation

Imputation<sup>[9]</sup> is a process used to add and identify new and often times more rare genetic variants in the form of SNPs based upon more generally identified genetic variants. During imputation, SNP information from a reference set is used to fill in gaps in a study set resulting in a more comprehensive set of SNPs. Liming Ling et al<sup>[15]</sup> performed an analysis on 14,177 eQTLs in lymphoblastoid cell lines and found that through imputation more eQTL effects can be identified.

## Materials & Methods

### Materials

During the course of this project multiple data sources have been utilized to perform the eQTL mapping and the various enrichments. Below, the used data sources are described.

#### Gene expression data

Gene expression levels were measured with expression probes in blood tissue for 1,240 human individuals. Gene expression levels were measured using Illumina HT12v3 micro-arrays. These arrays contain probes that are able to specifically measure gene expression levels. The resulting dataset contains gene expression measurements for each individual. The gene expression data is correct with 40 principal components (PCs) with a principal component analysis (PCA). The gene expression data was in Genome Build 36.

#### Genotype data

For the same 1,240 human individuals, genotype data is available. This genotype data is a dataset that identified 294,767 SNPs through SNP calling from Illumina genotyping arrays. For each SNP, various data is calculated such as the minor and major allele frequency and which specific bases have been found for that SNP.

#### Genome of the Netherlands imputed genotype data

The Genome of the Netherlands (GoNL)<sup>[9]</sup> is a project established by the Dutch national network of bio banks called Biobanking and Biomolecular Research Infrastructure-Netherlands (BBMRI-NL). The GoNL project aims to characterize the genetic variation in the Dutch population by sequencing 250 Dutch families, totaling 769 individuals. SNP Calling has been completed for all 769 individuals and has been improved through imputation resulting in 19,562,004 SNPs in total. This data has been used to form the GoNL imputed blood genotype dataset.

#### Permutation data

In the permutation data, SNPs with an eQTL effect and probes are randomized. SNPs and probe combinations are shuffled. This data was used for all performed enrichments to determine whether eQTLs were enriched for transcription factor binding sites or certain transcript types for example.

#### RegulomeDB

RegulomeDB<sup>[10]</sup> is a database that resulted from a project focused on identifying and annotating regulatory SNP variants. This was done by combining various data sources such as transcription factor binding site data from the ENCODE project, DNaseI hypersensitivity data, computational predictions and eQTL data. Regulatory variants in RegulomeDB are awarded scores depending on their annotation. Variants with a low score are better annotated than variants with a high score, which are less well annotated. The highest score is 1a, whereas the lowest score is 7.

### **Gencode Annotation data**

The GENCODE annotation data<sup>[11]</sup> is a data source consisting of annotated genomic loci. The GENCODE data contains the genomic locations of many different transcript types. These transcript types not only include protein coding transcripts, but also transcripts such as long intergenic noncoding RNAs (lincRNAs), miRNAs and pseudogenes. Currently, the latest version of the GENCODE annotation data is version 18, which was used during this project.

### **Retrotransposon data**

The in-house retrotransposon data consisted of Alu, SINE and LINE elements identified using the repeat masker from UCSC. This masker identified repetitive sequence regions and returned their chromosomal positions as well as the classification.

### **DBRPIP data**

The Database of Retrotransposon Insertion Polymorphisms (dbRIP)<sup>[11]</sup> is a database that merges data for Alu, LINE1 and SVA polymorphisms. For each retrotransposon insertion site, the database offered the chromosomal location of the retrotransposon as well as the classification of the retrotransposon. The retrotransposon classification within the database consists of the class, family and subfamily. If the retrotransposon insertion site was associated with a disease, the disease was also added to the entry.

### **ENCODE Histone mark data**

The histone mark data from the ENCODE project<sup>[12]</sup> is data obtained from ChIP-seq experiments for various cell lines, such as cell line K562, and different histone marks such as H3K4me3. For many genomic regions, the datasets contain peaks which indicate the signal received from the ChIP-seq experiment. This project used all available Gm and K562 cell lines and available histone marks for these cell lines. These cell lines were used because they are present in blood tissue.

## **Methods**

### **Genome build conversion**

A Genome Build refers to the digital version of a reference genome. Each new genome build consists of better annotations and contains fewer gaps in the sequence. The used eQTL Mapping pipeline outputted the genomic positions of the eQTL effects in Genome Build 36. Therefore, the chromosomal positions in all data sources were converted from Genome Build 37 to Genome Build 36 before performing the enrichments. Conversion of the data sources was done using the UCSC liftover tool and the hg19tohg18 chain file.

## eQTL Mapping

An eQTL mapping was performed on a non-imputed and GoNL 4 imputed blood genotype dataset, containing 1,240 samples, using the eQTL Mapping pipeline developed by Harm-Jan Westra and Lude Franke.<sup>[2]</sup> Both mappings used the same gene expression data. For both mappings, the False Discovery Rate (FDR) was set to 0.05, meaning that only 5% of the returned eQTL effects were expected to be false positives. The maximum number of FDR significant eQTL effects to return was set to one million. Furthermore, the call rate was set to 0.95, the Minor Allele Frequency to 0.05 and the Hardy Weinberg Threshold was set to 0.0001. During the mapping 100 permutation rounds were performed.

The used eQTL mapping pipeline generated P-Value and Z-Score for each eQTL. The P-Value denoted the significance of the eQTL. The Z-Score indicated the direction and could be a positive or negative value.

## Effect of imputation

Before performing any enrichment analyses, the effect of imputation on the eQTL mapping process was determined: the top eQTL effects detected for expression probes present in both non-imputed and imputed were compared. The Linkage Disequilibrium (LD) between the top effects per shared probe was determined using the GoNL version 4 imputed blood genotype data as this dataset contained more SNPs than the non-imputed dataset. For some probes, the LD between top effects could not be determined as either the imputed or not imputed top eQTL effects were not present in the genotype data. These probes were therefore excluded from further analysis.

The  $-\log_{10}$  transformed p-values (denoting the significance for each effect) of both sets were then plotted against each other per shared probe. The P-Values denote the significance of the eQTL effect. Smaller P-Values indicate more significant eQTL effects. The differences in Z-Scores between the top effects of non-imputed and imputed were plotted against the LD between the two top effects.

Afterwards, the top effects per shared probe for non-imputed and imputed were annotated using RegulomeDB. The eQTL entries from RegulomeDB category six were excluded from the analysis to avoid bias as these entries contained no functional annotation. The annotation results were plotted in a bar chart for visualization. This annotation was also performed using the proxy SNPs (nearby SNPs with a strong linkage to the top eQTL effect SNP) of each top eQTL effect. The proxy SNPs had to have an FDR 0.05 significant eQTL effect and had to be associated with the same probe as the top eQTL effect. Furthermore, proxies were filtered using LD cutoffs of 0.8, 0.9 and 0.99. Per probe only the best annotation was retained and plotted in a bar chart.

## Enrichment analyses

Various enrichment analyses were performed using the identified eQTL effects and the permutation data obtained from the eQTL mapping on the imputed dataset. Before performing any enrichment analysis, the eQTL results and permutation data were filtered to include only expression probes that were present in both sets. First, the top effects per probe were identified in the eQTL data. The LD between the top effect and other effects linked to the same probe were then identified. Effects that had an LD of 0.8 or higher with the top effect were retained. The same was done for the permutation data. For each probe the top effects and other effects with an LD of 0.8 or greater were identified and constituted the permutation data.

## Transcription factor binding site enrichment

To check whether transcription factor binding sites were significantly enriched for eQTLs, categories 1 through 5 of RegulomeDB were used. From these categories, only entries that contained one or more transcription factor binding site annotations were retained. Categories 6 and 7 from RegulomeDB were excluded as these two categories did not contain transcription factor binding site annotations. For each transcription factor it was determined how many eQTLs and permutations were located in its binding sites. Determining whether an eQTL or permutation was located in a binding site was based upon chromosomal position.

For each transcription factor with at least some eQTLs or permutations located in its binding sites, a Fisher Exact test was performed to determine whether the binding sites of each transcription factor were enriched or depleted for eQTLs. A Fisher Exact test was used because it is a non-parametric test and is better suited for smaller sample sizes than the chi-squared test. It was also determined whether the binding sites of transcription factors were significantly enriched or depleted for eQTLs by using Bonferroni correction. The Bonferroni correction was determined by dividing 0.05 by the number of transcription factors whose binding sites were enriched or depleted. The transcription factors that were found to be significantly enriched or depleted were used to search the GeneMANIA website<sup>[23]</sup> to assess whether these transcription factors might have some interaction, co-expression or shared protein domains with each other or other proteins.

## Transcript enrichment

Version 18 of the GENCODE annotation data was used to determine whether certain genomic transcript types such as protein coding transcripts, lincRNA and miRNA were enriched or depleted for eQTLs. For each transcript type it was determined how many eQTLs and permutations were located in transcripts of that type. For each transcript type a Fisher Exact test was performed to see which transcript types were enriched or depleted. A Bonferroni correction was then used to determine if a transcript type was significantly enriched or depleted for eQTLs. The Bonferroni correction was determined by dividing 0.05 by the number of transcript types for which an enrichment or depletion was found.

## Retrotransposon enrichment

The enrichment on the retrotransposons was performed in the same way as the enrichment performed on the GENCODE annotation data. For the retrotransposons it was determined whether a specific class of retrotransposons such as LINE, SINE or Alu elements were enriched or depleted for eQTLs by performing a Fisher Exact test for each of the retrotransposon class. The first enrichment for retrotransposons was performed on the in-house data obtained from the UCSC repeat masker. The second enrichment for retrotransposons was performed on the dbRIP data.

## Histone mark enrichment

The enrichment performed on histone marks was performed by determining whether eQTLs and permutations were located in histone mark sites based upon chromosomal position. If an eQTL or permutation was located in a histone mark, the peak value of that site was obtained. This eventually resulted in two lists of peak values, one for eQTLs and one for permutations. On these two lists of peak values, a Wilcoxon test was performed to determine whether there was a significant difference between the two groups. A Wilcoxon test was used because this is a non-parametric test which does not require that the data is in a normal distribution. Also, with the Wilcoxon test,

the two sets may differ in size, something that was often the case with this enrichment. This enrichment was performed for Gm cell lines GM06990, GM12864, GM12865, GM12866, GM12875 and GM12878 and the K562 cell line on all available histone marks.

### **Identifying independent eQTL effects**

After the first mapping on the imputed data, independent eQTL effects were identified by iteratively regressing out the most significant effects found in previous mappings and performing new eQTL mappings. The first iteration removed the top effects previously identified in the primary mapping before performing a new eQTL Mapping. The second iteration removed the top eQTL effects identified during the primary eQTL mapping and the top eQTL effects identified during the first iteration before running the eQTL Mapping. In total ten iterations were performed.

### **Second transcription factor binding site enrichment**

A second transcription factor binding site enrichment was performed the same way as the first enrichment. This enrichment however, combined the eQTLs from the first mapping and iterations and combined all permutation data from the first mapping and regression iterations.

## Results

### eQTL Mapping

The eQTL mapping on the non-imputed dataset identified 67,564 significant *cis*-eQTL effects for 6,039 unique probes (FDR < 0.05). The eQTL mapping on the imputed dataset identified one million significant *cis*-eQTL effects for 5,990 unique probes (FDR < 0.05). In total 5,517 probes have one or more eQTL effects in both non-imputed and imputed datasets. The imputed dataset has more eQTL effects per shared probe than non-imputed.

### Effect of imputation

Figures 2 through 4 show the effect that imputation has on the eQTL mapping process. Figure 2 displays two plots. Plot A shows the P-Values for the top effects for imputed and non-imputed plotted per shared probe. The P-Value denotes the eQTL effect. Smaller P-Values indicate stronger eQTL effects, whereas larger P-Values indicate weaker eQTL effects. The P-values in the plot A of Figure 2 are  $-\log_{10}$  transformed, therefore, smaller P-Values are larger values in this plot. As can be seen the plot is skewed towards imputed which has smaller P-Values for most of the shared probes, indicating that in the imputed dataset, *cis*-eQTLs on the same probe often have a lower P-Value and thus a larger significance. A more significant eQTL effect is closer to the causal variant. Plot B in this figure displays the difference in absolute Z-Scores between imputed and non-imputed datasets per shared probe (Delta Z-Score), plotted against the LD between each top effect per shared probe. The Z-Score can be calculated from the P-Value using the normal distribution. Higher Z-scores mean larger eQTL effects and lower p-values. The plot shows that top eQTL effects for imputed and non-imputed that have a high LD value (strong linkage) show less difference in Z-Scores than top eQTL effects that are barely linked. For some shared probes the top effect of non-imputed had a higher absolute Z-Score than imputed. These points have a Delta Z-Score smaller than 0. It is important to note that most top eQTL effects for non-imputed and imputed are strongly linked to each other as can be seen in Figure 3.



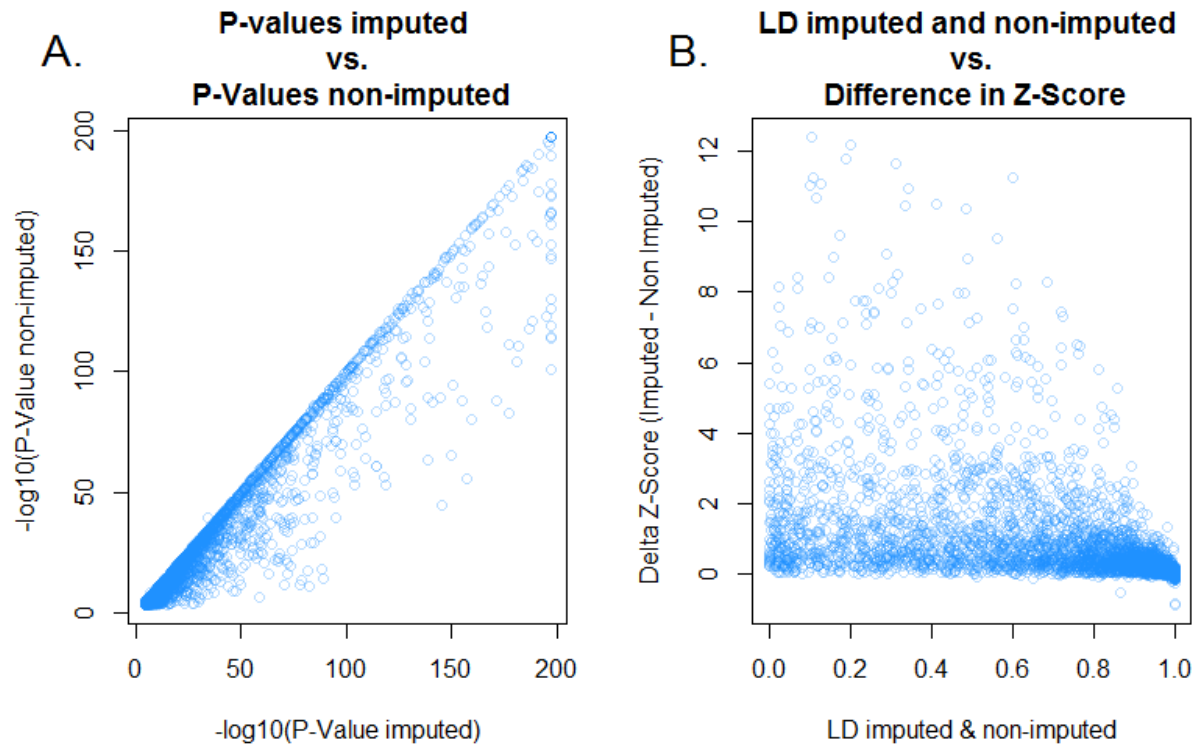
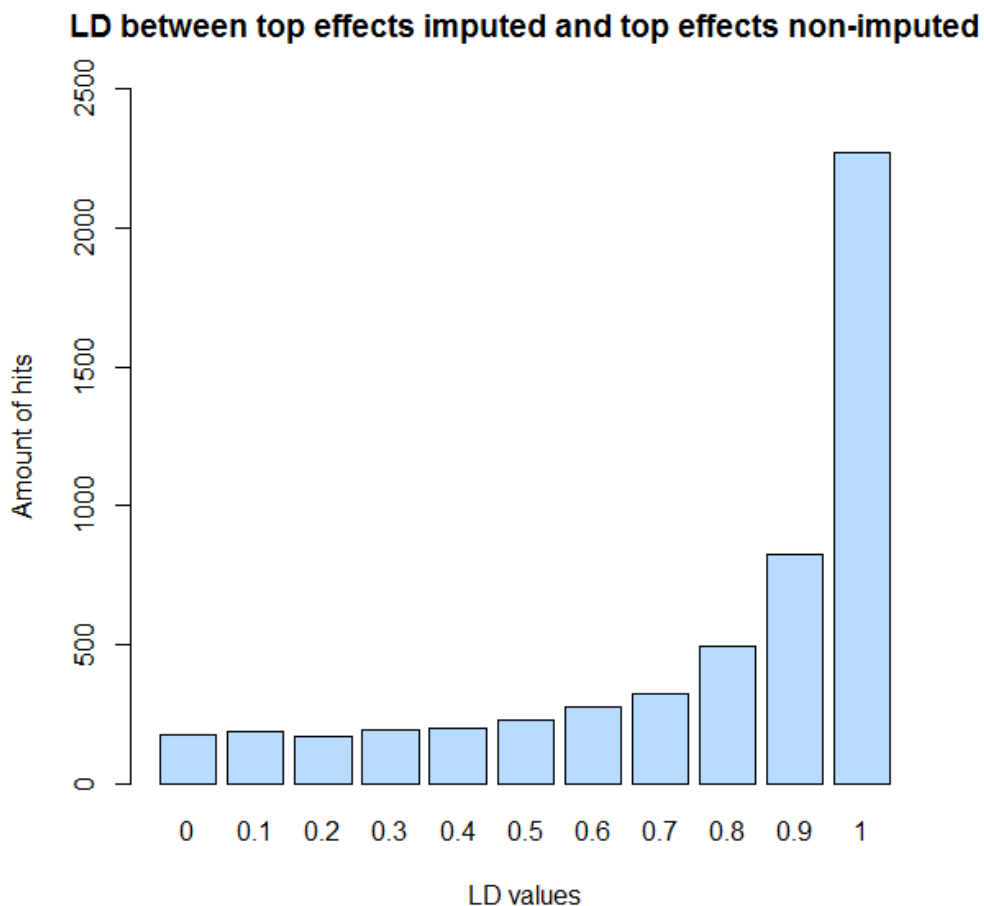


Figure 2: Plot A displays the P-Values of non-imputed plotted against the P-Values of imputed per shared probe. The P-Values are  $-\log_{10}$  transformed. Smaller P-Values therefore resemble higher values in this plot. The plot is skewed towards imputed, generally having smaller P-Values per shared probe. Plot B shows the difference in absolute Z-Scores between non-imputed and imputed. Z-Scores are directly linked to P-Values and can be converted from one to the other. Strongly linked eQTL effects have smaller differences in absolute Z-Scores.

Figure 3 displays the linkage between the top effects for non-imputed and imputed per shared probe. The vast majority of top effects are strongly linked to each other as most top effects per shared probe have a high LD value. About 2200 SNP are in complete or nearly complete LD. Only half of this amount, about 1100 is in LD of 0.5 or smaller, indicating most top effect are linked closely together and constitute the same effect. Most of these strongly linked top eQTL effects have a small but noticeable difference in Z-Score as can be seen in the plot B in Figure 2.



**Figure 3: Bar plot displaying the LD between top eQTL effects for non-imputed and imputed per shared probe. Most top eQTL effects for shared probes are strongly linked as high LD values are mostly found as can be seen by the high bar at the far right end of the barplot.**

Figure 4 shows the annotation results of the RegulomeDB annotation is visualized in a bar plot. In RegulomeDB, lower scores mean that a SNP has a better functional annotation. When using only the top effects, the annotation in low scoring annotations barely differ between non-imputed and imputed as can be seen in barplot A. Only the annotation in intermediate scores differs between non-imputed and imputed. When using proxy SNPs with an eQTL effect however, a shift from higher scores to lower scores can be seen for imputed. This shift is already present even when using a stringent cutoff value of 0.99 as can be seen in barplot B in Figure 4. In this barplot many annotations in scores 6 and 7 have shifted to lower scores. The shift from high scores to lower scores continues when slightly lower, yet still strict LD cutoff values are being used as seen in barplots C and D. In barplot D many hits from scores 6 and 7 have shifted to lower scores, mainly towards score 2a.

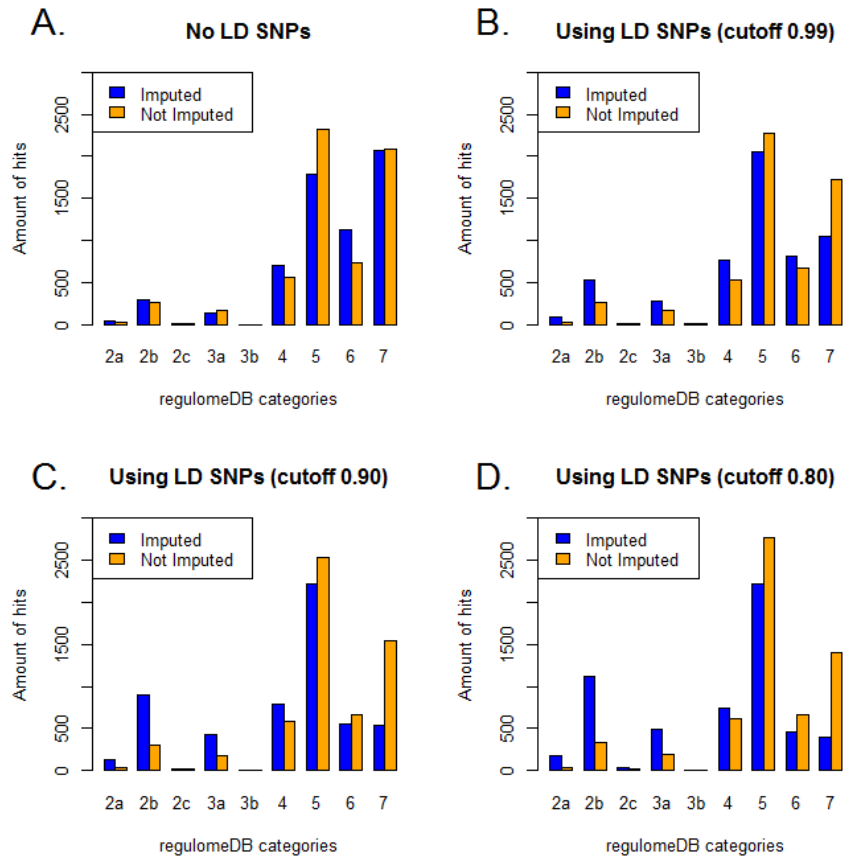


Figure 4: Barplots A through D displays the effect of imputation on imputation by performing an annotation in RegulomeDB. Barplot A displays the annotation when only the top eQTL effects per shared probes for non-imputed and imputed are used. Using proxy SNPs shifts the annotation from high scores to low scores mainly for imputed. This effect is already noticeable for the stringent LD cutoff of 0.99. The shift from high RegulomeDB scores to lower scores is more observable using slightly lower LD cutoff values.

## Transcription factor binding site enrichment

The enrichment analysis on transcription factor binding sites for the primary mapping on the imputed dataset returned hits for binding sites of 114 different transcription factors. The binding sites for nine of the 114 transcription factors are found to be significantly enriched or depleted for eQTLs. Table 1 displays the nine transcription factors of which the binding sites are enriched or depleted for eQTLs. Enriched transcription factors contain more eQTLs effects in their binding sites than expected by chance, whereas depleted transcription factors contain less eQTLs effects than expected by chance. Most of the significant transcription factors are expressed in blood and at least a few other tissue types such as kidney and liver tissues. Transcription factor NR4A1 is only expressed in blood<sup>[20]</sup>. As can be seen in Table 1, the binding sites of this transcription factor are enriched for eQTLs. Interestingly, POLR2A which encodes the largest subunit of RNA Polymerase II, is the most significant enriched factor. This is not surprising, POLR2A is responsible for mRNA transcription.

Transcription Factor	Fisher Exact Test P-Value	Enrichment/Depletion	Tissue
<b>POLR2A</b>	5.51E-16	Enrichment	Many tissues
<b>GATA1</b>	3.84E-10	Enrichment	Blood, Brain, Prostate cancer
<b>NR4A1</b>	3.27E-09	Enrichment	Blood
<b>CTCF</b>	8.41E-09	Depletion	Many tissues
<b>RFX3</b>	2.21E-08	Depletion	Blood, colon, bone prostate
<b>CEBPB</b>	2.77E-06	Enrichment	Many tissues
<b>MYC</b>	8.74E-05	Enrichment	Many tissues
<b>GATA2</b>	1.05E-04	Enrichment	Blood, Kidney, Prostate
<b>MAFK</b>	3.12E-04	Depletion	Blood, Kidney, Cervix

**Table 1: Displays 14 transcription factors with the most enriched or impoverished binding sites for eQTLs. Entries in green are found to be bonferroni significant. Enrichment indicates that more eQTLs are found for binding sites of that transcription factor than expected by chance whereas impoverishment indicates that less eQTLs are found for binding sites of that transcription factor.**

Figure 5 displays the nine significant transcription factors displayed in a gene network obtained from the GeneMANIA website. Factors such as GATA1 and GATA2 show that they contain shared protein domains with a variety of other, usually very similar proteins. At the bottom of Figure 5, a cluster containing six of the nine significant transcription factors can be seen. Some of these transcription factors are co-expressed with each other or have physical interactions. When these six factors are searched on the GeneMANIA website, it is revealed that they have co-expression with more proteins than what seemed to be the case in Figure 5. These other proteins are either not found or not significantly enriched or depleted for eQTLs. Many of the direct co-expression or interaction between the transcription factors remains the same however.

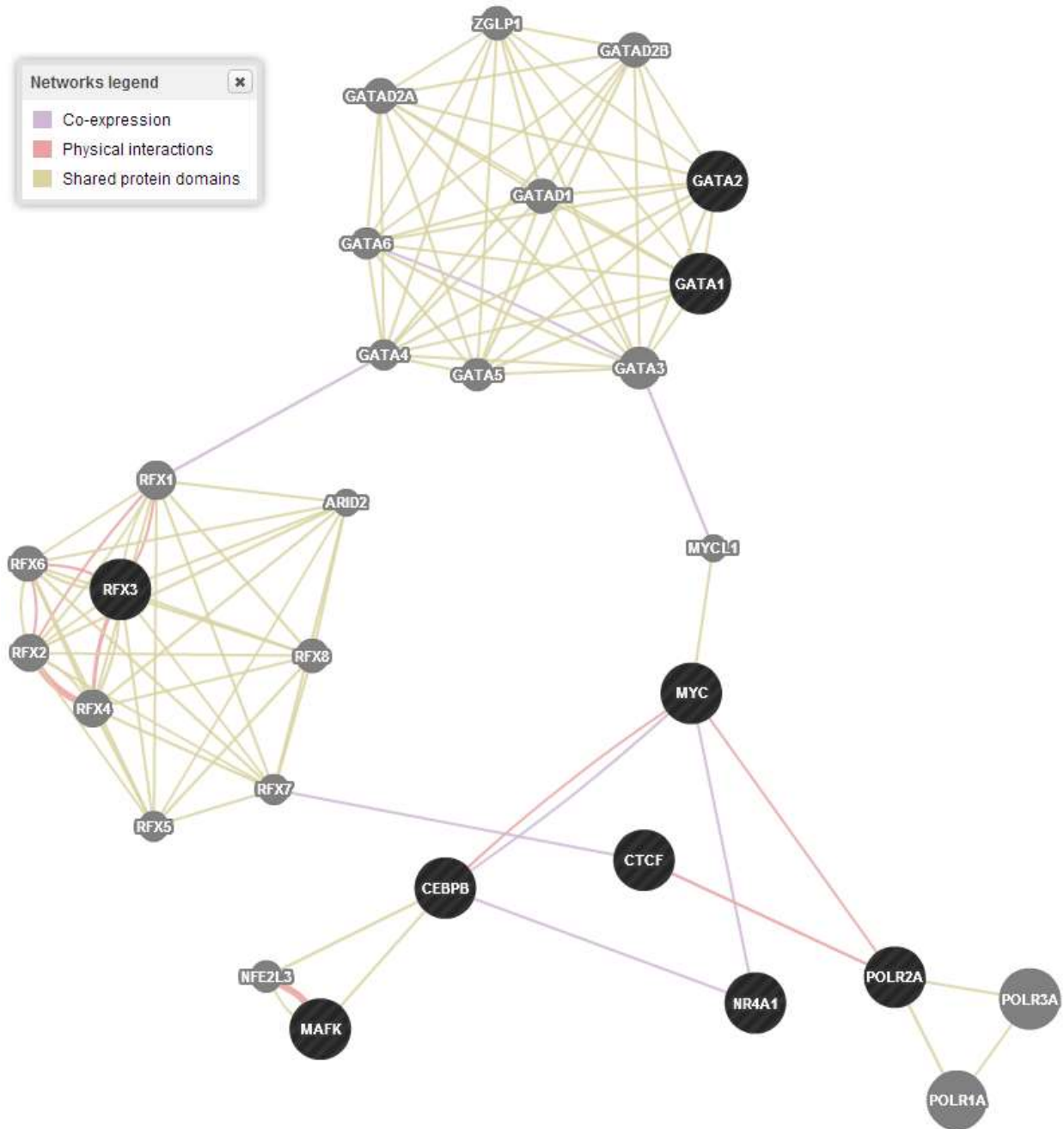


Figure 5: The nine transcription factors whose binding sites are significantly enriched or depleted for eQTL effects visualized in a gene network on GeneMANIA. Most (six of nine) of the transcription factors are co-expressed or have physical interactions. Some transcription factors such as GATA1 and GATA2 have shared protein domains with other similar proteins.

Figure 6 inspects the cluster of six transcription factors a bit closer and shows other genes that interact with these transcription factors.

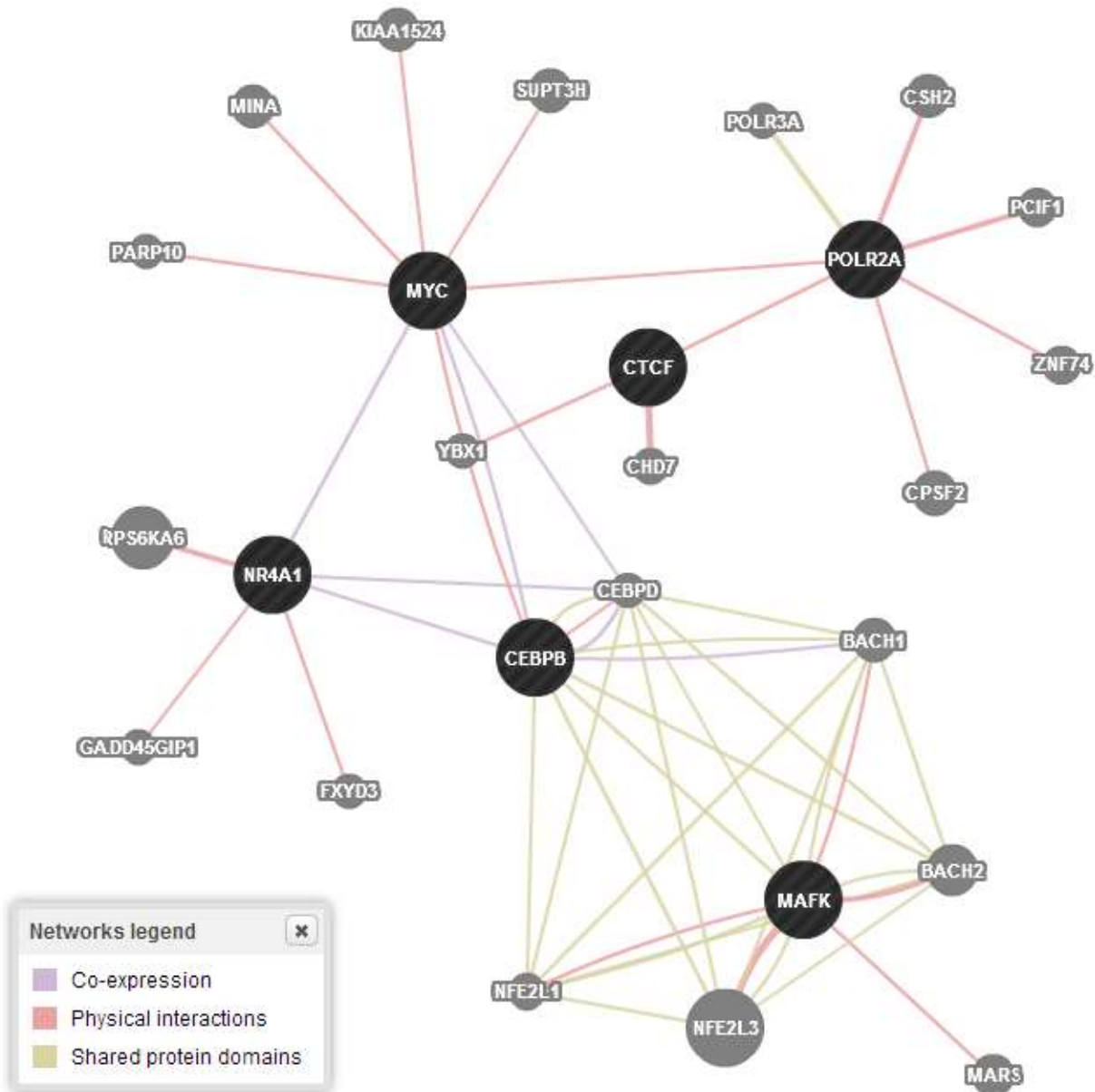


Figure 6 Close-up look at a cluster of six of nine transcription factors. A closer look to these six transcription factor reveals they are connected to more genes. Direct co-expression or physical interactions between transcription factors as seen in Figure 4 generally remain the same.

## Transcript enrichment

The enrichment analysis on the GENCODE annotation data shows enrichment and depletion for a total of ten different transcript types. Four of ten transcript types are significantly enriched or depleted for eQTLs and are displayed in Table 2. Protein coding transcripts are the most significantly enriched type of transcripts followed by antisense transcripts which are alternatively spliced variants. The strong enrichment for protein coding corresponds to earlier experiments. eQTLs showed the strongest significant depletion for pseudogenes, followed by antisense transcripts.

Transcript type	Fisher Exact Test P-Value	Enrichment/Depletion
Protein coding	5.70E-06	Enrichment
Pseudogene	6.50E-05	Depletion
Antisense	2.17E-04	Depletion
Retained intron	6.57E-04	Enrichment

Table 2: Top 5 enriched or impoverished transcript types from the GENCODE version 18 annotation data. eQTLs are found much more often in protein coding and retained intron transcripts than expected by chance.

## Retrotransposon enrichment

The performed enrichment on the repeat region data identified did not find an enrichment or depletion on any retrotransposon class or type. Also no single type of transcript was reported. Much like the enrichment performed on the in-house retrotransposon data, the enrichment on the dbRIP data neither found an enrichment or depletion for any retrotransposon class or type.

## Histone mark enrichment

For the GM (GM06990, GM12864, GM12865, GM12866, GM12875 and GM12878) and K562 cell lines no significant histone binding site enrichments were found for any of the available types of histone mark.

## Identifying independent eQTL effects

By iteratively regressing out previously identified top eQTL effects and performing new eQTL mappings, new independent eQTL effects were identified in the imputed genotype data. Table 3 displays the number of top eQTL effects and total FDR 0.05 significant eQTL effects for each iteration. Note that in the seventh, eighth and ninth iterations, eQTL effects are only found for one probe. This probe belongs to the KRT1 gene. At the tenth iteration, no new independent eQTLs effects are found.

Regression	Probes with eQTL effects	Number of eQTL effects
1 <sup>st</sup> Iteration	1,679	261,210
2 <sup>nd</sup> Iteration	539	64,759
3 <sup>rd</sup> Iteration	158	18,047
4 <sup>th</sup> Iteration	37	2,896
5 <sup>th</sup> Iteration	15	532
6 <sup>th</sup> Iteration	8	439
7 <sup>th</sup> Iteration	1	27
8 <sup>th</sup> Iteration	1	36
9 <sup>th</sup> Iteration	1	73
10 <sup>th</sup> Iteration	0	0

**Table 3:** Shows the amount of probes with eQTL effects and the total number of eQTLs identified after each iteration. After each iteration, the number of found eQTLs and probes with eQTLs drops. During the seventh, eighth and ninth iteration eQTLs are only found for one probe that is associated with the KRT1 gene.

## Second transcription factor binding site enrichment

Hits for the binding sites of 134 transcription factors were found. Of these 134 transcription factors, the binding sites of 20 transcription factors were significantly enriched for eQTLs and the binding sites of 25 other transcription factors were significantly depleted for eQTLs. Table 4 shows the 20 transcription factors of which the binding sites are found to be significantly enriched for eQTL effects. In total eight transcription factors are mainly expressed in blood tissue (usually also expressed in Liver and Kidney tissues), two of which are expressed only in blood tissue. The other transcription factors are expressed in many different tissues. For this enrichment, the binding sites of POLR2A are the most enriched for eQTLs. Most of the transcription factors found during this enrichment analysis were also found in the first transcription factor binding sites enrichment analysis but with different P-Values. Some transcription factors were only found in this enrichment en some only in the first enrichment analysis.



Transcription factor	Fisher Exact test P-Value	Tissues
POLR2A	0	Many tissues
TAF1	3.44E-56	Many tissues
ELF1	5.92E-23	Many tissues
GABPA	1.25E-22	Many tissues
E2F1	9.18E-17	Blood, Liver and Kidney tissues
NRF1	3.14E-14	Many tissues
E2F4	1.17E-11	Many tissues
HEY1	2.07E-09	Blood and Kidney tissues
YY1	3.65E-09	Blood, Kidney and Liver tissues
HMG3	2.37E-08	Many tissues
ELK4	4.68E-08	Blood tissue only
SP2	8.30E-08	Blood and Kidney tissues
RFX5	5.14E-07	Many tissues
SIN3A	4.17E-05	Many tissues
ATF3	8.79E-05	Blood and Kidney tissues
CHD2	1.14E-04	Many tissues
CCNT2	1.26E-04	Blood, Kidney and Liver tissues
IRF1	1.34E-04	Blood tissue only
E2F6	1.97E-04	Many tissues
GTF2B	2.43E-04	Many tissues

Table 4: The 20 transcription factors whose transcription factor binding sites are significantly enriched for eQTLs. As in the previous transcription factor binding site enrichment, POLR2A is the most significant factor. Factors ELK4 and IRF1 are only expressed in blood tissue.

## Conclusion & Discussion

### Effect of Imputation

As can be seen in figures 1 through 4, imputation has a positive effect on identification of eQTLs. Although the mapping on the imputed dataset identified effects for fewer probes, the top eQTLs found for shared probes are overall more significant. The P-Values of these top effects are smaller in imputed meaning they have a stronger effect. Imputation allows the identification of more eQTL effects per probe than when a non-imputed dataset is used. The identified SNPs provide better functional annotation, especially when proxies with eQTL effects are used as well. Better annotation of eQTLs leads to a better understanding how eQTLs influence gene expression and increase the chance of finding the causal variant of the change in gene expression.

### Transcription factor binding site enrichment

Transcription factor binding sites are important underlying mechanisms for eQTLs as many transcription factors were found. The observation that the binding sites of a small portion of the found transcription factors are enriched for eQTLs seems to indicate that specific transcription factors are enriched for eQTLs.

The cluster which is formed by six of the nine significant transcription factors in Figure 5 is interesting, especially since binding sites for two of the transcription factors are found to be depleted for eQTLs. Upon closer inspection of these six genes it becomes clear that other genes are also involved with each transcription factor. Most of these genes are also mainly expressed in blood, liver and kidney tissues. Some genes are also expressed in heart and brain tissue. Although no definite conclusions can be drawn from Figures 5 and 6 it can be interesting to determine if these genes are active in the same pathways. When using all identified eQTL effects (primary and secondary eQTL effects) the binding sites of many more transcription factors are enriched as would be expected when using more eQTL effects.

It is interesting to observe that the POLR2A binding sites are the most significantly enriched for eQTLs. Since POLR2A is the major subunit of RNA Polymerase II, which is the main RNA polymerase responsible for transcription of mRNA<sup>[18]</sup>, you would expect this protein to bind at every binding site to initiate transcription of genes. It might be that the eQTLs directly influence binding of the POLR2A protein. Transcription of genes however consists of multiple transcription factors and RNA polymerase II. Therefore, the eQTLs might not influence POLR2A directly but indirectly through other transcription factors. It could be that other independent effects not found during the eQTL mapping are actually responsible for the effect and most SNPs located in POLR2A binding sites are in strong LD with this other independent effect.

### Transcript enrichment

Protein coding transcripts are found to be strongly enriched for eQTLs as found in other studies. Somewhat interesting to observe is that “retained intron” transcript types are also found to be strongly enriched for eQTLs. Retained intron transcripts are alternative spliced variants of a gene. Most likely, the eQTLs found in protein coding and retained intron transcript types are located in exons. The eQTLs located in exons of protein coding transcripts are also located in the exons of the alternatively spliced variant of the same protein coding transcripts. This explains why these two transcript types are strongly enriched for eQTL effects. Retained intron transcript types are slightly less enriched for eQTLs than protein coding transcripts. The most likely explanation of this observation is

that some eQTLs are located in exons in the protein coding transcript that are removed, or spliced out, in the alternatively spliced transcripts. Whether this is the case can be determined using RNA-Seq data.

It is also interesting to observe that antisense transcripts are depleted for eQTLs. Antisense transcripts can regulate the gene to which it belongs. It might be that these transcripts are located just outside of the protein coding transcripts or in regions that generally do not contain eQTLs. Pseudogenes are also found to be significantly depleted for eQTLs. Pseudogenes are thought to be non-functional genes, so the fact that they are depleted for eQTLs might not be that surprising.

## Retrotransposon enrichment

Although it is suggested that retrotransposons might have a regulatory role, they are not enriched for eQTLs as no enrichments have been found at all. Observing that pseudogenes, which are a class of retrotransposons, are actually depleted for eQTLs clearly indicates that eQTLs do not regulate gene expression by means of retrotransposons. This, however, does not mean that retrotransposons might not regulate gene expression at all. It might be that no SNP are or can be called in retrotransposons or because probe did not map to these regions.

## Histone mark enrichment

During this project, no significant enrichments were found for eQTLs in histone marks. This is surprising as histone marks, especially those associated with active promoters and enhancers were supposed to be enriched for eQTLs. A possible explanation that no significant enrichments were found for histone marks might be that many histone marks fell in between the eQTLs. Or, because of the complexity created by LD of the genetic data, the permutation data might not have been correct.

## Identifying independent eQTL effects

More eQTL effects are identified by iteratively regressing out previous top effects. A strange observation however was that after six iterations, all significant eQTL effects originate from the same expression probe mapped to the KRT1 (keratine) gene. Whether these eQTL effects are significant or a technical artifact will have to be determined through multiple mappings in other datasets or other tissues.

## Second transcription factor binding site enrichment

Combining the eQTLs found in the first mapping and the eQTLs found iteratively regressing out top effects increases the number of transcription factors whose binding sites are significantly enriched and depleted for eQTLs. The fact that POLR2A is the most significantly enriched factor in this enrichment as well is not surprising, as discussed above. It is interesting to see that the binding sites of much more transcript factors are enriched or depleted when combining all effects. This could be because there are a lot more eQTL effects available and thus more capabilities to discover new enriched or depleted sites. Perhaps it might be the

Imputation has a positive effect on the eQTL mapping process: it identifies more significant effects that yield better functional annotations. *Cis*-eQTLs are enriched in transcription factor binding sites, protein coding transcripts and alternatively spliced variant transcripts but not in retrotransposons. Histone marks were also not found to be enriched. By iteratively regressing out previous top effects and performing new mappings more independent eQTL effects can be identified. Combining all these effects identifies more transcription factor binding sites enriched for eQTLs.

The methods used to perform the enrichments in this project work well in identifying which regions and factors of the genome are enriched for *cis*-eQTLs. This method can be extended to likely yield even more results. Instead of looking at only the positions of the SNPs with an eQTL effect, it can be more informative to treat SNPs as a region. In this approach, SNPs with an eQTL on the same gene and are strongly linked can be treated as a genomic region with an eQTL effect. These regions can then be used to determine which regulatory site is located in such a formed eQTL region. Regulatory sites that before were located in between SNPs may now be identified as well.

## References

1. Gaffney DJ (2013) Global Properties and Functional Complexity of Human Gene Regulatory Variation. *PLoS Genet* 9(5): e1003501. doi:10.1371/journal.pgen.1003501
2. Jingyuan Fu, Marcel G. M. Wolfs, Patrick Deelen et al., Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression. *PLoS Genet* 8; 2012
3. Fehrmann RSN, Jansen RC, Veldink JH, Westra H-J, Arends D, et al. (2011) Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. *PLoS Genet* 7(8): e1002197. doi:10.1371/journal.pgen.1002197
4. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet* 6(4): e1000888. doi:10.1371/journal.pgen.1000888
5. Gaffney et al.: Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology* 2012 13:R7.
6. Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet* 9(8): e1003649. doi:10.1371/journal.pgen.1003649
7. Chunlei Wu, David L. Delano, Nico Mitro et al., Gene Set Enrichment in eQTL Data Identifies Novel Annotations and Pathway Regulators. *PLoS Genet* 4; 2008
8. Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko et al., Systematic identification of trans eQTLs as putative drivers of known disease associations, *Nature Genetics* 45, p.1238-1243; 2012
9. Yun Li, Cristen Willer, Serena Sanna et al., Genotype Imputation, *Annual Review Genomics Human Genetics* 10, p.387-406; 2009
10. Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., et al. (2013). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics*. doi:10.1038/ejhg.2013.118.
11. Alan P. Boyle, Eurie L. Hong, Manoj Hariharan et al., Annotation of functional variation in personal genomes using RegulomeDB, *Genome Research* 22, p.1790-1797;2012
12. Jennifer Harrow, France Denoeud, Adam Frankish et al., GENCODE: producing a reference annotation for ENCODE, *Genome Biology* 7; 2006
13. The ENCODE Project Consortium, A User's Guide to the Encyclopedia of DNA Elements (ENCODE), *PLoS Biology* 9; 2010
14. Jianxin Wang, Lei Song, Deepak Grover et al., dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans, *Human Mutation* 27(4), p.323-329; 2006

15. Liming Liang, Nilesch Morar, Anna L. Dixon et al., A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines, *Genome Research* 23, p716-726; 2013
16. Dennis Wang, Augusto Rendon and Lorenz Wernish, Transcription factor and chromatin features predict genes associated with eQTLs, *Nucleic Acids Research* Vol 41 No. 3, p.1450-1463; 2012
17. Yoontae Lee, Minju Kim, Jinju Han et al., MicroRNA genes are transcribed by RNA polymerase II, *The EMBO Journal* 23, p.4051-4060; 2004
18. David S. Latchman, Transcription factors: an overview, *International Journal of Experimental Pathology* 74, p.417-422; 1993
19. David E. Reich, Michelle Cargill, Stacey Bolk et al., Linkage disequilibrium in the human genome, *Nature* 411, p.199-204; 2001
20. <http://www.genecards.org/cgi-bin/carddisp.pl?gene=NR4A1>
21. Han, J. S. and Boeke, J. D., LINE-1 retrotransposons: Modulators of quantity and quality of mammalian gene expression?, *Bioessays*, 27, p.775–784; 2005
22. Masashi Ikeno, Nobutaka Suzuki, Megumi Kamiya et al., LINE1 family member is negative regulator of HLA-G expresion, *Nucleic Acids Research*; 2012
23. <http://www.genemania.org>