Analyzing *cis*-eQTLs for Population Specific Impact of Genetic Variation on Gene Expresion

Author: Madeline LaScola-Hutcherson

ABSTRACT

Interpreting and characterizing the functional effects of genetic variants in humans is key to the future of predicting and understanding a plethora of genetic diseases. Genome sequencing projects that provide a deep analysis of mRNA are helping push the field towards the discovery of millions of genetic variants in humans. In this project we use 421 individuals from the 1000 Genomes Project and perform RNA sequencing to explore the effect of genetic variation on the regulation of genes. By identifying and mapping *cis* eQTLs in individuals and across 5 different populations we are able to infer casual varaints for various disease-associated loci with population-specificity. This study will provide deeper insight into functional variants in the human genome as well as the cellular mechanisms behind them in order to better understand genetic disease.

INTRO

Understanding genetic variants and what physiological effects they may cause is one of the outstanding obstacles in human genomics today(1). There is a pressing need to identify the molecular mechanisms behind genetic risk for certain complex traits and diseases as understanding this correlation could help predict genetic risk early on. One method of doing this is through the mapping of expression quantitative trait loci (eQTLs) (7) which are loci that partly explain the variation in gene expression phenotypes. Past studies have sought to address this issue through analyzing variants' effect on gene expression (4,5,6,7,8) which is known to have an effect on a variety of human diseases and traits. In this analysis we focus on *cis* eQTL, with a variant being identified as *cis* if it is located within 500 kb (megabase) of the transcription start site (TSS). Using genome-wide association studies (GWAS) (8) that test single-nucleotide polymorphisms (SNPs) for association with a disease have uncovered that these variants are likely involved in gene regulation and can be found without any prior knowledge of specific cis regulatory regions.

While there have been several studies dedicated to this topic there is still limited understanding of the biological mechanisms behind the detected causal variants. Studies thus far indicate that regulatory control take place in close proximity to the gene of interest, as large numbers of genes have been found to have significant *cis* eQTLs (9,10,11). As stated, GWAS has shown that these variants are predominantly found in non-coding regions of the genome where regulatory functions are largely uncharacterized. GWAS studies have been able to link

genetic loci to their resulting phenotypes but there is still improvement that can be done in understanding how these loci affect cellular phenotype (2,3).

In this study we bridge the gap in understanding of population specificity of *cis*-eQTLs by characterizing functional variation in human genomes by RNA-sequencing on lymphoblastoid cell line (LCL) of 421 individuals from the 1000 Genomes Project (1). We utilize previously published RNA-seq data across five populations: the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI) and mapped *cis*-eQTLS to transcriptome traits of mRNA. We found 476 *cis*-eQTLs located on all genes genome-wide and 178 *cis*-eQTLs when looking at population-specificty across all five populations.

Results and Discussion

Rs4275 was the most associated SNP identified (P = 4.244101e-22) and was located on gene ENSG00000260065.1. Figure 1 shows the GWAS region in which the SNP was located with the purple diamond representing rs4275. The left axis represents the -log10P, the right axis represents the recombination rate and the x-axis shows the position in Mb on the genome. Given that there are several variants in high LD with our SNP it would require further investigation to prove causation as one of the other several SNPs above the line of threshold could potentially be the true causal SNP.

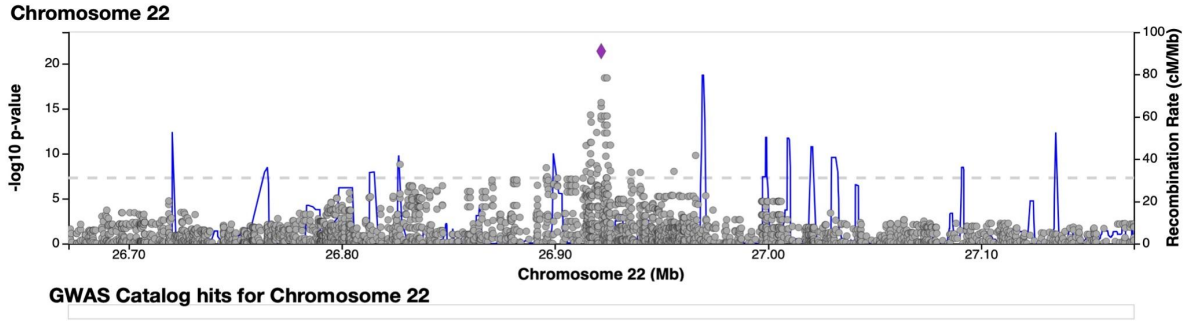

Figure 1

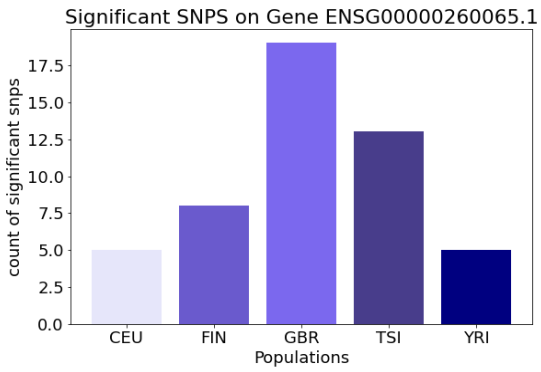

Figure 2

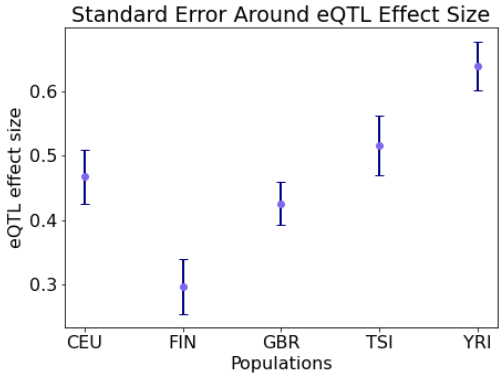

Figure 3

Looking at this gene further we are able to recognize how different each population is in as far as which variants are actually significant ($P < 5x10^{-8}$). In figure 2 we see the number of significant SNPs located on gene ENSG00000260065.1 and there is clearly a vast difference between the 5 populations. This extreme difference in associated SNPs points to the fact that genes from different populations do not act exactly similar to one another. In order to fully determine risk we must diversify research and compare causal variants across populations. Looking at SNP rs4275 in figure 3 we see the effect size and the standard error associated with each of the populations. The effect of rs4275 varies greatly between populations, again raising awareness that it is not enough to identify causal variants in just one population if we wish to assess risk on a broader scale.

Expanding to the data set as a whole, amongst the 476 cis-eQTLs discovered 178 of those were unique to a specific population. Figure 4 shows the count of cis-eQTLs found on chromosome 22 from each population. There is clearly a stark contrast between the amount of eQTLs discovered between the populations which again raises the concern that different populations express genes in very different ways. We can see this even more clearly looking at Figure 5 which depicts the number of unique eQTLs found in each population. This signifies that each population has its own subset of unique cis-eQTLs. This could indicate you cannot simply look for the presence of one variant to indicate risk in an individual, physicians would have to look for different variants depending on the patient's race and background. Knowing this, it is clear that diverse research must be done in order to understand what significant eQTLs might be casual in each population.

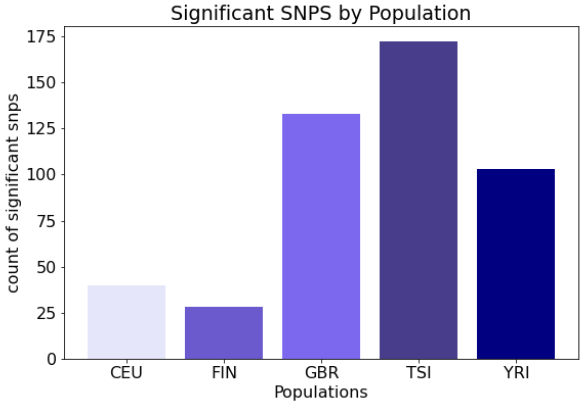

Figure 4

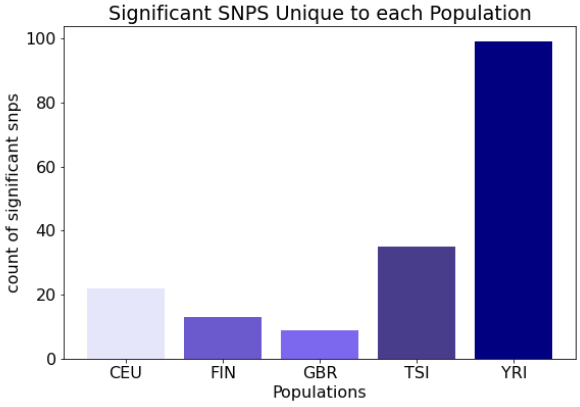

Figure 5

DATA

In order to conduct our analyses we used the phase1integrated_v3 data from the 1000 Genomes Project. Given the format of the genetic data it required a significant amount of preprocessing before the analyses could be conducted. We decided to run our analyses on

chromosome 22 and so we first filtered our vcf to only include genes and SNPs located on chromosome 22. This processing was done largely using *plink* which took in our vcf and filtered it to our specifications then outputted bed/bim/fam files which contained the SNPs and genotype information. When filtering we strictly kept SNPs, filtering out any variants with multiple Ref/Alt alleles. To further reduce the number of SNPs we are investigating we filtered for SNPs with a minor allele frequency(MAF) of $< .05$. After collecting the SNPs we would be using for analysis we combined this information with the GD462 gene expression data for each individual. This allowed us to conduct a linear regression based on the values associated with each individual. We also used the ALL_1000G_phase1integrated_v3 sample data to draw comparisons between each of the populations within our data.

METHODS

In order to identify the *cis*-eQTLs a linear regression was run on the data and a p-val was extracted to examine its significance. In order for an eQTL to be considered significant we set a threshold of $P < 5 \times 10^{-8}$. First, we established a baseline for how many cis-eQTLs are present in our data by looking at all genes genome wide. As mentioned above we found 476 cis-eQTLs located on all genes genome-wide. To run this regression, the scipy.stats library was used and every snp-gene pair was run through the regression model keeping only significant SNPs that had a p-value lower than the significance threshold.

After exploring the results of our baseline analysis we split our data up by the 5 super populations CEU, FIN, GBR, TSI, and YRI. We then ran the same regression on each isolated population to see how greatly the genome-wide results differed at a population level. 178 unique cis-eQTLs were discovered amongst the populations. Meaning that of those 178 eQTLs none were shared amongst populations but they were unique snp-gene pairs to that specific population. The counts for each population were as follows; CEU:22, FIN:13, GBR:9, TSI:35, YRI:99. Using this information we utilized Locuszoom and python to generate the plots present in this paper.

Citations

1. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65 (2012)
2. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678 (2007)
3. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012)
4. Emilsson, V. et al. Genetics of gene expression and its effect on disease. *Nature* 452, 423–428 (2008)
5. Stranger, B. E. et al. Population genomics of human gene expression. *Nature Genet.* 39, 1217–1224 (2007)
6. Grundberg, E. et al. Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nature Genet.* 44, 1084–1089 (2012)
7. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci. 2013 May 6;368(1620):20120362. doi: 10.1098/rstb.2012.0362. PMID: 23650636; PMCID: PMC3682727.
8. Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 10.1056/NEJMra0905980
9. Goring HH, et al. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208–1216 10.1038/ng2119
10. Schadt EE, et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107. 10.1371/journal.pbio.0060107
11. Dixon AL, et al. 2007. A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 10.1038/ng2109