Karthik Guruvayurappan, Madeline Lascola-Hutcherson, Andrew Li, Michael Nodini
DSC 180B
Quarter 2 Report

Comparing the Effects of Increased Sequencing Coverage on Analyses of Human Genetic Variation

Abstract
Genome-wide association studies determine associations between variants at individual genomic positions, called single nucleotide polymorphisms (SNPs) and various phenotypes such as height or diabetes. Expression quantitative trait loci (eQTLs) are SNPs that are significantly associated with gene expression for a certain gene. The link between genetic variation and observable phenotypes is poorly understood. eQTL analyses link genetic variation to gene expression, and these associations can then be used to inform how gene expression affects downstream phenotypes. These associations can be further refined using downstream analyses like fine-mapping or functional annotation enrichment which can pinpoint the location of causal variants across the genome. However, sequencing coverage, which is the amount of times an individual base pair is sequenced on average in a sequencing experiment, can have major implications affecting statistical power to detect SNPs, indels, and rare variants. In this analysis, we combine eQTL analysis with fine-mapping to compare high-coverage and low-coverage sequencing datasets for a cohort of individuals from the 1000 Genomes Project.

Introduction
Genome-wide association studies (GWAS) are a popular method for identifying inherited genetic variants associated with risk for certain phenotypes or disease traits.[13] These studies first survey the genome to determine which nucleotides individuals have at which genomic positions. The variations in these individual positions, called single nucleotide polymorphisms (SNPs) are then associated with diseases or phenotypes like lung cancer or height.[13] Genetic variations in non-protein coding regions of the genome can modulate expression levels of proteins while genetic variations in protein coding regions of the genome can alter proteins themselves. Both of these changes can affect functions of cells inside the body.[13] These small variations are responsible for phenotypes ranging from obesity to autoimmune diseases.[13] For example, sickle cell anemia stems from a mutation on a gene responsible for encoding the hemoglobin molecule on red blood cells.[14] Due to the mutation, red blood cells are formed in a sickle shape and can block blood flow leading to fatigue, pain, and organ dysfunction.[14] If disease risk is discovered early on in one's life preventative measures can be taken to decrease the likelihood of developing the disease or additional screenings can be scheduled to monitor an individual and catch the disease before it causes untreatable damage.[14] Understanding how genetic expression translates to the observable traits is a necessary step towards providing insights for disease risk, prevention, and heritability.

Expression quantitative trait loci (eQTLs) are single nucleotide polymorphisms (SNPs) that are significantly associated with gene expression. Previous studies have applied linear models to associate genotype with gene expression and call eQTLs.[1] The numerous resulting associations are significantly confounded by linkage disequilibrium (LD), and are therefore insufficient to determine causal relationships between SNPs and gene expression.[2] Linkage disequilibrium causes proximal genetic variants to be inherited together, meaning that individual variants are not inherited independently.[2] Further methods development is required to determine functional significance from eQTL mapping. Fine-mapping provides a suite of methods to determine high-confidence SNPs likely to include causal variants by incorporating linkage disequilibrium information, and can be performed with software tools such as susieR.[11]

There have been recent advances in the 1000 Genomes Project leading to a high-coverage whole-genome sequencing dataset for the 1000 Genomes Project cohort.[7] Sequencing coverage could have a significant

impact on the detection of SNPs and allele frequencies used for eQTL analysis. Reanalyzing eQTLs and causal variants with an increased coverage dataset could result in improved eQTL identification and better location of causal SNPs.

Results

To call eQTLs, we obtained genotype data from the 1000 Genomes Project for both standard coverage and 30x coverage for chromosome 22.[7] We then obtained previously published gene expression data from a subset of individuals in the 1000 Genomes Project.[1] Both the genotype and gene expression data were sequenced from lymphoblastoid cell lines.[1,7]

When calling significant cis-eQTLs by testing all possible SNPs within 500kb of a gene on the lower coverage dataset, we performed 1,802,930 association tests and identified 4,035 significant cis-eQTLs (p-value < 5e-8, Fig 1A). When performing a parallel analysis on the higher coverage dataset, we performed 2,151,453 association tests and identified 4,321 significant cis-eQTLs (p-value < 5e-8, Fig 1B).

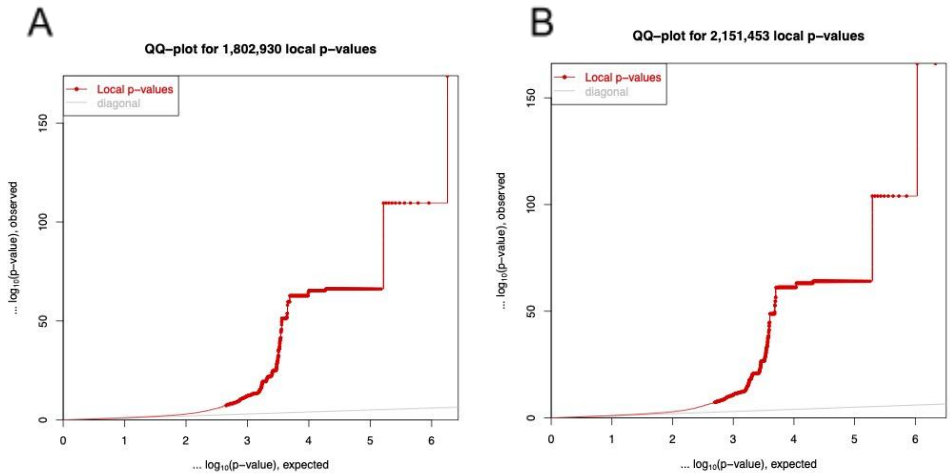

Figure 1. A) QQ-plot of p-values for 2,151,453 association tests performed using the higher coverage genotype data. 4,321 significant cis-eQTLs were identified (p-value < 5e-8). B) A QQ-plot of p-values for 1,802,930 association tests performed using the lower coverage genotype data. 4,035 significant cis-eQTLs were identified (p-value < 5e-8).

One locus with multiple significant cis-eQTLs was the *FAM118A* gene locus, which has previously been associated with inflammatory bowel disease.[15] To further investigate this locus, we analyzed p-values from univariate regression from SNPs within 500 kilobases of the *FAM118A* gene. In both the low coverage and high coverage dataset, we identify a cluster of SNPs that are significantly associated with *FAM118A* gene expression (-log10(p-value) > 8, Fig. 2A, Fig.2B).

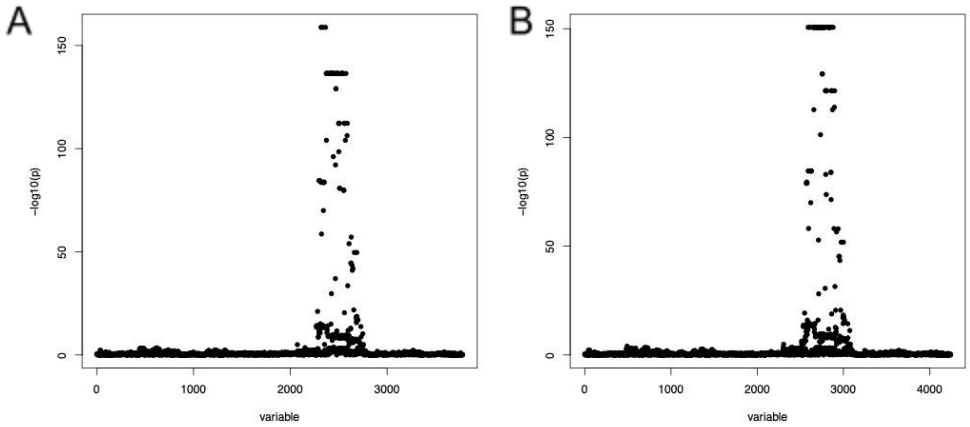

Figure 2. A) Distribution of p-values for association tests between SNPs within 500 kilobases of the *FAM118A* gene, ordered by genomic position, for the low coverage dataset. B) Distribution of p-values for association tests between SNPs within 500 kilobases of the *FAM118A* gene, ordered by genomic position, for the high coverage dataset.

We next fine-mapped these SNP-gene associations for *FAM118A* using the susieR R package (Methods).[11] When identifying 95% credible sets with these SNP-gene associations, one credible set was identified for both the lower coverage and higher coverage datasets. For the lower coverage dataset, three candidate causal SNPs were identified (Fig 3A). For the higher coverage dataset, 35 candidate causal SNPs were identified (Fig 3B).

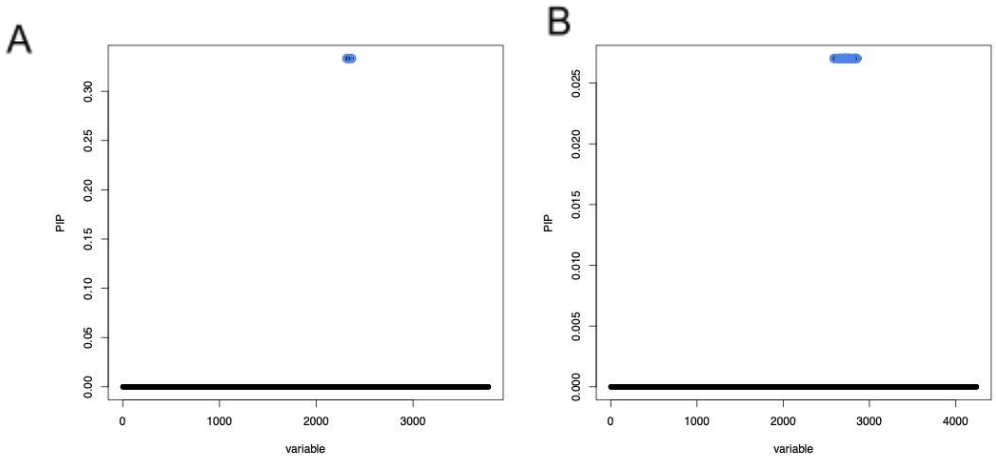

Figure 3. A) Fine-mapped posterior inclusion probabilities for SNPs proximal to the *FAM118A* gene for *FAM118A* gene expression for the lower coverage data. Blue dots indicate candidate causal SNPs in the 95% credible set. B) Fine-mapped posterior inclusion probabilities for SNPs proximal to the *FAM118A* gene for *FAM118A* gene expression for the higher coverage data. Blue dots indicate candidate causal SNPs in the 95% credible set.

Discussion and Future Directions

In this work, we develop a framework for determining significant SNP-gene associations and fine-mapping these associations to identify candidate causal variants for gene expression. We apply this framework to both low and high coverage genotype datasets obtained from the 1000 Genomes Project. In this work, we identify a greater number of eQTLs using higher coverage data than lower coverage data.

When performing fine-mapping for SNPs near the *FAM118A* gene locus, we find that a higher coverage dataset identifies a larger set of potential causal SNPs than the lower coverage dataset. However, both analyses lead to candidate SNPs with the same posterior inclusion probability, indicating that a single causal SNP could not be identified from both analyses.

This general framework could be applied to genome-wide eQTL and fine-mapping studies to determine how experimental design parameters like sequencing coverage can impact downstream analyses of genetic variation, and can be weighed against factors like experimental cost.

One of the major drawbacks of our analysis is the presence of population-level structures within our data. These population-level structures could confound our gene expression information, and therefore affect the cis-eQTL associations that we discover. To improve this analysis, in both our eQTL mapping and fine-mapping, we could include PEER factors to account for these structures.[1,17]

Another future direction for this work is to compare our fine-mapping results with functional enrichment analyses using ATAC-seq data from the Roadmap Epigenomics Consortium, and relating eQTLs to biological pathways using gene set enrichment analysis (GSEA).[6,10]

Methods

Raw VCF files were downloaded from the 1000 Genomes Project website, using VCFs from GRCh38 and GRCh38 with 30x coverage.[7] The normalized RNA-sequencing matrix was obtained from the European Bioinformatics Institute BioStudies website associated with the 1000 Genomes RNA-seq paper.[1]

Raw VCF (variant call file) processing was performed using the plink package.[8] For processing the VCF files, filtering steps were applied to only include biallelic variants and variants above a minor allele frequency threshold of 0.05. The VCF files were then used to produce a genotype matrix for Matrix eQTL and susieR input using the 'recodeA' option in plink.

Gene coordinates were converted from the GRCh37 genome assembly to the GRCh38 genome assembly using UCSC LiftOver (https://genome.ucsc.edu/index.html)[12] LiftOver was run using the 'liftOver' script with the hg19 to hg38 chain file, available through the UCSC Genome Browser website.

eQTL calling was performed using the Matrix eQTL R package.[9] The eQTL calling follows a linear regression. The formula for linear regression is as follows: $y = Xb + \epsilon$ where X represents a genotype matrix containing 0s, 1s, or 2s, to indicate an allele frequency for a given SNP for a given individual. b is a vector of coefficients that are the effect sizes associated with a specific SNP. y is a vector of gene expression values for each sample in the dataset for a given gene. For calling eQTLs a p-value threshold of 5e-8 was used, and all SNPs within 500kb of a gene were tested for significant associations.

Fine-mapping for causal variants was performed using the susieR package.[11] SusieR applies a method named "Iterative Bayesian Stepwise Selection" using a sum of single effects regression approach to finemap SNP-gene associations. With genotypes directly provided, susieR infers linkage disequilibrium statistics from the genotype matrix.[11]

Code Availability

Code for this project is available through our GitHub repository: https://github.com/somet3000/1kgp-coverage-analysis

References

1. Lappalainen, T., Sammeth, M., Friedländer, M. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013). https://doi.org/10.1038/nature12531
2. Pritchard, J. K., and M. Przeworski. 2001. "Linkage Disequilibrium in Humans: Models and Data." *American Journal of Human Genetics* 69 (1): 1–14.
3. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014 Oct;198(2):497-508. doi: 10.1534/genetics.114.167908. Epub 2014 Aug 7. PMID: 25104515; PMCID: PMC4196608.
4. Amariuta T, Luo Y, Gazal S, Davenport EE, van de Geijn B, Ishigaki K, Westra HJ, Teslovich N, Okada Y, Yamamoto K; RACI Consortium, GARNET Consortium, Price AL, Raychaudhuri S. IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. Am J Hum Genet. 2019 May 2;104(5):879-895. doi: 10.1016/j.ajhg.2019.03.012. Epub 2019 Apr 18. PMID: 31006511; PMCID: PMC6506796.
5. Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel... [et Al.]* 109 (January): 21.29.1–21.29.9.
6. Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences* 102 (43): 15545–50.
7. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, Fairley S, Runnels A, Winterkorn L, Lowy E; Human Genome Structural Variation Consortium; Paul Flicek, Germer S, Brand H, Hall IM, Talkowski ME, Narzisi G, Zody MC. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022 Sep 1;185(18):3426-3440.e19. doi: 10.1016/j.cell.2022.08.004. PMID: 36055201; PMCID: PMC9439720.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75. doi: 10.1086/519795. Epub 2007 Jul 25. PMID: 17701901; PMCID: PMC1950838.
9. Shabalin, Andrey A. 2012. "Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations." *Bioinformatics* 28 (10): 1353–58.
10. Roadmap Epigenomics Consortium., Kundaje, A., Meuleman, W. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015). https://doi.org/10.1038/nature14248
11. Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society: Series B (Statistical Methodology).
12. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.
13. Uffelmann, E., Huang, Q.Q., Munung, N.S. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9
14. Kato, G., Piel, F., Reid, C. *et al.* Sickle cell disease. *Nat Rev Dis Primers* 4, 18010 (2018). https://doi.org/10.1038/nrdp.2018.10
15. Robinson, P., Leo, P., Pointon, J. *et al.* Exome-wide study of ankylosing spondylitis demonstrates additional shared genetic background with inflammatory bowel disease. *npj Genomic Med* 1, 16008 (2016). https://doi.org/10.1038/npjgenmed.2016.8
16. Stegle, O., Parts, L., Piipari, M. *et al.* Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 7, 500–507 (2012). https://doi.org/10.1038/nprot.2011.457