Using integrative estimates of deleteriousness and regulatory potential to identify causal eQTLs

Ronald J. Hause Jr., Benjamin Weaver, Martin Kircher, Jay A. Shendure, and Gregory Cooper
1 Department of Genome Sciences, The University of Washington, Seattle, WA, 98195, USA, 2 HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35006, USA, 3Department of Genetics, University of Alabama-Birmingham, 35232, USA.

Expression quantitative trait loci (eQTL) mapping has identified DNA variation associated with differences in mRNA expression and has previously been used as an effective strategy for prioritizing genetic variation that may play a role in complex phenotypes. However, current eQTL analyses are limited in their ability to identify causal variants, because most single nucleotide variants (SNVs) identified as eQTLs are likely tag SNVs in linkage disequilibrium (LD) with the true functional variant altering gene expression. Furthermore, eQTLs can be context specific, and understanding the determinants distinguishing “global” from tissue-specific eQTLs is critical to better interpret GWAS results in their most relevant contexts. Recent studies have demonstrated that incorporating regulatory annotations can improve power to identify putatively functional candidate variants and reproducible eQTLs across studies and between tissues. We have recently developed a framework, Combined Annotation Dependent Depletion (CADD), that integrates different genomic annotations into a single C-score predicting the deleteriousness of any particular SNV. These annotations include metrics of chromatin accessibility, conservation scores, and gene model information, among others. C-scores are associated with annotations of functional importance, pathogenicity, and expression data to measure regulatory effects. Despite the substantial improvements made to prioritize eQTL discovery and effectively model replication across studies and tissues, many strong eQTLs still fail to replicate in independent cohorts. Therefore, we hypothesized that prioritizing variants on our composite metric of genome functionality could help to further enrich for causal variants involved in gene regulation.

Introduction

The goals of this project were to 1) further examine metrics related to eQTL reproducibility across studies and tissues and 2) to assess whether C-scores can improve power to detect reproducible and/or tissue-specific eQTLs along with cell type-specific cis-regulatory element information from ENCODE. To achieve this goal, we systematically analyzed six datasets where eQTLs have previously been performed.

<table>
<thead>
<tr>
<th>Study</th>
<th>Tissue</th>
<th>Samples</th>
<th>Genes</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pickrell</td>
<td>LCL</td>
<td>60</td>
<td>20863</td>
<td>RNA-Seq</td>
</tr>
<tr>
<td>Stranger</td>
<td>LCL</td>
<td>60</td>
<td>15757</td>
<td>Illumina WG-6 v1</td>
</tr>
<tr>
<td>Myers²</td>
<td>Brain</td>
<td>193</td>
<td>11707</td>
<td>Sentrix HumanV6x1</td>
</tr>
<tr>
<td>Zhang²</td>
<td>Brain</td>
<td>153</td>
<td>17377</td>
<td>Affymetrix HG-U133 v2</td>
</tr>
<tr>
<td>Innocenti²</td>
<td>Liver</td>
<td>206</td>
<td>16236</td>
<td>Agilent 4x44K</td>
</tr>
<tr>
<td>Kathiresan¹</td>
<td>Liver</td>
<td>60</td>
<td>10877</td>
<td>Illumina HumanMethylation450 v2</td>
</tr>
</tbody>
</table>

Data:
- DNase hypersensitivity sites from GM12878, SK-N-SH, and H4G2 cell lines
- TRFS clusters (V1) from ENCODE data uniformly processed by ENCODE Analysis Working Group
- GWAS SNPs from the NHGRI and curated variants from the NHGRI ClinVar database

Methods:
- Expression and genotype data were QCed and SNPs imputed against 1000 Genomes (1kG) variants
- Association mapping was performed by regressing surrogate variable-corrected, inverse normal-transformed gene-level expression measurements against minor allele frequency (MAF) copy number for each individual for all SNPs within 1MB of the transcription slack site (TSS)
- Pickrell, Myers, and UC studies were used as eQTL discovery cohorts, and the Zhang, UC, and UW studies were used as the corresponding eQTL replication cohorts
- eQTL replication is here defined as concordant effect direction and nominal significance for the subset of genes and common SNPs overlapping between studies and tissues
- Fisher’s exact tests were used for enrichment analyses and wilcoxon rank sum tests for differences between CADD score distributions
- C-scores and eQTLs

Previously, our lab has demonstrated that CADD scores are more strongly correlated with experimentally measured effects on transcriptional regulatory function of two enhancers and one promoter as measured by saturation mutagenesis and massively parallel reporter assays than other scores (upper left), indicating their relevance in potentially predicting genome functionality. This functional relationship was also observed for common eQTLs, with C-scores for eQTLs in LCLs are approximately 1.6 times higher than randomly selected, MAF- and location-matched SNVs from 1000 Genomes (P < 2.2x10^-10) (upper right).

Conclusions

- In summary, we observed that eQTLs that reproduce across studies and tissues are enriched to lie close to the TSS, overlap with DNase sites and TFS, and have overall higher integrative measures of deleteriousness as predicted by CADD.
- Additionally, prioritizing variants that are both eQTLs and predicted to be deleterious by CADD dramatically increased the odds of being GWAS-associated variant by four-fold relative to using eQTLs alone.
- After conditioning on distance to TSS or overlap with cis-regulatory elements, C-score improves eQTL reproducibility, indicating that the additional constraints that are put into CADD can help researchers in further informing not only likely eQTLs but also reproducible in eQTLs that can be further dissected into their underlying eQTNs.
- Our approach will be useful for not only prioritizing the most likely causal variants underlying eQTLs but also for better elucidating the functional mechanisms underlying GWAS results and eQTLs in a variety of contexts.

Future Directions

- Incorporate C-score as a prior to inform posterior eQTN probabilities in a Bayesian framework
- Integrate median CADD scores across LD-blocks associated with eQTLs
- Functionally validate predicted causal eQTNs using CRISPR/Cas9 or site-directed mutagenesis
- Jointly estimate global vs. tissue-specific and reproducible eQTLs rather than thresholding
- Create classifier incorporating C-scores to predict global and tissue-specific eQTLs
- Extend into other endophenotypes (eQTLs, miQTs, etc.)

References and Acknowledgements


We would like to thank Darren Cusanovich, Aaron McKenna, and Evan Boyle for helpful discussions. Our work was supported by grant DPH5K077613 from the NHGRI (S.J.).