# Are Genetic Interactions Influencing Gene Expression Evidence for Biological Epistasis or Statistical Artifacts?

Alexandra E. Fish[1], John A. Capra[1,2,3], William S. Bush[3,4*]

1. Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee, United States of America
2. Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, United States of America
3. Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, United States of America
4. Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, United States of America

*Corresponding Author

E-mail: willbush@case.edu

**Abstract**

Interactions between genetic variants, also called epistasis, are pervasive in model organisms; however, their importance in humans remains unclear because statistical interactions in observational studies can be explained by processes other than biological epistasis. Using statistical modeling, we identified 1,093 interactions between pairs of *cis*-regulatory variants impacting gene expression in lymphoblastoid cell lines. Factors known to confound these analyses (ceiling/floor effects, population stratification, haplotype effects, or single variants tagged through linkage disequilibrium) explained most of these interactions. However, we found 15 interactions robust to these explanations, and we further show that despite potential confounding, interacting variants were enriched in numerous regulatory regions suggesting potential biological importance. While genetic interactions may not be the true underlying mechanism of all our statistical models, our analyses discover new signals undetected in standard single-marker analyses. Ultimately, we identified new complex genetic architectures regulating 23 genes, suggesting that single-variant analyses may miss important modifiers.

## Introduction

The vast majority of variants identified by genome-wide association studies (GWAS) are non-protein-coding, which implies variants impact disease risk by altering regulatory DNA regions that control gene expression levels (Hindorff et al. 2009; Schaub et al. 2012). Several functional analyses of the mechanisms underlying SNP-trait associations identified by GWAS illustrate this principle. For example, a variant associated with low density

lipoprotein (LDL-C) levels creates a functional transcription factor binding site that alters the expression of *SORT1* (Musunuru et al. 2010). *Sort1* levels causally regulate LDL-C levels in mice (Musunuru et al. 2010). Similarly, a variant associated with BMI alters an enhancer that regulates *IRX3* levels, which causally impact BMI in mice (Smemo et al. 2014). As similar mechanisms likely result in other clinical phenotypes, there have been numerous studies of genetic variants associated with changes in gene expression, called expression quantitative loci (eQTL). By understanding the genetic underpinnings of gene regulation, we may elucidate disease mechanisms, identify at-risk populations, and identify novel targets for treatment.

On a molecular level, transcription is an intricate process that requires multiple transcription factors to assemble upon regulatory DNA regions (e.g., promoters and enhancers) that must act together to regulate gene expression levels. Most eQTL studies exclusively analyze variants within the *cis*-regulatory region of the regulated gene, called *cis*-eQTL (Veyrieras et al. 2008; Grundberg et al. 2012; Price et al. 2011). However, the statistical models used for these analyses do not capture the complexity of transcription because they analyze each variant individually, which assumes that the greater genomic context does not play a role in how a variant impacts the phenotype. Relationships between variants and their genomic context can be taken into account by analyzing multiple genetic variants within the same statistical model. This is a critical improvement when the effect of a genetic variant is dependent on the on the presence or absence of other variants, a phenomenon referred to as a statistical interaction (Cordell 2002), or epistasis.

There is strong evidence for interactions in model organisms: approximately half of all transcripts appear to be regulated by interactions in both *Saccharomyces cerevisiae* (Brem et al. 2005) and *Drosophila melanogaster* (Gibson et al. 2004). Rare diseases in humans are caused by interactions; for instance, fatal familial insomnia and familial Cruetzfeldt-Jakob disease are both caused by the D178N substitution in *PRNP*, but which disease develops is determined by a common variant at position 129 (Capellari et al. 2011). However, evidence of interactions for common traits in humans is elusive. While many have been reported, the majority of these studies do not attempt to replicate their findings, fail to replicate their findings, or can be explained by a variety of factors unaccounted for in the study (Wei, Hemani, and Haley 2014). Regardless, they do not find evidence for interactions on the scale reported in model organisms. This discrepancy may be explained if interactions are only observable within tightly

controlled genetic and experimental conditions that are not feasible for studies of most human phenotypes, but are feasible for gene expression analyses in human-derived cell lines. Thus, the study of interactions between genetic variants influencing gene expression (ieQTL) is a unique opportunity to investigate both the prevalence of interactions within humans and to better understand the genetic etiology of gene regulation.

While prior studies on ieQTL in humans have been conducted, they have not sufficiently accounted for alternative explanations for statistical interactions. It has long been known that statistical interactions can be produced by processes other than true biological epistasis, including: type I errors, technical artifacts (i.e., ceiling/floor effect), statistical artifacts (i.e., population stratification), and other biological mechanisms (i.e., haplotype effects). Some of these issues were not explicitly addressed in the ieQTL studies conducted by Turner and Bush (2011), Becker et al. (2012), and Fitzpatrick et al. (2015); furthermore, these studies did not replicate or otherwise functionally validate their findings. Hemani et al. (2014) and Brown et al. (2014) used stringent experimental designs to take precaution against many of these known issues in their study of ieQTL, and identified and replicated ieQTL in humans. In response to these findings, Wood et al. demonstrated that a non-epistatic mechanism was capable of producing statistical interactions and accounted for the vast majority of ieQTL identified by Hemani et al. (Wood et al. 2014) Essentially, Wood et al. demonstrated that two interacting variants, while in low linkage disequilibrium with one another, could jointly tag a single variant eQTL through LD. As a result of these studies there is growing interest in ieQTL, but also growing concern that identified associations are not representative of any underlying biological epistasis but are instead artifacts of the statistical models used and complex LD patterns.

In this study, we investigated the evidence for ieQTL after accounting for mechanisms other than biological epistasis that are capable of producing statistical interactions. We first identified 1,093 interactions between variants within the *cis*-regulatory region for 11,465 genes with expression data in lymphoblastoid cell lines (LCLs). We then determined whether the interactions could be explained by a nuanced form of population stratification, ceiling/floor effects, haplotype effects, or the tagging of *cis*-eQTL through LD. Ultimately, 15 interactions could not be accounted for by any of these mechanisms, which suggests they represent true biological epistasis. This is a lower bound on the number of ieQTL, since interactions consistent with multiple explanations may be caused by biological epistasis. Indeed, we used functional genomics to provide corroborative evidence for the biological plausibility of

many additional ieQTL. Interacting variants were enriched in promoters, enhancers, and numerous transcription factor binding sites, including CTCF and cohesin. Furthermore, we demonstrate that many of the interactions consistent with other biological explanations represent complex genetic architectures that would have gone undetected in a single-marker analysis using a genome-wide significance threshold. Given this evidence, we conclude that interaction analyses identify novel biological associations; however, careful experimental design and examination of results is required before inferring statistical interactions represent biological epistasis.

## Results

### Discovery and replication of genetic interactions that impact gene expression levels.

We identified interactions between nominal *cis*-eQTL that were significantly associated with gene expression levels. Our analysis was conducted using 210 individuals from the HapMap Project, Phase II, on whom both genotyping (Frazer et al. 2007) and gene expression data within LCLs (Stranger et al. 2007) were available. The overall workflow for the analysis is shown in Figure 1. For each gene with expression data (n=11,465), we identified common SNPs (MAF > 5%) within its *cis*-regulatory region, defined as 500 kb upstream to 500 kb downstream of the gene. To increase power, we only considered variants nominally associated with the genes expression ($p < 0.05$) in a single-marker analysis (Veyrieras et al. 2008). We analyzed all pair-wise combinations of these variants for each gene, resulting in over 21 million SNP pairs. We then performed a likelihood ratio test (LRT) comparing a full model, which contains covariates, main effects, and interaction terms, to a reduced model, containing only the covariates and main effects, to determine which interactions significantly improved model fit (Cordell 2002).

Given the complex nature of the analysis, the appropriate strategy for multiple testing correction required careful consideration. A Bonferroni correction ensures that the probability of a single false positive amongst all performed association tests is $\leq 0.05$, which is appropriate when very few loci are anticipated to have an association with the phenotype (Storey and Tibshirani 2003). Given the prevalence of ieQTL in prior studies, this is an inappropriate assumption for our analysis. We therefore calculated a false discovery rate, which uses the discrepancy between the observed distribution of p-values and the expected null distribution to estimate the proportion of true positives. We calculated an FDR of 5% ($p \leq 1.328$x10-5) using the p-values from all LRT performed in the discovery analysis using Storey's method (2015). We considered all interactions passing this threshold significant.

LD between variants complicates the interpretation of the interaction models. We addressed two types of LD in significant interaction models: within-pair LD, defined as the LD between the variants in the same interaction model, and between-pair LD, defined as the LD between variants in different interaction models. Modest within-pair LD indicates the variants are may be on the same haplotype, which may carry other variants that drive the association with gene expression; consequently, we removed pairs in modest LD with one another ($r^2 > 0.6$) from the remainder of the analysis. 5,439 interaction models were both significant and passed the within-pair LD filtering criteria; they were significantly associated with the expression of 165 unique genes (Dataset S1). The median $r^2$ between variants in these interaction models was 0.06 (Figure 1 – Figure Supplements 1). We then calculated between-pair LD, or the correlation of variants in different interaction models. Highly correlated interaction models were grouped together (Methods, Figure 1) because they likely represent the same pair of interacting genomic loci, as evidenced by their very similar statistical models (Figure 1 – Figure Supplements 2). The 5,439 interaction models represented 1,093 pairs of interacting genomic loci (Dataset S1). The interaction model with the most significant p-value in the discovery analysis was selected to represent the entire group in all subsequent analyses, unless specifically stated otherwise.

Next, we performed a replication analysis using an independent dataset of 232 unrelated individuals from the 1000 Genomes Project who had both whole-genome sequencing (The 1000 Genomes Project Consortium, 2012) data and gene expression levels in LCLs (Stranger et al. 2012) available. All ieQTL composed of variants that were common (MAF > 5%) and had available genotyping data were tested for significant interactions with the same procedure used in the discovery analysis. Of the 778 ieQTL able to be tested, 335 had p-values < 0.05 and 90 passed a Bonferroni multiple testing correction for all tests performed in the replication analysis. We considered all ieQTL models with LRT p-values < 0.05 to be successfully replicating.

**Exploration of Alternative Mechanisms Capable of Producing Interactions**

Statistical interactions can be produced from a variety of processes other than biological epistasis, including technical artifacts, statistical artifacts, and other biological processes captured through LD patterns. Technical artifacts are caused limitations of the data itself; for instance, limitations in the dynamic range of measureable gene

expression can result in interactions being identified through the ceiling/floor effect. Statistical artifacts are caused by improper applications of statistical methodology; for example, analyzing multiple ethnicities together can produce spurious associations known as population stratification. Technical and statistical artifacts are especially troubling since they are unlikely to represent real biological association between the loci and phenotype.

Other biological phenomena, namely haplotype effects and *cis*-eQTL effects, can be captured by interaction analyses due to LD patterns. We investigated whether the observed significant ieQTL models could be explained by each of these phenomena.

*Limitations in dynamic range may produce statistical interactions.*

The gene expression data used in this analysis was collected using microarrays. Microarray technology has a limited dynamic range, meaning that the upper and lower bound on the level of gene expression that microarrays can detect does not cover the full range observed in nature. When the observed range of gene expression values is limited due to technical constraints, variants with sufficiently large main effects may mask the main effects of other variants in the model if their combined effect exceeds the range limitation. This phenomenon, referred to as the ceiling/floor effect, may result in the identification of spurious interactions. Interactions caused by the ceiling/floor effect have a characteristic pattern, in which the main effects of both variants have the same direction of effect and the interaction terms are in the opposite direction. For example, both main effects may increase gene expression, but the interactions will decrease gene expression. An example of an interaction putatively caused by the ceiling effect is shown in Figure 2. Of 1,093 locus pairs, 99 exhibited a pattern consistent with the ceiling/floor effect. Since transcript production may also have a true biological ceiling, it is possible that true genetic interactions could product this pattern; consequently, we consider this an upper bound of the influence of ceiling/floor artifacts within our analysis.

*Population specific eQTLs may produce statistical interactions.*

In our discovery and replication analyses we analyzed multiple ethnicities together, which raises the concern of spurious interaction signals due to population stratification. Traditionally, population stratification refers to the spurious results identified when two ethnicities with differences in both the distribution of genotypes and phenotypes are analyzed together. In the population normalization procedure applied to the gene expression data, we

removed systematic differences in the expression of each gene between ethnicities. This enables to us analyze multiple ethnicities together without incurring spurious results from the traditional conception of population stratification, and is an approach used by several other studies (Becker et al. 2012; Veyrieras et al. 2008). We also controlled for the top three principal components in our analysis to adjust for residual ethnicity-dependent effects. Even though we have protected against population stratification, we performed a stratified analysis in our discovery dataset. We tested each of the 1,093 ieQTL pairs for significant interactions within each of the three discovery ethnicities (CEU, YRI, and CHB+JPT) separately. Despite a substantial reduction in power to detect effects, 826 of 1,093 ieQTL were nominally significant ($p < 0.05$) in at least one population, demonstrating that they are not attributable to population stratification.

Given our precautions against population stratification, it was surprising that 267 ieQTL did not remain nominally significant in the stratified analysis. We further investigated these ieQTL and identified a more nuanced mechanism through which population stratification could produce spurious results in interaction testing. We found that many of these interactions were between variants that were population-specific *cis*-eQTL (Stranger et al. 2012), meaning they were present in all populations but operated as a *cis*-eQTL in only a subset. The systematic differences within populations between the main effect of each variant and the frequency of two-locus genotype combinations resulted in the identification of spurious interactions. An example of how population-specific *cis*-eQTL produced statistical interactions is provided in Figure 3. The population-specific *cis*-eQTL mechanism could account for 238 of the 267 that failed in the stratified analysis. This does not impact the 826 ieQTL that remained nominally significant within the stratified analysis, as the interaction was observed within at least one population.

*IeQTL may capture haplotype effects through LD.*

In some LD architectures, a combination of two variants can identify haplotypes. While there is evidence to suggest haplotypes form in response to biological interactions between variants (Lappalainen et al. 2011), haplotypes may simply be carrying other variants that additively regulate gene expression. Figure 4 illustrates how additional variants carried on the haplotype may result in statistical interactions. Consequently, interactions between variants on the same haplotype cannot be used to demonstrate the existence of ieQTL in this analysis. As previously stated, we removed all interaction models composed of variants in modest LD with one another as assessed by $r^2$ ($r^2 < 0.6$)

from all portions of the study. We additionally investigated whether or not variants within the same interaction model were in modest LD with one another as measured by D' (D' < 0.6). Of the 1,093 interacting loci, 776 had D' values < 0.6. The distribution of LD statistics, both $r^2$ and D', for interaction models is shown in Figure 1 – Figure Supplements 1.

*Single cis-eQTL may be tagged by statistical interactions.*

Wood et al. recently demonstrated that all ieQTL identified by Hemani et al. could be explained by the effects of *cis*-eQTL (Wood et al. 2014). This can occur when the two interacting SNPs together tag a single *cis*-eQTL, which is possible even if the interacting SNPs are in low LD with one another. We addressed this concern by conditioning all interactions on all nominal *cis*-eQTL identified for the regulated gene. We identified *cis*-eQTL in a subset of individuals from our discovery dataset (n=174) with sequencing data available through the 1KG project to ensure we had the most comprehensive list of *cis*-eQTL. All common variants (MAF > 5%) within the *cis*-regulatory region that were nominally associated ($p < 0.05$) with gene expression were considered *cis*-eQTL. We then created all pairs of *cis*-eQTL and ieQTL for the same gene. We performed a conditional analysis for each of these combinations, in which the additive and dominant main effect for the *cis*-eQTL were incorporated into both the full and reduced model used in the LRT to determine the significance of the interaction. 130 of the 958 testable ieQTL remained significant ($p < 0.05$) in all conditional analyses performed, indicating that these interactions cannot be explained by *cis*-eQTL. Interactions may not have been significant in the conditional analysis if they were tagging *cis*-eQTL (illustrated in Figure 5), as suggested by Wood et al., or if power to detect their effects was substantially reduced due to the simultaneous addition of covariates and reduction in sample size.

*IeQTL cannot be entirely accounted for by alternative mechanisms*.

Finally, we assessed the cumulative impact of alternative explanations on interaction models (Dataset S2). Of the 1,093 interacting genomic loci identified, 355 had statistical characteristics consistent with either technical or statistical artifacts. If these interactions are caused by artifacts, they may not represent any biological process at this locus. 179 of 738 remaining ieQTL successfully replicated; these represent robust signals that are likely tagging some biological process. Biological explanations other than epistasis – namely haplotype effects of the tagging of *cis*-eQTL – could account for 164 of the 179 remaining interactions. Ultimately, 15 interactions (Table 1) replicated

and could not be explained by the ceiling/floor effect, population stratification, haplotype effects, or the tagging of *cis*-eQTL. Notably, each alternative explanation removed unique interaction models, highlighting that all these issues need to be considered to future interaction analyses.

**IeQTL analyses identify biological effects that would not be detected in single-marker analyses.**

Since the vast majority of observed statistical interactions between variants can be explained by either artifacts or other biological processes, it is natural to question the utility of performing ieQTL analyses. However, ieQTL analyses may still be useful if they can identify biological associations between loci and phenotypes that would have been undetected in a single-marker analysis. This certainly occurs when true biological epistasis is present; however, it may also occur if the single-marker effects are too nominal to be detected or if other complex genetic architectures underlie the association. To investigate the utility of interaction analyses, we looked exclusively at the 170 interactions which successfully replicated and were inconsistent with either technical or statistical artifacts. As previously stated, we found evidence that 15 of these ieQTL represent true biological epistasis; the remaining 155 could be accounted for by either haplotype effects or *cis*-eQTL in the conditional analysis. To determine if these other biological phenomena would have been detected in a single-marker analysis, we identified the *cis*-eQTL that most accounted for the interaction in the conditional *cis*-eQTL analysis. We then determined its significance in a single-marker analysis (Methods), and plotted this against the significance of the interaction term in the conditional analysis (Figure 6).

The resulting figure can be divided into four quadrants, based on both the significance of the *cis*-eQTL in the single-marker analysis and the significance of the interaction term in the conditional analysis. Interactions in the top left and top right quadrant still explained a significant portion of variability in gene expression when the *cis*-eQTL was taken into account. Those in top right quadrant were most mediated by a *cis*-eQTL that would have been detected using a genome-wide Bonferroni threshold for multiple testing ($p < 5x10^{-8}$), and those in top left quadrant were most mediated by *cis*-eQTL that did not reach genome-wide significance. In addition to the 15 interactions that likely represent true biological epistasis, these two quadrants contained 8 additional interactions that remained significant but were composed of variants in modest LD ($D' > 0.6$) with one another, and therefore consistent with haplotype effects. Interactions in the bottom right quadrant were completely eliminated by a highly significant *cis*-eQTL that

would have been detected using a genome-wide Bonferroni correction. These interactions may be capturing the effects of strong *cis*-eQTL through LD patterns, as suggested by Wood et al. Finally, the 39 interactions in the bottom left quadrant could be entirely accounted for by a *cis*-eQTL that would not have been detected using a single-marker analysis with a genome-wide Bonferroni correction. For these interesting models, it is unclear if the main effect of the *cis*-eQTL or the interacting SNP pair are the true causal factors. Overall, the interaction analysis identified 23 associations between loci and gene expression that were not fully accounted for by single-marker analyses, and an additional 39 associations between loci and gene expression that, while perhaps mediated by a single-variant, would not have been identified in a typical single-variant analysis. Altogether, interactions regulated the expression of 23 unique genes that would have not been detected in a traditional single marker analysis.

**IeQTL variants occur within known regulatory elements and may impact chromatin looping.**

Many ieQTL were consistent with multiple explanations; unfortunately, with current models, we cannot definitely determine the causal explanation statistically. Functional genomics data, however, offers an independent insight into whether or not interacting variants plausibly regulate gene expression. All 5,439 interaction models were used in this analysis rather than the representative 1,093 interaction models because we do not know which specific SNP-SNP interaction is causal, and while the statistical properties of ieQTL grouped together are very similar, they each have different regulatory annotations. We found that interacting variants were enriched (compared to all nominal *cis*-eQTL tested) in known regulatory regions identified by the ENCODE project in LCLs (Dataset S3) (Kellis et al. 2014), including: regions of open chromatin identified by DNase I hypersensitivity (OR: 1.85; $p = 2.73x10^{-29}$) and FAIRE peaks (OR: 2.32; $p = 4.80x10^{-44}$); predicted promoters (OR: 2.45; $p=5.57x10^{-91}$) and enhancers (OR: 1.30; $p=3.00x10^{-5}$); and within the binding peaks for 32 of 60 transcription factor assays. Notably, the most significantly enriched transcription factors have known functions in LCLs: RFX5 (OR: 8.02; $p = 3.04x10^{-187}$) activates transcription at MHC class II promoters; POU2F2 (OR: 4.07; $p=1.19x10^{-232}$) regulates immunoglobin genes; and STAT3 (OR: 7.84; $p=4.81x10^{-80}$) is essential for T-cell differentiation and the interferon response. There was a very significant enrichment of ieQTL SNPs within RNA polymerase II (POL2) binding peaks, a trend observed across all five POL2 assays performed within LCLs by the ENCODE project (OR ranged from 2.26 to 5.28). Interacting variants were also enriched within the binding sites for cohesin (RAD21 and SMC3) and CTCF, which co-localize

to regulate chromatin looping. Thus, interacting variants are enriched within regulatory elements, which may physically interact with one another through chromatin looping to produce epistatic effects.

**Discussion**

The impact of genetic interactions on complex human phenotypes has been the subject of much speculation and study. The systematic investigation of genetic interactions impacting a low-level phenotype has been enabled by the availability of human cell lines with comprehensive genetic and gene expression data. Several studies have replicated interactions influencing gene expression; however, they have faced scrutiny for legitimate reasons: statistical interaction models are vastly more complicated to interpret compared to single-variant associations, and are subject to confounding factors that limit their inference to true biological mechanisms. In this study, we performed a focused *cis*-regulatory genetic interaction analysis and attempted to comprehensively account for confounding factors that have not been addressed by other published studies. After they had been accounted for, we still observed evidence supporting the existence of interactions influencing gene expression in humans.

Confounding processes had a profound effect on the results from our analysis - the vast majority of interactions we identified were consistent with at least one other alternative explanation in addition to biological epistasis. Moreover, interactions consistent with these alternative explanations (i.e., haplotype effects, population stratification, ceiling/floor effect, and tagging *cis*-eQTL) often replicated. Thus, we emphasize that the replication of interactions, long held as the gold standard of genetic association studies, does not necessarily indicate a true biological effect – even with replication, additional analyses are needed to explicitly address these confounding processes. While not all of the confounding phenomena discussed here may be applicable to all future interaction analyses, their principles generalize. For instance, the idea that subpopulations with specific eQTLs may result in spurious interactions when analyzed together is a broader concept that could apply to case/control studies in addition to ethnicities, and typical corrections for population stratification do not eliminate this issue. Ultimately, we urge caution in the interpretation of interaction studies even when they demonstrate replicating effects — they may not necessarily be driven by the direct interaction of the genetic variants specified in the model. By explicitly accounting for these confounding processes, future studies can bolster support for putative biological interactions and ensure spurious results are not reported within the literature.

It is also critical to select an appropriate statistical model to represent interactions. Most interaction analyses assume an additive main effect of each variant (Turner and Bush 2011; Becker et al. 2012; Fitzpatrick et al. 2015), which is an intuitive choice—eQTL are presumed to behave in an allele-dose manner. However when modeling an interaction between two additively-encoded variants, any deviation of the main effect from additivity by either variant can be partitioned into the interaction term. This leads to a characteristic pattern, wherein the main effects are both in one direction and the interaction term is in the opposite direction. This issue is not inconsequential; it accounted for all significant interactions identified by Turner and Bush (2011). By incorporating dominant main effects into the interaction model, this issue is avoided. We, as well as Hemani et al. (2014) and Brown et al. (2014), have used a complex interaction model containing both additive and dominant main effects for each variant (Cordell 2002). We recommend all studies with sufficient power use interaction models with both additive and dominant main effects to prevent spurious interaction associations.

In this study, we identified genetic interactions that regulate gene expression in humans, which most likely represent true biological epistasis, after systematically accounting for confounding processes capable of producing statistical interactions. We identified these interactions using strict criteria – if an interaction could be accounted for by a confounding process we did not consider it evidence for biological epistasis, even though we could not discern the causal mechanism. It is difficult to fully understand the directionality of the confounding for some of these models—for instance, the single *cis*-eQTL could be tagging the multi-locus genotypes, especially when the *cis*-eQTL accounting for the interaction has a nominal main effect. Orthogonal support from functional genomics data makes it difficult to exclude the possibility of biological epistasis for many of the interactions consistent with multiple explanations. Regardless of their biological interpretation, we have demonstrated that *cis*-regulatory interaction analyses can discover new association models. By performing focused interaction analysis in addition to single-marker association analyses, we can step closer to capturing the complex regulatory architecture of gene expression, and by extension may explain additional disease liability not captured by the analytic methods used by GWAS and sequencing studies.

## Methods

### Genotyping & Gene Expression Data

The discovery dataset was comprised of individuals ascertained as part the International HapMap Project, PhaseI+II (Frazer et al. 2007). The discovery dataset consists of 210 unrelated individuals with genotyping data (Phase I+II, release 24). For each of these individuals, Stranger et al. collected and normalized gene expression levels from immortalized LCLs using the Sentrix Human-6 Expression Bead Chip, v1 (Stranger et al. 2007). We applied a population normalization procedure, described by Veyrieras et al. (2008), to the gene expression values that enabled us to combine all ethnicities in our analysis. Our replication dataset consists of 232 unrelated individuals from the 1000 Genomes Projects, for whom gene expression in LCLs was available. These individuals had been sequenced at low coverage as part of the 1KG project (The 1000 Genomes Project Consortium 2012); we used genetic data from phase I, version 3. Stranger et al. also collected and normalized gene expression levels in LCLs for these individuals using Illumina Sentrix Human-6 Expression BeadChip, v2 (Stranger et al. 2012). We applied the same population normalization procedure (Veyrieras et al. 2008) to these data. Both the discovery and replication dataset are multiethnic; the sample composition by ethnicity is shown in Table 2.

### Generating SNP Pairs for Interaction Testing

To generate SNP-pairs for each gene, we first identified all common SNPS within the gene's *cis*-regulatory region. To be considered common, variants had to have a MAF $\geq$ 5% when all ethnicities were combined. Based on *cis*-eQTL analyses (Veyrieras et al. 2008), the *cis*-regulatory region was defined as starting 500 kb upstream of the gene's start and ending 500 kb downstream of the gene's stop (including the gene itself); gene boundaries were taken from ENSEMBL. Previously, these variants were individually tested for association with the gene's expression level in the discovery dataset by Veyrieras et al. (2008). Based on this analysis, we filtered out SNPs whose marginal effects were not nominally associated with gene expression (excluded $p > 0.05$), under the hypothesis that nominally associated variants may represent weak marginal effects from a true underlying interaction. We then created all possible SNP-pairs amongst the remaining variants. Once this was done for each gene, over 21 million SNP-pairs were generated for interaction-testing.

**Interaction Model**

Each SNP pair was tested for interactions significantly associated with the expression of the gene for which it was generated. The following interaction model (Equation 1) (Cordell 2002) was used:

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2 + PC_1 + PC_2 + PC_3$$

(Equation 1)

where $y$ represents gene expression, $x_1$ and $x_2$ use additive encoding to represent the genotype at SNP A and SNP B respectively, $z_1$ and $z_2$ use Cordell's dominant encoding (2002) to represent the genotype at SNP A and B respectively, $a_1$ and $d_1$ are estimated coefficients representing the additive and dominant effects of SNP A, $a_2$ and $d_2$ are estimated coefficients representing the additive and dominant effects of SNP B, and $i_{aa}, i_{ad}, i_{da}$ and $i_{dd}$ are estimated coefficients representing both additive and dominant interaction effects. The top three principal components were also included as covariates ($PC_{1-3}$). To determine the significance of interactions, this model was compared to a reduced model lacking the four interaction terms using a LRT (Equation 2).

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + PC_1 + PC_2 + PC_3 \quad \text{(Equation 2)}$$

This test was implemented using the program INTERSNP (Herold et al. 2009). We calculated an FDR of 5% using the qvalue package in R (Storey 2015).

**Identification of representative ieQTL models for distinct pairs of interacting genomic loci.**

Some ieQTL models identified in the discovery analysis were redundant due to LD. As the variants within these models are essentially redundant, these models likely represent the same signal. For two ieQTL models to be considered redundant, each SNP within one significant ieQTL model had to be in high LD ($r^2 \geq 0.9$) with a SNP within the second ieQTL model, and vice versa. By using this criterion, the pairs were effectively correlated at $r^2 \geq 0.8$, the threshold typically used for tag-SNP selection. The redundant SNP-pairs have very similar betas for all parameters (Figure 1 – Figure Supplement 2), indicating they represent the same signal from a pair of interacting genomic loci. Redundant ieQTL models were grouped together. The model with the most significant LRT p-value in the discovery analysis was used to represent the entire group in most analyses, so that each pair of interacting genomic loci was equally represented. A visual schematic of this process is provided in Figure 1.

**Investigation of Artifacts**

We only used the representative ieQTL model for each pair of interacting genomic loci (n=1,093) in the all analyses pertaining to the investigation of alternative explanations for statistical interactions. This ensured that each pair of interacting genomic loci was equally represented.

We looked for statistical patterns characteristic of a ceiling/floor effect to determine an upper bound of its prevalence within our results. First, we identified the significant (β±SE could not contain zero) variables in the model. All interactions were then categorized as having 0, 1, or 2 SNPs with a significant main effect - either additive or dominant main effects counted; if both additive and dominant main effects were significant for the same variant, the one with the largest effect size was used to represent the main effect. For interactions were both variants had at least one significant main effect, we determined whether or not they had a concordant direction of effect. For those pairs with concordant directions of effect, we compared the significant interaction term with the largest absolute effect size to determine if it was discordant with the main effects. If this was the case, the interaction had a pattern consistent with a ceiling/floor effect.

We also investigated whether or not ieQTL could be attributable to population stratification artifacts by performing a stratified analysis. We divided the discovery dataset into three groups based on ancestry (CHB+JPT, YRI, CEU). We then tested each interaction in the three ethnicities separately, using the same methodology used in the discovery analysis. If an interaction was nominally significant ($p < 0.05$) in at least one population, we considered it not attributable to population stratification. For interactions which were not significant in any of the populations, we then determined if the interacting variant were population-specific *cis*-eQTL using the following model (Equation 3):

$$y = \mu + a_1 x_1 + d_1 z_1 + PC_1 + PC_2 + PC_3 \qquad \text{(Equation 3)}$$

where $y$ represents gene expression, $x_1$ uses additive encoding to represent the genotype for the variant, $z_1$ uses Cordell's (2002) dominant encoding to represent the genotype, and the top three principal components were included as covariates ($PC_{1\text{-}3}$). Variants with nominally significant ($p < 0.05$) main effects were considered *cis*-eQTL. If a variant was identified as a *cis*-eQTL in only a subset of populations, it was considered population-specific.

**Conditional *cis*-eQTL Analysis**

To determine if interaction-eQTL pairs were tagging a *cis*-eQTL as suggested by Wood et al. (Wood et al. 2014), we first identified all nominal *cis*-eQTL ($p < 0.05$) for genes with significant ieQTL. To identify all nominal *cis*-eQTL, we used a subset of the discovery analysis individuals (n=174) who were also sequenced as part of the 1KG Project (The 1000 Genomes Project Consortium 2012). We used the called genotypes from Phase III, v5. The same gene expression data previously described for the discovery set was used. Within this subset, we performed a single-marker *cis*-eQTL analysis for each common variant (MAF > 5%) within the *cis*-regulatory region using Equation 4:

$$y = \mu + a_1 x_1 + PC_1 + PC_2 + PC_3 \qquad \text{(Equation 4)}$$

where $y$ represents gene expression, $x_1$ uses additive encoding to represent the genotype for the variant, and the top three principal components were included as covariates ($PC_{1\text{-}3}$). Variants with nominal significant ($p < 0.05$) main effects were considered *cis*-eQTL.

To determine if any of these *cis*-eQTL could account for the interaction, we created all pairs of *cis*-eQTL and ieQTL for the same gene. We incorporated each *cis*-eQTL into each interaction model (Equation 5) as shown below.

$$y = \mu + a_1 x_1 + d_1 z_1 + a_2 x_2 + d_2 z_2 + a_3 x_3 + d_3 z_3 + i_{aa} x_1 x_2 + i_{ad} x_1 z_2 + i_{da} z_1 x_2 + i_{dd} z_1 z_2 + PC_1 + PC_2 + PC_3 \qquad \text{(Equation 5)}$$

where $y$ represents gene expression, $x_1$ and $x_2$ use additive encoding to represent the genotype at interacting SNPs A and B respectively, $z_1$ and $z_2$ use Cordell's dominant encoding to represent the genotype at interacting SNPs A and B respectively, $a_1$ and $d_1$ are estimated coefficients representing the additive and dominant effects of SNP A, $a_2$ and $d_2$ are estimated coefficients representing the additive and dominant effects of SNP B, and $i_{aa}, i_{ad}, i_{da}$ and $i_{dd}$ are estimated coefficients representing both additive and dominant interaction effects. The main effect of the *cis*-eQTL is represented with additive encoding by $x_3$ and with dominant encoding by $z_3$; the estimated coefficients corresponding to the main effects are $a_3$ and $d_3$ respectively. The top three principal components were also included as covariates ($PC_{1\text{-}3}$). We then performed a LRT comparing this model to a reduced model lacking the interaction terms (Equation 6).

$$y = \mu + a_1 x_1 + d_1 z_1 + a_2 x_2 + d_2 z_2 + a_3 x_3 + d_3 z_3 + PC_1 + PC_2 + PC_3 \qquad \text{(Equation 6)}$$

If the LRT p-value of an interaction was nominally significant ($p < 0.05$) for all conditional analyses, we considered this evidence that the interaction and *cis*-eQTL represented independent signals.

To determine if ieQTL analyses could identify novel signals, we identified which *cis*-eQTL most accounted for the interaction in the conditional analysis. In other words, we identified the *cis*-eQTL reduced the significance for the interaction the most. We then determined the significance of this *cis*-eQTL using Equation 4.

**Functional Genomics Analysis**

Functional annotations were downloaded from the ENCODE website (http://genome.ucsc.edu/ENCODE/downloads.html). We downloaded all DNase-seq peaks (FDR = 0.01), FAIRE peaks (FDR = 0.01), histone peaks, transcription factor binding site peaks (called with PeakSeq), and combined genome segmentations that were specific to LCLs (i.e., collected within GM12878) (Kellis et al. 2014). In total, 83 distinct functional annotations were downloaded (Dataset S3). We characterized enrichment of ieQTL SNPs within functional regions by first classifying every SNP-pair tested as having either significant interactions or not. Then, we classified each SNP within the pair as overlapping a region of DNA with the functional annotation or not using BEDTools (Quinlan 2002). This generated a 2x2 contingency table (axes corresponding to significance of the interaction and presence within the annotation), which we used to conduct a odds ratio test to determine if there was a significant difference in the proportion of ieQTL SNPs within the functional region versus non-ieQTL within the functional region. We used a Bonferroni multiple testing correction (n=83) to determine significance.

**Acknowledgements**

We thank Laura Wiley for normalizing gene expression values within the replication dataset. We also thank Jacob Hall, Corinne Simonti, and R. Michael Sivley for their help and advice on this project.

**References**


Becker, Jessica, Jens R Wendland, Britta Haenisch, Markus M Nöthen, and Johannes Schumacher. 2012. "A Systematic eQTL Study of Cis-Trans Epistasis in 210 HapMap Individuals." *European Journal of Human Genetics EJHG* 20 (1). Macmillan Publishers Limited: 97–101. doi:10.1038/ejhg.2011.156.

Brem, Rachel B, John D Storey, Jacqueline Whittle, and Leonid Kruglyak. 2005. "Genetic Interactions between Polymorphisms That Affect Gene Expression in Yeast." *Nature* 436 (7051): 701–3. doi:10.1038/nature03865.

Brown, Andrew A, Alfonso Buil, Ana Viñuela, Tuuli Lappalainen, Hou-feng Zheng, John B Richards, Kerrin S Small, Timothy D Spector, Emmanouil T Dermitzakis, and Richard Durbin. 2014. "Genetic Interactions Affecting Human Gene Expression Identified by Variance Association Mapping." *eLIFE*.

Capellari, Sabina, Rosaria Strammiello, Daniela Saverioni, Hans Kretzschmar, and Piero Parchi. 2011. "Genetic Creutzfeldt-Jakob Disease and Fatal Familial Insomnia: Insights into Phenotypic Variability and Disease Pathogenesis." *Acta Neuropathologica* 121 (1): 21–37. doi:10.1007/s00401-010-0760-4.

Cordell, Heather J. 2002. "Epistasis : What It Means , What It Doesn ' T Mean , and Statistical Methods to Detect It in Humans" 11 (20): 2463–68.

Fitzpatrick, Darren J, Colm J Ryan, Naisha Shah, Derek Greene, Cliona Molony, and Denis C Shields. 2015. "Genome-Wide Epistatic Expression Quantitative Trait Loci Discovery in Four Human Tissues Reveals the Importance of Local Chromosomal Interactions Governing Gene Expression." *BMC Genomics* 16 (1): 1–13. doi:10.1186/s12864-015-1300-3.

Frazer, Kelly a, Dennis G Ballinger, David R Cox, David a Hinds, Laura L Stuve, Richard a Gibbs, John W Belmont, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164): 851–61. doi:10.1038/nature06258.

Gibson, Greg, Rebecca Riley-Berger, Larry Harshman, Artyom Kopp, Scott Vacha, Sergey Nuzhdin, and Marta Wayne. 2004. "Extensive Sex-Specific Nonadditivity of Gene Expression in Drosophila Melanogaster." *Genetics* 167 (4): 1791–99. doi:10.1534/genetics.104.026583.

Grundberg, Elin, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, et al. 2012. "Mapping Cis- and Trans-Regulatory Effects across Multiple Tissues in Twins." *Nature Genetics* 44 (10): 1084–89. doi:10.1038/ng.2394.

Hemani, Gibran, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, et al. 2014. "Detection and Replication of Epistasis Influencing Transcription in Humans." *Nature* 508 (7495): 249–53. doi:10.1038/nature13005.

Herold, Christine, Michael Steffens, Felix F. Brockschmidt, Max P. Baur, and Tim Becker. 2009. "INTERSNP: Genome-Wide Interaction Analysis Guided by a Priori Information." *Bioinformatics* 25 (24): 3275–81. doi:10.1093/bioinformatics/btp596.

Hindorff, Lucia a, Praveen Sethupathy, Heather a Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri a Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits." *Proceedings of the National Academy of Sciences of the United States of America* 106 (23): 9362–67. doi:10.1073/pnas.0903103106.

Kellis, Manolis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, et al. 2014. "Defining Functional DNA Elements in the Human Genome." *Proceedings of the*

*National Academy of Sciences of the United States of America* 111 (17): 6131–38. doi:10.1073/pnas.1318948111.

Lappalainen, Tuuli, Stephen B Montgomery, Alexandra C Nica, and Emmanouil T Dermitzakis. 2011. "Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease." *American Journal of Human Genetics* 89 (3). The American Society of Human Genetics: 459–63. doi:10.1016/j.ajhg.2011.08.004.

Musunuru, Kiran, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, et al. 2010. "From Noncoding Variant to Phenotype via SORT1 at the 1p13 Cholesterol Locus." *Nature* 466 (7307). Nature Publishing Group: 714–19. doi:10.1038/nature09266.

Price, Alkes L, Agnar Helgason, Gudmar Thorleifsson, Steven a McCarroll, Augustine Kong, and Kari Stefansson. 2011. "Single-Tissue and Cross-Tissue Heritability of Gene Expression via Identity-by-Descent in Related or Unrelated Individuals." *PLoS Genetics* 7 (2): e1001317. doi:10.1371/journal.pgen.1001317.

Quinlan, Aaron. 2002. *BEDTools: The Swiss-Army Tool for Genome Feature Analysis*. *Current Protocols in Bioinformatics*. doi:10.1002/0471250953.

Schaub, Marc a, Alan P Boyle, Anshul Kundaje, and Kelly a Frazer. 2012. "Linking Disease Associations with Regulatory Information in the Human Genome Toward Mapping the Biology of the Genome," 1748–59. doi:10.1101/gr.136127.111.

Smemo, Scott, Juan J Tena, Kyoung-Han Kim, Eric R Gamazon, Noboru J Sakabe, Carlos Gómez-Marín, Ivy Aneas, et al. 2014. "Obesity-Associated Variants within FTO Form Long-Range Functional Connections with IRX3." *Nature* 507 (7492): 371–75. doi:10.1038/nature13138.

Storey, John D, and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences of the United States of America* 100 (16): 9440–45. doi:10.1073/pnas.1530509100.

Storey, John D. 2015. "qvalue: Q-value estimation for false discovery rate control." R package version 2.0.0,http://qvalue.princeton.edu/, http://github.com/jdstorey/qvalue.

Stranger, Barbara E, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, et al. 2012. "Patterns of Cis Regulatory Variation in Diverse Human Populations." *PLoS Genetics* 8 (4): e1002639. doi:10.1371/journal.pgen.1002639.

Stranger, Barbara E, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, et al. 2007. "Population Genomics of Human Gene Expression." *Nature Genetics* 39 (10): 1217–24. doi:10.1038/ng2142.

The 1000 Genomes Project Consortium. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 135 (V): 0–9. doi:10.1038/nature11632.

Turner, Stephen D, and William S Bush. 2011. "Multivariate Analysis of Regulatory Snps: Empowering Personal Genomics by Considering Cis-Epistasis and Heterogeneity." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 276–87. doi:9789814335058_0029 [pii].

Veyrieras, Jean-Baptiste, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K Pritchard. 2008. "High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation." *PLoS Genetics* 4 (10): e1000214. doi:10.1371/journal.pgen.1000214.

Wei, Wen-Hua, Gibran Hemani, and Chris S. Haley. 2014. “Detecting Epistasis in Human Complex Traits.” *Nature Reviews Genetics* 15 (11). Nature Publishing Group: 722–33. doi:10.1038/nrg3747.

Wood, Andrew R, Marcus a Tuke, Michael Nalls, Dena Hernandez, Andrew Singleton, David Melzer, Luigi Ferrucci, Timothy M Frayling, and Michael N Weedon. 2014. “An Alternative Explanation for Apparent Epistasis.” *Nature* 514 (Ld). Nature Publishing Group: 1–7. doi:10.1038/nature13691.

**Figures**

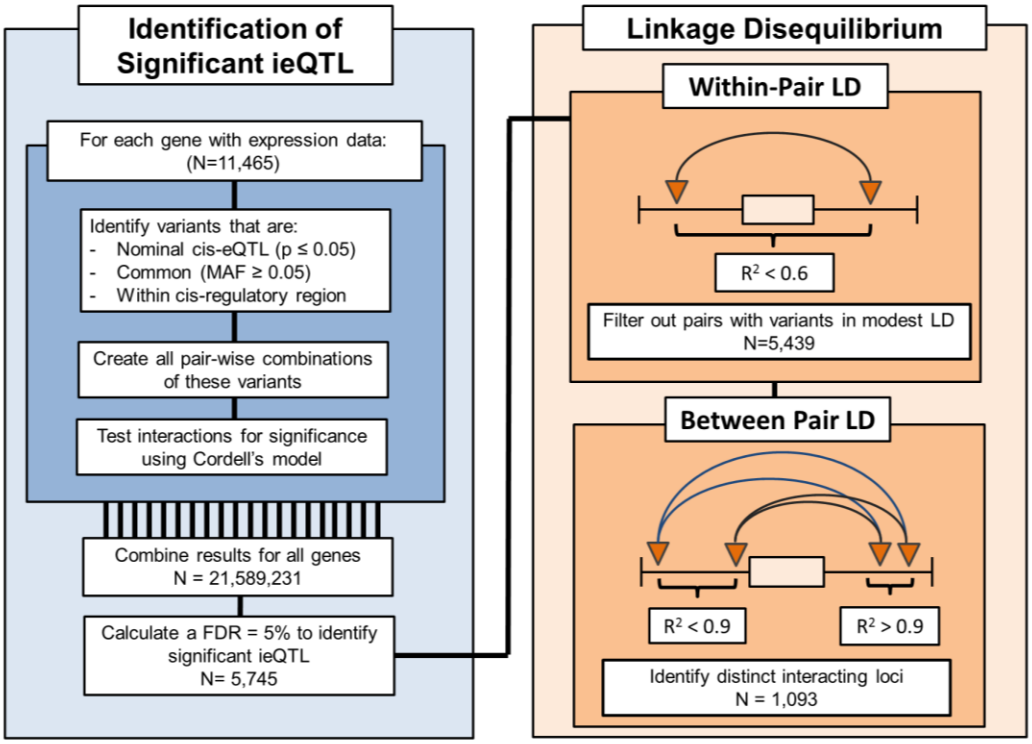


**Figure 1. Workflow used to identify ieQTL and group ieQTL together.** In the discovery analysis, nominally significant *cis*-eQTL (denoted by triangles) were paired together and tested for interactions significantly associated with gene expression levels (denoted by arcs). The within-pair LD was then calculated (Figure 1 – Figure Supplements 1), and interactions composed of variants in modest LD ($r^2 > 0.6$) with one another were removed from the remainder of the analysis. Some of the remaining interactions represented the same pair of interacting genomic loci (Figure 1 – Figure Supplements 2), and were grouped into distinct groups (denoted by the arc color). For two ieQTL models to be grouped together, each SNP within one significant ieQTL model had to be in high LD ($r^2 \geq 0.9$) with a SNP within the second ieQTL model, and vice versa.

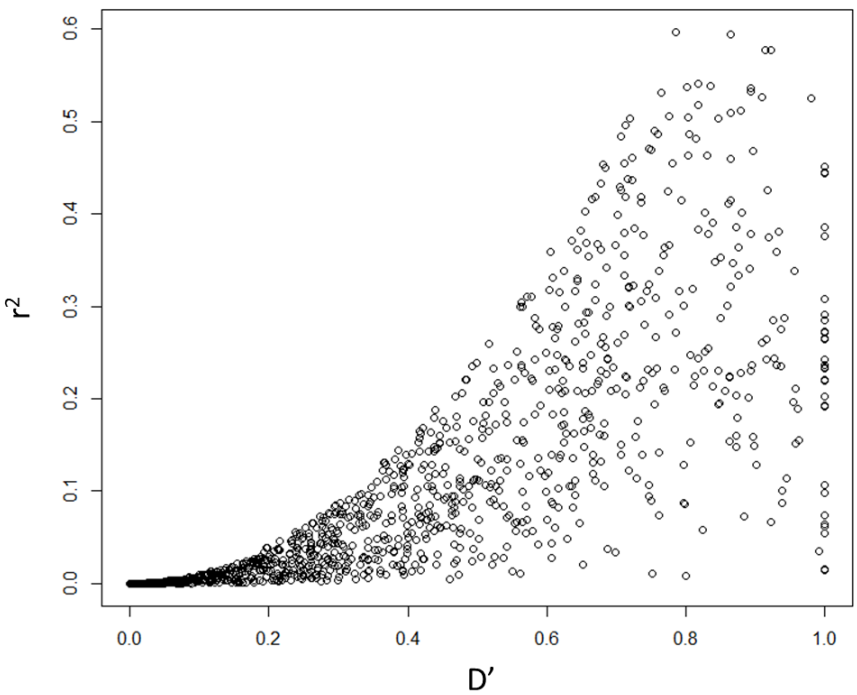


**Figure 1, Figure Supplements 1. Linkage disequilibrium between interacting variants.** We calculated LD between interacting variants using both $r^2$ and D' to determine if they may be on the same haplotype. Interactions between variants in modest LD ($r^2 > 0.6$) had been removed from all stages of the analysis, and hence are not shown here.

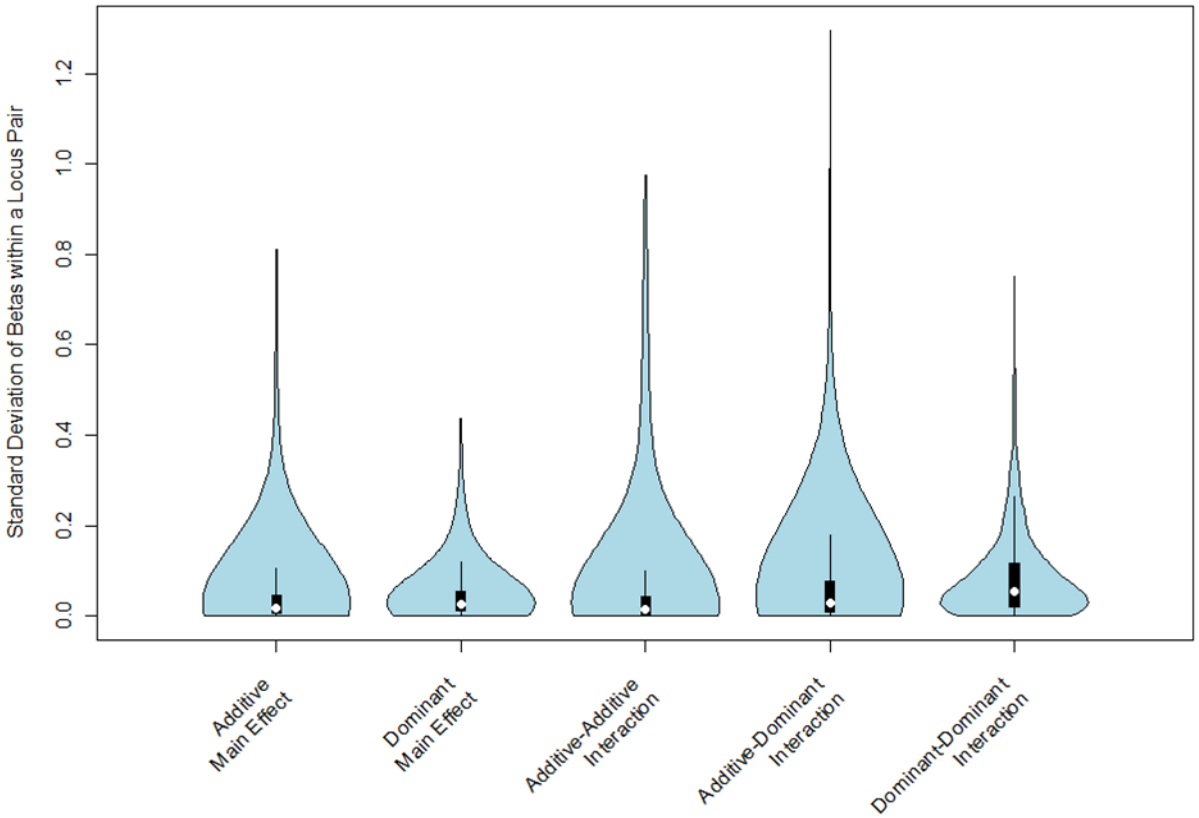


**Figure 1 – Figure Supplement 2. Redundant SNP-pairs have very similar parameter estimates.** We grouped together all pairs of interacting SNPs (n=5,439) identified as being redundant through LD measures. For each group, we identified all terms that were significant in at least one of the associated interactions ($p < 0.05$). We extracted the betas for these significant terms from all interactions within the group. We then calculated the standard deviation of the betas for each significant term within each group to determine how similar the parameter estimates were across all interactions in the same group. The distribution of these standard deviations, categorized by type of variable, is shown above.

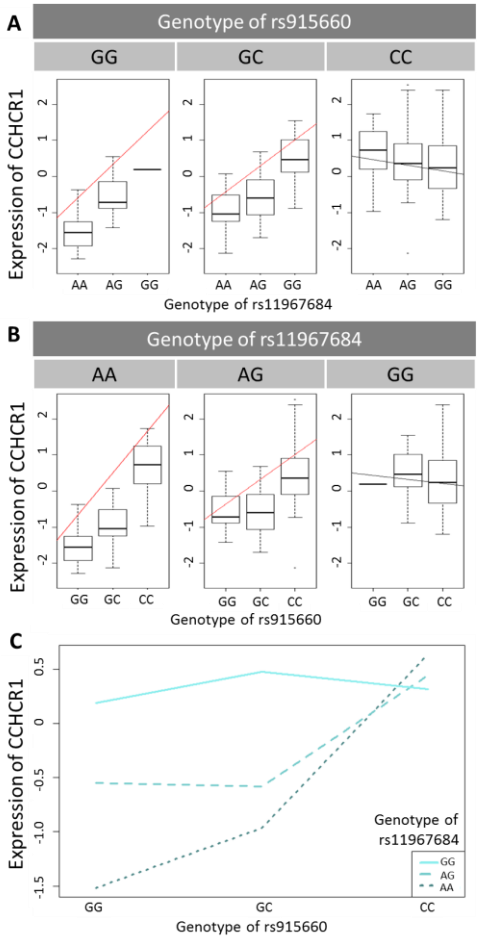


**Figure 2. The interaction between rs915660 and rs11967684 associated with the expression of *CCHCR1* may be a ceiling effect.** The ceiling effect, caused by limitations in the detectable range of gene expression, has a hallmark pattern – both variants have main effects with concordant direction of effect, and the interaction term has a discordant direction. Here, we illustrate that major allele of rs11967684 (G) increases the expression of *CCHCR1* (A). The overlaid regression line represents the association between the additive effect of rs11967684 on each background of rs915660 and the expression of *CCHCR1* (red signifies significance, $p < 0.05$). The major allele of rs915660 (C) also increases the expression of *CCHCR1* (B), which meets the first criteria for a ceiling effect. However, when the major allele of one variant, which should increase expression, co-occurs with two major alleles of the second variant, it no longer significantly increases gene expression (A,B). The interaction term captures this, and consequently has the opposite direction of effect, which fulfills the requirements for a ceiling effect. The interaction plot (C), which depicts the mean gene expression for all individuals with the specified genotype combination, shows the ceiling of gene expression for *CCHCR1* is ~0.5 standard deviation increase in gene expression.

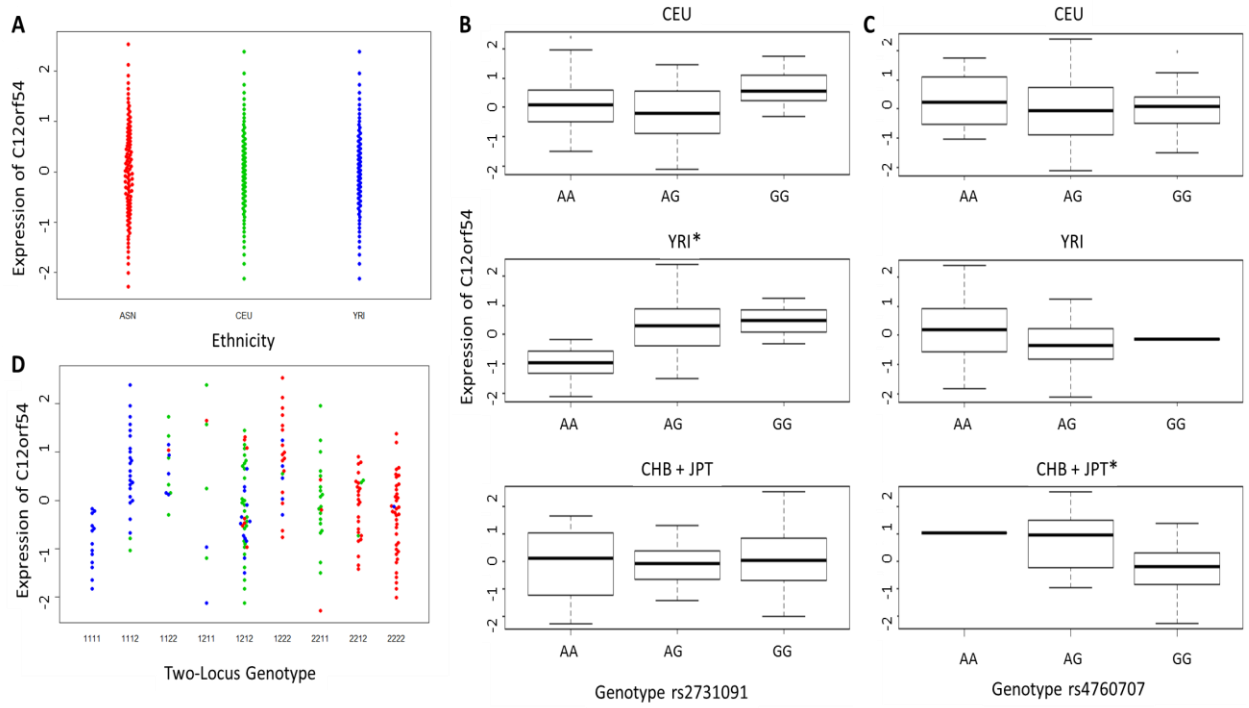


**Figure 3. Population specific eQTLs may underlie ieQTL regulating *C12orf54*.** An interaction between rs2731091 and rs4760707 regulating *C12orf54* was identified, replicated, and was inconsistent with the ceiling/floor effect; however it was not nominally significant ($p < 0.05$) in any population in the stratified analysis. There are not systematic differences in the expression of *C12orf54* between populations (A); however we found that each variant was a population-specific *cis*-eQTL (B,C). rs2731091 significantly regulated gene expression as a *cis*-eQTL in YRI($p = 7.28 \times 10^{-6}$), but not CEU ($p = 0.14$) or CHB+JPT ($p=0.84$). rs4760707 was a *cis*-eQTL in CHB+JPT ($p=7.25 \times 10^{-6}$), but not in YRI ($p=0.17$) or CEU ($p=0.96$). There are clear population differences in the frequency of two-locus genotypes between populations (D); in combination, it appears the population differences in two-locus genotypes and population specific *cis*-eQTL produced a nuanced form of population stratification.

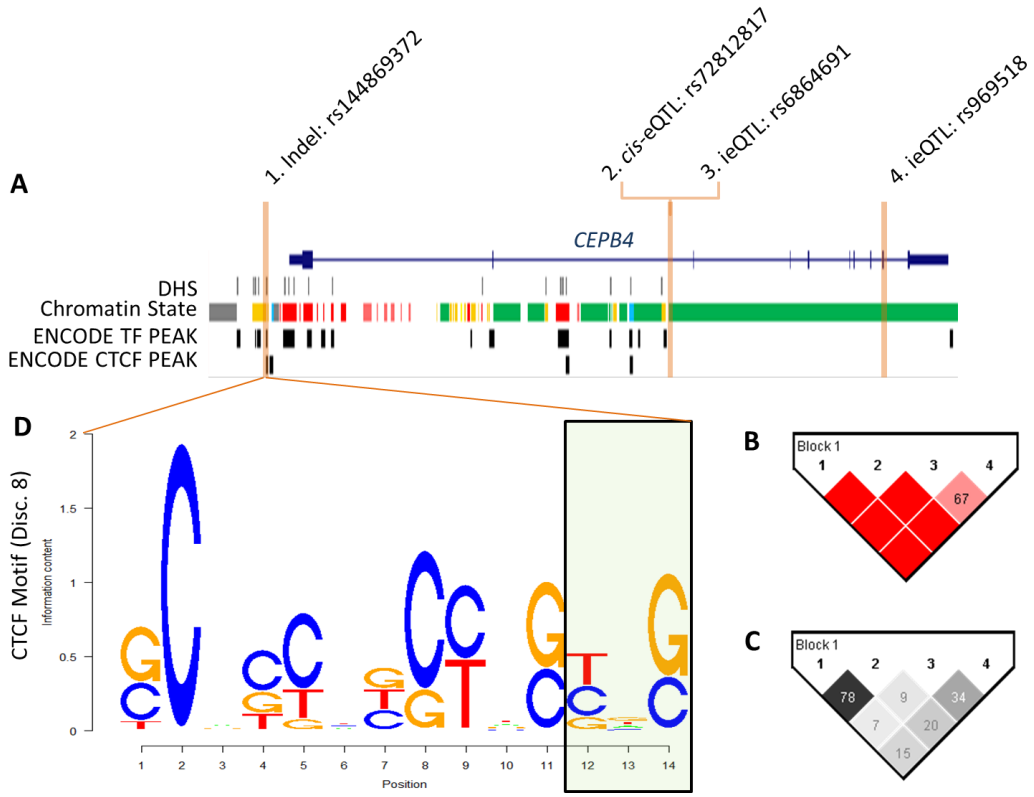


**Figure 4. Interactions impacting the expression of *CPEB4* may represent haplotype effects.** A significant interaction between rs6864691 and rs969518 regulating the expression of *CPEB4* was identified that replicated and was inconsistent with artifacts. The *cis*-eQTL rs72812817 mediated this interaction in the conditional analysis; however none of these variants were within putative regulatory elements in GM12878 assayed by the ENCODE Project (A). However, an indel, rs144869372, always occurred on the background of the *cis*-eQTL (D' = 1, B). In fact, the indel and *cis*-eQTL formed a haplotype with the interacting variants (B) based on D', despite modest r2 values as shown in the heatmap (C). The structural variant occurs within both a ChromHMM strong enhancer (yellow) and a CTCF binding peak in GM12878. Notably, the structural variant is predicted to alter the binding of CTCF (D) by HaploReg, by altering the last three nucleotides in the binding motif. Given the functional genomics evidence, the indel may be the causal variant, which is detected through interactions that tag the haplotype the indel is carried on.

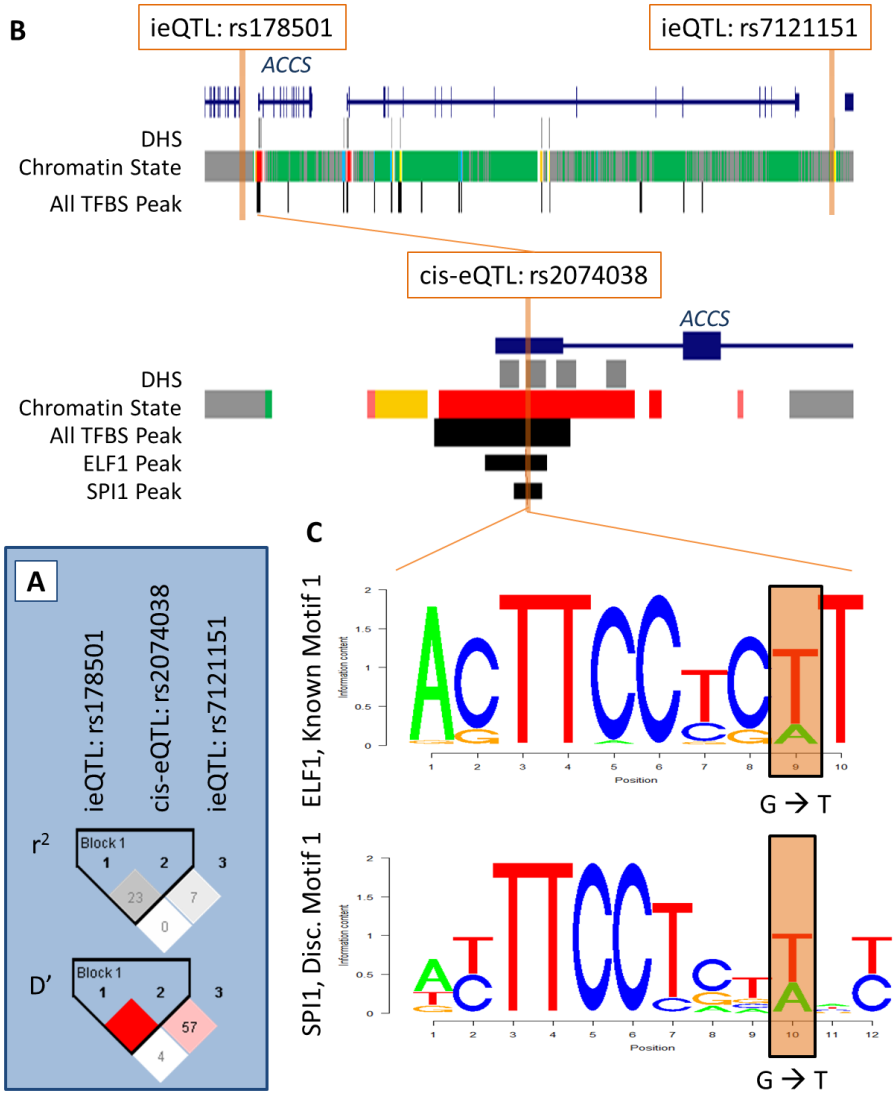


**Figure 5. The interacting SNPs regulating *ACCS* are likely tagging a single-variant *cis*-eQTL through linkage disequilibrium.** The interaction between rs178501 and rs7121151 is mediated by the *cis*-eQTL rs2074038 in the conditional analysis (interaction p-value > 0.05). While the interacting variants are in low LD with the *cis*-eQTL based on $r^2$, their high D' indicates they often occur on the same haplotype (A). The interacting variants are not located within DHS, predicted chromatin states with a regulatory function, or any of the uniform binding peaks identified for all transcription factors tested in GM12878 by ENCODE (B); however, the *cis*-eQTL is located within the canonical promoter for *ACCS*, a DHS, and numerous transcription factor binding peaks identified in GM12878 by ENCODE (B). Notably, the *cis*-eQTL occurs within a binding peak for both ELF1 and SPI1 in GM12878 (B), and also alters the binding motifs of these transcription factors at the position highlighted in orange (C). Thus, the *cis*-eQTL rs2074038 is likely the causal variant, and the interaction is simply capturing its effect through LD.

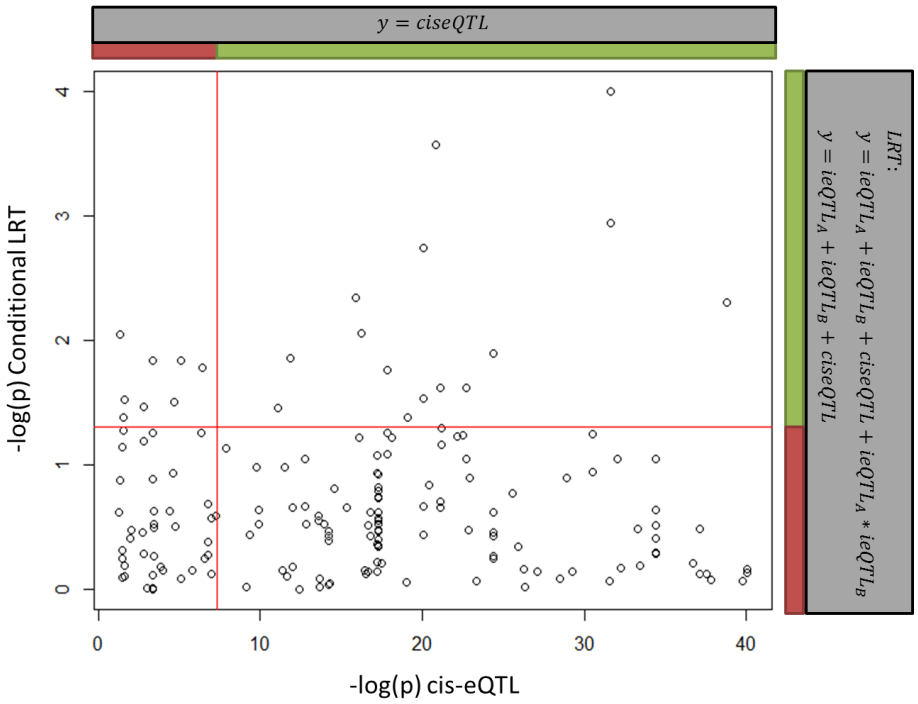


**Figure 6. Interaction analyses identify novel associations.** In the conditional analysis, we identified the *cis*-eQTL whose inclusion most accounted for the interaction (i.e., most reduced the significance of the interaction term in the LRT). The significance of the interaction term when this variant is conditioned on, represented by the formula on the right, is plotted along the Y axis. Interactions above the horizontal line remained at least nominally significant ($p > 0.05$) when the *cis*-eQTL was taken into account. To determine if the effect of the *cis*-eQTL could have been identified in a single marker analysis, we determined its significance in a single marker association test (represented by formula at the top), which is plotted along the X-axis. *Cis*-eQTL to the left of the vertical line would not have been identified using the standard GWAS significance threshold ($p > 5x10^{-8}$). Thus, the graph can be divided into four quadrants, representing the significance of the *cis*-eQTL and the significance of the ieQTL when conditioned on the *cis*-eQTL.

**Tables**

**Table 1 | Interactions that replicate and are not accounted for by other explanations.**

| Gene | SNP 1 | SNP 2 | Discovery‡ | Replication‡ |
|---|---|---|---|---|
| *EFCAB2* | rs2050632 | rs4072545 | 12.93 | 6.23 |
| *EFCAB2* | rs586013 | rs4072545 | 11.43 | 6.72 |
| *EFCAB2* | rs675033 | rs4658693 | 9.75 | 2.17 |
| *MYRFL* | rs1240251 | rs1240276 | 8.98 | 2.43 |
| *HMBOX1* | rs2221894 | rs2279446 | 7.98 | 2.23 |
| *HMBOX1* | rs2137471 | rs17059658 | 7.90 | 4.87 |
| *HMBOX1* | rs12674777 | rs17059658 | 7.75 | 2.31 |
| *MYRFL* | rs1262808 | rs11615099 | 7.41 | 2.69 |
| *EFCAB2* | rs651773 | rs4658693 | 6.95 | 2.29 |
| *EFCAB2* | rs1538165 | rs6664895 | 6.95 | 2.77 |
| *HMBOX1* | rs6558085 | rs2279446 | 6.10 | 2.58 |
| *MYOM2* | rs12681998 | rs17064818 | 5.64 | 1.55 |
| *TOMM34* | rs4812873 | rs6103899 | 5.39 | 1.75 |
| *NQO2* | rs6920099 | rs9378763 | 5.09 | 1.74 |
| *APIP* | rs1549791 | rs7115749 | 5.00 | 2.54 |

‡ $-\log_{10}$ *P* values for 4 d.f. LRT interaction test.

| Analysis | Total Sample Size | Ethnicity | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CHB | CEU | GIH | JPT | LWK | MXL | MKK | YRI |
| Discovery | 210 | 45 | 60 | - | 45 | - | - | - | 60 |
| Replication | 232 | 34 | - | - | 35 | 80 | 38 | - | 45 |

**Table 2. Dataset Composition by Ethnicity.** This provides the breakdown of ethnicities comprising each stage of the analysis.

**Supplemental Information**

**Supplemental Dataset 1. Significant interactions identified in the discovery analysis.** This file provides all 5,439 interactions identified in the discovery analysis. When these interactions appeared to represent the same signal, due to LD, they were placed into groups (n = 1,093) and a representative interaction was chosen. We provide the group identifier for each of the interactions, and the group's representative interaction.

**Supplemental Dataset 2. Alternative explanations for significant interactions identified in the discovery analysis.** We examined whether or not the 1,093 interactions could be explained by confounding processes. Here, we present which alternative explanations could account for each interaction.

**Supplemental Dataset 3. Enrichment of interacting SNPs within regulatory regions identified by ENCODE.** This file provides the enrichment (as measured by an odds ratio test) of interacting variants within functional annotation tracks (n=83) in LCLs from ENCODE as compared to all nominal cis-eQTL evaluated for interactions in this analysis. All 5,439 interactions were used in this analysis.