A gene-level test for directional selection on gene expression

Laura L. Colbran[1,*] and Iain Mathieson[1,*]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
[*]Correspondence to LLC (laura.colbran@pennmedicine.upenn.edu) or IM (mathi@pennmedicine.upenn.edu)

Abstract

Human phenotypes and evolutionary fitness are influenced by both coding and non-coding genetic variants. However, most variants associated in genome-wide association studies and. scans for selection are non-coding. Interpretation of these variants' effects and understanding of the way in which they contribute to phenotypic variation and adaptation is therefore limited by our understanding of gene regulation and by the difficulty in confidently linking non-coding variants to genes. To overcome this, we developed a gene-by-gene test for population-specific selection based on combinations of regulatory variants.

Specifically, we extended the Q_X test for polygenic selection to gene expression models trained using joint-tissue imputation, a transcriptome-wide association method trained on paired genotype and RNA-seq data. We used effect sizes and variants from those models to calculate Q_X statistics for 17,388 protein-coding genes based on allele frequencies in the twenty six 1000 Genomes populations. We identified 45 genes with significant evidence (FDR < 0.1) for selection, including *FADS1*, *KHK*, *SULT1A2*, *ITGAM*, and genes in the HLA region. We further confirm that significant selection signals do correspond to plausible population-level differences in predicted expression.

Our gene-level Q_X score is independent of other methods for detecting selection based on genomic variation, may therefore be useful when used in combination with more traditional selection tests to specifically identify selection on regulatory variation. However, we find that very few (0.2%) genes have strong evidence for directional, population-specific selection on the component of their expression that is predicted by *cis*-regulatory variants. While this is consistent with most *cis*-regulatory variation evolving under genetic drift or stabilizing selection, it is also possible that any effects are smaller than we can detect, or that population-specific selection is driven by tissue-specific or *trans* effects. Overall, our results demonstrate the utility of one approach to combining population-level information with functional data to understand the evolution of gene expression.

Keywords

human evolution; gene regulation; selection

Introduction

Natural selection is one process by which populations respond to their environment. Therefore, identifying phenotypes, genes and variants influenced by selection is an important aspect of understanding how organisms and populations adapt. In humans, there has been an increasing interest in identifying population-specific adaptations. This is usually in the hopes of better understanding the mechanisms underlying the phenotype (Crawford et al., 2017; Ilardo et al., 2018; Simonson et al., 2010) and using that information to improve human health. However, linking genomic signals of selection to specific phenotypes and evolutionary pressures remains challenging. It is believed that changes in gene expression

underlie most recent evolution (Corradin et al., 2016; King and Wilson, 1975), and are therefore the most likely changes to underlie selection on complex traits. On the other hand, across species gene expression seems to largely be under stabilizing selection or evolving neutrally (Chen et al., 2018; Rohlfs et al., 2014; Signor and Nuzhdin, 2018). While there are some genes with evidence for directional selection on their expression (Blekhman et al., 2008), these effects are tissue-specific, and the extent to which selection plays a role on genome-wide gene regulation remains poorly characterized (Price et al., 2022).

One approach to testing for selection on complex traits is to start with an observed trait difference, then to test whether that difference is greater than expected compared to genetic difference (Whitlock, 2008), and work backwards to understand the mechanism. The limitation of this approach is that it is difficult to account for the environmental component of the phenotypic variance. Another approach, typical for genome-wide scans for selection, is to identify individual outlier haplotypes based on allele frequency or linkage disequilibrium (LD) patterns, then working forward to understand which variant is the causal one, and what it might be influencing (Voight et al., 2006). Since variants rarely act in isolation, many traits are polygenic and any signals of selection on complex traits could therefore be spread across many variants across the genome. This can be identified by looking for enrichment of these locus-specific selection signals (Field et al., 2016). Somewhat intermediate to these approaches, the Q_X statistic (Berg and Coop, 2014) tests for polygenic selection by using genome-wide association results to test for systematically divergent allele frequencies among all independent variants associated with a phenotype, in theory capturing only genetic contributions to the phenotypic variance. However, in practice even this approach can be confounded due to poorly-controlled population stratification in the underlying GWAS (Berg et al., 2019). It is therefore unclear to what extent polygenic selection is relevant for human adaptation.

Most variants associated with complex traits through genome-wide association studies (and therefore those most likely to be subject to selection) are non-coding and likely operate through changes in gene expression. Therefore directional polygenic selection on complex traits, if it exists, may involve directional selection on gene expression. We aim to test for selection on the expression of specific genes, reasoning that this phenotype might reflect the effects of natural selection more clearly than other complex traits. However, gene regulation is extremely complicated, and mapping particular variants to their effects and genes is challenging Benton et al. (2019); Gasperini et al. (2020). In parallel to the development of GWAS methodology, there has also been a proliferation of methods and data to associate variants with gene expression. Single-variant eQTL studies are common, however genes often have multiple eQTL acting in concert to modulate expression. In addition, it is often prohibitively expensive to obtain the RNA-seq data needed to study expression directly in very large samples. Joint-tissue Imputation (JTI) is a machine-learning method that was developed to fill that gap by using expression and functional genomics data across dozens of tissues to predict gene expression based on combinations of genetic variants (Zhou et al., 2020a). These models and similar ones can be used in a transcriptome-wide association study (TWAS) to identify gene-level associations with complex traits (Petty et al., 2019; Zhu and Zhou, 2020), and we have used them to study predicted differences between ancient populations (Colbran et al., 2021). However whether predicted differences between populations in fact reflect real differences in expression and if so whether they are the result of directional selection are still open questions.

The goal of our study is to use the Q_X test with eQTL instead of GWAS data to test for population-specific directional selection on combinations of variants that are associated with expression of specific genes. Done genome-wide, this results in a gene-level test for selection on regulatory variation, which we believe will be more specific than polygenic selection scans on higher-level traits and more interpretable than variant-level scans. Overall, this work identifies dozens of genes whose regulation has been influenced by population-specific selection, and demonstrates the utility of incorporating mechanistic data into genome-wide scans for selection.

Results

Testing for Selection on Regulatory Variants

The first step in characterizing selection on gene regulation is to identify the relevant genetic variants to include. To do this, we used published statistical models of gene regulation built using Joint Tissue Imputation (JTI) (Zhou et al., 2020a). These models were trained using the genotypes and transcriptomes from 49 tissues from the GTEx project (Aguet et al., 2017), including variant selection and quantification of independent linear effects of combinations of variants on expression of each gene in each tissue. Because these models borrow information across tissues, they are often correlated with each other, particularly for genes with shared regulatory patterns. We wanted to use one model per gene, and to focus on models that capture accurate information about individual-level gene expression. Therefore, we compared predicted expression to observed expression in Lymphoblastoid Cell Lines (LCLs) for 5 populations from 1000 Genomes (1kG; N = 447) (Lappalainen et al., 2013). We found that the rank correlation between predicted and actual expression for the gene models trained in GTEx LCLs was highly variable by gene (Fig. 1A; median Spearman $\rho = 0.12$, maximum $\rho = 0.93$). It was also significantly correlated with model performance during training (Fig. 1B; ρ with model $R^2 = 0.58$, $P = 2.1 \times 10^{-686}$), indicating that the training R^2 is a useful proxy for out-of-sample performance in tissues we have not measured directly. In order to maximize power to detect selection without running large numbers of highly correlated tests, we decided to use the model for each gene with the highest training R^2, regardless of the primary tissue it was trained in (the "best models"). Unsurprisingly , these models showed lower agreement between observed expression in LCLs and predicted expression (Supp. Fig. 2; median $\rho = 0.026$, maximum $\rho = 0.93$). However, this agreement remained significantly correlated with the training R^2 of the model ($\rho = 0.22$, $P = 2.1 \times 10^{-190}$). Unless otherwise specified, we focused on these best models for our genome-wide scan for selection, but suggest focusing on relevant tissues when testing specific hypotheses.

To test for selection on gene regulation, we adapted the Q_X test for polygenic selection (Berg and Coop, 2014) (Fig. 1C). We calculated a per-gene Q_X score for 2,504 individuals from 26 populations from the high coverage hg38 1kG data (Byrska-Bishop et al., 2022; The 1000 Genomes Project Consortium, 2015) by using the effect sizes from the corresponding JTI model in place of the GWAS associations, and replicated our analyses in 929 individuals from 7 populations from the Human Genome Diversity Panel (Bergström et al., 2020). This allows us to test for overdispersion in genetic values of gene regulation, or coordinated differences in allele frequencies of regulatory variants (Methods). Theoretically the Q_X statistic follows a χ^2 distribution with degrees of freedom one fewer than the number of populations under consideration (Supp. Fig. 1). In practice however, it can be over- or under-dispersed for reasons other than selection, such as population stratification or stabilizing selection.

One way to control for this is to calculate an empirical distribution. In Berg and Coop (2014) this was done by resampling allele frequencies for the variants in question genome-wide (Supp. Fig. 3A). In our case, doing a similar process results in extremely badly calibrated P-values (Supp. Fig. 4); this is primarily because our gene regulation models break the assumption of independence between variants (discussed in more detail in Methods). While the effect sizes fit by the models are independent and variants were pruned for very high LD ($r^2 > 0.8$), it is reasonably common for a models to have varying numbers of variants in lower LD ($r^2 \approx 0.5$) with each other. We were unable to devise a test that would fairly randomize allele frequencies while maintaining the LD structure of each gene.

We therefore instead calculated P-values by fitting a gamma distribution to the Q_X distribution, which results in a highly similar ordering of genes as the frequency permutation approach (Spearman $\rho = 0.99$) but better calibrated P-values. We identified 45 genes under significant selection (FDR $<$ 0.1; Fig. 2A-B). Because predicted expression of nearby genes shared regulatory haplotypes, these corre-

sponded to 20 visible peaks. These included several loci known to have experienced population-specific selection (e.g. *FADS1* and the TLR and HLA loci; Mathieson et al., 2015). However, these P-values are potentially still inflated by uncorrected population stratification, and are additionally correlated with both the number of variants in a model as well as its training R^2 (Supp. Fig. 5). These are technical aspects of the model training we do not necessarily expect to influence patterns of selection, but probably influence power. These characteristics make it difficult to identify which gene in a peak is most likely to be the one directly under selective pressure, rather than merely influenced by the resulting allele frequency shifts.

We therefore also calculated a P-value for each gene by resampling the effect sizes of the variants in each model from the distribution of effects across all models (Supp. Fig. 3B). This allows us to control for the stratification-induced allele frequency correlations within a model, and for technical characteristics of the model (Supp. Fig. 5), while testing specifically for coordination in the effect sizes of the variants with large allele frequency differences that would lead to a directional shift in the regulation of that gene. While robust, this test is very conservative, and does not identify genes such as *FADS1* for which a single haplotype is under selection. Essentially, the Q_X statistic uses information about both allele frequency differences and effect sizes and direction but when computing P-values using this permutation approach we condition on the allele frequency difference and use only information about effect sizes and direction relative to those differences. Although 23 genes have significant (FDR < 0.1) P-values, overall the QQplot shows only a mild inflation over the null and little evidence of strong outliers (Fig. 2C-D). While there was no overlap in the significant genes in each P-value scheme, gene ordering was relatively highly correlated (Spearman $\rho = 0.71$).

We also ran the scan in HGDP. For the gamma-corrected P-values, 48 genes were significant at FDR <0.1, and 4 peaks (HLA, *KHK*, *KAT8*, and *ACO2*) overlapped genes identified in 1kG (Supp. Fig. 6). For the more conservative effect size permuted P-values, no genes passed that significance threshold, though *SAMD10* (also identified in 1kG) did have the smallest P-value. Overall these results suggest that directional selection on expression of individual genes is relatively rare, or at least is not driven by the *cis*-regulatory variation captured by eQTL studies.

Q_X is independent of other selection metrics

While our Q_X analysis did identify several previously-known genes, we wanted to know whether including gene regulation information was generally giving us more information than other, less-specific tests for selection. We therefore calculated the correlation between gene-level Q_X and a variety of other measures of selection (Fig. 3A). We find that the gamma-corrected P-values are not strongly correlated with either Loss-of-Function intolerance or PhyloP, both of which are metrics of evolutionary constraint (Lek et al., 2016; Pollard et al., 2010), or with two haplotype-based tests for more recent selection (iHS, nSL, averaged across a window for each gene; highest Spearman $\rho = -0.032$ between P and PhyloP), although as expected these other four metric do show some correlations with each other (highest Spearman $\rho = 0.38$ between iHS and nSL). The same is true of the effect size permuted P-values.

We also tested whether different classes of genes were enriched for signals of directional regulatory selection. LoF-intolerant genes are somewhat depleted among the genes with the smallest gamma-corrected P-values (e.g. OR $= 0.281$, $P = 0.0011$ for the top 100 genes; Supp. Fig. 7), suggesting that genes under strong constraint on their protein sequence also tend to be more constrained on their regulatory variation. Surprisingly, housekeeping genes, a broadly-expressed class of genes responsible for basic cellular functions that we might expect to be similarly constrained, are somewhat enriched among genes with more evidence for selection (OR $= 2.64$, $P = 2.8 \times 10^{-4}$ for top 500 gamma-corrected genes; OR $= 5.92$, $P = 1.5 \times 10^{-3}$ for top 2- effect-permuted genes). This is consistent with patterns seen in

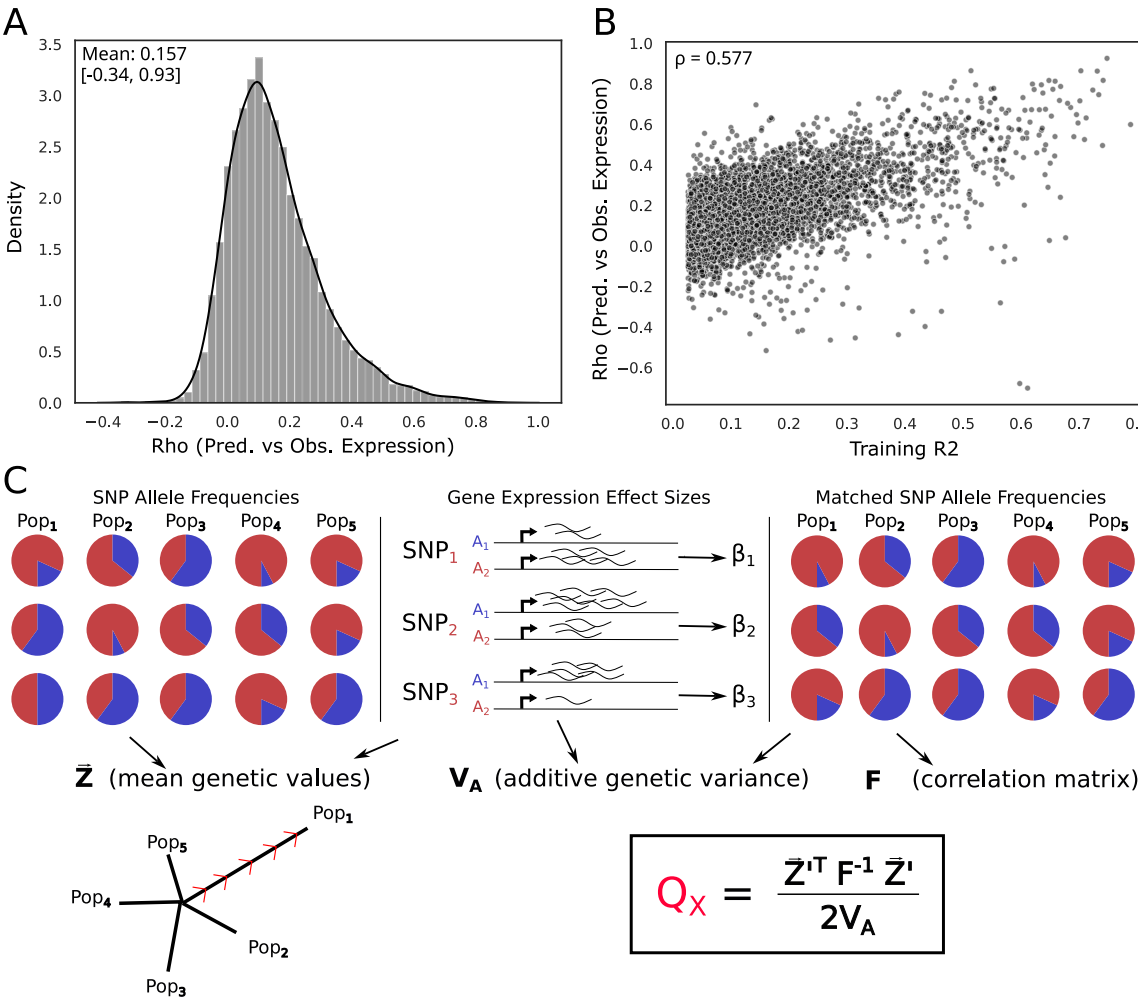


Figure 1: We adapted the Qx statistic to test for selection on regulatory variants.
A) Spearman ρ between observed and predicted expression in 1kG for 7,251 JTI models trained in GTEx LCLs. B) ρ between observed and predicted expression in 1kG for the best JTI models for 26,878 genes C) Schematic of Qx calculation as applied to PrediXcan models.

our previous study of regulatory differences between ancient populations (Colbran et al., 2021). We also tested for enrichment of genes of certain functional categories, such as viral-interacting proteins (Enard et al., 2016), immune genes that respond to interferon, and diet-related genes with known selection signals (Rees et al., 2020), as well as genes that have undergone stabilizing selection on expression across species (Chen et al., 2018), but found no significant trends for any of these categories for either set of P-values. Technically, the selected diet genes were significantly enriched among the top 10 gamma-corrected genes ($P = 0.021$); however that signal is driven entirely by *FADS1*, and is therefore uninformative about broader patterns.

Patterns of predicted expression among selected genes

We next wanted to confirm whether significant Q_X scores corresponded to differences in predicted expression between populations. To test this, we applied the best JTI models to predict expression in the same individuals we used in calculating the Q_X scores and summarized these predictions across populations. Indeed, we found that the genes showing significant selection differed between populations (Fig. 3B). For example, *LY6K* has a median predicted expression 3 standard deviations higher in Japanese populations than in most others. This is true for genes identified as significant under both P-value schemes (Supp.

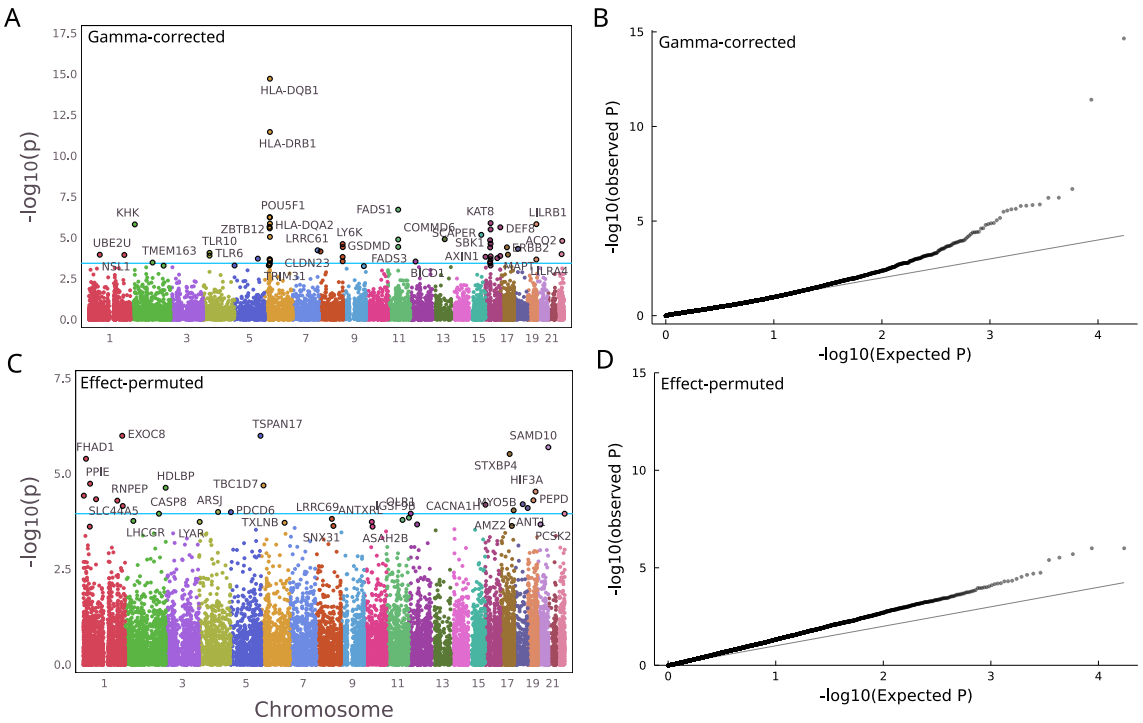


Figure 2: Genes under significant selection vary depending on P-value methodology
A) Manhattan plot and B) QQ-plot for gamma-corrected Q_X p-values per gene. C) Manhattan and D) QQ plots for effect size permuted P-values (Methods). The blue line in each Manhattan plot corresponds to the threshold where FDR < 0.1.

Fig. 8, suggesting both methods identify genes with overall predicted differences. These predicted patterns are also largely similar in HGDP, despite the decreased resolution of that data (Supp. Fig. 9). As expected, these predicted differences do not always agree with patterns of observed expression in LCLs ($P = 0.241$ for 1kG); this could be because of differing tissue expression patterns (we only have observed expression in LCLs), or because of differing environment or genetic background compared to the model training data (Supp. Fig. 10).

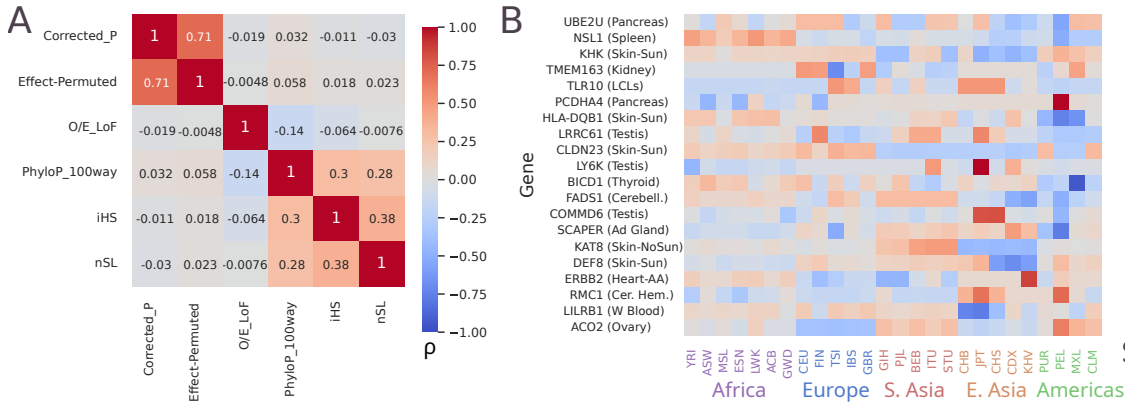


Figure 3: Our Q_X statistic is not correlated with other selection statistics.
A) Pairwise heatmap of Spearman rank correlations between Q_X P-values and various selection-related scores. B) Median predicted expression in each 1kG population for the top gene in each peak of the gamma-corrected P-values. For display purposes, for each gene is standardized across populations.

While there are 20 peaks of significant selection in the gamma-corrected P-values calculated in 1kG, it is unlikely that every gene in each peak is actually under selection. Instead, it is likely that the regulation of one gene has an impact on fitness, and the expression of other genes with shared or linked

regulatory variation hitchhikes along with the selected gene. For example, *FADS1* is a well-established example of a selected haplotype whose effect is correctly modeled and detected (tagged by rs174549; Fig. 4A). However the nearby genes *FEN1* and *FADS3* are significant in this analysis as well, as they are also influenced by the selected haplotype. Additional lines of evidence are therefore necessary to understand which gene in a peak is the cause of the selection, and which are side-effects. We focused on the 4 peaks that replicate in HGDP (the HLA locus, *ACO2*, *KHK*, and *KAT8*). The HLA region is another well-established locus, but the other three are novel. *ACO2* and *KHK* are both the sole genes in their peaks that replicate, so are the most likely candidates in each. *ACO2* is a mitochondrial gene that plays an important role in the TCA cycle (Gruer et al., 1997), and is predicted to be relatively downregulated in Europeans, and upregulated in Peruvians from Lima (PEL) and other Native American populations. Unlike for *FADS1*, these predictions are result of multiple SNP effects, although dominated by the allele frequencies of rs5758389 (Fig. 4B). *KHK* is a gene responsible for catabolizing dietary fructose (Bonthron et al., 1994), and is predicted to be downregulated in Asian populations (JPT, CHS, CDX, KHV, GIH, PJL, STU). While each of these genes was relatively isolated, the *KAT8* peak had 7 genes that replicated between 1kG and HGDP (*KAT8*, *ZNF668*, *ITGAM*, *STX1B*, *SNF646*, *SBK1*, and *SULT1A2*). A closer look shows that this peak in fact represents two independent signals (Supp. Fig. 13), with *SBK1* and *SULT1A2* showing strong predicted differences in PEL, and the other 5 genes showing predicted differences among Asian populations. Each has at least one potential candidate for selection; *ITGAM* is an integrin that is part of the innate immune system (Ramírez-Bello et al., 2019), while *SULT1A2* is important for metabolizing hormones, drugs, and other xenobiotic compounds (Glatt et al., 2001). Both are involved in responding to the environment, and are therefore the most likely to be subject to population-specific selection.

In contrast, the effect size-permuted P-values did not show a tendency to form peaks. Only 23 genes were significant in 1kG, and none of these replicated in HGDP. These 23 significant genes perform a variety of different functions that are also potentially interesting in the context of population-specific selection, including the regulation of insulin secretion (*STXBP4*, the binding of HDL cholesterol (*HDLBP*), and viral replication (PPIE) (Wang et al., 2011). *SAMD10* is the gene that comes closest to replicating in 1kG and HGDP ($P = 2.0 \times 10^{-6}$ in 1kG, $P = 1.2 \times 10^{-5}$ in HGDP), and is a plasma membrane protein that is most highly expressed in the Cerebellum and in LCLs (Aguet et al., 2017). Compared to Europeans (the primary population used in training the JTI models), it is predicted to be upregulated in African and South Asian populations, particularly Gujarati, Indian Telugu and Sri Lankan Tamil, and downregulated in most East Asian populations (Supp. Fig. 8).

Discussion

In this study, we applied the Q_X test for polygenic selection to regulatory variants identified using JTI expression imputation models to test for population-specific selection on gene regulation in 26 human populations. We identified 45 genes with significant regulatory selection. These included loci such as *FADS1*, *TLR*, and the HLA region that have been previously identified, as well as novel loci such as *KHK*, *SULT1A2*, and *ITGAM*. It was common for nearby genes to share high Q_X scores likely reflecting some combination of linkage disequilibrium and shared regulatory variants. We also used a more conservative approach, which uses only the magnitude and direction of effects (conditioning on allele frequency differences). This version only has power to detect genes with coordinated changes across multiple variants and found few genes with evidence of selection. Some of the exceptions include genes associated with metabolism (*HDLBP*, *STXBP4*) and immunity (*PPIE*). Despite correctly identifying some well-established examples, our gene-level Q_X score is not highly correlated with other metrics of

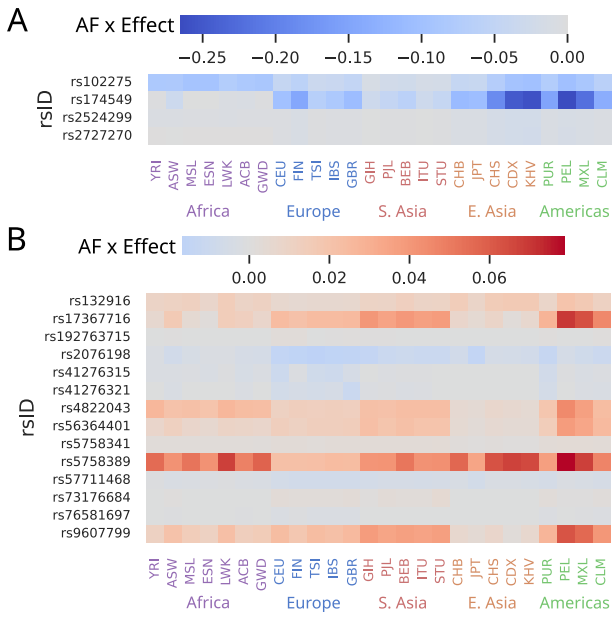


Figure 4: Different combinations of variant effects can drive predicted differences
A) for *FADS1* there is one primary haplotype (tagged by rs174549), while for B) *ACO2* there are 3 variants driving the upregulation in PEL. Cells are coloured by the product of JTI effect size times effect allele frequency in each population.

selection, suggesting that it captures independent information.

There are some caveats with this approach. First, we are limited to testing cis-eQTL identified in the predominantly European GTEx data used to train the JTI models. We therefore cannot test for selection on trans-regulatory effects, or on any population-specific eQTL that were rare or absent in GTEx. Our analysis is also potentially vulnerable to confounding due to population stratification in the gene expression data, which is difficult to correctly account for. Our gamma-controlled controlled analysis is likely susceptible to similar problems seen in the original Q_X studies (Berg et al., 2019). By resampling effect sizes, we controlled for those effects, but at the cost of much-reduced power. In addition, while a high Q_X does correspond to population-level differences in predicted expression in a particular tissue, these differences are not always reflective of actual differences in gene expression. More work in diverse populations and environments will be needed to confirm which of our specific results are true changes in gene expression or merely regulatory turnover.

In addition, our gamma-corrected approach highlights the fact that nearby genes are often co-expressed and share regulatory regions (Delaneau et al., 2019), making it difficult to determine which gene is actively subject to selection. Calculating P-values by permuting the effect sizes does control for these possibilities, but it also severely limits the type of effects we have power to detect. While our approach does allow us to identify genes influenced by potentially selected regulatory haplotypes, this is analogous to the issues with overlapping eQTL and GWAS studies (reviewed by Cano-Gamez and Trynka, 2020). A combination of LD and shared regulatory structure means signals often encompass multiple genes, and the tissue- and context-specificity of gene regulation means that our study, while genome-wide, is not exhaustive. Further lines of evidence will be required to disentangle these associations.

Despite these caveats, we do confirm several known instances of selection. In the case of *FADS1*, our method correctly identifies the known regulatory haplotype, and correctly predicts the direction of differences in expression between populations (Ameur et al., 2012; Buckley et al., 2017; Mathieson and Mathieson, 2018; Ye et al., 2017). The *TLR* locus is the site of a putative case of adaptive Neanderthal introgression (Quach et al., 2016), and our results suggest that this haplotype alters expression of all

three genes in the locus. We identified several novel signals for genes involved in pathways that are most likely to influence fitness in different environments. *KHK*, *ACO2*, *HDLBP*, *STXBP4*, and *SULT1A2* are all genes involved in various aspects of metabolism, whether directly diet-related or further downstream, while *PPIE* and *ITGAM* are both involved in innate immunity.

Overall our work suggests that strong, coordinated, population-specific selection on regulatory variation across multiple haplotypes is relatively rare among human genes and that patterns of variation in *cis*-regulation of gene expression across populations are largely explained by genetic drift. While it is possible that recent selection acts on regulatory variants we do not consider here (e.g. via trans effects), it is also possible that population-specific selection is not particularly common at strengths we are currently able to detect. Finally, our approach demonstrates that biologically-informed tests for selection can contribute orthogonal information to those based around LD patterns or other information, and therefore could be integrated with the results of other selection scans to increase interpretability.

Methods

Regulatory Variant Selection

To select regulatory variants and effect sizes, we started with the published Joint Tissue Imputation (JTI) gene expression models, which were trained in 49 tissues in GTEx v8 using all common variants (MAF > 0.05 in GTEx) (Zhou et al., 2020a). We used these models to predict expression in Lymphoblastoid Cell Lines (LCLs) for 447 individuals from 1000 Genomes Project (Geuvadis; Lappalainen et al., 2013; The 1000 Genomes Project Consortium, 2015). These individuals represented 4 populations of European ancestry (GBR, FIN, CEU, TSI) and one of African ancestry (YRI). We calculated TPM for ENA project PRJEB3366 using EMBL-EBI's REST API (accessed March 8, 2022). We compared the predicted to observed expression for these individuals by calculating a Spearman rank correlation for 7,251 gene models trained in LCLs across all individuals. We compared the agreement of the models to the variance explained by the models in the training data by calculating the rank correlation across all genes between the model R^2 and the predicted/observed correlation. For each gene, we selected the "best" model by choosing the model with the highest R^2 across all tissues. This resulted in 26,878 genes for which we could compare observed and predicted expression.

Qx score adaptation

We obtained effect sizes for variants by filtering the best models to include only protein coding genes (based on the "protein_coding" annotation in GenCode v26). The resulting 17,388 genes had a median of 12 (maximum 101) variants with effect sizes to calculate Q_X. We calculated Q_X as described by Berg and Coop (2014), using the effect sizes from each gene expression model in place of GWAS effect sizes. This allowed us to calculated a per-gene Q_X score. For each gene, we chose 100 matched variants for each model variant by binning all variants into 25 bins based on alternate allele frequency in GTEx (i.e. a bin size of 0.02). We calculated Q_X across the 26 populations of 1kG, encompassing 2504 unrelated individuals (Byrska-Bishop et al., 2022). We replicated using 929 individuals from the Human Genome Diversity Panel (HGDP) (Bergström et al., 2020), grouped into 7 broadly regional 'populations' due to the small sample size in individual populations. We plotted Manhattan plots with OpenMendel (Zhou et al., 2020b).

P-value calculation

While Q_X was designed to be a test for polygenic selection testing genome-wide, independent sets of variants, our adaptation of it would be better described as a multi-variant test for selection. The set of possible variants for each gene were pruned for linkage disequilibrium (LD) at $r^2 > 0.8$ before the JTI models were trained (Zhou et al., 2020a), and models are built around independent, additive effects for the variants ultimately included. However, these variants are much closer together (within 2Mb), and models often include variants in moderate LD ($r^2 \approx 0.4$) with each other. This means that, while the effect sizes are independent, the allele frequencies are not necessarily, and the degree to which the frequencies are correlated with each other varies across genes. This makes calculating a P-value for gene-level Q_X statistics complicated.

We tried three different methods for calculating P-values. The first is the method used in the original Q_X study, wherein we construct a 'null' distribution of Q_X scores for each gene by permuting the allele frequencies of the variants in the model (abbreviated as "freqPerm"; Supp. Fig. 3). For each permutation, we drew a random frequency for each variant in the model, holding effect sizes constant, and repeated that 100,000 times (up to 1,000,000 times if $P < 10^{-4}$). As expected, because this permutation strategy breaks the LD structure present in many gene models, the resulting P-values are extremely poorly calibrated (Supp. Fig. 4).

We therefore calculated "corrected" P-values by instead fitting a gamma distribution to the Q_X scores (with degrees of freedom equal to the number of populations minus 1). These P-values are much less inflated (Fig. 2B), while the ordering of genes is highly correlated with the order the freqPerm P-values gave (Spearman $\rho = 0.993$). They are however, somewhat correlated with technical characteristics of the gene models (Supp. Fig. 5).

To control for the technical confounding, we also calculated p-values by permuting the effect sizes for each gene while holding allele frequencies of variants constant (abbreviated as "effPerm"). Specifically, we randomly sampled effect sizes from the distribution of all possible effect sizes in any model, while holding the effect direction for each variant constant. We sampled 100,000 times for each gene, up to 1,000,000 for those with $P < 10^{-4}$. While this did control for the technical variables (Supp. Fig. 5) and was still correlated with the corrected P-values (Spearman $\rho = 0.709$), it has the side effect of narrowing the hypothesis we were testing. Rather than a broad test for selection on regulatory variants, this permutation scheme emphasizes coordinated differences between populations (i.e. in the same effect direction) across multiple variants in a model. This means that true selection on a single regulatory haplotype (e.g. in the case of *FADS1*) is not identified.

Population Expression comparison

We predicted expression by applying the best JTI Models to the same individuals used to calculate Q_X (2504 unrelated individuals from 1kG, and 929 from HGDP). We summarized across populations by taking the median predicted expression within each population.

We also compared this predicted expression observed expression in LCLs for both datasets. For 1kG, this was the same data described in *Regulatory Variant Selection*. For HGDP, this included 45 individuals from 5 geographic regions– Africa, the Middle East, East Asia, South Asia, and the Americas. We calculated the significance of the agreement across genes in the analysis as following: For each gene, we calculated a Spearman correlation between the median observed and predicted expression across all populations. We then summed the correlations together. We calculated an empirical P-value for each plot by shuffling the medians 10,000 times and repeating the summation.

Other Selection Scores

We obtained loss-of-function (LoF) tolerance scores for each gene from Gnomad v2 (Lek et al., 2016), and used the observed/expected number of LoF variants as a measure of conservation on the protein sequence (where a low score indicates more constraint). We downloaded the phyloP100way track from the UCSC Genome Browser, and calculated the average score for each gene across the 200kb window centered around the gene using the bigWigAverageOverBed tool. Positive phyloP scores indicate greater conservation in the region.

We calculated iHS and nSL statistics in all 26 1kG populations using SelScan 2.0.0 (Szpiech, 2021). nSL was calculated using unphased genomes (Ferrer-Admetlla et al., 2014). For iHS, we used phased genomes (Voight et al., 2006), and polarized ancestral and derived alleles based on the Chimpanzee genome. We interpolated recombination maps for our sites from 1kG maps (Spence and Song, 2022). For both we focused on variants with MAF > 0.05 in the population in question, then took the mean statistic across the 200kb window around each gene.

Enrichment Tests

We tested for enrichment for particular gene sets by calculating a 'tiered' enrichment. We sorted the genes by p-value, then tested for enrichment in the top N (for N = 10, 20, 30, 50, 75, 100, 150, 200, 300, 500, 1000, 2000, 3000, 5000 and 10000). Enrichment is calculated as the proportion of genes in the top N divided by the overall proportion that are in the gene set in question. We used the Binomial test to calculate a P-value, and calculated a confidence interval for each N using the Agresti-Coull method (Agresti and Coull, 1998). We did this for 3788 housekeeping genes (Eisenberg and Levanon, 2013) and for 2899 LoF-intolerant genes, where a gene is LoF-intolerant if the upper bound of the 95% confidence interval of the observed/expected ratio is lower than 0.35 (Lek et al., 2016), as well as 5352 genes that have undergone balancing selection on expression across species (Chen et al., 2018). We additionally tested sets of genes based on their function. These included 20 diet genes with previously-identified signals of selection (Rees et al., 2020) as well as two sets of immune genes: 1257 virus-interacting proteins (Enard et al., 2016) and 128 interferon response genes (all products in the gene ontology term GO:0032606 and all child terms).

Data and Code Availability

Q_X scores and P-values are available as supplementary files here, and all other data are publicly available. Scripts for parsing data files and running analyses are available via Github.

Acknowledgements

We thank Ziyue Gao, ML Benton and Sarah Fong for helpful discussions, Christopher Adams for help with the 1000 Genomes data, and Lin Poyraz for a script to polarize iHS scores.

Funding

This project was supported by the National Human Genome Research Institute training grant T32HG009495 to the University of Pennsylvania (L.L.C.) and the National Institute of General Medical Sciences R35GM133708 (I.M). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

L.L.C. and I.M. designed the experiments and wrote the manuscript. L.L.C conducted all experiments and analyzed data.

Competing Interests

The authors report no conflicts of interest.

Bibliography

Alan Agresti and Brent A. Coull. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, 1998.

François Aguet, Kristin G. Ardlie, Beryl B. Cummings, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017. ISSN 0028-0836. doi: 10.1038/nature24277.

Adam Ameur, Stefan Enroth, Asa Johansson, et al. Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *American journal of human genetics*, 90(5):809–820, may 2012. ISSN 1537-6605. doi: 10.1016/j.ajhg.2012.03.014.

Mary Lauren Benton, Sai Charan Talipineni, Dennis Kostka, and John A Capra. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics*, 20 (1):511, 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-5779-x.

Jeremy J. Berg and Graham Coop. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics*, 10(8), 2014. ISSN 15537404. doi: 10.1371/journal.pgen.1004412.

Jeremy J. Berg, Arbel Harpak, Nasa Sinnott-Armstrong, et al. Reduced signal for polygenic adaptation of height in UK biobank. *eLife*, 8:1–47, 2019. ISSN 2050084X. doi: 10.7554/eLife.39725.

Anders Bergström, Shane A McCarthy, Ruoyun Hui, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484):eaay5012, mar 2020. doi: 10.1126/science.aay5012.

Ran Blekhman, Alicia Oshlack, Adrien E Chabot, et al. Gene Regulation in Primates Evolves under Tissue-Specific Selection Pressures. *PLOS Genetics*, 4(11):e1000271, nov 2008.

David T Bonthron, Nlcola Brady, lain A Donaldson, and Beat Steinmann. Molecular basis of essential fructosuria: molecular cloning and mutational analysis of human ketohexokinase (fructokinase). *Human Molecular Genetics*, 3(9):1627–1631, sep 1994. ISSN 0964-6906. doi: 10.1093/hmg/3.9.1627.

Matthew T Buckley, Fernando Racimo, Morten E Allentoft, et al. Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Molecular biology and evolution*, 34(6):1307–1318, jun 2017. ISSN 1537-1719 (Electronic). doi: 10.1093/molbev/msx103.

Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, 2022. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2022.08.004.

Eddie Cano-Gamez and Gosia Trynka. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, 11(May):1–21, 2020. ISSN 16648021. doi: 10.3389/fgene.2020.00424.

Jenny Chen, Ross Swofford, Jeremy Johnson, et al. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research*, 29(1):53–63, 2018. ISSN 1088-9051. doi: 10.1101/gr.237636.118.

Laura L Colbran, Maya R Johnson, Iain Mathieson, and John A Capra. Tracing the Evolution of Human Gene Regulation and Its Association with Shifts in Environment. *Genome Biology and Evolution*, 13 (11):evab237, nov 2021. ISSN 1759-6653. doi: 10.1093/gbe/evab237.

Olivia Corradin, Andrea J Cohen, Jennifer M Luppino, et al. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat Genet*, advance on, 2016. doi: 10.1038/ng.3674http://www.nature.com/ng/journal/vaop/ncurrent/abs/ng.3674.html#supplementary-information.

Nicholas G Crawford, Derek E Kelly, Matthew E B Hansen, et al. Loci associated with skin pigmentation identified in African populations. *Science*, 358(6365):eaan8433, nov 2017. doi: 10.1126/science.aan8433.

O Delaneau, M Zazhytska, C Borel, et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, 364(6439):eaat8266, may 2019. doi: 10.1126/science.aat8266.

Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10): 569–574, 2013. ISSN 01689525. doi: 10.1016/j.tig.2013.05.010.

David Enard, Le Cai, Carina Gwennap, and Dmitri A Petrov. Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5:e12469, 2016. ISSN 2050-084X. doi: 10.7554/eLife.12469.

Anna Ferrer-Admetlla, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular biology and evolution*, 31(5): 1275–1291, may 2014. ISSN 1537-1719 (Electronic). doi: 10.1093/molbev/msu077.

Yair Field, Evan A Boyle, Natalie Telis, et al. Detection of human adaptation during the past 2000 years. *Science*, 2016.

Molly Gasperini, Jacob M Tome, and Jay Shendure. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21(5):292–310, 2020. ISSN 1471-0064. doi: 10.1038/s41576-019-0209-0.

Hansruedi Glatt, Heiner Boeing, Christina E H Engelke, et al. Human cytosolic sulphotransferases: genetics, characteristics, toxicological aspects. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 482(1):27–40, 2001. ISSN 0027-5107. doi: https://doi.org/10.1016/S0027-5107(01)00207-X.

Megan J Gruer, Peter J Artymiuk, and John R Guest. The aconitase family: three structural variations on a common theme. *Trends in Biochemical Sciences*, 22(1):3–6, 1997. ISSN 0968-0004. doi: https://doi.org/10.1016/S0968-0004(96)10069-4.

Melissa A. Ilardo, Ida Moltke, Thorfinn S. Korneliussen, et al. Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, 173(3):569–573.e15, 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.03.054.

Mary Claire King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188 (4184):107–116, 1975. ISSN 00368075. doi: 10.1126/science.1090005.

Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501:506, sep 2013.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536:285, aug 2016.

Iain Mathieson, Iosif Lazaridis, Nadin Rohland, et al. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583):499–503, 2015. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature16152.

Sara Mathieson and Iain Mathieson. FADS1 and the timing of human adaptation to agriculture. *Molecular Biology and Evolution*, 35(12):2957–2970, 2018. ISSN 15371719. doi: 10.1093/molbev/msy180.

Lauren E. Petty, Heather M. Highland, Eric R. Gamazon, et al. Functionally oriented analysis of cardiometabolic traits in a trans-ethnic sample. *Human Molecular Genetics*, 00(00):1–13, 2019. ISSN 0964-6906. doi: 10.1093/hmg/ddy435.

Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, jan 2010. ISSN 1549-5469 (Electronic). doi: 10.1101/gr.097857.109.

Peter D Price, Daniela H Palmer Droguett, Jessica A Taylor, et al. Detecting signatures of selection on gene expression. *Nature Ecology & Evolution*, 6(7):1035–1045, 2022. ISSN 2397-334X. doi: 10.1038/s41559-022-01761-8.

Hélène Quach, Maxime Rotival, Julien Pothlichet, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*, 167(3):643–656.e17, 2016. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2016.09.024.

Julian Ramírez-Bello, Celi Sun, Guillermo Valencia-Pacheco, et al. ITGAM is a risk factor to systemic lupus erythematosus and possibly a protection factor to rheumatoid arthritis in patients from Mexico. *PloS one*, 14(11):e0224543, 2019. ISSN 1932-6203 (Electronic). doi: 10.1371/journal.pone.0224543.

Jasmin S Rees, Sergi Castellano, and Aida M Andrés. The Genomics of Human Local Adaptation. *Trends in Genetics*, 36(6):415–428, jun 2020. ISSN 0168-9525. doi: 10.1016/j.tig.2020.03.006.

Rori V Rohlfs, Patrick Harrigan, and Rasmus Nielsen. Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation. *Molecular Biology and Evolution*, 31(1):201–211, jan 2014. ISSN 0737-4038. doi: 10.1093/molbev/mst190.

Sarah A Signor and Sergey V Nuzhdin. The Evolution of Gene Expression in cis and trans. *Trends in Genetics*, 34(7):532–544, jul 2018. ISSN 0168-9525. doi: 10.1016/j.tig.2018.03.007.

Tatum S Simonson, Yingzhong Yang, Chad D Huff, et al. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science*, 329(5987):72 LP – 75, jul 2010.

Jeffrey P Spence and Yun S Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10):eaaw9206, oct 2022. doi: 10.1126/sciadv.aaw9206.

Zachary A Szpiech. selscan 2.0: scanning for sweeps in unphased data. *bioRxiv*, page 2021.10.22.465497, jan 2021. doi: 10.1101/2021.10.22.465497.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, 2015. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature15393.

Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A Map of Recent Positive Selection in the Human Genome. *PLOS Biology*, 4(3):e72, mar 2006.

Zengfu Wang, Xiaoling Liu, Zhendong Zhao, et al. Cyclophilin E Functions as a Negative Regulator to Influenza Virus Replication by Impairing the Formation of the Viral Ribonucleoprotein Complex. *PLOS ONE*, 6(8):e22625, aug 2011.

M. C. Whitlock. Evolutionary inference from qst. *Mol Ecol*, 17(8):1885–96, 2008. ISSN 1365-294X (Electronic) 0962-1083 (Linking). doi: 10.1111/j.1365-294X.2008.03712.x.

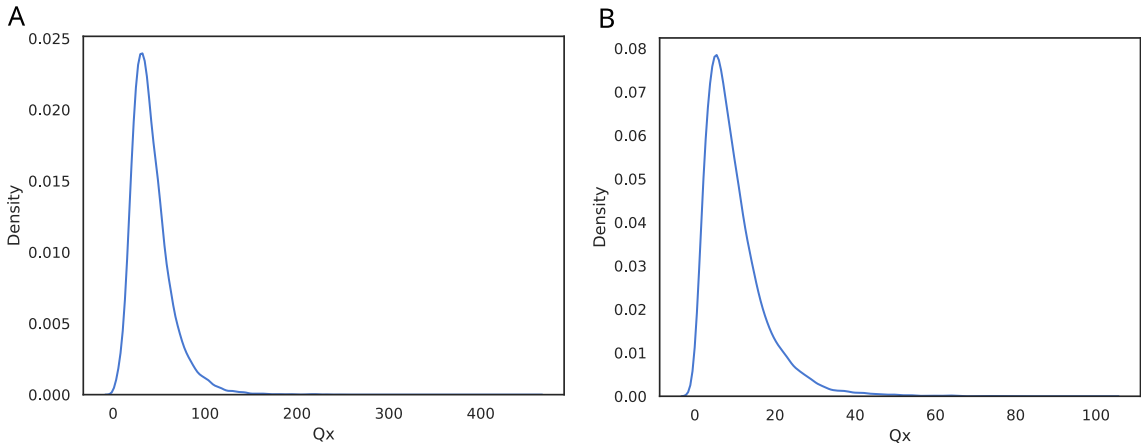
Kaixiong Ye, Feng Gao, David Wang, et al. Dietary adaptation of FADS genes in Europe varied across time and geography. *Nature Ecology & Evolution*, 1(7):167, 2017. ISSN 2397-334X. doi: 10.1038/s41559-017-0167.

Dan Zhou, Yi Jiang, Xue Zhong, et al. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nature genetics*, oct 2020a. ISSN 1546-1718 (Electronic). doi: 10.1038/s41588-020-0706-2.

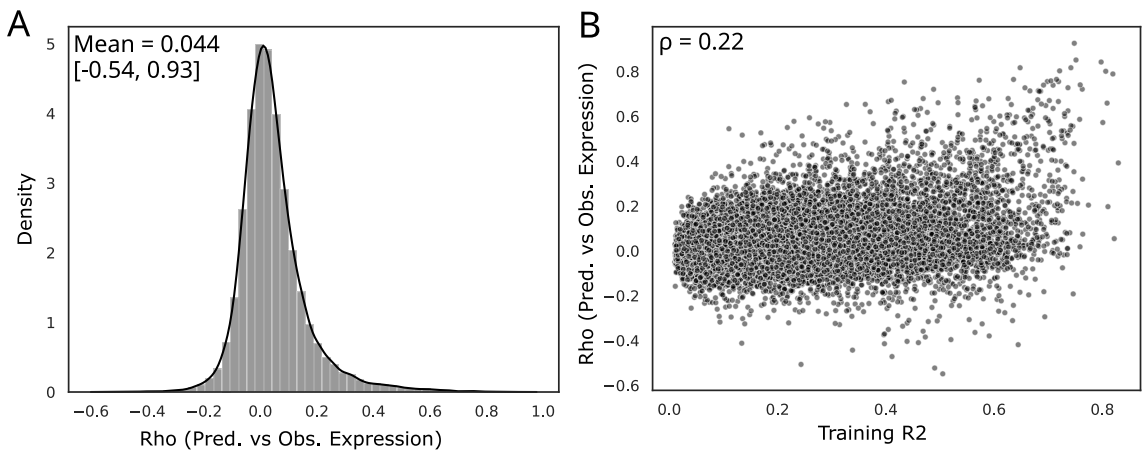
Hua Zhou, Janet S Sinsheimer, Douglas M Bates, et al. OpenMendel: a cooperative programming project for statistical genetics. *Human Genetics*, 139(1):61–71, 2020b. ISSN 1432-1203. doi: 10.1007/s00439-019-02001-z.

Huanhuan Zhu and Xiang Zhou. Transcriptome-wide association studies: a view from Mendelian randomization. *Quantitative Biology*, 2020. ISSN 2095-4697. doi: 10.1007/s40484-020-0207-4.

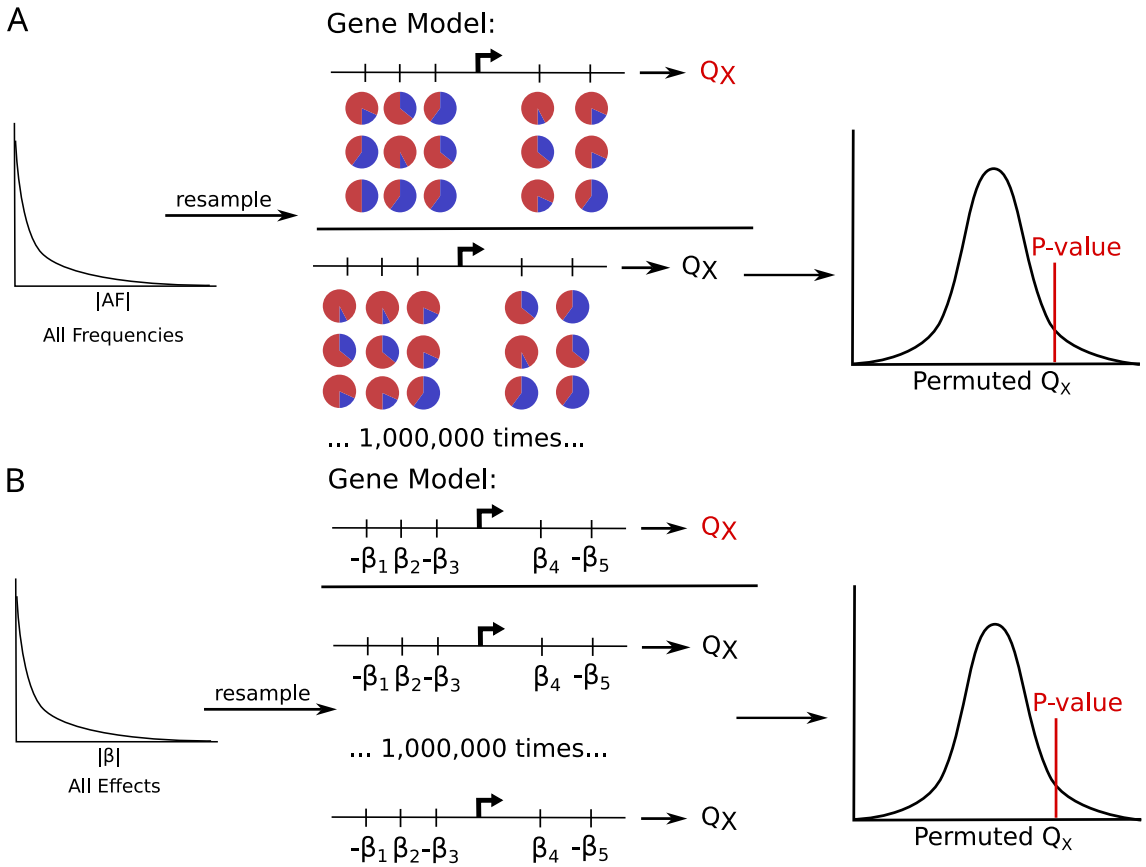
Supplementary Materials



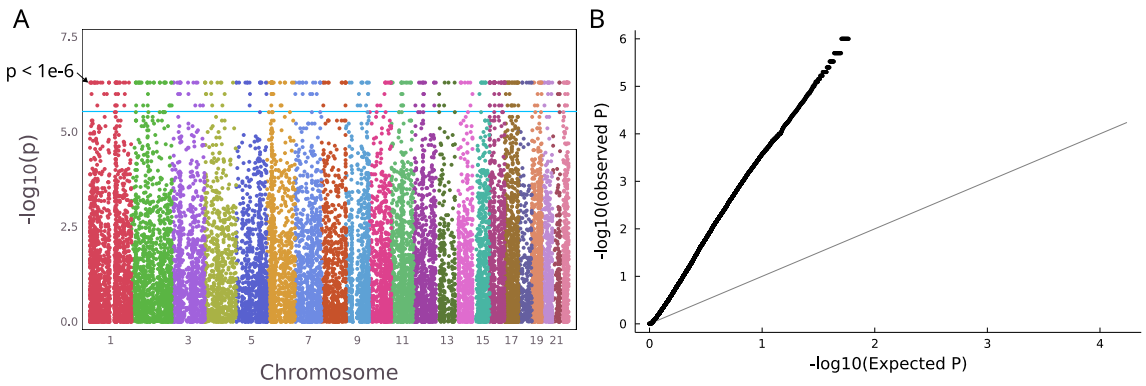
Supplementary Figure 1: Distributions of Q_X scores for the Best Models of all protein-coding genes in A) 26 1kG B) 7 HGDP populations.



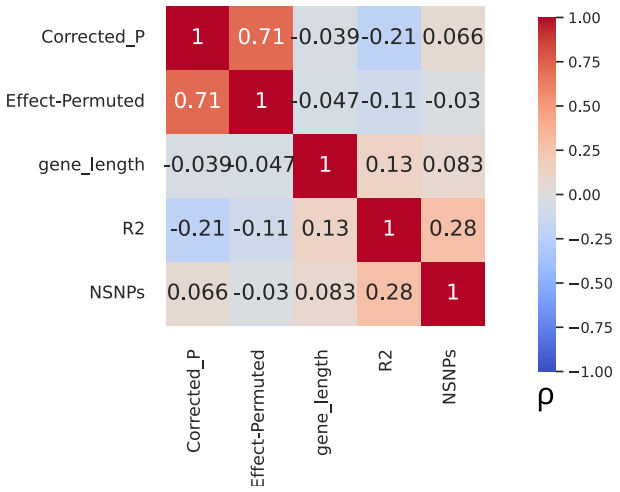
Supplementary Figure 2: A) Rank correlation between observed and predicted expression for the Best Models of all protein-coding genes in LCLs B) Scatterplot of that rank correlation vs R^2 from the JTI training process.



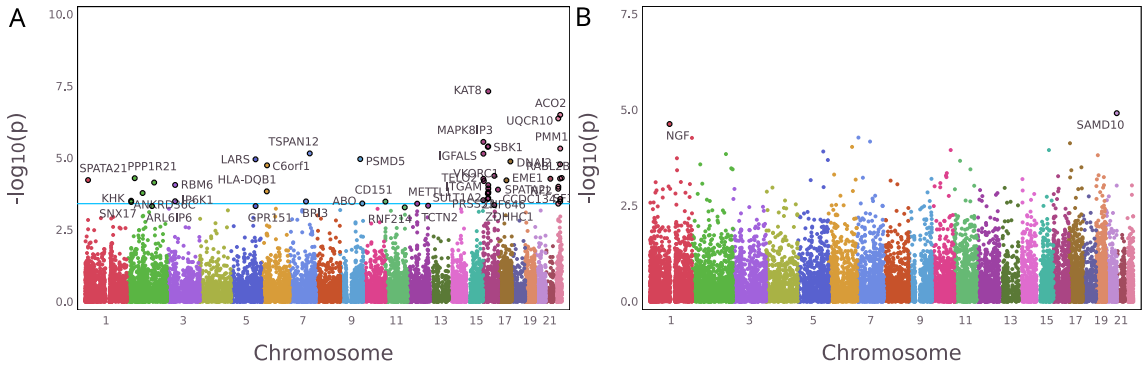
Supplementary Figure 3: Schematics describing the two permutations schemes we tried.
A) One option is to calculate an empirical P-value by permuting the variant allele frequencies across a model. B) The other option is to instead permute the effect sizes of the variants, while holding effect direction constant. In both, we resampled 100,000 times for each gene, extending to 1,000,000 for those with $P < 10^{-4}$.



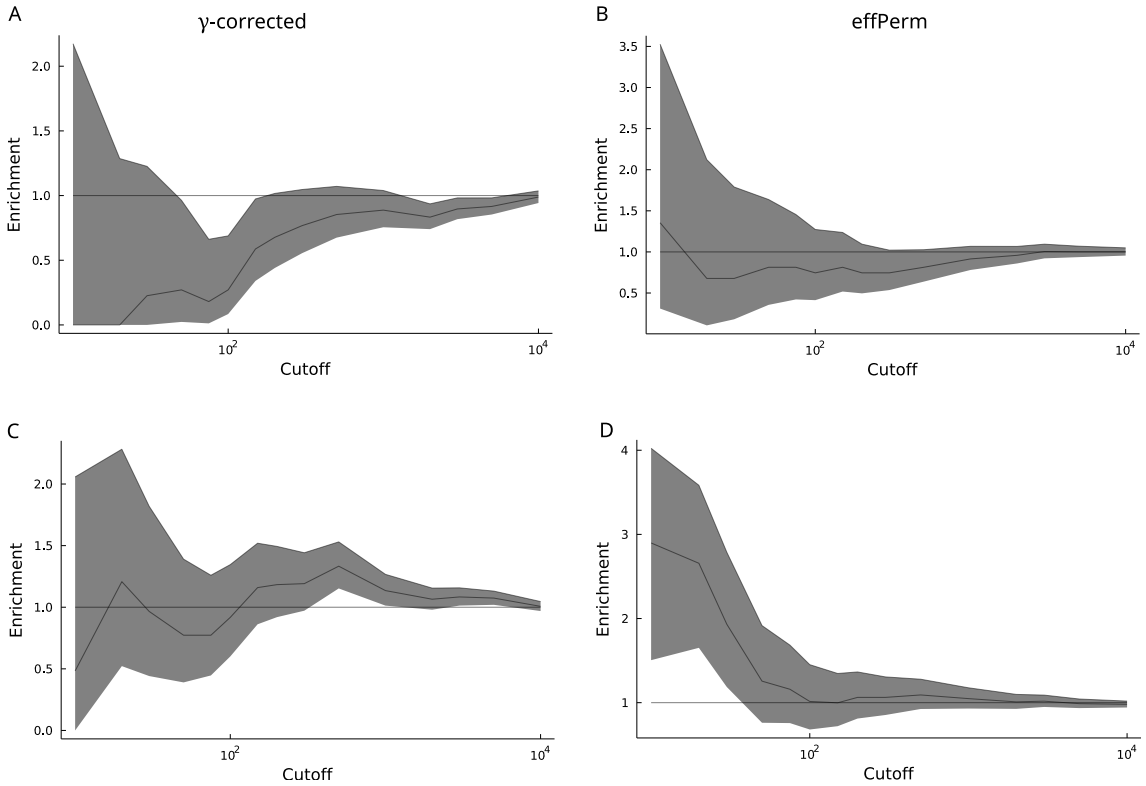
Supplementary Figure 4: P-values for the Best Model protein-coding genes calculated by permuting the allele frequencies of the model variants.
A) Manhattan plot; the maximum of $-log_{10}(P)$ is where $P < 1{,}000{,}000$. B) a QQ plot of the P-values, excluding those for which $P < 1{,}000{,}000$.



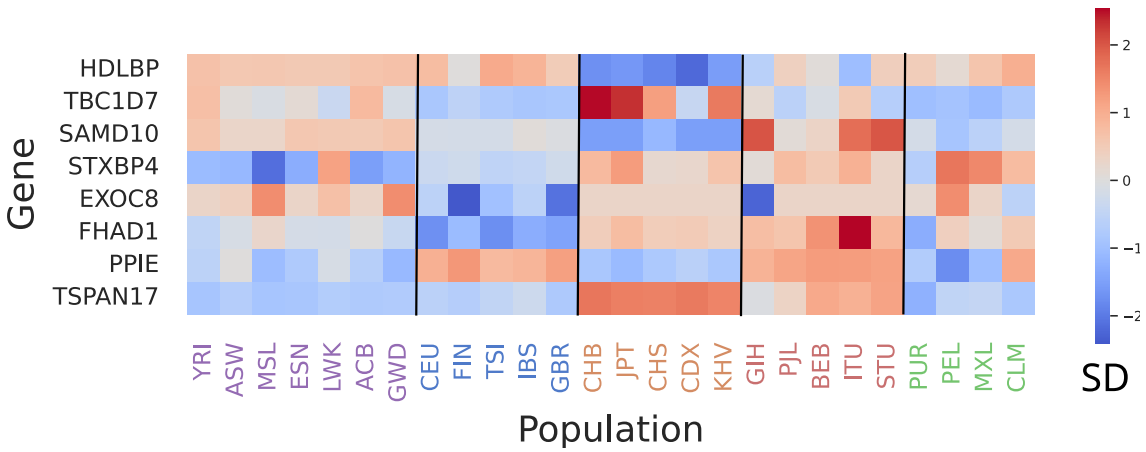
Supplementary Figure 5: Q_X is correlated with technical aspects of JTI models.
Pairwise Spearman rank correlations between gamma-corrected P-values and effect size permutation P-values with various technical variables. R^2 and $NSNPs$ refer to the training performance of the Best JTI models used to calculate Q_X



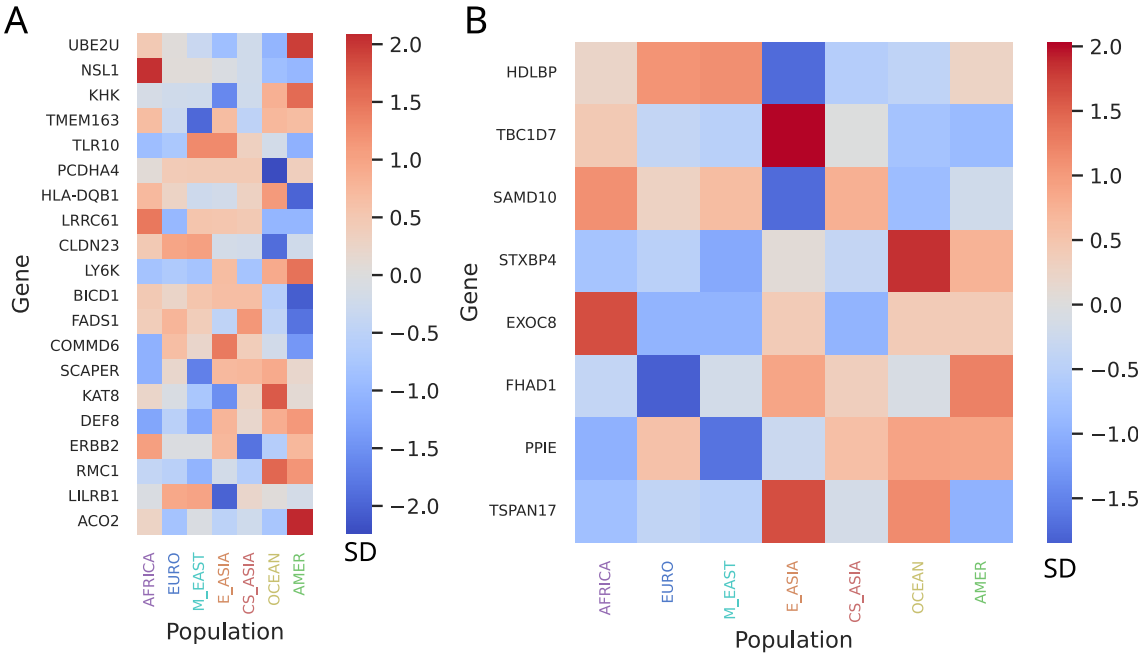
Supplementary Figure 6: HGDP analysis.
Manhattan plots of A) gamma-corrected P-values and B) effect size permuted P-values for Q_X calculated in HGDP. Blue line corresponds to FDR < 0.1; for B) the minimum FDR is 0.19.



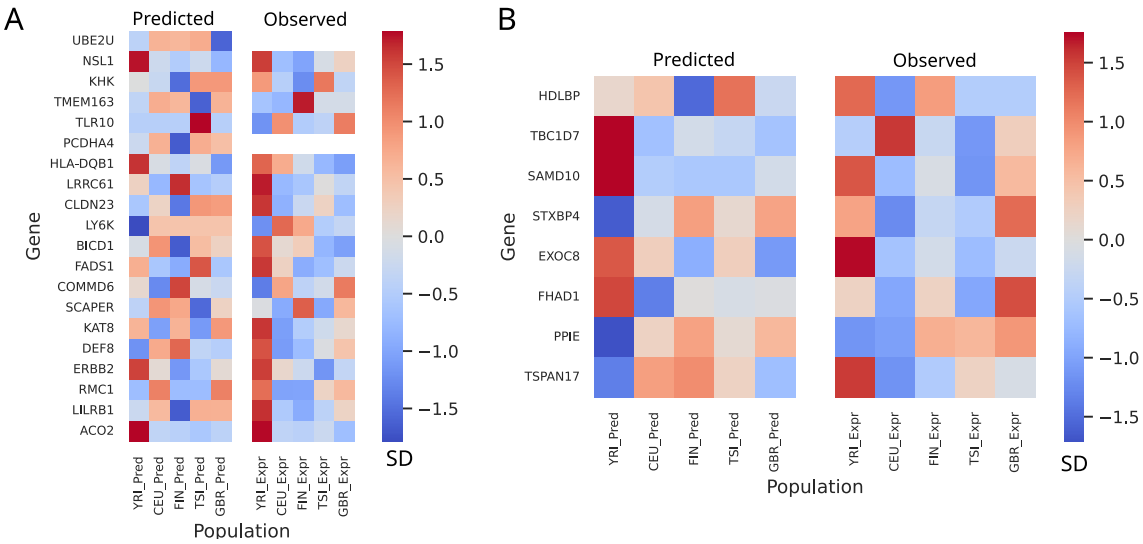
Supplementary Figure 7: Predicted regulation for genes with $FDR < 0.05$ using gamma-corrected and effect size permuted P-values.
A) and B) LoF intolerant, C) and D) Housekeeping genes



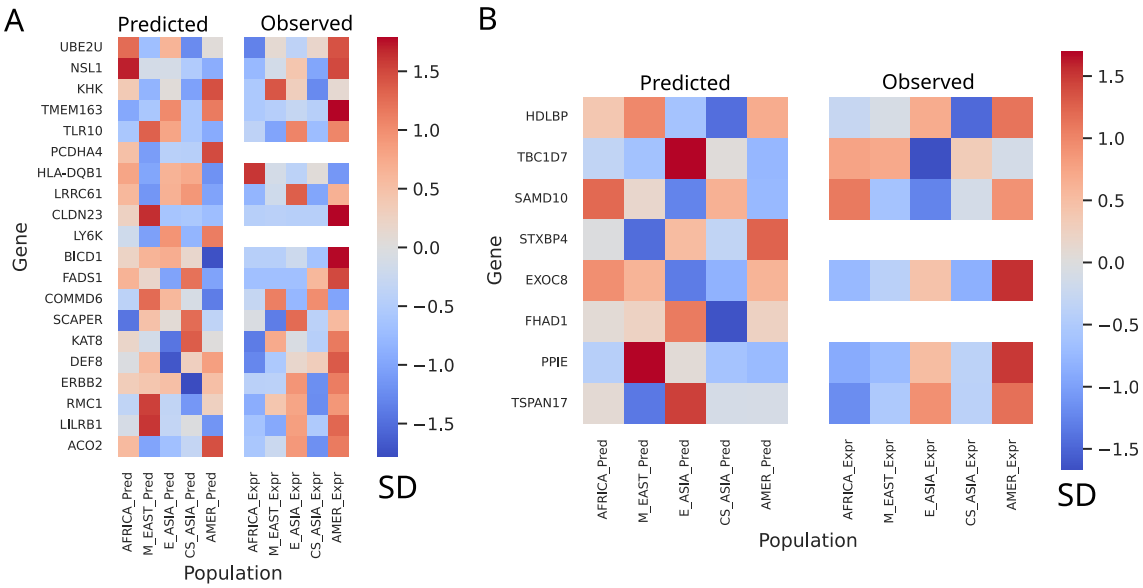
Supplementary Figure 8: Predicted expression for genes with $FDR < 0.05$ using effect size permuted P-values.
Each square is coloured by the median predicted expression for that gene in that population, and is standardized across the row.



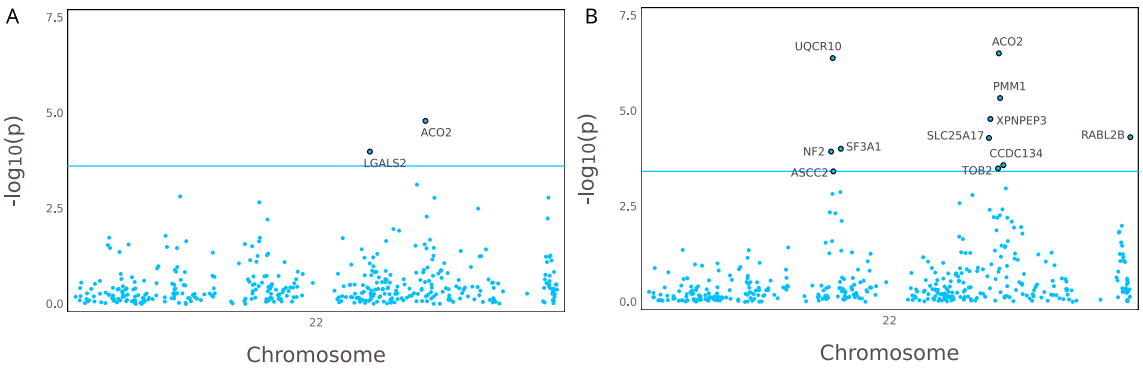
Supplementary Figure 9: HGDP Predicted expression patterns largely agree with 1kG across populations. Heatmap of median predicted expression in each HGDP population A) the top genes in each 'peak' based on the gamma-corrected P-values, and B) genes with $FDR < 0.05$ for the effPerm P-values in 1kG. For display purposes, values are standardized across populations for each gene.



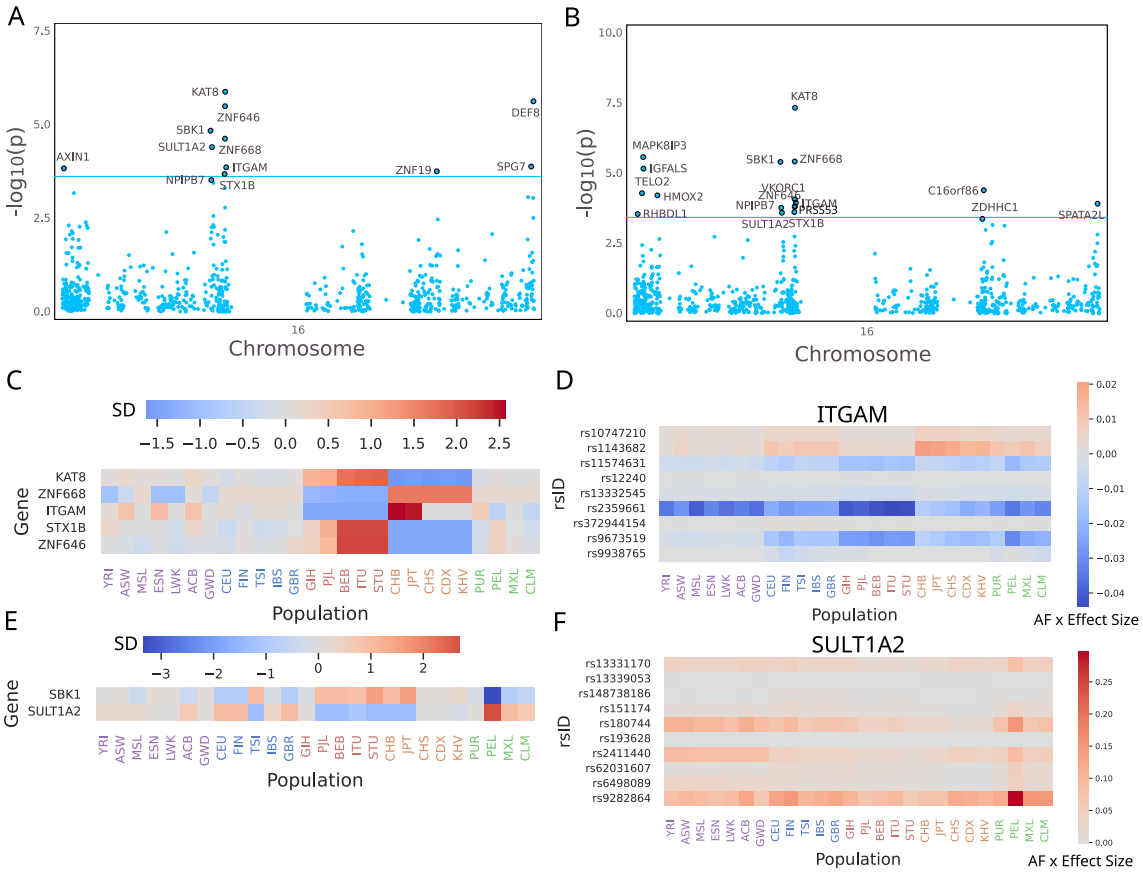
Supplementary Figure 10: Predicted expression pattern does not necessarily correlate with observed expression in LCLs.
Heatmaps of median predicted and observed expression in each 1kG population for A) the top genes in each 'peak' based on the gamma-corrected P-values ($P = 0.241$), and B) genes with $FDR < 0.05$ for the effPerm P-values ($P = 0.717$). For display purposes, values are standardized across populations for each gene.



Supplementary Figure 11: Predicted expression pattern does not necessarily correlate with observed expression in LCLs.
Heatmaps of median predicted and observed expression in each HGDP population for A) the top genes in each 'peak' based on the gamma-corrected P-values ($P = 0.612$), and B) genes with $FDR < 0.05$ for the effPerm P-values ($P = 0.697$). For display purposes, values are standardized across populations for each gene.



Supplementary Figure 12: *ACO2* is under selection in 1kG and HGDP
Zoomed Manhattan plots of chr22 for A) 1kG and B) HGDP.



Supplementary Figure 13: the *KAT8* peak is actually 2 independent signals.
Zoomed Manhattan plots of chr16 for A) 1kG and B) HGDP. C) Median Predicted expression for the 5 replicating genes near *KAT8*; D) Heatmap of the product of JTI effect size times effect allele frequency for variants in the *ITGAM* model. E) Median predicted expression for SBK1 and SULT1A2, which replicated in 1kG and HGDP, and F) Heatmap of the product of JTI effect size times effect allele frequency for variants in the *SULT1A2* model.