Transcriptome Sequencing of a Large Human Family Identifies the Impact of Rare Noncoding Variants

Xin Li,[1,*] Alexis Battle,[2,3,5] Konrad J. Karczewski,[2] Zach Zappala,[2] David A. Knowles,[3] Kevin S. Smith,[1] Kim R. Kukurba,[2] Eric Wu,[1] Noah Simon,[4] and Stephen B. Montgomery[1,2,3,*]

Recent and rapid human population growth has led to an excess of rare genetic variants that are expected to contribute to an individual's genetic burden of disease risk. To date, much of the focus has been on rare protein-coding variants, for which potential impact can be estimated from the genetic code, but determining the impact of rare noncoding variants has been more challenging. To improve our understanding of such variants, we combined high-quality genome sequencing and RNA sequencing data from a 17-individual, three-generation family to contrast expression quantitative trait loci (eQTLs) and splicing quantitative trait loci (sQTLs) within this family to eQTLs and sQTLs within a population sample. Using this design, we found that eQTLs and sQTLs with large effects in the family were enriched with rare regulatory and splicing variants (minor allele frequency < 0.01). They were also more likely to influence essential genes and genes involved in complex disease. In addition, we tested the capacity of diverse noncoding annotation to predict the impact of rare noncoding variants. We found that distance to the transcription start site, evolutionary constraint, and epigenetic annotation were considerably more informative for predicting the impact of rare variants than for predicting the impact of common variants. These results highlight that rare noncoding variants are important contributors to individual gene-expression profiles and further demonstrate a significant capability for genomic annotation to predict the impact of rare noncoding variants.

Introduction

Studies using deep and population-scale sequencing have reported large numbers of rare variants (minor allele frequency [MAF] $< 1\%$) present as a consequence of recent and rapid human population expansion.[1–6] However, interpreting the impact of rare variation remains an ongoing challenge. Several exome sequencing studies have suggested that rare variants are of broad importance with the finding that they represent the majority of potentially deleterious and damaging protein-coding alleles[2] and can contribute to complex disease risk.[7–11] In contrast, population-genetic models have indicated that rare alleles are unlikely to be large overall contributors to heritable variation for many complex diseases.[12] Indeed, large population studies of rare variants in autoimmune disorders have so far found negligible impact,[13] and analyses of personal genomes have reported multiple rare and protein-code-disrupting sites in presumably healthy individuals.[14,15] Further compounding the challenge of understanding the impact of rare variation has been that most studies have focused on only protein-coding alleles whose interpretation is facilitated by the genetic code. For rare variants in noncoding regions, there is no analogous code to aid in the prediction of their impact even though these regions harbor considerable complex-disease-associated variation[16,17] and most likely contain an abundance of important rare alleles.

Currently, genetic studies of gene expression provide a systematic means of identifying functional noncoding variants; such studies have identified noncoding variants associated with gene expression, splicing, and allele-specific expression (ASE).[18–20] However, insight into the impact of rare noncoding variants has been limited. Few studies have had the advantage of full genome sequencing data and, even when these data are available, they have only assayed unrelated individuals, providing limited power to describe rare-variant effects.[18,21,22] To overcome this challenge and provide more systematic insight into the impact of rare noncoding variants, we coupled high-quality genomes with transcriptomes within a large family ($n = 17$ individuals). The advantage of this design is that the large number of children ($n = 11$) provides high-confidence rare variants established through both deep sequencing and Mendelian segregation as well as sufficient power to test for *cis*-expression quantitative trait loci (eQTLs) present within a single human family. Furthermore, eQTLs from the family can be compared to eQTLs from a cell-type- and ethnicity-matched population sample recently reported by the Geuvadis Consortium,[18] providing the unique ability to identify large genetic effects specific to the family and test their relationship to rare variants.[23] Indeed, we report that rare regulatory variants are enriched near genes that exhibit large-effect *cis*-eQTLs for gene expression, splicing, and ASE within the family. Furthermore, the family eQTL genes are more evolutionarily constrained than comparable eQTL genes in the population, and several of the genes have established relationships with complex disease,

[1]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA; [2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [3]Department of Computer Science, Stanford University, Stanford, CA 94305, USA; [4]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
[5]Present address: Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
*Correspondence: xxli@stanford.edu (X.L.), smontgom@stanford.edu (S.B.M.)
http://dx.doi.org/10.1016/j.ajhg.2014.08.004.

indicating a potential for rare variants to further influence genetic risk.

In addition, as genome-interpretation approaches are becoming increasingly informed by diverse noncoding genome annotation,[24–26] genome and transcriptome analysis within a single large family provides unique insight into the predictive power of diverse noncoding annotation for rare variants. In our study, we demonstrated that the combination of variant location, epigenomic information, and evolutionary constraint is considerably more informative for predicting the impact of rare noncoding variants than for predicting the impact of common variants. Likewise, we observed equivalent increases in predictive strength for rare splicing variants. This suggests that many rare noncoding variants are likely to be interpretable via existing noncoding annotation and supports their more routine integration in rare-variant association studies.

Material and Methods

Cell Culture and RNA Sequencing

Epstein-Barr-virus-transformed peripheral blood B lymphocytes (catalog no. XC01463) from families from the CEU population (Utah residents with ancestry from northern and western Europe from the CEPH collection) were purchased from the Coriell Institute and grown in RPMI 1640 supplemented with 10% fetal calf serum and penicillin and streptomycin in humidified 5% CO_2 at a concentration of ~1 × 10^6 cells/ml. Total RNA was isolated with Trizol. RNA quality was assessed with the Agilent Bioanalyzer 2100, and RNA integrity numbers above 9 were used for cDNA production. One microgram of total RNA was used for isolating polyA-purified mRNA and subsequently used for cDNA-library construction with the Illumina TruSeq RNA Preparation Kit. Strand specificity was performed with 2′-deoxyuridine 5′-triphosphate during second-strand synthesis.[27] All samples were indexed with Illumina adapters and sequenced with an Illumina HiScanSQ. We subsequently sequenced each cDNA library on an Illumina HiSeq to obtain 30 million 75 bp paired-end reads per individual. We performed RNA sequencing (RNA-seq) for all 17 individuals (all three generations); however, for eQTL association, we only used the 11 children, and for ASE analysis, we used the two parents and 11 children. All RNA-seq data for all 17 individuals are freely available at the Gene Expression Omnibus under accession number GSE56961.

Quantification of Gene Expression, Splicing, and ASE

We used Tophat and Cufflinks to obtain gene-expression levels from RNA-seq. We used Tophat to map RNA reads to the human reference genome (UCSC Genome Browser, hg19) and Cufflinks to quantify transcript-expression levels. Gene-expression levels were the sum of transcript-expression levels. Gencode[28] v.12 was used as the input annotation for Cufflinks. We calculated transcript ratios to quantify alternative splicing patterns. Gene-expression and transcript-ratio data for Geuvadis samples were downloaded from the Geuvadis website; we used quantified gene-level reads per kilobase per million both before (for assessing effect sizes) and after (for eQTL mapping) normalization via probabilistic estimation of expression residuals.[29] We assessed ASE by counting RNA read depth at heterozygous sites. We performed multiple quality-control steps to reduce known technical artifacts (see Figure S2). We obtained read counts at each heterozygous site by using SAMtools[30] mpileup and our own ASE pipeline based on a binomial test modified for reference-mapping bias with a filter for observing at least five reads for each allele and a minimum read depth of 20× per site.[21,31] To assess the quality of ASE estimates, we compared ASE correlation between double-IBD (identical-by-descent) siblings, half-IBD siblings, and non-IBD siblings. Indeed, we observed an expected increase in correlation between degree of IBD and allelic ratio measured across all sites (Figures S28–S30).

Whole-Genome Sequencing Data

Whole-genome sequencing data for the family were downloaded from the Complete Genomics website. Family members were originally sequenced to an average genome-wide coverage of 80×. We used variants called by the Complete Genomics Analysis Pipeline (v.2.0.0). We performed an additional filtering step testing for Mendelian inconsistency to obtain a high-confidence set of variants, and we eventually retained 5,546,682 out of the original 6,181,281 SNPs. We further compared our selected variants to those assessed by long-fragment read (LFR) technology (N50s 400–1,500 kb).[32] LFR has a claimed error rate of 1 in 10 Mb. Our comparison showed that variant concordance between the 80× shotgun-sequencing approach and LFR technology was 99.91% to 99.95% (Table S2). In addition, the same family was also sequenced to 50× by Illumina Platinum Genomes, and the genotyping concordance with Complete Genomics was found to be 99.62% to 99.83% (Table S4).

Haplotyping and Verification by Long-Fragment Sequencing

We inferred recombination positions and haplotypes of the family by using our software tool Ped-IBD.[33] Haplotype blocks are defined by recombination positions. We identified a total of 813 recombination positions over 22 chromosomes. Haplotype blocks range in size from 0.02 to 12 Mb (90% interval) and have a median length of 1.65 Mb. We further confirmed haplotyping results with molecular haplotypes from the LFR technology in three individuals (NA12877, NA12885, and NA12886; Table S3). The comparison showed that phasing was 99.84% to 99.92% concordant between inferred and molecular haplotypes.

Linkage Mapping of *cis*-eQTLs in the Family

We used linear regression to evaluate correlation of gene-expression levels within local haplotype blocks. We measured effect size by using the regression slope, β, and the coefficient of determination, R^2. The linear model we used considers additive effects of two haplotype blocks. More specifically, for each block, the two parental haplotypes of each child are encoded with two covariates, p and m. The maternal haplotype m_i of child i, for example, is either 0 or 1, depending on which of the two possible maternal alleles is present. Then, an expression trait is regressed as the summation of effects of two parental haplotypes, $T_i \sim \mu + \beta_j p_i + \beta_k m_i$, where T_i is the trait of individual i, the effects of two parental alleles k and j are expressed by β_j and β_k, and μ is the intercept. Each sibling has two choices of parental haplotypes on each side—$p,m \in \{0,1\}$—to yield four total combinations. Gene expression T_i uses $\log_2$(FPKM [fragments per kilobase per million] values). For splicing quantifications, we used relative transcript abundances, which we calculated by dividing the FPKM of each

isoform by the FPKM of the whole gene (see Table S5). For *cis*-eQTLs, we only tested the local haplotypes containing the genes, which is sufficient for including most *cis*-eQTL signals (Figures S3–S5). Furthermore, we confirmed gene-expression levels and eQTL effect sizes with existing microarray data on the same family (Figures S6 and S7).

Comparison of *cis*-eQTL Effect Sizes between Population and Family

To compare *cis*-eQTL effect sizes, β, between the population and family, we sought to first correct for the overestimation of effect sizes (such discoveries exhibit characteristic regression to the mean). To address this in the population eQTLs, we divided the European-descended Geuvadis samples (n = 373) in half and partitioned them into discovery (n = 180) and replication (n = 193) panels. Within the discovery panel, we identified the strongest *cis*-associated variant per gene (by p value and within the same interval tested in the family). This allowed us to use the replication panel to more accurately measure the effect size of each *cis*-eQTL variant. However, to account for the difference in sample sizes between the replication panel (n = 193) and the family (n = 11), we further sought to estimate how much variance in effect-size measurements (β) could be obtained from sampling 11 people in the population at random. In this way we controlled for chance observations of larger effect sizes for some genes in the family. To achieve this, we repeatedly subsampled (100 times) 11 individuals from the replication panel while maintaining the exact same genotypes of the best associated variant between the subsample and the family. Figure S8 illustrates this subsample scheme. Effect sizes were then measured with the same regression formula, $T_i \sim \mu + \beta_p p + \beta_k m$, for both the family and the subsample; note that two regressors, p and m, match segregating patterns of both the haplotypes of the family and the best SNP of the population subsample. We note that estimation of β in the population was highly correlated independently of the use of a one- or two-regressor model (Figure S9). This allowed us to create a distribution of measured effect sizes that would be expected from randomly measuring the same number of individuals and genotypes in both the family and the population. Using this approach, we identified empirical p values representing how often measured effect sizes in the family were greater than that of the best associated SNP in the population. We also repeated this analysis by using fit (R^2) given that we observed differences in the distribution of raw β values between the family and population and also observed higher variance in gene expression in Geuvadis overall (see Figure S17).

We analyzed several features that could result in over- or underestimation of effect-size measurements between the family and population (see Figures S17–S19). First, because effect-size measurements can be influenced by differences in quantification pipelines, we repeated the experiment by using different quantification approaches (Tophat + Cufflinks and GEM[34] + Flux Capacitor;[31] Figures S13 and S14). Second, effect sizes in the population could potentially be underestimated if the best associated SNP in the discovery panel is not causal given that subsequent effect-size measurements, in the replication panel, might not accurately measure the largest effect. To address this, we examined different discovery-panel sizes (Table S6 and Figure S15) and different criteria (Figure S16) for selecting the best SNP from the population. In addition, we observed through permutation that levels of noise in measurements of effect size (β) were different between the family and the population (Figure S17). To better gauge confidence intervals (CIs) of family effect sizes, we estimated the degree of inflation through permutation and adjusted effect-size CIs by scaling. These adjusted CIs were only applied to comparisons of β values and are denoted by $CI_{adjusted}$ (see Figure S17–S19). For the main manuscript, we report only unadjusted CIs. Furthermore, without using subsampling or permutation, we also directly compared effect sizes with Welch's t test by applying analytic estimation of SEs of β. As a correctness check of the subsampling method, we compared and verified that analytic p values by Welch's t test and empirical p values by subsampling were concordant (Figure S19).

We applied the same subsampling method to identify large-effect splicing quantitative trait loci (sQTLs) and ASE. To compare ASE between the family and population, we focused on a subset of genes that had substantial data for the measurement and comparison of allelic ratios (n = 1,777 genes). For a gene to be included, allelic ratios at a single site had to be measurable for at least five siblings and at least 30 population samples. We tested each gene once and excluded genes that were not tested for eQTLs, such as pseudogenes or genes within high-complexity regions (human leukocyte antigen and immunoglobulin loci). For a site to be considered measurable, it needed to be covered by a minimum of 20 reads with at least five reads for each allele. We then took the maximum allelic ratio in the family and compared it with the maximum allelic ratio found in 1,000 subsamples of the Geuvadis; each subsample was matched to the number of heterozygous individuals found in the family for that site. This approach generated an empirical p value that we used to assess whether an ASE effect in the family was greater than that in the population. To account for ASE biases caused by differing read depths between the family and population, we downsampled (hypergeometric) Geuvadis reads by a factor of 1.97—we calculated this scaling factor by measuring the average level of read-depth differences between Geuvadis and family samples at those selected heterozygous sites for each gene. To exclude the possibility that large-effect ASE was due to technical artifacts such as mapping biases or sequencing errors, we also looked at ASE for the second-largest-effect siblings and IBD siblings (Figure S25).

Variant Annotation

We obtained annotations (missense, synonymous, regulatory, and splice region) by using the Variant Effect Predictor tool,[35] which queries annotation from the Ensembl website. ENCODE transcription factor (TF) binding and DNase I hypersensitivity peaks were obtained from RegulomeDB.[24] Conservation scores obtained from PhyloP[36] (phyloP100way) software were downloaded from the UCSC Genome Browser. Motif-disrupting sites were downloaded from HaploReg (v.2).[37] Variant allele frequency was based on phase 1 of the 1000 Genomes Project[38] as calculated across European populations.

Conservation and Network Annotation

We examined the conservation of family eQTL genes between humans and chimpanzees (*Pan Troglodytes*) by using the dN/dS ratios; dN measures the rate of amino acid substitutions, and dS measures the background rate of neutral DNA subsitutions.[39] The dN and dS values were obtained from BioMart[40] (Ensembl v.70), and the dN/dS ratios were computed. dN/dS is negatively correlated with the conservation status of a gene, so higher dN/dS ratios indicate

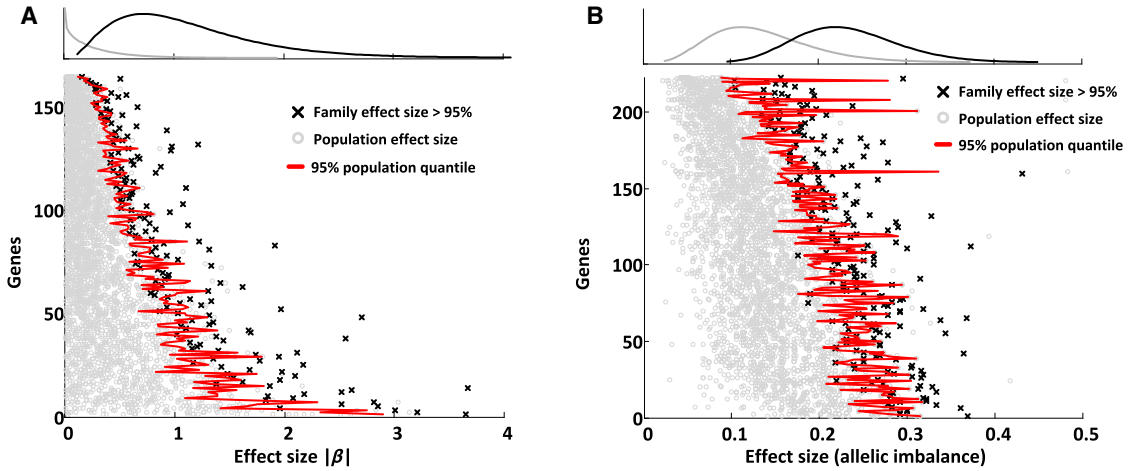

Figure 1. Large-Effect eQTLs and ASE in the Family
(A) Large-effect *cis*-eQTLs. Effect sizes are shown as β, the regression slope. The distribution of family effect sizes (black) is compared to the distribution of population effect sizes (gray). We show *cis*-eQTL genes for which family effect sizes are greater than 95% of population effect sizes. Here, we only plot the distribution of paternal effect sizes (maternal effects have a very similar distribution).
(B) Large-effect ASE genes. ASE effect sizes were assessed by allelic imbalance (0 is balanced, and 0.5 is monoallelic expression). We picked the maximum ASE effect out of 11 siblings and compared it to the maximum ASE effect out of the subsampled population. Plotted are family ASE effects greater than 95% of population ASE effects. To exclude outlier effects, we further tested this for the second-strongest ASE effect in the siblings (Figure S22).

lower conservation of a gene. We also compared centrality of eQTL genes by using the protein-protein interaction (PPI) network as another indication of the biological importance of the affected genes.[19] We computed connectivity of family and population eQTL genes in the PPI network. The PPI network was integrated from BioGRID,[41] the Molecular Interaction database,[42] the Human Protein Reference Database,[43] and IntAct,[44] all data obtained from the GeneMANIA[45] data repository (downloaded on January 4, 2012).

Rare-Variant Enrichment Analyses

To control for site discovery and genotyping differences between the population (1000 Genomes Project) and family (Complete Genomics) genomes, we performed enrichment analyses only for variants in the family genomes. Using these data, we calculated enrichment of rare variants at large-effect-size genes by dividing the proportion of large-effect-size genes with a rare variant by the proportion of all tested genes with a rare variant.

Results

We set out to develop an improved understanding of the impact and interpretability of rare noncoding variants. Our approach involved combining high-quality genomes and transcriptomes within a single large family to identify *cis*-eQTLs and compare these to *cis*-eQTLs discovered in a large population sample. Through the use of RNA-seq data, we were also able to conduct comparable analyses for alternative splicing and ASE. Our analyses focused on the enrichment of rare and potentially regulatory variants in large-effect eQTLs and sQTLs in the family, and we sought to identify the properties of genes that exhibit such effects. Furthermore, we investigated the degree to which family transcriptome data enable the detection of noncoding annotation relevant to interpreting rare noncoding variants genome-wide.

Family Transcriptome Sequencing Identifies Large-Effect *cis*-eQTLs

We hypothesized that rare variants acting either alone or in combination with common variants can cause an eQTL to exhibit a larger effect size in the family than in the population. To identify such cases, we applied a ranking scheme in which we compared gene-expression *cis*-eQTLs between the family and the population to find genes that exhibited larger effect sizes within the family (see Material and Methods). At CI > 0.95 (or empirical p value < 0.05), we found that 319 (including both paternal and maternal β measurements) of the 7,341 genes we tested had effect sizes exceeding that of the best population *cis*-eQTL SNP (false-discovery rate [FDR] $= 7{,}341 \times 0.05 \times 2 / 319 > 1$; Figure 1A). Using comparisons of β, we did not find more relatively large-effect eQTLs than we would expect by chance; however, we identified that this FDR is likely overconservative primarily because of differences in noise between the family and population (see Figures S17–S19), and we therefore also discuss less conservative estimates of FDR (see Figures S17–S19). It is important to note that FDR here measures whether there are more large effects in the family than in the population; however, ranking relative effect sizes by empirical p values is biologically meaningful whether there is an excess or a depletion. Such relative effects overlap (to a degree) genes measured only by absolute effect size in the family; for instance, when comparing genes at the 95% percentile for absolute β versus relative β, we observed an overlap of 52% (Figure S12). However, we chose to use in all subsequent

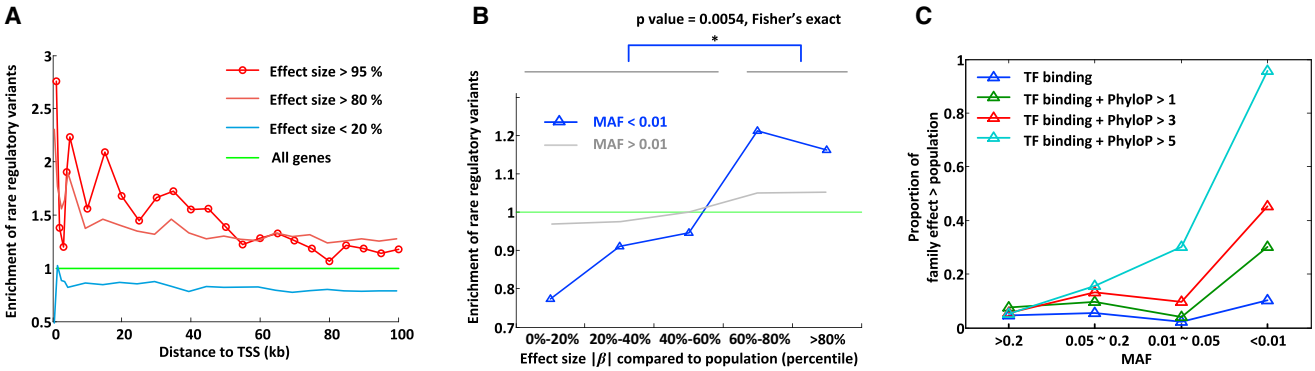

Figure 2. Enrichment of Rare Variants in Large-Effect eQTLs

(A) Enrichment of rare and potentially regulatory variants near the TSS of large-effect (β) *cis*-eQTL genes. Variants are restricted to those with a MAF < 0.01, within ENCODE TF binding and DNase I hypersensitivity peaks, and with a PhyloP score > 1. We observed increased enrichment of rare regulatory variants near the TSS of larger-effect-size genes in the family.

(B) Enrichment of potentially regulatory variants depends on allele frequency and relative effect sizes. We ranked genes (x axis) on the basis of how often their effect sizes in the family were greater than their effect sizes in the population subsamples, which is also 1 − their empirical p values (see Material and Methods). Variants are restricted to those within <100 kb of the TSS, within ENCODE TF binding and DNase I hypersensitivity peaks, and with a PhyloP score > 1. We observed that variant enrichment was dependent on whether the variant was rare (blue) or not (gray). We calculated enrichment by dividing the proportion of genes with such an annotated rare variant in each effect-size bin by the proportion of genes with an annotated rare variant across all effect-size bins.

(C) Conservation scores and allele frequency predict genes with a larger effect in the family than in the population. We restricted to variants within <100 kb of the TSS, within ENCODE TF binding and DNase I hypersensitivity peaks, and with different PhyloP thresholds. Proportions were computed by π_1 statistics on permutation-based p values of family effect larger than population effect. We observed that rare and highly conserved variants overlapping epigenomic data (light blue) were highly predictive of a larger effect in the family than in the population.

analyses the ranking of genes according to their relative effect sizes instead of absolute effect sizes because we hypothesized that the former might better inform family-specific effects. By instead measuring fit (R^2), we identified 577 *cis*-eQTLs that had a better fit in the family than the best population-level *cis*-eQTL variant (CI > 0.95; FDR $= 7{,}341 \times 0.05 / 577 = 63\%$; Figure S10). Among those genes that exhibited the largest effect sizes and fits in the family (both at a CI > 0.95), there was a significant overlap of 36.4% (Figure S11). To exclude the possibility of technical factors underlying effect-size differences, we repeated the analysis by using different quantification pipelines (Figures S13 and S14), population discovery-panel sizes (Table S6), and alternative methods for choosing the best SNP (Figure S16); we observed no significant difference in the discovery set of large-effect genes or on further downstream analyses (see Material and Methods).

We also identified genes that exhibited larger ASE effects in the family than in the population. We found that 223 of the 1,777 genes we tested had larger ASE effect sizes in the family (CI > 0.95, FDR $= 1{,}777 \times 0.05 / 223 = 40\%$; Figure 1B; Figure S25). We expected that on an individual basis, the family and population would actually have the same distribution of ASE effect sizes (no excess of large effects, FDR $= 1$). We controlled for some initially observed excess in the family by matching read depths via downsampling; however, this did not address all the excess in the family, and unknown factors still remained. We expected that any excess, however, would only add noise to subsequent rare-variant enrichment analyses, and we further validated large ASE effects by using evidence from IBD siblings (Figure S25). In addition, we applied ASE to support discoveries of *cis*-eQTLs in the family; by stratifying their degree of effect size relative to those in the population, we detected a proportionally increased enrichment of detectable ASE (significant ASE sites defined as allelic imbalance > 0.05, binomial test p value < 0.05; Figure S21). This relationship supports a potential regulatory role of rare variants because it indicates that large-effect *cis*-eQTLs in the family might be the consequence of heterozygous variants that manifest in ASE. This idea is further supported by our observation of a direct and simple linear relationship between *cis*-eQTL effect size among children and ASE effect size among parents (Figure S20).

Large-Effect *cis*-eQTLs in the Family Are Enriched with Rare Variants

We hypothesized that rare noncoding variants might be responsible for a considerable proportion of the large-effect-size *cis*-eQTLs in the family. Taking advantage of full genome data in the family, we assessed the enrichment of rare and potentially regulatory variants near the transcription start site (TSS) of genes with different magnitudes of relative effect sizes between the family and the population. Here, we used PhyloP to define potentially regulatory variants on the basis of ENCODE TF peaks, DNase I hypersensitivity peaks, and evolutionarily constrained regions across 99 vertebrate genomes; we will later further explore the relative importance of each of these annotations. We observed enrichment of rare and potentially regulatory noncoding variants in genes that had the largest effect sizes (CI > 0.95 and CI > 0.80; Figure 2A). This relationship

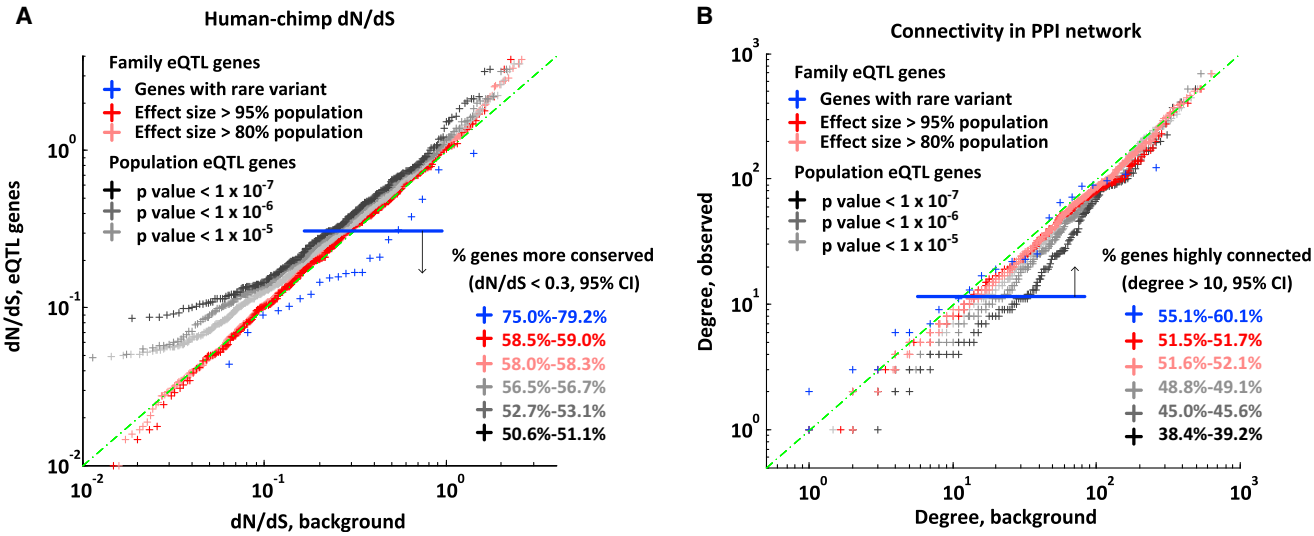

Figure 3. Large-Effect eQTLs Influence Essential Genes

(A) dN/dS ratio comparing large-effect family *cis*-eQTLs to population *cis*-eQTLs. We selected family eQTLs on the basis of their effect sizes relative to population eQTL effect sizes and plotted the distributions of dN/dS ratios. As a comparison, we show the distribution of dN/dS ratios for the most significant *cis*-eQTL genes identified only in the population (373 unrelated European individuals from the Geuvadis study) given different p value cutoffs. This is further compared to family-level genes that have rare and potentially regulatory variants (within 5 kb of the TSS, within ENCODE TF binding and DNase I hypersensitivity peaks, and with a PhyloP score > 1). We observed that for large-effect *cis*-eQTLs and family-level genes with a rare variant, a higher proportion were more conserved (described as the percentage of genes with a dN/dS < 0.3; lower dN/dS ratios indicate higher conservation).

(B) Comparison of centrality in the PPI network between large-effect *cis*-eQTLs in the family and population *cis*-eQTLs. Centrality is measured by the number of interacting proteins (degrees). Different groups of genes are defined in the same way as in (A). We show proportions of high-connectivity (hub) genes (degree > 10; higher degrees indicate more essential genes) among these groups. We observed that the proportion of high-connectivity genes was greatest for large-effect *cis*-eQTLs and family-level genes with a rare variant. This suggests that common regulatory variants are less likely to occur at conserved genes. In contrast, family-specific eQTL effects, because they arise from rare variants, can affect conserved genes.

was most pronounced within the first 5 kb close to the TSS and decayed as a function of distance. It was also related to the degree to which the family effect was larger than that detected in the population across the full distribution of measured effects (Figure 2B). Likewise, we tested both large-effect *cis*-eQTLs by fit (R^2) and large-effect ASE genes and observed similar strong enrichment of rare and potentially regulatory variants (Figures S23 and S25C).

We also evaluated the utility of known regulatory annotations in predicting eQTLs for rare variants. Comparing annotated rare variants with all rare variants, we observed strong enrichment (up to a 2-fold increase near the TSS) of annotated variants, indicating that annotation is highly informative in predicting eQTLs (Figure S22). Furthermore, we observed that the enrichment was higher in family-based eQTLs than in population eQTLs as a function of effect size (Figure S24). To test the contribution of different annotations to a large effect in the family, we further stratified by MAF and strength of evolutionary constraint. We observed that variants with lower MAF and with increasing degree of evolutionary constraint were the most informative factors indicative of large *cis*-eQTL effects in the family (Figure 2C).

Large-Effect *cis*-eQTLs in the Family Influence Essential Genes

It has been previously reported that *cis*-eQTLs based on population studies are depleted among essential genes.[19] We hypothesized that if rare variation was indeed responsible for large-effect *cis*-eQTLs in the family, reduced impact of purifying selection on rare variants would result in family eQTLs disproportionately affecting essential genes. We tested this hypothesis in two ways: defining gene essentiality by (1) its degree of evolutionary constraint and (2) its centrality within a PPI network. To assess evolutionary constraint, we used dN/dS ratios between humans and chimps to compare large-effect *cis*-eQTL genes in the family to *cis*-eQTL genes in the population. We observed that large-effect *cis*-eQTL genes in the family had significantly higher conservation status than population *cis*-eQTL genes (Figure 3A). This was even more pronounced for genes with a rare and potentially regulatory variant within 5 kb of the TSS. By contrast, *cis*-eQTL genes in the population were less constrained for increasingly stringent p values.

We next applied PPI networks with the premise that genes that are more central in the network or have more connections to other genes are more essential than less connected ones. We found significantly higher connectivity for large-effect *cis*-eQTL genes in the family than for *cis*-eQTL genes in the population (Figure 3B). Furthermore, this contrast became stronger when we focused only on those genes that also contained a proximal rare and potentially regulatory variant (Figure 3B).

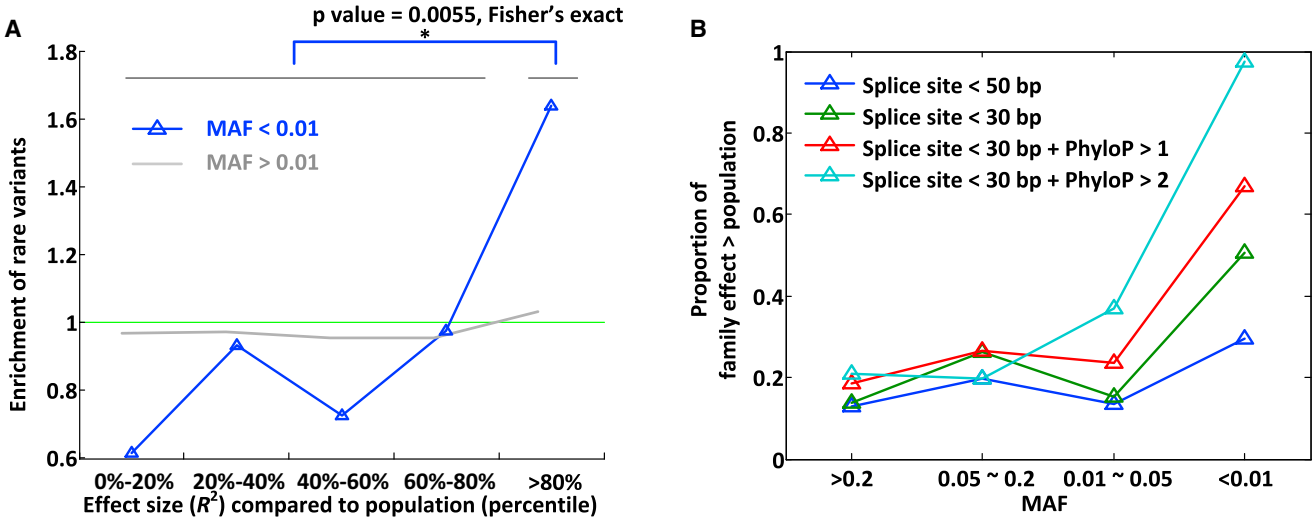

Figure 4. Large-Effect sQTLs in the Family

(A) Enrichment of rare variants at large-effect sQTL genes. We ranked genes (x axis) on the basis of how often their effect sizes in the family were greater than their effect sizes in the population subsamples (see Material and Methods). We restricted to variants within 30 bp of splice sites and with a PhyloP score > 1. As for *cis*-eQTLs (in Figure 2B), we observed that enrichment was dependent on allele frequency. We calculated enrichment by dividing the proportion of genes with such an annotated rare variant in each effect-size bin by the proportion of genes with an annotated rare variant across all effect-size bins.

(B) Conservation scores, the distance to splice site, and allele frequency predict genes with a larger effect in the family than in the population. We observed that rare and conserved variants near splice sites (light blue) were highly predictive of a larger splicing effect in the family than in the population.

Family Transcriptome Sequencing Identifies Large-Effect sQTLs

By comparing *cis*-sQTLs between the family and the population, we also ranked genes with larger relative effect sizes (measured as R^2) in the family than in the population (n = 726, >95% population, n = 5,622 genes; FDR = 39%). Differences in isoform-quantification pipelines probably overestimate the excess number of large-effect sQTLs because there is also more noise in isoform quantification in the population. However, as for large-effect eQTLs, we also observed enrichment of rare and potentially functional variants for large-effect sQTL genes in the family (Figure 4A). Furthermore, by stratifying on allele frequency, distance to splice sites, and evolutionary-constraint thresholds, we found that large-effect sQTLs in the family were much better predicted by rare variants than by common variants, especially for conserved regions near splice sites (Figure 4B). In addition to observing large effect sizes, we also found that sQTLs could exhibit very high heritabilities, nearly as high as those for Mendelian traits (examples in Figures S26 and S27).

Large-Effect *cis*-eQTLs in the Family Might Further Modify Complex-Disease-Associated Genes

There has been considerable interest in whether rare variants modify risk of complex disease.[46,47] Although we were unable to directly test disease associations within this family because of the anonymity of the individuals, we sought to quantify the number of genome-wide association study (GWAS) genes in which *cis*-eQTLs exhibited larger effects in the family than in the population. We identified 315 GWAS genes in which the known GWAS variant was an eQTL in the population (at an FDR of 5%), suggesting a regulatory basis to disease pathogenesis. Of these genes, we identified 65 with a larger-effect *cis*-eQTL in the family (>80th percentile). Of those, 17 (Table S9) were not polymorphic for the known GWAS SNP in the family, and two had a rare and potentially regulatory variant (within <100 kb of the TSS, within an ENCODE TF binding and DNase I hypersensitivity peak, and with a PhyloP score > 0) influencing genes implicated in body mass index, hypertension, and obesity. In addition, regardless of relative effect sizes of eQTLs between the family and population, we identified four GWAS genes (Table S10) in which the known GWAS SNP was an eQTL in the population and that had a rare and potentially regulatory variant (within <100 kb of the TSS, within an ENCODE TF binding and DNase I hypersensitivity peak, and with a PhyloP score > 3) in the family according to strong predictor variables. Although increased risk in this family is not known, the presence of rare and potentially regulatory variants in complex-disease-associated genes whose expression is implicated in disease pathogenesis suggests that complex traits and genes should be further studied with rare-variant association tests.

Functional Noncoding Annotations Are Informative of the Impact of Rare Noncoding Variants

Genome and transcriptome data from a single large family allowed us to test the utility of various noncoding annotations for predicting the impact of noncoding variants

on expression. Here, our goal was to identify those annotations that could inform a functional variant from genome sequence alone. We chose to include the following as potentially informative annotations: ENCODE TF binding, DNase I hypersensitivity peaks, evolutionary constraint, motif disruption as computed by HaploReg, and distance to the TSS. We identified that each noncoding annotation was more informative for predicting the impact of rare variants than the impact of common variants on expression (Figure 5A; Table S7). We observed that evolutionary constraint and distance to the TSS were the most informative for rare variants, and they further increased their utility with increasing strength of constraint and shorter distances, respectively. One potential concern we identified is that we might be only predicting a gene's ability to harbor an eQTL such that having a rare variant possessing specific annotation might indirectly inform genes tolerant of arbitrary functional variants (both common and rare). However, when assessing whether genes containing different annotations for rare variants were also more likely to have common eQTLs in the population, we saw no significant difference (Figure 5A, right panel). This demonstrates that particular species of rare noncoding variants might be interpretable from genome sequence data alone provided that there is sufficiently high-confidence genotyping of those rare variants. Furthermore, provided increasing availability of genome-interpretation methods, this method offers a means of determining and calibrating the efficacy of different approaches.

Through finer stratification of allele frequency, we were able to observe the degree to which genome annotation influenced predictions of *cis*-eQTLs. We observed that predictions of eQTLs were most informative for potentially regulatory variants when those variants were rare (Figure 5B). This was also the case for sQTLs: predictor variables such as evolutionary constraint and distance to splice sites were the most informative factors for predicting a sQTL when a variant was rare (Figure 5C).

Discussion

Our study combined high-quality genome sequencing and RNA-seq data for a 17 member, three-generation family, enabling us to investigate the role and interpretability of rare noncoding variants. In contrast to low-pass approaches, high-quality full-coverage genome sequencing and patterns of Mendelian segregation provided the ability to more confidently identify and genotype rare variants within the family. More importantly, the large number of children provided us with the ability to detect eQTLs caused by rare variants specific to the family. In contrast, the power of a design that includes many small families or trios would be reduced by the overall heterogeneity of causal rare variants in each family. A further advantage is that with matched cell type and population, we were able to compare family eQTLs to population eQTLs reported by the Geuvadis Consortium.[18] We identified genes that exhibit larger eQTL effect sizes in the family than in the population and demonstrated that these family-specific eQTLs are enriched with rare regulatory variants, influence more evolutionarily constrained and central genes, and are potential contributors to risk of complex disease.

One limitation of the study is that we did not observe many more large-effect eQTLs in the family than expected by chance; high FDRs were observed for all categories of large-effect eQTLs. This could suggest that there is not an overabundance of large-effect eQTLs specific to the family. It might also simply reflect limited power or imperfect comparison of effect sizes between cohorts, given that we explored by varying quantification pipelines, discovery-panel sizes, and methodologies for selecting testable variants. However, the enriched properties we identified for large-effect family eQTLs appear to be robust to such limitations, and we highlight that although there might not be a strong excess of large-effect eQTLs, the relative degree of effect between the family and population conveys meaningful properties of family eQTLs. For instance, as the degree of effect size increased in the family relative to the population, we observed an increasing enrichment of rare and potentially regulatory variants. Furthermore, such large-effect eQTLs in the family exhibited increasing enrichment in ASE, implicating a heterozygous causal variant. Additionally, the enrichment of family eQTLs among constrained and central genes was most extreme for the subset of genes in which a rare and potentially regulatory variant could be identified. These observations fit with population-genetic expectation given that rare variants can influence more essential genes because of a reduced impact of purifying selection. Furthermore, this is in contrast to the general properties of population eQTL genes; for increasing effect sizes, they have previously been shown to be less constrained and less central.[19] Taken together, these results implicate an important role of rare regulatory variants in large-effect eQTLs in the family.

We compared, in addition to gene expression, ASE and alternative splicing between the family and the population. As with gene expression, we observed enrichment in rare variants for large-effect ASE and sQTLs in the family. Furthermore, we observed that evolutionary constraint and distance to splice sites for rare splicing variants was significantly informative of large splicing effects in the family. With both large-effect eQTLs and large-effect sQTLs predicted by rare variants, this study highlights existing potential for routine integration of these variants in rare-variant association tests.

Ultimately, a principal goal in genome interpretation is to develop the ability to predict the impact of all variants, including those that are rare or novel. In our study, we were able to test the importance of diverse noncoding annotations for predicting the impact of noncoding

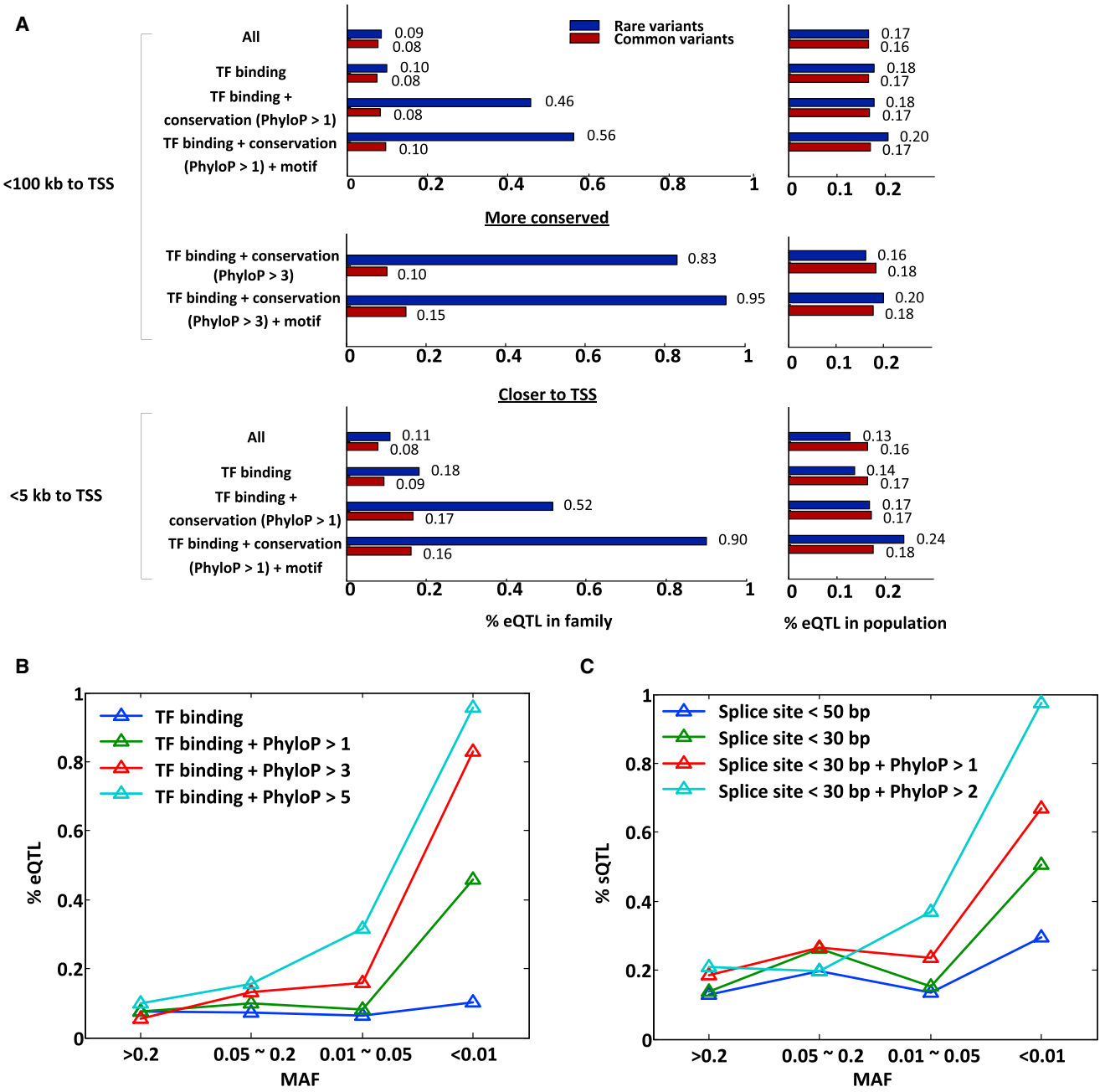

Figure 5. Predicting Rare and Common eQTLs

(A) Utility of diverse noncoding annotation for predicting rare and common eQTLs. We considered the enrichment in eQTLs (measured with the π_1 statistic) for rare (MAF < 0.01) and common (MAF > 0.01) variants overlapping the following different functional annotations: ENCODE TF binding and DNase I hypersensitivity peaks, distance to TSS, PhyloP conservation scores, and motif disruption (score change > 10); annotations were added one at a time. We found that these functional annotations were significantly more powerful for detecting an eQTL when intersecting rare variants rather than common variants. Furthermore, on the right, we demonstrate that none of the genes possessing rare variants overlapping the different categories of annotation were disproportionally enriched in their ability to also be eQTLs in the population. A full matrix summarizing intersections of these annotations is provided in Table S7.

(B) Conservation scores and allele frequency predict genes with an eQTL. We restricted to variants within 100 kb of the TSS, within ENCODE TF binding and DNase I hypersensitivity peaks, and with different PhyloP scores and allele frequencies to assess each variant class's enrichment in eQTLs. We observed that highly conserved and rare variants were strongly predictive of an eQTL.

(C) Conservation scores, the distance to splice site, and allele frequency predict genes with a sQTL. We considered different thresholds on distance to splice sites, PhyloP conservation scores, and allele frequencies. We observed that rare and conserved variants near splice sites (light blue) were highly predictive of a sQTL.

variants on gene expression. For rare variants, we identified that evolutionary constraint coupled with distance to the TSS and epigenomic information was highly informative in predicting eQTLs. For common variants, such annotations did not provide comparable predictive power. The likely reason for this difference is that

common variants, regardless of genomic annotation, are very likely to be neutral, whereas rare variants have a higher prior likelihood of functional impact that can be further informed by genomic annotation. Given that no previous analyses have had access to high-quality genomes and transcriptomes in a single large human family, this study provides data to support a much-needed framework for frequency-independent evaluation of genome interpretation for noncoding variants and suggests that the impact of many rare and causal noncoding variants might be easier to predict than expected.

Supplemental Data

Supplemental Data include 30 figures and 10 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2014.08.004.

Acknowledgments

We would like to thank Tomas Babak, Christopher Brown, Hunter Fraser, Arend Sidow, and members of the S.B.M. lab for critical review of this work and manuscript. This work was supported by the Edward Mallinckrodt, Jr. Foundation and the Li Ka Shing Foundation.

Received: May 16, 2014
Accepted: August 12, 2014
Published: September 4, 2014

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes Phase 1 Analysis Results, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/
Complete Genomics, 69 Genomes Data, http://www.completegenomics.com/public-data/69-Genomes/
Ensembl Genome Browser, http://www.ensembl.org
Ensembl Variant Effect Predictor, http://www.ensembl.org/info/docs/tools/vep/
GeneMANIA, http://www.genemania.org/
Geuvadis Data Browser, http://www.ebi.ac.uk/Tools/geuvadis-das/
Geuvadis RNA sequencing project, http://www.geuvadis.org/web/geuvadis/RNAseq-project
GWAS catalog, http://www.genome.gov/admin/gwascatalog.txt
HaploReg, http://www.broadinstitute.org/mammals/haploreg
Illumina Platinum Genomes, whole-genome sequencing data, http://www.illumina.com/platinumgenomes/
LFR data for family members, ftp://ftp2.completegenomics.com/
PhyloP conservation scoring, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way/
RegulomeDB, http://regulomedb.org/
UCSC Genome Browser, http://genome.ucsc.edu

Accession Numbers

The Gene Expression Omnibus accession number for the RNA-seq data of all 17 individuals reported in this paper is GSE56961.

References

1. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.
2. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
3. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.
4. Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. *1*, 131.
5. Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., et al.; 1000 Genomes Project (2011). The functional spectrum of low-frequency coding variation. Genome Biol. *12*, R84.
6. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
7. Flannick, J., Thorleifsson, G., Beer, N.L., Jacobs, S.B., Grarup, N., Burtt, N.P., Mahajan, A., Fuchsberger, C., Atzmon, G., Benediktsson, R., et al.; Go-T2D Consortium; T2D-GENES Consortium (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat. Genet. *46*, 357–363.
8. Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. *7*, e1002198.
9. Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J.F., de Rijk, P., Dewit, O., et al. (2011). Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat. Genet. *43*, 43–47.
10. Raychaudhuri, S., Iartchouk, O., Chin, K., Tan, P.L., Tai, A.K., Ripke, S., Gowrisankar, S., Vemuri, S., Montgomery, K., Yu, Y., et al. (2011). A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nat. Genet. *43*, 1232–1236.
11. Panoutsopoulou, K., Tachmazidou, I., and Zeggini, E. (2013). In search of low-frequency and rare variants affecting complex traits. Hum. Mol. Genet. *22* (R1), R16–R21.
12. Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. Nat. Genet. *46*, 220–224.
13. Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O., et al. (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature *498*, 232–235.

14. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.
15. MacArthur, D.G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. Hum. Mol. Genet. *19* (R2), R125–R130.
16. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
17. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
18. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
19. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24.
20. Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T.J., Sladek, R., and Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. Nat. Genet. *40*, 225–231.
21. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E.T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. *7*, e1002144.
22. Gaffney, D.J., Veyrieras, J.B., Degner, J.F., Pique-Regi, R., Pai, A.A., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. *13*, R7.
23. Lupski, J.R., Belmont, J.W., Boerwinkle, E., and Gibbs, R.A. (2011). Clan genomics and the complex architecture of human disease. Cell *147*, 32–43.
24. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.
25. Ritchie, G.R., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.
26. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
27. Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. *37*, e123.
28. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.
29. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
30. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
31. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.
32. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., et al. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190–195.
33. Li, X., Yin, X., and Li, J. (2010). Efficient identification of identical-by-descent status in pedigrees with many untyped individuals. Bioinformatics *26*, i191–i198.
34. Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat. Methods *9*, 1185–1188.
35. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.
36. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
37. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40* (Database issue), D930–D934.
38. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
39. Hurst, L.D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. *18*, 486.
40. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2012). Ensembl 2012. Nucleic Acids Res. *40* (Database issue), D84–D90.
41. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34* (Database issue), D535–D539.
42. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. Nucleic Acids Res. *35* (Database issue), D572–D574.
43. Goel, R., Harsha, H.C., Pandey, A., and Prasad, T.S.K. (2012). Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. Mol. Biosyst. *8*, 453–463.

44. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., et al. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res. *38* (Database issue), D525–D531.
45. Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. *38* (Web Server issue), W214–W220.
46. Gibson, G. (2011). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135–145.
47. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.

The American Journal of Human Genetics, Volume *95*

Supplemental Data

Transcriptome Sequencing of a Large Human Family Identifies the Impact of Rare Noncoding Variants

Xin Li, Alexis Battle, Konrad J. Karczewski, Zach Zappala, David A. Knowles, Kevin S. Smith, Kim R. Kukurba, Eric Wu, Noah Simon, and Stephen B. Montgomery

FIGURE S1. FAMILY STRUCTURE. 4
FIGURE S2. FLOWCHART OF GENOTYPE CALLING AND RNA-SEQ QUALITY CONTROL STEPS. 5
FIGURE S3. HAPLOTYPE / IDENTITY-BY-DESCENT (IBD) INFERENCE. 6
FIGURE S4. HISTOGRAM OF HAPLOTYPE LENGTHS. 7
FIGURE S5. DISTRIBUTION CIS-EQTL AND CIS-SQTL VARIANTS NEAR A GENE: LOCAL HAPLOTYPE BLOCKS HAVE THE LARGEST NUMBER OF EQTL (LEFT) OR SQTL (RIGHT) EFFECTS. 8
FIGURE S6. COMPARISON OF EQTL DISCOVERY BETWEEN RNA-SEQ AND MICROARRAY. 9
FIGURE S7. CONCORDANCE OF EQTL EFFECT SIZES (B) BETWEEN RNA-SEQ AND MICROARRAY. 10
FIGURE S8. IDENTIFYING LARGE-EFFECT CIS-EQTL GENES IN FAMILY COMPARED TO POPULATION. 11
FIGURE S9. EFFECT SIZE MEASURED BY THE ONE-REGRESSOR AND THE TWO-REGRESSOR MODEL. 13
FIGURE S10. LARGE-EFFECT FAMILY *CIS*-EQTL GENES. 14
FIGURE S11. OVERLAP OF B AND FIT (R^2) EFFECT SIZE OUTLIERS. 15
FIGURE S12. LARGE RELATIVE B VS. ABSOLUTE B. 16
FIGURE S13. EFFECT OF DIFFERENT QUANTIFICATION PIPELINES: COMPARISONS OF EFFECT SIZE B BETWEEN TOPHAT + CUFFLINKS AND GEM + FLUX PIPELINES. 17
FIGURE S14. ENRICHMENT OF RARE VARIANTS AT LARGE EFFECT SIZE B. 18
FIGURE S15. INFLUENCE OF DISCOVERY SAMPLE SIZES IN TAGGING CAUSAL SNPS. 19
FIGURE S16. INFLUENCE OF DIFFERENT CRITERIA IN SELECTING BEST SNP: SMALLEST *P*-VALUE OR LARGEST EFFECT SIZE. 20
FIGURE S17. ADJUSTMENT OF EFFECT SIZE EMPIRICAL *P*-VALUES: COMPARISON OF EFFECT SIZE CONFIDENCE INTERVALS (NOISE LEVELS) BETWEEN THE FAMILY AND THE POPULATION. 21
FIGURE S18. ADJUSTMENT OF EFFECT SIZE EMPIRICAL *P*-VALUES: DISTRIBUTION OF *P*-VALUES OF FAMILY VERSUS POPULATION EFFECT SIZES. 23
FIGURE S19. ADJUSTMENT OF EFFECT SIZE EMPIRICAL *P*-VALUES: COMPARISONS OF EMPIRICAL *P*-VALUE AND WELCH'S *T*-TEST. 25
FIGURE S20. CORRELATION OF EQTL EFFECT SIZE B AND ASE EFFECT SIZE (ALLELIC IMBALANCE). 26
FIGURE S21. ENRICHMENT OF ASE EFFECTS AT LARGE-EFFECT GENES. 27
FIGURE S22. RARE REGULATORY VARIANTS CONTRIBUTING TO LARGE-EFFECT EQTLS: ENRICHMENT OF RARE VARIANTS NEAR THE TSS OF LARGE-EFFECT (B) CIS-EQTL GENES, COMPARING ANNOTATED AND ALL RARE VARIANTS. 28
FIGURE S23. RARE REGULATORY VARIANTS CONTRIBUTING TO LARGE-EFFECT EQTLS. 29
FIGURE S24. ENRICHMENT OF RARE REGULATORY AT LARGE EFFECT GENES. 30
FIGURE S25. IDENTIFICATION OF LARGE ASE EFFECT. 31
FIGURE S26. MENDELIAN SEGREGATION OF ALTERNATIVE SPLICING PATTERNS. 33
FIGURE S27. EXAMPLES OF ALTERNATIVE SPLICING PATTERNS DETERMINED BY HAPLOTYPE GROUPS. 34
FIGURE S28. ASE HERITABILITY ANALYSIS. 35
FIGURE S29. ALLELIC RATIO CORRELATION WITH SIBLINGS, USING NA12879 AS REFERENCE. 36
FIGURE S30. ALLELIC RATIO CORRELATION BETWEEN DIFFERENT TYPES OF SIBLINGS. 37

TABLE S1. NUMBER OF VARIANTS SEGREGATING IN THE FAMILY. 38
TABLE S2. GENOTYPES CONFIRMED WITH COMPLETE GENOMICS LONG FRAGMENT READ[1] (LFR). 39
TABLE S3. PHASING CONFIRMED WITH MOLECULAR HAPLOTYPE BY LFR. 40
TABLE S4. GENOTYPES CONFIRMED WITH ILLUMINA PLATINUM GENOMES. 41
TABLE S5. LINKAGE ANALYSIS OF *CIS*-EQTL: SUMMARY OF EQTL AND SQTL GENES IDENTIFIED IN THE FAMILY.. 42
TABLE S6. EFFECT OF DIFFERENT DISCOVERY PANEL SIZES: NUMBER OF LARGE EFFECT B GENES GIVEN DIFFERENT DISCOVERY PANEL SIZES. 44
TABLE S7. PREDICTION OF EQTLS AT RARE VARIANTS GIVEN ANNOTATION: PROPORTION OF GENES BEING AN EQTL GIVEN A REGULATORY VARIANT NEAR TSS. 46
TABLE S8. EXAMPLES OF SQTL GENES. 47
TABLE S9. FAMILY-SPECIFIC CIS-EQTL MODIFYING COMPLEX TRAIT GENES. 49
TABLE S10. EXAMPLES OF RARE REGULATORY VARIANTS INFLUENCING GWA GENES. 50

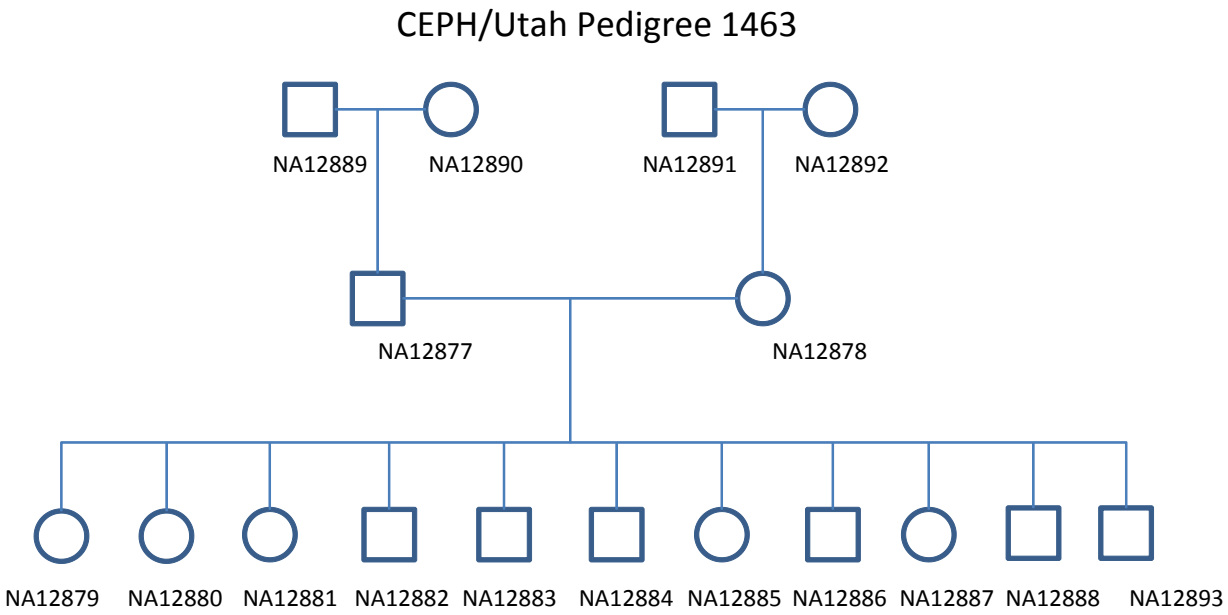

Figure S1. Family structure.

Four grandparents, two parents and eleven children. All family members are RNA-sequenced. Whole genome DNA-sequencing data of all family members were generated by Complete Genomics. Whole genome sequencing was also performed again by both Illumina Platinum Genomes and Complete Genomics Long Fragment Read[1] technology. All three sets of genome sequencing data are compared to confirm genotyping correctness (Table S2, Table S3, Table S4).

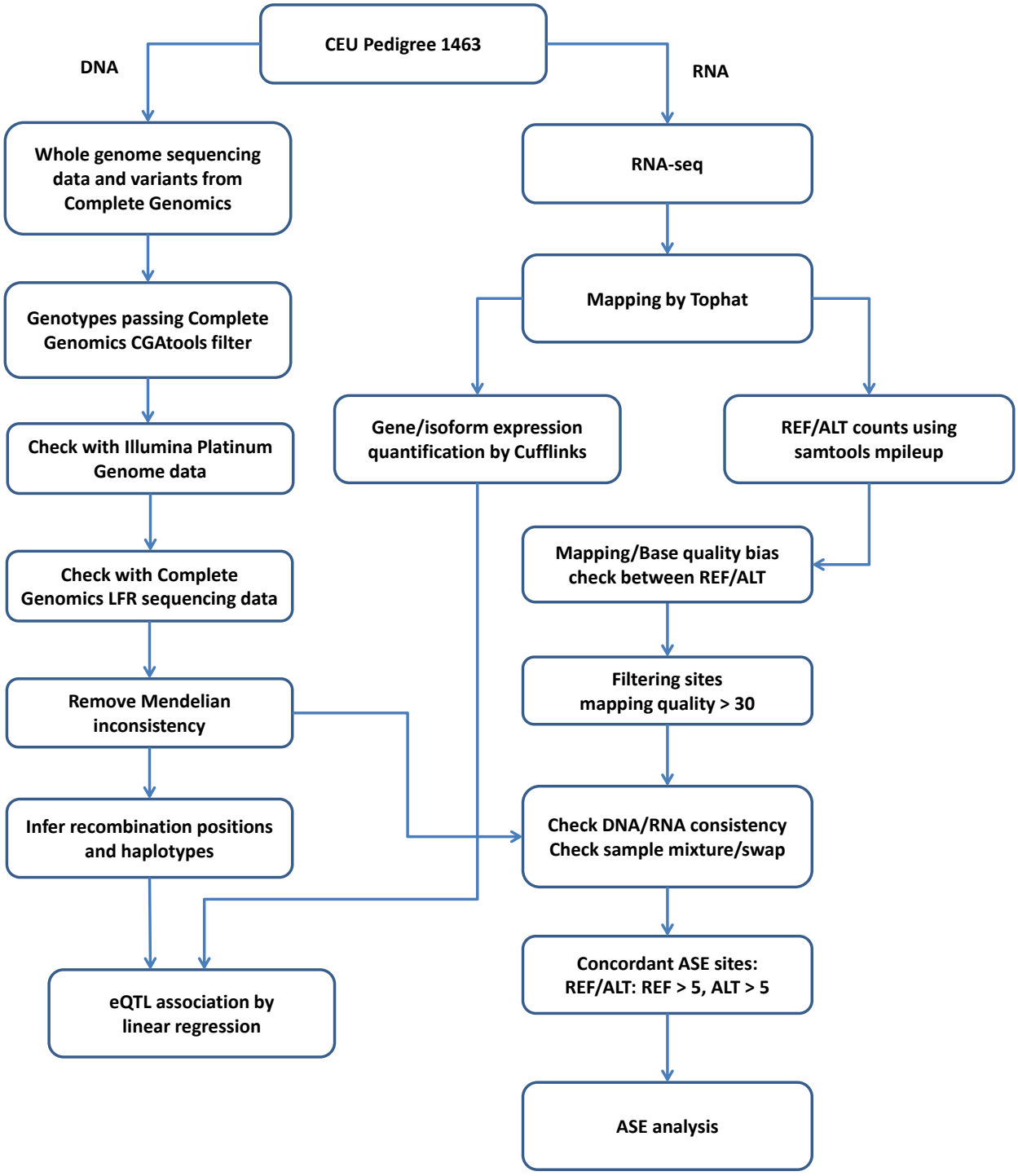

Figure S2. Flowchart of genotype calling and RNA-Seq quality control steps.

Genotyping data were confirmed across three sequencing platforms (Table S2, Table S3, Table S4) to guarantee correctness especially at rare variants. We further filtered variants by stringent Mendelian consistency throughout the whole family. RNA/DNA concordance was checked at heterozygous sites to avoid sample mixture/swap.

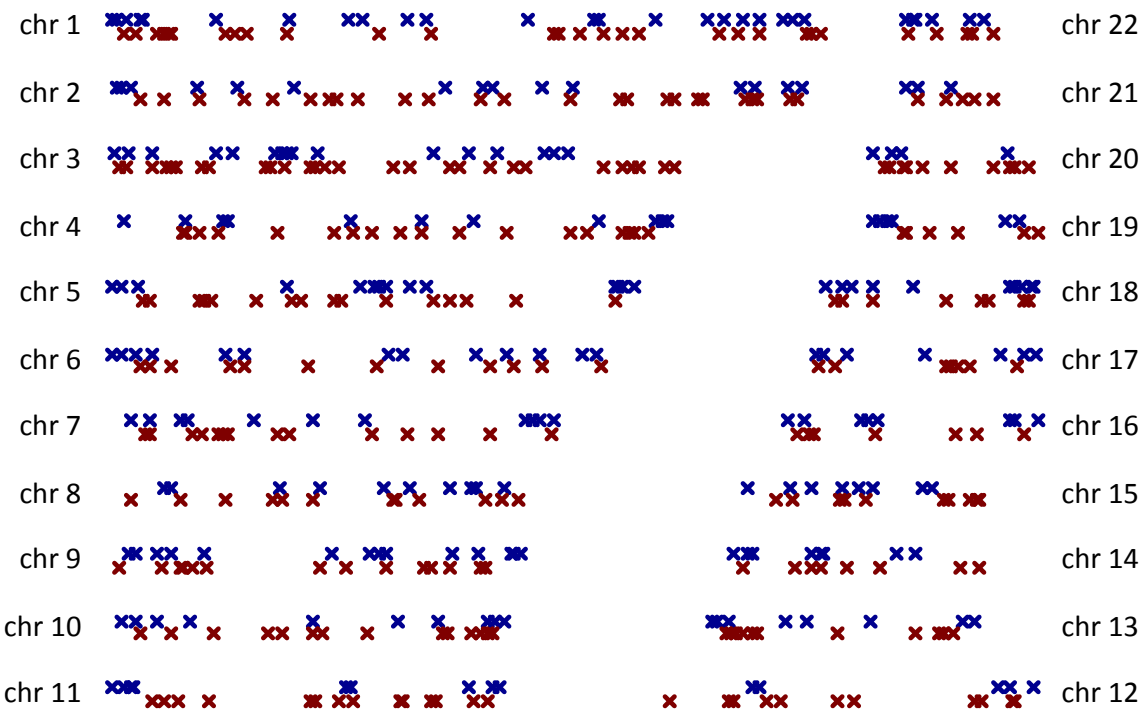

Figure S3. Haplotype / identity-by-descent (IBD) inference.

Distribution of recombination breakpoints. Red: maternal recombinations, Blue: paternal recombinations. We inferred 813 recombination positions in CEU family 1463. We partition chromosomes into haplotypes according to these recombination positions.

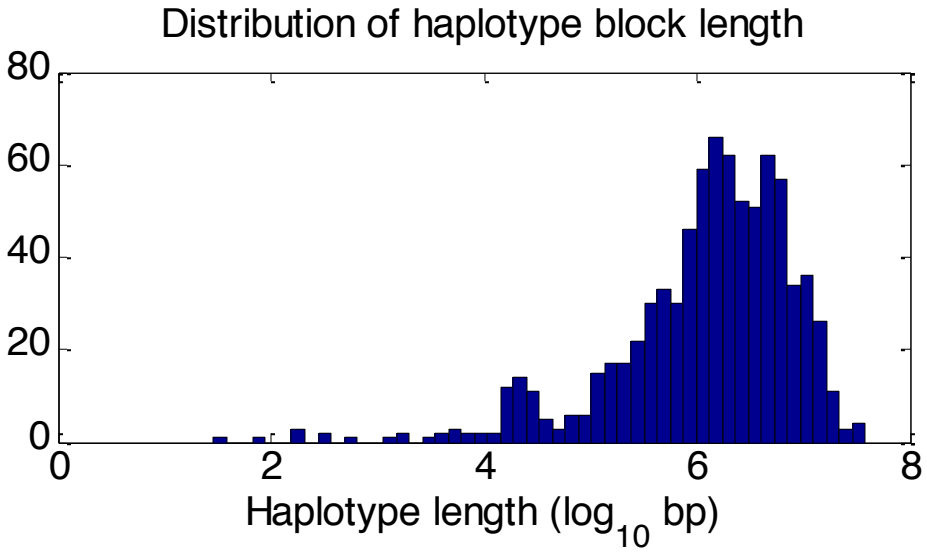

Figure S4. Histogram of haplotype lengths.

Haplotype blocks are defined by recombination positions as shown in Figure S3. Majority of haplotype blocks are long enough to include the most intensive cis-regulatory regions of a gene (100kb near TSS). The median haplotype length is 1.65Mb, and 90% of haplotype blocks range from 0.02Mb to 12Mb.

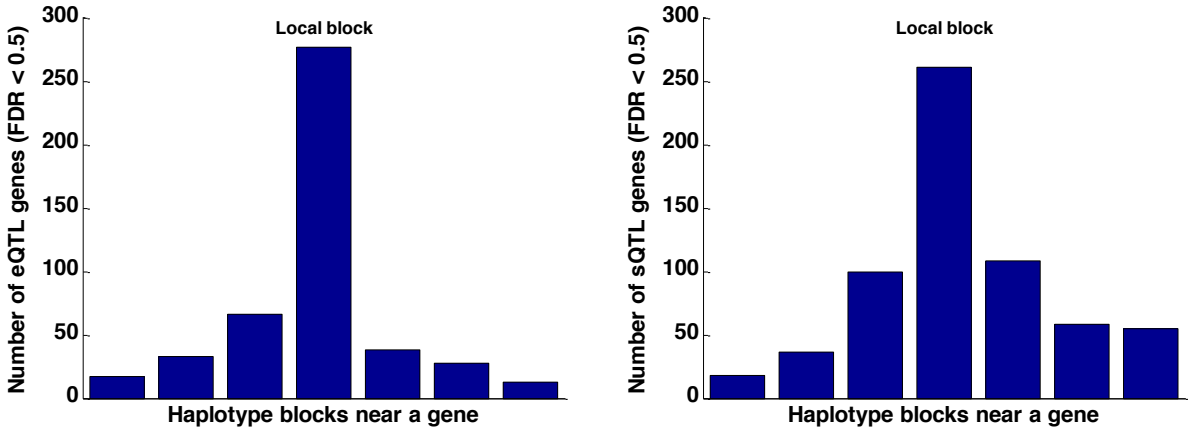

Figure S5. Distribution cis-eQTL and cis-sQTL variants near a gene: Local haplotype blocks have the largest number of eQTL (left) or sQTL (right) effects.

Compared with three nearby blocks, local haplotypes shows substantially larger number of eQTL / sQTL effects, compared to up and downstream haplotype blocks. We tested the local haplotype block containing each gene and three nearby haplotype blocks for eQTL linkage. As we expected, local haplotypes that contain the tested gene show the largest number of eQTL associations compared with nearby blocks. Local haplotypes also show largest number of sQTL associations compared with nearby blocks. The result suggests that most *cis*-acting expression or splicing QTL variants are located in the local haplotype blocks.

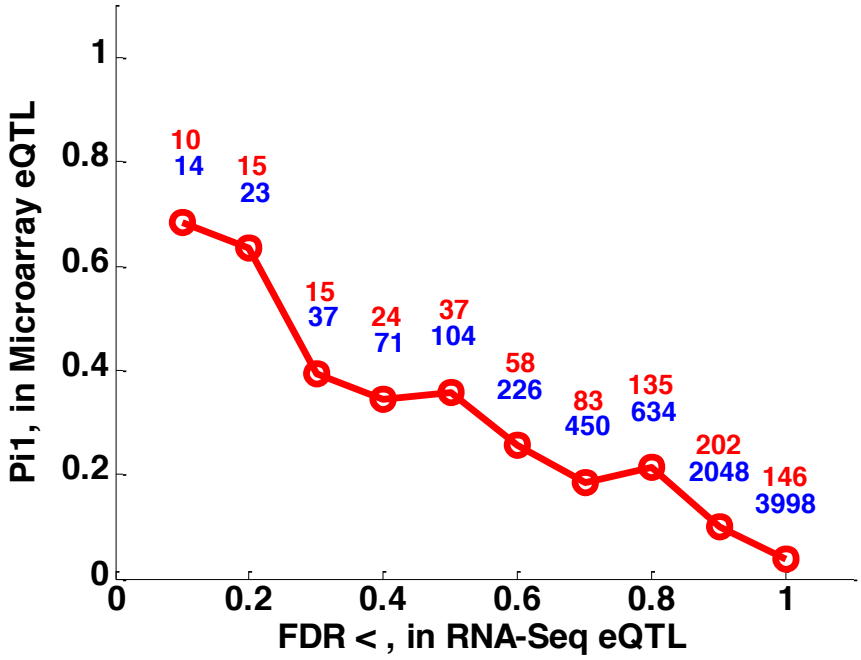

Figure S6. Comparison of eQTL discovery between RNA-Seq and microarray.

We tested eQTLs within the same family quantified in published microarray studies[2] (only seven of the siblings are available from this microarray data). We measured the number of eQTLs from RNA-Seq data that can also be detected using microarray. Blue numbers are total number of eQTL genes detected by RNA-Seq passing that FDR cutoff, red numbers are number of genes also showing eQTL effects by microarray as indicated by π_1[3]. We can observe that given more stringent FDRs that the two approaches give more concordant discoveries. Furthermore, both eQTL discoveries (Figure S6) and effect sizes (Figure S7) show concordant patterns between the two studies.

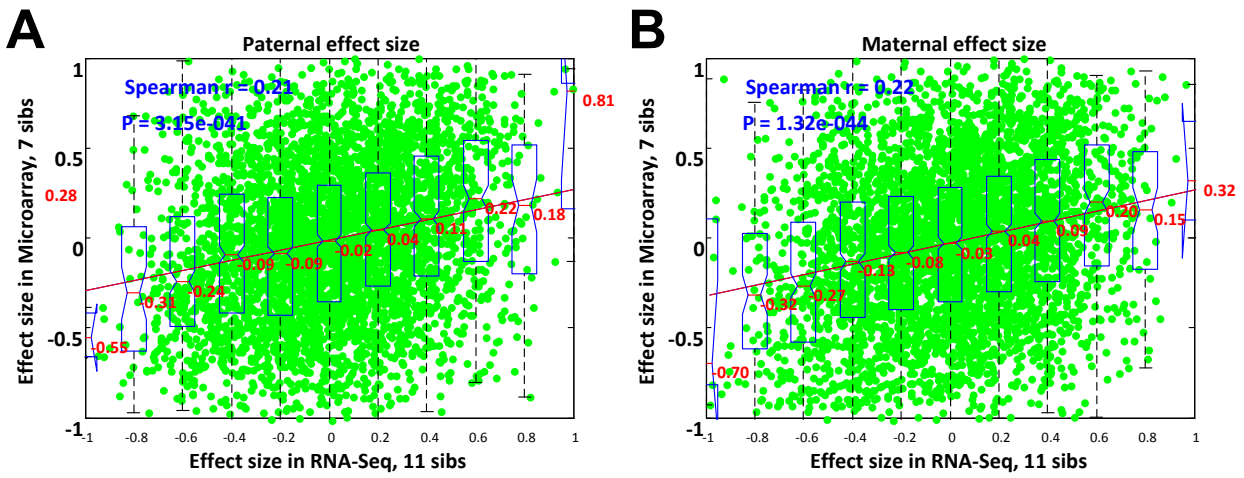

Figure S7. Concordance of eQTL effect sizes (β) between RNA-Seq and microarray.

(A) Paternal effect sizes. (B) Maternal effect sizes. We report effect sizes of eQTL as measured from RNA-Seq data or microarray data. Sign of effect size indicates whether the paternal haplotype of a parent (father or mother) increases or decreases expression in children. Red numbers are medians of each box. Effect sizes measured between microarray and RNA-Seq quantification are modestly concordant.

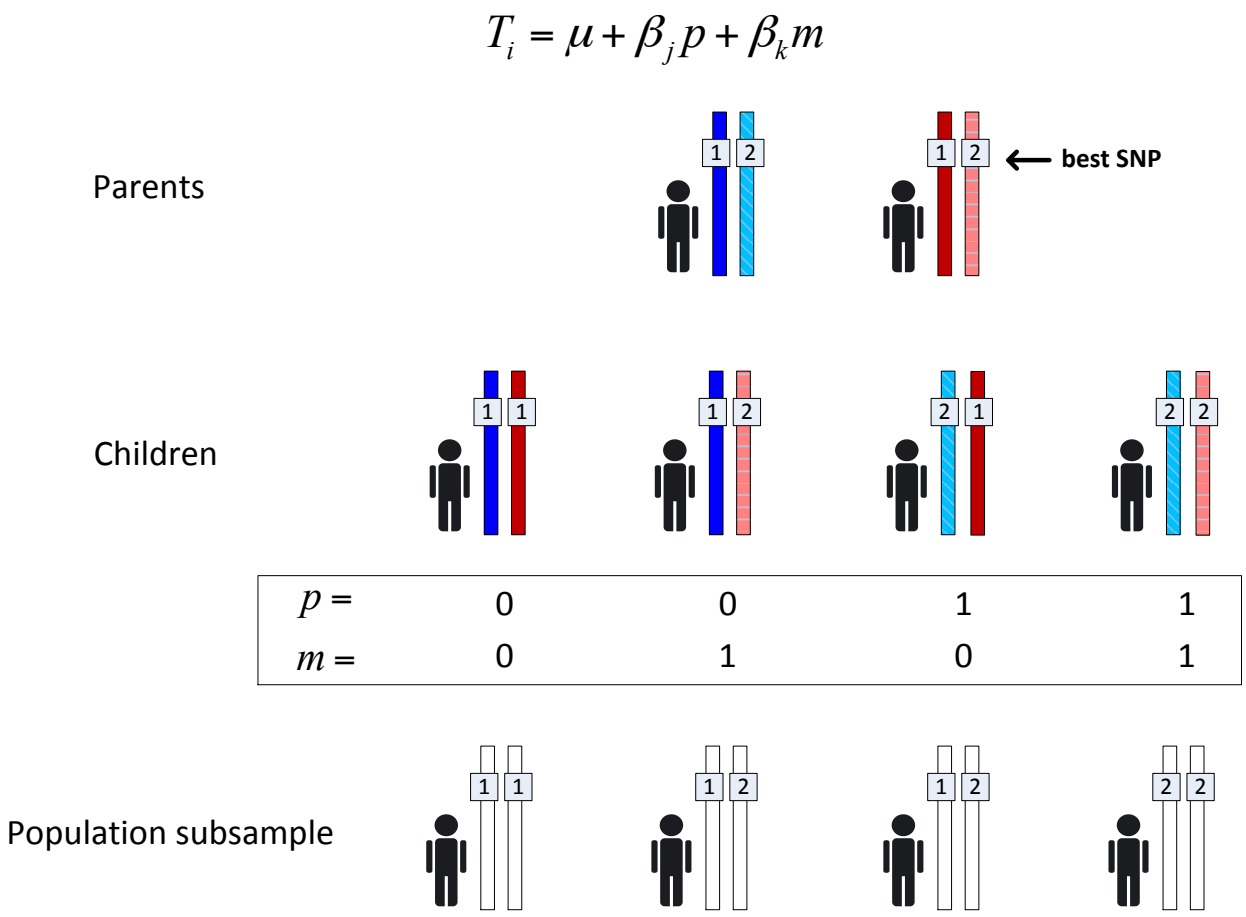

Figure S8. Identifying large-effect cis-eQTL genes in family compared to population

Effect sizes measured in the family are compared to those measured in genotype-matched population subsamples. We have two βs for both the family and the population data to avoid effect size inflation due to more regressors in the family than the population. We use the same regression to measure effect sizes for both the family and the population data: $T_i \sim \mu + \beta_j p + \beta_k m$, p, m are two regressors indicating paternal and maternal haplotypes in the family. We can use the same regression formula because genotypes are matched exactly between family and population subsample at the best associated SNP, so the two regressors p, m match segregating patterns of the best SNP in both the family and the population subsample. If we assume only the best SNP is functioning in both the family and the population samples, β_j, β_k are expected to be the same between the family and the population subsample. For population heterozygotes (with identical, unphased SNP genotypes) the maternal and paternal alleles are assigned arbitrarily from the two possible options, as needed, to match family genotypes. However this extra information does not influence the measure of effect sizes on either side.

Subsequently, large-effect outlier genes are identified by comparing effect size (β or R^2) of genes in the family to those in the population. Effect sizes (β) can be directly compared using analytical tests (Welch's *t*). However, to explore the behavior of effect sizes under different

sample sizes, we applied a subsampling approach among the population individuals to re-generate the expected effect size distribution of the best associated SNP among 11 individuals.

In specific, the best eQTL SNP is discovered in a separate 180-individual discovery panel to avoid bias of multiple selections (a phenomenon otherwise known as regression to the mean or winner's curse). Effect sizes of genes in the population are then assessed by subsampling the same number (N=11) of individuals from the 193-individual replication panel (of 373 Geuvadis European samples) to account for potential biases due to different sample sizes. β or R^2 are regression slope and coefficient of determination (fit) measured by linear regression. The method is illustrated in Figure S8. We then generated the population effect size distribution by subsampling down to 11 matched individuals multiple (100) times from the population data. For each gene, we generated an empirical *p*-value of observing a larger effect in the family, by counting how often the effect size in the family is larger than those from the population subsamples. We estimated the total number of genes exceeding population effect sizes using the π_1 statistic, based on empirical *p*-values of all genes. The estimated π_1 for large-effect cis-eQTLs is 0.0611 (for eQTLs selected by R^2). For β effect size estimates, we noticed there is difference in noise levels between the population and the family estimates which result in a *p*-value distribution significantly skewed towards 1; we discuss an adjustment for such differences in Figure S17, Figure S18 and Figure S19.

We do not use β and only use R^2 when comparing splicing QTL effect sizes, since the estimation noise of β is too large for transcript ratios in Geuvadis data. The estimated π_1 for large-effect cis-sQTL is 0.0749.

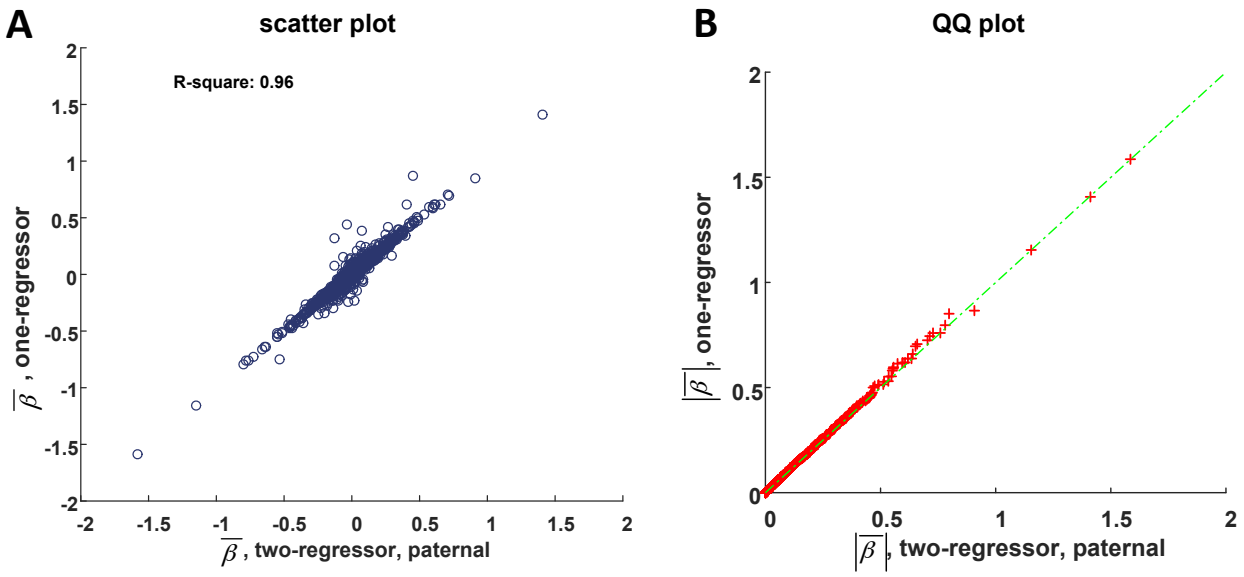

Figure S9. Effect size measured by the one-regressor and the two-regressor model.

For the family, in order to measure the effect of the whole haplotype, we need to use a two-regressor model. For the population subsample, however we chose to use the same two-regressor model to avoid possible effect size (β or R^2) inflation caused by the use of more regressors. This makes a fairest comparison between the family and the population subsample as they are now measured on exactly the same model with the same number of regressors, and effect sizes differences are truly due to biological factors specific to the family instead of different regression methods.

In order to match the regressors of the family, we actually implicitly phased the SNP of the population subsample according to the family (Figure S8), this information is arbitrary for the population subsample however this arbitrary splitting of one regressor into two regressors does not actually influence the measure of effect sizes. The two regressors, p and m, which indicate transmissions from either parents are statistically independent of each other or, in terms of linear relationships, orthogonal: $p \perp m$, $E(p \cdot m) = 0$, therefore each will capture their own effect without interfering with one another. Figure S9 shows the comparison of the actual effect sizes measured by the one-regressor and two-regressor models, which verifies that the two-regressor model unbiasedly captures the same effect sizes ($\bar{\beta}$) as the one-regressor model.

Panel A shows effect sizes measured using the mean from 100 subsamples out of the population. Using the one-regressor model, we simply regress on a single SNP (considering 00, 01, 10, 11 to be 0, 1, 1, 2). Using the two-regressor model, we regress on both the paternal haplotype and the maternal haplotype. Here, we show the effect size β measured on the paternal haplotype, oriented according to the SNP's phase on the father (to match the sign of the SNP β). Panel B shows the QQ plot of effect sizes comparing effect sizes (absolute values) measured by the one-regressor and the two-regressor models. However, the dispersion ($var(\hat{\beta})$) of a two-regressor model is expected to be substantially larger and is a reason why we emphasize the usage of the same model between the family and the population.

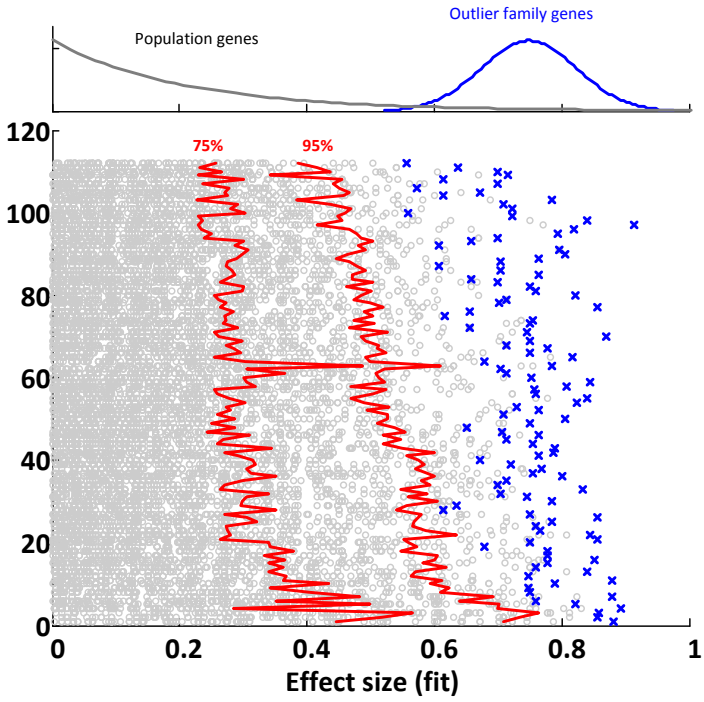

Figure S10. Large-effect family *cis*-eQTL genes.

Effect sizes are compared to population by fit (R^2) of linear regression. Shown are family eQTL genes (blue) with effect sizes greater than the 0.99 quantile (empirical *p*-value < 0.01) of population effect sizes (grey). The magnitude of the outlier proportion has been extended on the top to illustrate the range of effect sizes for measured large-effect *cis*-eQTLs.

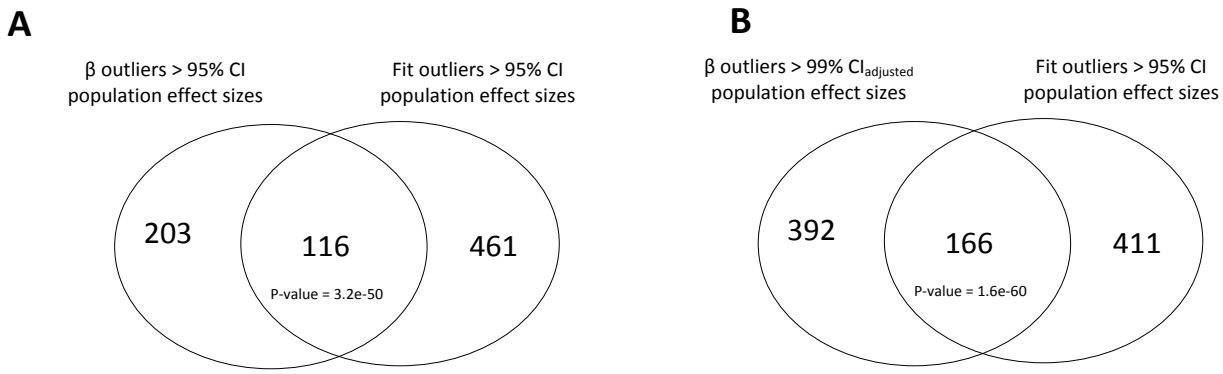

Figure S11. Overlap of β and fit (R^2) effect size outliers.

(A) There are 319 β outlier genes and 577 fit outlier genes with effect sizes greater than 95% quantile (empirical *p*-value < 0.05) of the population. The overlap is 116 genes. The overlap is statistically significant by Fisher's exact test, indicating shared effects they are capturing. (B) After adjustment of confidence intervals of β (described in Figure S17-Figure S19), there are 558 β outlier genes with effect sizes greater than 99% $CI_{adjusted}$ of the population. The overlap with 577 (> 95% CI) fit outlier genes is 166 genes.

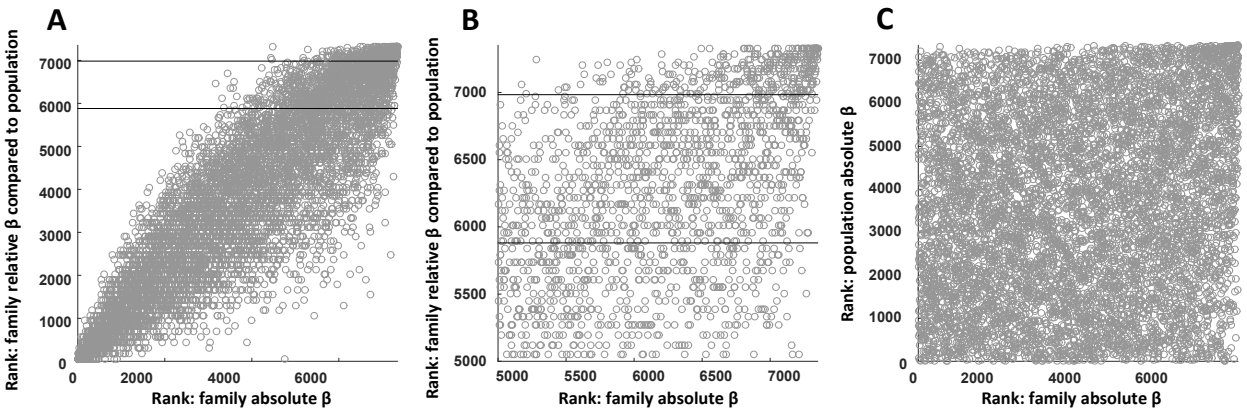

Figure S12. Large relative β vs. absolute β.

We tested the properties of large-effect genes compared to the population to see whether the comparison to the population adds additional information to the ranking of genes. Figure S12 shows that large absolute β are not necessarily highly ranked in the relative scale and vice versa. This does indicate that we are gaining novel information by ranking genes according to their relative effect sizes (empirical *p*-values) instead of just ranking them by their absolute β. (A) Ranks of absolute β, compared to ranks of relative β. Relative β is the empirical *p*-values comparing the family with the population effect sizes. Absolute β is just the original effect size β yielded by the linear regression. (B) Zoom-in of upper right rectangle of (A). The two lines indicate the top 5% and 20% of genes. The figure shows that by comparing to the population, the ranks of genes are not the same as simply ranking the genes by their original β. Of the top 5% of genes by each metric, the overlap is 52%. (C) Ranks of family absolute β compared to ranks of population absolute β. Family effect sizes and population effect sizes are not correlated in general.

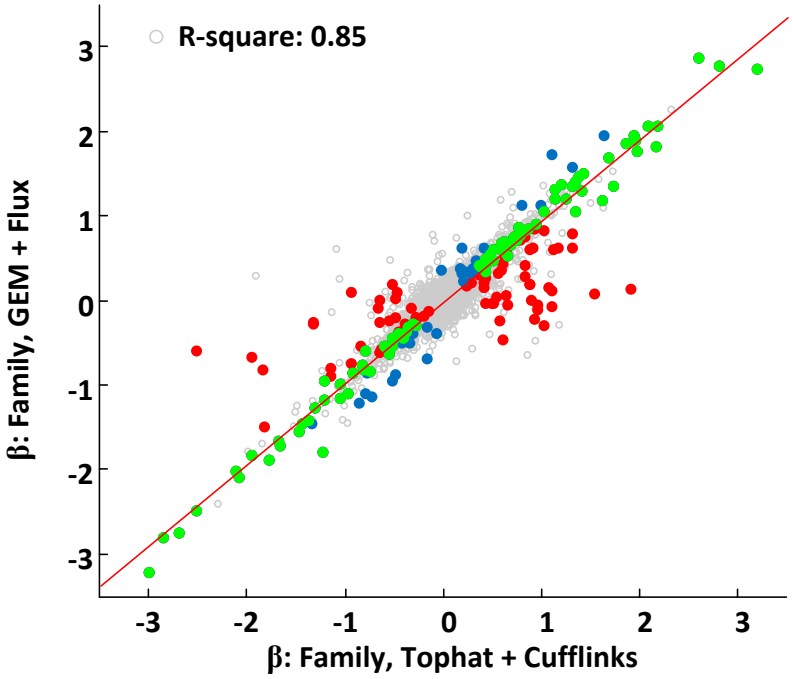

Figure S13. Effect of different quantification pipelines: comparisons of effect size β between Tophat + Cufflinks and GEM + Flux pipelines.

Effect sizes are highly correlated between two quantification methods. Here we plot only paternal side β, maternal side β patterns are very similar. Discovery of large effect size genes (> 95 CI of population, paternal side only) are: 165 genes by Tophat + Cufflinks (red), 125 genes by GEM + Flux (blue) and 90 genes of their intersection (green). Geuvadis expression values were based on a different quantification pipeline than used in the family data. To exclude the possibility that large-effect eQTL genes are due to technical differences between the family and population data, we compared discovery of large effect genes and enrichment of rare variants in the family using the same pipeline (GEM + Flux) as Geuvadis. For the family data, the effect size estimates are highly correlated between two pipelines. We observe a similar discovery set of large effect genes and also similar patterns of rare variant enrichment as Tophat + Cufflinks pipeline (Figure S13, Figure S14). The confidence intervals used for β effect sizes in Figure S13-Figure S16 are raw (if not otherwise specified), without further adjustment (described in Figure S17-Figure S19).

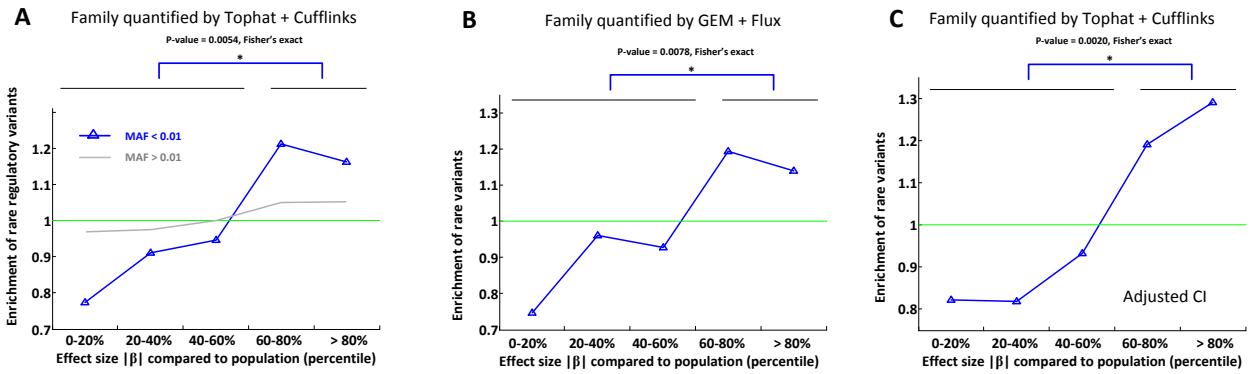

Figure S14. Enrichment of rare variants at large effect size β.

The enrichment pattern is similar when using GEM + Flux (A) pipeline compared to Tophat + cufflinks pipeline (B) and (C). X-axes in (A) and (B) are raw CIs, (C) is adjusted CI. We ranked effect sizes of genes based on 1 – their empirical *p*-values: how often their effect sizes in the family are larger than effect sizes among the population subsamples. The distribution of effect sizes in the population was generated by repeatedly subsampling 11 individuals from the population. Rare variants are defined as those with MAF < 0.01, within Encode TF binding + DNase peaks and PhyloP score > 1. Enrichment is defined as proportions of genes with such an annotated rare variant in each effect size bin divided by proportions of genes with such an annotated rare variant across all effect size bins.

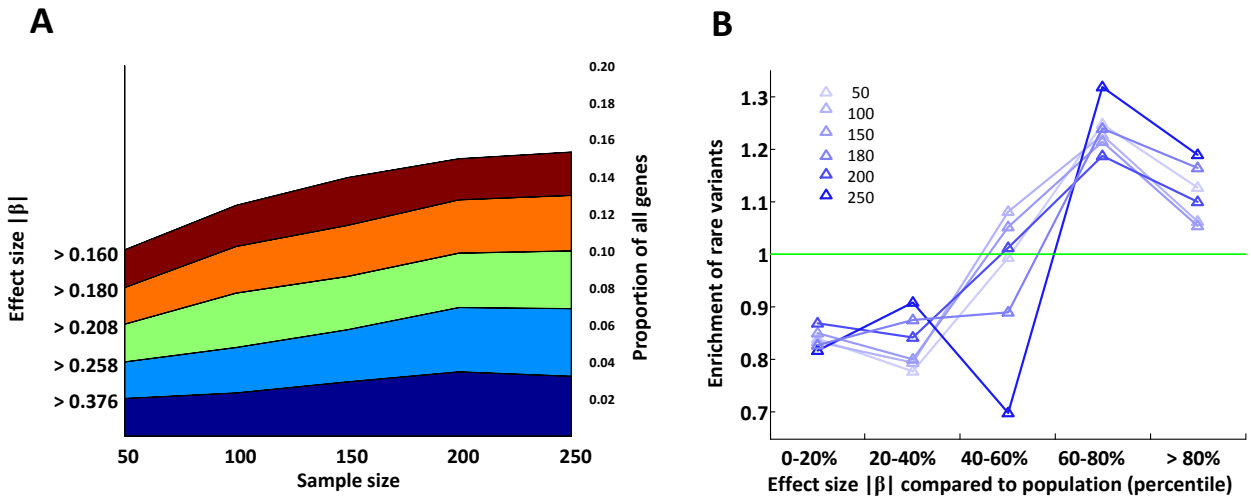

Figure S15. Influence of discovery sample sizes in tagging causal SNPs.

As we are using whole genomes for both the family and the population data, we have the ability to test all SNPs including the causal SNP in our association models. However, a general rule of statistics indicates that the power to capture the true largest effect or causal SNP in the population also depends on sample sizes. A small discovery panel may result in poor choice of SNP and deflation of effect sizes in the population, which can potentially over-estimate the number of large effect genes in the family. We analyze how sample sizes of the population discovery panel influence our identification of large effect genes in the family. We observe continuously increasing number of large effect eQTLs discovered in the population given larger sample sizes (Figure S15A), which indicates that large sample sizes do increase chance of tagging a true causal SNP. We consider this a very important effect suggesting the necessity of large sample sizes to accurately measure effect sizes. However, given our particular application, as largest effect genes are likely to saturate first, increasing sample sizes does not have a significant influence on our discovery of family large effect genes (Table S6). The enrichment of rare variants at large-effect genes is also comparable given different discovery panel sizes (Figure S15B).

(A) Number of large effect genes discovered in the population given larger sample sizes. Best SNPs are discovered in the discovery panel of varied size. We re-measured the effect sizes of those SNP in the replication panel. There are increased numbers of large effect SNPs discovered given a larger discovery panel size. Note that y-axis is piled inversely, with largest effect sizes stacked at the bottom. (B) Enrichment of rare variants at large effect family genes given different population discovery panel sizes. We ranked effect sizes of genes based on 1 - their empirical *p*-values: how often effect sizes in the family are larger than those of the population subsamples. Rare variants are defined as those with MAF < 0.01, within Encode TF binding + DNase peaks and PhyloP score > 1. Enrichment is defined as proportions of genes with such an annotated rare variant in each effect size bin divided by proportions of genes with such an annotated rare variant across all effect size bins.

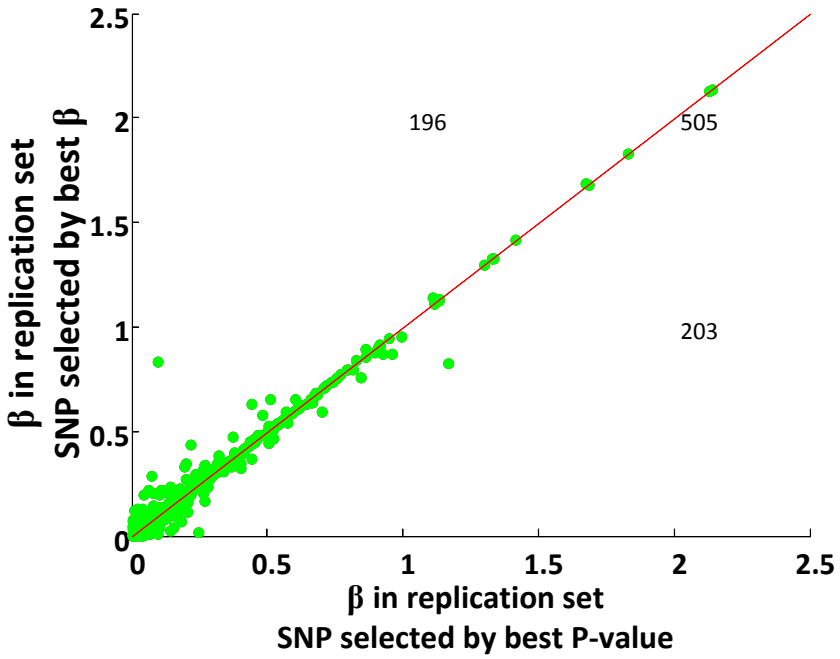

Figure S16. Influence of different criteria in selecting best SNP: smallest *p*-value or largest effect size.

X axis: re-measured β in replication panel if SNP is selected by best P value in the discovery panel. Y axis: re-measured β if SNP is selected by best β in the discovery panel. Shown here are 904 genes with a best SNP < 1e-5 in discovery panel. For each gene, we choose a best β among all SNPs within 10 * *p*-value of the best SNP. 203, 196 and 505 are genes with X > Y, X < Y and X = Y.

For each gene, we select the best SNP based on *p*-value in the discovery panel. However, this SNP is not necessarily truly the largest effect SNP as *p*-value is an indicator of best fit (R^2) instead of largest β, such that we may possibly miss a secondary effect SNP with larger β. To test the possibility that we miss larger effect SNPs and under-estimate the effect sizes in the population, we analyzed the differences of choosing the best SNP by *p*-value or β. For each gene, we first find a best SNP with smallest *p*-value, we then pick another SNP of largest β (could be the same one) among all SNPs with a *p*-value no more than an order of magnitude less significant than that of the best *p*-value, we then re-measure their effect sizes in the replication set. We observe that approximately half of the time, the best *p*-value and best β SNP is the same SNP. Further, even when they are not the same SNP, the measured effect size in the replication set is very similar. This suggests that most effect size differences near the best SNP is due to random noise, the existence of a secondary effect SNP with even larger effect size is not significant.

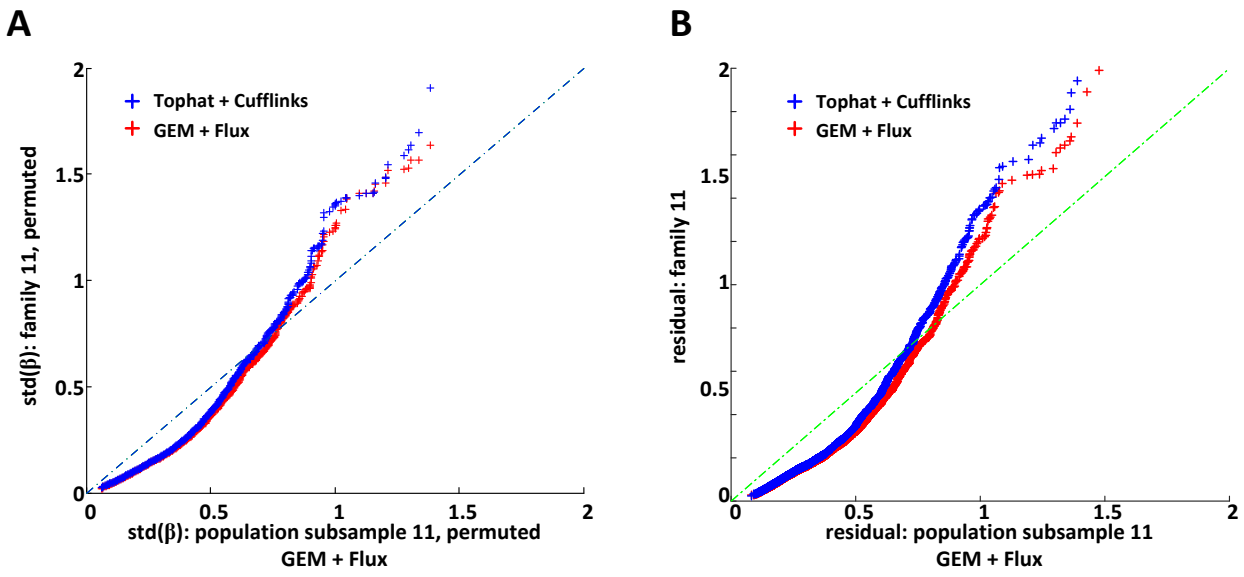

Figure S17. Adjustment of effect size empirical *p*-values: comparison of effect size confidence intervals (noise levels) between the family and the population.

(A) QQ plot of standard deviation of measured effect size $\hat{\beta}$, comparing the family and the population. Data are permuted such that $E(\hat{\beta}) = 0$, $std(\hat{\beta})$ measures the noise levels of measures of β. Population (Geuvadis) samples are quantified using GEM + Flux pipeline. Family data are quantified using both GEM + Flux and Tophat + Cufflinks pipelines respectively. Standard deviation of $\hat{\beta}$ is significantly larger in the population no matter which pipeline is used in family. (B) QQ plot of analytical estimation of standard error of β: $\hat{S}_\beta$, comparing the family and the population. $\hat{S}_\beta$ is computed by $\frac{1}{\sqrt{\sum(x_i-\bar{x})}}\sqrt{\frac{\sum \varepsilon_i{}^2}{11-3}}$, where ε_i is the residuals in linear regression $T_i \sim \mu + \beta_j p + \beta_k m + \varepsilon_i$, x is either p or m.

The degree of noise in estimated β is different between the population and the family even if we match the sample size and quantification methods. The noise in estimated effect sizes is significantly smaller in the family than in the population. This difference reflects both the fact that family members are more homogeneous (sharing more covariates such as genetic, environment, lifestyles and etc., thus having tighter fit to the regression slopes) and also the possible existence of other technical factors, which we cannot tell apart.

The discovery of such differences is actually biologically informative, however our subsampling scheme is not intended to reflect and calibrate the noise of effect size estimates of the family members. Therefore, regardless of the source of these differences, they have undesirably shifted the empirical *p*-values (see comparisons to analytic *p*-values Figure S18A and C, Figure S19A).

Here, we explored two methods to adjust different noise levels between effect size estimations, which yield empirical *p*-values closer to analytical *p*-values and less conservative estimates of FDR. However as there is not a robust way to precisely calculate this FDR, we

leave the over-conservative empirical *p*-values unadjusted for all main analyses. It is important to note that "FDR" here measures overall excess (FDR < 1) of large effect sizes between the family and the population, it does not mean that the ordering of effect sizes (empirical *p*-values) are all due to random chance regardless of the outcome of this FDR. Though we did not estimate an accurate FDR here, the relative ranks of genes according to their effect sizes compared to the population (empirical *p*-values) are not affected, which are still valid and biologically meaningful.

To correct for additional noise in population subsamples, we measured the standard deviation of β of randomly permuted data ($\bar{\beta} = 0$) in both the family and population. We estimated that the standard deviation of β estimates of the family is 0.55 times the size of that of the population: $std(\hat{\beta}_{\text{family}}) = 0.55 * std(\hat{\beta}_{\text{population}})$ (Figure S17A). Such difference will make the confidence intervals which are measured from subsampled population to be larger than the actual noise of β estimates in the family. To adjust for such differences, we narrowed the empirical distribution of a gene by moving each subsampled effect size in the population towards their mean: $\beta_{\text{adjusted}} = \bar{\beta} + 0.55 * (\beta - \bar{\beta})$. This adjustment shrinks the distribution of population effect sizes (and consequently reduces the empirical *p*-values) but retains the estimates of $\bar{\beta}$ of that gene from the population. After adjustment, the distribution of empirical *p*-values testing whether the family effect size is bigger than the subsampled population is more uniform (Figure S18). As the adjustment of confidence intervals mainly influences calculation of the FDR, the enrichment pattern of rare variants was very similar under the adjusted confidence intervals (comparing Figure S14, B and C).

Alternatively, we also directly estimated analytic standard errors (confidence intervals) of $\hat{\beta}$ without using permutation. Assuming residuals are normally distributed, from linear model theory the standard error of $\hat{\beta}$ is estimated (MLE) by $\frac{1}{\sqrt{\sum(x_i - \bar{x})}}\sqrt{\frac{\sum \varepsilon_i{}^2}{11-3}}$, where ε_i is the residuals in linear regression $T_i \sim \mu + \beta_j p + \beta_k m + \varepsilon_i$, x is either p or m. We compared the difference of standard errors of $\hat{\beta}$ between the family and the population subsamples (Figure S17B). The estimated global difference of the standard error of $\hat{\beta}$ follows $\hat{S}_{\beta_{\text{family}}} = 0.55 * \hat{S}_{\beta_{\text{population}}}$. The scaling factor is very similar to that inferred by permutation.

Here, both standard error tuning methods make the assumption that the noise in the family for each gene is approximately a constant scaling factor less than the noise in the population. The tuning factor is obtained by matching ranks (*U* statistic of Wilcoxon rank-sum test) of two distributions of standard errors of $\hat{\beta}$, until the test is not significant.

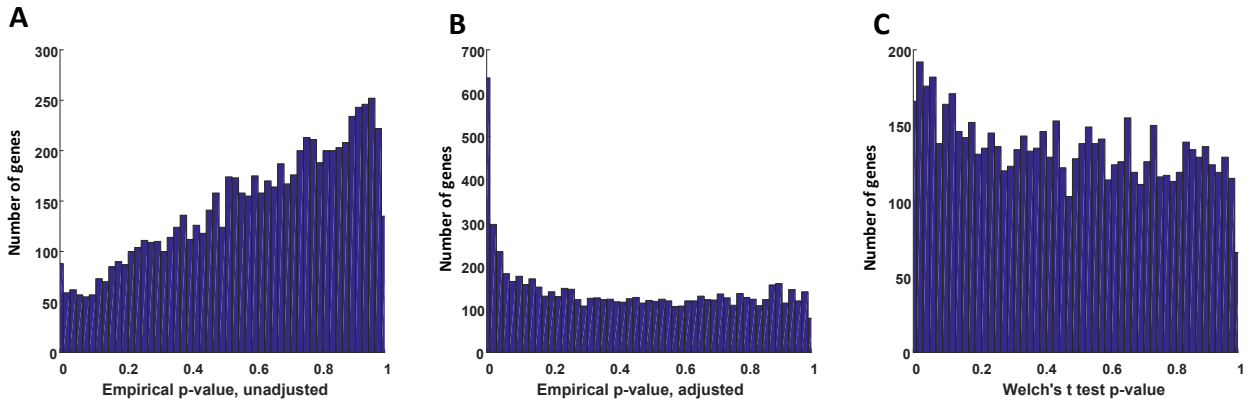

Figure S18. Adjustment of effect size empirical *p*-values: distribution of *p*-values of family versus population effect sizes.

Empirical *p*-values generated by subsampling: (A) before noise adjustment and (B) after noise adjustment. For each gene, we compute how often (1 – empirical *p*-values) the family effect size are larger than the effect sizes of the population subsamples. Distribution of effect sizes in the population is generated by subsampling 100 times from the population. After adjustment of their different noise levels, empirical *p*-values are more evenly distributed. (C) Welch's *t*-test *p*-values. Welch's *t*-test is performed by directly using a *t*-test between two regression slopes (two β's), with standard errors estimated analytically.

We computed the straightforward analytical *p*-values (Welch's *t*-test) without using any subsampling (single-SNP regression over whole replication panel), which provide a bottom-line theoretical control of the empirical *p*-values. The adjusted empirical *p*-values lie much closer to theoretical *p*-values than the raw empirical *p*-values. Here, we can simply use a pure analytic test to compare regression slopes β_{family} and $\beta_{\text{population}}$ without either subsampling or permutation by applying Welch's *t* test: $\frac{\beta_{\text{family}}-\beta_{\text{population}}}{\sqrt{\left(\hat{S}_{\beta_{\text{family}}}\right)^2+\left(\hat{S}_{\beta_{\text{population}}}\right)^2}}$. Under normality assumption of regression residuals, this test statistic follows *t* distribution, the standard errors are analytic estimations from regression residuals: $\hat{S}_{\beta_{\text{family}}} = \frac{1}{\sqrt{\sum(x_i-\bar{x})}}\sqrt{\frac{\sum \varepsilon_i{}^2}{11-3}}$, $\hat{S}_{\beta_{\text{population}}} = \frac{1}{\sqrt{\sum(x_i-\bar{x})}}\sqrt{\frac{\sum \varepsilon_i{}^2}{373-180-2}}$, degree of freedom is $\frac{\left(\left(\hat{S}_{\beta_{\text{family}}}\right)^2+\left(\hat{S}_{\beta_{\text{population}}}\right)^2\right)^2}{\left(\hat{S}_{\beta_{\text{family}}}\right)^4/(11-3)+\left(\hat{S}_{\beta_{\text{population}}}\right)^4/(373-180-2)}$, 373-180 = 193 is the size of replication panel where effect size of the best associated SNP is re-measured.

$FDR_{adjusted}$ of those large effect genes at a given empirical *p*-value (1 – $CI_{adjusted}$) cutoff is calculated as (total number of genes * *p*-value cutoff) / number of discoveries. At *p*-value < 0.01 ($CI_{adjusted}$ > 0.99), there are 558 larger effect β in the family compared to the population, $FDR_{adjusted}$ = 7341 * 0.01 * 2 / 558 ~ 0.26 (* 2 because we combined paternal and maternal discoveries). It is important to note that while our reported FDR is conservative, the adjusted FDR may be permissive if the differences in the variance reflect meaningful biological differences. By Welch's t-test, there are 320 larger effect β (*p*-value < 0.01) in the family compared to the population, FDR = 7341 * 0.01 * 2 / 320 ~ 0.46. We conclude that there is

definitely a significant excess of large effect cis-eQTLs in the family than in the population, however as there is not yet a very robust estimation of this proportion, we choose to state the conservative FDR in the main text.

It is also important to note that this “FDR” measures whether there is overall excess of large effect genes in the family. The ranking of empirical *p*-values which reflects the positioning of effect sizes in a population spectrum is still biologically meaningful regardless of this excess. The downstream analysis based on rankings of effect sizes does not rely on this estimation of FDR.

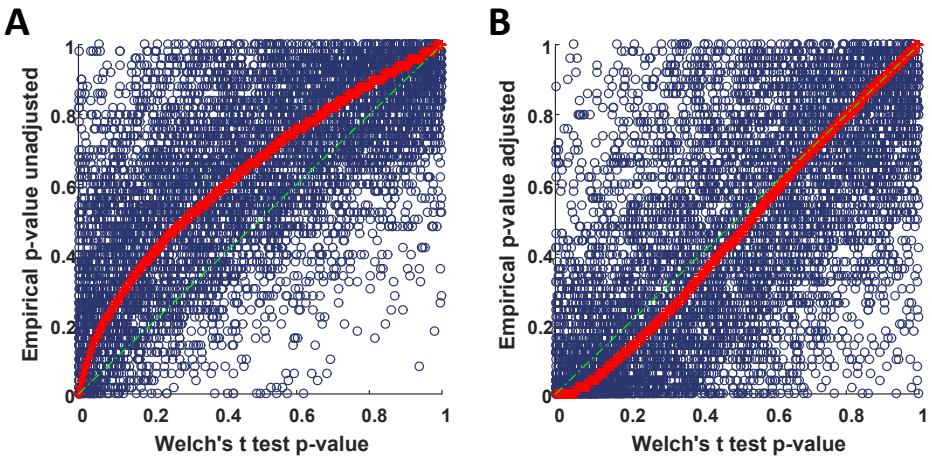

Figure S19. Adjustment of effect size empirical *p*-values: comparisons of empirical *p*-value and Welch's *t*-test.

Unadjusted empirical *p*-values (A) are significantly conservative than Welch's *t*-test, while adjusted *p*-values (B) are more optimistic than Welch's *t*-test. Here we only show empirical *p*-values and Welch's *p*-values measuring the difference of the paternal side β_j in the family and single-regressor $\beta_{\text{population}}$ of the population, the maternal side is similar.

As effect size (β) can be directly compared using analytical tests, to gain a theoretical control of the correctness of the subsampling scheme, we performed the conventional analytical test (Welch's *t*) to compare effect sizes. Here, population β is just a one-regressor (the best SNP) straightforward measurement of effect sizes over the whole replication panel without subsampling or implicit phasing. The analytical test can be used to gauge the overall soundness of empirical *p*-values. Comparing Welch's *t*-test with subsampling + permutation (empirical *p*-value) based test, these three *p*-values are mostly concordant with each other, however the empirical *p*-values are more conservative than Welch's *t* test before adjustment but more optimistic than Welch's *t* test after adjustment (Figure S19).

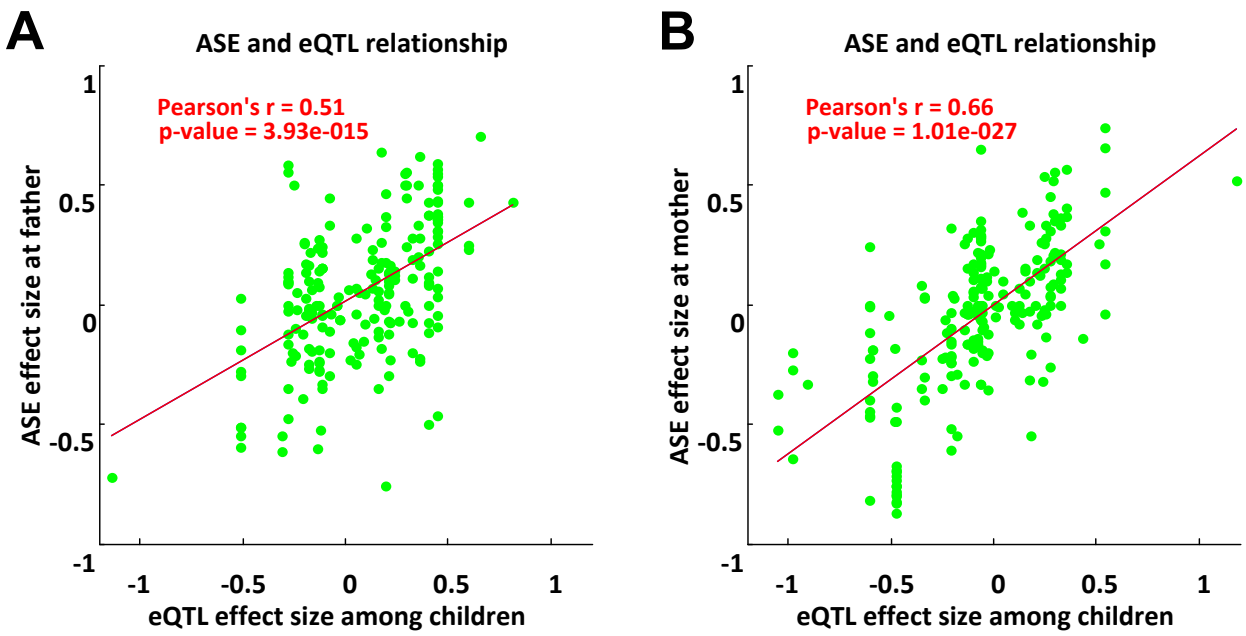

Figure S20. Correlation of eQTL effect size β and ASE effect size (allelic imbalance).

(A) Paternal β and allelic imbalance in father. (B) Maternal β and allelic imbalance in mother. ASE effect sizes in the parents and *cis*-eQTL effect sizes among the children should have a simple linear relationship. We computed correlation between ASE (quantified by allelic imbalance) in the parents and eQTL effect sizes (quantified by linear regression β) among the children. Indeed, we observed a linear relationship between ASE effect sizes in the parents and eQTL effect sizes among the children. In other words, the difference between two homologous alleles in a parent will exhibit as between-individual differences among the children, as expected by Mendelian segregation. For example, expression level difference between two homologous alleles (ASE) of the parent NA12878 (a1,a2): a1 – a2 is proportional to expression level differences between her offspring (a1,*) – (a2, *) depending on which haplotype they inherit. We observed that when a haplotype is highly expressed in a parent as indicated by ASE, children inheriting that haplotype also have higher expression levels. ASE effect size at a heterozygous site is represented by (paternal reads – maternal reads)/(paternal reads + maternal reads), i.e., 2 * (paternal allelic imbalance – 0.5). *cis*-eQTL effect size is defined as the difference of gene expression levels between children inheriting different haplotypes (which is simply β of linear regression: $T_i \sim \mu + \beta_j p + \beta_k m$). We can observe that β is linearly determined by allelic imbalances.

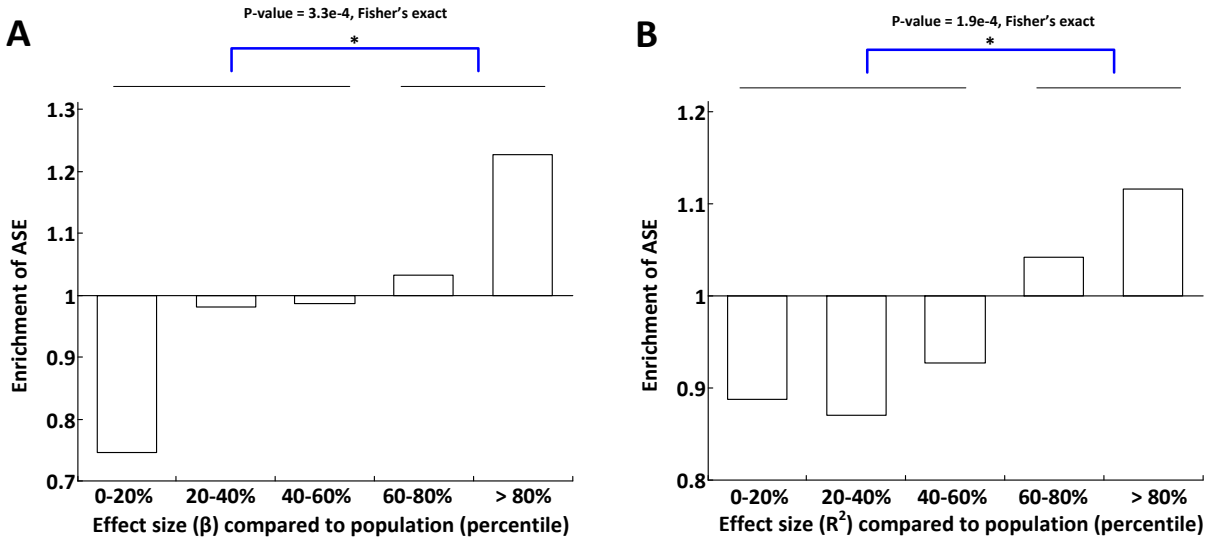

Figure S21. Enrichment of ASE effects at large-effect genes.

To further confirm that identified large-effect genes are potentially due to rare and heterozygous variants instead of technical artefacts, we assessed enrichment of ASE effects for large-effect eQTLs. ASE effects are evaluated in both parents of the family. In theory, large-effect eQTLs among siblings should also exhibit as ASE effects among at least one of the parents. For both β and fit, we observed increasing incidence of ASE effects at larger effect eQTLs. (A) Enrichment of ASE at large-effect (measured by β) genes. ASE effects (measured upon two parents) are defined by those passing binomial test *p*-value < 0.01 and allelic imbalance > 0.05. (B) Enrichment of ASE at large-effect (measured by R^2) genes. We ranked effect sizes of genes in the family based on the how often (x-axis, 1 – empirical *p*-values) their measured effect in the family was greater than in the population subsamples. Enrichment is defined as the proportion of genes exhibiting ASE in each effect size bin divided by the proportion of genes exhibiting ASE across all effect size bins. We only consider genes testable for ASE, i.e., with heterozygous sites in RNA covered regions.

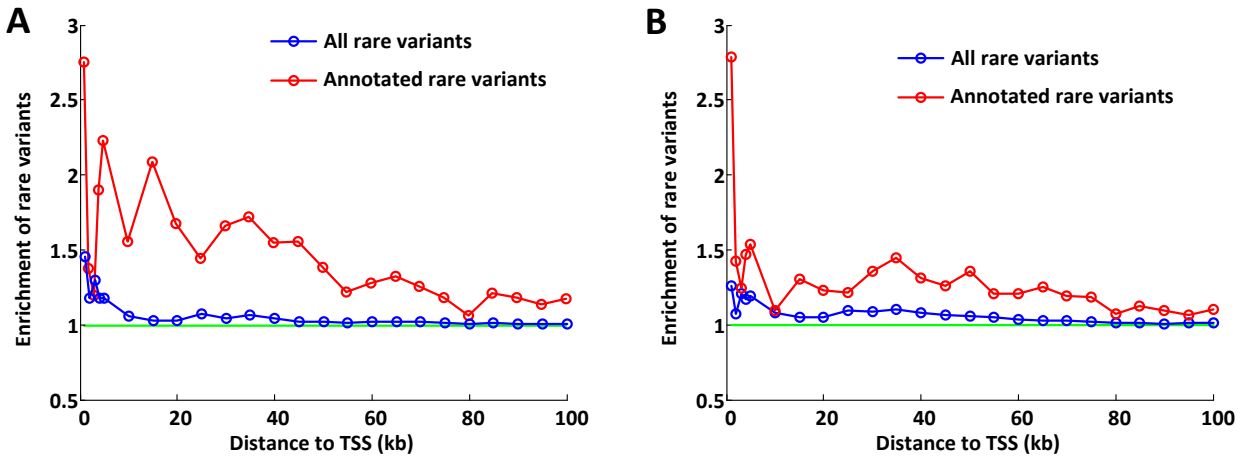

Figure S22. Rare regulatory variants contributing to large-effect eQTLs: enrichment of rare variants near the TSS of large-effect (β) cis-eQTL genes, comparing annotated and all rare variants.

We examined enrichment of rare variants near the transcription start site (TSS) of eQTL genes. Allele frequencies are based on the Phase 1 release of the 1000 Genomes Project European populations. We narrowed variants by multiple functional annotations such as conservation score (PhyloP) and regulatory features from annotations in RegulomeDB[4] indicating Encode[5] TF binding and DNaseI hypersensitivity peaks.

We observed an increasing enrichment of rare variants at larger effect size genes. Likewise, given a rare variant in an annotated regulatory region, we also see a significantly increased proportion of large effect genes. The enrichment is stronger in the immediately proximity of the TSS but also spreads across the 100kb regions. The enrichment is also much stronger among annotated regulatory sites than all other sites. Enrichment is defined as proportions of genes with such an annotated rare variant in each effect size bin divided by proportions of genes with such an annotated rare variant across all effect size bins.

(A) 319 genes CI > 0.95. (B) 558 genes $CI_{adjusted}$ > 0.99. The plot shows enrichment for all rare variants (MAF < 0.01, 100kb near TSS) and annotated rare variants (MAF < 0.01, 100kb near TSS, within Encode TF binding and DNaseI hypersensitivity peaks and with PhyloP score > 1). We observed increased enrichment of rare variants near the TSS of larger family effect size genes. Enrichment is stronger for annotated rare variants.

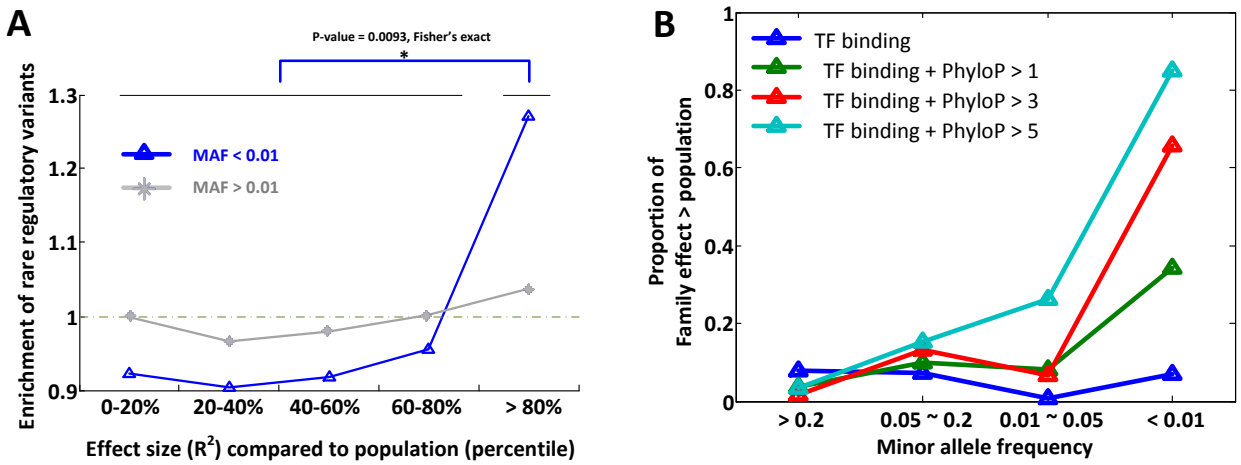

Figure S23. Rare regulatory variants contributing to large-effect eQTLs.

(A) Enrichment of rare variants at large-effect genes. Effect sizes are measured by fit (R^2) and binned by comparing to population effect sizes. We ranked genes according to their effect sizes in the family as percentile (x-axis, 1 – empirical *p*-values) in the population. Rare variants are defined as those of MAF < 0.01, Encode TF + DNase peak, PhyloP > 1 and within 100kb near TSS. (B) Utility of rare variants in predicting a larger effect in family than population. Enrichment is defined as proportions of genes with such an annotated rare variant in each effect size bin divided by proportions of genes with such an annotated rare variant across all effect size bins. (R^2). Rare variants are restricted to those in Encode TF + DNase peaks and different PhyloP score cutoffs. We estimate the proportion of family effects larger than population effects using π_1 statistics. π_1 is estimated from empirical *p*-values of whether a family effect size is larger than population by counting the number of times a family effect size is greater than subsampled population.

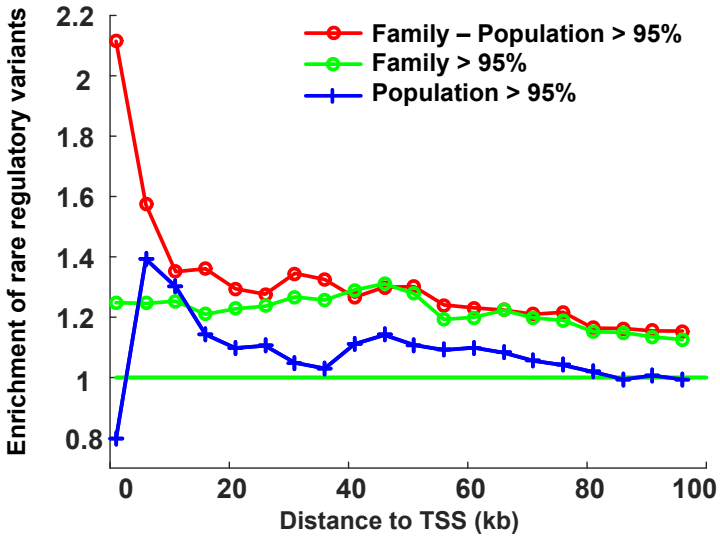

Figure S24. Enrichment of rare regulatory at large effect genes.

One might ask whether enrichment of rare variants is a general property of large-effect eQTL genes regardless of whether they are family specific or not (due to potentially larger number of regulatory elements near those genes). To explore such possibilities, we evaluated enrichment of rare regulatory variants at three categories of genes: genes whose effect sizes are larger in the family than in the population, genes whose effect sizes are large in the family regardless of whether they are larger than the population and genes whose effect sizes are large in the population. Here we consider three types of genes: genes with larger effect in the family than the population (red), genes with large-effect in the family regardless of effect sizes in the population (green) and genes with large effects in the population (blue). We consider the top 5% of the genes in each category. We only observe enrichment in the former two categories. This indicates that enrichment of rare variants is only at those family-specific large effect genes, it is not due to general enrichment of regulatory elements near large-effect genes.

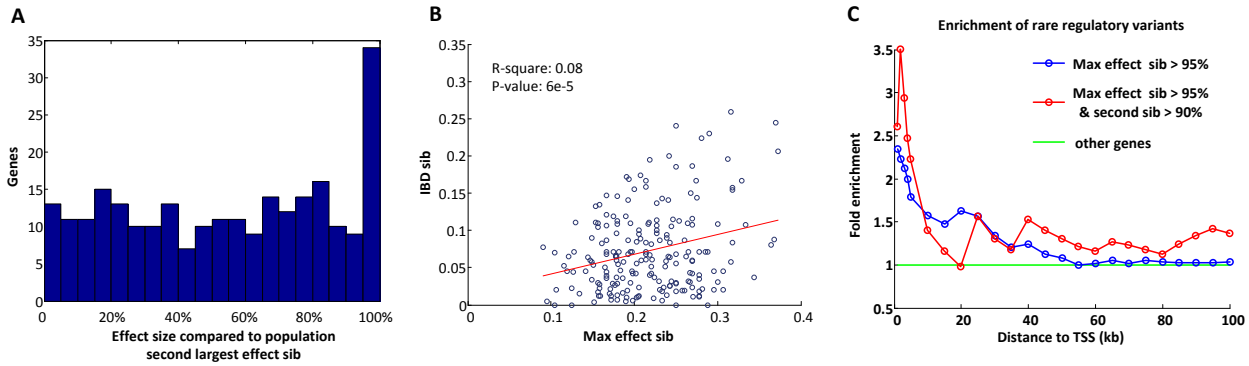

Figure S25. Identification of large ASE effect.

Genes are chosen where largest ASE effect sibling is > 95% quantile of population ASE effects. We checked ASE effects at a second largest effect sibling and IBD siblings to further confirm those large effects. (A) Here we show distribution of effect sizes (percentile as compared to population, 1 – empirical *p*-values) of second largest ASE effect sibling at outlier genes. (B) ASE effect size (allelic imbalance) between largest effect sibling and its double-IBD (identical-by-descent) sibling. (C) Enrichment of rare regulatory variants near TSS at those genes where both the first and second largest effect sibs are significantly larger than population. At each TSS cutoff, enrichment is defined as the proportion of large ASE effect genes with an annotated rare variant divided by proportions of genes with an annotated rare variant for all genes which are testable for ASE.

As ASE effects are evaluated at individual heterozygous sites, we wanted to exclude the possibility that large-effect ASE is due to technical artifacts such as mapping biases or sequencing errors. To achieve this, we looked at ASE for the second largest effect sibling. Our rationale being that the second sibling would be less likely to be a coincidental artifact than the first. We observe that the ASE effect at the second largest sibling is also significantly enriched for larger effect sizes (Figure S25A). Furthermore, we also looked at an identical-by-descent sibling of the maximum effect sibling. We observe that large effects are repeated at the IBD sibling (Figure S25B). When looking at genes with both first and second largest ASE effects greater than population, we observed strong enrichment of rare variants at those genes (Figure S25C).

ASE discovery. The following are additional notes on discovering large ASE effects and its FDR. We applied a similar method to identify large-effect ASE in the family. We use allelic ratio as a measure for ASE. Large-effect ASE genes are detected by comparing maximum allelic imbalance among 11 siblings and ASE in subsampled population data. We subsampled 11 individuals from the population (373 European individuals from Geuvadis data) and take the maximum allelic imbalance. We calculated empirical *p*-values of family effect sizes according to the effect size distribution of the population subsamples. To account for the differences in read depths between the family and population data, we further down-sampled the population data by a ratio of 1.97.

We discover 223 large effect genes at CI > 0.95 (with empirical *p*-value < 0.05), which yields an FDR = 1777 * 0.05 / 223 = 40%. We do not expect to see more large ASE genes in the family than in a population subsample. Unlike eQTLs, there is equivalent statistical power in the family and a population subsample to detect ASE effects either arising from rare or common variants. The excessive number of large effect genes mainly reflects read depth differences (lower read depth leads to larger allelic imbalance) between two datasets we have not yet corrected out. We are trying to correct out this factor by using a uniform down-sampling factor of 1.97 which reflects the global read depth difference between two datasets. However as there are substantial variability of read depth between individuals and sites, this global correction cannot remove all the technical differences.

It is very important to mention that by theory FDR for ASE should be inherently 1. However this “FDR” is a measurement of the excess of large ASE in the family compared to the population (which should be zero), it does not mean that large effect sizes are out of random chance. Empirical *p*-values here are not just random noises; they have biological meanings individually, reflecting the positioning of ASE effect sizes of one individual among the natural spectrum of all individuals. Therefore, the ranking of those genes by their empirical *p*-values are still biologically meaningful regardless of whether there is overall excess of larger effect sizes. As our purpose is not to estimate whether there is excess or not but to obtain an ordering of genes by their relative ASE effect sizes, it is therefore critical to emphasize the meaning of this FDR here.

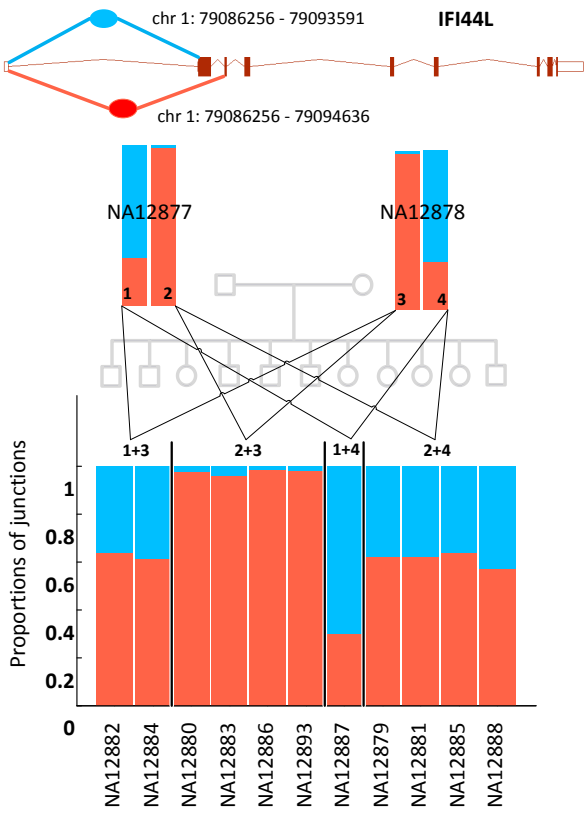

Figure S26. Mendelian segregation of alternative splicing patterns.

Alternative splicing patterns determined by haplotype groups. 1,2,3,4 are paternal grandfather, grandmother and maternal grandfather and grandmother haplotypes. We observed that transcript ratios can exhibit as Mendelian segregation in the family. We use JunctionsTK (junction toolkit, a tool developed by our group) to quantify such segregation patterns using splicing reads. JunctionsTK uses reads spanning splice junctions from junctions.bed files produced by TopHat. It calculates proportions of junction reads from one donor exon to different acceptors (or different donors to a same acceptor). Compared to transcript abundance reported by Cufflinks, splice junction reads, as they are specific to each alternative transcript, more directly inform alternative splicing differences between individuals. The figure shows segregation of splicing junction usage of these genes by different ancestral haplotypes. We show differentially expressed transcripts among four groups of siblings (depending on which two grandparental alleles are inherited), each group is divided by vertical bars. Here, the y-axis shows proportional usage of each junction site, from the same donor exon to different acceptor exons (or different donors to a same acceptor). We can observe usage of splicing junctions is highly consistent within same ancestral haplotypes, while distinct between different ancestral haplotypes.

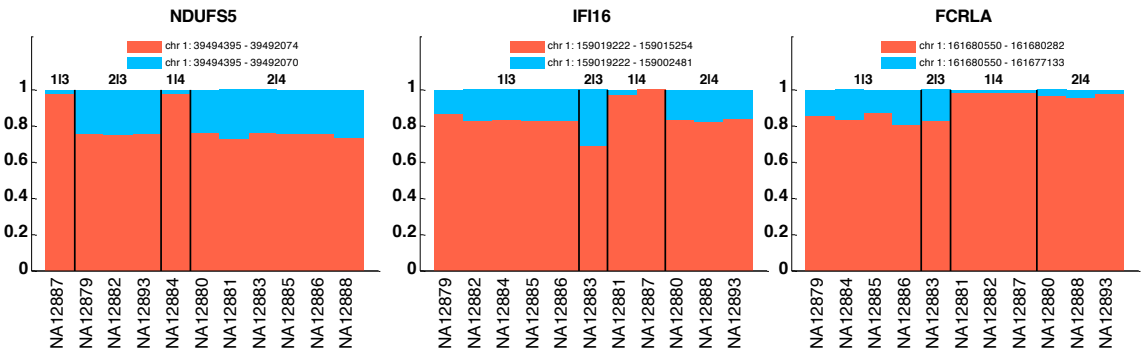

Figure S27. Examples of alternative splicing patterns determined by haplotype groups.

1, 2, 3 and 4 are paternal grandfather, grandmother and maternal grandfather and grandmother haplotypes, respectively. Explanations of segregation patterns are provided in Figure S26. Additional information about those genes is provided in Table S8.

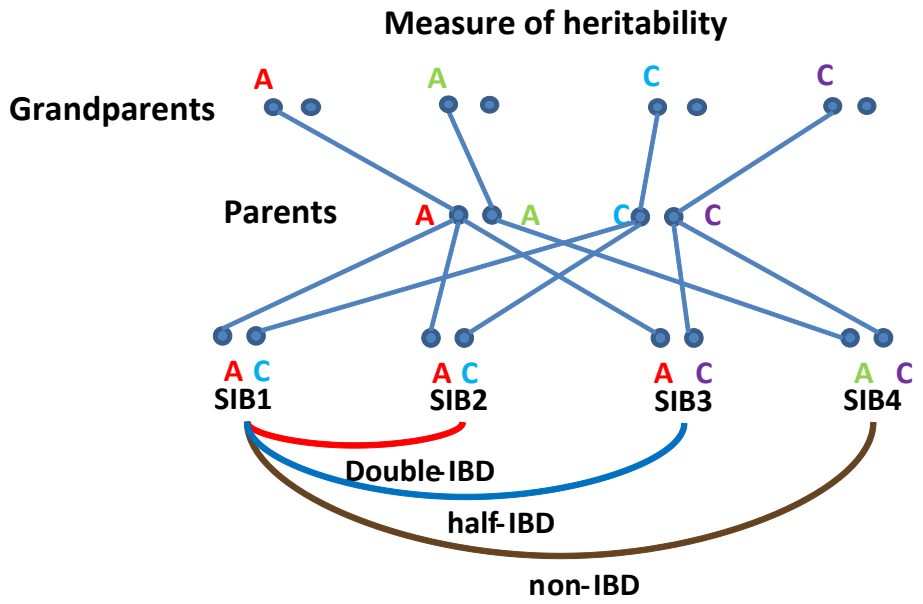

Figure S28. ASE heritability analysis.

Method of measuring heritability in family. IBD means Identical-by-descent, descending from the same ancestral haplotype. Sib1 and Sib2 are double-IBD siblings as they share both haplotypes. Sib1 and Sib3 are half-IBD as they share only one haplotype. Sib1 and Sib4 are non-IBD as they share neither of their haplotypes. We use each child as a reference and calculated the correlation of allelic ratios with their double-IBD, half-IBD and non-IBD siblings. We repeat this for each of 11 children.

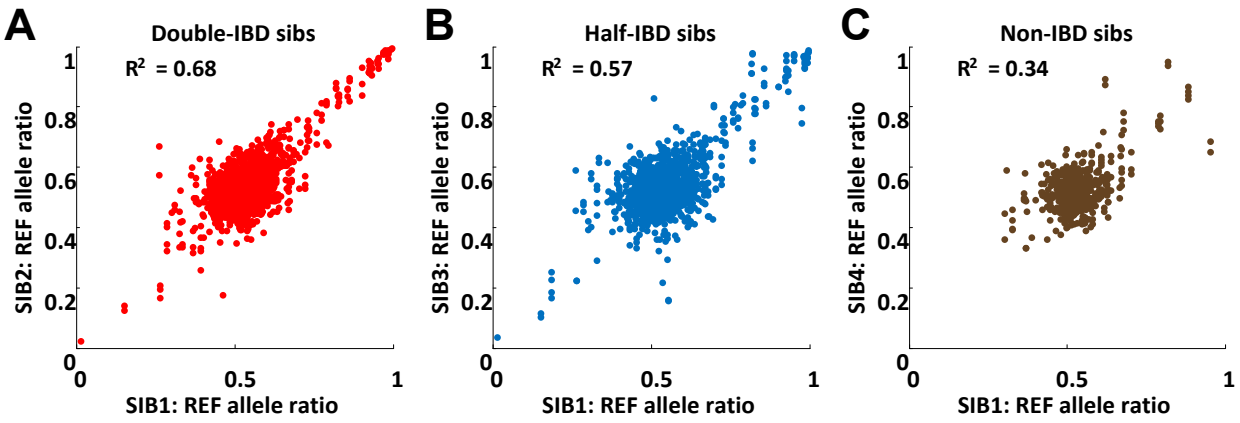

Figure S29. Allelic ratio correlation with siblings, using NA12879 as reference.

The figure shows measured correlation for each experiment. (A) Double-IBD, (B) half-IBD and (C) non-IBD are defined as sharing both, only one or neither haplotypes. To reduce random sampling noise, the result is based on sites of depths greater than 100.

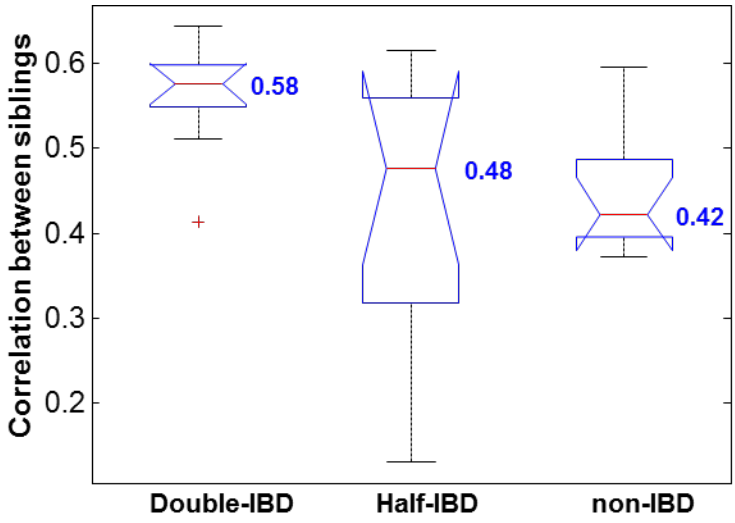

Figure S30. Allelic ratio correlation between different types of siblings.

From left to right, measured correlation between double-IBD, half-IBD and non-IBD siblings using each sibling as reference. Median correlation coefficients for double-IBD, half-IBD and non-IBD siblings are 0.58, 0.48, and 0.42 respectively.

Supplemental Tables

	Total segregating SNPs	MAF in 1000 Genomes European			
		≤ 0.05		≤ 0.01	
CEU family	2,936,403	345,232	11.76%	91,882	3.13%

Table S1. Number of variants segregating in the family.

Segregating variants among children are those variants that are heterozygous in at least one of the parents and both alleles are transmitted to the children. We define rare variants as variants with minor allele frequency below 0.01 (or otherwise specified in the paper) in the Phase 1 release of the 1000 Genomes Project European populations. As calling of rare variants is especially susceptible to genotyping errors, we use stringent Mendelian constraints to reduce these errors. We require all called variants to be completely consistent along the IBD inheritance across the whole family.

	NA12877	NA12885	NA12886	NA12891	NA12892
Sites genotyped in both LFR and original sequencing	1835141	1711513	1754935	1716804	1763045
Concordant sites % concordant	1834123 0.999445	1710681 0.999514	1753403 0.999127	1716076 0.999576	1762294 0.999574

Table S2. Genotypes confirmed with Complete Genomics Long Fragment Read[1] (LFR).

Comparison of genotypes between original Complete Genomics sequencing and LFR technology. We have five individuals in the family that were sequenced again using LFR technology which can generate molecular haplotypes. LFR can place 92% of heterozygous SNP into long ~500kb contigs. It has very high genotyping accuracy, with an error rate of 1 in 10Mb. Genotypes used in this study were found to be more than 99.9% concordant between original sequencing data and LFR data. Further, we confirmed called genotypes with both new sequencing data from both Complete Genomics Long Fragment Read[1] technology and from Illumina Platinum Genomes (Table S4).

		NA12877	NA12885	NA12886
Total phased het sites	PEDIBD	1913154	1903914	1922039
	LFR	1876780	1747734	1795687
	Overlap	1834122	1710681	1753402
	Phase concordant % concordant	1831234 0.998425	1709182 0.999124	1752028 0.999216

Table S3. Phasing confirmed with molecular haplotype by LFR.

Comparison of inferred haplotypes by Ped-IBD and molecular haplotypes generated by LFR technology. Phasing results are confirmed to be 99.9% concordant with molecular haplotypes generated by LFR technology.

	Variant sites genotyped by Complete Genomics	Variant sites genotyped by Illumina	Genotyped by both	Concordance
NA12889	3232336	2862355	2703495	0.9983
NA12890	3237426	2798769	2672122	0.9981
NA12891	3216575	2816832	2678531	0.9982
NA12892	3231991	2869083	2711768	0.9983
NA12877	3228441	2828846	2681250	0.9976
NA12878	3220524	2802242	2667261	0.9979
NA12879	3220425	2796242	2662748	0.9977
NA12880	3231151	2806653	2670442	0.9980
NA12881	3236471	2802998	2670859	0.9979
NA12882	3270994	2825547	2684441	0.9881
NA12883	3234742	2796940	2667074	0.9982
NA12884	3226121	2785306	2653732	0.9966
NA12885	3233256	2781640	2657075	0.9982
NA12886	3202270	2847260	2668380	0.9961
NA12887	3209866	2825088	2672224	0.9972
NA12888	3234370	2816855	2675737	0.9977
NA12889	3245741	2825098	2686108	0.9976

Table S4. Genotypes confirmed with Illumina Platinum Genomes.

Comparison of genotypes between Illumina Platinum Genomes and Complete Genomics sequencing data. The same CEU family samples were also sequenced by Illumina as part of Illumina Platinum Genomes. All 17 members were sequenced to 50x depth on a HiSeq 2000 system. We compared genotypes called by Complete Genomics and Illumina (passing genotype filter and quality score = 99). On average, Complete Genomics data cover more sites than Illumina and include the majority (> 95%) of Illumina sites. Genotype concordance between the two platforms at overlapping sites ranged from 0.9966 to 0.9983 across individuals.

	Tested genes	Number of eQTL genes	
eQTL genes	8,974 genes	π_1=0.078	~698
		FDR < 50%	274
sQTL genes	7,954 genes	π_1=0.079	~624
		FDR < 50%	261

Table S5. Linkage analysis of *cis*-eQTL: summary of eQTL and sQTL genes identified in the family.

Total number of eQTL genes π_1[3] and numbers of genes below FDR 0.5. Numbers of haplotypes blocks holding these eQTL or sQTL genes are also listed.

Gene expression quantification. We used the Tophat/Cufflinks to quantify expression levels of whole genes and each transcript from RNA-Seq data (Figure S2). We performed eQTL discovery using linear regression of gene expression levels with local haplotype blocks. We identified *cis* expression QTLs (*cis*-eQTLs) by restricting association to the haplotype block that contains the tested gene. We only considered protein-coding genes, and to minimize possible technical artifacts in quantification we also exclude all pseudogenes, all immunoglobulin and HLA genes where there is an increased potential for mapping biases and sequencing errors. We required an average FPKM greater than 2 and at least 3 individuals with FPKM greater than 1 for a gene to be tested. Setting this threshold, we tested 8,974 genes for eQTLs. For *cis*-splicing QTLs (sQTLs), we additionally require each gene to have two or more quantified alternative transcripts (N=7,954 genes).

eQTL and sQTL discoveries. To detect eQTLs in the family, we used a two-variable (paternal and maternal haplotypes) linear regression to test for gene expression ~ haplotype association. For each haplotype block, the two parental haplotypes of each child are encoded using two variables, *p* and *m*. The maternal haplotype m_i of a child *i*, for example, is either 0 or 1, depending on which of the two possible maternal alleles is present. Then, an expression trait is regressed as the summation of effects of two parental haplotypes, $T_i \sim \mu + \beta_j p_i + \beta_k m_i$, where T_i is the trait of individual i, the effects of two parental alleles k and j are expressed by β_j and β_k and μ is the intercept. For sQTL, we selected the most significant *p*-value among all transcripts for each gene. *P*-values are further adjusted using permutation as described below.

Empirical *p*-values were generated using permutation by swapping phenotypes across individuals. We performed 10000 permutations at each gene and computed *p*-values by counting how many times permuted *p*-values fell below the nominal *p*-value.

To quantify effects for common variants, we used linear regression to test common variants among 373 unrelated European individuals from Geuvadis study[6]. To ensure

discoveries in the population were relevant to the family, we only test variants which were also polymorphic in the family.

Sample size	50	100	150	200	250
50	312	257	243	247	252
100	257	321	243	245	254
150	243	243	308	241	245
200	247	245	241	315	249
250	252	254	245	249	366

Table S6. Effect of different discovery panel sizes: number of large effect β genes given different discovery panel sizes.

On the diagonal are numbers of genes with family effect sizes > 95% CI population effect sizes. Off diagonal cells show their intersections.

MAF	Distance to TSS	TF binding + DNase peak	PhyloP score	Motif	# of genes	% eQTL in family	% eQTL in population
< 0.01	100 kb	-	-	-	7912	0.0858	0.1662
< 0.01	100 kb	Yes	-	-	3123	0.0990	0.1775
< 0.01	100 kb	Yes	> 1	-	367	0.4577	0.1759
< 0.01	100 kb	Yes	> 1	Yes	135	0.5627	0.2049
< 0.01	5 kb	-	-	-	1525	0.1099	0.1807
< 0.01	5 kb	Yes	-	-	386	0.1815	0.1743
< 0.01	5 kb	Yes	> 1	-	41	0.5151	0.1622
< 0.01	5 kb	Yes	> 1	Yes	17	0.8999	0.2000
< 0.01	100 kb	-	-	-	7912	0.0869	0.1284
< 0.01	100 kb	Yes	-	-	3123	0.0968	0.1371
< 0.01	100 kb	Yes	> 3	-	88	0.8303	0.1688
< 0.01	100 kb	Yes	> 3	Yes	30	0.9528	0.2400
> 0.01	100 kb	-	-	-	8312	0.0785	0.1647
> 0.01	100 kb	Yes	-	-	8186	0.0757	0.1652
> 0.01	100 kb	Yes	> 1	-	6110	0.0820	0.1676
> 0.01	100 kb	Yes	> 1	Yes	2456	0.0965	0.1688
> 0.01	5 kb	-	-	-	7525	0.0801	0.1684
> 0.01	5 kb	Yes	-	-	6092	0.0941	0.1755
> 0.01	5 kb	Yes	> 1	-	1359	0.1676	0.1833

> 0.01	5 kb	Yes	> 1	Yes	413	0.1616	0.1777
> 0.01	100 kb	-	-	-	8312	0.0781	0.1647
> 0.01	100 kb	Yes	-	-	8186	0.0762	0.1652
> 0.01	100 kb	Yes	> 3	-	1542	0.1012	0.1717
> 0.01	100 kb	Yes	> 3	Yes	457	0.1504	0.1762

Table S7. Prediction of eQTLs at rare variants given annotation: proportion of genes being an eQTL given a regulatory variant near TSS.

We assessed the utility of different variant annotations in predicting eQTLs. We incrementally add annotations for minor allele frequency, distance to TSS, transcription factor binding, DNase sites, conservation score and transcription factor motif.

We selected those annotations which are previously found to be informative in predicting eQTLs: distance to TSS, transcription factor binding, DNaseI sites, conservation and transcription factor motifs. Encode transcription factor binding and DNaseI hypersensitivity peaks were obtained from RegulomeDB database[4]. Conservation scores using PhyloP[7] (phyloP100way) software were downloaded from the UCSC genome browser (genome.ucsc.edu). Motif disrupting sites were downloaded from HaploReg database (v2)[8].

Gene name	transcript ratio *p*-value	splice junction *p*-value	function
NDUFS5	0.0030	2.51E-09	neurological disorders
IFI44L	0.0020	1.33E-08	response to viral infection
IFI16	0.0010	1.91E-06	response to viral infection
FCRLA	0.0001	9.30E-05	B-cell development

Table S8. Examples of sQTL genes.

These genes are identified as sQTL genes by both transcript ratios and splice junction read ratios. Transcript ratio *p*-values are based on quantification of transcript abundances by Cufflinks, splice junction *p*-values are based on quantification of splice junction reads by JunctionTK. Two methods are in general concordant with each other. Segregating patterns of those genes are illustrated in Figure S26 and Figure S27.

GENE	GWA SNP ID*	chr	bps	Distance of nearest rare regulatory variant to TSS (bps)	Trait
B4GALT1	rs10813960-T	9	33180362	-	Urate levels
BAK1	rs210134-A	6	33540209	-	Platelet counts
PHTF2	rs12234571-C	7	77549906	-	Obesity-related traits
BAK1	rs9469457-A	6	33489882	-	Obesity-related traits
TCFL5	rs17854409-G	20	61491494	-	Obesity-related traits
TRAF3IP2	rs3851228-T	6	111848191	123850	Inflammatory bowel disease
PHTF2	rs848452-?	7	77596812	-	Dental caries
EPB41L5	rs13401620-A	2	120513133	-	Breast size
BAK1	rs210142-C	6	33546837	-	Chronic lymphocytic leukemia
INSIG1	rs10263087-C	7	154970469	49137	Formal thought disorder in schizophrenia
BAK1	rs210134-G	6	33540209	-	Mean platelet volume
BAK1	rs210134-G	6	33540209	-	Platelet counts
ENTPD6	rs1044573-A	20	25206654	128236	Allergic rhinitis
ABHD12	rs7267979-G	20	25298087	-	Liver enzyme levels (alkaline phosphatase)
PLCL2	rs9821630-G	3	16970938	-	Multiple sclerosis
ZKSCAN5	rs11761528-T	7	99118801	-	Dehydroepiandrosterone sulphate levels
PLCL2	rs1372072-A	3	16955259	-	Primary biliary cirrhosis
TRAF3IP2	rs33980500-T	6	111913262	123850	Psoriasis
TRAF3IP2	rs33980500-T	6	111913262	123850	Psoriatic arthritis
TRAF3IP2	rs240993-A	6	111673714	123850	Psoriasis
PRMT7	rs7197653-C	16	68383047	-	Magnesium levels
DHCR7	rs12785878-?	11	71167449	158269	Vitamin D insufficiency

NADSYN1	rs12785878-?	11	71167449	158510	Vitamin D insufficiency
CST3	rs911119-?	20	23612737	6080	Chronic kidney disease
PHTF2	rs6465825-C	7	77416439	-	Chronic kidney disease
ALDH7A1	rs13182402-G	5	125918148	-	Osteoporosis
BAK1	rs210138-G	6	33542538	-	Testicular germ cell tumor
IL16	rs7172689-?	15	81533695	-	Inattentive symptoms

*Identified as an eQTL in[6], but not polymorphic in family

Table S9. Family-specific cis-eQTL modifying complex trait genes.

We assessed the number of GWA loci which were influenced by large-effect family eQTLs. Here, we identified family-specific eQTL for genes in the NHGRI GWA catalog17. We tested all those GWA SNPs in Geuvadis data and select those which are eQTLs (π_1[3]~ 0.3, 315 genes at an FDR < 0.05). We then sub-selected those where the associated GWA SNP was not polymorphic in the family, so the GWA SNP is not causing the eQTL in the family, and another regulatory variant for the same gene is likely to be present. This highlights the potential for rare regulatory variants manifesting as family-specific eQTLs to be modifying important complex disease associated genes. The table lists examples of large-effect (CI > 0.80) family eQTL that influence GWA genes. The GWA SNP for the trait is determined to be an eQTL SNP (FDR 0.05) but not polymorphic in the family. Rare regulatory variants are defined as those within Encode TF binding + DNase peaks, MAF < 0.01 and PhyloP > 0, within 200kb near TSS.

GENE	GWA SNP ID	chr	bps	Distance of nearest rare regulatory variant to TSS (bps)	Trait
IRF5	rs12531711-G	7	128617466	50878	Primary biliary cirrhosis
IRF5	rs10488631-C	7	128594183	50878	Primary biliary cirrhosis
IRF5	rs10488631-C	7	128594183	50878	Rheumatoid arthritis
IRF5	rs729302-A	7	128568960	50878	Systemic lupus erythematosus
IRF5	rs12531711-G	7	128617466	50878	Systemic lupus erythematosus
IRF5	rs10488631-C	7	128594183	50878	Systemic lupus erythematosus
IRF5	rs4728142-A	7	128573967	50878	Systemic lupus erythematosus
IRF5	rs12537284-A	7	128717906	50878	Systemic lupus erythematosus
IRF5	rs10488631-C	7	128594183	50878	Systemic lupus erythematosus
IRF5	rs10488631-C	7	128594183	50878	Systemic sclerosis
IRF5	rs4728142-A	7	128573967	50878	Ulcerative colitis
IRF5	rs4728142-A	7	128573967	50878	Ulcerative colitis
NAPRT1	rs2290416-?	8	144657600	46423	Attention deficit hyperactivity disorder
NT5E	rs494562-G	6	86117129	40975	Metabolic traits
TCF7	rs756699-A	5	133446575	63462	Multiple sclerosis

Table S10. Examples of rare regulatory variants influencing GWA genes.

The GWA SNP for the trait is determined to be an eQTL SNP (FDR 0.05). Rare regulatory variants are defined as those within Encode TF binding + DNase peaks, MAF < 0.01 and PhyloP > 3, within 200kb near TSS.

References

1. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., et al. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487, 190-195.
2. Cheung, V.G., Nayak, R.R., Wang, I.X., Elwyn, S., Cousins, S.M., Morley, M., and Spielman, R.S. (2010). Polymorphic cis- and trans-regulation of human gene expression. PLoS biology 8, 14.
3. Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 479-498.
4. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome research 22, 1790-1797.
5. Consortium, E.P., Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
6. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511.
7. Cooper, G.M., Stone, E.A., Asimenos, G., Program, N.C.S., Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome research 15, 901-913.
8. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research 40, D930-934.