Published in final edited form as:
Hum Genet. 2014 June ; 133(6): 727–735. doi:10.1007/s00439-014-1446-0.

Determining causality and consequence of expression quantitative trait loci

A.J. Battle[1,2,*] and **S.B. Montgomery**[2,3,*]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, 21218, USA

[2]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

[3]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA

Abstract

Expression quantitative trait loci (eQTLs) are currently the most abundant and systematically-surveyed class of functional consequence for genetic variation. Recent genetic studies of gene expression have identified thousands of eQTLs in diverse tissue types for the majority of human genes. Application of this large eQTL catalogue provides an important resource for understanding the molecular basis of common genetic diseases. However, only now has both the availability of individuals with full genomes and corresponding advances in functional genomics provided the opportunity to dissect eQTLs to identify causal regulatory variants. Resolving the properties of such causal regulatory variants is improving understanding of the molecular mechanisms that influence traits and guiding the development of new genome-scale approaches to variant interpretation. In this review, we provide an overview of current computational and experimental methods for identifying causal regulatory variants and predicting their phenotypic consequences.

Introduction

Characterizing the functional impact of human genetic variation is essential for understanding the molecular underpinnings of inherited disease risk. While human genome sequencing has enabled rapid and efficient cataloging of tens of millions of genomic variants, for the majority of these, we know little about their functional impact. This is particularly true for mutations in non-coding regions. Genetic studies of gene expression provide one means to interpret the functional impact of non-coding variants; these studies have identified expression quantitative trait loci (eQTLs) in different populations[1–3], tissues[4–8] and in response to different stimuli[9,10]. However, due to the presence of linkage disequilibrium and often incomplete resolution of genetic variation, the majority of eQTLs only inform the presence of some causal variant and not the precise causal variant itself. Now, in the wake of advances in genome and functional genomics sequencing, there is increased ability to directly identify specific causal variants that modulate gene expression. Such advances, however, require both the development of computational approaches that

*Corresponding authors: smontgom@stanford.edu or ajbattle@cs.jhu.edu.

integrate genomes with diverse functional genomic and population genetic data and the application of new high-throughput experimental approaches that validate subsequent predictions. These approaches and data offer the potential to expose the genomic properties of causal non-coding variants and interpret variant impact and phenotypic consequences from genome sequence alone. This need is particular acute as recent surveys of genetic variation in human population have highlighted extensive impactful rare variation whose effect is not well captured through association alone[11–13]. To begin to more completely understand how to infer causality and consequence of non-coding variation, we describe in this review recent statistical and experimental advances in characterizing causal non-coding variants after eQTLs have been identified.

Using expression quantitative trait studies to identify causal regulatory variants

Detecting causal non-coding variants through fine-mapping

Many eQTL studies have relied on genetic data obtained through genotyping arrays. While such data provides the means to detect eQTLs, they are limited in resolution of potential causal variants – the specific variants that underlie eQTLs. Now with the growing availability of high-density genotyping and genome sequencing data there is increased likelihood to directly observe genotypes for all candidate causal non-coding variants. Alternatively, variants not measured directly can be indirectly inferred through cost-effective imputation strategies using reference panels such as the 1000 Genomes Project[14] or HapMap[15]. The principle of imputation is that by exploiting patterns of linkage disequilibrium within populations, the genotypes of unobserved sites can be inferred. To achieve this, several tools are available including Impute2[16], Beagle[17] and Minimac[18]. Many of which come packaged with supporting haplotype data from reference panels. However, there are important caveats with imputation including low accuracy for rare variants, computational time and adequacy of the reference panel[19]. Once such approaches have been applied, the working hypothesis is that a candidate causal non-coding variant underlying an eQTL will be the individual variant that exhibits the best fit to expression level of all variants in the region (Figure 1).

In practice, it may be challenging to resolve a single causal variant through association alone, as several candidate variants may be in high linkage disequilibrium and exhibit equal fit to expression level[20]. In addition, there may be in fact more than one regulatory variant with a causal effect. And ultimately, even if one variant exhibits the strongest association, it may not actually be the causal variant and merely reflect the composite signal of another causal variant plus noise or inaccurate estimation of association from a small sample size. In our own work, we have compared the relative discovery of *cis*-eQTLs between high-density genotyping of one million SNPs from the HapMap3 project versus 5–7 million SNPs from the 1000 Genomes Project in 60 individuals[21]. We discovered that both platforms yielded an equal number of *cis*-eQTL across permutation thresholds. Likewise, Liang *et al*. explored relative power gains from imputation by comparing *cis*-eQTL discoveries from the Illumina 300k chip with imputed HapMap2 and 1000 Genomes data in two separate cohorts (N>200 individuals)[22]. In their work, they observed power increases of 6–7% and 5–8% at an FDR

of 5%, respectively. This suggests that total number of *cis*-eQTLs do not dramatically increase through imputation. This may be expected, as lower density platforms are sufficient for capturing the majority of common human haplotypes. However, in both studies the power of imputation aided to significantly refine identification of the variant with the strongest association. Liang *et al*. demonstrated localization of a *cis*-eQTL signal in TIMM22 to the 3' UTR by imputation of 1000 Genomes genotypes. We observed that 80% of the variants exhibiting the top associations for eQTLs would not have been identified with HapMap genotypes alone and were only discovered with full sequencing data[21]. A similar comparison made by Gaffney *et al.* noted that 20% of detected eQTLs exhibited a 1000 Genomes SNP with a *p*-value at least one order of magnitude lower than the best HapMap SNP[23]. These results indicate that low-density genotypes, while equally good at identifying genes with an eQTL, may miss the most likely causal variants, and that high-density genotyping is bringing us closer to capturing these variants.

Replication of eQTLs across studies and populations to find causal non-coding variants

Replication of eQTLs across studies and populations can potentially increase the confidence that a studied locus harbors a causal variant. Unlike genome-wide association studies, however, many eQTL studies do not explicitly employ replication designs. Of the few studies that have tested replication, key factors that influenced replication rate include statistical power in both discovery and replication panels, association strength of the discovery, the eQTLs distance to the associated gene's transcription start site and the confounding presence of spatiotemporally-distinct or spurious eQTLs due unmatched technical, biological or environmental factors between the discovery and replication panels. One example where replication has led to functional fine-mapping of causal variants was performed by Innocenti *et al.*; the authors first investigated the reproducibility of eQTL results in primary liver and reported that up to 67% of *cis*-eQTL were replicated, and among those factors which correlated to replication rate were association strength, proximity to the transcription start site, the presence of array hybridization artifacts and the mean and variance of the gene's expression level[24]. However, by subsequently taking reproducible eQTLs, they were able to select and identify causal-variant containing sequences that exhibited *in vitro* functional effects for 3 of 14 genes tested. Analysis of eQTL replication between two mouse cross also reported high replication of *cis*-eQTLs (63%) yet low replication of *trans*-eQTLs (18%; LOD>4.3) with the replication rate increasing with association strength[25]. For causal regulatory variant detection, both studies suggest that proximity to transcription start size, well-powered study designs and well-matched discovery and replication panels may aid in localizing causal loci. For *trans*-eQTLs on the other hand, limited power and low-replication suggest that functional follow-up is best applied for eQTLs that demonstrate replication. Indeed, Westra *et al.* recently conducted a meta-analysis of seven eQTL studies comprising 5311 individuals with a replication panel of another 2775 individuals to systematically identify *trans*-eQTLs[26]. In total, they were able to identify and replicate 103 independent loci equivalent to 223 variants of 4542 tested. Replicating *trans*-eQTLs were enriched for miRNA binding sites and blood-specific enhancer regions exposing the likely causal mechanisms of non-coding variants that exert long-range effects.

Replication of eQTLs across multiple populations further confers a unique advantage for localizing causal non-coding variants. As tightly-linked variants may be indistinguishable in their association strength, integrating different population data can increase causal variant localization by exploiting population-specific differences in LD structure to effectively breaking up LD blocks and refine an expression or trait association signal. Indeed, this approach was evaluated by Zaitlen *et al.* were they highlighted that refinement of an association signal using multi-population data reduced the number of potentially causal variants that needed to be assayed[27]. However, they cautioned that the best study design was not always a multi-population design as, for population-specific causal variants, individual populations are maximally powered. In fact, we have seen that by integrating multiple populations for eQTL discovery, we find replicated effects are more tightly-distributed around the transcription start site[1]. This suggests that causal regulatory variants are more closely coupled to a gene's proximal regulatory machinery than informed from single population analyses.

Mapping causal regulatory variants with allele-specific expression

Another emerging route for replicating eQTLs and resolving causal non-coding variants is through analysis of allele specific expression (ASE). Measurements of ASE are now routinely accessible from RNA-sequencing data and provide qualitatively different information from traditional eQTL approaches; eQTL analyses are based on genotypic association with total expression levels across individuals while ASE is measurable between alleles within an individual. Detecting ASE is based on selecting protein-coding heterozygous sites and subsequently measuring the skew in read counts for each allele; using this approach one is able to identify patterns of ASE which may inform the presence of a *cis*-linked regulatory variant within an individual. Indeed, targeted studies of allele-specific expression have revealed that the majority of ASE can be explained by nearby genetic variants[28]. We have also applied sharing of ASE from RNA-Seq across individuals to detect potentially causal regulatory variants by focusing on just those sites that share heterozygosity in the presence of the ASE effect[21]. This approach allowed us to identify variants close to genes which were highly correlated to the allelic effect. Similarly, in a recent study with 922 individuals, we tested for association between heterozygosity of non-coding variants and allelic imbalance to confirm the regulatory impact of a large number (641) *cis*-eQTL SNPs[29]. Further demonstrating the specificity of ASE in detecting potential causal variants, Lappalainen *et al*. recently demonstrated that variants which are jointly best associated to expression level and ASE are significantly enriched in diverse epigenetic annotation from ENCODE[30]. Likewise, a statistical method developed for RNA-Seq data has demonstrated increases in discovery power of eQTLs when both expression level and ASE are tested together[31]. These results suggest that the combination of genotypic and allelic association significantly aids in refining causal non-coding variants. However, while testing association using ASE is intuitively straightforward it has its own limitations; it requires that the number of heterozygotes testable for ASE to be sufficient, the ASE effect is detectable from overlapping read data and an assumption of a single or few causal regulatory variants.

Statistical methods for identifying causal regulatory variants

The strategies described above offer the potential to refine the localization of causal variants directly from expression data, but in most cases full disambiguation will not be possible from eQTL associations alone. For instance, Gaffney *et al.* report, in an eQTL study leveraging dense genotyping that for 80% of the significant eQTLs identified, ambiguity remained between at least two SNPs with p-values differentiated by less than one order of magnitude[23]. In such cases, further disambiguation may be possible through computational and statistical methods, specifically methods using additional sources of data and genomic properties of the associated loci. Analysis of sequence conservation is one strategy for assessing which regions of the genome are likely to have deleterious consequences, and has been applied to non-coding regions of the genome[32,33]. Overlap has been demonstrated between conserved regions and eQTLs[34], and between conserved regions and regulatory elements such as enhancers[35,36], but the utility of conservation scores in disambiguating individual causal regulatory variants remains complex as not all important regulatory elements are well-conserved[37,38]. A second category of approaches includes computational methods that predict consequences of sequence variation on specific regulatory mechanisms. Based on high-throughput SELEX, ChIP sequencing data, and protein binding microarrays, position weight matrices (PWMs) have now been described for many human transcription factors supporting predictions of the impact of genetic variants on TF binding genome-wide[39–41]. *In-silico* methods have also been developed for prediction of other potential regulatory effects of sequence including RNA binding protein motifs[42] and DNAshape[43]. Complementing these approaches there is a growing wealth of high-throughput data providing regulatory element annotation across a variety of tissues, developmental stages, and populations[44]. By intersecting eQTL data with high-throughput data or the results of computational analysis, we can identify sub-regions or even specific nucleotides within an eQTL-associated region that have specific evidence to support a functional role. Tools and databases that provide diverse regulatory element annotations to aid in assessment of functional roles include HaploReg[45], RegulomeDB[46], and ORegAnno[47]. Furthermore, enrichment of eQTLs within TF binding sites and other putative regulatory elements has demonstrated that such annotation is informative for detecting causal elements[23,48–50] and methods that integrate these sources of annotation to predict causal regulatory variants from eQTL data have been developed[23,51,52]. While specific modeling choices and training methods vary, each of these methods is built on a regression model which estimates the regulatory impact of each SNP based on available genomic annotations. The hierarchical Bayesian model of Gaffney *et al.* demonstrates good predictive accuracy; testing their method on eQTLs with unambiguous association signals, this model ranks the best SNP among the top 10% of candidates in over 70% of test cases[23]. The results obtained with our own method, LRVM, demonstrated that integration of diverse annotation allows prediction of impact on ASE, which as discussed above, may bring us closer to true causal variants[29]. Methods trained on other data types may also be applied to interpretation of eQTLs. For instance, a recent tool called GWAVA[53] uses random forests based on genomic annotation features, but is trained on known regulatory variants implicated in disease from the Human Gene Mutation Database[54] to predict functional variants in non-coding regions of the genome. Challenges remain, however, in the assessment of accuracy of causal regulatory

variant models; in general, the research community does not have access to large-scale gold standard data with which to evaluate their performance.

Experimental methods for identifying and validating causal regulatory variants

While both expression data in the form of various QTLs and allelic effects can significantly aid in informing the presence and specific location of a regulatory variant, there are a growing number of novel assays which are illuminating the relationship between single variants and gene expression. Assays measuring sequence-specificity of transcription factor binding (and existing databases) can inform causality of eQTLs if a particular TF is suspected to be involved in the underlying mechanism[39,41]. Here, we cover in more detail two classes of recently developed assays that are able to inform a large fraction of eQTLs, either through direct measurement of changes to gene expression, or through measurement of epigenetic changes.

Epigenetic assays

ChIP-Seq data from diverse cell types and stages has been essential to highlighting broad or punctate regions containing transcription factor binding sites and allele-specific differences in transcription factor binding[44,55,56]. Such differences have been informative in assignment of function to disease-associated variants[57]. Furthermore, as for ASE measurements using RNA-Seq, by assaying which heterozygous sites exhibit a skewed balance of mapped ChIP-Seq reads for each allele, one is able to directly select putative causal regulatory variants. Here the hypothesis is that allele-specific binding (ASB) of a transcription factor will causally map to an allele-specific differences in expression. Indeed, Rozowsky *et al*. demonstrated concordant patterns of ASB and ASE[56]. However, it remains difficult to establish a causal link between ASB and effects on gene expression. Further, less than half of the 1500- 2000 human transcription factors have an experimentally characterized DNA binding motif[39,58,59], and antibody efficiencies for many TFs remain variable or poor. Considering these challenges, it is still technically infeasible to definitively survey the consequences of sequence variation on binding for all known transcription factors. Complementing ChIP-Seq assays, DNaseI hypersensitivity-sequencing offers a non-specific approach to identifying potential causal regulatory variants. For this assay, allele-specific binding and genetic association within DNaseI footprints can aid in pinpointing a potential regulatory element when the bound transcription factor is itself unknown. Indeed, when DNaseI hypersensitivity data was assessed as a quantitative trait in 70 Yoruban individuals, it was estimated that over half of all eQTLs were driven by genotypic differences in DNaseI sensitivity[49]. Here, the combination of allelic analyses in DNaseI footprints and transcription factor binding site prediction using PWMs supported the hypothesis that TF binding often drives changes to chromatin state that mediate effects on gene expression. Additional studies of TF binding, histone modification, and chromatin state provide further evidence of this causal relationship, using population, family and allele-specific analysis (rather than QTL detection) of each phenotype in a limited number of individuals[60–62]. However, only 16% of all DNaseI QTLs themselves were mapped to a change in gene expression suggesting that only a minority of binding differences definitively influence

expression. Partially explaining this apparent incomplete overlap of dsQTLs with variants affecting gene expression, in this analysis, Degner *et al* demonstrate that influence of dsQTLs on gene expression depends on genomic context including distance to the nearest TSS, intervening CTCF binding sites, and association of the variant with methylation (meQTLs)[63]. Uncertainty still remains, however, in completely determining the influence of epigenetic changes on expression of nearby candidate genes, from identifying the correct genes for potentially distant regulatory elements, to potential variability in strength and direction TF effect, to combinatorial regulatory mechanisms. Identification of dsQTLs, meQTLs or other epigenetic effects for an eQTL of interest therefore provides evidence of mechanism but not conclusive validation.

Massively parallel reporter assays

Advances in massively-parallel reporter assays (MPRA) through sequencing are achieving quantitative and high-throughput readouts of the impact of regulatory elements in vivo. One such study synthesized known Crx-bound ChIP-seq regions and placed them in front of a minimal promoter, reporter gene and unique barcode to assess their relative activities in explanted newborn mouse retinas[64]. Through direct sequencing of the barcode the authors were able to determine the relative ability of their synthesized sequences in enhancing expression. In a large-scale MPRA study of 2,000 human enhancers, Kheradpour *et al.* verified sequence-specificity of both predicted repressors and predicted activators[65]. Using two human cell lines, their results also confirm tissue-specificity of enhancer activity. However, a challenge with the MPRA approach is that it requires the synthesis of assayed targets. Complementing this approach, an assay called STARR-seq facilitates high-throughput screening by measuring randomly sheared DNA's ability to promote its own expression[66]. By placing random sequences downstream of a minimal promoter, Arnold *et al.* assayed which sequences enhance their own expression in different cellular contexts. More complex application of these approaches is also unlocking new information about transcription factor interaction and organization. Smith *et al.* recently created a library of ~5,000 synthetic promoters containing 12 liver-specific transcription factors to identify features of transcription factor organization in HepG2 cells[67]. Among their discoveries was the observation that there were multiple, but non-generalizable arrangements of motifs which support strong and weak expression. If such observations are themselves generalizable, it would argue that the impact on gene expression by causal variants will remain challenging to infer from sequence context. Furthermore, an important caveat to these studies is that they all work in a heterologous context and putative regulatory elements may in fact be endogenously silenced. Despite this however, they provide a quantitative and cell-type specific readout of regulatory activity that maps well to enhancer-associated histone modifications and offers new potential for the assessment of regulatory architecture and causal regulatory variant impacts.

Genome editing

While MPRAs and related approaches allow high-throughput testing of sequence variants using reporter constructs, other methods are appropriate for testing variants in their own genomic context *in vivo*. Previously, gene knockout or knockdown experiments have been employed to validate the regulatory effects of entire genes, for example providing evidence

to support *trans*-eQTLs[68]. More recently, new high-precision methods of genome editing have become available, including Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)[69] and transcription activator-like effector nucleases (TALENs)[70]. As a result, editing of single loci to introduce specific sequences into the genome (e.g. mutations, GFP), using Cas9 together with a homologous recombination donor bearing the mutation of interest, is now straightforward[71–73]. Modified cell lines can be assayed for gene expression using targeted PCR or genome-wide methods, providing validation for both *cis*- and *trans*-eQTLs. In addition, epigenetic assays and other forms of cellular phenotyping could also be applied to provide further evidence of the specific mechanisms relevant to each tested eQTL. For example, in a recent study Bauer *et al.* used TALENs to disrupt an intronic enhancer in *BCL11A*, and demonstrated changes to expression of *BCL11A* along with alteration to the predicted target phenotype (increased embryonic globin protein levels)[74]. While validation of eQTLs through CRISPR and related techniques has not yet been applied on a large scale, this method holds great promise for directly testing the effects of individual regulatory variants on gene expression and other cellular traits.

Connecting regulatory variants to traits

Fully characterizing regulatory variation would include identifying the downstream consequences to the cell and organism. In this discussion, we focus primarily on the effects of regulatory variation on *complex* traits. For these traits the majority of associated variants occur outside of coding regions of the genome, suggesting that regulatory variants play a significant role[75,76]. Furthermore, through integrated analyses of eQTLs with trait-associated variants it has been possible to identify specific causal mechanisms. Indeed, it has been demonstrated that trait-associated variants are enriched for eQTLs[77], with *trans*-eQTLs shown to be particularly enriched for GWAS variants[78]. In addition RNA-sequencing has recently broadened the scope of eQTL analysis and thus the potential for investigating disease variants, enabling the inclusion of genetic effects on splicing, novel transcripts, alternative polyadenylation and other expression phenotypes like non-coding RNAs not typically available in eQTL studies based on microarrays. For instance, in a recent study, we identified 159 known disease variants with evidence of splicing QTL associations ($p < 1e\text{-}7$)[29]. However, the simple overlap between eQTL signal and a disease-associated locus cannot establish a causal relationship, due to confounding effects including linkage disequilibrium and correlated environmental factors. Here, Gagneur *et al.* have discussed the possible pathway relationships that may underlie eQTLs and trait-association and demonstrated that causal effects on higher-level traits are more likely to arise for certain classes of eQTLs[79]. Specifically, they found that environment-dependent eQTLs are much less likely to reveal causal genes than eQTLs shared across multiple environments. To address LD, statistical approaches have been developed aiming to determine the overlap between eQTLs and GWAS signals by measuring the correlation of the relative association signals across multiple proximal and partially-linked markers[80]. Furthermore, we have recently developed a test based on ASE to integrate both rare and common regulatory effects underlying trait-associated variants[81]. The hypothesis of this test is that heterozygotes for trait predisposing variants will exhibit more ASE than homozygotes thereby indicating an enrichment of either single or multiple distinguishing causal variants, where the controlled

comparison of alleles within an individual is less sensitive to environmental confounders than eQTLs based on expression levels across individuals. Additionally, the intersection of disease variants with tissue-specific eQTL data[4,5,82] may allow us to pinpoint the specific cell type or tissue where a disease variant acts or has the largest effect, and improve our ability to design appropriate follow up experiments. Thus, an increasingly complete understanding of disease variants may be possible, beginning with the simple connection suggested by existing eQTL data, combined with methods that predict or experimentally validate causal regulatory mechanism described in the previous two sections. However, while many trait-associated variants exhibit as eQTLs, the majority of known eQTLs have not yet been demonstrated to be associated with any disease.

With the prevalence of eQTLs, evidently even among common genetic variants, what are the consequences of this variation to the organism, and specifically to human health? The measurement of additional intermediate phenotypes, including histology reports and cellular traits such as protein levels[83–85], may shed light on these questions. Initial studies of individual variation in protein levels using mass-spectrometry indicate that eQTLs are likely to manifest as protein QTLs as well[85]. Building on this complexity to measure the impact of these effects on cellular and endophenotypes will remain an important and ongoing challenge. Ultimately, direct validation from methods including MPRAs[64,65] and genome editing assays[69] measuring both expression and additional downstream traits will provide the strongest evidence of causality.

Conclusions and future directions

We are currently at the point where we can refine eQTLs significantly based on high-density genotyping, population, tissue, condition-specific expression data and diverse high-throughput functional genomic data. Application of these approaches requires ongoing advances in integrative computational methods. However, high-throughput reporter assays also suggest that there will be limits to the types of regulatory architectures that can be easily predicted. For instance, beyond the interaction of specific transcription factors, the mechanistic properties of long-range effects like locus-control regions, enhancers or *trans*-regulators remain to be explored and integrated into a more complete model of gene regulation. Challenges with interrogating the complete developmental and condition-specificity of eQTLs due to the vast number of testable environmental perturbations and the diversity of cell types, of which many are routinely inaccessible, remain to be systematically encapsulated to reveal the most impactful gene regulatory networks and variations influencing traits and disease. The diversity of regulatory effects that are being elucidated also remains to be better connected to disease-associated variants as increasingly variants are being associated to alternative splicing, RNA degradation, splicing efficiency, poly-adenylation sites and miRNA recognition elements; each of which will add to the complexity of understanding gene regulation. Furthermore, as rare variants are increasingly recognized as abundant in human populations, building statistical and experimental systems that can integrate the impact of both rare and common alleles remain to be developed. It is expected that given current human population size and mutation rate, any regulatory variant that does not dramatically impact an individual's fitness, will in fact be present in some individuals suggesting that the ultimate goal is understanding the potential function of every

single non-coding base in the genome. However, among these challenges many exciting innovations remain to be developed including new approaches for integrating diverse expression and epigenomic datasets, advancements to the characterization of genetic regulatory networks and *trans*-eQTLs and ultimately, building models and experimental tools which can identify and integrate impactful non-coding variants to predict an individual's genetic disease risk.

References

1. Stranger BE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012; 8:e1002639. [PubMed: 22532805]
2. Spielman RS, et al. Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet. 2007; 39:226–231. [PubMed: 17206142]
3. Storey JD, et al. Gene-expression variation within and among human populations. Am J Hum Genet. 2007; 80:502–509. [PubMed: 17273971]
4. Dimas AS, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009; 325:1246–1250. [PubMed: 19644074]
5. Nica AC, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011; 7:e1002003. [PubMed: 21304890]
6. Myers AJ, et al. A survey of genetic human cortical gene expression. Nat Genet. 2007; 39:1494–1499. [PubMed: 17982457]
7. Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452:423–428. [PubMed: 18344981]
8. Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]
9. Barreiro LB, et al. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A. 2012; 109:1204–1209. [PubMed: 22233810]
10. Grundberg E, et al. Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. PLoS Genet. 2011; 7:e1001279. [PubMed: 21283786]
11. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012; 336:740–743. [PubMed: 22582263]
12. Tennessen JA, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337:64–69. [PubMed: 22604720]
13. Nelson MR, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012; 337:100–104. [PubMed: 22604722]
14. Genomes Project, C. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
15. International HapMap, C. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
16. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5:e1000529. [PubMed: 19543373]
17. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009; 84:210–223. [PubMed: 19200528]
18. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012; 44:955–959. [PubMed: 22820512]
19. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010; 11:499–511. [PubMed: 20517342]

20. Lawrence R, et al. Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. Genome Res. 2005; 15:1503–1510. [PubMed: 16251460]
21. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET. Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. 2011; 7:e1002144. [PubMed: 21811411]
22. Liang L, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res. 2013; 23:716–726. [PubMed: 23345460]
23. Gaffney DJ, et al. Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. 2012; 13:R7. [PubMed: 22293038]
24. Innocenti F, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. 2011; 7:e1002078. [PubMed: 21637794]
25. van Nas A, et al. Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. Genetics. 2010; 185:1059–1068. [PubMed: 20439777]
26. Westra HJ, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]
27. Zaitlen N, Pasaniuc B, Gur T, Ziv E, Halperin E. Leveraging genetic variability across populations for the identification of causal variants. Am J Hum Genet. 2010; 86:23–33. [PubMed: 20085711]
28. Zhang K, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nat Methods. 2009; 6:613–618. [PubMed: 19620972]
29. Battle A, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014; 24:14–24. [PubMed: 24092820]
30. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. [PubMed: 24037378]
31. Sun W, Hu Y. eQTL Mapping Using RNA-seq Data. Stat Biosci. 2013; 5:198–219. [PubMed: 23667399]
32. Goode DL, et al. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Res. 2010; 20:301–310. [PubMed: 20067941]
33. Katzman S, et al. Human genome ultraconserved elements are ultraselected. Science. 2007; 317:915. [PubMed: 17702936]
34. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol. 2009; 26:649–658. [PubMed: 19091723]
35. King DC, et al. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res. 2005; 15:1051–1060. [PubMed: 16024817]
36. Pennacchio LA, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006; 444:499–502. [PubMed: 17086198]
37. Consortium EP, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
38. Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJ. A survey of genomic properties for the detection of regulatory polymorphisms. PLoS Comput Biol. 2007; 3:e106. [PubMed: 17559298]
39. Jolma A, et al. DNA-binding specificities of human transcription factors. Cell. 2013; 152:327–339. [PubMed: 23332764]
40. Portales-Casamar E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010; 38:D105–D110. [PubMed: 19906716]
41. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006; 24:1429–1435. [PubMed: 16998473]
42. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA. 2010; 16:1096–1107. [PubMed: 20418358]

43. Zhou T, et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 2013; 41:W56–W62. [PubMed: 23703209]
44. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
45. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–D934. [PubMed: 22064851]
46. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22:1790–1797. [PubMed: 22955989]
47. Griffith OL, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 2008; 36:D107–D113. [PubMed: 18006570]
48. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
49. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–394. [PubMed: 22307276]
50. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
51. Veyrieras JB, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4:e1000214. [PubMed: 18846210]
52. Lee SI, et al. Learning a prior on regulatory potential from eQTL data. PLoS Genet. 2009; 5:e1000358. [PubMed: 19180192]
53. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014
54. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. [PubMed: 19348700]
55. Wang J, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012; 22:1798–1812. [PubMed: 22955990]
56. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011; 7:522. [PubMed: 21811232]
57. Karczewski KJ, et al. Systematic functional regulatory assessment of disease-associated variants. Proc Natl Acad Sci U S A. 2013; 110:9607–9612. [PubMed: 23690573]
58. Brivanlou AH, Darnell JE Jr. Signal transduction and the control of gene expression. Science. 2002; 295:813–818. [PubMed: 11823631]
59. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009; 10:252–263. [PubMed: 19274049]
60. Kilpinen H, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013; 342:744–747. [PubMed: 24136355]
61. McVicker G, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013; 342:747–749. [PubMed: 24136359]
62. Kasowski M, et al. Extensive variation in chromatin states across humans. Science. 2013; 342:750–752. [PubMed: 24136358]
63. Bell JT, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011; 12:R10. [PubMed: 21251332]
64. White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci U S A. 2013; 110:11952–11957. [PubMed: 23818646]
65. Kheradpour P, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013; 23:800–811. [PubMed: 23512712]
66. Arnold CD, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013; 339:1074–1077. [PubMed: 23328393]
67. Smith RP, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet. 2013; 45:1021–1028. [PubMed: 23892608]

68. Cheung VG, et al. Polymorphic cis- and trans-regulation of human gene expression. PLoS Biol. 2010; 8
69. Gilbert LA, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013; 154:442–451. [PubMed: 23849981]
70. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol. 2013; 14:49–55. [PubMed: 23169466]
71. Mali P, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339:823–826. [PubMed: 23287722]
72. Jinek M, et al. RNA-programmed genome editing in human cells. Elife. 2013; 2:e00471. [PubMed: 23386978]
73. Ran FA, et al. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013; 8:2281–2308. [PubMed: 24157548]
74. Bauer DE, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science. 2013; 342:253–257. [PubMed: 24115442]
75. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]
76. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; 363:166–176. [PubMed: 20647212]
77. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]
78. Fehrmann RS, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011; 7:e1002197. [PubMed: 21829388]
79. Gagneur J, et al. Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. PLoS Genet. 2013; 9:e1003803. [PubMed: 24068968]
80. Nica AC, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010; 6:e1000895. [PubMed: 20369022]
81. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet. 2013; 92:126–130. [PubMed: 23246294]
82. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–585. [PubMed: 23715323]
83. Melzer D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008; 4:e1000072. [PubMed: 18464913]
84. Khan Z, et al. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013; 342:1100–1104. [PubMed: 24136357]
85. Wu L, et al. Variation and genetic control of protein abundance in humans. Nature. 2013; 499:79–82. [PubMed: 23676674]

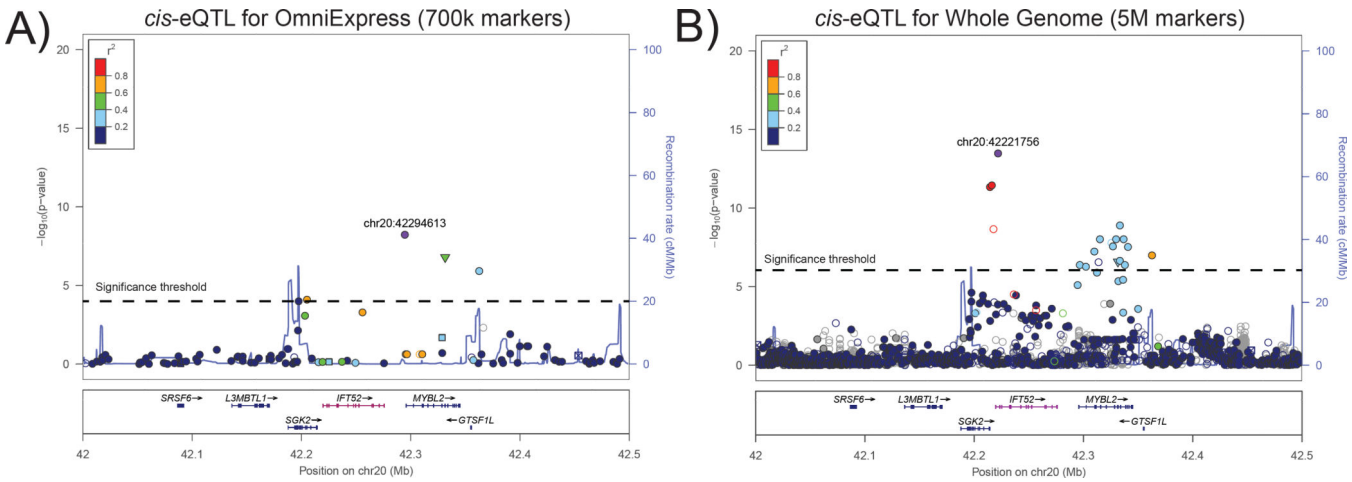

Figure 1. Fine-mapping of a cis-eQTL for IFT52 using whole genome sequencing

(A) Shows association between markers and gene expression in a European population using a combination of microarrays[50] and genetic markers typed by the OmniExpress (700k markers, genome-wide). The multiple-testing significance level is marked by a horizontal dashed line. Here, the top associated SNP (purple) is 3' of *IFT52*. (B) Rerunning the *cis*-eQTL association using whole genome sequencing data (5M markers) identifies a new, more significantly associated variant at the transcription start site of *IFT52*. Furthermore, this variant is in weak LD (r^2 between 0.2 and 0.4, light blue) with multiple 3' variants suggesting that the original top SNP detected in panel A was not in fact the causal variant but was associated due to its linkage with the causal variant now more likely located at the transcription start site. It is also important to note that the multiple-testing significance level has become more stringent when testing eQTL in whole genomes due to testing more markers such that variants near *SGK2* which were significant in the OmniExpress analysis are no longer equally significant in the whole genome analysis.