Quantifying the regulatory effect size of *cis*-acting genetic variation using allelic fold change

Pejman Mohammadi,[1,2] Stephane E. Castel,[1,2] Andrew A. Brown,[3,4,5] and Tuuli Lappalainen[1,2]

[1]*New York Genome Center, New York, New York 10013, USA;* [2]*Department of Systems Biology, Columbia University, New York, New York 10032, USA;* [3]*Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211, Switzerland;* [4]*Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, 1211, Switzerland;* [5]*Swiss Institute of Bioinformatics, Geneva, 1211, Switzerland*

Mapping *cis*-acting expression quantitative trait loci (*cis*-eQTL) has become a popular approach for characterizing proximal genetic regulatory variants. In this paper, we describe and characterize log allelic fold change (aFC), the magnitude of expression change associated with a given genetic variant, as a biologically interpretable unit for quantifying the effect size of *cis*-eQTLs and a mathematically convenient approach for systematic modeling of *cis*-regulation. This measure is mathematically independent from expression level and allele frequency, additive, applicable to multiallelic variants, and generalizable to multiple independent variants. We provide efficient tools and guidelines for estimating aFC from both eQTL and allelic expression data sets and apply it to Genotype Tissue Expression (GTEx) data. We show that aFC estimates independently derived from eQTL and allelic expression data are highly consistent, and identify technical and biological correlates of eQTL effect size. We generalize aFC to analyze genes with two eQTLs in GTEx and show that in nearly all cases the two eQTLs act independently in regulating gene expression. In summary, aFC is a solid measure of *cis*-regulatory effect size that allows quantitative interpretation of cellular regulatory events from population data, and it is a valuable approach for investigating novel aspects of eQTL data sets.

[Supplemental material is available for this article.]

Noncoding genetic variation affecting gene regulation and other cellular phenotypes has a key role in phenotypic variation and disease susceptibility (Albert and Kruglyak 2015). One of the most commonly used methods to characterize genetic variants that affect gene expression is expression quantitative trait loci (eQTL) mapping (Schadt et al. 2003; Lappalainen et al. 2013; The GTEx Consortium 2015), which identifies genetic loci where genotypes of genetic variants are significantly associated to gene expression in a population sample. Genes and variants with significant associations are often called eGenes and eVariants, respectively, and the eVariant with the best *P*-value in a given locus usually used as the proxy for the causal variant. The association between genotype and gene expression is typically tested by regressing gene expression on the number of alternative alleles using a linear model, and the significance of the regression slope is used to measure significance of the eQTL (Shabalin 2012; Ongen et al. 2016). eQTLs can occur either in *trans* through altering diffusible factors that affect gene expression distally or in *cis* through allelic, physical interactions on the same chromosome typically <1 Mb away from the eGene, which are the focus of this study. The allelic effect of *cis*-regulation leads to unequal expression of the two haplotypes (allelic imbalance) in individuals that are heterozygous for a *cis*-acting eVariant (Fig. 1A).

The effect size of an eQTL describes the magnitude of the effect that it has on gene expression and is an important statistic for characterizing the nature of regulatory variants. Estimating the relative effect of eQTL alleles on expression levels has applications in computational functional genetics analysis, as well as in analysis of genetic regulatory variants by experimental assays such as genome editing (Arnold et al. 2013; Canver et al. 2015; Vockley et al. 2015; Tewhey et al. 2016; Ulirsch et al. 2016; Wright and Sanjana 2016). However, thus far there has been no consensus definition for eQTL effect size, with each study defining its own measure for quantifying regulatory effect size. The most widely used measure of effect size is the linear regression slope, a readily available statistic from eQTL calling tools (Shabalin 2012; Gutierrez-Arcelus et al. 2013; Lee et al. 2015; Tung et al. 2015). Linear regression has also been utilized on log-transformed (Flutre et al. 2013; Battle et al. 2014, 2015) or *z*-scored expression data (Lappalainen et al. 2013) to derive slope estimates that do not depend on expression levels. Other statistics include the observed difference between genotype classes, such as the mean difference in expression between heterozygous and the more common homozygote class, sometimes with log transformation or scaling by mean (Gutierrez-Arcelus et al. 2015; Josephs et al. 2015). The proportion of expression variance in the population explained by an eQTL is a widely used statistic that is informative of population variance but not of the molecular effect of an eQTL (Grundberg et al. 2012; Wright et al. 2014; Kirsten et al. 2015). A recent method, developed simultaneously and independently from our work, uses the ratio between the slope and intercept of the linear regression in a variance stabilized model (Palowitch et al. 2016). While all these approaches provide estimates that are generally correlated with *cis*-regulatory effect of a given variant and have a specific

Corresponding authors: pmohammadi@nygenome.org, tlappalainen@nygenome.org

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.216747.116. Freely available online through the *Genome Research* Open Access option.

27:1872–1884 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

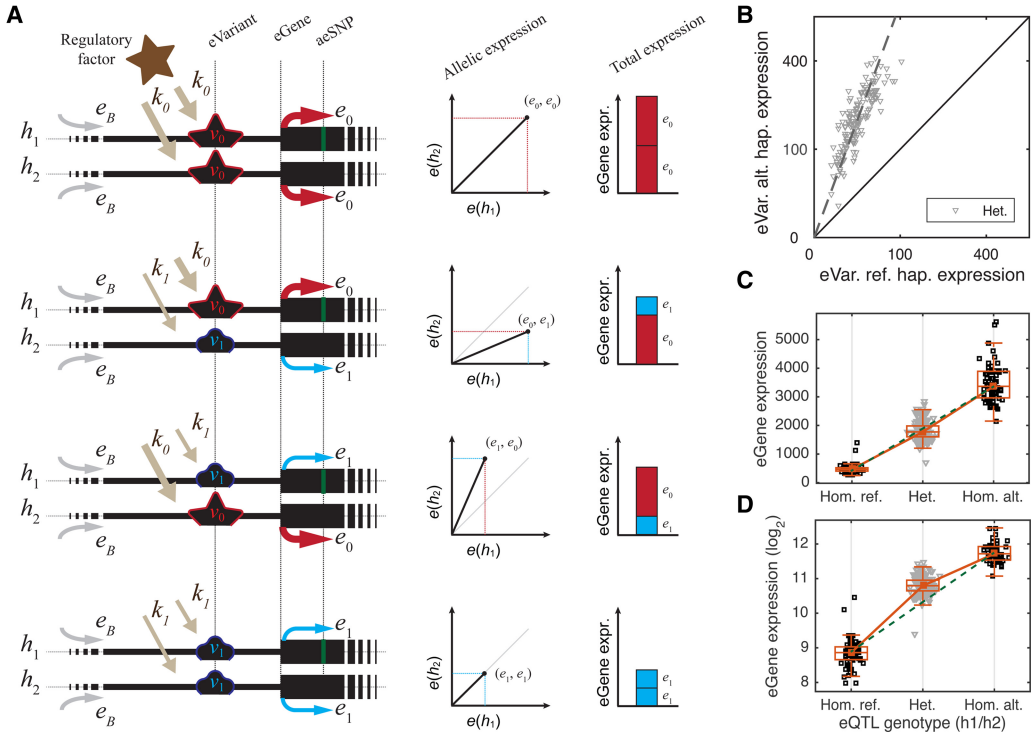

Figure 1. (*A*) Schematic representation of *cis*-regulatory eQTL model in Equations 1 and 2. (*B*) Example of allelic expression associated with each of the alleles of a *cis*-eQTL (eVariant Chr 5: 96252589 T/C; eGene *ERAP2*) in GTEx adipose subcutaneous. Each dot corresponds to allelic imbalance in one individual heterozygous for the eVariant, measured using reads that overlap heterozygous SNPs (aeSNP) in the eGene. Phasing between the aeSNP and the eQTL SNP is utilized to associate the measured allelic expression with each of the eQTL alleles. (*C*,*D*) eGene expression for the same example eQTL. The green dashed line connects the median expression of the two homozygous classes. Expression is linear with number of alternative alleles (*C*), but the linearity is lost after log transformation (*D*).

statistical interpretation in the context of eQTL data, they often lack a straightforward biological interpretation in the greater context of *cis*-regulation that is comparable across different studies, conditions, or data types. Furthermore, many of these easily accessible statistics systematically depend on other variables such as gene expression level, allele frequency, or the amount of technical variation or other noise, which hinders their broad usability across different studies. Finally, a group of statically involved *cis*-eQTL calling methods include regulatory effect size as one of the many parameters for the models that map regulatory variants using both allele-specific expression (ASE) and total gene expression data (Pickrell et al. 2010; Sun 2012; Hu et al. 2015; van de Geijn et al. 2015; Kumasaka et al. 2016), but these methods are distinct from standard, commonly used methods for *cis*-eQTL mapping.

In this study, based upon the mechanistically justified model of *cis* genetic effects on gene expression, we advocate using the log-ratio between the expression of the haplotype carrying the alternative allele to the one carrying the reference allele, the log *allelic fold change* (aFC), as a biologically interpretable and mathematically convenient measure of *cis*-regulatory effect size, applicable to eQTLs discovered by standard eQTL calling methods. This measure is equivalent to the expected log-fold expression ratio of the individuals homozygous for the alternative allele to those homozygous for the reference allele of an eQTL. We provide a thorough description of the derivation and properties of aFC and its generalizations, present practical guidelines and tools for calculating aFC from eQTL as well as allelic expression data, and demonstrate how the extended aFC model can be applied to study more complex regulatory scenarios.

Results

Model

Additive model of regulation

For a given gene and a given *cis*-regulatory variant, v, with two alleles in the population, v_0 and v_1, we define allelic expressions e_0, and e_1 as the amount of transcript produced from the gene when it is located on the same chromosome copy as alleles v_0, and v_1, respectively. We assume that the allelic expression is determined by a shared basal expression of the gene, e_B, driven by the cellular regulatory environment, and allele-specific factors k_0, $k_1 \geq 0$, which represent distinctive effect of the allele v_0, and v_1 on transcription, respectively (Fig. 1A):

$$\begin{aligned} e_0 &= k_0 e_B, \\ e_1 &= k_1 e_B. \end{aligned} \tag{1}$$

Under the *cis*-regulatory model, the regulatory effect of an allele does not depend on the genotype on the other chromosome copy, and $e_{i,j}$, the total expression of the gene in an individual with alleles v_i and v_j on the first and second haplotype is

$$e_{i,j} = (k_i + k_j) e_B, \;\; i, j \in \{0, 1\}. \tag{2}$$

By using $\delta_{i,j} = k_i/k_j$ in Equation 1, the expression of haplotype carrying the alternative allele v_1 is given as

$$e_1 = \delta_{1,0} e_0, \tag{3}$$

relative to e_0, the expression of the haplotype carrying the

reference allele. Similarly, the total relative expression of the gene is

$$e_{i,j} = (\delta_{i,0} + \delta_{j,0})e_0,\ \ i, j \in \{0, 1\}. \quad (4)$$

For a given *cis*-acting eVariant, we define log aFC, $s_{1,0} = \log_2 \delta_{1,0}$, as the relative *cis*-regulatory strength of the allele v_1 versus the reference allele v_0. This quantity is similar to the widely used log expression fold change of differentially expressed genes, but defined between two alleles of a genetic variant. The aFC of a biallelic eVariant can be directly quantified from allelic gene expression in heterozygous individuals (Fig. 1A,B; Supplemental Fig. S1; Box 1) or from summary statistics of standard eQTL linear regression between genotypes and total expression levels (Fig. 1C; Box 2). A further challenge in eQTL effect size estimation is the heteroscedasticity of noise in expression data, which violates the data normality assumptions of linear regression. Although different RNA measurement platforms such as RNA sequencing, microarrays, and other techniques have distinct technical variation profiles, biological variation in gene expression data is generally considered to be log-normally distributed (Tu et al. 2002; Whitehead and Crawford 2006; Anders and Huber 2010). However, after the commonly used variance stabilization by log transformation, gene expression is no longer a linear function and, as such, cannot be characterized efficiently (Fig. 1D; Methods). Thus, we introduce an efficient approximation method to estimate aFC from log-transformed total gene expression data in linear time (Box 3; Methods). The method generates a set of four candidate aFC estimates: The first three estimates are calculated by using only two out of the three eQTL genotype classes at a time. The fourth estimate is derived using log-linear regression, utilizing the fact that log-transformed eQTL data approach a linear function in weak eQTLs as log aFC goes to zero ($s_{1,0} \to 0$; Methods). The candidate aFC that minimizes the residual variance in log-transformed data is reported as the final estimate (Methods).

Generalization to multiple eVariants with multiple alleles

Beside clear biological interpretation, log aFC has several convenient mathematical properties that facilitate downstream analysis of the values (Box 4, Supplemental Methods) and allow generalization to analysis of multiallelic genetic variants, as well as to joint analysis of multiple independent eQTLs for the same eGene. Here we consider the case of N eVariants, $v1$, …, vn, …vN acting on the same eGene independently with m_1, … m_n, … m_N alleles, respectively. Let $i_1 \ldots i_n \ldots i_N$ denote a haplotype carrying the i_n-th allele of the vn; the relative expression on this haplotype is

$$e_{i_1 \ldots\, i_n \ldots i_N} = e_0 \prod_{n=1}^{N} \delta^{vn}_{i_n,0}, \quad (5)$$

where $\delta^{vn}_{i_n,0}$ denotes the aFC associated with allele i_n, at the nth eVariant vn versus its reference allele 0, and e_0 is the reference expression associated with the case $e_{0 \ldots\, 0 \ldots 0}$, where the haplotype carries reference alleles for all eVariants. Thus, the log allelic fold difference between two haplotypes $i_1 \ldots\, i_n \ldots i_N$ and $j_1 \ldots\, j_n \ldots j_N$ is

$$s_{i_1 \ldots\, i_n \ldots i_N,\ j_1 \ldots\, j_n \ldots j_N} = \sum_{n=1}^{N} s^{vn}_{i_n, j_n}, \quad (6)$$

where $s^{vn}_{i_n,j_n}$ denotes the log aFC associated with two alleles i_n and j_n, at the nth eVariant. The total expression of the eGene given the genotype is

$$e_{i_1 \ldots\, i_n \ldots i_N,\ j_1 \ldots\, j_n \ldots j_N} = e_0 \left(\prod_{n=1}^{N} \delta^{vn}_{i_n,0} + \prod_{n=1}^{N} \delta^{vn}_{j_n,0} \right). \quad (7)$$

Following the *cis*-regulatory model, this inherently takes specific configuration of the alleles on each of the two haplotypes into account. The last two equations can be used to simultaneously estimate effect sizes of N eVariants from allelic expression or transcription profiles of genotyped individuals, respectively.

Calculating aFC

We used simulation to evaluate how our three alternative methods for calculating aFC perform under a realistic expression noise level: M1, linear method that uses linear regression coefficients from eQTL data as benchmark for speed (Box 2); M2, nonlinear method that directly solves the regression problem in Equation 17 using a standard nonlinear least square optimization tool (Methods) as a benchmark for accuracy; and M3, nonlinear approximation that solves the nonlinear regression problem from Equation 17 in linear time (Box 3, Fig. 2C). In this simulation, we used simulated data of 10,000 eQTLs with varying allele frequencies and effect sizes (Equations 3, 4), with noise added to the expression levels at 40% coefficient of variation within genotype groups ($\log_{10} \varepsilon_n \sim norm[0, \sigma = 0.17]$; Equation 17) similar to what is observed in real data from GTEx (Supplemental Fig. S2). We found that at this level of noise all three methods provide highly accurate and similar estimates (Fig. 2). All estimates, especially the linear method (M1), deteriorate in eQTLs in which the lower expressed allele has also a low frequency (Fig. 2B). This problem is inherent to *cis*-eQTL data and is expected to occur regardless of the expression measurement platform. Overall, all three methods achieved comparable performances. Specifically, the aFC estimates from the nonlinear

BOX 1. Calculating aFC from allelic expression data. Allelic expression associated with each of the eQTL alleles can be measured in individuals that are heterozygous for the eQTL and that are heterozygous for at least one variant in the eGene (aeSNP). Since allelic expression is measured at the aeSNPs, haplotype phasing data are utilized to obtain the allelic expression from each of the eQTL alleles (Supplemental Fig. S1).

Input:

- Allelic expression of the haplotypes carrying the reference ($c_{0,n}$), and the alternative allele ($c_{1,n}$) of an eQTL in N individuals: ($c_{0,n}$, $c_{1,n}$), where $n \in \{1,2,\ldots, N\}$

1. Get median ratio of the allelic counts:

$$\delta_{1,0} = \underset{n=1 \ldots N}{\text{median}} \frac{c_{1,n}}{c_{0,n}}.$$

Output: Report effect size: $s_{1,0} = \log_2 \delta_{1,0}$

BOX 2. Calculating aFC from gene expression data (for derivations, see Methods).

Input:

- eGene expression in N individuals: $y_1 \ldots y_N$, where $y_n \in [0, +\infty)$
- Number of alternative alleles in each individual: $t_1 \ldots t_N$, where $t_n \in \{0,1,2\}$

1. Use simple linear regression to model expression as a function of t_n:

$$y_n = b_0 + b_1 t_n + \text{noise}.$$

2. Use the slope b_1 and intercept b_0 to calculate

$$\delta_{1,0} = \frac{2b_1}{b_0} + 1.$$

Output: Report effect size: $s_{1,0} = \log_2 \delta_{1,0}$

model (M2) provided the lowest root mean squared deviation (RMSD) from the true values. The linear model was 84 times faster than the nonlinear model but provided 64% higher RMSD. Finally, the nonlinear approximation (M3) presented a trade-off between the speed and accuracy, providing only 10% higher RMSD than the nonlinear model at only 1.8 times the runtime of the linear model.

Next, we applied the three methods for effect size estimation to the *cis*-eQTLs discovered in the Genotype Tissue Expression (GTEx) (The GTEx Consortium 2013, 2015) v6p data set, with eQTL data from 44 tissues (70 to 361 individuals per tissue) (The GTEx Consortium 2017), calculating aFC for all the reported eQTLs in each tissue using the eVariant with the best *P*-value for each eGene. aFCs were estimated from both ASE (Box 1) and eQTL data (Boxes 2–3), independently. For ASE data, we used haplotypic expression at eGenes calculated by summing allelic expression from all phased heterozygous SNPs within the gene (Supplemental Fig. S1). aFC was reported for an average of 57% of eGenes per tissue, requiring haplotypic coverage of at least 10 reads in at least five individuals (The GTEx Consortium 2017). For eQTL-based aFC estimates, we log transformed normalized read counts and corrected for significant linear effects by confounding factors identified using PEER (Stegle et al. 2012) and the top three principal components of the genotype matrix (see Methods, Equations 23, 24). The log aFCs for the eQTLs were calculated using the three models as in the simulation study and constrained to ±log 100. All three eQTL methods provided highly similar aFC estimates with high concordance to ASE-based estimates (Fig. 3A,C). The effect sizes were more discordant between ASE- and eQTL-based estimates when the rare allele was the lower expressed allele, as predicted by the simulation study (Fig. 3B). The nonlinear model provided the best estimates as evaluated by RMSD from ASE-based estimates, and was closely trailed by the nonlinear approximation method (Fig. 3C). Thus, for the rest of the analyses, we used only the nonlinear approximate method as it provided both high accuracy and speed.

Accounting for confounding variation by methods such as PEER is commonly used to improve the statistical power in eQTL calling. Next, we evaluated the effect of this correction on aFC estimates from eQTL data (Supplemental Fig. S3). We found that it has a minimal impact on the aFC estimates (Pearson R = 0.96). However, correcting for confounding sources of variation leads to narrower confidence intervals for the aFC estimates, which is consistent with the increased power in eQTL calling. Finally, we tested the effect of quantile normalization that enforces log-normality of expression data within each genotype. While this is commonly used to avoid outlier effects, we did not observe improvement of the effect size estimates (Fig. 3D).

Comparison of aFC to slope of linear regression

Linear regression slope is the most common measure used for estimating *cis*-eQTL effect size. aFC is closely related to this familiar statistic. From an analytical point of view, the aFC estimation method presented in Box 2 (M1) is a normalization technique to appropriately account for gene expression level. Furthermore, the nonlinear eQTL model provided for estimating aFC from log-transformed gene expression data (M2; Equation 17) is well approximated by log-linear regression for weak eQTLs (for proof, see Supplemental Information). In this case, the regression slope is approximately half of the log aFC; a property we used in the nonlinear approximation method provided in Box 3 (M3) to derive one of the four candidate aFC estimates (Fig. 2C,D).

We used the simulations with realistic expression noise described above to compare the slope of linear regression to aFC. In

BOX 3. Linear time algorithm for estimating aFC from log-transformed gene expression data (for derivations, see Methods).

Input:

- eGene expression in N individuals in $\log_2$ scale: $z_1 \ldots z_N$, where $z_n \in [-\infty, +\infty)$
- Number of alternative alleles in each individual: $t_1 \ldots t_N$, where $t_n \in \{0,1,2\}$

1. Calculate m_0, m_1, m_2 as geometric mean of expression for individuals with $t_n = 0$, 1, and 2, respectively.
2. Calculate the following three candidate estimates:

$$\delta^{*1}_{1,0} = \frac{m_2}{m_0}$$

$$\delta^{*2}_{1,0} = \left(2\frac{m_1}{m_2} - 1\right)^{-1}$$

$$\delta^{*3}_{1,0} = 2\frac{m_1}{m_0} - 1$$

3. Use simple linear regression to model $\log_2$ expression as a function of t_n:

$$z_n = c_1 t_n + c_0 + \text{noise}.$$

4. Use the slope c_1 times two as the fourth candidate estimate:

$$\delta^{*4}_{1,0} = 2^{2c_1}.$$

5. Use each of the four estimates $\delta^{*i}_{1,0}$, $k = 1 \ldots 4$ to calculate

$$r_n(i) = z_n - \log_2\left[(2 - t_n) + t_n \delta^{*i}_{1,0}\right],$$

where $(2 - t_n) + t_n \delta^{*i}_{1,0}$ is predicted gene expression in nth individual using the ith estimate.

6. Pick the estimate that provides the lowest variance in the residuals:

$$\delta_{1,0} = \delta^{*\mathrm{I}}_{1,0}, \quad \mathrm{I} = \underset{i \in 1 \ldots 4}{\operatorname{argmin}} \mathrm{V}[r(i)].$$

Output: Report effect size: $s_{1,0} = \log_2 \delta_{1,0}$

BOX 4. Mathematical properties of log aFC as a relative measure of *cis*-regulatory effect size (for proofs, see Supplemental Methods).

1. Zero log aFC indicates the absence of regulatory difference: $s_{i,i} = 0$.
2. Choice of reference allele only affects the sign of log aFC: $s_{i,j} = -s_{j,i}$.
3. Log aFC is additive:

$$s_{i,k} = s_{i,j} + s_{j,k}.$$

4. Log aFC associated with joint effect of independent regulatory variants, $v1 \ldots vN$ is sum of their individual aFCs:

$$s_{i_1 \ldots i_n \ldots i_N,\, j_1 \ldots j_n \ldots j_N} = \sum_{n=1}^{N} s^{vn}_{i_n, j_n},$$

where $i_1 \ldots i_n \ldots i_N$ and $j_1 \ldots j_n \ldots j_N$ are the set of present alleles on each of the haplotypes.

5. Absolute value of log aFC, $d_{i,j} = |s_{i,j}|$, is a pseudometric:

(i) $d_{i,j} \geq 0$,

(ii) $d_{i,i} = 0$,

(iii) $d_{i,j} = d_{j,i}$,

(iv) $d_{i,k} \leq d_{i,j} + d_{j,k}$.

addition to using linear regression on untransformed expression data, we considered log transformation and *z*-scoring as two common approaches used for eliminating the effect of gene expression level on regression estimates. The results demonstrated that the two transformations largely remove the effect of gene expression level from regression slopes and yield estimates that are highly correlated with aFC estimates (Fig. 4A–C; Supplemental Fig. S4). However, in both cases the transformation introduces systematic biases in the effect size estimates that manifest as distinct deviation patterns from the simulated aFC with respect to allele frequency and eQTL strength. Specifically, in log-transformed data, the slope of linear regression is skewed proportional to frequency of the lower expressed allele, and in *z*-scored data, the slope is inflated as allele frequency deviates from 50% (Fig. 4A–C; Supplemental Fig. S5). We used GTEx data from adipose subcutaneous to see if these biases can be observed in real data using aFC estimates as a baseline. This analysis recapitulated the patterns observed in simulations (Fig. 4D–F). Altogether, these results show that while regression slope is a useful statistic for many purposes, its direct use as eQTL effect size leads to suboptimal results compared with aFC.

Application to GTEx eQTLs

Next, we used GTEx data to explore empirical properties and general trends in eQTL effect size data measured by log aFC. We found that the distributions of aFCs for eQTLs detected in different GTEx tissues are highly dependent on the sample size, due to the fact that tissues with lower sample size lack the power to detect weak eQTLs (Fig. 5A). The effect size estimates from eQTL and ASE data are highly similar but overall 6.35% smaller (CI: [4.6, 8.1]; estimated by errors-in-variables linear regression fit) across the tissues when estimated from eQTL data (Fig. 5B,C). However, this pattern is reversed in effect sizes involving weaker eQTLs, which is consistent with potential winner's curse in the eQTL calling stage (Fig. 5D). This highlights the added value of ASE-based estimates alongside eQTL data. We next analyzed the correlation of

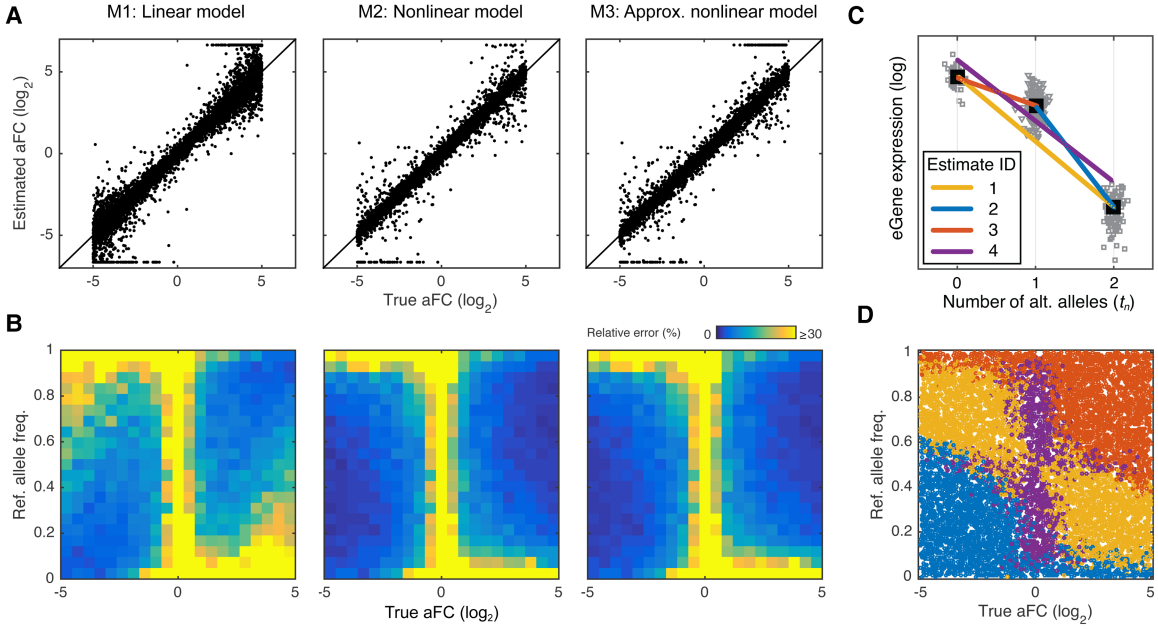

Figure 2. Comparison of the aFC estimation methods using simulated data. We simulated 10,000 eQTLs with noise (40% coefficient of variation), and uniformly selected $\log_2$ aFC (range: [−5,5]), and reference allele frequency (range: [0,1]). (*A*) True aFC used in simulation versus identified values using linear model (M1), nonlinear model (M2), and the nonlinear model approximation (M3). At this level of noise, M2 performed the best, with M1 and M3 having RMSDs of 164% and 110% of M2. (*B*) Quality of the effect size estimates as a function of allele frequency and the true effect size, evaluated by average error relative to the true $\log_2$ aFC. All three estimates, and particularly M1, deteriorate when the lower expressed allele is the minor allele. (*C*,*D*) Schematic representation of the nonlinear model approximation method (Box 3) based on four different candidate estimates (*C*), and the selected estimate with minimum residual variance for each simulated eQTL as a function of reference allele frequency and the true aFC (*D*).

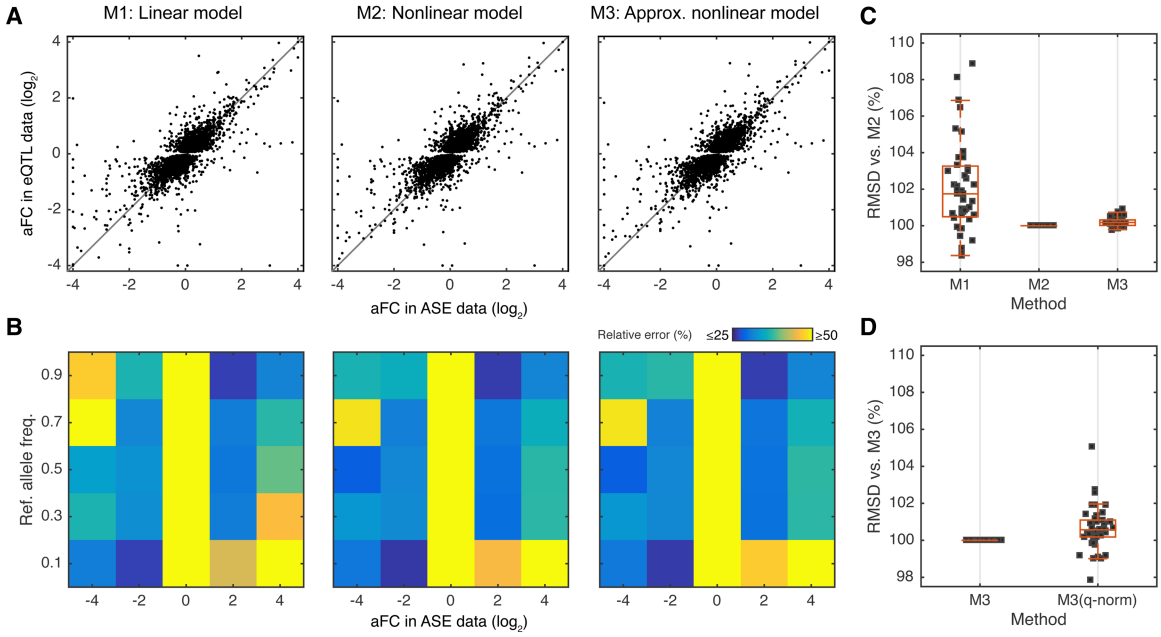

Figure 3. Comparison of the methods for estimating aFC using GTEx data. (*A*) aFC as estimated from ASE data versus estimates from eQTL data using linear model (M1), nonlinear model (M2), and the nonlinear model (M3) approximation for all top eQTLs in adipose subcutaneous. All three estimates are ~75% correlated with estimates from ASE data. (*B*) Quality of the eQTL estimates as a function of allele frequency and the aFC estimate from allelic expression data, evaluated by average relative error between aFC from ASE data and from eQTL estimates. (*C*) Concordance between the estimates from allelic expression and eQTL data as evaluated by RMSD between the most accurate method, M2, and the other two methods. Each dot represents one tissue in GTEx. (*D*) Concordance between the estimates from ASE and eQTL data as evaluated by RMSD, comparing M3 to M3 applied after quantile normalization within each genotype group. Each dot represents one tissue in GTEx.

aFC with other properties of the eVariant or eGene. Low-frequency eVariants tend to have higher effect sizes (Fig. 5E), likely a compound effect of increased selection pressure on stronger eQTLs as well as reduced statistical power in calling weak low-frequency eQTLs with limited data. eGenes with high expression levels, expression in multiple tissues, and high coding region conservation measured by RVIS (Petrovski et al. 2013) have lower effect sizes (Fig. 5F–H), which suggests that genes under strong selective constraints are less likely to tolerate regulatory variants with high effect sizes. Further biological interpretation of effect sizes across eVariants in different annotations, eGenes of different biotypes, and eQTLs that are tissue specific or shared is described by The GTEx Consortium (2017). In these and other downstream analyses of eQTL effect sizes, it is important to correct for correlated factors such as sample size and allele frequency. Even though our simulations demonstrate that aFC is highly robust to key confounders, differences in the power of eQTL mapping will always affect the properties of discovered eQTLs, including effect size distribution.

The aFCs of GTEx eQTLs are provided in the GTEx portal (http://gtexportal.org). Additionally, we implemented the linear model (M1) and the nonlinear approximation model (M3) in a python script (see Software Availability) that takes as input the standard file formats used also by the FastQTL software for eQTL calling. This makes calculation of aFC for other eQTL data sets straightforward and fast.

Application to genes with two distinct eQTL signals in GTEx

Iterative greedy procedures have been utilized to find multiple distinct eQTLs signals for each eGene in the GTEx data (Methods) (The GTEx Consortium 2017). We used GTEx eGenes with two distinct eQTLs to demonstrate how the aFC calculation can be extended to gain mechanistic insight into more complex eQTL patterns. The expression model in Equation 7 written for two biallelic eVariants was used in a nonlinear regression to simultaneously estimate the aFC associated with both eQTLs (Fig. 6A; Supplemental Table S1). These estimates were used to predict the relative expression of the two haplotypes between the 16 possible haplotypic combinations. We found that the predicted values from eQTL data correlate well with the observed values in ASE data across the genotypes (median $r = 0.81$) (Fig. 6B–D). Our generalized expression model inherently accounts for specific arrangement of the alleles for the two eVariants on haplotypes ($e_{11,00} > e_{10,01}$) (Supplemental Fig. S6A). Specifically, according to the model, in individuals that are heterozygous for both eQTLs, the eGene is expected to have higher expression when the two higher expressed alleles occur on the same haplotype ($e_{HH,LL} > e_{HL,LH}$) (Supplemental Fig. S6B). By using eGenes with two eQTLs in adipose subcutaneous, we found that this predicted effect of haplotype arrangement on eGene expression is consistent with the observed expression data ($r = 0.43$, $P = 10^{-25}$) (Supplemental Fig. S6C–F).

Next, we considered the modeling assumption that the two eVariants act independently. Under this assumption, regulatory activity of the alleles from the first eQTL does not depend on the genotype at the second eQTL site and vice versa; therefore, the change in expression of the haplotype carrying the alternative allele for both eVariants is the multiplication of the two aFCs for each individual eVariant ($e_{11} = e_{00}\delta_{1,0}^{v1}\delta_{1,0}^{v2}$; Equation 5). In order to analyze how well the data are described assuming the independence of the two eVariants, we relaxed this assumption to allow for interactions by defining the joint genotype of the two eVariants as the

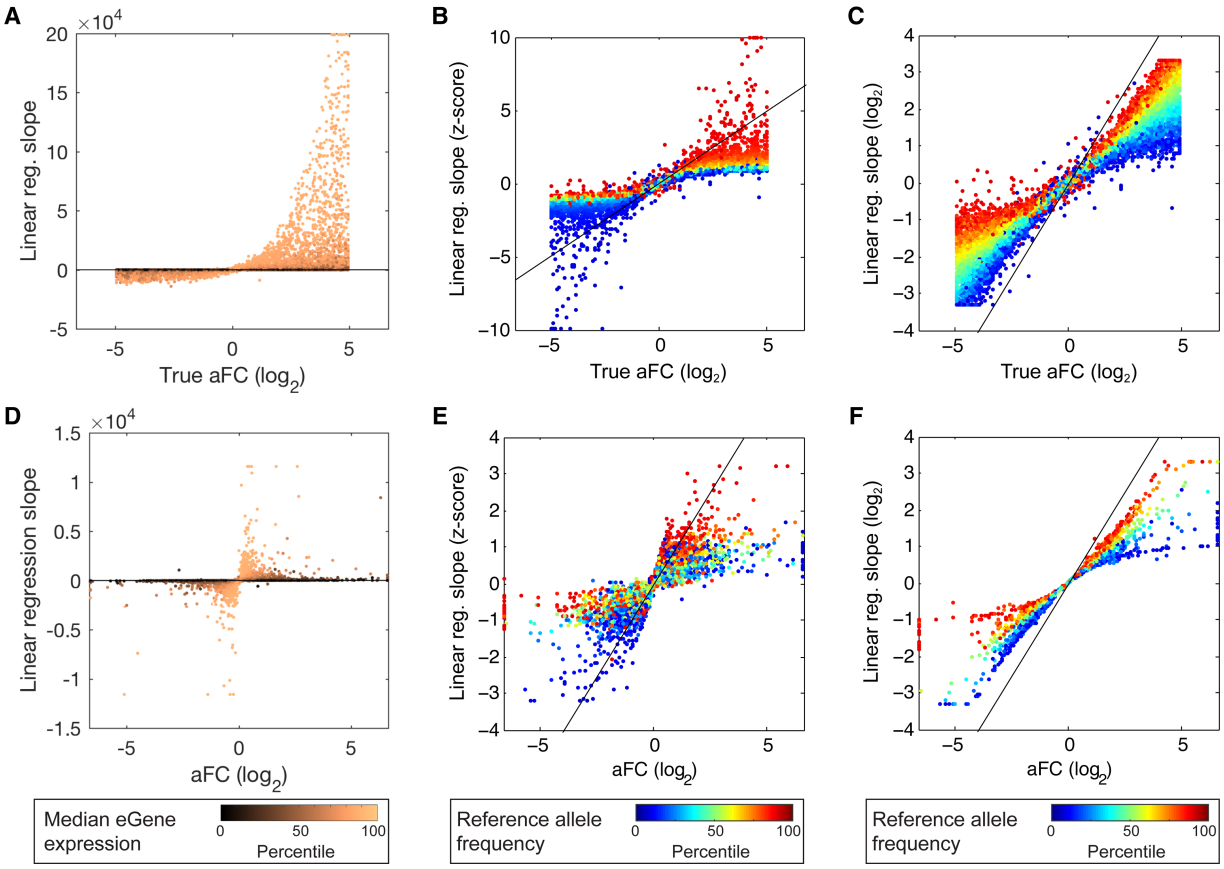

Figure 4. aFC compared with linear regression slope. (*A*–*C*) Slope of linear regression from 10,000 simulated eQTLs generated similarly to data shown in Figure 2. The true aFC value is compared with regression slopes from raw (*A*), *z*-scored (*B*), and $\log_2$ transformed (*C*) data. The color code represents median eGene expression (*A*) and reference allele frequency (*B*,*C*) with alternative color-coding for the same plots in Supplemental Figure S4. (*D*–*F*) Regression slope compared with aFC values estimated using GTEx eQTLs data from adipose subcutaneous.

genotype of a hypothetical variant with four possible alleles. We used Equation 7 written for one four-allelic eVariant to separately estimate the aFC associated with each of the two eVariants, as well as the aFC of their co-occurrence. We found that the estimates from the two models generally agree very well (Fig. 6C). We used the Bayesian information criterion (BIC) within a bootstrapping scheme to decide if relaxing the regulatory independence assumption provides a significantly better description of the data. This could be a sign of biological mechanisms such as epistasis or dosage compensation, as well as confounding factors such as linkage disequilibrium or expression quantification artifacts (Brown et al. 2014; Hemani et al. 2014; Wood et al. 2014; Fish et al. 2016). After accounting for the increased model complexity and uncertainty associated with sampling distribution, we found that only in 0.2% (range across tissues [0, 0.42]) of the two eQTLs for the same gene in GTEx data does the regulatory independence model fail to provide an adequate fit (Fig. 6D; Supplemental Fig. S7; Supplemental Table S1). This finding suggests that distinct eQTL signals identified using the iterative approach are largely driven by independent regulatory mechanisms. We note that the popular iterative discovery approach may be biased toward better discovery of independently acting eQTLs, and future work applying our method to distinct eQTLs discovered by other methods will be required to fully quantify the joint effects of *cis*-regulatory variants in human populations.

Discussion

Despite over a decade of eQTL analysis and its increasingly widespread use in functional and medical genetics, eQTL effect size has lacked a consensus definition that is founded upon molecular interpretation of *cis*-regulation and is analytically convenient for broad use. Here, we described log aFC, a generalizable measure of *cis*-regulatory effect size that captures the mechanistic regulation of haplotype expression in *cis*. Log aFC is consistent across expression levels and allele frequencies and holds mathematically convenient properties that facilitate its application for downstream analysis. We show that aFC model for a single biallelic eQTL SNP is analytically equivalent to linear regression under the additive noise assumption, and therefore, it can be used to obtain effect sizes for eQTLs discovered with standard eQTL calling methods, as well as confidence intervals for aFC estimates that are consistent with eQTL significance. In addition to the aFC that captures the molecular effect, the proportion of expression variation explained by an eQTL in population data remains useful as a complementary measure valuable for describing population-level effect of an eQTL. aFC provides uniform estimates from both allelic expression and *cis*-eQTL data, and replication of *cis*-eQTLs using orthologous ASE data from the same samples can complement classical replication with an independent sample. Furthermore, estimating aFC from ASE and from eQTL data can prove useful in other scenarios.

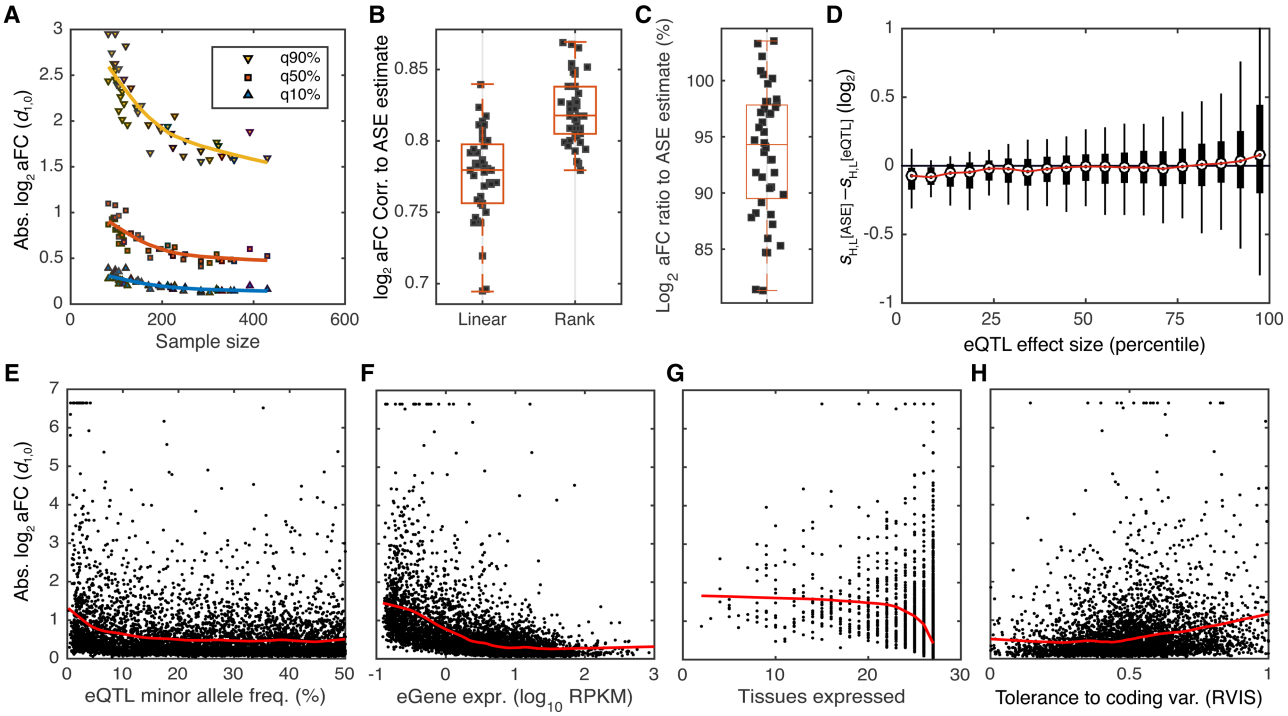

Figure 5. Empirical properties of the aFC distributions in GTEx data. All aFC values are calculated with the nonlinear approximation method (M3). (*A*) Distribution of absolute $\log_2$ aFC across tissues as a function of sample size. Each point represents a tissue in GTEx data, and 90%, 50%, and 10% quantiles of absolute aFC across a tissue are shown. (*B*,*C*) Correlation of $\log_2$ aFC estimates (*B*), and the ratio of the estimates (*C*) derived from eQTL and ASE data. Each point corresponds to one GTEx tissue. (*D*) Difference between the aFC estimates from allelic expression (s^{ASE}) and eQTL (s^{eQTL}) as a function of absolute average aFC ($|s_+^{ASE}\ s^{eQTL}|/2$), with H and L referring to higher and lower expressed alleles of each eQTL in adipose subcutaneous, respectively. Estimated effect size form ASE data tend to be smaller in weak eQTLs and larger for stronger eQTLs as compared to those derived using eQTL data. (*E*–*H*) Distribution of absolute $\log_2$ aFCs calculated from GTEx adipose subcutaneous as function of minor allele frequency (*E*), gene expression level (*F*), number of tissues where the gene is expressed >0.1 RPKM in 10 or more individuals (*G*), and logistic-transformed RVIS, a measure of each gene's tolerance to variation in the coding region (*H*) (Petrovski et al. 2013). Red line shows fit by robust locally weighted scatterplot smoothing.

For instance, ASE-based estimation allows for exploring effects of *cis*-regulatory variation in single individuals, while this is not possible using total expression data (Kukurba et al. 2014; Rivas et al. 2015; The GTEx Consortium 2015).

While the correlation between effect sizes estimated from ASE and eQTL data is high, this is still likely an underestimate and could be improved by using methods that produce more accurate measures of haplotypic expression (Castel et al. 2016). The two alternative aFC calculation methods provided use untransformed and log-transformed eQTL data to account for additive and multiplicative noise, respectively. We showed that the estimates that utilize log-transformed data are generally better. However, both methods perform well, and the preferred noise model can vary depending on the expression measurement platform and upstream preprocessing pipelines that have been utilized. We benchmarked aFC for RNA sequencing data, the most popular platform for expression level quantification that provides both ASE and eQTL data, but aFC is a general measure, and the presented methods provided for eQTL data can be directly applied to data from other quantification platforms such as microarray and qPCR. Systematic extension of aFC-based model of *cis*-regulation to multiple alleles and multiple eQTLs, as demonstrated for the eGenes with two eQTLs in GTEx, allows investigating more complex problems while maintaining mechanistic interpretability of the results. By using the extended model, we showed that the haplotypic arrangement of the alleles of two distinct eQTLs affecting the same gene is important for accurate estimation of gene expression. We also found that the overwhelming majority of distinct eQTLs for the same gene found using the popular iterative eQTL discovery approach are likely to be driven by independent regulatory mechanisms, although future work is needed to study whether this applies to *cis*-regulatory variants in general. Finally, we introduced practical guidelines and a tool for estimating aFC from real data and provided a catalog of *cis*-eQTL effect sizes across all GTEx tissues as a resource for future studies.

A biologically interpretable and well-defined eQTL effect size estimate enables diverse downstream applications. By using the GTEx data set (The GTEx Consortium 2017), we have investigated differences in effect sizes among eGene types, eVariant annotations, and eQTL tissue specificity. Even though aFC itself is unbiased with respect to allele frequency and expression level, we showed here that it is essential for all downstream analyses to take into account factors that indirectly confound the effect size distribution via differences in eQTL discovery power. eQTL effect size quantification will be valuable for making quantitative comparisons between effects on gene expression and other phenotypes at the cellular and physiological level. Indeed, our method is generally applicable to estimating effect size of *cis*-regulatory variants affecting other cellular traits such as methylation, chromatin state, and protein levels as long as the general *cis*-regulatory assumption underlying the model remains realistic. Furthermore, due to the additive nature of log aFC, the magnitude of difference between two effect sizes is a readily interpretable statistic. This feature makes log aFC a useful tool for future characterization of variation

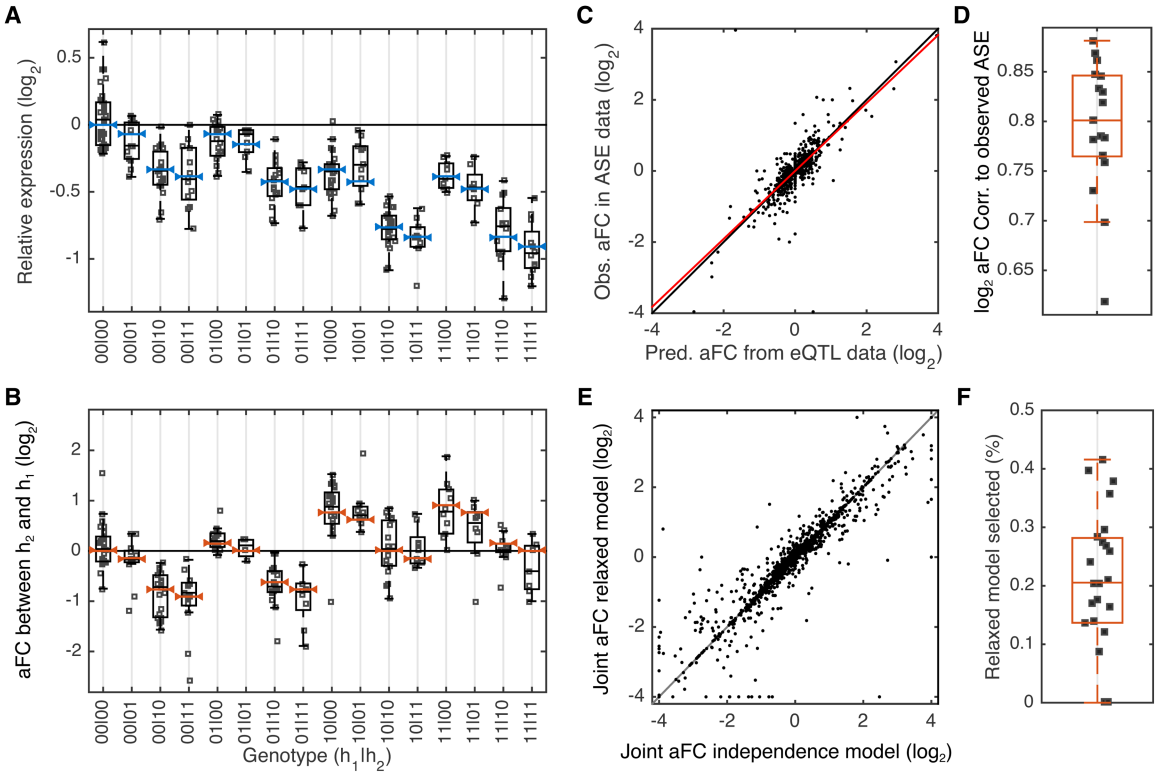

Figure 6. Joint analysis of aFCs for GTEx eGenes with two eQTLs. (*A*) An example of relative expression of eGene *ZC3H3* and the model fits for different genotype groups of its two eQTLs (eVariant1: Chr 8: 144633728 A/G; eVariant2: Chr 8: 144556836 G/A) in GTEx adipose subcutaneous. The effect size of the first and the second eQTLs are −0.77 and −0.14 as measured by $\log_2$ aFC. Each dot represents observed expression in one individual, scaled relative to the expression at all-reference genotype. The blue bars show model fits from the two-eQTL model based on regulatory independence assumption. Reference and alternative alleles are denoted by 0 and 1, respectively, and haplotypes are separated by "|" sign (e.g., 10|11 corresponds to the cases that one haplotype carries alternative and reference alleles of eVariant1 and eVariant2, respectively, and the other haplotype carries the alternative allele of both eVariants). (*B*) Expression of the second haplotype relative to the first haplotype, observed in ASE data. The red bars show expected haplotype expression ratios based on the model in panel *A*, learned on the eQTL data. (*C*) aFC between two haplotypes as predicted from eQTL data compared with median aFC observed in ASE data for all eGenes with two eQTLs in adipose subcutaneous. Each dot represents one randomly selected genotype for one eGene. Red line indicates the robust linear fit ($y = 0.9x + 0.002$). (*D*) Predicted and observed median aFC for all eGenes with two eQTLs calculated from eQTL and ASE data, respectively, in each tissue with more than 200 eGenes with two eQTLs. (*E*) *cis*-Regulatory effect size associated with co-occurrence of the alternative alleles of the two eQTLs, as predicted under regulatory independence model or learned using the relaxed model. (*F*) Percentage of the two eQTLs that are not well described using the independent regulatory assumption across all tissues with more than 200 eGenes with two eQTLs.

in eQTL activity across cellular or environmental contexts. For disease-associated eQTLs, understanding the relationship between the quantitative expression effect in the cells and disease risk will be important for understanding molecular mediators of disease risk. Finally, the recent development of experimental approaches such as MPRA (Tewhey et al. 2016; Ulirsch et al. 2016), STARR-seq (Arnold et al. 2013; Vockley et al. 2015), and CRISPR genome editing assays (Canver et al. 2015; Wright and Sanjana 2016) has created demand for translating summary statistics of eQTL mapping to quantifications that are interpretable as reflecting molecular events in the cell. Our biologically interpretable estimates of *cis*-eQTL effect sizes from population data can be directly compared with in vitro quantification of regulatory variant effects.

Methods

Estimating *cis*-regulatory effect of an eVariant from allelic expression data

Standard RNA sequencing reads can be used to measure the expression of each of the two gene copies, via allelic counts in individuals carrying a heterozygous SNP (aeSNP) inside the transcribed region of the gene (Castel et al. 2015). Allelic counts provide measurement of the true allelic expression e_0 and e_1 from Equation 1 in a given sample on a relative scale (Supplemental Fig. S1). Since both measurements are drawn from the same sample, they share the same basal expression (e_B in Equation 1), and thus in absence of noise, the ratio between the two allelic counts directly reflects the effect of the *cis*-regulatory variant. Given allelic expression data from a set N of individuals heterozygous for an eVariant of interest, the aFC can therefore be robustly estimated as

$$\delta_{1,0} = \underset{n=1\ldots N}{\text{median}} \frac{c_{1,n}}{c_{0,n}}, \tag{8}$$

where $c_{0,n}$ and $c_{1,n}$ are the allelic counts in the nth individual for haplotype carrying reference and alternative allele for the *cis*-regulatory variant, respectively. Here we assume phasing between the regulatory alleles and the aeSNP alleles are known. In cases when phasing information is not available, the magnitude of the regulatory effect size can be calculated as

$$d_{1,0} = |\log_2 \delta_{1,0}| = \underset{n=1\ldots N}{\text{median}} \left|\log_2 \frac{c_{1,n}}{c_{0,n}}\right|. \tag{9}$$

However, this estimate without phasing information is more sensitive to noise and will systematically overestimate the effect size,

particularly in cases where the true effect size is small in magnitude and the variation in allelic counts is dominated by measurement noise.

Estimating *cis*-regulatory effect of an eVariant from gene expression data

Gene expression is linear with the number of alternative alleles for biallelic eVariants

By using Equation 4, we can derive gene expression in an individual as function of the number of alternative alleles, t:

$$e(t) = [(2 - t) + t\delta_{1,0}]e_0, \tag{10}$$

where t is 0, 1, and 2 for individuals homozygous for reference allele, heterozygous, and homozygous for alternative allele, respectively. This equation can be written as

$$e = b_1 t + b_0, \tag{11}$$

where

$$b_0 = 2e_0, \tag{12a}$$

$$b_1 = e_0(\delta_{1,0} - 1), \tag{12b}$$

showing that total gene expression under a *cis*-regulatory model is linear for the number of alternative alleles of the variant (Fig. 1C). For estimating the aFC from expression data, we consider two cases of noise distribution: additive and multiplicative noise.

Estimating aFC from eQTL data with additive noise

Under an additive noise model, the measured gene expression in the nth individual, y_n, is the true expression, $e(t)$, plus a normally distributed noise, ε_n, with zero mean and unknown variance. By using $e(t)$ from Equation 10,

$$y_n = [(2 - t_n) + t_n\delta_{1,0}]e_0 + \varepsilon_n, \tag{13}$$

where t_n is the number of alternative allele in the individual. Similar to Equation 10, Equation 13 can be written in linear form:

$$y_n = b_1 t_n + b_0 + \varepsilon_n. \tag{14}$$

Maximum likelihood estimates for b_0 and b_1 can be derived efficiently using ordinary least squares, and solving Equations 12a and 12b, for $\delta_{1,0}$, the aFC is

$$\delta_{1,0} = \frac{2b_1}{b_0} + 1. \tag{15}$$

Estimating aFC from eQTL data with multiplicative noise

Assuming a multiplicative noise model, the measured gene expression in the nth individual, y_n, is the true expression, $e(t)$, multiplied by a noise, ε_n, such that $\log \varepsilon_n$ is normally distributed with zero mean and unknown variance. Substituting $e(t)$ from Equation 10 again,

$$y_n = [(2 - t_n) + t_n\delta_{1,0}]e_0\varepsilon_n. \tag{16}$$

Due to the multiplicative noise, this equation can no longer be solved as a simple linear regression problem. Applying log transformation to both sides,

$$z_n = \log_2 y_n = \log_2[(2 - t_n) + t_n\delta_{1,0}] + \log_2 e_0 + \log_2 \varepsilon_n. \tag{17}$$

The noise is captured by $\log_2 \varepsilon_n$, which is additive and normally distributed, but the right side of the equation is no longer linear for the number of alternative alleles (Fig. 1D). By using nonlinear least squares optimization, Equation 17 can be solved to derive maximum likelihood estimates for the effect size $\delta_{1,0}$ directly.

Efficient approximation of aFC from eQTL data with multiplicative noise

Nonlinear least squares optimization needed for solving regression problem in Equation 17 is done using iterative numerical optimization that is a relatively slow procedure and not always straightforward to implement. In order to improve efficiency, we use four simplified linear models to derive four candidate estimates of the effect size and choose the one that provides the highest likelihood of the data. First, we derive three estimates of the regulatory effect size using the ratio of the expressions between each of the two genotypes:

$$\delta^{*1}_{1,0} = \frac{m_2}{m_0}, \tag{18a}$$

$$\delta^{*2}_{1,0} = \frac{1}{2\dfrac{m_1}{m_2} - 1}, \tag{18b}$$

$$\delta^{*3}_{1,0} = 2\frac{m_1}{m_0} - 1, \tag{18c}$$

where m_0, m_1, and m_2 are the geometric means of expression in the samples homozygous for reference allele ($t_n = 0$), heterozygous ($t_n = 1$), and homozygous for the alternative allele ($t_n = 2$), respectively (see Supplemental Methods). When the *cis*-regulatory effect size approaches zero, the log-transformed gene expression is linear with number of alternatives alleles (See Supplemental Methods). Therefore, the nonlinear model in Equation 17 can be well approximated with linear regression in cases where the effect size is small ($\log_2\delta_{1,0} \to 0$). We regress log-transformed expressions on the genotype,

$$z_n = c_1 t_n + c_0 + \log_2 \varepsilon_n, \tag{19}$$

and calculate the fourth effect-size estimate as (see Supplemental Methods)

$$\delta^{*4}_{1,0} = 2^{2c_1}. \tag{20}$$

Residual of the fit, r_n, in the nth sample for a given effect size estimate, $\delta^{*k}_{1,0}$, is

$$r_n(k) = z_n - \log_2[(2 - t_n) + t_n\delta^{*k}_{1,0}]. \tag{21}$$

The estimate with lowest variance of the residuals among the four candidates is reported:

$$\delta_{1,0} = \delta^{*\mathrm{I}}_{1,0}, \quad \mathrm{I} = \underset{i\in 1..4}{\operatorname{argmin}}\, \mathrm{V}[r(i)]. \tag{22}$$

Simulation experiment

The simulated data set includes 200 individuals and 10,000 eGenes each associated to exactly one eQTL. Each eQTL has two alleles; frequency of the reference allele, f_0, was drawn from a uniform distribution for each eQTL ($f_o \sim uniform[0,1]$). The eQTL genotype in each individual was decided using two Bernoulli trials. Reference and alternative alleles induce expressions e_0 and $e_1 = \delta_{1,0}\, e_0$ in the eGene in *cis*, respectively (Equations 1, 2). The expression e_0 is generated for each eGene randomly across four orders of magnitude ($\log_{10}\, e_0 \sim uniform[0,4]$). Similarly, the aFC, $\delta_{1,0}$, was assumed to be uniformly distributed in logarithmic scale ($\log_2\, \delta_{1,0} \sim uniform$ [−5,5]) across simulated eQTLs. In order to choose a realistic noise level, we used data from all eGenes associated with eQTLs in GTEx. For each eQTL genotype class, expression mean and variance of

the associated eGene was calculated. As expected, gene expression was highly heteroskedastic with the mean–variance relationship resembling that of multiplicative noise by log-normal distribution (Supplemental Fig. S2). We used average within genotype standard deviation of $\log_{10}$-transformed gene expression to add log-normal noise in the simulation ($\log_{10}\varepsilon_n \sim norm[0, \sigma = 0.17]$; Equation 17).

Estimating aFC for GTEx eQTLs

ASE-based estimates

Haplotypic counts were generated as described by The GTEx Consortium (2017). Briefly, allelic counts for each sample were generated from uniquely aligned RNA-seq reads for all heterozygous SNPs from OMNI Array imputed genotypes using the GATK ASEReadCounter tool (Castel et al. 2015). SNPs covered by less than eight reads, those that showed bias in mapping simulations (Panousis et al. 2014), those that had a UCSC 50-mer mappability lower than one, or those without evidence for heterozygosity (Castel et al. 2015) were excluded. The expression associated with each eQTL allele haplotype was obtained by summing up allelic counts within a gene using population phasing relative to the eQTL variant (eVariant) for each sample. All individuals that are heterozygous for the eVariant were used in Equation 8 to calculate eQTL effect size from haplotypic counts. Bias-corrected and accelerated bootstrap was applied to infer 95% confidence intervals for the aFC estimates (Efron 2012).

eQTL-based estimates

For eQTL data, expression counts were scaled for the total library size, and one pseudocount was added to smooth the normalized counts. Log-transformed expression data were corrected for confounding factors identified using PEER (Stegle et al. 2012) and the three top principal components of the genotype matrix uniformly for all three tested methods: linear, nonlinear, and nonlinear approximation. The correction was done in two steps: First, the log-transformed expression profile of the eGene in nth sample, z_n, was modeled using linear regression:

$$z_n = \mu + \alpha C_n + \beta_{t_n} + \varepsilon_n, \tag{23}$$

where C_n is the nth column of the matrix $C_{M\times N}$ containing M confounding factors, and $t_n \in \{0, 1, 2\}$ indicates the number of alternative alleles in the nth sample. All nonsignificant columns, for which the 95% confidence interval of the regression coefficient in α overlapped zero, were discarded from C. In the second step, the regression was repeated using the reduced covariate matrix, and corrected expression was derived as

$$\hat{z} = z - \alpha C. \tag{24}$$

The corrected expression vector, $\hat{z}$, was used for effect size calculations. For direct estimation of aFC from Equation 17 (the nonlinear method, M2, in Figs. 3, 4), we used the Matlab generic nonlinear least square solver (*lsqnonlin*). The effect size estimates used in Figure 5, as well as those published on GTEx portal (http://gtexportal.org), were calculated using the nonlinear approximation method (M3), and the 95% confidence intervals for the aFC estimates were calculated using the bias-corrected and accelerated bootstrap (Efron 2012). The full data of the GTEx V6p release are available in dbGaP (study accession phs000424.v6.p1), and eQTL summary statistics, including the effect size estimates for the top eVariant–eGene pair per tissue, are available from the GTEx portal (http://gtexportal.org).

Mapping multiple eQTL signals per eGene

Multiple distinct signals for a given expression phenotype were identified by forward stepwise regression followed by a backward selection step. The gene-level significance threshold was set to be the maximum beta-adjusted *P*-value (correcting for multiple testing across the variants) over all eGenes in a given tissue. At each iteration, we performed a scan for *cis*-eQTLs using FastQTL (Ongen et al. 2016), correcting for all previously discovered variants and all standard GTEx covariates. If the beta-adjusted *P*-value for the lead variant was not significant at the gene-level threshold, the forward stage was complete and the procedure moved on to the backward stage. If this *P*-value was significant, the lead variant was added to the list of discovered *cis*-eQTLs as a distinct signal and the forward step moves on to the next iteration. The backward stage consisted of testing each variant separately, controlling for all other discovered variants. To do this, for an eGene with n eVariants, we ran n *cis* scans (in effect $n-1$ *cis* scans, as one replicates the final stage of the forward analysis). For each *cis* scan, we control for all covariates and all but one of the discovered eVariants (the one dropped is the genetic signal that is being tested, conditioned on the full model). If no variant was significant at the gene-level threshold, the variant in question was dropped, otherwise the lead variant from this scan, which controls for all other signals found in the forward stage, was chosen as the variant that represents the signal best in the full model.

Joint analysis of two eQTLs

Regulatory independent model

Let us assume two biallelic eVariants, v_1 and v_2, regulating expression of the same eGene in *cis* (Supplemental Fig. S6A). This is a special case of Equations 5 through 7 where $N = 2$ and $m_1 = m_2 = 2$. Under the independence assumption, the regulatory effect of each eVariant allele on the expression of the carrying haplotype does not depend on the present allele for the other eVariant, and therefore, the expression of a haplotype carrying alleles i_1 and i_2 for the two eVariants is

$$e_{i_1 i_2} = e_0 \delta^{v1}_{i_1,0} \delta^{v2}_{i_2,0}, \tag{25}$$

where indices $i_1, i_2 \in \{0, 1\}$ indicate reference (zero) and the alternative allele (one); $\delta^{v1}_{i_1,0}$, and $\delta^{v2}_{i_2,0}$ are the aFCs associated with the present alleles relative to the reference allele for v_1 and v_2, respectively; and e_0 is the expression of a haplotype carrying reference allele for both eVariants. Under this model, the log ratio between the expressions of the two haplotypes is

$$s_{i_1 i_2, j_1 j_2} = \log_2 \frac{e_{i_1 i_2}}{e_{j_1 j_2}}, \tag{26}$$

where indices $i_1, i_2 \in \{0, 1\}$ and $j_1, j_2 \in \{0, 1\}$ indicate the present alleles on the first and second haplotype, respectively. From definition of Afc,

$$\delta_{i,0} = \delta_{i,j}\delta_{j,0}; \tag{27}$$

thus after substituting haplotypic expressions from Equation 25 in Equation 26, the log ratio between the expressions of the two haplotypes is

$$s_{i_1 i_2, j_1 j_2} = \log_2\left(\delta^{v1}_{i_1,j_1}\delta^{v2}_{i_2,j_2}\right) = s^{v1}_{i_1,j_1} + s^{v2}_{i_2,j_2}. \tag{28}$$

This equation presents the expected log aFC for a given genotype. Therefore, under the regulatory independence model, the joint effect of the two alternative alleles is sum of their individual effects:

$$s_{11,00} = s^{v1}_{1,0} + s^{v2}_{1,0}. \tag{29}$$

Under the *cis*-regulatory model, total expression of the eGene for each genotype is the some of the individual haplotype expressions:

$$e_{i_1 i_2, j_1 j_2} = e_{i_1 i_2} + e_{j_1 j_2}. \tag{30}$$

Substituting haplotypic expressions from Equation 25, we can use measured expression profiles of genotyped individuals to estimate aFC associated with the two eVariants. The observed expression value for the eGene in the nth sample after log transformation is

$$z_{i_{n,1} i_{n,2}, j_{n,1} j_{n,2}} = \log e_0 + \log\left(\delta^{v1}_{i_{n,1},0}\delta^{v2}_{i_{n,2},0} + \delta^{v1}_{j_{n,1},0}\delta^{v2}_{j_{n,2},0}\right) + \alpha C_n + \varepsilon_n, \tag{31}$$

where indices $i_{n,1}$, $i_{n,2}$, $j_{n,1}$, $j_{n,2} \in \{0, 1\}$ indicate the present alleles, and C_n is the provided column vector of the confounding factors for the sample. The nonlinear regression problem can be solved to estimate reference expression e_0, individual aFC effects $\delta^{v1}_{1,0}$, $\delta^{v2}_{1,0}$, and the cofactor weight vector α (by definition $\delta^{v1}_{0,0}$ and $\delta^{v2}_{0,0}$ are each equal to 1).

In order to estimate aFCs for eGenes with two eQTLs in GTEx data, we used PEER (Stegle et al. 2012) and top three principal components of the genotype matrix as the confounding factors in matrix C. Generic nonlinear least square optimizer in Matlab (*lsqnonlin*) was used to derive parameter estimates for the Equation 26 regression problem. Confidence intervals of the parameters were derived using the *t*-statistic estimated via Jacobean matrix calculated at the optimal function values (Matlab function: *nlparci*). Predicted aFCs for regulatory independence model presented in Figure 6, B through E, and Supplemental Figure S7C (blue bars) were derived using Equation 28. The prediction of haplotype arrangement effects in Supplemental Figure S6 were derived using Equations 30 and 31.

Relaxed model

In this model, we relax the regulatory independence assumption, allowing the regulatory effect associated with co-occurrence of the two alternative alleles to be potentially different from sum of their individual effects. In contrast to Equation 25, haplotype expression is

$$e_{i_1 i_2} = e_0 \delta_{i_1 i_2, 00}, \tag{32}$$

where $\delta_{i_1 i_2, 00}$ is the aFC associated to copresence of the alleles i_1 and i_2 of the eVariants v_1 and v_2 compared with a haplotype carrying reference allele for both eVariants. This model is equivalent to a special case of models in Equations 5 through 7, where $N = 1$ and $m_1 = 4$. From the aFC definition,

$$\delta_{i_1 i_2, 00} = \delta_{i_1 i_2, j_1 j_2}\delta_{j_1 j_2, 00}, \tag{33}$$

and the log ratio between the expressions of the two haplotypes is

$$s_{i_1 i_2, j_1 j_2} = \log_2 \delta_{i_1 i_2, j_1 j_2} = s_{i_1 i_2, 00} - s_{j_1 j_2, 00}. \tag{34}$$

The total expression is the sum of the individual haplotypic expressions (Equation 30); thus, the observed expression value for the eGene in the nth sample under the relaxed regulatory model after log transformation is

$$z_{i_{n,1} i_{n,2}, j_{n,1} j_{n,2}} = \log e_0 + \log\left(\delta_{i_{n,1} i_{n,2}, 00} + \delta_{j_{n,1} j_{n,2}, 00}\right) + \alpha C_n + \varepsilon_n, \tag{35}$$

where indices $i_{n,1}$, $i_{n,2}$, $j_{n,1}$, $j_{n,2}$ indicate the present alleles and C_n the covariates as described in Equation 31. The nonlinear regression problem can be solved for reference expression e_0, joint aFC effects $\delta_{10,00}$, $\delta_{01,00}$, $\delta_{11,00}$, and the cofactor weight vector α (by definition $\delta_{00,00}$ is equal to one).

To estimate aFCs in GTEx data, regression parameters and their confidence intervals were estimated as described for the regulatory independence model. Predicted aFCs for the relaxed model presented in Figure 6E and Supplemental Figure S7C (red bars) were derived using Equation 34.

Model comparison

In order to compare the two models of *cis*-regulation, the independence and the relaxed model, we calculated total data likelihood for each of the models under the log-normality assumption:

$$\mathrm{L}(\underline{z}|\mathrm{M}) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{r_n^2}{2\sigma^2}}, \tag{36}$$

where $\underline{z}$ is the vector of N samples, r_n is the fit residual at the nth sample using the model considered M, and σ is the standard deviation of the fit residuals. Bayesian information criterion (BIC) for each of two models was calculated:

$$\mathrm{BIC}(\mathrm{M}) = -2\log \mathrm{L}(\underline{z}|\mathrm{M}) + \lambda \log N, \tag{37}$$

where λ, the number of parameters in each model, is the number of cofactor coefficients plus three and plus four for the regulatory independence and the relaxed model, respectively. We used bias-corrected and accelerated bootstrap (Efron 2012) to estimate confidence intervals for $\Delta\mathrm{BIC} = \mathrm{BIC}(\text{Relaxed model}) - \mathrm{BIC}(\text{Independence model})$ in cases where $\Delta\mathrm{BIC}$ is negative. The relaxed model was selected in cases where the upper bound for the 95% confidence interval for $\Delta\mathrm{BIC}$ fell below zero, and for the rest of the cases, the independence model that has fewer parameters was deemed adequate. The calculated aFCs for all eGenes in GTEx with two associated eQTLs are provided in Supplemental Table S1.

Software availability

Software for calculating aFC from standard eQTL data is provided in Supplemental Software S1 and is available online on GitHub (https://github.com/secastel/aFC).

Acknowledgments

The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000 029C) to The Broad Institute. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by supplements to University of Miami grants DA006227 and DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 and MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina–Chapel Hill (MH090936 and MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v6.p1 on 05/23/2016. T.L.

and P.M. are supported by NIH grant R01M H106842, T.L. is supported by the NIH grant UM1HG008901, and T.L. and S.E.C. are supported by the NIH contracts HHSN268201000029C and R01MH101814. The multiple eQTL mapping was performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics (SIB).

Author contributions: P.M. and T.L. designed the study. P.M. developed the statistical models and the MATLAB toolbox and analyzed the data. S.E.C. developed the Python package and analyzed the data. A.A.B. provided the independent eQTL data. P.M. and T.L. wrote the manuscript with contributions from all the authors. All the authors read and approved the final manuscript.

References

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16:** 197–212.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, 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.

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2015. Impact of regulatory variation from RNA to protein. *Science* **347:** 664–667.

Brown AA, Buil A, Viñuela A, Lappalainen T, Zheng H-F, Richards JB, Small KS, Spector TD, Dermitzakis ET, Durbin R. 2014. Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3:** e01381.

Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE. 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527:** 192–197.

Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16:** 195.

Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. 2016. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun* **7:** 12817.

Efron B. 2012. Better bootstrap confidence intervals. *J Am Stat Assoc* **82:** 171–185.

Fish AE, Capra JA, Bush WS. 2016. Are interactions between cis-regulatory variants evidence for biological epistasis or statistical artifacts? *Am J Hum Genet* **99:** 817–830.

Flutre T, Wen X, Pritchard J, Stephens M. 2013. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* **9:** e1003486.

Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang T-P, Meduri E, Barrett A, et al. 2012. Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat Genet* **44:** 1084–1089.

The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45:** 580–585.

The GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348:** 648–660.

The GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550:** 204–213.

Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2:** e00523.

Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, et al. 2015. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet* **11:** e1004958.

Hemani G, Shakhbazov K, Westra H-J, Esko T, Henders AK, McRae AF, Yang J, Gibson G, Martin NG, Metspalu A, et al. 2014. Detection and replication of epistasis influencing transcription in humans. *Nature* **508:** 249–253.

Hu Y-J, Sun W, Tzeng J-Y, Perou CM. 2015. Proper use of allele-specific expression improves statistical power for *cis*-eQTL mapping with RNA-seq data. *J Am Stat Assoc* **110:** 962–974.

Josephs EB, Lee YW, Stinchcombe JR, Wright SI. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc Natl Acad Sci* **112:** 15390–15395.

Kirsten H, Al-Hasani H, Holdt L, Gross A, Beutner F, Krohn K, Horn K, Ahnert P, Burkhardt R, Reiche K, et al. 2015. Dissecting the genetics of the human transcriptome identifies novel trait-related *trans*-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum Mol Genet* **24:** 4746–4763.

Kukurba KR, Zhang R, Li X, Smith KS, Knowles DA, How Tan M, Piskol R, Lek M, Snyder M, MacArthur DG, et al. 2014. Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet* **10:** e1004304.

Kumasaka N, Knights AJ, Gaffney DJ. 2016. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48:** 206–213.

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501:** 506–511.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47:** 955–961.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32:** 1479–1485.

Palowitch J, Shabalin A, Zhou Y, Nobel AB, Wright FA. 2016. Estimation of interpretable eQTL effect sizes using a log of linear model. arXiv:1605.08799 [stat.ME].

Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, Lappalainen T. 2014. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol* **15:** 467.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9:** e1003709.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, et al. 2015. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348:** 666–669.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422:** 297–302.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28:** 1353–1358.

Stegle O, Parts L, Piipari M, Winn J, 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.

Sun W. 2012. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68:** 1–11.

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S. 2016. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165:** 1519–1529.

Tu Y, Stolovitzky G, Klein U. 2002. Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci* **99:** 14031–14036.

Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. 2015. The genetic architecture of gene expression levels in wild baboons. *eLife* **4:** e04729.

Ulirsch JC, Nandakumar SK, Wang L, Giani FC. 2016. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165:** 1530–1545.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12:** 1061–1063.

Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe WL, Reddy TE. 2015. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res* **25:** 1206–1214.

Whitehead A, Crawford DL. 2006. Neutral and adaptive variation in gene expression. *Proc Natl Acad Sci* **103:** 5425–5430.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46:** 1173–1186.

Wright JB, Sanjana NE. 2016. CRISPR screens to discover functional noncoding elements. *Trends Genet* **32:** 526–529.

Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou Y-H, et al. 2014. Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46:** 430–437.

Received September 30, 2016; accepted in revised form June 5, 2017.