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], Tuuli Lappalainen[1, 2]

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

Keywords: eQTL, allele-specific expression, regulatory variant, effect size, *cis*-regulation, GTEx companion

Abstract

Mapping *cis*-acting expression quantitative trait loci (*cis*-eQTL) has become a popular approach for characterizing proximal genetic regulatory variants. However, measures used for quantifying the effect size of *cis*-eQTLs have been inconsistent and poorly defined. In this paper, we describe log allelic fold change (aFC) as a biologically interpretable and mathematically convenient unit that represents the magnitude of expression change associated with a given genetic variant. This measure is mathematically independent from expression level and allele frequency, applicable to multi-allelic variants, and generalizable to multiple independent variants. We provide tools and guidelines for estimating aFC from eQTL and allelic expression data sets, and apply it to 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 these eQTLs are independent in their regulatory activity. 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.

Introduction

Non-coding 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 eQTL mapping (Schadt et al. 2003; Lappalainen et al. 2013; 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 less than 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 (Tewhey et al. 2016; Ulirsch et al. 2016; Vockley et al. 2015; Arnold et al. 2013; Canver et al. 2015; Wright and Sanjana 2016). However, thus far there has been no consensus definition for eQTL effect size, let alone one that allows a direct biological interpretation, with previous eQTL studies using different units and approaches. The most widely used measure of effect size is simply the regression slope, a readily available statistic from eQTL calling tools (Shabalin 2012; Gutierrez-Arcelus et al. 2013; Tung et al. 2015; Lee et al. 2015). Other statistics include slope of linear regression based on log-transformed expression (Flutre et al. 2013; Battle et al. 2014), and estimation of the 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 (Wright et al. 2014; Kirsten et al. 2015; Grundberg et al. 2012). 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, they often lack a well-defined unit that enables biological interpretation of the effect size. Furthermore, many of these statistics are confounded by nuisance variables such as genotype frequency, gene expression level or technical or environmental variation. These limitations can confound downstream analysis.

cis-acting regulatory variation is known to be reflected in both allele-specific expression (ASE), and total gene expression data as incorporated in previous statically involved *cis*-eQTL calling methods (Pickrell et al. 2010; Sun 2012; van de Geijn et al. 2015; Hu et al. 2015; Kumasaka et al. 2016). In this study, based upon the mechanistically justified model of additive *cis* genetic effects on gene expression, we define 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. We provide a thorough description of the derivation and properties of this measure, including its generalizations that enable analysis of multi-allelic genetic variants and joint modeling of multiple *cis*-regulatory variants. We make the calculation of eQTL effect sizes accessible to the wide eQTL community by practical guidelines and tools, and provide effect sizes for all *cis*-eQTLs in the GTEx data set (co-submitted -(Aguet et al. 2016)). We characterize the empirical trends across the effect sizes in GTEx data, demonstrating a good fit between the empirical results and simulations, and describing factors correlated with observed allelic fold changes. Finally, we demonstrate application of *cis*-regulatory model extended for joint analysis of allelic fold changes in eGenes with two eQTLs in GTEx data.

Results

1. Model

1.1 Additive model of regulation

For a given gene and a given *cis*-regulatory variant, v, with two alleles in 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, $k_1 \geq 0$, that represent distinctive effect of the allele v_0, and v_1 on transcription, respectively (**Fig. 1A**):

$$e_0 = k_0 e_B$$
$$e_1 = k_1 e_B \qquad (1)$$

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

$$e_{i,j} = (k_i + k_j) e_B, \quad i,j \in \{0,1\} \qquad (2)$$

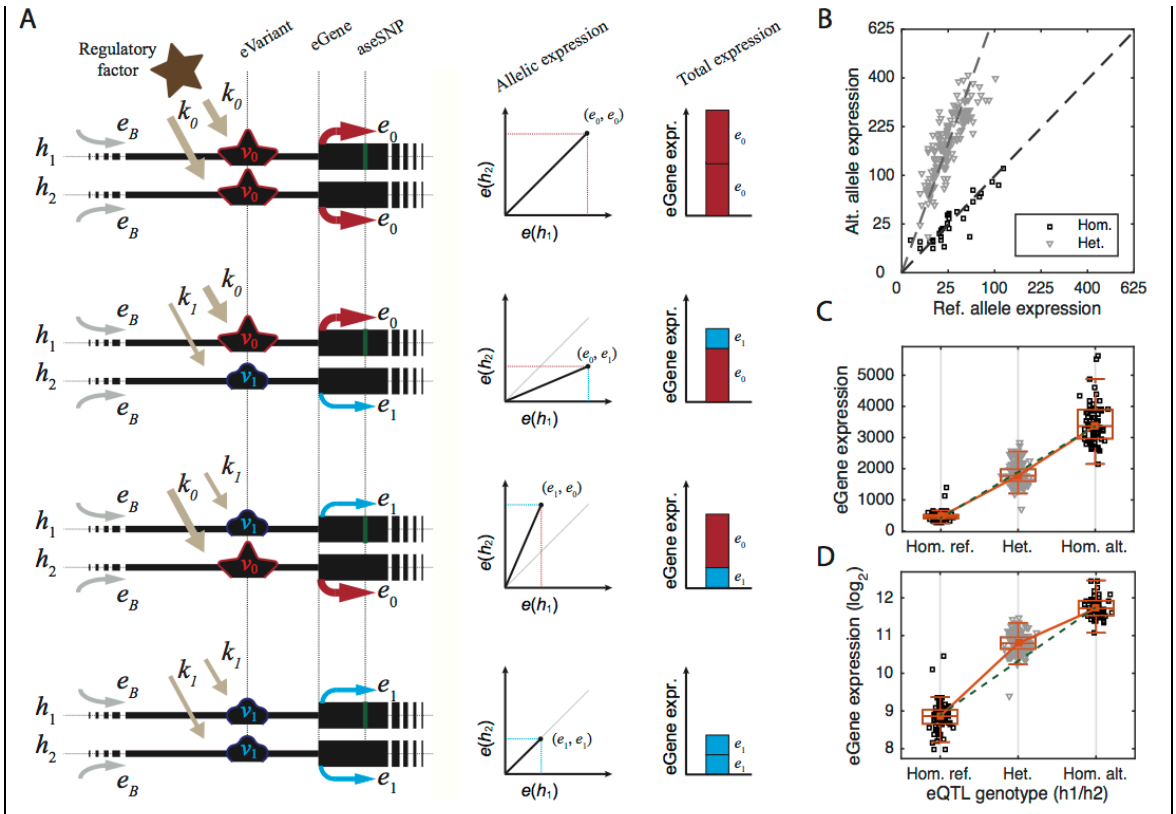

Figure 1. A) Schematic representation of *cis*-regulatory eQTL model in **Eq. 1, 2**. B) Example of allelic expression (eVariant chr5:96252589 T/C; eGene *ERAP2*) in GTEx Adipose Subcutaneous. C-D) eGene expression for the same example eQTL. The green dashed line connects the median expression of the two homozygous classes. Expression is linear with number of alternative alleles (C), but the linearity is lost after log-transformation (D).

However, observational population data generally includes only relative expression quantifications. Using $\delta_{i,j}=k_i/k_j$ in **Eq.1**, the expression of haplotype carrying the alternative allele v_1 is given as

$$e_1 = \delta_{1,0} e_0 \quad (3)$$

relative to e_0, the expression of the haplotype carrying the reference allele. Similarly, the total relative expression of the gene is

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

For a given *cis*-acting eVariant, we define log allelic fold change, $s_{1,0} = \log_2 \delta_{1,0}$, 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. Allelic fold change of a biallelic eVariant can be directly quantified from allelic gene expression in heterozygous individuals (**Fig. 1A-B; 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 solved efficiently (**Fig. 1D; Methods**). Thus, we introduce an efficient heuristic method to estimate allelic fold change from log-transformed gene expression data in linear time (**Box 3**). 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 loglinear regression, utilizing the fact that log-transformed eQTL data approaches a linear function in weak eQTLs as log allelic fold change 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**).

1.2 Generalization to multiple eVariants with multiple alleles

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

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

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

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

Where $s^{vn}_{i_n,j_n}$ denotes the log allelic fold change associated with two, alleles i_n and j_n, at the n^{th} eVariant. The total expression of the eGene given the genotype is

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

Following the *cis*-regulatory model, this inherently takes into account the independent expression of the two haplotypes according to the alleles that they carry, which is different from a simple additive model of multiple eQTLs that ignores their haplotype configuration. 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.

Input:

- Allelic expression in N individuals heterozygous for the top eVariant of an eQTL of interest: $(c_{0,1}, c_{1,1}) \dots (c_{0,N}, c_{1,N})$

1. Get median ratio of the allelic counts:

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

where $(c_{0,n}, c_{1,n})$ are allelic counts from the 1st, and 2nd haplotype in the n^{th} individual.

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

Box 1: Calculating aFC from allelic expression data.

Input:

- eGene expression in N individuals: $y_1 \dots y_N$, where $y_n \in [0, +\infty)$
- Number of alternative alleles in each individual: $t_1 \dots t_N$, 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: $s_{1,0}=\log_2 \delta_{1,0}$

Box 2: Calculating aFC from gene expression data (see **Methods** for derivations).

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:

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

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

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

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:

$$\delta_{1,0}^{*4} = 2^{2c_1}$$

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

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

where $(2 - t_n) + t_n \delta_{1,0}^{*i}$ is predicted gene expression in n^{th} individual using the i^{th} estimate.

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

$$\delta_{1,0} = \underset{i \in 1 \ldots 4}{\operatorname{argmin}} \, \mathrm{V}[r(i)]$$

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

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

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, *v*1…*v*N is sum of their individual aFCs:

$$s_{\langle i_1 \dots i_n \dots i_N \rangle, \langle j_1 \dots j_n \dots j_N \rangle} = \sum_{n=1}^{N} s_{i_n, j_n}^{vn}$$

where $\langle i_1 \dots i_n \dots i_N \rangle$ and $\langle j_1 \dots j_n \dots j_N \rangle$ 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 pseudo-metric:
   i) $d_{i,j} \geq 0$
   ii) $d_{i,i} = 0$
   iii) $d_{i,j} = d_{j,i}$
   iv) $d_{i,k} \leq d_{i,j} + d_{j,k}$

Box 4: Mathematical properties of log aFC as a relative measure of *cis*-regulatory effect size (see **Supplemental methods** for proofs).

2. Simulation without noise

Regression slope is probably the most common measure used for estimating *cis*-eQTL effect size. However, compared to aFC, it lacks a clear biological interpretation, and is prone to systematical biases introduced by expression level and allele frequency. We demonstrate this by simulation of *cis*-eQTLs without noise (**Eq. 3, 4**), and comparing estimates of effect size by log aFC (**Box 2, 3**; since there is no noise, both methods yield identical results) linear regression slope (**b_1 in Box 2),** and regression slope after log-transformation of the expression data (c_1 in **Box 3**). We consider two eQTLs: one with four times higher expression of the alternative than the reference allele and another the opposite. The three measures of effect size were calculated for a fixed reference allele frequency of 50% and varying gene expression levels (**Fig. 2A**), and for a fixed gene expression level and varying allele frequency (**Fig. 2B**). Our results show that linear expression slope varies with gene expression levels, and the loglinear slope varies with allele frequency, and neither provides a quantitative estimate of the four-fold expression difference between alleles. The log aFC estimate remains insensitive to both confounding factors, and yields the correct estimate of the eQTL effect size.

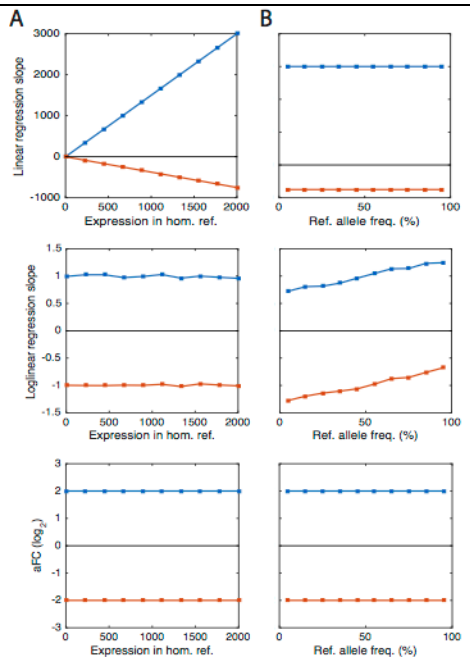

Figure 2: Comparison of three eQTL effect size measurements in simulations without noise. Log aFC is compared to linear and loglinear regression slopes for simulated eVariants with the alternative allele expressed four times higher compared to the reference (blue) and vice versa (red), for varying expression levels of the homozygous reference genotype (A) and varying allele frequencies (B). The results demonstrate that log aFC quantifies the biological expression difference between alleles, and is robust to changing allele frequencies and expression levels.

3. Noise distribution in eQTL data and simulation with realistic noise

Next, 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 **Eq. 17** using a standard nonlinear least square optimization tool (**Methods**) as a benchmark for accuracy; M3) Nonlinear approximation that solves the nonlinear regression problem from **Eq. 17** using our heuristic solution (**Box 3, Fig. 3C**). In this simulation, we used simulated data of **10,000** eQTLs with varying allele frequencies and effect sizes (**Eq. 3, 4**), with noise added to the expression levels at **40%** coefficient of variation within genotype groups ($\log_{10} \varepsilon_n \sim norm[0, \sigma = 0.17]$, **Eq. 17**) similar to what is observed in real data from GTEx (**Supplementary Fig. 1**). We found that at this level of noise all three methods provide highly accurate and similar estimates (**Fig. 3**). All estimates, especially the linear method (M1), deteriorate in eQTLs in which lower expressed allele has also a low frequency (**Fig. 3B**). This problem is inherent to *cis*-eQTL data and is expected to occur regardless of the expression measurement platform. Overall, the aFC estimates from the nonlinear model (M2) provided the lowest root mean squared deviation (RMSD) from the true values, with the its approximation (M3) providing only 10% worse RMSD than the nonlinear model at 1.8 times the runtime of the linear model. The linear model was 84 times faster than the nonlinear model but provided 64% higher RMSD.

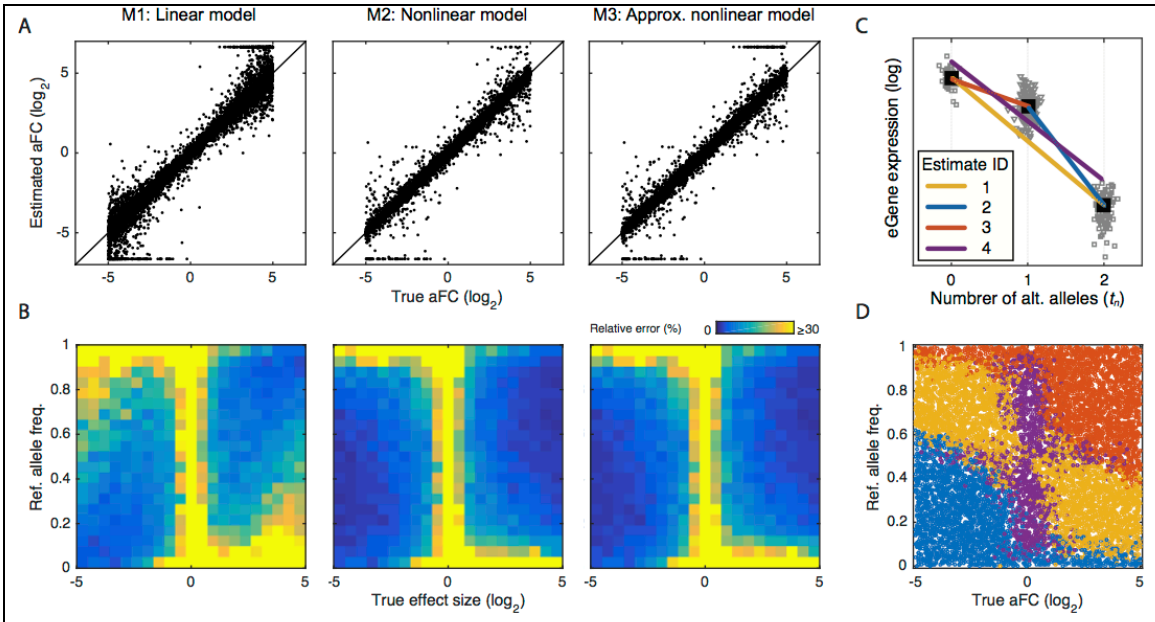

Figure 3: Comparison of the aFC estimation methods using simulated data. We simulated 10,000 eQTLs with noise (40% coefficient of variation), and uniformly selected $\log_2$ aFC (range: [-5,5]), and reference allele frequency (range: [0,1]). A) True aFC used in simulation versus identified values using 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).

4. Application to GTEx eQTLs

Next, we applied our methods for effect size estimation to the *cis*-eQTLs discovered in the Genotype Tissue Expression (GTEx) (GTEx Consortium 2013; 2015) v6p dataset, with eQTL data from 44 tissues (70-361 individuals per tissue; co-submitted (Aguet et al. 2016)), calculating aFC for all the reported eQTLs in each tissue, using the eVariant with the best p-value for each eGene. Allelic fold changes were estimated from both allele-specific expression (ASE; **Box 1**) and eQTL data (**Box 2-3**). For ASE data, we used haplotypic expression at eGenes calculated by summing allelic expression from all phased heterozygous SNPs within the gene. aFC was reported for an average of 57% of eGenes per tissue, requiring haplotypic coverage of at least 10 reads in at least 5 individuals (co-submitted (Aguet et al. 2016)). For eQTL-based aFC estimates, we log transformed normalized read counts, and corrected for confounding factors identified using PEER (Stegle et al. 2012) and the top three principal components of the genotype matrix (see methods, **Eqs. 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. 4A, 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. 4B)**. 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. 4C**). Thus, for the rest of the analyses we used only the nonlinear approximate method as it provided both high accuracy and speed. 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. 4D**).

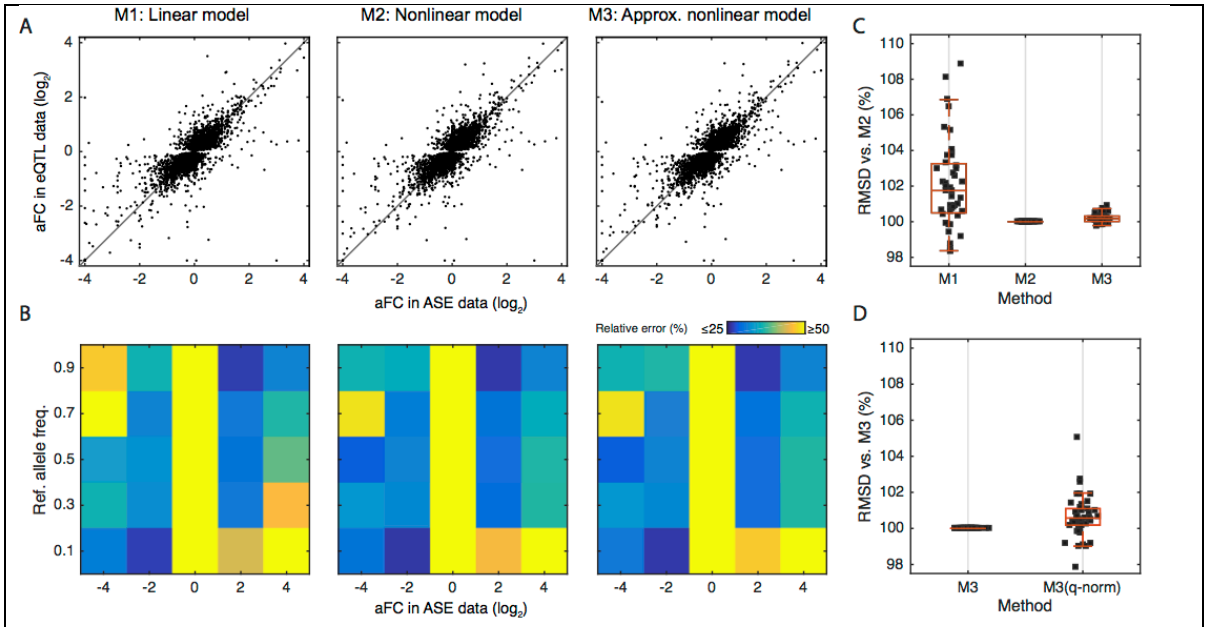

Figure 4: Comparison of the methods for estimating aFC using GTEx data. A) Allelic fold change as estimated from ASE data versus estimates from eQTL data using linear model (M1), nonlinear model (M2), and the nonlinear model approximation (M3) for all top eQTLs in Adipose Subcutaneous. All three estimates are ~75% correlated with estimates form 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.

The empirical distributions of aFCs for eQTLs detected in different GTEx tissues are highly dependent on the sample size, since tissues with lower sample size lack power to detect weak eQTLs (**Fig. 5A**). The effect size estimates from eQTL and ASE data are highly similar, but on average 1.45% (CI: [1.3, 1.6]) smaller across the tissues when estimated from ASE data (**Fig. 5B, C**). This mild overestimation of the effect size involving weaker eQTLs 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 aFC with other

properties of the eVariant or eGene. Low-frequency eVariants tend to have higher effect sizes (**Fig. 5E**), likely due to differences in power to detect eQTLs and other statistical artifacts. 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 constraint are less likely to tolerate regulatory variants with high effect sizes. Further biological interpretation of effect sizes across eVariants in different annotations and eGenes of different biotypes and eQTLs that are tissue-specific or shared is described in ((Aguet et al. 2016); co-submitted). 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.

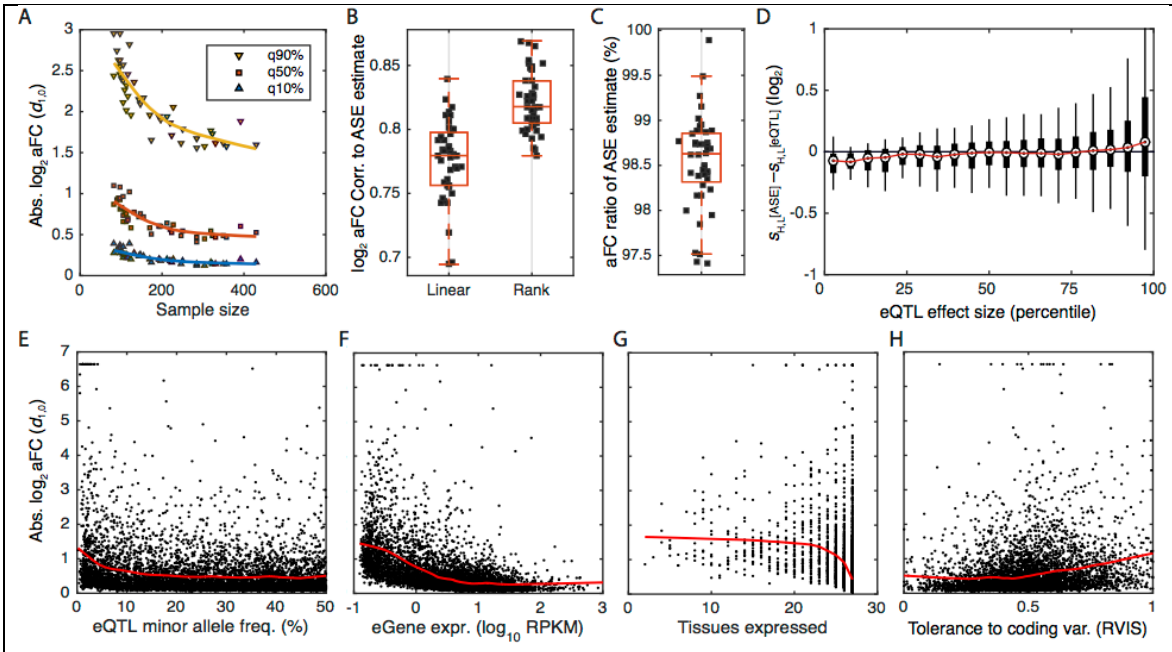

Figure 5: Empirical properties of the aFC distributions in GTEx data. All aFC values are calculated with the nonlinear approximation method (M3). A) Distribution of absolute $\log_2$ aFC across tissues as a function of sample size. Each point represents a tissue in GTEx data, and 90%, 50%, and 10% quantiles of absolute aFC across a tissue are shown. B-C) Correlation of $\log_2$ aFC estimates (B), and the median ratio of aFC 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 individuals (G), and logistic-transformed RVIS, a measure of each gene's tolerance to variation in the coding region ((Petrovski et al. 2013); H). Red line shows fit by robust locally weighted scatterplot smoothing.

The allelic fold changes of GTEx eQTLs are provided in the GTEx portal (see Data Access section). Additionally, we implemented the linear model (M1) and the nonlinear approximation model (M3) in a python script (see **Data Access**) 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 datasets straightforward and fast.

5. Independent eQTLs in GTEx

Iterative greedy procedures have been utilized to find multiple independent eQTLs signals for each eGene in the GTEx data (co-submitted (Aguet et al. 2016)). We used GTEx eGenes with two independent eQTLs to demonstrate how the aFC calculation can be extended to gain mechanistic insight into more complex eQTL patterns. The expression model in **Eq. 7** written for two biallelic eVariants was used in a nonlinear regression to simultaneously estimate the aFC associated with both eQTLs (**Fig. 6A; Methods**). 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**). While the used model accounts for specific arrangement of the alleles for the two eVariants on haplotypes ($e_{\langle 11\rangle,\langle 00\rangle} \neq e_{\langle 10\rangle,\langle 10\rangle}$, **Eq. 7**), it assumes that the two eVariants act independently ($e_{\langle 11\rangle} = e_{\langle 00\rangle}\delta_{1,0}^{v1}\delta_{1,0}^{v2}$, **Eq. 5**). In order to analyze how well the data is described assuming the independence of the two eVariants, we relaxed this assumption by defining the joint genotype of the two eVariants as the genotype of a hypothetical variant with four possible alleles. We used **Eq. 7** written for one four-allelic eVariant to separately estimate the aFC associated with each of the two eVariants, and 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 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 the regulatory independence model fails to provide an adequate fit **(Fig. 6D, Table S1)**.

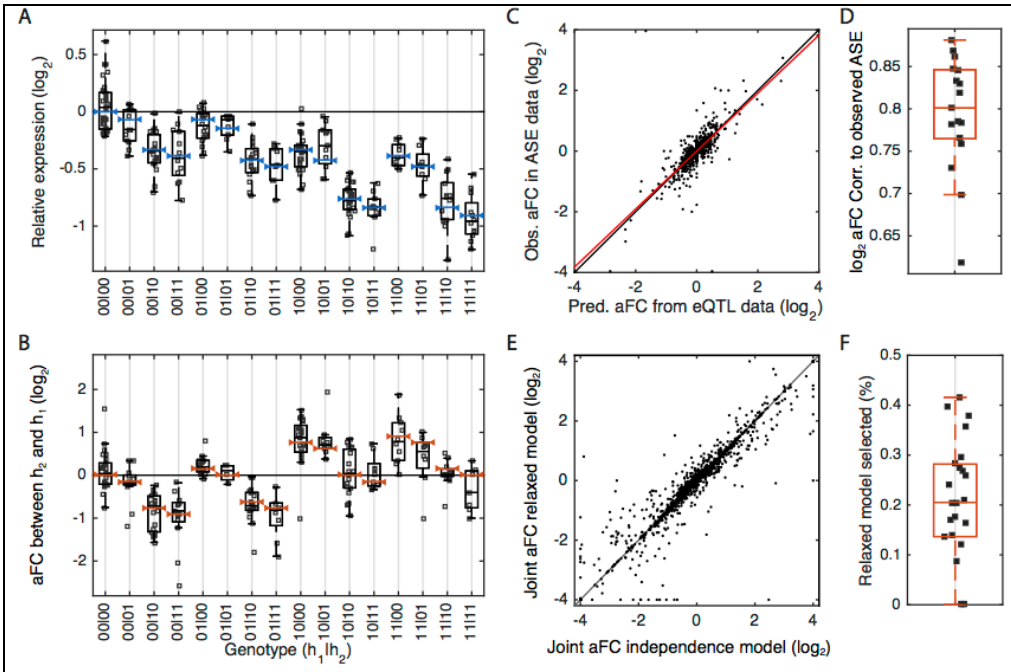

Figure 6: Joint analysis of aFCs for GTEx eGenes with two eQTLs. A) An example of relative expression of eGene *ZC3H3* and the model fits for different genotype groups of its two eQTLs (eVariant1: chr8:144633728 A/G and eVariant2: chr8:144556836 G/A) in GTEx Adipose Subcutaneous. The effect size of the 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 for 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 to aFC observed in ASE data for all eGenes with two eQTLs in Adipose Subcutaneous. Each dot represents one randomly selected genotype for on eGene. Red line indicates the robust linear fit (y=0.9x+0.002). D) Predicted and observed allelic fold change 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.

Discussion

Despite over a decade of eQTL analysis and its increasingly widespread use in functional and medical genetics, eQTL effect size has lacked a clear, biologically interpretable, and computationally feasible definition. Here, we described log allelic fold change, a generalizable measure of *cis*-regulatory effect size that captures the independent regulation of haplotype expression in *cis*. Log aFC is consistent across expression levels, allele frequencies and holds mathematically convenient properties that facilitate its application for downstream analysis. 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. 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, but aFC is a general measure and the presented methods 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. 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 better understanding of the effects of regulatory variants at many levels. In downstream analyses of GTEx effect sizes (co-submitted (Aguet et al. 2016)), we investigate differences in effect sizes between 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 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. Furthermore, the additive nature of log aFC makes it a useful tool for characterization of variation in eQTL activity across cellular or environmental contexts in the future. 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 (Vockley et al. 2015; Arnold et al. 2013) 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 to *in vitro* quantification of regulatory variant effects.

Methods

1. 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 (aseSNP) 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 **Eq.1** in a given sample on a relative scale. Since both measurements are drawn from the same sample they share the same basal expression (e_B in **Eq. 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 allelic fold change 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 n^{th} individual for haplotype carrying reference and alternative allele for the *cis*-regulatory variant respectively. Here we assume phasing between the regulatory alleles and the aseSNP alleles are known. In cases when phasing information is not available the magnitude of the regulatory effect size can be calculated as

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

However, this estimate without phasing information is more sensitive to noise, and will systematically overestimate the effect size in cases where the true effect size is small in magnitude and the variation in allelic counts is dominated by measurement noise.

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

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

Using **Eq. 4** we can derive gene expression in an individual as function of the number of alternative alleles, t:

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

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

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

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

showing that total gene expression under a *cis*-regulatory model is a linear with 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.

2.2 Estimating aFC from eQTL data with additive noise

Under an additive noise model, the measured gene expression in the n^{th} individual, y_n, is the true expression, $e(t)$, plus a normally distributed noise, ε_n, with zero mean and unknown variance. Using $e(t)$ from **Eq. 10**:

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

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

$$y_n = b_1 t_n + b_0 + \varepsilon_n \quad (14)$$

Maximum likelihood estimates for b_0 and b_1 can be derived efficiently using ordinary least squares, and solving **Eqs. 12a** and **12b**, for $\delta_{1,0}$, the allelic fold change is:

$$\delta_{1,0} = \frac{2b_1}{b_0} + 1 \quad (15)$$

2.3 Estimating aFC from eQTL data with multiplicative noise

Assuming a multiplicative noise model the measured gene expression in the n^{th} individual, y_n, is the true expression, $e(t)$, multiplied by a noise, ε_n, such that log ε_n is normally distributed with zero mean and unknown variance. Substituting $e(t)$ from **Eq. 10** again:

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

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

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

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 with the number of alternative alleles (**Fig. 1D**). Using nonlinear least squares optimization **Eq. 17** can be solved to derive maximum likelihood estimates for the effect size $\delta_{1,0}$ directly.

2.4 Efficient approximation of aFC from eQTL data with multiplicative noise

Nonlinear least squares optimization needed for solving regression problem in **Eq. 17** is done using iterative numerical optimization that is 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:

$$\delta_{1,0}^{*1} = \frac{m_2}{m_0} \quad (18a)$$

$$\delta_{1,0}^{*2} = \frac{1}{2\frac{m_1}{m_2} - 1} \quad (18b)$$

$$\delta_{1,0}^{*3} = 2\frac{m_1}{m_0} - 1 \quad (18c)$$

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 **Eq. 17** can be well approximated with linear regression in cases where the effect size is small. We regress log-transformed expressions on the genotype:

$$z_n = c_1 t_n + c_0 + \log_2 \varepsilon_n \quad (19)$$

and calculate the fourth effect-size estimate as (See **Supplemental methods**)

$$\delta_{1,0}^{*4} = 2^{2c_1} \quad (20)$$

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

$$r_n(k) = z_n - \log_2\left[(2 - t_n) + t_n\delta_{1,0}^{*k}\right] \quad (21)$$

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

$$\delta_{2,1} = \underset{k\in 1\ldots 4}{\operatorname{argmin}} \mathrm{V}[r(k)] \quad (22)$$

3. Simulation experiment

The simulated dataset 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 ($f_0 \sim$ *uniform*[0,1]). eQTL genotype in each individual was decided using two Bernoulli trials. Reference and alternative allele induce expressions e_0, and $e_1 = \delta_{1,0}\, e_0$ in the eGene in *cis*-, respectively (**Eqs. 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 mean-variance relationship resembling that of multiplicative noise by log-normal distribution (**Fig. S1**). 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, σ = 0.17], **Eq. 17**).

4. Estimating aFC for GTEx eQTLs

Haplotypic counts were generated as describe in ((Aguet et al. 2016) co-submitted). 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 8 reads, those that showed bias in mapping simulations (Panousis et al. 2014), had a UCSC 50-mer mappability lower than 1, or those without evidence for heterozygosity (Castel et al. 2015), were filtered. Haplotypic counts were generated by summing allelic counts within each gene using population phasing. For eQTL data, expression counts were scaled for the total library size, and one pseudo-count was added to smooth the normalized counts. Log transformed expression data was 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, log transformed expression profile of the eGene in n^{th} sample, z_n, was modeled using linear regression:

$$z_n = \mu + \alpha C_n + \beta_{t_n} + \varepsilon_n \quad (23)$$

where, C_n is the n^{th} 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 n^{th} sample. All non-significant columns, for which 95% confidence interval of the regression coefficient in α overlapped zero, were discard from C. In the second step, the regression was repeated using the reduced covariate matrix and corrected expression were derived as

$$\hat{z} = z - \alpha C \quad (24)$$

Corrected expression vector, $\hat{z}$, was used for effect size calculations. For direct estimation of aFC from **Eq. 17** (the Nonlinear method, M2, in **Fig. 3, 4**), we used Matlab generic nonlinear least square solver (*lsqnonlin*). The effect size estimates used in **Fig. 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).

5. Independent eQTL calling

Multiple independent signals for a given expression phenotype were identified by forward stepwise regression followed by a backwards 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 an independent signal and the forward step moves on to the next iteration. The backwards 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.

6. Joint analysis of two eQTLs

6.1 Regulatory independent model

Let us assume two biallelic eVariants, v_1 and v_2 regulating expression of the same eGene in *cis*. This is a special case of **Eq. 5-7** where N=2, and m_1=m_2=2. Under independence assumption, 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 therefor, the expression of a haplotype carrying alleles i_1 and i_2 for the two eVariants is

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

where indices $i_1, i_2 \in \{0,1\}$ indicate reference (zero) and the alternative allele (one), and $\delta^{v1}_{i_1,0}$, 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

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

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 on the second haplotype, respectively. From definition of aFC

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

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

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

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

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

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

$$e_{\langle i_1 i_2\rangle,\langle j_1 j_2\rangle} = e_{\langle i_1 i_2\rangle} + e_{\langle j_1 j_2\rangle} \quad (30)$$

Substituting Haplotypic expressions from **Eq. 25**, we can use measured expression profiles of genotyped individuals to estimate aFC associated with the two eVariants. Observed expression value for the eGene in the n^{th} sample after log transformation is

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

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 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 **Eq. 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 **Fig. 6B-E**, and **Fig. S2C** (blue bars) were derived using **Eq. 28**.

6.2 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 **Eq. 25**, haplotype expression is

$$e_{\langle i_1 i_2\rangle} = e_0 \delta_{\langle i_1 i_2\rangle,\langle 00\rangle} \quad (32)$$

where, $\delta_{\langle i_1 i_2\rangle,\langle 00\rangle}$ is the aFC associated to co-presence of the alleles i_1 and i_2 of the eVariants v_1 and v_2 as compared to a haplotype carrying reference allele for both eVariants. This model is equivalent to a special case of models in **Eq. 5-7** where N=1, and m_1=4. From aFC definition

$$\delta_{\langle i_1 i_2\rangle,\langle 00\rangle} = \delta_{\langle i_1 i_2\rangle,\langle j_1 j_2\rangle}\delta_{\langle j_1 j_2\rangle,\langle 00\rangle} \quad (33)$$

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

$$s_{\langle i_1 i_2\rangle,\langle j_1 j_2\rangle} = \log_2 \delta_{\langle i_1 i_2\rangle,\langle j_1 j_2\rangle} = s_{\langle i_1 i_2\rangle,\langle 00\rangle} - s_{\langle j_1 j_2\rangle,\langle 00\rangle} \quad (34)$$

Total expression is the sum of the individual haplotypic expressions (**Eq. 30**), thus, the observed expression value for the eGene in the n^{th} sample under the relaxed regulatory model after log transformation is

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

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 **Eq. 31**. The nonlinear regression problem can be solved for reference expression e_0, joint aFC effects $\delta_{\langle 10\rangle,\langle 00\rangle}$, $\delta_{\langle 01\rangle,\langle 00\rangle}$, $\delta_{\langle 11\rangle,\langle 00\rangle}$ and the cofactor weight vector α (By definition $\delta_{\langle 00\rangle,\langle 00\rangle}$ is equal to 1).

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 **Fig. 6E**, and **Fig. S2C** (red bars) were derived using **Eq. 34**.

6.3 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 log-normality assumption:

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

where, $\underline{z}$, is the vector of N samples, and r_n is the fit residual at the n^{th} 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:

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

where λ, the number of parameters in each model, is the number of cofactor coefficients plus 3 and plus 4 for the regulatory independence, and the relaxed model, respectively. We used bias-corrected and accelerated bootstrap (Efron 2012) to estimate confidence intervals for ΔBIC = BIC(Relaxed model) – BIC(Independence model) in cases where ΔBIC negative. The relaxed model was selected in cases were the upper bound for the 95% confidence interval for ΔBIC fell below zero, and for the rest of the cases the independence model that has fewer parameters was deemed adequate. Calculated aFCs for all eGenes in GTEx with two associated eQTLs are provided in **Table S1**.

Data access

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 [to be released at publication], are available from the GTEx Portal (http://gtexportal.org). Software for calculating allelic fold change from standard eQTL data is available in GitHub (https://github.com/secastel/aFC).

Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. 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, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. 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 a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & 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. TL and PM are supported by NIH grant R01MH106842, TL is supported by the NIH grant UM1HG008901, and TL and SEC are supported by the NIH contract HHSN2682010000029C 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.

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.

Disclosure declaration

The authors declare no competing financial interests.

References

Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV, Zappala Z, Abell NS, Fresard L, Gamazon ER, et al. 2016. *Local genetic effects on gene expression across 44 human tissues*. Cold Spring Harbor Labs Journals.

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.

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*.

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*.

Fish AE, Capra JA, Bush WS. 2016. Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? *The American Journal of Human Genetics*.

Flutre T, Wen X, Pritchard J, Stephens M. 2013. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues ed. G. Gibson. *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.

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

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

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. ed. C.D. Brown. *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 USA* 201503027.

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.

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. ... *genetics*.

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.

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 ed. S.M. Williams. *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.

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*.

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

Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. 2015. Author response ed. E.T. Dermitzakis. *eLife* **4**: 1061.

Ulirsch JC, Nandakumar SK, Wang L, Giani FC. 2016. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*.

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 USA* **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 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.

Wright JB, Sanjana NE. 2016. CRISPR Screens to Discover Functional Noncoding Elements. *Trends in Genetics* **32**: 526–529.

Supplemental methods

Derivations and proofs

1. **Proof: Log-transformed eGene expression is linear in the number of alternative alleles as the *cis*-regulatory effect size approaches zero.**

   Let α_1 and α_2 be the slope of the line connecting eGene expressions from reference homozygous to heterozygous, and from heterozygous to homozygous alternative genotype, respectively, in the piecewise linear model of log-transformed eQTL data (**Fig. 1C**):

   $$\begin{aligned}\alpha_1 &= \log e_{1,0} - \log e_{0,0} \\ \alpha_2 &= \log e_{1,1} - \log e_{1,0}\end{aligned} \tag{S1}$$

   In a linear model α_1 is equal to α_2. Substituting the allelic expressions from the main text **Eq. 4**, the ratio between the two slopes for weak eQTLs is

   $$\lim_{s_{1,0}\to 0} \frac{\alpha_1}{\alpha_2} = \lim_{s_{1,0}\to 0} \frac{\log(2^{s_{1,0}}+1) - \log 2}{\log 2 + \log 2^{s_{1,0}} - \log(2^{s_{1,0}}+1)} \tag{S2}$$

   where, $s_{1,0} = \log_2 \delta_{1,0}$ is the eQTL effect size. Since, the limit value for both nominator and the denominator is 0, we apply L'Hôpital's rule

   $$\lim_{s_{1,0}\to 0} \frac{\alpha_1}{\alpha_2} = \lim_{s_{1,0}\to 0} \frac{\alpha_1'}{\alpha_2'} = \lim_{s_{1,0}\to 0} \frac{\frac{1}{2^{s_{1,0}}+1}}{\frac{1}{2^{s_{1,0}}} - \frac{1}{2^{s_{1,0}}+1}} = \frac{\frac{1}{1+1}}{\frac{1}{1} - \frac{1}{1+1}} = 1 \tag{S3}$$

   Thus, the two slopes, α_1 and α_2 are equal in weak eQTLs as $s_{1,0} \to 0$.

2. **Derivations: Approximate nonlinear model for aFC estimation**

   Let us assume t_n is the number of alternative allele in n^{th} sample, and 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. First, we use the expression ratio between each of the two genotype classes to estimate aFC. From **Eq. 17**, the expected log-transformed expression at each eQTL genotype class is

   $$\mathrm{E}[z_n|t_n = 0] = \log_2 e_0 + 1 \tag{S4a}$$
   $$\mathrm{E}[z_n|t_n = 1] = \log_2 e_0 + \log_2(\delta_{1,0} + 1) \tag{S4b}$$
   $$\mathrm{E}[z_n|t_n = 2] = \log_2 e_0 + \log_2 \delta_{1,0} + 1 \tag{S4c}$$

   Using **Eqs. S4a**, and **S4c**, the $\log_2$ aFC is

   $$\mathrm{E}[z_n|t_n = 2] - \mathrm{E}[z_n|t_n = 0] = \log_2 \delta_{1,0} \tag{S5}$$

   Substituting observed geometric means $m_t = 2^{\mathrm{E}[z_n|t_n=t]}$, and exponentiating both sides of the equation, the aFC is

   $$\delta_{1,0} = \frac{m_2}{m_0} \tag{S6}$$

   Next, we use **Eqs. S4b**, and **S4c**:

   $$\mathrm{E}[z_n|t_n = 2] - \mathrm{E}[z_n|t_n = 1] = \log_2 \delta_{1,0} + 1 - \log_2(\delta_{1,0} + 1) \tag{S7}$$

   Exponentiating the both sides we have

   $$\frac{2^{\mathrm{E}[z_n|t_n=2]}}{2^{\mathrm{E}[z_n|t_n=1]}} = \frac{2\delta_{1,0}}{\delta_{1,0}+1}$$

   after substituting geometric means and rearranging the terms, the aFC is given:

   $$\delta_{1,0} = \frac{1}{2\frac{m_1}{m_2} - 1} \tag{S8}$$

   Using **Eqs. S4a**, and **S4b**

$$\mathrm{E}[z_n|t_n = 1] - \mathrm{E}[z_n|t_n = 0] = \log_2(\delta_{1,0} + 1) - 1 \quad \text{(S9)}$$

aFC can be similarly derived:

$$\delta_{1,0} = 2\frac{m_1}{m_0} - 1 \quad \text{(S10)}$$

As a fourth estimate, we use loglinear regression to derive another aFC estimate. This is an accurate model for weak eQTLs where the piece-wise linear eQTL model approaches linearity (see **Eqs. S1-3**). The regression line passes $\mathrm{E}[z_n|t_n = 0]$ at t_n= 0, and $\mathrm{E}[z_n|t_n = 2]$ at t_n= 2, therefore the slope, c_1, of the line is

$$c_1 = \frac{\mathrm{E}[z_n|t_n = 2] - \mathrm{E}[z_n|t_n = 0]}{2 - 0} = \frac{\log_2 \delta_{1,0}}{2} \quad \text{(S11)}$$

Thus aFC is given as

$$\delta_{1,0} = 2^{2c_1} \quad \text{(S12)}$$

It is worth noting that under the *cis*-regulatory model of **Eqs. 4a-c**, the expression in the heterozygous class is at least half of that of the higher expressed homozygous class, taking place when the weak allele is effectively zero expressed, thus:

$$-\infty \geq \mathrm{E}[z_n|t_n = 2] - \mathrm{E}[z_n|t_n = 1] \geq 1 \quad \text{(S13a)}$$

$$-\infty \geq \mathrm{E}[z_n|t_n = 0] - \mathrm{E}[z_n|t_n = 1] \geq 1 \quad \text{(S13b)}$$

In practice, the observed expression of the genotype classes, m_0, m_1, and m_2, can occasionally fall outside these boundaries due to noise or other confounding biological factors beyond the considered *cis*-regulatory model. Therefore, the ratios $\frac{m_1}{m_0}$ and $\frac{m_1}{m_2}$ in **Eqs. S8** and **S10** should be bound to be ≥0.5 to avoid negative aFC estimates.

3. Mathematical properties of log aFC

Recalling log aFC definition:

$$\begin{aligned} s_{\mathrm{i,j}} &= \log_2 \delta_{\mathrm{i,j}} \\ &= \log_2 e_1 - \log_2 e_0 \\ &= \log_2 k_i - \log_2 k_j \end{aligned}$$

We show that the following statements are true:

a. Zero log aFC indicates the absence of regulatory difference: $s_{i,i} = 0$

$$s_{i,i} = \log_2 k_i - \log_2 k_i = 0$$

b. Choice of reference allele only affects the sign of log aFC: $s_{i,j} = -s_{j,i}$

$$\begin{aligned} s_{i,j} &= \log_2 k_i - \log_2 k_j \\ &= -(\log_2 k_j - \log_2 k_i) \\ &= -s_{j,i} \end{aligned}$$

c. Log aFC is additive: $s_{i,k} = s_{i,j} + s_{j,k}$

$$\begin{aligned} s_{i,k} &= \log_2 k_i - \log_2 k_k \\ &= \log_2 k_i - \log_2 k_k + \log_2 k_j - \log_2 k_j \\ &= (\log_2 k_i - \log_2 k_j) + (\log_2 k_j - \log_2 k_k) \\ &= s_{i,j} + s_{j,k} \end{aligned}$$

d. aFC associated with joint effect of independent regulatory variants, *v*1…*v*N is sum of their individual log aFCs:

$$\mathbf{s}_{\langle i_1 \ldots i_n \ldots i_N \rangle, \langle j_1 \ldots j_n \ldots j_N \rangle} = \sum_{n=1}^{N} \mathbf{s}_{i_n, j_n}^{v\mathbf{n}}$$

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

Assuming that variants affect gene expression independently, haplotype expression in **Eq. 1** in the main text can be written for N eVariants as

$$e_{\langle i_1 \dots i_n \dots i_N \rangle} = e_B \prod_{n=1}^{N} k_{i_n}^{vn}$$

where $k_{i_n}^{vn}$ denotes the regulatory effect on the eGene expression specific to allele i_n of the n^{th} eVariant. Therefore, the joint aFC is

$$\begin{aligned} s_{\langle i_1 \dots i_n \dots i_N \rangle, \langle j_1 \dots j_n \dots j_N \rangle} &= \log_2 \frac{e_{\langle i_1 \dots i_n \dots i_N \rangle}}{e_{\langle j_1 \dots j_n \dots j_N \rangle}} \\ &= \log_2 \frac{e_B \prod_{n=1}^{N} k_{i_n}^{vn}}{e_B \prod_{n=1}^{N} k_{i_n}^{vn}} \\ &= \log_2 \prod_{n=1}^{N} \frac{k_{i_n}^{vn}}{k_{j_n}^{vn}} \\ &= \sum_{n=1}^{N} \log_2 \frac{k_{i_n}^{vn}}{k_{j_n}^{vn}} \\ &= \sum_{n=1}^{N} s_{i_n, j_n}^{vn} \end{aligned}$$

e. Absolute value of log aFC, $d_{i,j} = |s_{i,j}|$, is a pseudo-metric:

i. $d_{i,j} \geq 0$

ii. $d_{i,i} = 0$

iii. $d_{i,j} = d_{j,i}$

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

The first condition is met by definition and the second and third conditions are trivial considering the aFC properties **S.I** and **S.II** shown above. In order to demonstrate the truth of the fourth condition we consider two cases:

1) When $s_{i,j}$ and $s_{j,k}$ are both positive or both negative; in such cases due to additivity of log aFC (Statement **S.III**), $s_{i,k}$ will also have the same sign, and therefore, $d_{i,k} = d_{i,j} + d_{j,k}$ is trivial.

2) When $s_{i,j}$ and $s_{j,k}$ have different signs; Let us assume $s_{i,j} \geq 0$ and $s_{j,k} \leq 0$, from **S.III**:

$$\begin{aligned} s_{i,k} &= s_{i,j} + s_{j,k} \\ &= d_{i,j} - d_{j,k} \\ &\leq d_{i,j} + d_{j,k} \end{aligned}$$

Additionally,

$$\begin{aligned} -s_{i,k} &= -(s_{i,j} + s_{j,k}) \\ &= d_{j,k} - d_{i,j} \\ &\leq d_{i,j} + d_{j,k} \\ &\Rightarrow s_{i,k} \geq -(d_{i,j} + d_{j,k}) \end{aligned}$$

Combining the last two statements $d_{i,k} = |s_{i,k}| \leq -d_{i,j} + d_{j,k}$. The opposite case where $s_{i,j} \leq 0$ and $s_{j,k} \geq 0$, is the same.

Supplemental figures

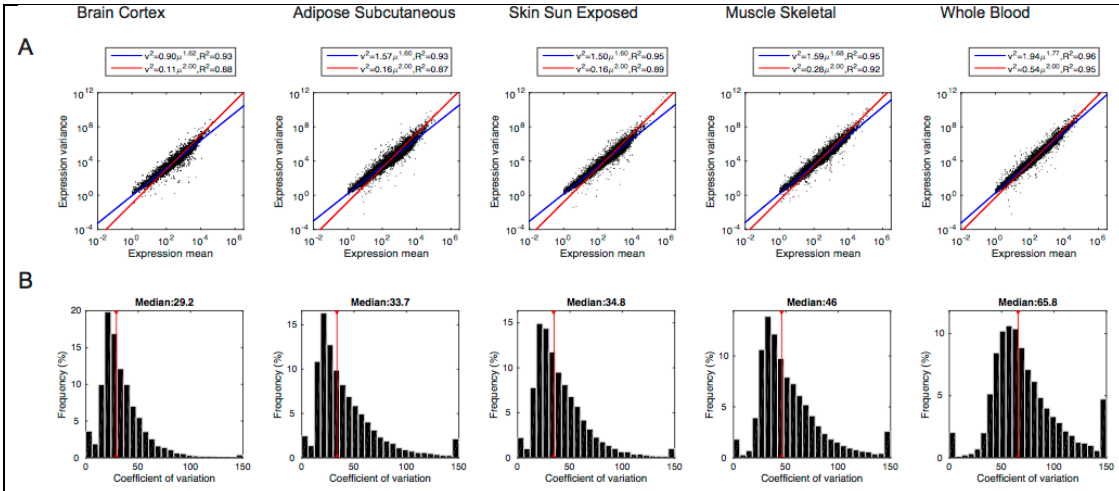

Figure S1: Gene expression noise distribution in GTEx data. A) Mean and variance of eGene expression within genotype classes of the top eQTL for five example tissues in GTEx data. Each dot corresponds to data from one eGene within an eQTL genotype class. Red line indicates the best expected mean-variance dependence from lognormal distributed data, and blue lines shows the optimal linear regression line. This pattern shows that variance structure eQTL data is highly similar to lognormal distribution. B) Coefficient of variation, the ratio between the standard deviation and mean, for eGene expression within eQTL genotype classes for the same tissues.

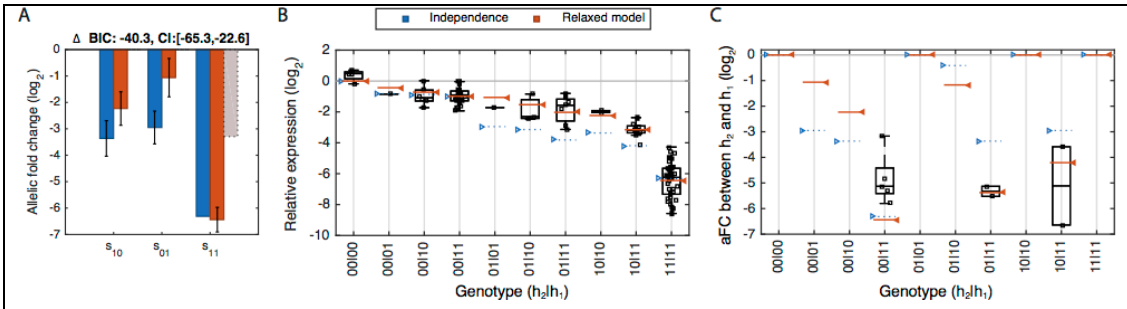

Figure S2: An example of two eQTLs regulating the expression of the same gene that is not well described by their individual regulatory effects acting independently (eGene: *HLA-DQB1-AS1*; eVariant1: chr6:32627082 A/G and eVariant2: chr6:32609813 T/C; Tissue: LCL) A). Estimated log aFC associated with the alternative alleles for the first (s_{10}), and the second eSNP (s_{01}) individually, along with the estimated log aFC associated with co-occurrence of the alternative alleles (s_{11}). The independent regulation model is shown in blue, where s_{10} , s_{01} and s_{11} are estimated from the data with the constraint of $s_{11} = s_{10} + s_{01}$. The red bars show estimates from the alternative, relaxed model which allows for non-independence or epistatic-like interaction between the two eVariants, and s_{10} , s_{01} and s_{11} are estimated without assuming $s_{11} = s_{10} + s_{01}$. The support for non-independent effects comes from the difference between this estimated s_{11} to the sum of s_{10} and s_{01} from the relaxed model (gray dashed bar), which represents the expected joint effect of the two alternative alleles had they acted independently. B) Relative expression of the eGene and the model fits for the different genotype classes. Each dot is the expression observed in one individual, and expression levels are shown relative to the all-reference genotype. The blue and red bars show best fits achieved with and without the regulatory independence assumption, respectively. The model assuming regulatory independence between the two eVariants fails to adequately describe the observed data as measured by significance of BIC difference. C) Expression of the second haplotype relative to the first haplotype shown for different genotype groups. The dots indicate the observed values in ASE data and the blue and red bars show predicted values from the model fitted on eQTL data (as shown in panel B) using regulatory independence and the relaxed model, respectively. Genotypes in panel B and C are labeled following the notation in **Fig. 6**, and classes identical with regard to the *cis*-regulatory model are collapsed together in each panel.