Haplotype-based eQTL mapping finds evidence for complex gene regulatory regions poorly tagged by marginal SNPs

Robert Brown ^{1*}, Sriram Sankararaman ^{1,2}, Bogdan Pasaniuc ^{1,2,3}

¹ *Department of Computer Science, UCLA, Los Angeles, 90242, USA*

² *Department of Human Genetics, UCLA, Los Angeles, 90242, USA*

³ *Pathology and Laboratory Medicine, UCLA, Los Angeles, 90242, USA*

Abstract

Motivation: Expression quantitative trait loci (eQTLs), variations in the genome that impact gene expression, are identified through eQTL studies that test for a relationship between single nucleotide polymorphisms (SNPs) and gene expression levels. These studies typically assume an underlying additive model. Non-additive tests have been proposed, but are limited due to the increase in the multiple testing burden and are potentially biased by filtering criteria that relies on marginal association data. Here we propose using combinations of short haplotypes instead of SNPs as predictors for gene expression. Essentially, this method looks for genomic regions where haplotypes have different effect sizes. The differences in effect can be due to multiple genetic architectures such as a single SNP, a burden of rare SNPs, multiple SNPs with independent effect or multiple SNPs with an interaction effect occurring on the same haplotype.

Results: Simulations show that when haplotypes, rather than SNPs, are assigned non-zero effect sizes, our method has increased power compared to the marginal SNP method. In the GEUVADIS gene expression data, our method finds 101 more eGenes than the marginal method (5,202 vs. 5,101). The methods do not have full overlap in the eGenes that they find. Of the 5,202 eGenes found by our method, 707 are not found by the marginal method—even though it has a lower significance threshold. This indicates that many genes have regulatory architectures that are not well tagged by marginal SNPs and demonstrates the need to better model alternative architectures.

1. Introduction

Expression quantitative trait loci (eQTLs) are genetic variants that regulate gene expression. Association scans to find eQTLs have been successfully applied to multiple datasets to find many eGenes, genes with at least one eQTL regulating their expression. These studies have shown that much of the variation in gene expression is heritable [1, 2, 3, 4] and that the genetic architectures that regulate expression are often found near the gene they regulate [5, 6]. The GEUVADIS [7] and GTEx [8] projects have publically provided genotype and expression data for hundreds of individuals across multiple tissues. This data has been successfully used to probe how genetic variation influences complex diseases through gene expression regulation [9, 10, 11, 12].

The standard test of association assumes an underlying additive model [13, 14] that relates genotype to expression level using a marginal single nucleotide polymorphism (SNP) test. eQTLs found through studies using this model may have independent additive effects, but could also have more complicated interactions that are only tagged by the marginal SNP test. The marginal SNP test does not account for the presence of multiple eQTLs existing for the same eGene [15] or for interactions between eQTLs for the same gene [16]. Recent work [17, 18, 19, 20] has searched for evidence of eQTLs arising from SNPxSNP interactions where one SNP moderates the effect of another. Such interactions are known to exist in yeast [21], but their relevance to gene expression in humans has been difficult to ascertain [22, 23, 24]. Other work has found evidence of haplotype effects when using a likelihood model for short haplotypes comprised of 10 SNPs [25] and for haplotypes defined by regions with a high likelihood of affecting gene expression [26]. However, these methods work for regulatory regions and must leverage functional predictions in order to reduce the computational burden to a small fraction of the genome. This makes their extension to complex phenotype studies problematic since they are restricted to only testing haplotypes within a megabase of genes.

In this work we present a new method (HapSet) for investigating haplotype effects on phenotype. While our method can be applied genome-wide with only a small increase in the computational burden, we apply it to gene expression data in order to test it against thousands of phenotypes. We hypothesize that short haplotypes each have their own specific independent

effect on gene expression. The HapSet method divides haplotypes from a 10 kb region into all possible two set combinations and looks for a difference in the average effect size between the sets. A short region is selected so that the population can be characterized by a small number of haplotypes. This test allows each haplotype to have its own genetic regulatory architecture and effect size and only assumes that there is no interaction between haplotypes. In order to use the marginal SNP method as a subset of our method, we force our method to include haplotype sets defined by the presence or absence of an alternate allele at specific SNP locations. Our method is not biased by filtering or variant selection that utilize marginal test statistics since it determines haplotype sets independent of the expression data. We compute significance thresholds for the haplotype set tests in order to control the family-wise error rate (FWER) at desired levels.

We simulate gene expression assuming five underlying architectures: single causal SNPs, multiple causal SNPs, SNPxSNP interactions, haplotypes with non-zero effects and a null model. Our simulations show that both methods maintain a 0.05 FWER under the null model. With the common SNP-based simulated architectures, the marginal SNP method has slightly higher power due to a less stringent significance threshold. We expect this result since the SNP method is a subset of the HapSet method. For example, when a single common SNP accounts for 5% of the variance in phenotype, the SNP method has 78% power while the HapSet method has 71%. However, when the underlying model is based on a random set of haplotypes assigned the same non-zero effect size, the HapSet method has 71% power compared to the SNP method that only has 56%. This demonstrates that there is insufficient SNP density to tag many of the possible haplotype combinations.

We apply our method to find eGenes with the GEUVADIS data using expression and genotype data from 373 Europeans individuals and find evidence for many genes regulated by complex genetic architectures. Of the 18,621 genes in the data, the HapSet method finds 707 eGenes where the marginal SNP method had no significant association despite having a lower significant threshold. This demonstrates that for these 707 genes, there is likely to be a complex regulatory architecture that is poorly tagged by marginal SNPs. In these genes, the top HapSet association on average explains 6% of the variance in gene expression while the top (non-significant) marginal SNP only explains 3% of the variance. The marginal SNP method finds 606 eGenes not identified by the HapSet method. Since the marginal SNP method is a subset of the HapSet method, the HapSet method did not identify these

74 eGenes due to its more stringent significance threshold. However, the top as-
 75 sociations both explained 5% of the variance in expression. In addition to the
 76 eGenes found by only one of the methods, there are 4,495 eGenes where both
 77 HapSet and the marginal SNP method find significant associations with the
 78 top associations on average each explaining 10% of the variance in expression.
 79

80 2. Methods

81 A SNP represents variation at one specific location in the genome. Hap-
 82 lotypes represent variation across a set of successive SNPs on a single chro-
 83 mosome. Generally, a small number of haplotypes are representative of the
 84 majority of the haplotypes in a sample when looking at a small region. Each
 85 of these haplotypes can have different effect sizes depending on the genetic
 86 architectures occurring on the haplotypes. Our proposed method seeks to
 87 maximize the difference between the frequency-weighted mean effect size on
 88 gene expression of haplotypes within a defined set (a HapSet) and those not
 89 in the set. Once a set is defined, haplotypes from individuals are identified
 90 as either in or out of the set, and a pseudo-genotype, defined as the number
 91 of haplotypes an individual carries that are within the set, is assigned. Gene
 92 expression can then be regressed on this pseudo-genotype to estimate the
 93 mean difference in effect size of haplotypes in and out of the set.

94 2.1. Haplotype effect model

95 We divide the genome up into S equal length sections. We assume that
 96 the i^{th} haplotype in the s^{th} section has its own specific and independent effect
 97 β_{si} on gene expression y

$$y = \sum_{s=1}^S \sum_i \beta_{si} h_{si} + \varepsilon \quad (1)$$

98 Here y is the gene expression for an individual, h_{si} indicates the number
 99 of the i^{th} haplotype in the s^{th} section that an individual carries (either 0, 1 or
 100 2) and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. When only interested in finding associations between y
 101 and any of the haplotypes in a region s , we can rewrite our model (Equation
 102 2). Here ε now includes the variance due to all of the haplotypes not in
 103 region s as well as environmental noise.

$$y = \sum_i \beta_i h_i + \varepsilon \quad (2)$$

104 2.2. Marginal SNP method

105 The marginal SNP method for identifying an association between a SNP
106 and expression fits the following additive model

$$y = \beta_{g_j} g_j + \varepsilon \quad (3)$$

107 where g_j is the number of alternate alleles an individual carries at the j^{th}
108 SNP. Assuming the model in Equation 2, β_{g_j} will be the difference between
109 the frequency weighted average effect of haplotypes containing an alternate
110 allele at SNP j and the average effect of haplotypes that carry the reference
111 allele:

$$\beta_{g_j} = \sum_i \frac{\beta_i f(h_i) I(h_i(g_j) = 1)}{f(h_i) I(h_i(g_j) = 1)} - \sum_i \frac{\beta_i f(h_i) I(h_i(g_j) = 0)}{f(h_i) I(h_i(g_j) = 0)} \quad (4)$$

112 Here $I(h_i(g_j) = 1)$ is equal to 1 if h_i contains an alternate allele at
113 the j^{th} SNP (g_j) otherwise it is 0. Similarly, $I(h_i(g_j) = 0)$ is equal to 1 if h_i
114 contains a reference allele at the j^{th} SNP (g_j) otherwise it is 0. The frequency
115 of haplotype h_i is given by $f(h_i)$.

116 2.3. Haplotype set method

117 The haplotype set (HapSet) method fits an additive model similar to
118 Equation 3

$$y = \beta_H g_H + \varepsilon \quad (5)$$

119 where g_H is a pseudo-genotype for the number of haplotypes an individual
120 carries that are members of a set H of haplotypes (see Section 2.5). β_H is
121 the difference between the average effect size of haplotypes in set H and the
122 average effect size of haplotypes not in set H (i.e. in set H^c)

$$\beta_H = \sum_i \frac{\beta_i f(h_i) I(h_i(H))}{f(h_i) I(h_i(H))} - \sum_i \frac{\beta_i f(h_i) I(h_i(H^c))}{f(h_i) I(h_i(H^c))} \quad (6)$$

123 Here $(h_i(H))$ and $(h_i(H^c))$ are equal to 1 if haplotype h_i is a member of
124 set H (or H^c) and 0 otherwise.

125 2.4. Testing for significant associations

126 A standard measure of association is the Wald statistic for expression
127 levels y of n individuals and genotypes (or pseudo-genotypes) g with σ_g^2
128 variance

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{cov(g, y)\sqrt{n}}{\sqrt{\sigma_g^2\sigma_e^2}} \quad (7)$$

129 which asymptotically follows a normal distribution with variance 1 and a
130 non-centrality parameter given by

$$\lambda\sqrt{n} = \beta\sqrt{n\frac{\sigma_g^2}{\sigma_e^2}} \quad (8)$$

131 Since the $\mathbb{E}(\hat{\beta}) = \beta$, under the marginal SNP method, testing the SNP
132 with the largest β_g after standardizing g_j will result in the largest association
133 statistic for the marginal SNP test. Likewise, for the HapSet method, a set
134 H that maximizes β_H for a standardized g_H will maximize the association
135 statistic for the HapSet method.

136 2.5. Determining haplotype sets

137 Since there is no way to know *a priori* the H that maximizes the differ-
138 ence, we test all possible H s from all 10 kb windows within 500 kb of a gene's
139 transcript start site. To determine the H s, we remove all SNPs with minor
140 allele frequency < 0.05 and find all the haplotypes in a 10 kb region. We
141 choose a small region so that a small number of haplotypes will be represena-
142 tive of the majority of the haplotypes in the sample. We treat all haplotypes
143 with frequency < 0.05 as a single haplotype. If there are more than 8 hap-
144 lotypes in the group, we use the 7 most frequently occurring haplotypes and
145 treat all the remaining haplotypes as the 8th haplotype. We then compute
146 all possible ways to form two sets (H and H^c). Using the phased data pro-
147 vided, for each H we encode for each individual the number of haplotypes
148 they carry that are contained in H as g_H . We compute a g_H for each H of
149 every 10 Kb region in the genome.

150 We force the marginal SNP method to be a subset of the HapSet method.
151 For each SNP with minor allele frequency > 0.05 , we define all haplotypes
152 with an alternate allele at the SNP location to be in set H and all haplotypes
153 with the reference allele at the SNP location to be in set H^c .

154 2.6. Power simulations

155 We simulate four causal architectures using genotype data from chromo-
 156 some 22 of the 373 Europeans in the GEUVADIS project. We use the non-
 157 transformed genotype data. We simulate using the following models where
 158 the proportion of the variance due to the underlying architecture is h^2 : null,
 159 single causal SNP, two causal SNPs, SNPxSNP interaction and a haplotype
 160 set model (Equations 9, 10, 11, 12 and 13 respectively). Causal SNPs (g_{C1}
 161 and g_{C2}) are randomly chosen from SNPs within 500 Kb of a transcription
 162 start site. All SNPs have minor allele frequency > 0.05 except when using
 163 a rare single causal SNP with frequency between 0.01 and 0.05. We fix all
 164 SNP or HapSet effect sizes to be 1 ($\beta_{C1} = \beta_{C2} = \beta_H = 1$) and add noise to
 165 achieve the desired h^2 . The haplotype set model assumes that a random set
 166 of haplotypes H all have the same non-zero effect size. We draw the causal
 167 H from one of the H s determined in section 2.5 that is within 500 kb of
 168 a transcription start site. When analyzing how the methods perform when
 169 the simulated causal haplotypes are not present in the data, (masked causal
 170 haplotypes), we do not test any SNPs or H s from the 10kb region that con-
 171 tains the causal haplotypes. We then regress the simulated expression on all
 172 genotypes and g_{HS} within 500 kb of the transcription start site. We simulate
 173 each trait 65,000 times for each h^2 and the null model 650,000 times.

$$y \sim \mathcal{N}(0, 1) \quad (9)$$

$$y = \beta_{C1}g_{C1} + \varepsilon$$

$$\text{where } \varepsilon \sim \mathcal{N}(0, \text{var}(\beta_{C1}g_{C1})\frac{1-h^2}{h^2}) \quad (10)$$

$$y = \beta_{C1}g_{C1} + \beta_{C2}g_{C2} + \varepsilon$$

$$\text{where } \varepsilon \sim \mathcal{N}(0, \text{var}(\beta_{C1}g_{C1} + \beta_{C2}g_{C2})\frac{1-h^2}{h^2}) \quad (11)$$

$$y = \beta_{C1C2}g_{C1}g_{C2} + \varepsilon$$

$$\text{where } \varepsilon \sim \mathcal{N}(0, \text{var}(\beta_{C1C2}g_{C1}g_{C2})\frac{1-h^2}{h^2}) \quad (12)$$

$$y = \beta_Hg_H + \varepsilon$$

$$\text{where } \varepsilon \sim \mathcal{N}(0, \text{var}\beta_Hg_H)\frac{1-h^2}{h^2}) \quad (13)$$

3. Results

3.1. Data for simulations and analysis

Our analyses are based on publically available genotype and lymphoblastoid expression data of 373 European individuals provided by the GEUVADIS [7] project. Following the GEUVADIS project, we only use expression from genes that had > 0 quantifications in $> 90\%$ of individuals. We center and standardize the RPKM and PEER normalized gene expression levels.

3.2. Controlling the family-wise error rate

Using all of the g_H values calculated from the chromosome 22 genotypes of the 373 European individuals in the GEUVADIS data and the genotype data for SNPs with minor allele frequency above 0.05, we compute significance thresholds using SLIDE [27]. Since eQTL testing only looks at predictors within 500 kb of each gene, we ensure a family-wise error rate (FWER) level for each gene by scaling the threshold for a 1 Mb region. For a 0.05 desired FWER, this results in a 2.0×10^{-5} significance threshold for our HapSet method. Running the same procedure using the genotype data for the SNP method, we estimate a per megabase significance threshold of 5.8×10^{-5} . The results from the different FWER levels (see Table 1) suggest that the HapSet method needs to correct for approximately 3 times as many independent tests in comparison to the marginal SNP method.

	Marginal SNP Method	HapSet Method
0.01 FWER	9.8×10^{-6}	3.3×10^{-6}
0.05 FWER	5.8×10^{-5}	2.0×10^{-5}
0.10 FWER	1.2×10^{-4}	4.0×10^{-5}

Table 1: Per megabase significance thresholds estimated with SLIDE. For the 0.01, 0.05 and 0.1 FDRs analyzed, the HapSet method performs approximately 3 times as many independent tests as the SNP method.

3.3. Power analysis

Our simulations show that the marginal SNP method and HapSet method are both well calibrated under the null with 0.043 and 0.046 discovery rates, respectively, when controlling for a 0.05 FDR. The power of both methods increases as the variance due to the underlying genetic architecture increases.

199 The discovery rate of the SNP method is slightly superior to that of the
 200 HapSet method when the underlying model is based on common SNPs (see
 201 Figure 1 and Table 2). However, if the underlying model is a single "rare"
 202 SNP with minor allele frequency between 0.01 and 0.05, the HapSet method
 203 outperforms the marginal SNP method. When the underlying model is a
 204 random set of haplotypes H , the HapSet method shows a substantial increase
 205 in power over the marginal SNP method (see Figure 2 and Table 2). For
 206 example, for an $h^2 = 0.05$ and a haplotype set architecture, the SNP method
 207 has 56% power compared to 71% for the HapSet method (see Table 2). When
 208 the set of causal haplotypes is masked, the HapSet method still outperforms
 209 the marginal SNP method.

	Maringal SNP Method	HapSet Method
Null ($h^2 = 0$)	0.043±0.0006	0.046±0.0006
Common SNP	0.78±0.0032	0.71±0.0036
Two SNPs	0.35±0.0036	0.30±0.0036
SNPxSNP Interaction	0.75±0.0034	0.7±0.0034
Rare SNP	0.20±0.0032	0.22±0.0032
Haplotype Set	0.56±0.0038	0.71±0.0036
Haplotype Set Masked	0.43±0.0038	0.47±0.0040

Table 2: Discovery rate ($\pm 2SE$) of the marginal SNP and HapSet methods when $h^2 = 0.05$. The marginal SNP method outperforms the HapSet method when the underlying genetic architecture is based on SNPs. The exception is when the underlying architecture is based on a 'rare' SNP with allele frequency between 0.01 and 0.05. When the architecture is based on haplotype sets, the HapSet method strongly outperforms the SNP method. When the haplotype sets and SNPs within 10kb of the simulated casual haplotype set are not included in the data for testing (masked), the HapSet method still outperforms the SNP method. This indicates that there is not enough SNP density to tag some haplotype combinations.

210 3.4. Analysis of GEUVADIS data

211 We apply the two methods to the GEUVADIS gene expression data. Fol-
 212 lowing the GEUVADIS analysis, we regress expression on the genotypes or
 213 HapSets (both standardized to mean 0 and variance 1) while controlling for
 214 the top three genotype-based PCs [7]. We use the 18,621 genes that passed
 215 the filtering criteria. We report in Table 3 the number of eGenes found by the

Figure 1: Discovery rate of the marginal SNP and HapSet methods with SNP-based simulated architectures. The proportion of variance due to the underlying genetic architecture is given by h^2 . The HapSet method slightly outperforms the SNP method for causal SNPs with minor allele frequencies between 0.01 and 0.05. Since the HapSet method has a more stringent significance threshold to control the FWER at 0.05, this indicates that the HapSet method is better tagging the causal SNP than the SNP method. For all other SNP-based causal architectures, the HapSet method performs slightly below the SNP method. Since the SNP method is a subset of the HapSet method, this is due only to the difference in significance thresholds.

Figure 2: Discovery rate of the marginal SNP and HapSet methods with haplotype-based simulated architectures. When a random set of haplotypes has a non-zero effect size, the HapSet method has a large power advantage over the marginal SNP method. When the set of causal haplotypes is masked, the HapSet method slightly outperforms the SNP method.

216 SNP method and the HapSet method while controlling for different FWER
217 levels.

218 When controlling for a 0.05 FWER, the two methods found 4,495 overlap-
219 ping eGenes (see Table 3). The HapSet method also found 707 eGenes that
220 were not detectable using genotype data. Since the SNP method is a subset
221 of the HapSet method, yet has a less stringent significance threshold, this in-
222 dicates that marginal SNPs poorly tag the regulatory architecture for these
223 genes. Conversely, the 606 eGenes found only by the SNP method represent
224 eGenes that had p-values smaller than 5.8×10^{-5} (the SNP method threshold
225 for a 0.05 FWER) but larger than 2.0×10^{-5} (the HapSet threshold for a
226 0.05 FWER). In total, there are 101 more eGenes identified by the HapSet
227 method than by the marginal SNP method. Using a more stringent 0.01
228 FWER shows that the marginal SNP test identifies 20 more eGenes than
229 the HapSet method. However, by allowing for a looser FWER of 0.1, the
230 HapSet method again outperforms the marginal SNP method and identifies
231 112 more eGenes.

232 We evaluate the average squared effect sizes for the top SNP or HapSet
233 for each eGene when controlling for a 0.05 FWER. The 4,495 eGenes found
234 by both methods had mean squared effect sizes of 0.097 and 0.099 for the
235 SNP and HapSet method, respectively. The 606 eGenes only found by the
236 SNP method had effects sizes of 0.045 (SNP) and 0.045 (HapSet). The SNP
237 method is a subset of the HapSet method but has a less stringent significance
238 threshold, which explains why these eGenes were not found by the HapSet
239 method but have the same mean squared effect size. The mean squared
240 effect sizes for the eGenes found only by the HapSet method have the largest
241 difference of 0.034 (SNP) and 0.055 (HapSet). This difference may indicate
242 that there are underlying causal architectures that the HapSet method is
243 capable of effectively tagging but the marginal SNP method cannot identify.
244 We regress out the effect of the top associated HapSet for the 707 eGenes
245 found only by the HapSet method and compute the mean squared effect of
246 the same top (though not significant) SNP associations with the residuals.
247 We observe a mean squared effect of 0.013. This indicates that the top SNPs
248 were largely picking up on signal better captured by the HapSet method.

249 4. Discussion

250 In this work we present a new framework for finding regions of the genome
251 with complex regulatory architectures that are poorly captured by a marginal

	Only SNP	Both Methods	Only HapSet
0.01 FWER	348	3,408	328
0.05 FWER	606	4,495	707
0.10 FWER	874	5,157	986

Table 3: Number of eGenes identified by the marginal SNP and HapSet methods while controlling for a given FWER. The marginal SNP method identifies 20 more eGenes than the HapSet method when using a 0.01 FWER, but with the 0.05 and 0.1 FWERs, the HapSet method identifies 101 and 112 more eGenes respectively.

252 SNP-based regression method. We assume that the two haplotypes an indi-
253 vidual carries have additive non-interacting effects, but make no assumptions
254 about the form of the regulatory architecture on each haplotype. We imple-
255 ment our method and demonstrate that it is well calibrated under the null
256 hypothesis. We also show that it has little loss of power due to an increase
257 in the multiple testing burden for common SNP architectures, but it has
258 increased power for SNPs with < 0.05 minor allele frequency and haplotype-
259 based architectures.

260 When we apply our method to real gene expression data, we see increased
261 power to detect eGenes as compared to the standard marginal SNP method.
262 Most importantly, we observe a large number of eGenes that are detected
263 by the HapSet method yet undetected by the SNP method. Since the SNP
264 method is a subset of the HapSet method, but with a less stringent signifi-
265 cance threshold, this indicates that in many genes there are complex regula-
266 tory effects that are poorly captured by marginal SNP association methods.

267 Our method gives no indication of the specific causal architectures un-
268 derlying each of the eGenes found with the HapSet method. Further analy-
269 ses must be performed in order to identify if there are multiple independent
270 causal SNPs, a SNPxSNP interaction or a more complicated haplotype based
271 effect. It is possible that our method may be identifying rare SNPs or un-
272 typed genetic variation. Confounding from such variables has been observed
273 in tests associating SNPxSNP interactions with gene expression [22, 23, 28].

274 Future work can implement a version of our method for identification
275 of associated loci in complex traits. Many complex traits have GWAS loci
276 that overlap known eQTLs [29], and applying the HapSet method to complex
277 trait data is a natural extension. The second direction is to apply the HapSet
278 method to fine mapping. Modeling more than 2 or 3 causal SNPs in a region

for fine mapping can be computationally challenging [30, 31, 32, 33, 34, 35, 36]. However, knowing which combinations of haplotypes are associated, and to what degree, may improve efficient selection of SNPs and interactions between SNPs for testing.

Funding

Research reported in this publication was supported by the National Institutes of Health under Awards R01-HG009120 (BP) and 5R00-GM111744 (SS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

- [1] V. G. Cheung, L. K. Conlin, T. M. Weber, M. Arcaro, J. Kuang-Yu, M. Morley, R. S. Spielman, Natural variation in human gene expression assessed in lymphoblastoid cells, *Nature genetics* 33 (2003) 422.
- [2] F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y.-H. Zhou, et al., Heritability and genomics of gene expression in peripheral blood, *Nature genetics* 46 (2014) 430–437.
- [3] S. Yang, Y. Liu, N. Jiang, J. Chen, L. Leach, Z. Luo, M. Wang, Genome-wide eqtls and heritability for gene expression traits in unrelated individuals, *BMC genomics* 15 (2014) 13.
- [4] A. L. Price, A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong, K. Stefansson, Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals, *PLoS genetics* 7 (2011) e1001317.
- [5] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, J. T. Burdick, Mapping determinants of human gene expression by regional and genome-wide association, *Nature* 437 (2005) 1365.
- [6] J.-B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, J. K. Pritchard, High-resolution mapping of expression-qtls yields insight into human gene regulation, *PLoS genetics* 4 (2008) e1000214.

- 311 [7] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. ACt Hoen, J. Mon-
312 long, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G.
313 Ferreira, et al., Transcriptome and genome sequencing uncovers func-
314 tional variation in humans, *Nature* 501 (2013) 506.
- 315 [8] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad,
316 R. Hasz, G. Walters, F. Garcia, N. Young, et al., The genotype-tissue
317 expression (gtex) project, *Nature genetics* 45 (2013) 580–585.
- 318 [9] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx,
319 R. Jansen, E. J. De Geus, D. I. Boomsma, F. A. Wright, et al., Integra-
320 tive approaches for large-scale transcriptome-wide association studies,
321 *Nature genetics* 48 (2016) 245–252.
- 322 [10] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaafari, K. Aquino-
323 Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J.
324 Cox, et al., A gene-based association method for mapping traits using
325 reference transcriptome data, *Nature genetics* 47 (2015) 1091–1098.
- 326 [11] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, M. Lathrop, Mapping
327 complex disease traits with global gene expression, *Nature reviews.*
328 *Genetics* 10 (2009) 184.
- 329 [12] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W.
330 Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, et al., Inte-
331 gration of summary data from gwas and eqtl studies predicts complex
332 trait gene targets, *Nature genetics* 48 (2016) 481–487.
- 333 [13] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt,
334 E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, J. K. Pritchard,
335 Understanding mechanisms underlying human gene expression variation
336 with rna sequencing, *Nature* 464 (2010) 768.
- 337 [14] B. E. Stranger, A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Bea-
338 zley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, et al., Population
339 genomics of human gene expression, *Nature genetics* 39 (2007) 1217.
- 340 [15] C. D. Brown, L. M. Mangravite, B. E. Engelhardt, Integrative modeling
341 of eqtls and cis-regulatory elements suggests mechanisms underlying cell
342 type specificity of eqtls, *PLoS genetics* 9 (2013) e1003649.

- 343 [16] T. Zeller, P. Wild, S. Szymczak, M. Rotival, A. Schillert, R. Castagne,
344 S. Maouche, M. Germain, K. Lackner, H. Rossmann, et al., Genetics
345 and beyond—the transcriptome of human monocytes and disease suscep-
346 tibility, PloS one 5 (2010) e10693.
- 347 [17] H. J. Cordell, Detecting gene-gene interactions that underlie human
348 diseases, Nature reviews. Genetics 10 (2009) 392.
- 349 [18] S. Prabhu, I. Pe’er, Ultrafast genome-wide scan for snp–snp interactions
350 in common complex disease, Genome research 22 (2012) 2230–2240.
- 351 [19] J. P. Lewinger, J. L. Morrison, D. C. Thomas, C. E. Murcray, D. V.
352 Conti, D. Li, W. J. Gauderman, Efficient two-step testing of gene-gene
353 interactions in genome-wide association studies, Genetic epidemiology
354 37 (2013) 440–451.
- 355 [20] G. Hemani, K. Shakhbazov, H.-J. Westra, T. Esko, A. K. Henders, A. F.
356 McRae, J. Yang, G. Gibson, N. G. Martin, A. Metspalu, et al., Detection
357 and replication of epistasis influencing transcription in humans, Nature
358 508 (2014) 249.
- 359 [21] R. B. Brem, J. D. Storey, J. Whittle, L. Kruglyak, Genetic interactions
360 between polymorphisms that affect gene expression in yeast, Nature 436
361 (2005) 701.
- 362 [22] G. Hemani, K. Shakhbazov, H.-J. Westra, T. Esko, A. K. Henders, A. F.
363 McRae, J. Yang, G. Gibson, N. G. Martin, A. Metspalu, et al., Hemani
364 et al. reply., Nature 514 (2014) E5–6.
- 365 [23] A. R. Wood, M. A. Tuke, M. A. Nalls, D. G. Hernandez, S. Bandinelli,
366 A. B. Singleton, D. Melzer, L. Ferrucci, T. M. Frayling, M. N. Weedon,
367 et al., Another explanation for apparent epistasis/hemani et al. reply,
368 Nature 514 (2014) E3.
- 369 [24] A. A. Brown, A. Buil, A. Viñuela, T. Lappalainen, H.-F. Zheng, J. B.
370 Richards, K. S. Small, T. D. Spector, E. T. Dermitzakis, R. Durbin, Ge-
371 netic interactions affecting human gene expression identified by variance
372 association mapping, Elife 3 (2014) e01381.
- 373 [25] S. Garnier, V. Truong, J. Brocheton, T. Zeller, M. Rotival, P. S. Wild,
374 A. Ziegler, T. Munzel, L. Tired, S. Blankenberg, et al., Genome-wide

- haplotype analysis of cis expression quantitative trait loci in monocytes,
PLoS genetics 9 (2013) e1003240.
- [26] D. Ying, M. J. Li, P. C. Sham, M. Li, A powerful approach reveals
numerous expression quantitative trait haplotypes in multiple tissues,
Bioinformatics (2018).
- [27] B. Han, H. M. Kang, E. Eskin, Rapid and accurate multiple testing
correction and power estimation for millions of correlated markers, PLoS
genetics 5 (2009) e1000456.
- [28] A. E. Fish, J. A. Capra, W. S. Bush, Are interactions between cis-
regulatory variants evidence for biological epistasis or statistical arti-
facts?, The American Journal of Human Genetics 99 (2016) 817–830.
- [29] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, N. J.
Cox, Trait-associated snps are more likely to be eqtls: annotation to
enhance discovery from gwas, PLoS genetics 6 (2010) e1000888.
- [30] G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L.
Price, P. Kraft, B. Pasaniuc, Integrating functional data to prioritize
causal variants in statistical fine-mapping studies, PLoS genetics 10
(2014) e1004722.
- [31] S. L. Spain, J. C. Barrett, Strategies for fine-mapping complex traits,
Human molecular genetics 24 (2015) R111–R119.
- [32] J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, J. M.
Howson, A. Auton, S. Myers, A. Morris, et al., Bayesian refinement of
association signals for 14 loci in 3 common diseases, Nature genetics 44
(2012) 1294–1301.
- [33] L. L. Faye, M. J. Machiela, P. Kraft, S. B. Bull, L. Sun, Re-ranking se-
quencing variants in the post-gwas era for accurate causal variant iden-
tification, PLoS genetics 9 (2013) e1003609.
- [34] F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identify-
ing causal variants at loci with multiple signals of association, Genetics
198 (2014) 497–508.

- 405 [35] J. K. Pickrell, Joint analysis of functional genomic data and genome-
406 wide association studies of 18 human traits, *The American Journal of*
407 *Human Genetics* 94 (2014) 559–573.
- 408 [36] D. J. Gaffney, J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai,
409 G. E. Crawford, M. Stephens, Y. Gilad, J. K. Pritchard, Dissecting
410 the regulatory architecture of gene expression qtls, *Genome biology* 13
411 (2012) R7.