Title

Enhanced methods to detect haplotypic effects on gene expression

Journal

Bioinformatics, 33(15)

ISSN

1367-4803

Authors

Brown, Robert
Kichaev, Gleb
Mancuso, Nicholas
et al.

Publication Date

2017-08-01

DOI

10.1093/bioinformatics/btx142

Peer reviewed

Bioinformatics, 33(15), 2017, 2307–2313
doi: 10.1093/bioinformatics/btx142
Advance Access Publication Date: 22 March 2017
Original Paper

OXFORD

Gene expression

Enhanced methods to detect haplotypic effects on gene expression

Robert Brown,[1,*] Gleb Kichaev,[1] Nicholas Mancuso,[2] James Boocock[1] and Bogdan Pasaniuc[1,2,3,*]

[1]Bioinformatics IDP, University of California Los Angeles, Los Angeles, CA, USA, [2]Department of Pathology and Laboratory Medicine and [3]Department of Human Genetics, Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on May 19, 2016; revised on March 9, 2017; editorial decision on March 11, 2017; accepted on March 20, 2017

Abstract

Motivation: Expression quantitative trait loci (eQTLs), genetic variants associated with gene expression levels, are identified in eQTL mapping studies. Such studies typically test for an association between single nucleotide polymorphisms (SNPs) and expression under an additive model, which ignores interaction and haplotypic effects. Mismatches between the model tested and the underlying genetic architecture can lead to a loss of association power. Here we introduce a new haplotype-based test for eQTL studies that looks for haplotypic effects on expression levels. Our test is motivated by compound heterozygous architectures, a common disease model for recessive monogenic disorders, where two different alleles can have the same effect on a gene's function.

Results: When the underlying true causal architecture for a simulated gene is a compound heterozygote, our method is better able to capture the signal than the marginal SNP method. When the underlying model is a single SNP, there is no difference in the power of our method relative to the marginal SNP method. We apply our method to empirical gene expression data measured in 373 European individuals from the GEUVADIS study and find 29 more eGenes (genes with at least one association) than the standard marginal SNP method. Furthermore, in 974 of the 3529 total eGenes, our haplotype-based method results in a stronger association signal than the standard marginal SNP method. This demonstrates our method both increases power over the standard method and provides evidence of haplotypic architectures regulating gene expression.

Availability and Implementation: http://bogdan.bioinformatics.ucla.edu/software/

Contact: rob.brown@ucla.edu or pasaniuc@ucla.edu

1 Introduction

Expression quantitative trait loci (eQTLs) are genetic variants, typically single nucleotide polymorphisms (SNPs), associated with gene expression levels. eQTLs are found through association scans that test for an additive effect of SNPs on expression (Pickrell *et al.*, 2010; Stranger *et al.*, 2007). In addition to additive effects, effects from interacting SNPs can moderate gene expression (Cordell, 2009; Hemani *et al.*, 2014a; Lewinger *et al.*, 2013; Prabhu and Pe'er, 2012). Some types of cis-interactions can only be captured by phase-aware methods (Buil *et al.*, 2015; Dimas *et al.*, 2008). However, many estimated interaction effects are explained by untyped variants or confounders that cast doubt on the importance and prevalence of interactions in humans (Fish *et al.*, 2016; Hemani *et al.*, 2014b; Wood *et al.*, 2014). Despite this, marginal SNP tests and SNP interaction tests cannot represent the full range of possible genetic architectures that can influence gene expression.

Studies of monogenic disorders (i.e. diseases caused by damaging mutations in a single gene) have been particularly successful in determining the genetic mechanisms responsible for disease. Recessive monogenic disorders can have an underlying compound heterozygous architecture of causal mutations that are usually loss-of-function (LOF) (Gilissen *et al.*, 2011, 2012; Ng *et al.*, 2010). These architectures arise when a gene is heterozygous at two different positions for LOF variants on different haplotypes. LOF compound heterozygous architectures are known to be important in complex traits as well (Lim *et al.*, 2013), but are challenging to detect due to multiple testing issues (Gibson, 2011). Intuitively, for fully penetrant recessive disorders, additional LOF mutations on the same haplotype have no additional effect since the gene function has already been disrupted. With widespread evidence of compound heterozygote architectures in monogenic disorders, in this work we extend such ideas to finding their effects on gene expression.

Transcriptional processes are controlled through multiple layers of genome organization (Grubert *et al.*, 2015; Kilpinen *et al.*, 2013; Koch, 2015; Waszak *et al.*, 2015). We hypothesize that specific sets of SNP alleles have cis-acting effects (Larson *et al.*, 2015) on transcriptional processes. Specifically, in this work we assume the effect of having one of the alleles on a haplotype is the same as having multiple. For example, having either of two alleles on a haplotype may have the same effect on an epigenetic state affecting expression from that haplotype as having both alleles on that same haplotype (ENCODE Project Consortium *et al.*, 2007; Ernst *et al.*, 2011; Guenther *et al.*, 2007; McVicker *et al.*, 2013; Taudt *et al.*, 2016). As an alternative example under this model, if alleles of two SNPs can each alone disrupt the function of an enhancer, then having both alleles on a haplotype will have the same effect as just having one of either allele on that same haplotype. To test this hypothesis, we define compound regulatory predictors (CRPs) that encode the number of haplotypes in each individual carrying at least one alternate allele from a predefined set of SNPs and test for association between the CRPs and gene expression levels. This does not preclude SNPs not in the set from having other independent effects for the same gene. We restrict our analysis to looking at CRPs composed from pairs of two SNPs.

Using simulations of multiple causal architectures, we demonstrate our method is better able to capture the signal from underlying CRP architectures leading to an increased number of eGenes discovered after controlling the false discovery rate (FDR). Importantly, the combined SNP and CRP method has no loss of power relative to the marginal SNP test to detect single causal SNPs.

To investigate the extent of CRPs in real data, we apply our method to data from the GEUVADIS eQTL study (Lappalainen *et al.*, 2013). We find that 2222 of the 3529 identified eGenes (genes with at least one association) contain both a SNP eQTL and a CRP eQTL. Of these genes, 822 have more of the expression variance captured with a CRP eQTL than a SNP eQTL. Of all eGenes, 974 (27.6%) have a CRP as the top association. There are 153 genes with a CRP eQTL but no SNP eQTL. Our combined SNP and CRP test finds 29 (0.8%) more eGenes than the marginal SNP test despite a larger multiple testing burden. Although this is only a small increase in overall power, the results as a whole demonstrate that some underlying genetic architectures affecting expression are better captured using a CRP model.

2 Materials and methods

We start with an overview of our proposed approach. We first regress gene expression on marginal SNP genotypes (the SNP test) in a 1 Mb window centered on a transcription start site. We then re-encode genotypes so that the alternate allele is positively associated with expression levels. This way alternate alleles forming CRPs will have the same effect direction on expression levels. We then encode CRPs as the number of haplotypes in an individual with at least one alternate allele at either of two SNP positions ($g_{CRP} \in \{0, 1, 2\}$). Last, we regress gene expression on g_{CRP}. To avoid a large increase in the number of tests, we limit multiple testing through a SNP pair selection process.

We illustrate the importance of the CRP model with a toy example in Figure 1. Alternate alleles, encoded as g_1 and g_2, can each affect a transcriptional process in such a way as to completely prevent gene expression from the haplotype(s) carrying the alternate allele(s). Since most eQTLs have small to modest effect sizes(Aguet *et al.*, 2016; GTEx Consortium, 2015; Lappalainen *et al.*, 2013), full loss of expression due to a SNP allele is an extreme example for illustrative purposes and not assumed by our model. Both the SNP test and the SNP × SNP interaction test (Cordell, 2009; Lewinger *et al.*, 2013) have reduced power since neither g_1, g_2 nor g_1g_2 are perfectly correlated with g_{CRP}. Since the alternate alleles have a cis-acting effect on the transcriptional process, gene expression is dependent both on the genotypes and the phase of the alleles in the special case of $(g_1,g_2) = (1,1)$.

2.1 The CRP model

A general additive two-SNP haplotype model in which each possible haplotype has an effect on the phenotype y is

$$y = \beta_{00}h_{00} + \beta_{10}h_{10} + \beta_{01}h_{01} + \beta_{11}h_{11} + \varepsilon \quad (1)$$

Here h indicates the number (0, 1 or 2) of each of the four possible haplotypes carried by an individual, β is the effect size of each haplotype, and $\varepsilon \sim N(0, \sigma_e^2)$. The subscripts specify the allele combinations for each haplotype. We focus on the model in which alternate alleles form a CRP ($\beta_{CRP} = \beta_{10} = \beta_{01} = \beta_{11} \neq 0$). We introduce a new variable ($g_{CRP} = h_{01} + h_{10} + h_{11}$) to indicate the number of haplotypes containing at least one alternate allele. We rewrite the model in terms of g_{CRP} as

$$y = \beta_{CRP}\, g_{CRP} + \varepsilon \quad (2)$$

Given genotype data g_i for SNP_i (or g_{CRP}) and phenotype data y for n individuals, a standard measure of association is the Wald statistic:

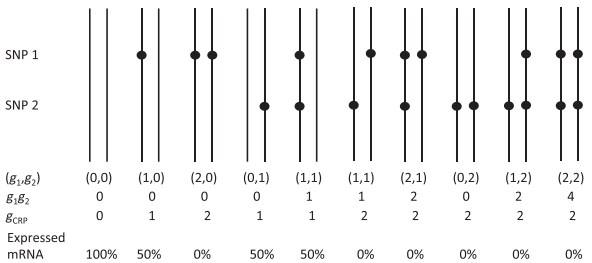

Fig. 1. Example of a causal CRP architecture. Each pair of vertical bars represents a maternal and paternal haplotype (unordered). A dot represents an alternate allele with g_1 and g_2 denoting the genotypes of the SNPs for an individual. The number of haplotypes carrying at least one alternate allele is given by g_{CRP}. The example phenotype, expressed mRNA, represents the percentage of the maximum amount of mRNA that can be produced and is linearly dependent on g_{CRP}. Full loss of expression due to the alleles is an extreme example for illustrative purposes. The term g_1g_2 represents the product of the two genotypes. The example shows two instances of $(g_1,g_2) = (1,1)$ where the phase will lead to different values for g_{CRP} and expression

$$z_i = \frac{\widehat{\beta}_i}{\mathrm{SE}\left(\widehat{\beta}_i\right)} = \frac{\mathrm{Cov}(g_i, y)\sqrt{n}}{\sqrt{\mathrm{Var}(g_i)\sigma_e^2}} \tag{3}$$

which asymptotically follows a normal distribution with variance 1 and a non-centrality parameter (NCP) given by

$$\lambda_i\sqrt{n} = \frac{\beta_i\sqrt{\mathrm{Var}(g_i)}}{\sigma_e}\sqrt{n}. \tag{4}$$

The NCP governs the power of rejecting the null hypothesis that there is no association between g_i and the phenotype at a specified family wise error rate (FWER). Non-causal SNPs ($\beta = 0$) have an induced-NCP if they are in linkage disequilibrium (LD) with a causal SNP (Hormozdiari *et al.*, 2014; Kichaev *et al.*, 2014; Kostem *et al.*, 2011; Pritchard and Przeworski, 2001; Wang *et al.*, 2015; Zaitlen *et al.*, 2010). Similarly, an induced-NCP can exist for SNPs comprising or tagging a causal CRP. We let x and y^* represent mean 0 and variance 1 transformed genotypes and phenotypes and β^* represent the β for the transformed data. We obtain an estimate for each β^* in a linear additive model.

$$\begin{bmatrix} \widehat{\beta}_i^* \\ \widehat{\beta}_j^* \\ \widehat{\beta}_{\mathrm{CRP}}^* \end{bmatrix} = \frac{1}{n}\begin{bmatrix} x_i^{\mathrm{T}} & x_j^{\mathrm{T}} & x_{\mathrm{CRP}}^{\mathrm{T}} \end{bmatrix} y^* \tag{5}$$

$$= \frac{1}{n}\begin{bmatrix} x_i^{\mathrm{T}} & x_j^{\mathrm{T}} & x_{\mathrm{CRP}}^{\mathrm{T}} \end{bmatrix}\left(\begin{bmatrix} x_i \\ x_j \\ x_{\mathrm{CRP}} \end{bmatrix}\beta^* + \varepsilon\right) \tag{6}$$

$$= \begin{bmatrix} 1 & r_{i,j} & r_{i,\mathrm{CRP}} \\ r_{i,j} & 1 & r_{j,\mathrm{CRP}} \\ r_{i,\mathrm{CRP}} & r_{j,\mathrm{CRP}} & 1 \end{bmatrix}\beta^* + \frac{1}{n}\begin{bmatrix} x_i^{\mathrm{T}} & x_j^{\mathrm{T}} & x_{\mathrm{CRP}}^{\mathrm{T}} \end{bmatrix}\varepsilon \tag{7}$$

$$= V\beta^* + \frac{1}{n}\begin{bmatrix} x_i^{\mathrm{T}} & x_j^{\mathrm{T}} & x_{\mathrm{CRP}}^{\mathrm{T}} \end{bmatrix}\varepsilon \tag{8}$$

We rewrite the $\widehat{\beta}^*$ estimates as random variables drawn from a multivariate normal distribution with means given by $V\beta^*$ and variance $\sigma_e^2 V n^{-1}$, in which V is the correlation matrix of the standardized genotypes.

$$\widehat{\beta}^* \sim \mathrm{MVN}\left(V\beta^*, \frac{\sigma_e^2}{n}V\right) \tag{9}$$

For a causal CRP architecture in which $\beta^* = [\,0 \quad 0 \quad \beta^*{}_{\mathrm{CRP}}\,]^{\mathrm{T}}$, $\lambda = V\beta^*$ is the mean values of $\widehat{\beta}^*$,

$$\begin{bmatrix} \lambda_i \\ \lambda_j \\ \lambda_{\mathrm{CRP}} \end{bmatrix} = \begin{bmatrix} 1 & r_{i,j} & r_{i,\mathrm{CRP}} \\ r_{i,j} & 1 & r_{j,\mathrm{CRP}} \\ r_{i,\mathrm{CRP}} & r_{j,\mathrm{CRP}} & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ \beta^*{}_{\mathrm{CRP}} \end{bmatrix} = \begin{bmatrix} r_{i,\mathrm{CRP}}\lambda_{\mathrm{CRP}} \\ r_{j,\mathrm{CRP}}\lambda_{\mathrm{CRP}} \\ \lambda_{\mathrm{CRP}} \end{bmatrix} \tag{11}$$

Here a SNP i that comprises or tags the CRP will appear to have a mean effect size $\lambda_i = r_{i,\mathrm{CRP}}\,\lambda_{\mathrm{CRP}}$. The mean effect size λ gives the NCP when testing a SNP or CRP for association with a phenotype for a given sample size.

2.2 Correlation between SNPs and CRPs

Each 2-SNP haplotype is characterized by the presence or absence of an alternate allele at the first and second SNP position (h_1 and h_2). The variable h_{CRP} indicates if a haplotype carries either of the two alternate alleles. The allele frequencies (f_1 and f_2) and the linkage between the SNPs (D) govern the haplotype probability in a sample.

Table 1. Two-SNP haplotype characterization

Haplotype	h_1	h_2	h_{CRP}	Haplotype probability (p)
h_{00}	0	0	0	$(1-f_1)\,(1-f_2)+D$
h_{01}	0	1	1	$f_1\,(1-f_2)\,-D$
h_{10}	1	0	1	$(1-f_1)\,f_2-D$
h_{11}	1	1	1	f_1f_2+D

Each 2-SNP haplotype is characterized by the presence or absence of an alternate allele at the first and second SNP position (h_1 and h_2). The variable h_{CRP} indicates if a haplotype carries either of the two alternate alleles. The allele frequencies (f_1 and f_2) and the linkage between the SNPs (D) govern the haplotype probability in a sample.

We calculate the correlation $r_{i,\mathrm{CRP}}$ for SNPs from a hypothetical sample where the probability of each two-SNP haplotype is a function of the allele frequencies and linkage (D) (see Table 1). Here two SNPs define each haplotype, with h_1 and h_2 representing the presence of an alternate allele at the first and second SNP positions. The maximum linkage ($D_{\max}$) between SNPs is a function of their allele frequencies (f_1 and f_2) and puts an upper bound on their correlation (r) (Hill and Robertson, 1968).

$$D = r\sqrt{(1-f_1)(1-f_2)f_1f_2} \tag{12}$$

$$D_{max} = \begin{cases} \min\{f_1f_2\,, \quad (1-f_1)(1-f_2)\} & \text{when } D < 0 \\ \min\{f_1(1-f_2),\ (1-f_1)f_2\} & \text{when } D > 0 \end{cases} \tag{13}$$

Assuming that haplotypes are inherited independently, there are 16 possible maternal and paternal two-SNP haplotype combinations for each individual. The haplotype probability (p) is the probability of drawing a specific haplotype with replacement. For each pair of haplotypes (indexed with superscripts k and l) we can compute the probability of the haplotype pair as p^kp^l. The equations $g_i^{k,l} = h_i^k + h_i^l$ and $g_j^{k,l} = h_j^k + h_j^l$ give the genotypes of an individual at SNPs i and j who has one kth and one lth haplotype. The $g_{\mathrm{CRP}}^{k,l}$ term, given by $g_{\mathrm{CRP}}^{k,l} = h_{\mathrm{CRP}}^k + h_{\mathrm{CRP}}^l$ is the number of haplotypes in an individual with one kth and one lth haplotype that contain either alternate allele. From these values, the correlation between g_i and g_{CRP} is computed using the following relationships:

$$\begin{aligned} r_{i,\mathrm{CRP}} &= \frac{\sum_{k=1}^{4}\sum_{l=1}^{4} p^kp^l\left(g_{\mathrm{CRP}}^{k,l} - \mu_{\mathrm{CRP}}\right)\left(g_i^{k,l} - \mu_i\right)}{\sigma_{CRP}\sigma_i} \\ \sigma_i^2 &= \sum_{k=1}^{4}\sum_{l=1}^{4} p^kp^l\left(g_i^{k,l} - \mu_i\right)^2 \ \text{ where } \mu_i = \sum_{k=1}^{4}\sum_{l=1}^{4} p^kp^l g_i^{k,l} \\ \sigma_{\mathrm{CRP}}^2 &= \sum_{k=1}^{4}\sum_{l=1}^{4} p^kp^l\left(g_{\mathrm{CRP}}^{k,l} - \mu_{CRP}\right)^2 \ \text{ where } \mu_{\mathrm{CRP}} = \sum_{k=1}^{4}\sum_{l=1}^{4} p^kp^l g_{\mathrm{CRP}}^{k,l} \end{aligned} \tag{14}$$

2.3 Power analysis to detect CRP effects

Using the model given in Equation (2) with the phenotype standardized to have mean 0 and variance 1 ($\sigma_Y^2 = 1$) we compute the power to reject the null hypothesis with a 0.05 significance threshold for a given sample and effect size. Let f_{CRP} be the frequency of risk haplotypes ($f_{\mathrm{CRP}} = E[g_{\mathrm{CRP}}]/2$).

$$\sigma_Y^2 = 2f_{\mathrm{CRP}}(1 - f_{\mathrm{CRP}})\beta_{\mathrm{CRP}}^2 + \sigma_e^2 \tag{15}$$

Let $\sigma_{\mathrm{CRP}}^2 = 2f_{\mathrm{CRP}}(1 - f_{\mathrm{CRP}})\beta_{\mathrm{CRP}}^2$ such that $\sigma_Y^2 = \sigma_{\mathrm{CRP}}^2 + \sigma_e^2 = 1$ where σ_{CRP}^2 is the variance of the phenotype explained by the CRP. We can then estimate the variance of $\widehat{\beta}_{\mathrm{CRP}}$ and approximate the NCP for the Wald statistic.

$$\mathrm{Var}\left(\widehat{\beta}_{\mathrm{CRP}}\right) = \frac{\sigma_e^2}{n\mathrm{Var}(g_{\mathrm{CRP}})} \approx \frac{\sigma_e^2}{2nf_{\mathrm{CRP}}(1-f_{\mathrm{CRP}})} \quad (16)$$

$$\lambda_{\mathrm{CRP}}\sqrt{n} \approx \frac{\beta_{\mathrm{CRP}}}{\sqrt{\mathrm{Var}\left(\widehat{\beta}_{\mathrm{CRP}}\right)}} = \sqrt{n\frac{\sigma^2_{\mathrm{CRP}}}{1-\sigma^2_{\mathrm{CRP}}}} \quad (17)$$

This result is identical to that of testing a single SNP for association with a phenotype, but uses g_{CRP} as the predictor as opposed to a SNP genotype. Assuming the true causal architecture is a CRP, SNPs will have induced-NCPs given by Equation (11). We calculate the power of a test to have a significant association given that the true causal architecture is a CRP:

$$\text{Power} = \Phi(\Phi^{-1}(\alpha/2) + \lambda\sqrt{n}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda\sqrt{n}) \quad (18)$$

Here λ can be either the NCP of the CRP (λ_{CRP}) or the induced-NCP at SNP$_i$ ($r_{i,\mathrm{CRP}}\ \lambda_{\mathrm{CRP}}$). $\alpha = 0.05/M$ is the desired FWER, where M is the number of tests performed for each gene.

2.4 CRPs in gene expression data

In order to search for SNP and CRP eQTLs in both real and simulated gene expression, we begin with the SNP test that regresses expression levels on centered and standardized SNP genotypes in a 1 Mb window around the gene's transcription start site. Following the GEUVADIS analysis, we included the top three genotype-based principal components as covariates as well as a binary variable denoting whether individuals were originally obtained from the 1000 Genomes (1000 Genomes Project Consortium *et al.*, 2012) Phase 1 or imputed. We only use SNPs with estimated maf >0.05. We re-encode the genotype data so that alternate alleles are positively associated with expression levels.

We limit the number of tests by only performing the CRP test on selected SNP pairs. To select SNP pairs, we look at all possible pairs of SNPs in the window being tested; when both SNPs in the pair pass a suggestive 0.4 significance threshold (Bonferroni corrected based on the number of SNP tests performed) and when each SNP in the pair has $|r_{i,\mathrm{CRP}}| < 0.8$, we test the CRP formed by the SNP pair for association with expression. This process looks for CRPs primarily in genes that already have a significant or near significant marginal association, so it is not expected that this test will greatly increase power.

For real data analysis we determine an empirical P-value using an adaptive permutation procedure following the GTEx approach (GTEx Consortium, 2015). We perform at least 1000 permutations and at most 10 000 permutations. After the first 1000 permutations, an exit criteria is reached if 15 permutations have a stronger association than the observed association. Therefore, all P-values are estimated with at least 1000 permutations. For each gene, we permute the expression levels and then rerun the entire SNP test as well as the entire SNP and CRP test including the SNP selection. We then control for a 0.05 FDR across genes using the Benjamini-Hochberg procedure. In the real data analysis the largest significant P-value after FDR control was 0.0095, which indicates that all significant genes required more than 1000 permutations before reaching the exit criteria.

For simulated data, we permute each gene on chromosome 22 10 000 times and use the resulting null distribution of association statistics for each gene to determine the P-values for simulated genes.

2.5 Simulations for multiple causal architectures

We base our simulations on the chromosome 22 genotypes of Europeans ($n = 373$) from the GEUVADIS study(Lappalainen *et al.*, 2013). We ran Beagle 4.1(Browning and Browning, 2007, 2016) to impute and phase missing or unphased genotypes for both the simulations and for the real expression analysis. Simulations draw either 0, 1 or 2 SNPs to be causal from a 1 Mb window centered on a randomly drawn transcription start site. After simulating a phenotype (see below), we run the tests as explained in Section 2.4. We also run an interaction test where we use an f-test to compare the model containing just the top marginal association to the model the contains the top marginal association as well as the product of the SNP genotypes for the same SNPs that are being evaluated as a CRP. For each causal architecture, we simulate 200 sets of 18 000 genes and report the mean number of genes with a significant association after controlling for the FDR.

Phenotypes are simulated using an additive model so that the causal genetic architecture explains a fixed $\sigma_g^2 = 0.08$ proportion of the variance in expression. We simulate five underlying causal architectures using either common SNPs with maf > 0.05 or rare SNPs with $0.01 <$ maf < 0.05: (i) We randomly choose a single common or rare SNP to be causal. (ii) Two causal common SNPs are randomly chosen and each explains half of σ_g^2 after accounting for linkage disequilibrium. (iii) A causal CRP formed by two randomly chosen SNPs with either both common or both rare. We also simulate CRPs with two common SNPs but require that the SNPs are correlated either with $r^2 > 0.8$ or <0.2. The high LD simulation replicates conditions likely seen in a regulatory element where SNPs are often strongly linked. (iv) The genotypes of two randomly chosen common SNPs are multiplied to form a causal interaction effect. (v) A null model where the phenotype is simply a draw from a normal distribution. We run the simulations using either masked or unmasked causal SNPs to determine how the methods will perform with un-typed variation and confounders.

2.6 Real data analysis

We re-analyzed data from the GEVUDADIS project (Lappalainen *et al.*, 2013). Following the original work (Lappalainen *et al.*, 2013), we filter out non-autosomal genes and genes that did not have >0 quantification in >90% of individuals resulting in 18 621 genes. We standardize the RPKM and PEER normalized gene expression levels sampled from human lymphoblastoid cells after removing non-European data. Last, we run the tests as described in Section 2.4. We compute a centered and standardized g_{CRP} from the phased genotype data for SNP pairs selected for the CRP test.

To determine if the top CRP eQTL is confounded due to correlation with the top SNP eQTL, we perform conditional regression that removes the effect of the top SNP eQTL if it is has an empirical P-value < 0.05. We then re-run the CRP analysis. Significance of the CRP is determined using the permutation method described earlier.

3 Results

3.1 Underlying SNPs poorly tag CRPs

The test statistic (z_i) is drawn from a normal distribution with a mean given by the NCP or induced-NCP. Under certain frequency and linkage conditions, the induced-NCPs at SNPs that comprise a causal CRP can be significantly lower than the CRP's NCP (see Figs 2 and 3). For example, two SNP genotypes (g_1 and g_2) each with maf = 0.5 and under no LD ($r_{1,2} = 0$) each have a correlation ($r_{1,\mathrm{CRP}}$ and $r_{2,\mathrm{CRP}}$) of 0.58 with the CPR (g_{CRP}). In this case, if the

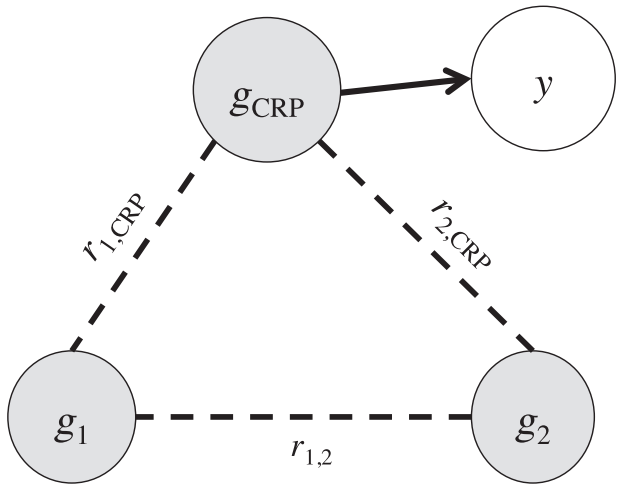

Fig. 2. Correlation structure between the SNP genotypes (g_1 and g_2) and the CRP (g_{CRP}). The phenotype y is dependent on the CRP

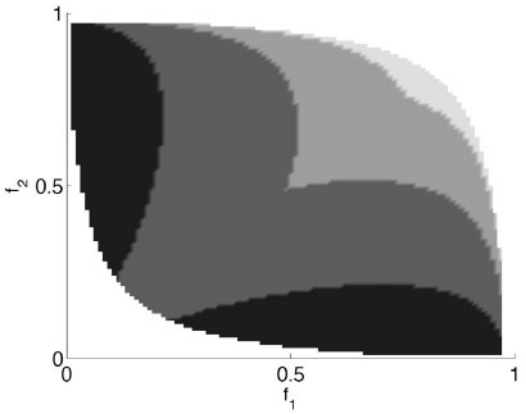

Fig. 3. The correlation between SNPs and CRPs. The greyscale represents the absolute maximum of $r_{1,CRP}$ and $r_{2,CRP}$ given the SNP frequencies indicated by the x- and y-axis and a correlation ($r_{1,2}$) of -0.2 between the SNPs. From darkest to lightest, the greyscale represents absolute maximum $r_{i,CRP}$ from (1,0.75), (0.75,0.5), (0.5,0.25) and (0.25,0)

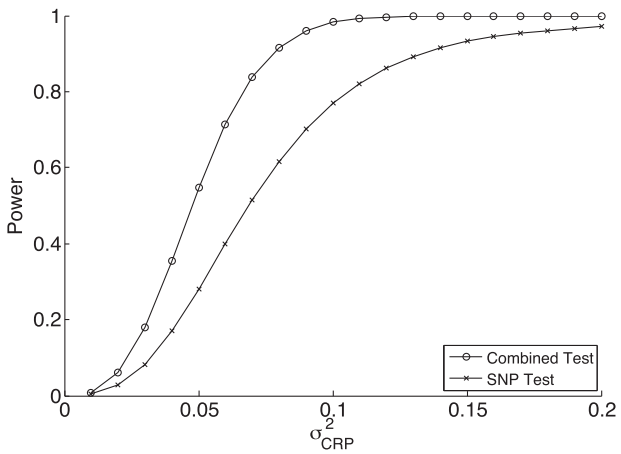

Fig. 4. Power to detect a causal CRP with 373 individuals and a 0.05 Bonferroni corrected significance level

CRP were causal, the SNP test's induced-NCP is only 58% the NCP of the CRP. This could result in a significant loss of power.

3.2 Power to detect CRPs

We computed the power at a 0.05 Bonferroni corrected significance level of the SNP test and the combined SNP and CRP test to have an association with a trait having an underlying CRP architecture. Because this does not take SNP selection into account for testing CRPs, the combined test results represent the power achievable if testing the causal CRP directly. We correct using the number of SNP (2266) or SNP and CRP (3064) tests performed on average per gene in the real gene expression data. As the variance explained due to a causal CRP decreases, the combined test outperforms the SNP test (see Fig. 4). The combined test has 92% maximum possible power to detect a CRP with $\sigma^2_{CPR} = 0.08$, assuming the CRP is directly tested, as opposed to 62% power with the SNP test.

We simulate gene expression under different causal architectures to evaluate the effect of confounders and to determine how the SNP selection process affects the power of the combined SNP and CRP test versus the marginal SNP test (see Table 2). In our simulations, we fix the percentage of phenotypic variance due to the underlying architecture at $\sigma^2_g = 0.08$ and simulate 200 sets of 18 000 genes under each architecture.

Using the null simulations, we observe that the SNP test, combined SNP and CRP test, and the combined SNP and interaction tests have mean false positive rates of 4.97, 4.96 and 4.97% respectively, each having a standard error of 0.01%. This indicates that the three tests are well calibrated under the null.

We evaluate the power of each test by comparing the average number of genes that have a significant association after controlling FDR at 0.05 in each 18 000 gene set using Benjamini-Hochberg (see

Table 2. Average number of eGenes identified after controlling the FDR for different underlying causal genetic architectures

Causal architecture	SNP test	SNP and interaction test		SNP and CRP test	
Null	0	0	*(0)*	0	*(0)*
		Unmasked			
SNP (c)	17 404	17 404	*(0)*	17 405	*(7)*
CRP (c)	14 421	14 422	*(1)*	14 514*	*(117)*
CRP (c) low LD	14 402	14 403	*(1)*	14 502*	*(124)*
CRP (c) high LD	16 135	16 136	*(0)*	16 135	*(9)*
2 SNPs (c)	14 529	14 529	*(0)*	14 640*	*(134)*
G_1G_2 (c)	10 477	10 485	*(9)*	10 538*	*(124)*
		Masked			
SNP (c)	16 395	16 395	*(0)*	16 400	*(18)*
SNP (r)	3298	3303†	*(6)*	3278	*(150)*
CRP (c)	13 565	13 565	*(0)*	13 670*	*(132)*
CRP (r)	2863	2864	*(2)*	2909*	*(211)*
2 SNPs (c)	13 701	13 701	*(0)*	13 824*	*(153)*

This table reports the mean number of simulated genes with at least one significant association for a given test and simulated causal architecture after controlling the FDR at 0.05. The * represents a significant difference in the number of eGenes discovered between the SNP and CRP test and both other tests and the † represents a significant difference between the SNP and interaction test and the SNP and CRP test (using a *t*-test with a significance threshold of 0.05/22). The SNP and interaction test was never significantly different form the SNP test. The italicized values in parentheses represent the number of genes found by the specified combined test but not included in the set of eGenes found by the SNP test. The (c) and (r) represent architectures using common or rare SNPs.

Table 2). No genes from the Null simulations are significant after controlling the FDR. When the underlying causal model is a single causal SNP, the three tests find approximately the same number of eGenes: 17 404 (SNP test), 17 404 (SNP and interaction test) and 17 405 (SNP and CRP test). Even though the tests find only one difference in the total number of eGenes, the sets of eGenes found by each test are not subsets of the most powerful test. The SNP and CRP test on average finds seven unique eGenes not included in the set of eGenes found by the SNP test. Interestingly, the rare single SNP causal model is the only simulated architecture where the combined SNP and CRP model is outperformed by the other models, indicating that CRPs poorly tag SNPs with maf < 0.05.

For simulated CRP architectures, the combined SNP and CRP test significantly outperforms the SNP test (and the SNP and interaction test) by finding 93 (and 92) more eGenes. However it found 117 unique eGenes not found by the SNP test indicating that in those genes marginal SNPs were much poorer tags of the underlying CRP. The most extreme example of this is found when looking at CRPs formed by two rare SNPs where the SNP and CRP test finds 211 eGenes not found by the SNP test even though it only finds 46 total more eGenes.

The combined SNP and CRP test has increased power over the SNP test when SNPs forming a CRP are in low LD. In this case the combined test finds 100 more eGenes than the SNP test, 124 being unique to the SNP and CRP test. Conversely, for the high LD CRP simulations, the combined test finds the same number of eGenes with 9 being unique. There is no increase in the total number of eGenes discovered since CRPs formed by high LD SNPs are very well tagged by single SNPs (see Figs 2 and 3). This also explains why the number of unique eGenes found by the SNP and CRP test is similar to what was found with the single causal SNP architecture.

For the two causal SNP and interaction architectures, the combined SNP and CRP test significantly outperforms the SNP and combined SNP and interaction tests. This is likely due to the CRP test being able to tag combinations of haplotypes poorly tagged by single SNPs and the fact that the SNP selection method used for both the CRP and the interaction tests is optimized for finding CRPs.

3.3 CRPs in real gene expression data

Looking at all 18 621 genes that passed the filtering criteria, our combined SNP and CRP test identifies 3529 eGenes while the marginal SNP test only finds 3500. Of the 3529 eGenes, 1154 have a SNP eQTL but no CRP eQTL and 2222 have both a SNP and a CRP eQTL. In 37.0% of the 2222, the CRP eQTL has a larger effect size than the SNP eQTL. For these eGenes, the top CRP eQTL from the combined test on average captures 12.6% of the variance in expression, as opposed to 10.8% with the top SNP eQTL. Finally, 153 identified eGenes have a CRP eQTL but no SNP eQTL. In these eGenes the top CRP eQTL captures 7.1% of the expression variance while the top (not significant) SNP captures 4.9%. These results demonstrate that the combined test is both more powerful than the marginal SNP test and in many eGenes better captures the signal from the genetic effect on expression.

The CRP model makes two predictions. The first is that the mean expression levels of individuals who are heterozygous at the two SNPs $(g_1,g_2) = (1,1)$ that form a CRP will depend on the phase of the alleles. The individuals will have $g_{CRP} = 1$ if the alleles are in phase or $g_{CRP} = 2$ if they are out of phase (see Fig. 1). The second prediction is that there should be no difference in mean expression levels between individuals with $(g_1,g_2) = (2,0)$ and the individuals with $(g_1,g_2) = (0,2)$. Both of these groups of individuals will have $g_{CRP} = 2$.

There are 887 eGenes with a CRP eQTL where at least four individuals fall into each of the groups. Using a Hochberg-Benjamini FDR control ($\alpha = 0.05$ applied to a t-test, we find 11 CRPs where there is a significant difference in mean expression levels between individuals with $(g_1,g_2) = (1,1)$ and $(g_{CRP} = 1)$ and individuals with $(g_1,g_2) = (1,1)$ and $(g_{CRP} = 2)$ but found no significant difference between individuals with $(g_1,g_2) = (0,2)$ and those with $(g_1,g_2) = (2,0)$.

In order to determine if the top CRP eQTL tags the top SNP eQTL, we condition gene expression on the top SNP eQTL with an empirical P-value < 0.05 and then re-run the CRP analysis and permutations. This results in 3218 eGenes where there is only a SNP eQTL, 158 eGenes that contain both a SNP and a CRP eQTL and 38 eGenes that contain only a CRP eQTL. This analysis shows that while the top CRP eQTLs are highly correlated with the top SNP eQTLs for most genes, in some cases the CRP eQTLs are capturing signal not included with the top SNP eQTL.

After running the SNP test and the combined SNP and CRP test, there are three SNPs of interest: g_m is the SNP that has the strongest marginal association with gene expression, $g_{CRP,1}$ and $g_{CRP,2}$ are the two SNPs that form the top CRP (g_{CRP}). We use Akaike information criterion (AIC) to compare eight models that use the following predictors: (1) g_m, with $k = 3$ (2) $g_{CRP,1}$ with $k = 3$, (3) $g_{CRP,2}$ with $k = 3$, (4) g_{CRP} with $k = 3$, (5) $g_{CRP,1}{}^*g_{CRP,2}$ with $k = 3$, (6) g_m and $g_{CRP,1}{}^*g_{CRP,2}$ with $k = 4$ (7) g_m and g_{CRP} with $k = 4$ (8) $g_{CRP,1}$ and $g_{CRP,2}$ and $g_{CRP,1}{}^*g_{CRP,2.}$ with $k = 5$.

Using AIC we determine if a CRP (g_{CRP}) effect is more likely than a SNP interaction ($g_{CRP,1}{}^*g_{CRP,2}$) by comparing models (6) and (7) that each include the main marginal effect as well (g_m). For the 2222 eGenes with both a SNP eQTL and a CRP eQTL, the model with the CRP is 100 times more likely than the model with the interaction effect for 263 of the eGenes. When looking at the 153 eGenes that only have a CPR eQTL, the model with the CRP is 100 times more likely than the model with the interaction effect in 21 of the eGenes.

We then compare the model that only includes the CRP effect (model 4) to all other models. For the 2222 eGenes, the CRP only model is most likely compared with all other models in 154 of the eGenes. When looking at the 153 eGenes that only have a CPR eQTL, the CRP only model is most likely in 31 of the eGenes.

4 Discussion

In this work we introduce a new method to detect haplotype effects on gene expression. Motivated by monogenic disorders, we extend ideas behind compound heterozygotes to gene expression through a CRP. Our method performs almost identically to the standard marginal SNP methods when the underlying architecture is a single causal, but outperforms it when there are more complex underlying architectures.

The main limitation of our combined test is that it only allows for CRPs composed of two SNPs. It is possible that any number of SNPs affects a transcriptional process or tag haplotypes with similar effect size. Due to the conservative SNP selection process, our method is best able to find CRP associations in genes that already have significant or close to significant associations. It is underpowered to find CRPs when the CRPs are poorly tagged by all marginal SNPs.

Though the gain in power of the combined test over the marginal SNP test is small in real data, the eGenes identified using the CRP model suggest that marginal tests of common SNPs do not fully tag the genetic architectures that influence gene expression. Without comprehensive functional analysis, it is impossible to know if a CRP eQTL causally changes expression levels, or if it is simply tags an

un-typed causal variant, interaction or a more complex causal mechanism.

Given the stronger CRP eQTL signals seen in many genes, our model may be useful for imputing gene expression that can then be leveraged in transcript-wide association studies (Gamazon *et al.*, 2015; Gusev *et al.*, 2016) Finally, the CRP eQTLs motivate two future directions. First, the method can be adapted to increase the power of genome-wide association studies to find novel associated loci. Second, fine-mapping methods used to prioritize potentially causal variants may become more accurate by explicitly modeling CRP architectures.

Acknowledgements

We would like to acknowledge Valerie Arboleda, Kathryn Burch, Huwenbo Shi, Kikuye Koyanu, Malika Kumar and Megan Roytman for assisting in the article preparation.

Funding

Research reported in this publication was supported by the National Institutes of Health under Awards [R01-HG009120 to B.P.], [T32-HG002536 to R.B.], [R01-GM053275 to B.P.] and [T32-CA201160 to G.K.]. 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

1000 Genomes Project Consortium. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Aguet,F. *et al.* (2016) Local genetic effects on gene expression across 44 human tissues.

Browning,B.L. and Browning,S.R. (2016) Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.*, **98**, 116–126.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Buil,A. *et al.* (2015) Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Gene.*, **47**, 88–91.

Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Dimas,A.S. *et al.* (2008) Modifier effects between regulatory and protein-coding variation. *PLoS Genet.*, **4**, e1000244.

ENCODE Project Consortium. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Fish,A.E. *et al.* (2016) Are interactions between cis-regulatory variants evidence for biological epistasis or statistical artifacts? *Am. J. Hum. Genet.*, **99**, 817–830.

Gamazon,E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.

Gibson,G. (2011) Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145.

Gilissen,C. *et al.* (2012) Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.*, **20**, 490–497.

Gilissen,C. *et al.* (2011) Unlocking Mendelian disease using exome sequencing. *Genome Biol.*, **12**, 228.

Grubert,F. *et al.* (2015) Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, **162**, 1051–1065.

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

Guenther,M.G. *et al.* (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.

Gusev,A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.

Hemani,G. *et al.* (2014a) Detection and replication of epistasis influencing transcription in humans. *Nature*, **508**, 249–253.

Hemani,G. *et al.* (2014b) Hemani et al. reply. *Nature*, **514**, E5–E6.

Hill,W.G., and Robertson,A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.

Hormozdiari,F. *et al.* (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.

Kichaev,G. *et al.* (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.*, **10**, e1004722.

Kilpinen,H. *et al.* (2013) Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science*, **342**, 744–747.

Koch,L. (2015) Genomics: Adding another dimension to gene regulation. *Nat. Rev. Genet.*, **16**, 563–563.

Kostem,E. *et al.* (2011) Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. *Genetics.*, **188**, 449–460.

Lappalainen,T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.

Larson,N.B. *et al.* (2015) Comprehensively evaluating cis-regulatory variation in the human prostate transcriptome by using gene-level allele-specific expression. *Am. J. Hum. Genet.*, **96**, 869–882.

Lewinger,J.P. *et al.* (2013) Efficient two-step testing of gene-gene interactions in genome-wide association studies. *Genet. Epidemiol.*, **37**, 440–451.

Lim,E.T. *et al.* (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron*, **77**, 235–242.

McVicker,G. *et al.* (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747–749.

Ng,S.B. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.

Pickrell,J.K. *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Prabhu,S., and Pe'er,I. (2012) Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res.*, **22**, 2230–2240.

Pritchard,J.K., and Przeworski,M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.

Stranger,B.E. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

Taudt,A. *et al.* (2016) Genetic sources of population epigenomic variation. *Nat. Rev. Genet.*, **17**, 319–332.

Wang,Z. *et al.* (2015) Gene-Gene Interactions Detection Using a Two-stage Model. *J. Comput. Biol.*, **22**, 563–576.

Waszak,S.M. *et al.* (2015) Population variation and genetic control of modular chromatin architecture in humans. *Cell*, **162**, 1039–1050.

Wood,A.R. *et al.* (2014) Another explanation for apparent epistasis. *Nature*, **514**, E3–E5.

Zaitlen,N. *et al.* (2010) Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.*, **86**, 23–33.