Acharya *et al. BMC Bioinformatics* (2017) 18:455
DOI 10.1186/s12859-017-1856-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Mapping eQTL by leveraging multiple tissues and DNA methylation

Chaitanya R. Acharya[1,2], Kouros Owzar[2] and Andrew S. Allen[1,2*]

Abstract

Background: DNA methylation is an important tissue-specific epigenetic event that influences transcriptional regulation of gene expression. Differentially methylated CpG sites may act as mediators between genetic variation and gene expression, and this relationship can be exploited while mapping multi-tissue expression quantitative trait loci (eQTL). Current multi-tissue eQTL mapping techniques are limited to only exploiting gene expression patterns across multiple tissues either in a joint tissue or tissue-by-tissue frameworks. We present a new statistical approach that enables us to model the effect of germ-line variation on tissue-specific gene expression in the presence of effects due to DNA methylation.

Results: Our method efficiently models genetic and epigenetic variation to identify genomic regions of interest containing combinations of mRNA transcripts, CpG sites, and SNPs by jointly testing for genotypic effect and higher order interaction effects between genotype, methylation and tissues. We demonstrate using Monte Carlo simulations that our approach, in the presence of both genetic and DNA methylation effects, gives an improved performance (in terms of statistical power) to detect eQTLs over the current eQTL mapping approaches. When applied to an array-based dataset from 150 neuropathologically normal adult human brains, our method identifies eQTLs that were undetected using standard tissue-by-tissue or joint tissue eQTL mapping techniques. As an example, our method identifies eQTLs by leveraging methylated CpG sites in a LIM homeobox member gene (LHX9), which may have a role in the neural development.

Conclusions: Our score test-based approach does not need parameter estimation under the alternative hypothesis. As a result, our model parameters are estimated only once for each mRNA - CpG pair. Our model specifically studies the effects of non-coding regions of DNA (in this case, CpG sites) on mapping eQTLs. However, we can easily model micro-RNAs instead of CpG sites to study the effects of post-transcriptional events in mapping eQTL. Our model's flexible framework also allows us to investigate other genomic events such as alternative gene splicing by extending our model to include gene isoform-specific data.

Keywords: eQTL, Multiple tissues, Tissue-specificity, DNA methylation, CpG islands, Gene expression, SNP, Score test, Monte Carlo simulations, Brain

Background

It has been long established that regulatory regions in higher eukaryotes activate gene transcription in a tissue-specific manner [1, 2]. These regulatory regions, which affect the binding affinities of transcription factors, are susceptible to both genetic variation and epigenetic modifications that play a coordinated role in regulating tissue-specific gene expression [3–7]. One form of epigenetic variation is DNA methylation that targets non-methylated and noncoding GC-rich and CpG-rich regions of the DNA sequence, which constitute approximately 70% of all annotated promoters [8]. DNA methylation is linked to transcriptional silencing, and many CpG island promoters are active in a tissue-specific manner. Previous studies have shown that inter-individual variation in DNA methylation at distinct CpG sites has been consistently linked to genetic variation such as single nucleotide polymorphisms (SNPs), known as methylation eQTLs

*Correspondence: asallen@duke.edu
[1]Program in Computational Biology and Bioinformatics, Duke University, 2424 Erwin Road, Suite 1104, 27710 Durham, NC, USA
[2]Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Suite 1104, 27710 Durham, NC, USA

(mQTLs) [9–11]. Since an increased DNA methylation at any of the distinct CpG sites located in the promoter regions necessitate chromatin remodeling and subsequent decrease in gene expression, any DNA sequence variation within the CpG-rich regions that disrupts the methylation process may have an opposite effect on gene expression.

Even though, mechanisms which regulate DNA methylation are unclear, it is clear that there is some association between genetic variation and quantitative changes in methylation levels [12]. For example, Catechol-O-methyltransferase (COMT) gene, which is implicated in schizophrenia has a SNP, $Val^{158}Met$ (rs4680) that is associated with differential COMT expression across regions of the brain during the course of the illness [13]. More specifically, the substitution of a methionine (Met) for a valine (Val) at position 158 results in reduced activity of the COMT enzyme due to reduced protein stability. Methylation of CpG islands associated with the aforementioned variant affect the region-specific expression of COMT [13]. Identifying and studying the mechanisms through which genetic variation, DNA methylation and gene expression interact may provide us yet another clue to understanding regions within the genome that are associated with complex disease phenotypes (Fig. 1a).

We have previously proposed a score test-based approach to map multi-tissue eQTLs where we model tissue-specificity as a random effect and investigated an overall shift in the gene expression combined with tissue-specific effects due to genetic variants [14]. Current approaches to delineate the role played by both genetic and epigenetic variation in gene expression are limited to identifying statistically significant pairs of mRNA - SNPs and CpG - SNPs by performing independent eQTL and mQTL analyses, respectively, within a tissue-by-tissue (TBT) framework [4, 11, 15]. These pairs are then expanded to combinations of mRNA transcript, CpG site and a SNP wherever the SNP was significantly correlated with either mRNA or CpG site of the mRNA - CpG pair. First, any such TBT analyses have been shown to fall short in fully exploiting patterns across the tissues thus impacting eQTL or mQTL discovery [14, 16, 17]. Second, independent eQTL and mQTL analyses do not reveal any underlying effects of genetic variation on tissue-specific gene expression due to DNA methylation. Consequently, we propose to map eQTLs by leveraging DNA methylation and testing for any higher order interactions among methylation, genotype and tissues. We extend this framework to include methylation-specific effects and model

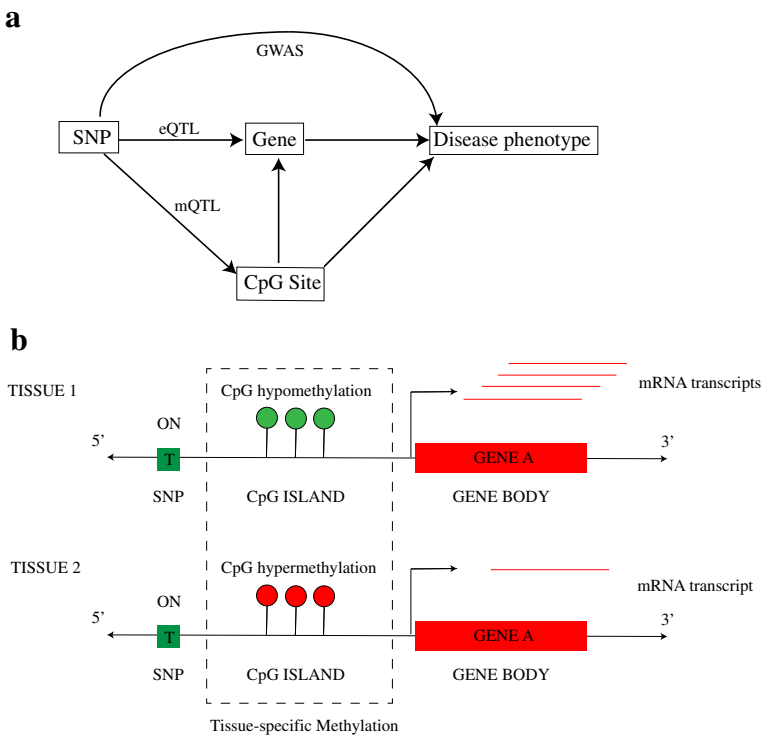

Fig. 1 Tissue-specific gene expression is controlled by genetic, epigenetic and transcriptional regulatory mechanisms. **a** Figure illustrating the idea that identifying and studying the mechanisms through which genetic variation, DNA methylation and gene expression interact may provide us with clues to understanding regions within the genome that are associated with complex disease phenotypes. **b** Figure illustrating the role played by tissue-specific methylation patterns and a genetic variant in regulating gene expression

the combined effect of genetic and epigenetic variation on gene expression (Fig. 1b).

Our main objective is to improve eQTL discovery by accounting for epigenetic effects such as DNA methylation. We show using Monte Carlo simulations that our joint score test is more powerful in mapping eQTLs by controlling for methylation than any TBT approach that uses methylation as a covariate (TBTm-eQTL). We also show that the new joint score test is better at identifying eQTLs in the presence of DNA methylation than our previously proposed multi-tissue eQTL and TBT methods. Finally, we show that in cases where the interaction effects of DNA methylation are absent, our approach is slightly less powerful but remains competitive. We demonstrate the effectiveness of our method by applying it to a publicly available expression, methylation and SNP array datasets from normal adult human brains [4] and show that by jointly analyzing multiple brain regions (or tissues), we identify eQTLs that may otherwise be not identified by multi-tissue eQTL methods.

Results and discussion

Evaluating our new score test using Monte Carlo simulations

We evaluate our approach through extensive simulation studies. Briefly, each Monte Carlo simulated dataset was comprised of data from a single locus and a single gene, whose expression is measured across 5 tissues in 500 observations. For a given mRNA - SNP pair, the genotypes at each SNP in all the individuals were simulated as Binomial(2,0.3), i.e. a minor allele frequency 0.30 and assuming Hardy-Weinberg equilibrium. Methylation data for all tissues was generated from a multivariate normal distribution with a positive definite variance-covariance matrix. Since all the tissue-specific effects are modeled as random effects, a test of whether there are any tissue-specific effects is equivalent to testing whether the variances of the random effects (γ and δ) are zero. Thus, our model involves testing four scalar parameters (β, ϕ, γ and δ). Simulations under the null hypothesis confirm that our method has the correct type 1 error (see Additional file 1). Since we model the effects of both epigenetic and genetic variation, we evaluated any power loss in identifying mRNA - SNP associations in the absence of any epigenetic effect. This was accomplished by comparing our method's performance with TBT-eQTL approach by keeping all the parameters associated with methylation in Eq. 5 at zero (i.e. $\lambda = \phi = \delta = \theta = 0$). We also compared our method with a previously proposed multi-tissue eQTL method, implemented in our software JAGUAR [18], which is made available at Comprehensive R Archive Network (CRAN) repository. Briefly, JAGUAR implements an approach that jointly models the overall shift in the gene expression due to genotype together with tissue-specific interaction with genotype in order to efficiently identify multi-tissue eQTL. From Fig. 2a, we see that JAGUAR outperforms both TBT-eQTL and our new joint score test. This loss of power, though not substantial, may be attributed to testing for an inexistent methylation effect. However, in the presence of a methylation effect our method outperforms both TBT-eQTL and JAGUAR as evidenced by Fig. 2b. When the number of tissues is increased from 5 to 10, the same pattern in statistical power was observed (see Additional file 1 section for figures).

We also compared our joint score test to a TBT-eQTL approach that included methylation as a baseline covariate [15], henceforth referred to as TBTm-eQTL analysis, using the following linear regression model –

$$Y = M\alpha + G\beta + GM\phi + \xi \quad (1)$$

where Y is a nt-dimensional matrix of expression levels in t tissues and n individuals, α is a fixed effect representing the tissue-specific intercepts, G is a nt-dimensional matrix of genotypes, β is a fixed effect of genotype across all tissues, M is an nt-dimensional matrix of methylation information and ϕ is genotype $\times$ methylation interaction effect (fixed effect). Minimum p value from the TBTm-eQTL analysis across all the tissues is computed for power calculations. Table 1 shows that our method significantly outperforms TBTm-eQTL approach showing a clear statistical advantage in using our joint score test over the TBTm-eQTL approach.

See Additional file 1 methods for more information on the description of various null hypotheses being tested.

Region-specific DNA methylation impacts eQTL mapping in adult human brains

In order to demonstrate the effectiveness of our method, we applied it to Gibbs et al. [4] dataset comprising of 150 individual data obtained from four regions of human brain. We performed data analyses that focused on only *cis* candidate regions i.e., the proximity of an eQTL to the transcription start site of a gene did not exceed 100 kilobase up- and down-stream of the transcription start site of a gene/mRNA transcript (*cis*-SNP). CpG islands that were less than 1.5 kilobase up- and down-stream of the transcription start site of the same mRNA transcript were paired with the mRNA transcripts. Each mRNA transcript was tested for an association with every *cis*-SNP in the presence of a (methylated or unmethylated) CpG site located in the promoter region.

Our joint score test method performed a total of 471,272 tests (totaling 11,076 mRNA transcripts, 14,244 CpG sites and 144,393 *cis*-SNPs). Each such mRNA - CpG pair is tested for an association with a *cis*-SNP. It is important to note that our method does not test any direct association between an mRNA transcript and its corresponding

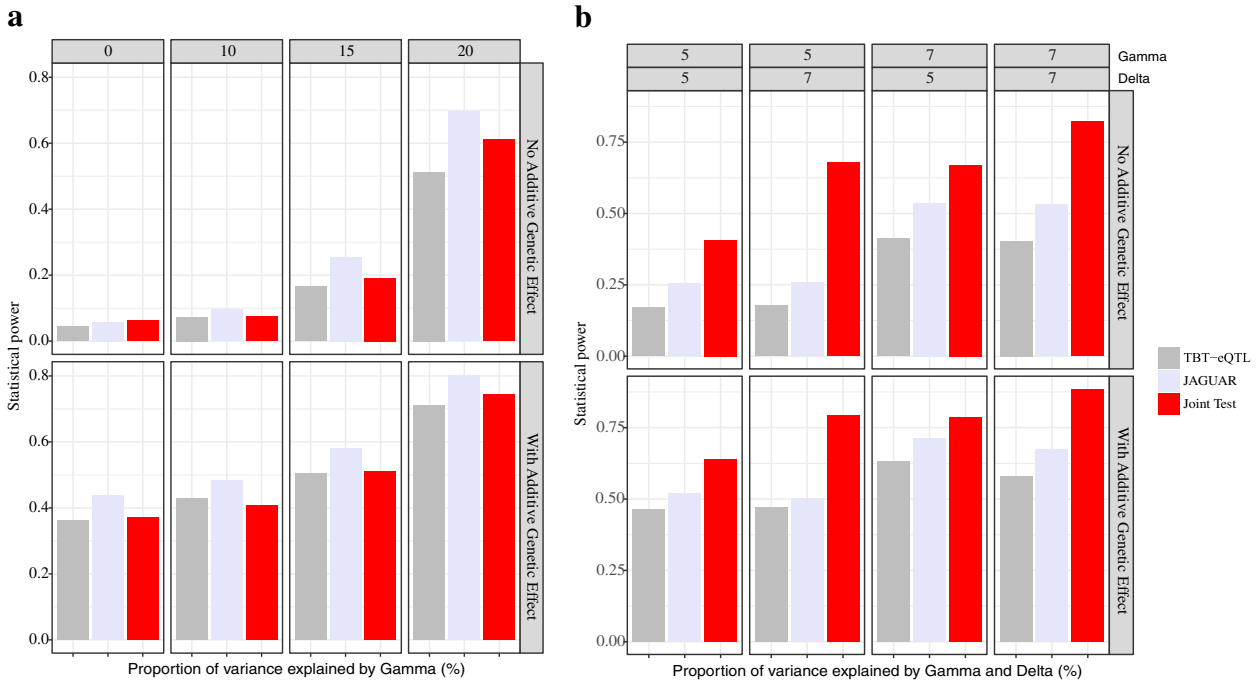

Fig. 2 eQTL identification using TBT-eQTL, JAGUAR and our method in the presence and absence of DNA methylation effects. **a** In the absence of methylation data, statistical power from the joint analysis of genotype and tissue-specific interaction using JAGUAR is marginally better than our joint score test. A tissue-by-tissue (TBT-eQTL) method is also used for comparison. The x-axis denotes the proportion of variance explained by the $G \times T$ effect and the y-axis denotes the statistical power. These data were generated from 1,000 simulations run on 500 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3). **b** In the presence of DNA methylation effect, our method out performs JAGUAR and tissue-by-tissue analyses. The top two rows in the figure indicate $PVE_{G\times T}$ and $PVE_{G\times M\times T}$, respectively, on the x-axis. Statistical power is denoted on the y-axis. These data were generated from 1,000 simulations run on 500 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3)

CpG site. Any resulting combinations of mRNA transcript, CpG site and a SNP would describe the relationship between the mRNA and SNP in the presence of the corresponding promoter CpG site, i.e. identify an eQTL. Our method identified a total of 5967 eQTLs that are statistically significant at 5% false discovery rate (FDR). In order to account for the number of traits being tested, the p values obtained from applying our joint score test were adjusted for multiple testing using an optimized FDR approach to obtain per-SNP q values (FDR adjusted p values) [19]. We observed that majority of these significant results are driven by a combination of additive genetic effect (93%) and $G \times T$ effect (81%) while the G × M and G × M × T effects were barely observed. This may be due to a lack of any distinct tissue-specificity in the methylation data, which we observed while preprocessing Gibbs et al. data (see "Methods" section). However, we expect that the aforementioned effects may be well pronounced across diverse tissue types such as the ones made available by the Genotype-Tissue Expression (GTEx) initiative [20].

We performed two region-by-region or TBT approaches on the same set of mRNA transcripts, CpG sites and SNPs as above, one with DNA methylation as a covariate (TBTm-eQTL) and the other with no methylation (TBT-eQTL) and compared the results with our approach. We estimated q values from each set of p values (originated from each region-by-region analysis) and minimum q value for a given mRNA - SNP pair across all the brain regions was computed, which indicates the presence of a statistically significant pair in at least one brain region. The number of significant associations in at least one brain region were then assessed at 5% FDR (q value $\leq \frac{0.05}{4}$ where 4 is the number of brain regions). TBT-eQTL approach identified a total of 5009 mRNA-*cis*-SNP pairs or *cis*-eQTLs significant in at least one region of the brain at 5% FDR. Roughly 79% of these TBT-eQTLs overlap with eQTLs identified using our method. On the other hand, TBTm-eQTL approach identified 5625 eQTLs with a 73% overlap with eQTLs identified using our method.

In order to assess the role of brain region-specificity on gene expression and the advantages in jointly modeling all the brain regions on mapping eQTLs, we compared our joint score test approach with a previously proposed multi-tissue eQTL mapping method [14] implemented by our software JAGUAR. JAGUAR identifies 7934 eQTLs (96% of them overlap with the TBT-eQTLs, 80% of them

Table 1 Table comparing the statistical power of our method and TBTm approach

Additive Genetic Effect	$G \times M$ Effect	$PVE_{G\times M\times T}$	$PVE_{G\times T}$	TBTm	Joint Score Test
NO	NO	0	0	0.041	0.045
NO	NO	7	0	0.139	0.141
NO	NO	10	0	0.415	0.433
NO	NO	0	7	0.097	0.172
NO	NO	7	7	0.234	0.332
NO	NO	10	7	0.472	0.552
NO	NO	0	10	0.218	0.433
NO	NO	7	10	0.341	0.546
NO	NO	10	10	0.547	0.721
NO	YES	0	0	0.351	0.171
NO	YES	7	0	0.511	0.337
NO	YES	10	0	0.719	0.598
NO	YES	0	7	0.388	0.363
NO	YES	7	7	0.565	0.501
NO	YES	10	7	0.708	0.679
NO	YES	0	10	0.525	0.605
NO	YES	7	10	0.653	0.694
NO	YES	10	10	0.782	0.816
YES	NO	0	0	0.155	0.244
YES	NO	7	0	0.296	0.371
YES	NO	10	0	0.543	0.601
YES	NO	0	7	0.229	0.357
YES	NO	7	7	0.389	0.513
YES	NO	10	7	0.57	0.702
YES	NO	0	10	0.425	0.606
YES	NO	7	10	0.522	0.692
YES	NO	10	10	0.708	0.819
YES	YES	0	0	0.487	0.423
YES	YES	7	0	0.627	0.572
YES	YES	10	0	0.753	0.708
YES	YES	0	7	0.536	0.563
YES	YES	7	7	0.69	0.689
YES	YES	10	7	0.78	0.801
YES	YES	0	10	0.648	0.719
YES	YES	7	10	0.761	0.807
YES	YES	10	10	0.821	0.856

This data were generated from 1,000 simulations run on 500 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3)

overlap with TBTm-eQTLs, and 94% of them overlap with the joint tests's eQTLs) at 5% FDR. All the eQTLs that overlap between JAGUAR and our new joint score test are mostly driven by the additive genetic effect and $G \times T$ effect and not higher order methylation interaction effects such as $G \times M$ and $G \times M \times T$. This absence of any pronounced region-specific DNA methylation effect explains the lower number of eQTLs identified by our joint test method. However, as we have shown using simulation data, in the presence of any region-specific interaction effects involving methylation, our joint score test is far more informative than the results from JAGUAR. Few of the eQTLs identified by our method were not detected by JAGUAR. This could be because a majority of these eQTLs were driven by $G \times M \times T$ interaction effect, which is not tested by JAGUAR. For example, let us consider a splice variant of LIM Homeobox protein coding gene (LHX9; Ensemble ID - ENSG00000143355), located on chromosome 1, which has 2 annotated *cis*-SNPs (SNP IDs: rs10922303 and rs2047541) possibly in LD with each other) and two promoter CpG sites (CpG IDs: cg07214572 and cg08008403) in our preprocessed datasets. Out of these 4 (number of mRNA - CpG pairs $\times$ the number of SNPs) combinations of mRNA transcript, CpG sites and SNPs and a possible 2 eQTLs, our method identified all of them to be statistically significant. None of them were found to be statistically significant by any TBT-based or the multi-tissue eQTL approaches (Fig. 3a). This is a good example of mapping eQTLs by leveraging effects due to DNA methylation since there is tissue-specific interaction effect clearly observed in Fig. 3b not captured by either JAGUAR or TBT methods. Of note, LHX9 is ubiquitously expressed in brain and are known to help in determining neuronal differentiation in humans [21]. On the other hand, we also see many instances of eQTLs that were observed to be statistically significant using JAGUAR but not our joint score test method due to the lack of any distinct tissue-specific DNA methylation effects. For example, JAGUAR method identified gene glutathione S-transferase mu 4 (GSTM4; Ensembl ID - ENSG00000168765), a gene that belongs to a superclass of glutathione S-transferases, which play a major role in the development of brain tumors [22], to have a statistically significant association with a promoter eQTL (SNP ID: rs524998), as illustrated by Fig. 4. However, we found that GSTM4 gene has two promoter CpG sites (CpG IDs: cg11903880 and cg15955341). Since there is no tissue-specific methylation effect, our joint test method was less powerful in detecting this eQTL . As seen in this figure, the lack of any tissue-specific methylation effects may have resulted in not being identified as a potential eQTL by our joint score test method.

To assess the biological relevance of the genes with eQTLs identified by TBT or multi-tissue methods including our new joint score test, we performed a KEGG pathway term enrichment analysis [23] for each set of results separately (see Additional file 1). KEGG pathways were

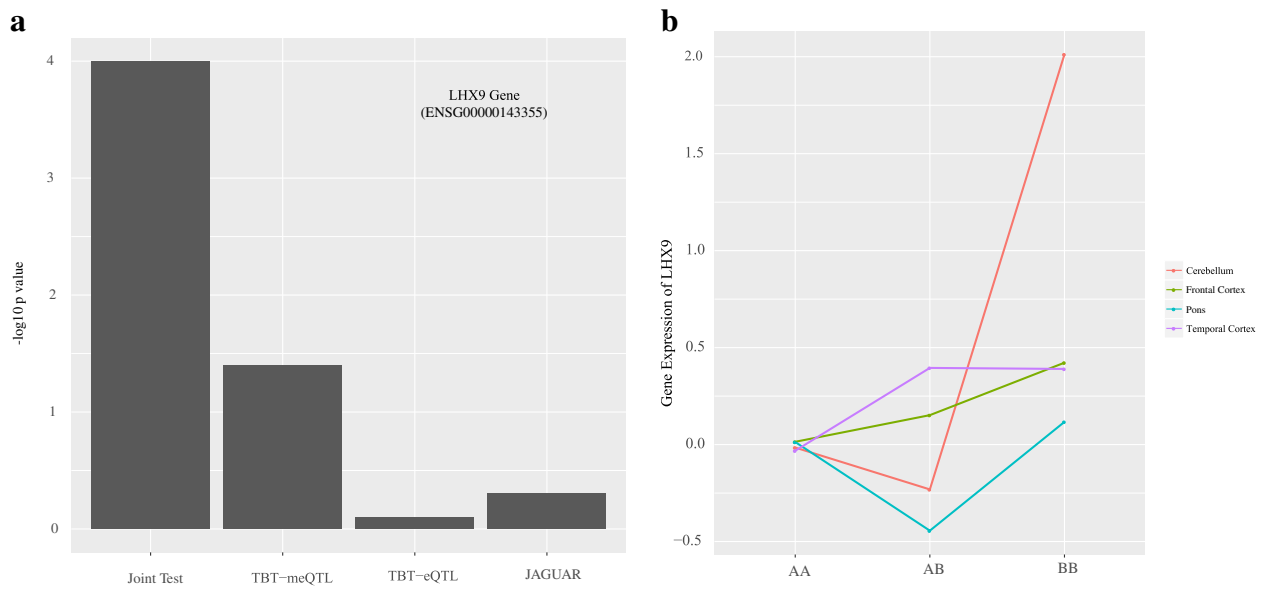

Fig. 3 An example of an eQTL for gene LHX9 identified as statistically significant by our joint score test method. **a** Barplot displaying all the statistics computed for LHX9 gene and SNP rs10922303 from our joint test, TBT-meQTL, TBT-eQTL and JAGUAR methods. The vertical axis represent -log10 *p* values. **b** Interaction plot illustrating tissue-specific genotypic effect on gene expression. Given that the lines are nonparallel, there is an interaction effect between tissue type and genotype

considered overrepresented if a set of at least three genes from different linked regions is observed to be overrepresented with an adjusted significance level of q value <0.05, calculated from a hypergeometric test. Our method identified 5 overrepresented pathways (Metabolic pathways, Ribosome, Fatty acid degradation, Purine and Pyramidine metabolism), JAGUAR identified 2 pathways while TBT-eQTL identified 1 overrepresented pathway. The overrepresented pathway, "Metabolic Pathways" (KEGGID: hsa01100) is the only common pathway between TBT-eQTL, JAGUAR and our method. On the basis of prior knowledge of function, the overrepresented pathways "Purine metabolism" (KEGGID: hsa00230) and "Pyramidine metabolism" (KEGGID: hsa00240) are plausible functional candidate pathways for schizophrenia [24]. These information can be used to guide genetic analyses by selecting these relevant pathways and genes associated with the pathways for schizophrenia.

Conclusion

Overall, our efforts are primarily directed to understanding two very specific aspects – 1) the overall effect of a

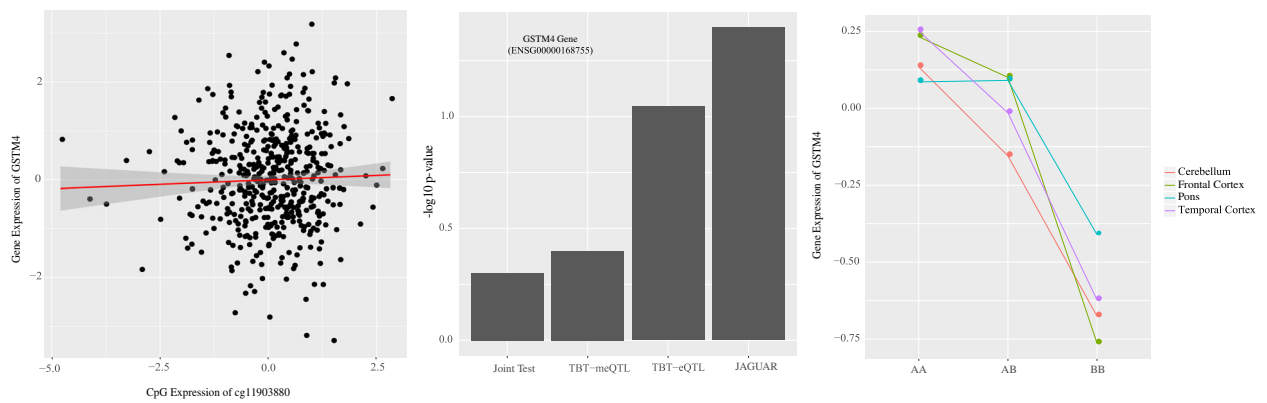

Fig. 4 An example of an eQTL for gene GSTM4 not identified as statistically significant by our joint score test method. Left panel displays a regression plot showing no association between DNA methylation of CpG site cg11903880 and gene expression of GSTM4. The middle panel shows all the statistics computed for GSTM4 gene and SNP rs524998. Right panel illustrates the interaction plot of tissue-specific eQTL

genetic variant on gene expression regulation by accounting for any changes in tissue DNA methylation levels, and 2) map eQTLs by leveraging tissue-specific methylation effects. Currently, there are no methods that jointly model the epigenetic and genetic control of tissue-specific gene expression. Many eQTL studies fail to account for the masking effect on a genetic variant due to DNA methylation, which may regulate gene expression across multiple tissues. Our method provides an efficient framework to integrate SNPs, DNA methylation and gene expression, and investigate how the different forms of variation interrelate.

The dataset examined here used genome-wide association (GWA) study SNP array platform to interrogate germline variation that includes an overwhelming number of common variants. Although GWA studies have been able to explain a small fraction of the genetic components of common human diseases, it is hypothesized that some of the missing heritability may be due to rare variation. Since standard common disease common variant approaches are severely underpowered to tease out any underlying variants that are moderate to extremely rare, there is an emphasis on large sample sizes and gene-based association tests in order to securely identify genetic risk factors that may otherwise be outside the range detectable by GWA studies [25]. One solution to the aforementioned issue would be to prioritize genetic variants in a non *ad-hoc* framework that preferentially weights genetic variants. Our method can provide a statistically disciplined weighting framework within which genetic variants can be either up- or down-weighted for any subsequent downstream analyses. Our method may also be useful in generating weights to any methods that use a reference data set in which both genome variation and gene expression levels have been measured to develop prediction models for gene expression [26].

The absence of strong tissue-specific methylation effects has an effect on mapping eQTLs using our joint test method. In the absence of any tissue-specific methylation effect, our method is less powerful while mapping eQTLs. One potential way to overcome such situations would be to run an omnibus test that identifies strongest evidence between JAGUAR and our joint test model. Specifically, we calculate the p value under each model, and then compute the minimum of the two p values and compare the observed minimum p value to its null distribution. Deriving the analytical null distribution of the minimum p value is not trivial considering the complex correlation structure between the statistics and due to the presence of higher order interaction effects (see Additional file 1 section). This approach is purely speculative and was not tested by us.

Since we are modeling the effects of non-coding regions (via CpG sites) on gene expression using our model, we can easily use micro-RNA (miRNA) data instead of CpG site methylation data and model post-transcriptional regulation of tissue-specific gene expression. miRNA expression, also a tissue-specific phenomenon, have been known to post-transcriptionally silence expression of mRNA transcripts. The presence of genetic variants such as SNPs may have an effect on the biogenesis and function of miRNA molecules leading to a downstream effect on gene expression [27]. This tissue-specific interaction between miRNA and SNP can be modeled in a similar fashion, analogous to modeling the interaction effects of tissue-specific DNA methylation and SNPs. The flexibility of our model also enables us to incorporate new information such as gene isoform data and accommodate the analysis of next-generation sequencing data (such as RNA-seq) by modeling gene transcripts in an analogous fashion to tissues in our current model formulation. This type of analysis would aggregate expression over all the splice variants of a gene across multiple tissues and inform us of tissue-specific alternative splice variant of a gene. These results become relevant to studying genetic effects on alternative splicing and its key role in important cellular networks.

Methods

Our model

For a given mRNA transcript, tissue-specific gene expression is modeled as a function of genotype and methylation –

$$Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi \quad (2)$$

where Y is an nt-dimensional vector of expression levels in t tissues and n individuals, α is a vector of tissue-specific intercepts, G is nt-dimensional vector of genotypes, β is a fixed effect of genotype across tissue, M is nt-dimensional vector of methylation levels, λ is an overall methylation-specific fixed effect, MG is nt-dimensional vector of the product of methylation and genotype, ϕ is the regression coefficient for genotype and methylation interaction (fixed effect), $u \sim N\left(0, \tau AA^T\right)$ is a vector of subject-specific random effect, $v \sim N\left(0, \gamma BB^T\right)$ is a vector of tissue-specific random effects, $w \sim N\left(0, \delta CC^T\right)$ is a vector of random effects that describes the interaction effect between genotype, methylation and tissue, $x \sim N\left(0, \theta DD^T\right)$ is a vector of random effects describing tissue-specific methylation effects and $\xi \sim N\left(0, \epsilon I_{nt}\right)$. The matrices J, A, B, C, and D are design matrices with B being a function of genotype, C is a function of both genotype and methylation data and finally, D is a function of just the methylation data. J is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. A is $nt \times n$ design matrix for the subject-specific intercepts. B is a $nt \times t$ design matrix of stacked genotypes.

C is a $nt \times t$ design matrix of stacked (product of) tissue-specific methylation and genotype data. D is $nt \times t$ design matrices of stacked tissue-specific methylation data. The parameters of interest are γ, δ, β and ϕ; $\alpha, \lambda, \tau, \theta$ and ϵ are nuisance parameters. Alternatively, we can represent the distribution of Y conditional on methylation and genotype as –

$$\left(Y|M = m, G = g\right) \sim N\left(J\alpha + G\beta + M\lambda + MG\phi, \Sigma\right)$$

From our model, the log-likelihood function of the parameters conditional on the genotype and methylation data is given by –

$$\begin{aligned}\ell\left(\Theta; Y|M = m, G = g\right) \\ = -c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(Y - J\alpha - G\beta - M\lambda - MG\phi\right)^T \\ \Sigma^{-1}\left(Y - J\alpha - G\beta - M\lambda - MG\phi\right)\end{aligned} \tag{3}$$

where Θ represents the vector of all the variance components involved in Σ and c is a constant. We test the null hypothesis that $H_0 : \beta = \phi = \gamma = \delta = 0$, i.e. the variant does not affect gene expression across any of the tissues. To do so, we compute the efficient scores for γ, δ, β and ϕ by projecting off components correlated with the nuisance parameters. The reduced model under the null is –

$$Y_{H_0} = J\alpha + M\lambda + Au + Dx + \xi$$

The efficient scores evaluated under the null are given by –

$$\begin{aligned}\text{Additive Genetic Effect} \coloneqq U_{\beta|H_0} &= \hat{Y}^T\hat{\Sigma}_n^{-1}\left(G - \bar{G}\right) \\ G \times M \text{ Effect} \coloneqq U_{\phi|H_0} &= \hat{Y}^T\hat{\Sigma}_n^{-1}\left(MG - \overline{MG}\right) \\ G \times T \text{ Effect} \coloneqq U_{\gamma|H_0} &= \frac{1}{2}\hat{Y}^T\hat{\Sigma}_n^{-1}BB^T\hat{\Sigma}_n^{-1}\hat{Y} \\ G \times M \times T \text{ Effect} \coloneqq U_{\delta|H_0} &= \frac{1}{2}\hat{Y}^T\hat{\Sigma}_n^{-1}CC^T\hat{\Sigma}_n^{-1}\hat{Y}\end{aligned}$$

where $\hat{Y}$ are the residuals from the model, $\bar{G}$ is an nt-dimensional vector of mean-centered genotypes, $\overline{MG}$ is an nt-dimensional vector of mean-centered product of genotypes and methylation, and $\hat{\Sigma} = \hat{\epsilon}I + \hat{\tau}ZZ^T + \hat{\theta}DD^T$. Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$). More on the individual components of our score test can be found in the Additional file 1 section.

We propose a weighted sum of the above components (under the null) to arrive at our joint score test statistic, U_ζ. Since U_β and U_ϕ are linear in Y while U_γ and U_δ are quadratic, we propose the following rule to combine them –

$$\begin{aligned}U_\zeta &\equiv \left(\boldsymbol{a_\beta}U_\beta^2 + \boldsymbol{a_\phi}U_\phi^2 + \boldsymbol{a_\gamma}U_\gamma + \boldsymbol{a_\delta}U_\delta\right) \\ &= \hat{Y}^T\hat{\Sigma}_n^{-1}\Big[a_\beta\left(G - \bar{G}\right)\left(G - \bar{G}\right)^T + a_\phi\left(MG - \overline{MG}\right) \\ &\quad \left(MG - \overline{MG}\right)^T + a_\gamma\frac{1}{2}BB^T + a_\delta\frac{1}{2}CC^T\Big]\hat{\Sigma}_n^{-1}\hat{Y}\end{aligned} \tag{4}$$

where a_β, a_ϕ, a_γ and a_δ are scalar constants chosen to minimize the variance of U_ζ. Under the null, U_ζ is distributed as a mixture of chi-square random variables. We use Satterthwaite method [28] to approximate the p values from a scaled χ^2 distribution by matching the first two moments as $U_\zeta \sim \kappa\chi_\nu^2$ where $\kappa = \frac{Var(U_\zeta)}{2E[U_\zeta]}$ and $\nu = \frac{2E[U_\zeta]^2}{Var(U_\zeta)}$.

Simulations

For a positive integer t that represents number of tissues, if $\mathbf{1}$ denotes a column vector of t ones and $\mathbb{I}$ denotes the corresponding $t \times t$ diagonal matrix, following the t-variate normal law denoted by $N_t[\mu, \Sigma]$ with mean $\mu \in \mathbb{R}^t$ and variance $\Sigma \in \mathbb{R}^{t\times t}$, expression levels of a target gene j at a single locus by using the following vectorized form of the linear mixed model –

$$\begin{aligned}y_{ij} = \alpha_j + \mathbf{1}\beta_j g_i + \mathbf{1}\lambda_j m_{ij} + \mathbf{1}\phi_j m_{ij} g_i + \mathbf{1}a_i + b_j g_i + c_j m_{ij} g_i \\ + d_j m_{ij} + \xi_{ij} \qquad \xi_{ij} \overset{i.i.d.}{\sim} N\left(0, \epsilon\mathbb{I}\right)\end{aligned} \tag{5}$$

where y_{ij} is a $t \times 1$ vector of gene expression data, α_t is the tissue-specific intercept ($\alpha_t \in \mathbb{R}^t$), β_j describes the main additive genotypic effect ($\beta_j \in \mathbb{R}^1$), λ_j describes the overall effect due to methylation ($\lambda_j \in \mathbb{R}^1$), ϕ describes the interaction effect between the overall methylation and genotype ($\phi_j \in \mathbb{R}^1$), g_i is the value of a bi-allelic genotype such that $g \in (0, 1, 2)$ represents the number of copies of the minor allele. The random effect $b_j \in \mathbb{R}^t$ represents tissue-specific effect of the genotype, $c_j \in \mathbb{R}^t$ represents tissue-specific interaction effect between methylation and genotype, $d_j \in \mathbb{R}^t$ represents tissue-specific methylation effect, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1(0, \tau)$, $b_j \sim N_t(0, \gamma\mathbb{I})$, $c_j \sim N_t(0, \delta\mathbb{I})$ and $d_j \sim N_t(0, \theta\mathbb{I})$. Methylation data was generated independently from a multivariate normal distribution with mean zero and positive definite variance-covariance matrix.

We use 1000 data replicates to evaluate the type I error and for power calculations. Simulations were performed by varying the following parameters- β (additive genetic effect), ϕ ($G \times M$ effect), the proportion of variation

explained by the $G \times T$ effect $\left(PVE_\gamma \equiv \left(\frac{\gamma}{\theta+\tau+\epsilon+\gamma+\delta}\right)\right)$ and the proportion of variation explained by the $G \times M \times T$ effect $\left(PVE_\delta \equiv \left(\frac{\delta}{\theta+\tau+\epsilon+\gamma+\delta}\right)\right)$. A linear mixed effects model was fit using the package *lme4* [29, 30] in the statistical environment R (R Core Team). The significance of an association between a mRNA - SNP pair in a tissue-by-tissue (TBT-eQTL) analysis is assessed by the p value obtained using *lm* function in R by fitting the following linear regression model.

For each mRNA - *cis*-SNP pair, TBT-eQTL analysis was performed using the following linear regression model –

$$Y = \beta_0 + \beta_1 G + \epsilon$$

where Y is either gene expression data and G represents genotypes encoded as the number of copies of minor allele. The test statistic is the minimum p value over the total number of tissues from linear regressions performed separately in each tissue for each mRNA - SNP pair. Statistical significance was determined at a nominal p value of 0.05 for all power simulations (in case of TBT-eQTL analysis, it is $\frac{0.05}{k}$ where k is the number of tissues).

Preprocessing Gibbs et al datasets

Data description

Fresh frozen tissue samples of the cerebellum (CRBLM), frontal cortex (FCTX), caudal pons (PONS) and temporal cortex (TCTX) were obtained from 150 neuropathologically normal samples [4]. Genotyping was performed using Infinium HumanHap550 beadchips (Illumina) to assay genotypes for 561,466 SNPs, from the cerebellum tissue samples. CpG methylation status was determined using HumanMethylation27 BeadChips (Illumina), which measure methylation at 27,578 CpG dinucleotides at 14,495 genes. Profiling of 22,184 mRNA transcripts was performed using HumanRef-8 Expression BeadChips (Illumina) The datasets are publicly available (GEO Accession Number: **GSE15745**; dbGAP Study Accession: **phs000249.v1.p1**).

Gene expression data

Gene expression on four brain regions are publicly available as rank-invariant [31] normalized gene expression data ("series matrix file"). All the negative values in the gene expression dataset are changed to a 1 and the entire dataset was then log2 transformed. Before generating the PCA plots, samples with African and Asian ancestry (n = 2) were removed from the analysis in order to keep the study a homogenous mixture of European-Caucasians. All the gene expression probes on sex chromosomes X and Y were removed from the analysis.

Each gene expression probe was then adjusted for known variation contributed by batch effects and biological covariates such as tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age as well as unknown variation using surrogate variable analysis (SVA) model [32].

$$\begin{aligned}\text{Gene Expression} \sim\ & \text{Biological Covariates} \\ & + \text{Known Batch Effects} \\ & + \text{Unkown Variation} \\ & + \text{Measurement Error}\end{aligned}$$

It has previously been shown that the number of *cis*-eQTL detected significantly improved when multiple PCs were removed from the expression data [33].

Methylation data

Methylation data, obtained as a "series matrix file" consisted of Beta-values, which represent the ratio of methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities) [34]. We followed the previously mentioned method to preprocess methylation data using the SVA model. The biological covariates here include tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age.

Genotype data

The genotype data was obtained from dbGAP database (**phs000249.v1.p1**) following requisite author permissions. The genotype data was recoded into a SNP matrix of values 0, 1 and 2 representing minor allele counts. Samples with African and Asian ancestry were removed from the analysis in order to keep the data relatively homogeneous with patients of European-Caucasian ancestry. These SNPs were filtered on the missing-ness of the individual data and the SNP data (excluded SNPs with missing values), followed by MAF (included SNPs with MAF $\geq$ 0.05)and Hardy-Weinberg equilibrium (HWE; p-values $\leq$ 0.001) in the same order using PLINK [35] software. The resulting dataset has 400,097 SNPs after preprocessing.

Additional file

Additional file 1: Supplementary material. Supplementary material expanding on 1) Our model, 2) Individual components of our joint score test statistic, 3) Description of various null hypotheses, 4) Null and power simulations of our joint score test statistic, 5) Gibbs et al. dataset preprocessing, 6) Design of our data analysis, 7) KEGG pathway analysis on the results from Gibbs et al brain data, 8) JAGUAR, 9) A potential strategy to combine two models to maximize eQTL discovery, and 10) Reproducibility. (PDF 635 kb)

Abbreviations

eQTL: Expression quantitative trait loci; FDR: False discovery rate; FWER: Family wise error rate; GWAS: Genome wide association study; KEGG: Kyoto encyclopedia of genes and genomes; mQTL: Methylation quantitative trait loci; PCA: Principal components analysis; SNP: Single nucleotide polymorphism; TBT-eQTL: Tissue-by-tissue

Acknowledgements
The authors acknowledge the Research Computing Center at Duke University for providing high performance computing resources that have contributed to the research results reported within this paper and wish to thank Thomas Milledge in particular for his help to use the Duke Shared Cluster Resource. We wish to thank Dr. Janice McCarthy at Duke Department of Biostatistics and Bioinformatics for useful discussions. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health.

Funding
This research was supported in part by Award number P01CA142538 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
All the gene expression and methylation data are publicly available and the de-identified genotype data is available on database of Genotypes and Phenotypes (dbGaP), which provides controlled access. All the R scripts for both data simulations and real data analyses are available at https://github.com/cramanuj/Epigen_Rcodes.

Authors' contributions
AA and CA designed the study. CA developed methods, derived the model, implemented software and analyzed data. AA, CA, and KO wrote, read and approved the final manuscript.

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 5 October 2016 Accepted: 2 October 2017
Published online: 18 October 2017

References

1. Ong CT, Corces V. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12:283–93.
2. Geyer PK, Green MM, Corces VG. Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila. EMBO J. 1990;9:2247–56.
3. Bell J, Pai A, Pickrell J, Gaffney D, Pique-Regi R, Degner J, Gilad Y, Pritchard J. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology. 2011;12(R10):.
4. Gibbs J, van der Brug M, Hernandez D, Traynor B, Nalls M, Lai SL, Arepally S, Dillman A, Rafferty I, Troncoso J, Johnson R, Zielke H, Ferrucci L, Longo D, Cookson M, Singleton A. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010;6(5):.
5. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010;20:883–9.
6. Wrzodek C, Büchel F, Hinselmann G, Eichner J, Mittag F, Zell A. Linking the Epigenome to the Genome: Correlation of Different Features to DNA Methylation of CpG Islands. Plos ONE. 2012;7(4):.
7. Lemire M, Zaidi S, Ban M, Ge Bea. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat Commun. 2014;6(6326):.
8. Deaton A, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010–22.
9. Wagner J, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15:.
10. Hellman A, Chess A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics Chromatin. 2010; 24(3):1.
11. Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery S, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, Falconnet E, Biesler D, Gagnebin M, Giger T, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis S, Dermitzakis E. Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. PLoS Genet. 2015.
12. Banovich N, Lan X, McVicker G, van de Geijn B, Degner J, Blischak J, Roux J, Pritchard J, Gilad Y. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. PLoS Genet. 2014;10(9):.
13. Swift-Scanlan T, Smith C, Bardowell S, Boettiger C. Comprehensive interrogation of CpG island methylation in the gene encoding COMT, a key estrogen and catecholamine regulator. BMC Med Genet. 2014.
14. Acharya C, McCarthy J, Owzar K, Allen A. Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics. 2016;17(257). doi:10.1186/s12859–016–1123–5.
15. Gutierrez-Arcelus M, Lappalainen T, Montgomery S, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Biesler D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis S, Dermitzakis E. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife. 2013;2:.
16. Flutre T, Wen X, Pritchard J, Stephens M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS Genet. 2013;9(5):.
17. Sul J, Han B, Ye C, Choi T, Eskin E. Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches. PLoS Genet. 2013;9(6):.
18. Acharya CR, Allen AS. JAGUAR: Joint Analysis of Genotype and Group-Specific Variability Using a Novel Score Test Approach to Map Expression Quantitative Trait Loci (eQTL). 2016. [https://CRAN.R-project.org/package=JAGUAR]. [R package version 3.0.1].
19. Storey J, Tibshirani R. Statistical significance for genome-wide experiments. PNAS. 2003;100(16):9440-445.
20. Lonsdale J, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.
21. Peukert D, Weber S, Lumsden A, Scholpp S. Lhx2 and Lhx9 Determine Neuronal Differentiation and Compartition in the Caudal Forebrain by Regulating Wnt Signaling. PLoS Biol. 2011;9(12):e1001218.
22. Lai R, Crevier L, Thabane L. Genetic polymorphisms of glutathione S-transferases and the risk of adult brain tumors: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2005;14(7):1784–90.
23. Yu G, Wang L, Yan G, He Q. DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis. Bioinformatics. 2014;31(4):608–9.
24. Micheli V, Camici M, Tozzi M, Ipata P, Sestini S, Bertelli M, Pompucci G. Neurological disorders of purine and pyrimidine metabolism. Curr Top Med Chem. 2011;11(8):.
25. McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, Ioannidis J, Hirschhorn J. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356–69.
26. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium G, Nicolae DL, Cox NJ, Im HK. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8. [http://dx.doi.org/10.1038/ng.3367].
27. Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis D, Sommer S, Rossi J. SNPs in human miRNA genes affect biogenesis and function. RNA. 2009;15: 1640–51.
28. Satterthwaite F. An approximate distribution of estimates of variance components. Biom Bull. 1946;2(6):110–4.
29. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. 2014. [http://CRAN.R-project.org/package=lme4]. [R package version 1.1-7].
30. Bates D, Maechler M, Bolker BM, Walker S. lme4: Linear mixed-effects models using Eigen and S4. 2014. [http://arxiv.org/abs/1406.5823]. [ArXiv e-print; submitted to *Journal of Statistical Software*].

31. Schmid R, Baum P, Ittrich C, Fundel-Clemens K, Huber W, Brors B, Eils R, Weith A, Mennerich D, Quast K. Comparison of normalization methods of Illumina BeadChip HumanHT-12 v3. BMC Genomics. 2010;11:.
32. Leek J, Storey J. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.
33. Fu J, Wolfs M, Deelen P, Westra H, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 2012;8:.
34. Du P, Zhang X, Huang C, Jafari N, Kibbe W, Hou L, Lin S. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:.
35. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool-set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.