Published in final edited form as:
Ann Appl Stat. 2016 ; 10(4): 2254–2273. doi:10.1214/16-AOAS973.

A STATISTICAL MODEL TO ASSESS (ALLELE-SPECIFIC) ASSOCIATIONS BETWEEN GENE EXPRESSION AND EPIGENETIC FEATURES USING SEQUENCING DATA

Naim U. Rashid[*], **Wei Sun**[†], and **Joseph G. Ibrahim**[*]

[*]University of North Carolina at Chapel Hill

[†]Fred Hutchinson Cancer Research Center

Abstract

Sequencing techniques have been widely used to assess gene expression (i.e., RNA-seq) or the presence of epigenetic features (e.g., DNase-seq to identify open chromatin regions). In contrast to traditional microarray platforms, sequencing data are typically summarized in the form of discrete counts, and they are able to delineate allele-specific signals, which are not available from microarrays. The presence of epigenetic features are often associated with gene expression, both of which have been shown to be affected by DNA polymorphisms. However, joint models with the flexibility to assess interactions between gene expression, epigenetic features and DNA polymorphisms are currently lacking. In this paper, we develop a statistical model to assess the associations between gene expression and epigenetic features using sequencing data, while explicitly modeling the effects of DNA polymorphisms in either an allele-specific or nonallele-specific manner. We show that in doing so we provide the flexibility to detect associations between gene expression and epigenetic features, as well as conditional associations given DNA polymorphisms. We evaluate the performance of our method using simulations and apply our method to study the association between gene expression and the presence of DNase I Hypersensitive sites (DHSs) in HapMap individuals. Our model can be generalized to exploring the relationships between DNA polymorphisms and any two types of sequencing experiments, a useful feature as the variety of sequencing experiments continue to expand.

keywords and phrases

Bivariate binomial logistic-normal (BBLN) distribution; bivariate Poisson log-normal (BPLN) distribution; DNase-seq; genetics; genomics; RNA-seq

Department of Biostatistics, University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, North Carolina 27599, USA, naim@unc.edu, ibrahim@bios.unc.edu
Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, Washington 98109, USA, wsun@fredhutch.org

SUPPLEMENTARY MATERIAL

Supplement to "A Statistical model to assess (allele-specific) associations between gene expression and epigenetic features using sequencing data" (DOI: 10.1214/16-AOAS973SUPP; .pdf). Contains details on numerical maximization procedures for the BBLN and BPLN models.

1. Introduction

Gene expression regulation is an essential biological process by which static genetic information gives rise to dynamic organismal phenotypes [Jaenisch and Bird (2003)]. Multiple epigenetic features are involved in gene expression regulation, including DNase I hypersensitive sites (DHSs) [Song et al. (2011)], DNA methylation [Fang et al. (2012)] and histone modifications [Heintzman et al. (2009)]. DHSs, which delineate open chromatin regions, are among the most well-studied epigenetic features. DHSs often harbor regulatory DNA elements that can influence gene expression [Thurman et al. (2012)], and thus the presence or absence of DHSs is often associated with gene expression variation [Djebali et al. (2012)]. Both gene expression and DHSs are heritable [McDaniell et al. (2010)], and previous studies have found their variations are often associated with DNA variants such as single nucleotide polymorphisms (SNPs) [Degner et al. (2012), Pickrell et al. (2010)]. Characterizing these associations plays an important role in understanding how one's genotype modifies phenotype, such as in Cowper-Sal et al. (2012), where the authors systematically determined SNPs associated with breast cancer and found these SNPs are over-represented on the binding sites of a transcription factor FOXA1. They then confirmed that these SNPs modified the FOXA1 binding strength, which further leads to imbalance of downstream gene regulation.

Gene expression and epigenetic features are being routinely assessed by high-throughput sequencing solutions, and the results are quantified by the number of sequenced reads within certain genomic regions. For example, the number of RNA-seq reads within a gene provides a measure of gene expression, which can be further normalized by read depth (the total number of sequencing reads sampled per individual) and gene length to facilitate comparison across individuals and across genes. Sequencing data not only provide more comprehensive and more accurate assessments of genomic activity, but also reveal novel information that is not available from traditional microarrays, such as allele-specific signals. In a diploid genome, the DNA sequence at each autosomal locus has two copies (i.e., the maternal and paternal copy), and each copy is referred to as an allele.

Recently, allele-specific signals have been studied in various sequencing studies, including gene expression [Pickrell et al. (2010)], DNA methylation [Fang et al. (2012)], transcription factor binding [Rozowsky et al. (2011)] and chromatin accessibility [Degner et al. (2012)]. Such allele-specific signals can be used to distinguish *cis*-acting and *trans*-acting genetic effects [Sun (2012)]. A *cis*-acting DNA polymorphism only modifies expression of genes or epigenetic features that are located on the same haploid genome as the DNA polymorphism. In contrast, a *trans*-acting DNA polymorphism has the same effect on both alleles of its target. Therefore, an imbalance of Allele-Specific Read Counts (ASReCs) of the two alleles within one individual implies the presence of a *cis*-acting regulatory element, and the variation of the Total Read Count (TReC, summation of read count from either allele) across individuals can be due to either *cis*-acting or *trans*-acting regulations.

Previous studies have demonstrated the association between gene expression and epigenetic features using either TReC or ASReC and their associations with DNA polymorphisms. Unfortunately, no study has systematically assessed the joint associations between gene expression, epigenetic features and underlying genotype. Furthermore, no method exists to determine such associations with allele-specific sequencing data (ASReC). To address this issue, we develop a novel statistical method, which we refer to as BASeG (Bivariate Aassociation studies using Sequencing data, while accounting for shared Genetic effects). Specifically, we study the association of TReC and ASReC using Bivariate Poisson-Log-Normal (BPLN) regression and Bivariate Binomial-Logistic-Normal (BBLN) regression, respectively. We demonstrate BASeG's utility in simulations and a study of the association between gene expression (measured by RNA-seq) and DHSs (measured by DNase-seq). BASeG is general enough to be applied to study the associations between any two types of sequencing data, such as gene expression (by RNA-seq) vs. DNA methylation measured by bisulfite sequencing or histone modifications measured by ChIP-seq (Chromatin Immunoprecipitation followed by sequencing).

2. Model

2.1. Bivariate Poisson-log-normal regression for Total Read Count (TReC)

Assume we are interested in the RNA-seq TReC of a particular gene, denoted by T_R, and the DNase-seq TReC within a particular genomic region (e.g., a 250-bp window in the promoter of the gene of interest), denoted by T_C in the ith sample. For notational simplicity, we drop sample subscript i for now. We assume the expected value of T_R is associated with a genetic variable Z_R and some other covariates X_R, and, similarly, the expected value of T_C is associated with a genetic variable Z_C and some other covariates X_C. Such covariates may include the log of the sequencing depth for each sample (the log transformation is due to the fact that our model of TReC has a log link function), as well as demographic variables and/or batch effects. We also assume the genetic effect is additive such that Z_R or Z_C equals 0, 1 or 2, which is the number of nonreference (alternative) alleles of the SNP. In this study, the reference allele of a SNP is defined based on the 1000 Genomes Project SNP annotation file and this definition is applied consistently across samples. Without loss of generality, we also assume that this genetic effect jointly impacts each data type (i.e., gene expression or DHSs), allowing us to assess whether the observed correlation of gene expression and DHSs is due to a joint effect of a single SNP. It is straightforward to define other types of genetic effects (e.g., dominant or co-dominant) if desired. We model the joint distribution of T_R and T_C by a bivariate Poisson-log-normal (BPLN) distribution:

$$f_{\mathrm{BPLN}}(T_R, T_C) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathrm{P}}(T_R;\mu_R) f_{\mathrm{P}}(T_C;\mu_C)\phi(\varepsilon_R, \varepsilon_C; \sum\nolimits_1)\, d\varepsilon_{Ri} d\varepsilon_{Ci}, \quad (2.1)$$

where $f_{\mathrm{P}}(;\mu)$ denotes the Poisson distribution probability mass function with mean μ. For RNA-seq and DNase-seq data, we assume $\log(\mu_R) = X_R\beta_R + Z_R b_R + \varepsilon_R$ and $\log(\mu_C) = X_C\beta_C + Z_C b_C + \varepsilon_C$, respectively, where ε_R and ε_C are two random variables following a bivariate normal distribution with mean $\mathbf{0}$ and covariance Σ_1, denoted by the bivariate normal probability density function $\phi(\varepsilon_R, \varepsilon_C; \Sigma_1)$,

$$\sum_1 = \begin{pmatrix} \sigma_R^2 & \rho_1\sigma_R\sigma_C \\ \rho_1\sigma_R\sigma_C & \sigma_C^2 \end{pmatrix},$$

and −1 ≤ ρ_1 ≤ 1 is a correlation parameter. Therefore, in this BPLN distribution, the correlation, in the absence of a shared genetic effect, between T_R and T_C is induced by the correlation ρ_1 between ε_R and ε_C. We compare our model with that of a generalized mixed linear model framework with heterogeneous variances in the discussion section of this manuscript.

The probability mass function of (T_R,T_C) is obtained by integrating out the random effects ε_R and ε_C. To efficiently approximate this integral computationally, we utilize a multivariate form of adaptive Gauss-Hermite quadrature [Liu and Pierce (1994)]:

$$f_{\mathrm{BPLN}}(T_R, T_C) \approx \sum_{j=1}^{s}\sum_{k=1}^{s} w_j^* w_k^* f_{\mathrm{P}}\left(T_R; \mu_R^*\right) f_{\mathrm{P}}\left(T_C; \mu_C^*\right) \phi\left(\varepsilon_j^*, \varepsilon_k^*; \sum_1\right), \quad (2.2)$$

where the s quadrature nodes ε_j^* and ε_k^* are chosen with respect to the mode of the integrand and are scaled according to the estimated curvature at the mode, and weights w_j^* and w_k^* are utilized as defined in Section 1 of the Supplementary Material [Hartzel, Agresti and Caffo (2001), Rashid, Sun and Ibrahim (2016)]. Here $\log(\mu_R^*) = X_R\beta_R + Z_R b_R + \varepsilon_j^*$ and $\log(\mu_C^*) = X_C\beta_C + Z_C b_C + \varepsilon_k^*$. Adaptive quadrature approaches are typically utilized to increase the accuracy of an integral approximation while utilizing fewer quadrature points to control computational cost. Details regarding the adaptive quadrature procedure are given in the Supplementary Material. For all simulations and real data analyses in this manuscript we have used s = 10 quadrature points.

The log likelihood corresponding to all n samples can then be expressed as

$$l_{\mathrm{BPLN}}(T_R, T_C) = \sum_{i=1}^{n} \log[f_{\mathrm{BPLN}}(T_{Ri}, T_{Ci})].$$

The derivatives of this log likelihood can be factored into the form of (2.2), and thus maximization with respect to the parameters β_R,β_C, b_R, b_C,σ_R,σ_C and ρ_1 can be performed via quasi-newton methods such as L-BFGS-B. We provide further details of the maximization procedure in the Supplementary Material.

2.2. Bivariate Binomial-logistic-normal regression for Allele-specific Read Counts (ASReC)

Next we consider the statistical model for allele-specific read counts (ASReC). Similar to the previous section, we wish to assess conditional correlations after accounting for genetic effects. As before, we drop the subject subscript i for notational simplicity and describe the PMF for a single sample. For a gene of interest, we assume its two haplotypes are known, and denote them by h_1 and h_2, respectively. Let N_{R1} and N_{R2} be the number of allele-specific RNA-seq reads from haplotype h_1 and h_2, respectively, and let $N_R = N_{R1} + N_{R2}$. Analogously, we define N_{C1}, N_{C2} and N_C for the DNase-seq data. We exclude those samples with $N_C < u$ or $N_R < u$ for ASReC studies because allelic imbalance cannot be reliably estimated when there are few allele-specific reads. In the following real data studies, we set $u = 1$. For the remaining samples, we model the joint distribution of N_{R1} and N_{C1} by a Bivariate Binomial-Logistic-Normal regression model (BBLN), denoted by $f_{\mathtt{BBLN}}$:

$$f_{\mathrm{BBLN}}(N_{C1}, N_{R1}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathrm{B}}(N_{R1}; N_R, \pi_R) f_{\mathrm{B}}(N_{C1}; N_C, \pi_C) \phi(\xi_R, \xi_C; \sum\nolimits_2) \, d\xi_R d\xi_C,$$

where $f_{\mathrm{B}}(;N,\pi)$ denotes the binomial distribution probability mass function with N trials and probability of success π. In this scenario, success pertains to a read's alignment to haplotype h_1. We define π_R and π_C to be the success probabilities in the RNA-seq and DNase-seq data, respectively, given some possible underlying genetic effect. We model π_R and π_C such that $\log[\pi_R/(1 - \pi_R)] = v_R E_R + \xi_R$ and $\log[\pi_C/(1-\pi_C)] = v_C E_C + \xi_C$, where E_R or E_C describes the allele-specific effect of a SNP:

$$E_R \text{ (or } E_C) = \begin{cases} 0, & \text{if the SNP is homozygous}, \\ -1, & \text{if the SNP is heterozygous and its reference allele is on haplotype } h_1, \\ 1, & \text{if the SNP is heterozygous and its reference allele is on haplotype } h_2; \end{cases}$$

that is, the success probability in each data type may be related to an allele-specific effect of an underlying SNP. When the SNP is homozygous, it has the same allele in both haplotypes, and thus cannot lead to any allelic imbalance of gene expression. Therefore, E_R (or E_C) = 0 if the SNP is homozygous. When the SNP is heterozygous and it is responsible for allelic imbalance of gene expression, the higher expression haplotype may have either reference allele or alternative allele. The magnitude of this effect in each data type is conveyed by v_R and v_C. Thus, the definition of genetic effect relies on which haplotype has the reference allele. The confounding covariates X_R or X_C used for TReC model are ignored because such covariates' effects are often canceled out when we compare the expression of one allele vs. the other allele. It is straightforward to add such effects back into the model if needed. Similarly to the model for TReC data, we assume ξ_C and ξ_R follow a bivariate normal distribution: $\phi(\xi_C, \xi_R; \Sigma_2) \sim \mathcal{N}(\mathbf{0}, \Sigma_2)$, where

$$\sum\nolimits_2 = \begin{pmatrix} \kappa_R^2 & \rho_2 \kappa_R \kappa_C \\ \rho_2 \kappa_R \kappa_C & \kappa_C^2 \end{pmatrix},$$

and $-1 \le \rho_2 \le 1$ is the correlation parameter. Therefore, in the absence of a shared genetic effect, the dependence between the observed allele-specific read counts (N_{R1} and N_{C1}) is induced by the correlation parameter ρ_2 between ξ_C and ξ_R. We compare and contrast our model with that of a generalized mixed linear model framework with heterogeneous variances in the discussion section of this paper.

Finally, the joint log likelihood of ASReC for *n* individuals is

$$l_{\mathrm{BBLN}}(N_{Ci1}, N_{Ri1}) = \sum_{i=1}^{n} I\left(N_{Ri} \geq u \text{ and } N_{Ci} \geq u\right) \log\left[f_{\mathrm{BBLN}}(N_{R1i}, N_{C1i})\right],$$

where $I(\,)$ is an indicator function. We obtain the MLE (Maximum Likelihood Estimate) of the parameters similarly to the BPLN model for TReC data; see the Supplementary Material for details.

2.3. Testing framework using TReC or ASReC

Utilizing the MLE of the above models, we employ likelihood ratio tests (LRTs) with degree of freedom 1 to assess the correlation between gene expression and DHS site. Specifically, we will conduct the following four tests:

1. *Assess the correlation between RNA-seq and DNase-seq TReC in the presence of genetic effects*. Conduct the LRT using the TReC likelihood with H_0: $\rho_1 = 0$ vs. H_1: $\rho_1 \neq 0$.
2. *Assess the correlation between RNA-seq and DNase-seq TReC in the absence of genetic effects*. Conduct the LRT using the TReC likelihood with H_0: $b_R = b_C = \rho_1 = 0$ vs. H_1: $b_R = b_C = 0$, and $\rho_1 \neq 0$.
3. *Assess the correlation between RNA-seq and DNase-seq ASReC in the presence of genetic effects*. Conduct the LRT using the ASReC likelihood with H_0: $\rho_2 = 0$ vs. H_1: $\rho_2 \neq 0$.
4. *Assess the correlation between RNA-seq and DNase-seq ASReC in the absence of genetic effects*. Conduct the LRT using the ASReC likelihood H_0: $v_R = v_C = \rho_2 = 0$ vs. H_1: $v_R = v_C = 0$, and $\rho_2 \neq 0$.

It is also desirable to test the two null hypotheses $\rho_1 = 0$ and $\rho_2 = 0$ simultaneously as a two degree of freedom test. However, it is possible that only one of the null hypotheses is correct in certain situations. For example, if the association between gene expression and DHS is totally due to a common *cis*-acting SNP (i.e., $Z_C = Z_R$) and the SNP is heterozygous across all individuals, then without conditioning on SNP genotype, $\rho_1 = 0$ but $\rho_2 \neq 0$.

We conduct a genome-wide assessment of the dependency between gene expression and DHS in the following steps. First, for each gene, we only consider the DHSs that are local (e.g., within 2 kb) since distant DHSs are unlikely to influence gene expression and would increase the burden of multiple testing correction. Second, for each gene and each DHS, we only consider the SNPs that are close to either feature (e.g., within 2kb of either feature), which has been a common practice in previous eQTL studies [Sun (2012)]. Our method allows distinct SNPs to be associated with the RNA-seq and DNase-seq data, respectively. However, since our focus is to account for the case where the dependence between gene expression and DHS is induced by shared genetic effect, we choose to use the same SNP for RNA-seq and DNase-seq data (i.e., $Z_R = Z_C$). Another important motivation for this strategy is to reduce the multiple testing burden. For example, if there are 100 SNPs around a gene-

DHS pair, we correct for the multiple tests across 100 SNPs in the case of a common SNP effect $Z_R = Z_C$. However, if we allow two distinct SNPs to be associated with the RNA-seq and DNase-seq data ($Z_R \neq Z_C$), 10,000 SNP combinations will be evaluated, with much higher multiple testing burden and more complicated correlation structures among the 10,000 tests. We note that a SNP that is found to explain the correlation between two data types may not be the only possible SNP to do so, as we do not survey every single SNP in the genome for association. Furthermore, it is possible that two separate SNPs may jointly explain such correlation. However, given previous interest in searching for common SNPs with a joint effect [Degner et al. (2012)], we focus the rest of the manuscript assuming a joing SNP effect.

3. Results

3.1. Simulation studies

We use simulated data to evaluate the power and type I error of the tests in Section 2.3 for a triplet of gene expression, DHS and SNP. First, TReC data were simulated from $f_{\mathtt{BPLN}}$ under the combinations of the following situations:

- Sample size: $n = 50$, 100 or 300.
- SNP minor allele frequency: 0.5.
- SNP effect: $b_R = b_C = 0$, 0.05, 0.075, 0.1, 0.15 or 0.2.
- Four covariates. The first one is the intercept, the other three are simulated from uniform (0, 1) distribution. The coefficients are $\beta_C = (2.5, 0.5, 0.5, 0.5)$ and $\beta_R = (2.5, 1, 1, 1)$.
- Variance: $\sum = \begin{bmatrix} 0.1 & 0.1^2\rho_1 \\ 0.1^2\rho_1 & 0.1 \end{bmatrix}$, with $\rho_1 = 0$, 0.05, 0.1, 0.15, 0.2, 0.25, 0.35 or 0.5.

The simulation study results are summarized in Figure 1. We note that b_R and b_C represent the effect of the common SNP on read counts in each data type, whereby larger values of each induce more correlation in read counts. Therefore, if one accounts for the SNP effect in the BPLN model, the estimated correlation parameter will be much smaller in this model relative to the model that ignores the SNP effect. For testing $\rho_1 = 0$ in the presence of a shared genetic effect (Figure 1A), there is slight inflation of Type I error for small sample sizes ($n = 50$); however, such inflation disappears as sample size increases ($n = 100$ or 300). When shared genetic effects on RNA-seq and DNase-seq are ignored, testing the correlation between RNA-seq and DNase-seq TReC data has inflated Type I error, and such inflation increases as the genetic effects b_R and b_C increase (Figure 1B). This suggests the importance of accounting for genetic effects in our model, as the correlation between TReC counts may be induced by a shared genetic effect. We also find that the power for detecting the correlation between RNA-seq and DNase-seq increases greatly with sample size (Figure 1C). When the sample size is 50, we achieve approximately 80% power to detect correlation $\rho_1 = 0.5$. For $n = 300$, we achieve 80% power to detect correlation $\rho_1 = 0.2$. The power calculations in Figure 1C correspond to data simulated such that $b_R = b_C = 0$, while results for other values of b_R and b_C are similar. Reducing the MAF in our model from 0.5 to 0.1, we find that our power analysis with respect to ρ_1 is unchanged, as we utilize data from all subjects regardless of genotype to estimate ρ_1 (Supplementary Figure S1A).

Next, we simulated ASReC data from $f_{\mathsf{BBLN}}(N_{Ri1}, N_{Ci1})$ over the following situations:

- Sample size: n=50 or 100.
- SNP minor allele frequency: 0.5.
- SNP effect: $v_R = v_C = 0, 0.2, 0.3$ or 0.4.
- $N_R, N_C \sim$ Poisson(λ), $\lambda = 5, 20$ or 100.
- Variance: $\sum_2 = \begin{bmatrix} 0.1 & 0.1^2\rho_2 \\ 0.1^2\rho_2 & 0.1 \end{bmatrix}$, where $\rho_2 = 0, 0.05, 0.1, 0.15, 0.2, 0.25,$ 0.035 and 0.5.

The simulation results are shown in Figure 2. When we account for the shared genetic effect, testing for $\rho_2 = 0$ has little inflation of Type I error, regardless of the values of π_1 and π_2 or the total number of allele-specific reads (Figures 2A–B). Under model misspecification where we ignore genetic effects (i.e., assuming $v_R = 0$ and $v_C = 0$ or, equivalently, $\pi_{Ri} = \pi_{Ci} = 0.5$), type I error in testing for $\rho_2 = 0$ increases as π_R and π_C deviate from 0.5 (Figures 2C–D). In Figures 2E–F, we find that the power for testing for $\rho_2 = 0$ is mostly a function of the total number of allele-specific reads, while sample size has little effect on power. For example, doubling the sample size from $n = 50$ to $n = 100$ leads only to modest gains in power, mostly at lower levels of ρ_2. Notably, having only 5 total allele-specific reads per site has almost zero power to detect correlation. This observation justifies our suggestion of ignoring allele-specific read data when there are few allele-specific reads. Similar to the BPLN simulation, decreasing MAF to 0.1 does not have a large impact on our power to detect ρ_2 (Supplementary Figure S1B).

3.2. Real data analysis

We applied our method to study the DNase-seq and RNA-seq data of 60 HapMap YRI individuals [Degner et al. (2012), Pickrell et al. (2010)]. The data were downloaded from http://eqtl.uchicago.edu/. Given the results in simulation studies with respect to model misspecification, we seek to assess gene-DHS association in the presence of a common SNP effect.

3.2.1. Genotype data preparation—Among these 60 individuals, 42 have phased genotypes from the 1000 Genomes Project (TGP) Phase I Release Version 3 [1000 Genomes Project Consortium et al. (2012)] consisting of 36 million SNPs. For the remaining 18 individuals we obtained their corresponding HapMap r27 genotypes consisting of approximately 3 million SNPs, and imputed the genotypes and haplotypes on TGP SNPs using MACH 1.0 [Li et al. (2010)] with the TGP AFR (African population) reference panel. Prior to imputation, about 4000 HapMap SNPs whose rsIDs have changed between human genome build hg18 and hg19 were removed using the liftRsNumber tool (http://genome.sph.umich.edu/wiki/LiftRsNumber.py).

3.2.2. Tabulating TReC for RNA-seq and DNase-seq data—Raw data of paired-end RNA-seq reads were downloaded from http://eqtl.uchicago.edu/RNA_Seq_data/unmapped_reads/ and were mapped to human genome build hg19 using Tophat version 2.0.6 [Trapnell, Pachter and Salzberg (2009)] given Ensembl transcriptome annotation (GRCh37 release 66). All lanes of data pertaining to the same individual were merged subsequent to mapping.

We obtained the RNA-seq TReC for each gene by first counting the number of RNA-seq reads that overlap with exonic regions using R function `countReads` in R package `R/isoform` (http://research.fhcrc.org/sun/en/software/isoform.html) [Sun et al. (2014)]. To account for possible batch effects in the RNA-seq TReC data, we computed and retained the first 6 principal components from the TReC data matrix for later association analysis using TReC data. Specifically, the count data was first transformed such that $x_{ij}=\log[(y_{ij}+1/6)/(\sum_{j=1}^{P} y_{ij})]$, where y_{ij} is the original count for sample i, $i = 1 \dots n$ and feature j, $j = 1 \dots P$. P is the total number of features and n is the total number of samples. Mapped single-end DNase-seq reads were downloaded from http://eqtl.uchicago.edu/dsQTL_data/MAPPED_READS/ and were lifted over from build hg18 to hg19 to preserve the quality controls performed in a previous study [Degner et al. (2012)]. Total DNase-seq read counts were tabulated using BedTools v2.17 [Quinlan and Hall (2010)] for each of 1.5 million 100 bp candidate regions defined in Degner et al. (2012); and following Degner et al. (2012), we assigned a read to a candidate region based on the 5′ start position of each read. We also computed and retained the first 6 principal components from the DNase-seq TReC data matrix and used them as part of the association analysis using TReC data.

The allele-specific reads mapped to haplotype 1 and haplotype 2 in the RNA-seq data were extracted given the list of heterozygous SNPs per individual using R function `extractAsReads` in R package `R/asSeq` (http://research.fhcrc.org/sun/en/software/asSeq.html) [Sun (2012)]. The isolation of allele-specific DNase-seq reads was performed using the function `asCountsBED5` from the R package developed for this manuscript `BASeG`. Then the Allele-specific Read Count (ASReC) per gene and per haplotype was counted using R function `countReads`. As mentioned earlier, adjusting for confounding factors is often not necessary in the allele-specific analysis since the ASReC from one haplotype is directly compared to the other haplotype within an individual, serving as its own control, and thus we do not use any covariate other than genotype for association analysis using ASReC data. Other packages for TReC and ASReC read count tabulation may be utilized, as our method will accept any $n \times p$ table of counts as input for each data type, where n is the number of samples and p is the number of features being considered for a particular data type.

We performed some additional filtering before our analysis. We removed genes and DNase-seq candidate regions without enough TReC or ASReC. Specifically, we kept features for our allele-specific analysis that had ≥10 allele-specific reads in at least 10 individuals. For our TReC-based analysis, we kept genes that had an FPKM (Fragments Per Kilobase of sequence Per Million total reads) ≥ 3 in at least 15 individuals and DHSs with RPM (Reads Per Million total reads) ≥ 3 in at least 15 individuals, where total sequencing read depth was the sum of number of reads across all sites in an individual. We also removed SNPs with minor allele frequency (MAF) less than 0.05. During testing, a Gene-DHS pair was skipped if less than 10 individuals had at least 10 allele-specific reads for either type of the data. The final number of features utilized for testing in each data type and for each chromosome is given in Supplementary Tables S1 and S2 in the Supplementary Material. We only performed testing between genes and DHS candidate regions (DHS for short) that are within 2 Kb of each other, and only consider SNPs that are within 2 Kb of either feature. Using TReC data, we tested 9368 gene-DHS pairs (consisting of 2841 genes and 8689 DHSs), with 9.97 SNPs per gene-DHS pair on average. After removing results from gene-DHS-SNP trios that failed during testing (approximately 14%), we are left with 8689 gene-DHS pairs.

We summarized the results for each gene-DHS pair by three *p*-values:

- $p_{\texttt{uncond}}$: the *p*-value without conditioning on any SNP.
- $p_{\texttt{max}}$: the maximum of the *p*-values conditioning on each of the local SNPs.
- $p_{\texttt{min.corr}}$: the minimum of the *p*-values conditioning on each of the local SNPs, after multiple testing correction.

Suppose M_k local SNPs are considered as possible genetic factors of the *k*th gene-DHS pair, and denote the *p*-values conditioning on each of these SNPs by $(\varrho_1, \ldots, \varrho_{M_k})$. Then $p_{\texttt{min.corr}} = \min(1, \min(\varrho_1, \ldots, \varrho_{Mk})M_{k,\mathrm{eff}})$, where $M_{k,\mathrm{eff}}$ is the effective number of independent SNPs of the M_k SNPs [Nyholt (2004)]:

$$M_{k,\mathrm{eff}}=1+(M_k-1)\left(1-\frac{\mathrm{var}(\lambda_{\mathrm{obs}})}{M_k}\right),$$

and $\mathrm{var}(\lambda_{\mathrm{obs}})$ is the variance of the observed eigenvalues from the correlation matrix of the M_k SNPs. A precise correction for multiple testing correction for $p_{\texttt{max}}$ can be conducted as follows. First we can assume the *p*-values of the M_k SNPs follow a mixture distribution: $\pi_0 f_0 + (1 - \pi_0)f_1$, where f_0 is a distribution skewed to 0 and f_1 is a uniform distribution. Then we need to calculate the effective number of independent SNPs among those SNPs whose *p*-values follow uniform distribution. Denote this number as $M^{\max}_{k,\mathrm{eff}}$. Then the multiple testing corrected *p*-value is $p_{\max}^{M^{\max}_{k,\mathrm{eff}}}$. The rationale of this formula is as follows. Suppose we have $M^{\max}_{k,\mathrm{eff}}$ independent *p*-values, denoted by $\varrho_1, \ldots, \varrho M^{\max}_{k,\mathrm{eff}}$, which follow the Uniform distribution, then $P(\max_u \varrho_u \le \tau)=\prod_{u=1}^{M^{\max}_{k,\mathrm{eff}}} P(\varrho_u \le \tau) \le \tau^{M^{\max}_{k,\mathrm{eff}}}$. Due to limited SNPs around a gene-DHS pair and their strong correlation, it is difficult to estimate $M^{\max}_{k,\mathrm{eff}}$, and thus we use a conservative choice of $M^{\max}_{k,\mathrm{eff}}=1$.

These three *p*-values are further converted to *q*-values using the R package `qvalue` [Dabney and Storey (2015)] to account for multiple testing across the gene-DHS pairs, and we denote the *q*-values by $q_{\texttt{uncond}}$, $q_{\texttt{max}}$ and $q_{\texttt{min.corr}}$, respectively. As illustrated in Figure 3, the significant unconditional association of many gene-DHS pairs disappears after conditioning one of the local SNPs. The tables in Figure 3C provide a summary in terms of number of significant findings at *q*-value cutoff 0.1. Our method detects significant unconditional associations for 80 gene-DHS pairs (0.92% of pairs), while only 10 of them remain significant after conditioning on local SNPs. A previous study testing for correlation between RNA-seq and DNase-seq data in this dataset found ~0.7% of all gene-DHS pairs tested (3587 out of 4,678,275 pairs) showed significant correlation, after scaling the TReC for each data type to account for possible confounding factors [Degner et al. (2012)]. The small proportion of gene-DHS pairs with significant unconditional association can be explained by the small sample size and low read depth, and thus low statistical power of this dataset. We estimated that about 8.1% of gene-DHS pairs are associated without conditioning on local SNPs by estimating the non-null proportion of the *p*-values across all the gene-DHS pairs [Dabney and Storey (2015)].

We further examine several significant associations between RNA-seq and DNase-Seq while accounting for the effect of a common SNP. In this context, adjusted TReC refers to the residuals that are calculated from the BBLN model from each data type. For example, for the RNA-seq data, the adjusted TReC is determined as $T_R - \exp(X_R\hat{\beta}_R)$, where T_R is the RNA-seq read count for a particular gene, X_R is the associated covariate matrix of factors for the model that was fit, and $\hat{\beta}_R$ is the estimate for the regression coefficients pertaining to the RNA-seq data from the fitted BBLN model. We similarly calculate the residuals for the DNase-seq data.

One example involves the RNA-seq TReC of SLFN5 and the DNase-seq TReC of a DHS site near an intron approximately 1.5 kb from its transcription start site. SLFN5 has been shown to play a role in melanoma and renal cell carcinoma, and is known to be inducible by interferon-α [Mavrommatis et al. (2013)]. Ignoring any possible joint SNP effect, we find that the correlation between the DNase-seq TReC and SLFN5 RNA-seq TReC is significant (Figure 4A, $q_{\mathtt{uncond}} = 5.9 \times 10^{-10}$). However, after adjusting for the additive genetic effects, such as nearby SNP rs11080327 (Figure 4C), we find this correlation is no longer significant (Figure 4E, $q_{\mathtt{max}} = 1.0$), indicating that the observed correlation between the RNA-seq TReC and DNase-seq TReC is induced by shared genetic factors. We also observe a significant correlation between the RNA-seq TReC from gene EGR1 and the DNase-seq TReC for a DHS located upstream of the gene (Figure 4B, $q_{\mathtt{uncond}} = 7.3\times10^{-3}$). This correlation remains significant after adjusting for nearby SNPs, for example, rs7735367 (Figure 4D, F, $q_{\mathtt{max}} = 0.084$). In fact, both RNA-seq and DNase-seq data show very weak associations with the genotype of rs7735367 (Figure 4D). We also reran our analysis without PCs and, after *p*-value correction, we found that there were approximately 50% fewer significant results after our *p*-value correction compared to when the PCs were utilized.

We also compared our method to the much simpler approach of computing correlations between the TReC observed in each of the gene-DHS pairs considered by our model. To adjust for read depth, we transformed the TReC from each data type to Counts Per Million (CPM). DNase-seq CPM was computed such that the DHS TReC for a given individual was divided by the total DNAse-seq read count for that individual, multiplied by one million. RNA-seq CPM for a given individual was computed similarly. We then computed three types of correlations based on the computed CPMs for each of the gene-DHS pairs considered by our model: Spearman correlation between the RNA-seq and DNase-seq CPM, Pearson correlation between log(CPM + 1) transformed RNA-seq and DNase-seq CPM, and

$\sqrt{\mathrm{CPM}}$ transformed RNA-seq and DNase-seq CPM. We then perform a correlation test between the CPM from each data type to assess the significance of the association, and *p*-values across the gene-DHS pairs were converted to *q*-values. The results are given in Supplementary Figure S3, where we observed 8 significantly associated gene-DHS pairs using Spearman correlation, 14 for the Pearson correlation of log(CPM + 1) transformed counts and 8 for Pearson correlation of $\sqrt{\mathrm{CPM}}$ transformed counts at an FDR threshold of 0.1. We attribute the lower sensitivity of the simple approach relative to our BBLN model to loss of power due to transformation, and also the inability to account for additional possible confounders affecting the data.

For our ASReC data, we did not observe many significant results after applying our *p*-value correction procedure (Supplementary Table S3). This is due to the fact that we could not find many sites with coverage in both alleles at sufficient depth (Supplementary Figure S2), leading to only 567 DHS-gene pairs being evaluated. Of those evaluated, there was a median of 17 allele-specific read counts for the DNase-seq data and 46 for the RNA-seq data. This also resulted in few individuals being utilized during testing for a given site, as many samples did not have enough ASReC in both data types to be included in the model. As the cost of high throughput sequencing drops, we expect this to be less of an issue in the near future.

4. Discussion

We introduce a new method to model relationships across three types of data: gene expression, epigenetic features and genetic variants. We demonstrate the utility and power of our method to test for bivariate correlation between RNA-seq and DNase-seq data while adjusting for a possible shared genetic effect. Our simulation results show that there is relatively low power to detect weaker associations at smaller sample sizes, such as $n = 50$, which may explain the limited number of findings from our real data study with sample size 60. While this is a limitation for this dataset, in the near future we expect to see larger sample sizes as the cost of sequencing decreases.

The univariate form of our model, the Poisson-Log-Normal model, has been long utilized as a model to handle overdispersed counts and has been applied in the contexts of species abundance analysis [Bulmer (1974)], prediction of highway crash counts [Ma, Kockelman and Damien (2008)] and many others. For the TReC data, our BPLN model is a bivariate generalization of the Poisson Log-Normal model and a special case of the multivariate version introduced by Aitchison and Ho (1989). These methods have similarly been applied to contexts involving multivariate overdispersed count data, such as multivariate crash count data [Park and Lord (2007)] and network inference in microRNA-seq interaction networks [Gallopin et al. (2013)]. The advantage of this approach is the flexibility in specifying the correlation structure between the bivariate counts via Σ_1. In addition, overdispersion in the RNA-seq and DNase-seq TReC is modeled via variances σ_R and σ_C, respectively, where larger variance corresponds to larger overdispersion. Most importantly, both positive and negative correlations are allowed between the bivariate counts using this approach. However, the numerical integration that is required to evaluate the BPLN likelihood and derivatives increases the complexity of the estimation procedure, and may become unstable for lower sample sizes and lower signal levels. The BBLN (Bivariate Binomial-Logistic-Normal) model for the ASReC data also shares similar flexibilities and computational issues as the BPLN model.

One alternative to the BPLN is the bivariate negative binomial distribution introduced by Famoye (2010). This model is simply the product of two marginal negative binomial distributions corresponding to each of the two random variables, plus a multiplicative term with an additional parameter λ controlling the correlation of the two random variables. This approach also allows for either positive or negative correlations between the two variables, and evaluation of the likelihood and derivatives of this distribution does not require numerical integration. However, the maximization of the corresponding likelihood with respect to λ is difficult in practice because the plausible values of λ are bounded and such bounds are not known a priori. When the mean of each marginal distribution is not modeled by covariates, these bounds can be derived analytically. However, in the regression setting such bounds are difficult to determine. For ASReC, a model analogous to the bivariate negative binomial distribution is the Bivariate Beta Binomial Distribution [Danaher and Hardie (2005)] and it suffers from similar problems. Our model also shares some similarities with the generalized linear mixed model framework with heterogeneous variances. However, we do not share any fixed effects covariates or intercepts between data types, complicating the specification of the model; that is, each data type has distinct sets of covariates and dynamic ranges of signal (TReC for genes are typically larger than TReC for short windows tabulating DNA-seq TReC).

Despite the computational complexity of the BPLN and BBLN models, our implementation proved to be robust and computationally efficient relative to alternative approaches of numerical integration (ns^2 operations per likelihood evaluations, where s is the number of quadrature points). Our software implementation is freely available as an R package accessible at https://github.com/naimrashid/BASeG. In our implementation, testing of 874 trios in chromosome 21 took 1.5 hours. This time can be greatly reduced by setting more lenient convergence criteria, however, we chose more stringent settings for this particular study. Sampling-based integration methods such as Monte Carlo integration could have been used to evaluate the BPLN and BBLN, however, the inherent randomness in such approaches may pose problems during maximization. Fully Bayesian approaches are not computationally efficient for our applications.

Given the size of the observed read counts, especially for the RNA-seq data, a logarithmic transformation would be merited, and simple correlations can be computed. However, for certain features, such as the DHS sites that we consider in our manuscript, such a transformation may not be appropriate, as these sites accumulate relatively smaller counts. Features such as DHS sites, which are on the order of 100 bp in our manuscript, naturally tend to capture relatively fewer sequencing reads relative to larger features like gene bodies. In addition, shorter genes may exhibit smaller counts relative to larger genes. More importantly, a logarithmic transformation with our BPLN model would be efficient only if we are modeling *total* read counts, not *allele-specific* read counts, which tend to be much lower. For these reasons, we chose to develop a general model that utilizes the count data directly instead of modeling the data independent of any transformations.

Our current model conditions the distribution of the observed read counts in each data type jointly on a common SNP, implying that the SNP impacts both gene expression and DNAse-I hypersensitivity; that is, we are assessing the following causal model: DHS signal ←SNP→Gene expression. If the causal model is instead SNP→DHS signal→Gene expression, we would still observe association between DHS and expression. Conditioning on the common SNP in this scenario may reduce our power to detect correlation between data types, but would allow for the detection of a direct instead of indirect relation between DHS signal and gene expression. One may further compare this conditional independence model DHS signal ← SNP → Gene expression versus the following two causal models SNP → DHS signal → Gene expression or SNP → Gene expression → DHS signal. These tasks can be accomplished by simply comparing the likelihoods of these models or by a non-nested likelihood ratio test [Sun, Yu and Li (2007)]. Our approach provides the likelihood model for such a comparison, though we did not further make such comparisons due to limitations of the real data, for example, sample size and read depth.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

Aitchison J, Ho C-H. The multivariate Poisson-log normal distribution. Biometrika. 1989; 76:643–653.

Bulmer MG. On fitting the Poisson lognormal distribution to species-abundance data. Biometrics. 1974:101–110.

Cowper-Sal R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, Moore JH, Lupien M, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet. 2012; 44:1191–1198. [PubMed: 23001124]

Dabney A, Storey JD. qvalue: Q-value estimation for false discovery rate control. R package Version 1.38.0. 2015

Danaher PJ, Hardie BGS. Bacon with your eggs? Applications of a new bivariate beta-binomial distribution. Amer Statist. 2005; 59:282–286.

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. DNaseI sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–394. [PubMed: 22307276]

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. Nature. 2012; 489:101–108. [PubMed: 22955620]

Famoye F. On the bivariate negative binomial regression model. J Appl Stat. 2010; 37:969–981.

Fang F, Hodges E, Molaro A, Dean M, Hannon GJ, Smith AD. Genomic landscape of human allele-specific DNA methylation. Proc Natl Acad Sci USA. 2012; 109:7332–7337. [PubMed: 22523239]

Gallopin M, Rau A, Jaffrézic F, Chen L. A hierarchical Poisson log-normal model for network inference from rna sequencing data. PLoS ONE. 2013:8.

Hartzel J, Agresti A, Caffo B. Multinomial logit random effects models. Stat Model. 2001; 1:81–102.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching Ka, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson Ja, Crawford GE, Kellis M, Ren B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–12. [PubMed: 19295514]

Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet. 2003; 33(Suppl):245–254. [PubMed: 12610534]

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816–834. [PubMed: 21058334]

Liu Q, Pierce DA. A note on Gauss-Hermite quadrature. Biometrika. 1994; 81:624–629.

Ma J, Kockelman KM, Damien P. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Anal Prev. 2008; 40:964–975.

Mavrommatis E, Arslan AD, Sassano A, Hua Y, Kroczynska B, Platanias LC. Expression and regulatory effects of murine Schlafen (Slfn) genes in malignant melanoma and renal cell carcinoma. J Biol Chem. 2013; 288:33006–33015. [PubMed: 24089532]

McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010; 328:235–239. [PubMed: 20299549]

Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004; 74:765–769. [PubMed: 14997420]

Park E, Lord D. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transp Res Rec. 2007; 2019:1–6.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]

Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

Rashid NU, Sun W, Ibrahim JG. Supplement to "A statistical model to assess (allele-specific) associations between gene expression and epigenetic features using sequencing data". 2016; doi: 10.1214/16-AOAS973SUPP

Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al. AlleleSeq: Analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011:7.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Gräf S, Huss M, Keefe D, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011; 21:1757–1767. [PubMed: 21750106]

Sun W. A statistical framework for eQTL mapping using RNA-seq data. Biometrics. 2012; 68:1–11. [PubMed: 21838806]

Sun W, Yu T, Li K-C. Detection of eQTL modules mediated by activity levels of transcription factors. Bioinformatics. 2007; 23:2290–2297. [PubMed: 17599927]

Sun W, Liu Y, Crowley JJ, Chen TH, Zhou H, Chu H, Huang S, Kuan PF, Li Y, Miller D, Shaw G, Wu Y, Zhabotynsky V, McMillan L, Zou F, Sullivan PF, Pardo-Manuel de Villena F. IsoDOT detects differential RNA-isoform usage with respect to a categorical or continuous covariate with high sensitivity and specificity. J Amer Statist Assoc. 2015; 110:975–986.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]

Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]

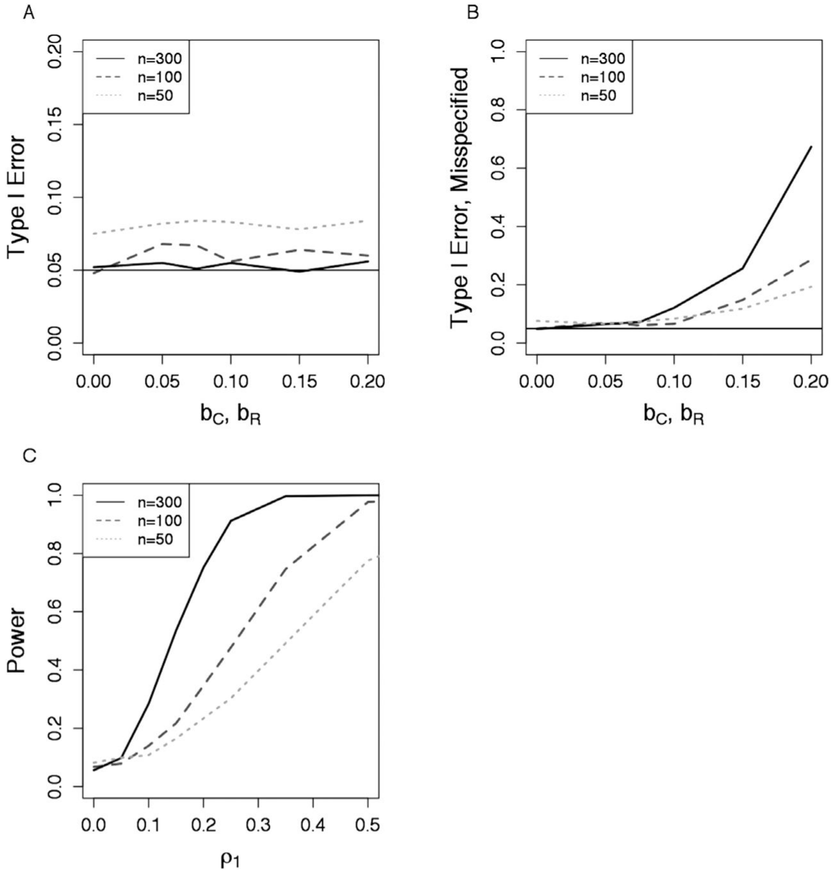


Fig. 1.

Simulation results for the BPLN (Bi-variate Poisson Log Normal) model. (A) Type I error in testing for $\rho_1 = 0$ given b_C and b_R. (B) Type I error in testing for $\rho_1 = 0$ under the assumption of $b_C = 0$ and $b_R = 0$ while the true values of b_C and b_R vary from 0 to 0.2. (C) Power in testing for $\rho_1 = 0$ with different sample sizes, given $b_C = 0$ and $b_R = 0$.

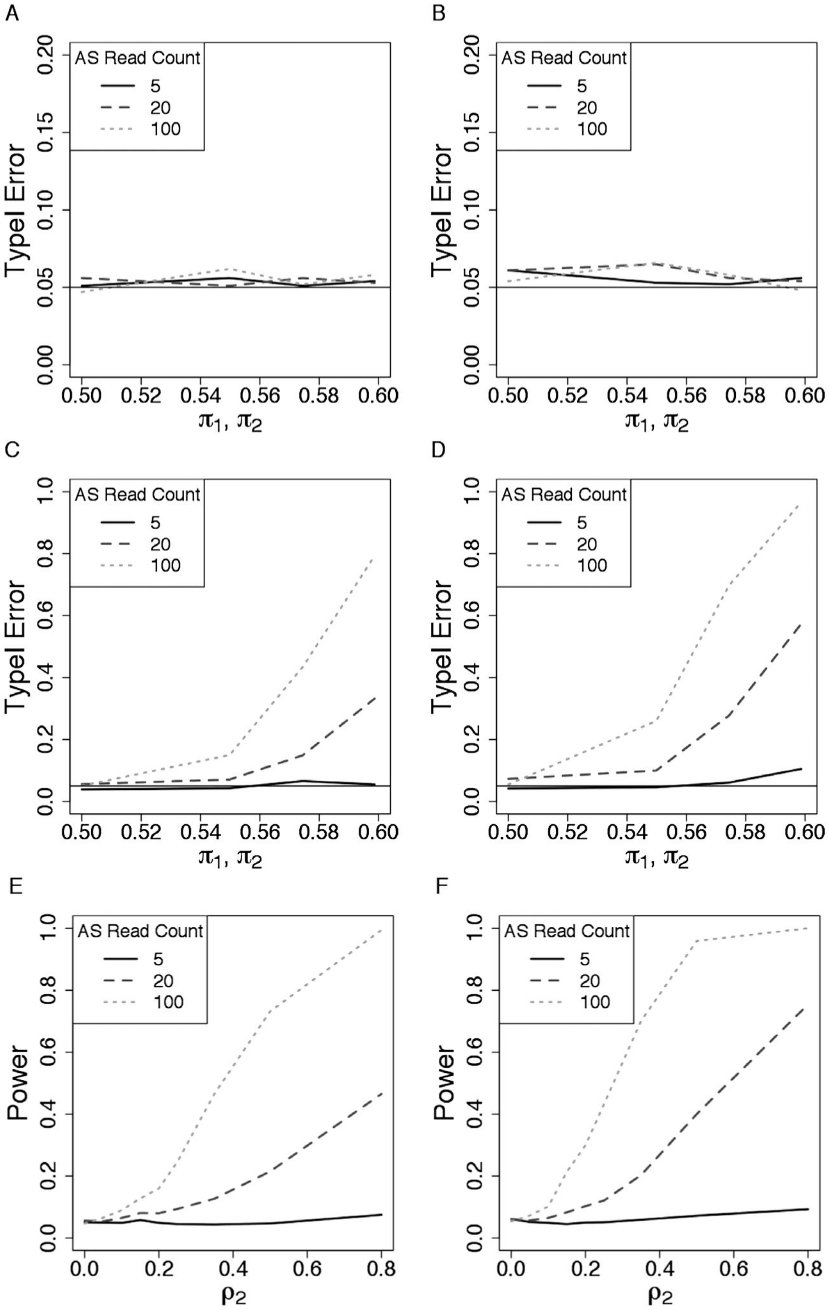
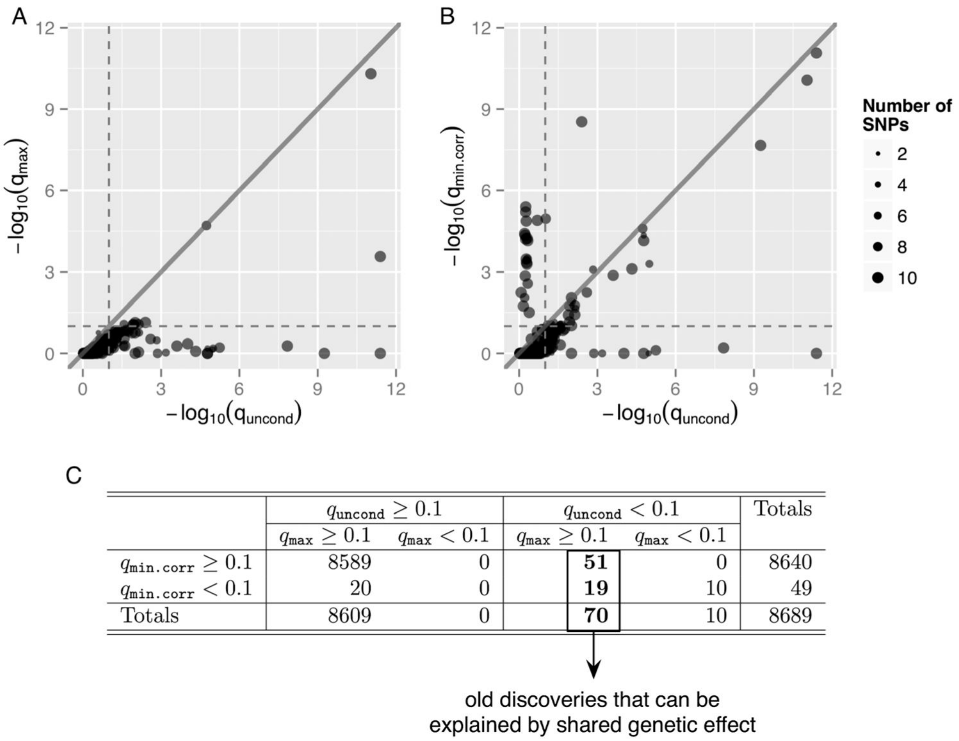


Fig. 2.

Simulation results for BBLN (Bi-variate Binomial Logistic Normal) model. (A) and (B): Type I error in testing for $\rho_2 = 0$ while accounting for genetic effects when $n = 50$ (A) or $n = 100$ (B). (C) and (D): Type I error in testing for $\rho_2 = 0$ while ignoring genetic effect (i.e., assuming $\pi_1 = 0.5$ and $\pi_2 = 0.5$) when $n = 50$ (C) or $n = 100$ (D). (E) and (F): Power in testing for $\rho_2 = 0$ when $n = 50$ (E) or $n = 100$ (F).



	$q_{\text{uncond}} \geq 0.1$		$q_{\text{uncond}} < 0.1$		Totals
	$q_{\max} \geq 0.1$	$q_{\max} < 0.1$	$q_{\max} \geq 0.1$	$q_{\max} < 0.1$	
$q_{\text{min.corr}} \geq 0.1$	8589	0	**51**	0	8640
$q_{\text{min.corr}} < 0.1$	20	0	**19**	10	49
Totals	8609	0	**70**	10	8689

old discoveries that can be explained by shared genetic effect

Fig. 3.

Panels (A) and (B) show the comparison between unconditional *q*-value (q_{uncond}) vs. (A) maximum (conditional) *q*-value ($q_{\max}$) and (B) multiple testing corrected minimum (conditional) *q*-value ($q_{\text{min.corr}}$). Note that multiple testing corrected minimum *p*-value $p_{\text{min.corr}}$ account for multiple testing across multiple SNPs of each gene-DHS pair, while calculation of *q*-value from *p*-values accounts for multiple testing across multiple gene-DHS pairs. The size of each point represents the number of conditioning SNPs for each gene-DHS pair, and it is truncated at 10. The dashed lines indicate q-value threshold 0.1 and the solid line is the diagonal line of $y = x$. Panel (C) demonstrates our findings by tables.

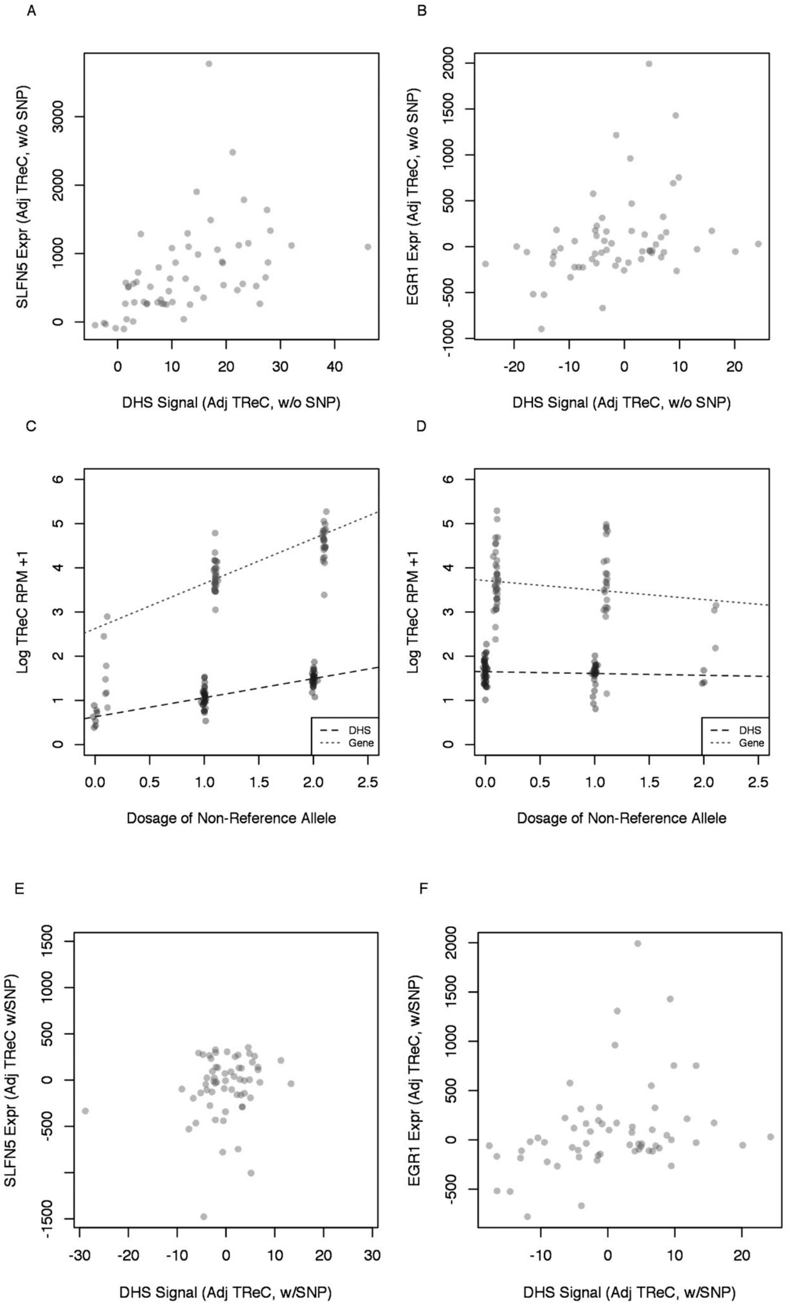


Fig. 4.

Illustrations of significant interactions between the TReC of select gene-DHS pairs, as well as the modulatory effects of nearby SNPs. In this context, adjusted TReC refers to the residuals that are calculated from the BBLN model from each data type. (A) Association between the adjusted TReC of SLFN5 expression and a DHS in intron 1 of SLF5, and (B) the adjusted TReC of EGR1 expression and a DHS in the upstream region of EGR1, after accounting for sequencing depth and PCs in the BBLN model. (C) The genotype of SNP rs11080327 is associated with both the SLFN5 gene expression and the nearby DHS. (D) The genotype of SNP rs7735367 is weakly associated with both the EGR1 gene expression and a nearby DHS. (E) The adjusted TReC of the SLFN5 expression and the nearby DHS is not associated after accounting for sequencing depth, PCs and SNP effect of rs11080327 in the BBLN model. (F) The adjusted TReC of the EGR1 expression and the nearby DHS are still associated after adjusting for sequencing depth, PCs and SNP effect of rs11080327.