Published in final edited form as:

Stat Med. 2019 April 15; 38(8): 1357–1373. doi:10.1002/sim.8040.

A likelihood-based approach to transcriptome association analysis

Jing Qian¹, Evan Ray², Regina L. Brecha², Muredach P. Reilly³, Andrea S. Foulkes^{*,2}

¹Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, School of Public Health and Health Sciences, MA, USA

²Department of Mathematics and Statistics, Mount Holyoke College, MA, USA

³Department of Medicine, Columbia University, College of Physicians and Surgeons, NY, USA

Abstract

Elucidating the mechanistic underpinnings of genetic associations with complex traits requires formally characterizing and testing associated cell and tissue-specific expression profiles. New opportunities exist to bolster this investigation with the growing numbers of large publicly available -omics level data resources. Herein we describe a fully likelihood-based strategy to leveraging external resources in the setting that expression profiles are partially or fully unobserved in a genetic association study. A general framework is presented to accommodate multiple data types, and strategies for implementation using existing software packages are described. The method is applied to an investigation of the genetics of evoked inflammatory response in cardiovascular disease research. Simulation studies suggest appropriate type-1 error control and power gains compared to single regression imputation, the most commonly applied practice in this setting.

Keywords

Transcriptome association analysis; EM algorithm; protein coding genes; regulatory elements; RNAseq

1 | INTRODUCTION

Large-scale genome-wide association (GWA) studies have contributed substantially to knowledge on the genetic underpinnings of complex disease traits, including the identification of a vast array of single nucleotide polymorphisms (SNPs), protein-coding genes and regulatory elements, that associate with clinically relevant outcomes. The biological mechanisms underlying these genetic associations, particularly the relationships among susceptibility loci, expression profiles in disease relevant tissues, and complex traits, however, remains largely uncharacterized. Emerging cell and tissue-specific transcriptome-wide association (TWA) data will likely play a decisive role in elucidating these

*Corresponding author name, Corresponding address. afoulkes@mtholyoke.edu.
Present Address 50 College Street, South Hadley, MA 01075

mechanisms. Moreover, the availability of large and comprehensive reference data panels, including sequencing data and genome-wide cell and tissue specific expression profiles, can serve as highly informative and tractable resources to enhance this investigation. Herein we propose a general approach to leveraging such reference multi-omic data in the context of genetic association studies for which transcriptome data are partially or fully unobserved.

The setting motivating this research is one in which two independent data resources are available: (1) GWA data, composed of complete genotype and trait information on each individual, in addition to the possible availability of expression profiles on a subset of cell types, and a subset of individuals; and (2) TWA data, composed of genotype and complete expression profiles, as measured by RNAseq, on multiple genes, several cell and tissue types and multiple isoforms, for an independent set of individuals. Analysis is based on all SNPs that are shared between the two data resources. Typically transcriptome data do not include a measured trait, although emerging resources for which trait data will be available, can additionally be incorporated into the proposed framework. The analytical goal is to leverage the transcriptome data in analysis of the GWA data to inform the mechanisms underlying gene-trait associations. For example, a GWA analysis may suggest that SNPs close to a particular gene are associated with a given trait. By leveraging transcriptome data, we can additionally unravel whether this association is mediated by expression of the same gene, or another gene or regulatory element.

The specific data example motivating this research is derived from the Genetics of Evoked Responses to Niacin and Endotoxemia (GENE) study, an NIH-sponsored investigation of the genomics of inflammatory and metabolic responses during low-grade endotoxemia^{1,2,3}. In this study, n=294 individuals were genotyped at baseline and monitored during a 48 hour hospital visit after an endotoxin challenge. Multiple clinical variables and five plasma biomarkers – tumor necrosis factor alpha (TNF α), interleukin-6 (IL-6), interleukin-1 receptor agonist (IL-1RA), Serum Amyloid A (SAA) and high-sensitivity C-reactive protein (CRP) – were recorded repeatedly over time in increments of 1 to 12 hours. In a recent study, we investigated the genetic factors associated with IL-6 and CRP levels, and identified several genes and long non-coding RNAs that associate with the change in these biomarkers from baseline to peak, as well as the non-linear biomarker trajectories over time⁴. The next crucial step of this investigation is to decipher the mechanistic relationships between these genomic classes and induced biomarker responses through targeted transcriptome analysis.

Expression data are generally unavailable or limited in GWA settings. For example, in the GENE study, whole-genome RNAseq data are available on a subset of individuals with extreme phenotypes and exclusively on one of many relevant blood cell types. Drawing inference in this setting is possible, however, through integration of reference transcriptome data. Several large reference panels that include genome-wide SNP data, as well as whole-genome RNAseq data on multiple cell and tissue types, are now publicly available offering new potential for integrated analysis platforms. Three well-described and comprehensive TWA resources are: (1) the Genotype Tissue Expression (GTEx) data V6p release, composed of RNAseq data on 53 tissues and whole-blood for a total of 449 donors⁵; (2) the European Bioinformatics Institute Geuvadis data, composed of RNAseq data from lymphoblastoid cell lines of 465 individuals from the 1000 Genomes project⁶; and (3) the

Depression Genes and Networks study (DGN) sequencing data composed of whole-blood RNAseq for 922 individuals⁷. In this manuscript we describe and evaluate a statistical strategy for integrating these available resources in the analysis of data arising from genetic association studies with limited to no observed transcriptome data to uncover mechanistic relationships.

A few notable publications describe methods for integrating reference transcriptome data for association studies^{8,9,10,11}. For example, Gamazon et al. propose a two-stage strategy, termed PrediXcan, that first singly imputes cell-specific gene expression using genetic profiles and publicly available reference transcriptome data resource, and second characterizes association between imputed expression and a measured trait⁸. The first-stage prediction algorithm relates genotype to expression using the elastic net¹², demonstrating similar performance to the least absolute shrinkage and selection operator (LASSO)¹³. Predicted cell-specific genetically regulated expression (GReX) is then treated as known and used in a generalized linear model to evaluate association with a binary outcome. The approach of Gusev et al. similarly involves unit imputation while alternative predictive models are used¹⁰, including a mixed effects model with random SNP-level effects and a Bayesian sparse linear mixed model (BSLMM)¹⁴. Perhaps the most notable statistical shortcoming of these approaches, which can be viewed as special cases of regression imputation, is the lack of consideration of the uncertainty in the prediction of genetically regulated expression. Additionally, well-established statistical research indicates that two-stage imputation approaches are generally less powerful than fully likelihood-based approaches that integrate prediction of unobserved data and association analysis^{15,16}.

As a robust alternative that addresses the limitations of current practice, we propose combining the two data resources, the GWA and TWA data, and framing the analysis in terms of a missing data problem. Specifically, we assume RNAseq data are only partially observed. We begin by presenting a unified strategy based on maximizing the observed data likelihood which is straightforward in the context of normalized expression data. The expectation-maximization (EM) algorithm is an established approach to estimation in the presence of missing data¹⁷ in the case that solving for the observed data likelihood is not tenable, and in Section 2 we develop a simple EM-based strategy for this setting. We then discuss opportunities for practical and efficient implementation, and describe strategies for statistical inference. Our fully likelihood-based approach appropriately accounts for the error in prediction (not addressed in current methods) by integrating the prediction of unobserved data and association analysis. Simulation studies and application to the GENE study data are also described to fully characterize method performance and compare to the more common application of regression imputation for this setting (Sections 3 and 4). We conclude with a discussion in Section 5.

2 | METHODS

2.1 | A unified likelihood-based approach for normalized expression data

We first define a generalized linear model of association between expression and a univariate biomarker response, and second propose a similar modeling framework for the relationship between genotype and expression. Let $\mathbf{y}$ be an $n \times 1$ vector representing a univariate

biomarker response on n individuals, and consider a generalized linear model of the form $g(E[y]) = \mathbf{W}\boldsymbol{\gamma}_0 + \mathbf{x}\boldsymbol{\gamma}_1$, where covariates are included in the design matrix $\mathbf{W}$, cell-specific expression (for a single cell type) is represented by the $n \times 1$ vector $\mathbf{X}$ and $g(\cdot)$ is a canonical link function. A general form of the conditional probability density function for $\mathbf{y}$, given $\mathbf{W}$ and $\mathbf{X}$, assuming an exponential family distribution, is given by ¹⁸:

$$f_{\Omega}(\mathbf{y}|\mathbf{W}, \mathbf{x}) = \prod_{i=1}^n \exp\{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\}, \quad (1)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specified functions, $\Omega = (\theta, \phi)$, θ depends on the linear predictor and ϕ is a scale parameter. In our example of Section 4, the biomarker is a quantitative measure of immune modulation, and so we let $g(\cdot)$ be the identity link and assume the conditional probability density function is given by a multivariate normal distribution. In this case, $\theta_i = \mathbf{w}_i^T \boldsymbol{\gamma}_0 + \mathbf{x}_i^T \boldsymbol{\gamma}_1$, $\phi = \sigma^2$, $i = 1, \dots, n$,

$$a(\phi) = \phi, \quad b(\theta_i) = \theta_i^2/2 \quad \text{and} \quad c(y_i, \phi) = -\frac{1}{2}\{y_i^2/\phi + \log(2\pi\phi)\}. \quad (2)$$

Gene expression as measured by RNAseq are count data and can also be modeled with an exponential family distribution of the form of Equation 1. Let $h_{\Theta}(\mathbf{x}|\mathbf{Z})$ represent the product of probability distribution functions for expression counts conditional on genotype $\mathbf{Z}$, where $\mathbf{Z}$ is defined explicitly. To begin we will apply LD pruning, as described in ¹⁹, in order to avoid high degrees of collinearity and an unidentifiable model. Alternative modeling strategies to account for moderate to large degrees of LD are described in Section 5. Untransformed RNAseq data are often assumed to follow a negative binomial distribution to account for over dispersion ^{20,21}, although normalized gene expression data, specifically Reads Per Kilobase of transcript, per Million mapped reads (RPKM), are typically modeled using a Gaussian model ⁸. In this manuscript, we begin by describing a likelihood based strategy based on a Gaussian model for expression data in Section 2.1. Secondly, we consider inference using the more general EM framework, which could be used with a wide variety of models; in Section 2.2 we develop an EM-type algorithm arising from a discrete exponential family distribution.

In our setting, the complete data are the combination of the primary data – composed of genotype and biomarker data – and the external data resource – composed of genotype and complete RNAseq data for an independent sample. Specifically we can think of the RNAseq data, represented by $\mathbf{x}$, as unobserved in the primary data. In this case, the complete data are given by $\mathbf{y}$ and $\mathbf{x}$, and the complete data likelihood is written:

$$L^c(\Phi|\mathbf{y}, \mathbf{W}, \mathbf{Z}, \mathbf{x}) = Pr(\mathbf{y}, \mathbf{x}|\mathbf{W}, \mathbf{Z}, \Phi) = f_{\Omega}(\mathbf{y}|\mathbf{W}, \mathbf{x})h_{\Theta}(\mathbf{x}|\mathbf{Z}), \quad (3)$$

where $\Phi = (\Omega, \Theta)$ and $f_{\Omega}(\mathbf{y}|\mathbf{W}, \mathbf{x})$ and $h_{\Theta}(\mathbf{x}|\mathbf{Z})$ are the conditional probability density functions of $\mathbf{y}$ and $\mathbf{x}$, respectively. We further let the complete data log likelihood be denoted $l_{\Phi}^c = \log L^c(\Phi|\mathbf{y}, \mathbf{W}, \mathbf{Z}, \mathbf{x})$.

In order to estimate the elements of Φ we aim to maximize the observed data log likelihood, denoted l_{Φ}^{obs} , where the observed data in this setting are given by $\{(y_i, \mathbf{z}_i, \mathbf{w}_i)_{i=1}^n; (x_j, \mathbf{z}_j)_{j=1}^m\}$.

The observed data likelihood is derived by integrating $L^c(\Phi | \mathbf{y}, \mathbf{W}, \mathbf{Z}, \mathbf{x})$ over the missing information. In the special case that $h_{\Theta}(\mathbf{x} | \mathbf{Z})$ is given by a multivariate normal distribution and the model relating expression to the biomarker is linear with normally-distributed errors, it is straightforward to show that the observed data log likelihood is given by:

$$l_{\Phi}^{\text{obs}} = \sum_{i=1}^n \left[-\frac{1}{2} \log(\sigma_{\delta}^2 \gamma_1^2 + \sigma_{\epsilon}^2) - \frac{(y_i - \mathbf{w}_i^T \gamma_0 - \mathbf{z}_i^T \beta \gamma_1)^2}{2(\sigma_{\delta}^2 \gamma_1^2 + \sigma_{\epsilon}^2)} \right] + \sum_{j=1}^m \left[-\log \sigma_{\delta} - \frac{(x_j - \mathbf{z}_j^T \beta)^2}{2\sigma_{\delta}^2} \right], \quad (4)$$

where $Y | \mathbf{w}, x \sim N(\mathbf{w}^T \gamma_0 + x \gamma_1, \sigma_{\epsilon})$ and $X | \mathbf{z} \sim N(\mathbf{z}^T \beta, \sigma_{\delta})$. As the corresponding estimating equations are difficult to solve, a numeric technique, such as the Newton-Raphson or Gauss-Legendre quadrature algorithm, can be applied to arrive at parameter estimates.

Corresponding standard errors for the parameter estimates are obtained using the inverse of the observed fisher information matrix (see Appendix A).

2.2 | An EM approach for expression data arising from a discrete exponential family distribution

The derivation of the observed data likelihood in Section 2.1 was facilitated by the use of normal models for the relationships between genotype and expression and expression and the trait. In general, deriving the observed data likelihood is not straight-forward. However, in the case of a discrete distribution function $h_{\Theta}(\mathbf{x} | \mathbf{Z})$, the conditional expectation of the complete data log likelihood is given by a weighted sum over possible gene expression patterns, where the weights correspond to posterior probabilities given the observed genotypes. Formally, we have:

$$E(l_{\Phi}^c | \mathbf{y}, \mathbf{W}, \mathbf{Z}) = \sum_{i=1}^n \sum_{x_i \in \chi_i} \pi_i(\Phi; x_i) l_{\Phi, i}^c, \quad (5)$$

where $l_{\Phi, i}^c$ is the contribution of the i th individual to the complete data log likelihood, and the weights are given by:

$$\pi_i(\Phi; x_i) = Pr(x_i | y_i, \mathbf{w}_i, \mathbf{z}_i) = \frac{f_{\Omega}(y_i | \mathbf{w}_i, x_i) h_{\Theta}(x_i | \mathbf{z}_i)}{\sum_{x_i \in \chi_i} f_{\Omega}(y_i | \mathbf{w}_i, x_i) h_{\Theta}(x_i | \mathbf{z}_i)}, \quad (6)$$

and χ_i is the set of all expression levels that are consistent with the observed genotypes for individual i .

For each iteration of the EM algorithm, in the E-step we substitute the current estimates of Φ into the calculation of the posterior expression probabilities, to arrive at the estimates

$\hat{\pi}_{\mathbf{x},i}^{(t)} = \pi_i(\hat{\Phi}^{(t)}; x_i)$. Then in the M-step, we update the parameter estimates by maximizing the conditional expectation. Maximization at the $(t+1)$ th iteration proceeds by taking the partial derivative of the expected complete data log-likelihood given by Equation 5 with respect to each parameter, setting this equal to 0 and solving for the corresponding parameter. That is, at the $(t+1)$ th iteration we aim to find the roots of the partial derivatives with respect to each parameter in $\Phi = (\phi_1, \dots, \phi_K)$. Formally, at the $(t+1)$ th iteration of the M-step, we aim to find the roots of the following system of equations for $k = 1, \dots, K$:

$$\frac{\partial}{\partial \phi_k} E(l_{\Phi}^c | \mathbf{y}, \mathbf{W}, \mathbf{Z}) = \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \hat{\pi}_{x_i}^{(t)} U_i(\phi_k; \Phi), \quad (7)$$

where $U_i(\phi_k; \Phi) = \partial l_{\Phi,i}^c / \partial \phi_k$ is the i th individual's contribution to the k th element of the score function for the complete data log likelihood (provided in Appendix B). Inference in this setting requires estimation of the covariance matrix, given by the inverse of the observed information matrix, which we approximate with the empirical observed information^{22,23}. That is, we estimate the variance of the model parameters as $\mathbf{I}^{-1}$ evaluated at the final iteration of the EM algorithm after a convergence criterion is met (see Appendix B), where:

$$\begin{aligned} \mathbf{I} = & \sum_{i=1}^n E_{\Phi}(-\partial^2 l_{\Phi,i}^c / \partial \Phi^2 | \mathbf{y}, \mathbf{W}, \mathbf{Z}) - \sum_{i=1}^n E_{\Phi}(U_i(\Phi) U_i(\Phi)^T | \mathbf{y}, \mathbf{W}, \mathbf{Z}) \\ & - \sum_{i \neq j} E_{\Phi}(U_i(\Phi) | \mathbf{y}, \mathbf{W}, \mathbf{Z}) E_{\Phi}(U_j(\Phi) | \mathbf{y}, \mathbf{W}, \mathbf{Z})^T. \end{aligned} \quad (8)$$

The EM approach can be applied in general to exponential family distributions. For example, under the negative binomial model, letting $\Gamma(n) = (n-1)!$ be the usual Gamma function, $\mu_i = \exp(\mathbf{z}_i^T \boldsymbol{\beta})$, $\theta_i = \log[(\mu_i/\alpha)/(1 + \mu_i/\alpha)]$, $\phi = \alpha$,

$$a(\phi) = 1; \quad b(\theta_i) = -\alpha \log(1 - \exp(\theta_i)) \quad \text{and} \quad c(x_i, \phi) = \log\left(\frac{\Gamma(x_i + \phi)}{\Gamma(x_i + 1)\Gamma(\phi)}\right). \quad (9)$$

where $\mathbf{z}_i$ is the vector of observed SNPs for individual $i = 1, \dots, n$ and $\boldsymbol{\Theta} = (\alpha, \boldsymbol{\beta})$. In the context of RNAseq data and a univariate real-valued outcome, the model specifications of Equations 2 and 9 are generally appropriate. In this case, the i th contribution to the complete data log likelihood is given by:

$$\begin{aligned} l_{i,\Phi}^c = & \frac{1}{2} \left\{ y_i (\mathbf{w}_i^T \boldsymbol{\gamma}_0 + x_i \gamma_1) - (\mathbf{w}_i^T \boldsymbol{\gamma}_0 + x_i \gamma_1)^2 / 2 \right\} - \frac{1}{2} \left\{ y_i^2 / \sigma^2 + \log(2\pi\sigma^2) \right\} + \alpha \log(\alpha) \\ & + x_i \log(\mu_i) - (\alpha + x_i) \log(\alpha + \mu_i) + \log \Gamma(x_i + \alpha) - \log \Gamma(x_i + 1) - \log \Gamma(\alpha). \end{aligned}$$

At the M-step, we find the roots of Equation 7, based on the corresponding elements of the score function.

Arriving at closed form solutions for γ_0 , γ_1 and σ^2 at each iteration is straightforward. In fact, solving for γ_0 and γ_1 results in the familiar weighted least squares solutions, as detailed in Section 2.2. In general, closed form solutions may not exist, as is the case for β and α , so an alternative numeric technique, as described for the Gaussian setting, can be applied to identify roots to the corresponding weighted score functions. As it is not straightforward to maximize the expectation simultaneously with respect to the mean and scale parameters, we apply the expectation conditional maximization approach of ²⁴ in which we condition on current scale parameter estimates to update the mean parameter estimates, and then condition on mean parameters to update scale parameters. A bias adjustment is also made for σ at each iteration of the algorithm. In the next section we consider computationally efficient strategies for estimation in this setting, taking advantage of existing theory and functionalities.

Practical implementation of the EM strategy—In practice, at each iteration of the EM algorithm, we are maximizing a weighted likelihood function where the data for each individual are expanded to a dimension that equals the number of possible values of the unobserved data. To see this, first let $\tilde{y} = y_i \mathbf{1}_{m_i}$ and $\tilde{\mathbf{W}}_i = \mathbf{w}_i \otimes \mathbf{1}_{m_i}$, where $\mathbf{1}_{m_i}$ is a $m_i \times 1$ vector of 1's, m_i is the number of possible values of x_i for individual i and $\otimes$ represents kronecker product. Additionally define $\tilde{x}_i$ as an $m_i \times 1$ vector with elements equal to the values of x_i that are consistent with the observed genotype data and let $\hat{\mathbf{p}}_i^{(t)}$ be a vector of weights with elements equal to the set of posterior probabilities, $\hat{\pi}_{\mathbf{x},i}^{(t)}$. Finally, the concatenated versions of $\tilde{y}_i$, $\tilde{\mathbf{W}}_i$ and $\tilde{x}_i$ across all individuals are denoted $\tilde{\mathbf{y}}$, $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{x}}$, respectively. Letting j index the combination of individual and possible values of unobserved data, $j = 1, \dots, N$ where $N = \sum_{i=1}^n m_i$, we can now write Equation 5 as:

$$E(l_{\Phi}^c | \mathbf{y}, \mathbf{W}, \mathbf{Z}) = \sum_{j=1}^N \hat{p}_j^{(t)} [\log f_{\Omega}(\tilde{\mathbf{y}}_j | \tilde{\mathbf{w}}_j, \tilde{x}_j) + \log h_{\Theta}(\tilde{x}_j | \tilde{\mathbf{z}}_j)]. \quad (10)$$

For example, in the case of γ_1 , the primary parameter of interest, the EM algorithm involves calculating:

$$\frac{\partial}{\partial \gamma_1} E(l_{\Phi}^c | \mathbf{y}, \mathbf{W}, \mathbf{Z}) = \sum_{j=1}^N \hat{p}_j^{(t)} [\tilde{x}_j (\tilde{y}_j - \tilde{\mathbf{w}}_j^T \gamma_0 - \tilde{x}_j \gamma_1) / \sigma^2], \quad (11)$$

and setting this equal to 0, yielding the familiar weighted least squares estimator:

$$\hat{\gamma}_1^{(t+1)} = (\tilde{\mathbf{x}}^T \hat{\mathbf{P}}^{(t)} \tilde{\mathbf{x}})^{-1} \tilde{\mathbf{x}}^T \hat{\mathbf{P}}^{(t)} (\tilde{\mathbf{y}} - \tilde{\mathbf{W}} \gamma_0), \quad (12)$$

where $\hat{\mathbf{P}}^{(t)} = \text{diag}(\hat{p}_j^{(t)})$.

Based on this result, it is judicious to consider application of standard generalized linear model fitting procedures that employ weights in the maximization step of the algorithm

described in Section 2.2. In order to assess the appropriateness of this strategy, we first note that the log likelihood for a generalized linear model with prior weights, as described in ²⁵, (Chapter 7, page 183) is given by:

$$\log f_{\Omega}(y_i | \mathbf{w}_i, x_i) = A_i(y_i \theta - b(\theta)) / a(\phi) + c(y_i, \phi / A_i), \quad (13)$$

where A_i are known prior weights. Notably, this formulation differs from the weights introduced in ¹⁸ (Chapter 2, page 29) for the normal distribution, in which Equation 1 is modified by setting $a(\phi)$ equal to ϕ / A_i but weights are not introduced into $c(y, \phi)$. We focus on the model given by Equation 13 as it is the basis for the prior weight specification for the `glm()` function in R ²⁵.

In order to see the relationship of this to our setting, note that the score of γ_1 , can be written:

$$\frac{\partial}{\partial \gamma_1} E(l_i^c | \mathbf{y}, \mathbf{W}, \mathbf{Z}) = \sum_{j=1}^N \hat{p}_{ij}^{(t)} \frac{\partial}{\partial \gamma_1} [A_i(y_i \theta_i - b(\theta_i)) / a(\phi)], \quad (14)$$

as $\log h_{\Theta}(\mathbf{x} | \mathbf{Z})$ and $c(y, \phi / A_i)$ do not depend on γ_1 . Setting this equation equal to 0 and solving for γ_1 is equivalent to maximizing Equation 13 with respect to γ_1 . This is similarly true for γ_0 and β . However, in terms of the scale parameter, ϕ , we need that:

$$\sum_{j=1}^N \hat{p}_j^{(t)} \frac{\partial}{\partial \phi} c(y, \phi) = \sum_{j=1}^N \frac{\partial}{\partial \phi} c(y, \phi / \hat{p}_j^{(t)}). \quad (15)$$

For the normal and negative binomial models this does not generally hold; however, it can be shown that a simple transformation, $\hat{\phi}^{(t+1)} = \hat{\phi}_p^{(t+1)} * N/n$ where $\hat{\phi}_p^{(t+1)}$ is the estimated variance component based on specifying weights within the `glm()` function in R (see Appendix C). A simple transformation is not as straightforward for the negative binomial scale parameter (see Appendix C). Moreover, while the documentation for the `glm.nb()` function in R to fit a negative binomial regression similarly references the weighted regression approach of ²⁵ and applies an iterative procedure that uses the `glm()` function, the likelihood defined in the R code is of the form given in our setting (Equation 10).

3 | SIMULATION STUDIES

Simulations are performed to evaluate type 1 error, power and convergence of the unified likelihood-based approach for normalized expression data (Section 3.1), and the EM strategy for count data (Section 3.2). Both approaches are compared to a standard application of two-stage regression imputation.

3.1 | Unified approach to normalized expression data

To begin we investigate bias by simulating normalized data based on the modeling results from the example provided in Table 5 of Section 4. In this study, we sample with replacement from the observed genotypes across 20 SNPs within Capping Protein Regulator And Myosin 1 Linker 1 (CARMIL1, previously LRRC16A) that are present in both the

GENE study and DGN data; these resampled observations serve as the independent data for the simulation. The minor allele frequencies of these SNPs range from 0.11 to 0.45 in the DGN data. LD pruning based on a threshold of 0.20 was applied to arrive at this subset of 20 independent signal SNPs. Parameter values for the simulations are $\gamma_0 = -0.435$, $\gamma_1 = 4.650$, $\sigma_\delta = 0.651$, $\sigma_\epsilon = 1.195$. The 20 SNP coefficients ranged from -0.915 to 0.675 and SNPs are assumed to influence gene expression according to an additive genetic model and an additive model of association.

The distributions of bias across model parameters based on 1000 simulations for a range of sample sizes are illustrated in Figure 1. The center panel most closely reflects the observed sample sizes of our data example in Section 4. In all cases, bias is appropriately centered around zero and the variability is similar between the unified and regression imputation strategies. While decreasing the sample size of the GWA data set (left-hand panel) increases the variability in the parameter estimates, the impact of reducing the TWA sample size is larger (right-hand panel) with greater bias and variability for both the unified and single regression imputation approaches. Corresponding numerical summaries are provided in Table 1.

Secondly we estimate type-1 error and power for a range of expression level effect sizes (γ/σ_ϵ). Here we generate 10,000 simulations per condition based on the sample sizes from our data example in Section 4 – nGWA=173 and nTWA=906 for the GENE and DGN data, respectively – after removing observations with missing genotypes. The observed data genotypes are used as input and again parameter values for the simulations are based on the results of our real data application (Table 5). The simulation results are illustrated in the left hand panel of Figure 2. Estimated type-1 error rates are 0.0539 and 0.0491 for the unified likelihood-based and regression imputation strategies, respectively. As expected power increases with increasing effect size and the unified approach is consistently slightly more powerful. Both approaches have approximately 80% power for a moderate effect size of 0.30.

We note that these simulations assume correct model specification and as the approach is likelihood-based, we expect it to be sensitive to model mis-specification. To evaluate this we first consider the case that y is incorrectly assumed to be normally distributed when in fact the errors arise from a log normal distribution. In this case, the type-1 error rates remained controlled at the nominal 0.05 level for both the unified and regression imputation approaches – for the unified approach, estimated errors based on 10,000 simulations are 0.0527, 0.0505 and 0.0494 for shift parameters (log mean) of 1, 2 and 3, respectively. On the other hand, if the expression data follow a log normal distribution and are incorrectly assumed to follow a normal distribution, then the type-1 error rate based on 10,000 simulations is estimated to reach 0.116 for a shift parameter of 3. This result is consistent with our expectation as correctly specifying the model is critical in likelihood-based strategies. We additionally note that the divergence from normality is readily apparent in these simulations by visual inspection of histograms and this level of extreme mis-specification can be avoided in practice.

Finally, we take a closer look at the the power differential between the unified and regression imputation strategies. Under the same simulation conditions just described and based on 10,000 simulations per condition, we estimated power for a range of TWA sample sizes. This is illustrated in the center panel of Figure 2 and suggests that the TWA sample size is not highly impactful. The unified strategy performs consistently slightly better with respect to power with a greater difference seen with a smaller TWA sample size. Interval estimates for the difference in proportions are illustrated in the right-hand panel of Figure 2 and in all cases the results of the simulation study provide statistically significant evidence that there is a non-zero difference in power for these methods, with an estimated difference of between 1.3 and 2.1%. We also considered inflating the GWA sample size by a factor of 2 in which case the power differential diminishes substantially, as shown by the darker lines in the center and right-hand panels of Figure 2.

3.2 | EM approach for count data

In order to evaluate the EM approach, data are generated using parameter values for the negative binomial model given by $\gamma_0 = 2.0$, $\sigma^2 = 1.0$, $\beta = (1.0, 0.4, 0.4)$ and $\phi = 1.2$. In this case, each of two SNPs are assumed to influence gene expression based on an additive genetic model and an additive model of association. Both SNPs are generated with a minor allele frequency of 0.20. Sample sizes of 150 and 500 are considered for the GWA data (nGWA) and sample sizes of 1000 and 100 are considered for the TWA data (nTWA). 1000 simulations are conducted per condition to estimate type-1 error and evaluate convergence when $\gamma = 0$ and to estimate power under a range of effect sizes. The convergence criterion for the EM is defined as a maximum absolute difference of 1×10^{-6} in the parameters estimates between consecutive iterations of the EM.

The distributions of the resulting estimates are illustrated in Figure 3 and corresponding numerical values are given in Table 2. The samples sizes in Figure 3(a) are approximately equal to the sample sizes for the data example of Section 4. In this case, the number of iterations until convergence ranged from 16 to 479 with a median of 54 and in inter-quartile range of 42 to 71. As seen in this figure, the estimates after convergence are centered around the true parameter values in the cases in which nTWA=1000, with the exception of α , the scale parameter from the negative binomial model, which tends to be slightly over estimated. This bias is accentuated in the setting of a small TWA sample size as shown in Figure 3(c) where greater bias is also observed for α , η_1 and η_2 . This result is expected as the maximum likelihood estimate of the scale parameter from a negative binomial model is known to be biased²⁶.

Estimated type-1 error rates and power for a range of γ values and sample sizes are illustrated in Figure 4. For comparison, we again implement regression imputation similar to the strategies described in⁸ for this setting with two notable differences: (1) a negative binomial model is assumed for the RNAseq data in place a Gaussian linear model; and (2) LD reduction is applied prior to model fitting to obviate the need for using penalized regression. These changes are made for consistency with the EM approach, and are not expected to change the relative performance of the two strategies. Based on these simulations, the EM strategy consistently performs better than regression imputation with

respect to power while the estimated type 1 error rates are comparable and close to the nominal 0.05 level. A reduction in the GWA sample size (nGWA) from 500 to 100, holding the TWA sample size (nTWA) fixed at 1000, results in a substantial reduction in power; however, a change in nTWA from 1000 to 100, with nGWA fixed at 500, corresponds to only a modest reduction in power for both the EM and regression imputation approaches. In-line with the bias described above, type-1 error rate is slightly inflated at 0.084 for the EM approach in the small TWA sample size setting.

4 | APPLICATION TO THE GENE STUDY WITH THE DGN REFERENCE PANEL

In a recent investigation involving the GENE study data⁴, we reported on three protein-coding genes whose genotypes associate ($p < 5 \times 10^{-6}$) with resting state or change from baseline to peak biomarker response after an endotoxin challenge, as indicated in Table 3. We now extend this analysis, interrogating these three genes further, by testing for association of gene-level *expression* in whole blood with biomarker response, using the DGN reference panel data. By testing for expression-trait association, we are investigating whether the corresponding gene expression is a potential mediator between the genotype and the biomarker. Due to differences in genotyping platforms between the GENE and DGN studies, only a subset of SNPs used in the original GENE study analysis are also available in the DGN data. The present analysis uses overlapping SNPs while alternative strategies for addressing inconsistencies across platforms are discussed in Section 5. LD pruning is applied based on this subset using a threshold of < 0.2 ¹⁹. The resulting numbers of SNPs are provided in the last three columns of Table 3.

Results of applying the EM strategy to the GENE study data (nGWA = 174), using DGN whole blood RNAseq as the reference panel (nTWA = 922), are provided in the top portion of Table 4. Here the top signal SNP is reported for each gene, corresponding to the column with heading η^* , and for all three genes, a strong SNP-level association with the corresponding gene whole-blood RNAseq ($p < 0.002$) is observed. Whole-blood expression-level association with biomarker, on the other hand, is not statistically significant for any of the three genes under consideration. For completeness, the results of applying the alternative regression imputation approach described in Section 3 are also provided in this table. The EM algorithm is known to converge slowly and is expected to be computationally less efficient than the regression imputation approach. In this application, using a single 3.2 GHz Intel Xeon W processor, the EM approach takes 1 minute 16 seconds, 4 minutes 35 seconds, and 2 minutes and 5 seconds for the analysis of COL5A2, PDIA5 and LRRC16A, respectively.

In this data example, the SNP-level associations with expression are generally consistent between the EM strategy and regression imputation. While inferences regarding the association between expression of PDIA5 and LRCC16A in whole blood and the biomarker are also consistent between the two approaches, regression imputation identifies a significant association ($p < 0.001$) between expression of COL5A2 in whole blood and baseline CRP that is not significant using the EM approach. The smaller estimated value of the negative

binomial scale parameter ($\hat{\alpha} = 1.16$) for this gene compared to corresponding parameter estimates for PDIA5 and LRCC16A ($\hat{\alpha} = 4.00$ and $\hat{\alpha} = 3.28$, respectively) suggests a greater rightward skew in the distribution of expression values. As reported in the simulation study based on $\hat{\alpha} = 1.2$, regression imputation, which is based on a mean value, performs less favorably with respect to type-1 error in this setting – 0.058 versus 0.032. In order to further investigate this result, we additionally evaluated bias for the parameter γ using regression imputation. For the purpose of comparison, we assume the same conditions as those indicated for the column labeled ‘(b)’ in Table 2, which are most similar to the real data example setting. Based on an additional 1000 simulations, the estimated bias is 0.0024 (se = 0.0585) which is ten-fold greater than what we observed for the EM setting (bias = 0.0002, se = 0.0412). Finally we apply the unified likelihood and regression imputation strategies using normalized whole blood expression of the LRRC16A gene. The resulting parameter estimates and inferences are similar between the unified and regression imputation strategies with slightly larger absolute test statistics for γ and α^* in the unified approach (Table 5). This result is consistent with the slight increase in power observed in the simulation studies.

5 | DISCUSSION

This manuscript describes a fully likelihood-based framework for transcriptome association analysis for normalized and count expression data. Similar to previously described regression imputation strategies, the likelihood-based approaches leverage both GWA data, which include genotype and trait information, and TWA data, which include genotype and cell or tissue specific expression profiles. The unified and EM approaches described in this work differ from regression imputation as they account for uncertainty in prediction of unobserved expression data, and additionally use available information on the trait to inform the missing data. Associated gains in power for expression-trait association analysis were observed in our simulation study (Section 3, Figures 4 and 2). Application of the two strategies to the GENE study and DGN data were generally consistent with the exception that regression imputation identified a significant association between COL5A2 expression and baseline CRP that was not identified using the EM – a result that may be attributable to a type-1 error. In general, the type-1 error rates are comparable between the likelihood-based and regression imputation approaches with both controlled at the 0.05 level, while more notable gains in power are observed using the likelihood-based strategies as compared to regression imputation.

In the data example provided, a normally distributed univariate trait and a single negative binomially distributed or normally distributed cell-specific gene expression was considered. The theory presented is based on a generalized linear model, and thus the framework is flexible for handling binary or repeatedly measured traits, as well as normally or Poisson distributed gene expressions. Additional extensions for survival outcomes can be achieved through inclusion of a Cox proportional hazards or fully parametric accelerated failure time model in place of Equations 1 and 2. Simultaneous consideration of multiple gene expressions, across cell or tissue types and/or across more than one gene or regulatory element, requires further extensions that account for the correlations among genes within a locus as well as expression profiles across compartments. In order to account for correlations

among SNPs within a single gene, an LD pruning strategy was applied. Alternatively, a penalized regression approach, such as the elastic net, could be integrated into the likelihood-based framework; however, an alternative to cross-validation within each iteration of the EM to identify an appropriate tuning parameter may be needed to reduce the associated computational burden. For the data example, analysis was limited to SNPs present in both the GWA and TWA data resources. A common approach to address inconsistencies across platforms is SNP imputation based on the 1000 genomes or HapMap reference panels^{15,27}. While existing regression imputation strategies, such as PrediXcan, simply treat these imputed SNPs as observed⁸, the likelihood framework offers a natural structure for additionally and appropriately layering in these missing data components. The expected power gains from incorporating additional SNPs across platforms will depend on the extent to which the unobserved SNPs explain additional variability in expression. Finally, we focus in this manuscript on frequentist strategies; however, a well-established alternative strategy for missing data is a fully Bayesian MCMC approach, in which unobserved covariates, expression profiles in our setting, are treated as random variables. Our choice to use a frequentist approach stems from interest in classical hypothesis testing, although further research is needed to explore the relative advantages of a fully Bayesian approach in this setting, particularly with respect to extensibility and computational efficiency.

Acknowledgements

Support for this research was provided by NIH R15GM126485 and R01GM127862.

APPENDIX A

Without loss of generality, we assume that $Y|w, x \sim N(\gamma_0 + x\gamma_1, \sigma_e^2)$ and $X|\mathbf{z} \sim N(\mathbf{z}^T \boldsymbol{\beta}, \sigma_\delta^2)$, where $w = 1$, x is a scalar and $\mathbf{z}$ is a vector. The observed fisher information matrix is based on the following calculations.

The log likelihood function is:

$$l = \sum_{i=1}^n \left[-\frac{1}{2} \log(\sigma_e^2 + \sigma_\delta^2 \gamma_1^2) - \frac{(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{2(\sigma_e^2 + \sigma_\delta^2 \gamma_1^2)} \right] + \sum_{j=1}^m \left[-\frac{1}{2} \log \sigma_\delta^2 - \frac{(x_j - \mathbf{z}_j^T \boldsymbol{\beta})^2}{2\sigma_\delta^2} \right]$$

The first order derivatives are:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) \gamma_1 \mathbf{z}_i}{\sigma_e^2 + \sigma_\delta^2 \gamma_1^2} + \sum_{j=1}^m \frac{(x_j - \mathbf{z}_j^T \boldsymbol{\beta}) \mathbf{z}_j}{\sigma_\delta^2}$$

$$\frac{\partial l}{\partial \gamma_0} = \sum_{i=1}^n \frac{(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)}{\sigma_e^2 + \sigma_\delta^2 \gamma_1^2}$$

$$\frac{\partial l}{\partial \gamma_1} = \sum_{i=1}^n \left[-\frac{\sigma_{\delta}^2 \gamma_1}{\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2} - \frac{(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)(-\mathbf{z}_i^T \boldsymbol{\beta})(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2) - \sigma_{\delta}^2 \gamma_1 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2)^2} \right]$$

$$\frac{\partial l}{\partial \sigma_{\delta}^2} = \sum_{i=1}^n \left[-\frac{1}{2} \frac{\gamma_1^2}{\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2} - \frac{-2\gamma_1^2 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{4(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2)^2} \right] + \sum_{j=1}^n \left[-\frac{1}{2} \frac{1}{\sigma_{\delta}^2} - \frac{-2(x_j - \mathbf{z}_j^T \boldsymbol{\beta})^2}{4\sigma_{\delta}^4} \right]$$

$$\frac{\partial l}{\partial \sigma_{\epsilon}^2} = \sum_{i=1}^n \left[-\frac{1}{2} \frac{1}{\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2} - \frac{-2(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{4(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2)^2} \right]$$

The second order derivatives are:

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} = \sum_{i=1}^n \frac{-\gamma_1^2 \mathbf{z}_i \mathbf{z}_i^T}{\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2} + \sum_{j=1}^m \frac{-\mathbf{z}_j \mathbf{z}_j^T}{\sigma_{\delta}^2}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \gamma_0} = \sum_{i=1}^n \frac{-\gamma_1 \mathbf{z}_i}{\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \gamma_1} = \sum_{i=1}^n \left[\frac{[(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) \mathbf{z}_i - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1 \mathbf{z}_i](\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2) - 2\sigma_{\delta}^2 \gamma_1^2 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) \mathbf{z}_i}{(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2)^2} \right]$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \sigma_{\delta}^2} = \sum_{i=1}^n \frac{-\gamma_1^3 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) \mathbf{z}_i}{(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2)^2} + \sum_{j=1}^m \frac{-(x_j - \mathbf{z}_j^T \boldsymbol{\beta}) \mathbf{z}_j}{\sigma_{\delta}^4}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \sigma_{\epsilon}^2} = \sum_{i=1}^n \frac{-(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) \gamma_1 \mathbf{z}_i}{(\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2)^2}$$

$$\frac{\partial^2 l}{\partial \gamma_0^2} = \sum_{i=1}^n \frac{-1}{\sigma_{\epsilon}^2 + \sigma_{\delta}^2 \gamma_1^2}$$

$$\frac{\partial^2 l}{\partial \gamma_0 \partial \gamma_1} = \sum_{i=1}^n = \frac{-\mathbf{z}_i^T \boldsymbol{\beta} (\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2) - 2\sigma_\delta^2 \gamma_1 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2}$$

$$\frac{\partial^2 l}{\partial \gamma_0 \partial \sigma_\delta^2} = \sum_{i=1}^n \frac{-\gamma_1^2 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2}$$

$$\frac{\partial^2 l}{\partial \gamma_0 \partial \sigma_\delta} = \sum_{i=1}^n \frac{-(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \gamma_1^2} &= \sum_{i=1}^n \frac{\sigma_\delta^2 \sigma_\epsilon^2 - \sigma_\delta^4 \gamma_1^2}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} - \frac{(\mathbf{z}_i^T \boldsymbol{\beta})^2 (\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2) + 2\sigma_\delta^2 \gamma_1 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) (\mathbf{z}_i^T \boldsymbol{\beta})}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} \\ &+ \frac{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2) \{ \sigma_\delta^2 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2 - 2\sigma_\delta^2 \gamma_1 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) (\mathbf{z}_i^T \boldsymbol{\beta}) \} - 4\sigma_\delta^4 \gamma_1^2 (\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2) (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^4} \end{aligned}$$

$$\frac{\partial^2 l}{\partial \gamma_1 \partial \sigma_\epsilon^2} = \sum_{i=1}^n \left[\frac{\sigma_\delta^2 \gamma_1}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} - \frac{(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1) (\mathbf{z}_i^T \boldsymbol{\beta})}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} + \frac{2\sigma_\delta^2 \gamma_1 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^3} \right]$$

$$\frac{\partial l}{\partial (\sigma_\delta^2)^2} = \sum_{i=1}^n \left[\frac{\gamma_1^4}{2(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} - \frac{\gamma_1^2 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^3} \right] + \sum_{j=1}^m \left[\frac{1}{2\sigma_\delta^4} - \frac{(x_j - \mathbf{z}_j^T \boldsymbol{\beta})^2}{\sigma_\delta^6} \right]$$

$$\frac{\partial^2 l}{\partial \sigma_\delta^2 \partial \sigma_\epsilon^2} = \sum_{i=1}^n \left[\frac{\gamma_1^2}{2(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} - \frac{\gamma_1 (y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^3} \right]$$

$$\frac{\partial^2 l}{\partial (\sigma_\delta^2)^2} = \sum_{i=1}^n \left[\frac{1}{2(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^2} - \frac{(y_i - \gamma_0 - \mathbf{z}_i^T \boldsymbol{\beta} \gamma_1)^2}{(\sigma_\epsilon^2 + \sigma_\delta^2 \gamma_1^2)^3} \right]$$

APPENDIX B

The elements of the score function are given by:

$$U_i(\gamma_1; \Phi) = x_i(y_i - \mathbf{w}_i^T \boldsymbol{\gamma}_0 - x_i \gamma_1) / \sigma^2,$$

$$U_i(\boldsymbol{\gamma}_0; \Phi) = \mathbf{w}_i^T (y_i - \mathbf{w}_i^T \boldsymbol{\gamma}_0 - x_i \gamma_1) / \sigma^2,$$

$$U_i(\sigma^2; \Phi) = -\frac{n}{2\sigma^2} + \frac{(y_i - \mathbf{w}_i^T \boldsymbol{\gamma}_0 - x_i \gamma_1)^T (y_i - \mathbf{w}_i^T \boldsymbol{\gamma}_0 - x_i \gamma_1)}{2\sigma^4},$$

$$U_i(\boldsymbol{\beta}; \Phi) = \mathbf{z}_i^T \left(\frac{x_i - \exp(\mathbf{z}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\beta}) / \alpha} \right),$$

and

$$U_i(\alpha; \Phi) = \Psi_0(x_i + \alpha) - \Psi_0(\alpha) - \frac{x_i + \alpha}{\exp(\mathbf{z}_i^T \boldsymbol{\beta}) + \alpha} - \log(\exp(\mathbf{z}_i^T \boldsymbol{\beta}) / \alpha + 1) + 1,$$

where $\Psi_0(x) = d[\log \Gamma(x)]/dx$ is the digamma function. Letting the expectation of the second derivative term in Equation 8 with respect to the parameters ϕ_1 and ϕ_2 , be given by:

$$\mathcal{B}_{\Phi, i}(\phi_1, \phi_2) = E_{\Phi} \left(-\partial^2 l_{\Phi, i}^c / \partial \phi_1 \partial \phi_2 \mid \mathbf{y}, \mathbf{w}, \mathbf{z} \right).$$

We have the following results:

$$\mathcal{B}_{\Phi, i}(\gamma_1, \gamma_1) = \sum_{x_i \in \mathcal{X}_i} \hat{\pi}_{x, i}^{(t)} x_i^2 / \sigma^2,$$

$$\mathcal{B}_{\Phi, i}(\gamma_1, \boldsymbol{\gamma}_0) = \mathcal{B}_{\Phi, i}(\boldsymbol{\gamma}_0, \gamma_1) = \sum_{x_i \in \mathcal{X}_i} \hat{\pi}_{x, i}^{(t)} x_i \mathbf{w}_i^T / \sigma^2,$$

$$\mathcal{B}_{\Phi, i}(\gamma, \sigma^2) = \mathcal{B}_{\Phi, i}(\sigma^2, \gamma) = \sum_{x_i \in \mathcal{X}_i} \hat{\pi}_{x, i}^{(t)} \frac{x_i (y_i - \mathbf{w}_i^T \boldsymbol{\gamma}_0 - x_i \gamma_1)}{\sigma^4},$$

$$\mathcal{B}_{\Phi, i}(\gamma_0, \gamma_0) = \sum_{x_i \in \chi_i} \hat{\pi}_{x, i}^{(t)} \mathbf{w}_i^T \mathbf{w}_i / \sigma^2,$$

$$\mathcal{B}_{\Phi, i}(\gamma_0, \sigma^2) = \mathcal{B}_{\Phi, i}(\sigma^2, \gamma_0) = \sum_{x_i \in \chi_i} \hat{\pi}_{x, i}^{(t)} \frac{\mathbf{w}_i (y_i - \mathbf{w}_i^T \gamma_0 - x_i \gamma_1)}{\sigma^4},$$

$$\mathcal{B}_{\Phi, i}(\sigma^2, \sigma^2) = \sum_{x_i \in \chi_i} \hat{\pi}_{x, i}^{(t)} \left[\frac{(y_i - \mathbf{w}_i^T \gamma_0 - x_i \gamma_1)^2}{\sigma^6} - \frac{1}{2\sigma^4} \right],$$

$$\mathcal{B}_{\Phi, i}(\beta, \beta) = \sum_{x_i \in \chi_i} \hat{\pi}_{x, i}^{(t)} \mathbf{z}_i^T \left[\frac{\alpha \mu_i (x_i + \alpha)}{(\alpha + \mu_i)^2} \right],$$

$$\mathcal{B}_{\Phi, i}(\beta, \alpha) = \mathcal{B}_{\Phi, i}(\alpha, \beta) = \sum_{x_i \in \chi_i} \hat{\pi}_{x, i}^{(t)} \mathbf{z}_i^T \left[\frac{\mu_i (\mu_i - x_i)}{(\alpha + \mu_i)^2} \right],$$

$$\mathcal{B}_{\Phi, i}(\alpha, \alpha) = \sum_{x_i \in \chi_i} \hat{\pi}_{x, i}^{(t)} \left[-\Psi_1(x_i + \alpha) + \Psi_1(\alpha) - \frac{(\mu_i^2 + x_i \alpha)}{\alpha(\mu_i + \alpha)^2} \right],$$

where x_i and $\mu_i = \exp(\mathbf{z}_i^T \beta)$ are scalars. The remaining six second derivatives are equal to 0. The empirical observed information matrix of Equation 8 can be estimated using these results replacing each parameter with its estimate at the final iteration of the EM algorithm.

APPENDIX C

Under the Gaussian model, we have the partial derivative of the expectation of the complete data log likelihood with respect to ϕ given by:

$$\begin{aligned} \frac{\partial}{\partial \phi} L_1 &= \sum_{j=1}^N \hat{p}_j^{(t)} \frac{\partial}{\partial \Phi} \left[(\tilde{y}_j \theta_j - \theta_j^2/2) / \phi - \frac{1}{2} (\tilde{y}_j^2 / \phi + \log(2\pi\phi)) \right] \\ &= \sum_{j=1}^N \left[\frac{\hat{p}_j^{(t)} (\tilde{y}_j \theta_j - \theta_j^2/2)}{\phi^2} + \frac{\hat{p}_j^{(t)} \tilde{y}_j^2}{2\phi^2} - \frac{\hat{p}_j^{(t)}}{2\phi} \right] \\ &= \sum_{j=1}^N \left[\frac{\hat{p}_j^{(t)} (\tilde{y}_j - \theta_j)^2}{2\phi^2} - \frac{\hat{p}_j^{(t)}}{2\phi} \right]. \end{aligned}$$

Setting this equal to 0 yields the maximum likelihood estimate for ϕ :

$$\hat{\phi}^{(t+1)} = \frac{\sum_{j=1}^N \hat{p}_j^{(t)} (\tilde{y}_j - \theta_j)^2}{\sum_{j=1}^N \hat{p}_j^{(t)}}.$$

On the other hand, in specifying the weights option for `glm()` in R, the partial derivative of the log likelihood with respect to ϕ is given by:

$$\begin{aligned} \frac{\partial}{\partial \phi} L_2 &= \sum_{j=1}^N \frac{\partial}{\partial \phi} \left[\hat{p}_j^{(t)} (\tilde{y}_j \theta_j - \theta_j^2/2) / \phi - \frac{1}{2} (\hat{p}_j^{(t)} \tilde{y}_j^2 / \phi + \log(2\pi\phi / \hat{p}_j^{(t)})) \right] \\ &= \sum_{j=1}^N \left[-\frac{\hat{p}_j^{(t)} (\tilde{y}_j \theta_j - \theta_j^2/2)}{\phi^2} + \frac{\hat{p}_j^{(t)} \tilde{y}_j^2}{2\phi^2} - \frac{1}{2\phi} \right] \\ &= \sum_{j=1}^N \left(\frac{\hat{p}_j^{(t)} (\tilde{y}_j - \theta_j)^2}{2\phi^2} \right) - \frac{N}{2\phi}. \end{aligned}$$

Setting this equal to 0 yields the weighted least squares estimator:

$$\hat{\phi}_P^{(t+1)} = \frac{\sum_{j=1}^N \hat{p}_j^{(t)} (\tilde{y}_j - \theta_j)^2}{N}.$$

Thus:

$$\hat{\phi}_P^{(t+1)} * \frac{N}{n} = \hat{\phi}^{(t+1)},$$

$$\text{where } n = \sum_{j=1}^N \hat{p}_j^{(t)}.$$

Under the negative binomial model, we have the left hand side of Equation 15 given by:

$$\begin{aligned} \sum_{j=1}^N \hat{p}_j^{(t)} \frac{\partial}{\partial \phi} c(\tilde{r}_j, \phi) &= \sum_{j=1}^N \hat{p}_j^{(t)} \frac{\partial}{\partial \phi} \log \left(\frac{\Gamma(\tilde{r}_j + \phi)}{\Gamma(\tilde{r}_j + 1) \Gamma(\phi)} \right) \\ &= \sum_{j=1}^N \hat{p}_j^{(t)} [\Psi_0(\tilde{r}_j + \phi) - \Psi_0(\phi)] \\ &= \sum_{j=1}^N \hat{p}_j^{(t)} \sum_{k=\phi+1}^{\tilde{r}_j + \phi - 1} \frac{1}{k}. \end{aligned}$$

The right hand side of Equation 15 on the other hand is given by:

$$\begin{aligned}
\sum_{j=1}^N \frac{\partial}{\partial \Phi} c(\tilde{r}_j, \phi / \hat{p}_j^{(t)}) &= \sum_{j=1}^N \frac{\partial}{\partial \Phi} \log \left(\frac{\Gamma(\tilde{r}_j + \phi / \hat{p}_j^{(t)})}{\Gamma(\tilde{r}_j + 1) \Gamma(\phi / \hat{p}_j^{(t)})} \right) \\
&= \sum_{j=1}^N \frac{\Psi_0(\tilde{r}_j + \phi / \hat{p}_j^{(t)}) - \Psi_0(\phi / \hat{p}_j^{(t)})}{\hat{p}_j^{(t)}} \\
&= \sum_{j=1}^N \frac{1}{\hat{p}_j^{(t)}} \sum_{k=\phi / \hat{p}_j^{(t)} + 1}^{\tilde{r}_j + \phi / \hat{p}_j^{(t)} - 1} \frac{1}{k}.
\end{aligned}$$

□

Abbreviations:

DGN	Depression Gene Network
EM	Expectation-Maximization
GENE	Genetics of Evoked Responses to Niacin and Endotoxemia
GWA	Genome-Wide Association
TWA	Transcriptome-Wide Association

References

1. Ferguson JF, Patel PN, Shah RY, et al. Race and gender variation in response to evoked inflammation. *J Transl Med* 2013;11:63. [PubMed: 23497455]
2. Ferguson JF, Ryan MF, Gibney ER, Brennan L, Roche HM, Reilly MP Dietary isoflavone intake is associated with evoked responses to inflammatory cardiometabolic stimuli and improved glucose homeostasis in healthy volunteers. *Nutr Metab Cardiovasc Dis* 2014;24(9):996–1003. [PubMed: 24875672]
3. Ferguson JF, Meyer NJ, Qu L, et al. Integrative genomics identifies 7p11.2 as a novel locus for fever and clinical stress response in humans. *Hum. Mol. Genet* 2015;24(6):1801–1812. [PubMed: 25416278]
4. Qian J, Nunez S, Kim S, Reilly MP, Foulkes AS A score test for genetic class-level association with nonlinear biomarker trajectories. *Stat Med* 2017;36(19):3075–3091. [PubMed: 28543585]
5. Lonsdale J, Thomas J, Salvatore M, Phillips R, al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet* 2013;45(6):580–585. [PubMed: 23715323]
6. Lappalainen T, Sammeth M, Friedlander MR, al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013;501(7468):506–511. [PubMed: 24037378]
7. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 2014;24(1):14–24. [PubMed: 24092820]
8. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet* 2015;47(9):1091–1098. [PubMed: 26258848]
9. Wheeler HE, Aquino-Michaels K, Gamazon ER, et al. Poly-omic prediction of complex traits: OmicKriging. *Genet. Epidemiol* 2014;38(5):402–415. [PubMed: 24799323]
10. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet* 2016;48(3):245–252. [PubMed: 26854917]

11. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet* 2016;48(5):481–487. [PubMed: 27019110]
12. Zou Hui, Hastie Trevor Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B* 2005;67:301–320.
13. Tibshirani Robert Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* 1994;58:267–288.
14. Zhou X, Carbonetto P, Stephens M Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* 2013;9(2):e1003264. [PubMed: 23408905]
15. Hu YJ, Lin DY Analysis of untyped SNPs: maximum likelihood and imputation methods. *Genet. Epidemiol* 2010;34(8):803–815. [PubMed: 21104886]
16. Little Roderick JA Regression with Missing X's: A Review. *Journal of the American Statistical Association* 1992;87(420):1227–1237.
17. Dempster AP, Laird NM, Rubin DB Maximum likelihood from incomplete data via the EM algorithm (C/R: p22–37). *Journal of the Royal Statistical Society, Series B, Methodological* 1977;39:1–22.
18. McCullagh P, Nelder JA Generalized Linear Models Chapman and Hall; 1989.
19. Laurie CC, Doheny KF, Mirel DB, al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol* 2010;34(6):591–602. [PubMed: 20718045]
20. Robinson MD, Smyth GK Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 2007;23(21):2881–2887. [PubMed: 17881408]
21. Robinson MD, McCarthy DJ, Smyth GK edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139–140. [PubMed: 19910308]
22. Meilijson Isaac A Fast Improvement to the EM Algorithm on Its Own Terms. *Journal of the Royal Statistical Society, Series B: Methodological* 1989;51:127–138.
23. Louis TA Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1982;44(2):226–233.
24. Meng Xiao-Li, Rubin Donald B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 1993;80:267–278.
25. Venables WN, Ripley BD Modern Applied Statistics with S New York: Springer; fourth ed.2002 ISBN 0–387-95457–0.
26. Saha K, Paul S Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* 2005;61(1):179–185. [PubMed: 15737091]
27. Marchini J, Howie B Genotype imputation for genome-wide association studies. *Nat. Rev. Genet* 2010;11(7):499–511. [PubMed: 20517342]

FIGURE 1. Simulation-based parameter distributions after convergence for normalized expression data

The distribution of parameter estimates minus their true values across 1000 simulations are illustrated. *A subset of the 20 SNPs analyzed are illustrated in this figure to reflect a range of effect sizes and minor allele frequencies: rs2690110 has a minor allele frequency of 0.36 and an estimated parameter value (based on the observed data) of 0.25; rs301371 has a minor allele frequency of 0.42 and an estimated parameter value of 0.049; and rs999039 has a minor allele frequency of 0.17 and an estimated parameter value of 0.053. While decreasing the sample size of the GWA data set (left-hand and center panels) increases the variability in the parameter estimates, the impact of reducing the TWA sample size is more impactful (right-hand panel) with greater bias and variability for both the fully likelihood-based and single regression imputation approaches. Numerical summaries are provided in Table 1

FIGURE 2. Power of unified likelihood and regression imputation strategies for normalized expression data

Power and type-1 error rates in the left-hand panel are estimated based 10,000 simulations per condition and a significance threshold of 0.05. The black horizontal lines have y-intercepts of 0.05 and 0.80. For the left hand panel, the observed data sample sizes of nGWA=173 and nTWA=906 are assumed and a range of effect sizes (μ/σ_c) are considered. For the center and right hand panel an expression effect size of 0.3 is assumed and the observed genotype data are used. Assumed parameter values are based on application to the CARMIL1, previously LRRC16A, gene (Table 5). A range of nTWA sample sizes are considered and the sample size of nGWA is assumed to be the observed sample size and then inflated by a factor of 2. Power remains relatively constant for nTWA sample sizes ranging from 250 to 1000. Power increases substantially with larger GWA sample size. Power of the likelihood-based approach is consistently higher than the regression imputation approach, though the difference is between 0 and 2% for the scenarios considered. This difference is statistically significant for the observed sample sizes.

FIGURE 3. Simulation-based distributions using EM-based and regression imputation strategies for expression as count data

The distribution of bias for parameter estimates across 1000 simulations are illustrated. In all three scenarios, the upward bias observed for the scale parameter (α) is consistent with the known bias associated with the maximum likelihood estimator from a negative binomial model. This is accentuated with the EM approach in the small TWA sample size setting (right-hand panel). The results in the larger TWA settings (left-hand and middle panels) are generally consistent between the EM and regression imputation approaches. This is summarized numerically in Table 2.

FIGURE 4. Power analysis for EM and regression imputation strategies

Power and type-1 error rates are estimated based 1000 simulations and a significance threshold of 0.05. The black horizontal dotted lines have y-intercepts of 0.05 and 0.80. Assumed parameter values are $\beta = 2.0$, $\sigma^2 = 1.0$, $\eta = (1.0, 0.4, 0.4)$ and $\alpha = 1.2$, where each of two SNPs are generated with a minor allele frequency of 0.20. Power of the EM approach is consistently higher than the regression imputation strategy. Similar to the unified likelihood approach described for normalized data, power increases substantially with increasing GWA sample size and is less impacted by TWA sample size.

TABLE 1

Simulation-based estimates of bias and standard error under varying sample sizes for normalized expression data

Parameter	Method	(a) [§]		(b) [§]		(c) [§]	
		bias [*]	se	bias [*]	se	bias [*]	se
β_0	Likelihood-based	-0.0033	0.1421	0.0001	0.1493	0.0203	0.5069
	Reg-Imp	-0.0030	0.1422	0.0002	0.1493	0.0233	0.5271
β_1 (rs2690110 [*])	Likelihood-based	0.0012	0.0374	-0.0010	0.0365	0.0020	0.1309
	Reg-Imp	0.0013	0.0376	-0.0010	0.0364	0.0013	0.1344
β_2 (rs301371 [*])	Likelihood-based	0.0001	0.0364	-0.0004	0.0373	0.0035	0.1276
	Reg-Imp	0.0002	0.0365	-0.0004	0.0372	0.0031	0.1302
β_3 (rs999039 [*])	Likelihood-based	-0.0000	0.0547	-0.0006	0.0545	0.0052	0.1924
	Reg-Imp	-0.0000	0.0549	-0.0007	0.0545	0.0055	0.1956
δ_0	Likelihood-based	0.0006	0.0546	0.0049	0.0996	0.0016	0.0554
	Reg-Imp	0.0004	0.0546	0.0047	0.0996	-0.0002	0.0548
δ_1	Likelihood-based	0.0004	0.0751	-0.0056	0.1388	-0.0012	0.0807
	Reg-Imp	-0.0021	0.0739	-0.0081	0.1364	-0.0241	0.0683
σ_δ	Likelihood-based	-0.0112	0.0212	-0.0122	0.0218	-0.1242	0.0818
	Reg-Imp	-0.0007	0.0212	-0.0016	0.0218	-0.0108	0.0808
σ_ϵ	Likelihood-based	-0.0026	0.0312	-0.0122	0.0594	-0.0047	0.0315
	Reg-Imp	0.0040	0.0307	0.0011	0.0579	0.0027	0.0312

^{*} A subset of the 20 SNPs analyzed are included in this table to reflect a range of effect sizes and minor allele frequencies: rs2690110 has a minor allele frequency of 0.36 and an estimated parameter value (based on the observed data) of 0.25; rs301371 has a minor allele frequency of 0.42 and an estimated parameter value of 0.049; and rs999039 has a minor allele frequency of 0.17 and an estimated parameter value of 0.053.

TABLE 2

Simulation-based estimates of bias and standard error under varying sample sizes for expression as count data

Parameter	Method	(a) [§]		(b) [§]		(c) [§]	
		bias*	se	bias*	se	bias*	se
α	EM	0.0149	0.0786	0.0102	0.0794	0.1506	0.2944
	Reg-Imp	0.0078	0.0778	0.0087	0.0753	0.0087	0.0753
β_0	EM	-0.0014	0.0489	-0.0032	0.0481	0.0007	0.1473
	Reg-Imp	0.0006	0.0484	-0.0025	0.0483	-0.0025	0.0483
B_1	EM	0.0005	0.0572	-0.0007	0.0571	-0.0503	0.1697
	Reg-Imp	-0.0027	0.0560	0.0019	0.0584	0.0019	0.0584
β_2	EM	-0.0037	0.0565	0.0016	0.0565	-0.0478	0.1587
	Reg-Imp	-0.0023	0.0574	-0.0022	0.0576	-0.0022	0.0576
γ_0	EM	-0.0006	0.1203	-0.0009	0.1810	-0.0005	0.1523
	Reg-Imp	0.0035	0.1329	0.0124	0.2454	0.0124	0.2454
γ_1	EM	0.0005	0.0286	0.0002	0.0412	-0.0002	0.0383
	Reg-Imp	-0.0011	0.0317	-0.0024	0.0585	-0.0024	0.0585
σ	EM	-0.0045	0.0348	-0.0039	0.0636	-0.0083	0.0390
	Reg-Imp	-0.0010	0.0306	-0.0028	0.0596	-0.0028	0.0596

[§]Sample sizes are as follows: (a) nGWA=500 and nTWA=1000; (b) nGWA=150 and nTWA=1000; and (c) nGWA=500 and nTWA=100.

* Bias is defined as $(E(\hat{\phi}_k) - \phi_k)$ for $k = 1, \dots, 7$ where ϕ_k represents the k th parameter. Bias is estimated by replacing $E(\hat{\phi}_k)$ with the mean of the corresponding parameter estimate across 1000 simulations.

TABLE 3

Candidate loci interrogated in transcriptome association analysis

Gene [§]	Chr	Start	Stop	Biomarker [○]	# SNPs [*]		
					(1)	(2)	(3)
COL5A2	1	189896640	190044605	CRP	7	2	1
PDIA5	6	122785855	122880953	CRP	38	12	6
LRRC16A	20	25279655	25620758	IL-6	341	63	20

[§]Previously identified genes associated with biomarkers at the $p < 5 \times 10^{-6}$ level in the GENE study include: Collagen Type V Alpha 2 Chain (COL5A2), Protein Disulfide Isomerase Family A Member 5 (PDIA5) and Capping Protein Regulator And Myosin 1 Linker 1 (CARMIL1, previously LRRC16A).

[○]Biomarker associations correspond to resting state (baseline) C-reactive protein (CRP) or change from baseline to peak after stimulus interleukin-6 (IL-6).

^{*}A subset of SNPs within each gene are used in the transcriptome association analysis. Here (1) represents the number of SNPs in the original GENE study data; (2) represents the number of SNPs in both the GENE study and the DGN data; (3) represents the number of SNPs remaining after LD pruning using a threshold of 0.20.

TABLE 4

Transcriptome-biomarker association analysis

		γ_0	γ_1	σ	β_0	β_1^*	α
EM [§] :							
COL5A2	estimate:	-0.377	-0.00914	1.118	3.126	-0.354	1.162
	std err:	0.179	0.00735	–	0.0346	0.0713	–
	test stat:	-2.113	-1.243	–	90.450	-4.970	–
	p-value:	0.0346	0.214	–	< 0.001	< 0.001	–
PDIA5	estimate:	-0.256	-0.00100	1.122	5.654	0.0904	4.003
	std err:	0.413	0.00128	–	0.0415	0.0285	–
	test stat:	-0.784	-0.619	–	136.104	3.176	–
	p-value:	0.536	0.433	–	< 0.001	0.002	–
LRRC16A	estimate:	4.362	0.00118	1.501	5.0286	-0.288	3.281
	std err:	0.346	0.00184	–	0.132	0.0370	–
	test stat:	12.609	0.656	–	38.049	-7.781	–
	p-value	< 0.001	0.523	–	< 0.001	< 0.001	–
Regression Imputation:							
CoL5A2	estimate:	1.966	-0.119	1.075	3.123	-0.343	1.160
	std err:	0.638	0.0297	–	0.0346	0.0726	–
	test stat:	3.083	-4.015	–	90.917	-4.726	–
	p-value:	0.002	< 0.001	–	< 0.001	< 0.001	–
PDIA5	estimate:	0.529	-0.00348	1.120	5.651	0.0913	4.002
	std err:	0.920	0.00290	–	0.0415	0.0284	–
	test stat:	0.575	-1.203	–	158.692	2.937	–
	p-value:	0.566	0.231	–	< 0.001	0.002	–
LRRC16A	estimate:	4.166	0.00228	1.491	5.039	-0.288	3.280
	std err:	0.530	0.00292	–	0.131	0.0370	–
	test stat:	7.859	0.780	–	38.509	-7.789	–
	p-value	< 0.001	0.436	–	< 0.001	< 0.001	–

[§]Coefficient estimates and standard errors are based on application of the EM strategy using a convergence criterion of $< 1 \times 10^{-6}$. Convergence is reached at 34 iterations for COL5A2, 71 iterations for PDIA5 and 34 iterations for LRCC16A.

^{*}The coefficient estimate and corresponding standard error is reported for the strongest signal SNP within each gene (rs9467471 for LRCC16A; rs836861 for PDIA5).

TABLE 5

Transcriptome-biomarker association analysis

		γ_0	γ_1	σ_ϵ	β_0	$\beta_1^*(rs6938645)$	σ_δ
Unified likelihood [§] :							
LRRC16A	estimate:	4.650	0.157	1.195	-0.435	-0.915	0.651
	std err:	0.092	0.127	—	0.153	0.0431	—
	test stat:	50.653	1.238	—	-2.851	-21.212	—
	p-value:	< 0.001	0.216	—	0.004	< 0.001	—
Regression Imputation:							
LRRC16A	estimate:	4.649	0.153	1.207	-0.431	-0.914	0.659
	std err:	0.092	0.126	—	0.155	0.0437	—
	test stat:	50.397	1.213	—	-2.789	-20.939	—
	p-value:	< 0.001	0.227	—	0.005	< 0.001	—

[§] Coefficient estimates and standard errors are based on application of the unified likelihood strategy.

* The coefficient estimate and corresponding standard error is reported for the strongest signal SNP within LRCC16A (rs6938645).