# Cross-tissue eQTL mapping in the presence of missing data via surrogate outcome analysis

Zachary R. McCaw,[1] Sheila M. Gaynor,[1] Ryan Sun,[2] Xihong Lin[1,3]

March 22nd, 2021

**Abstract**

Missing data are prevalent in the Genotype-Tissue Expression (GTEx) project, where measurements from certain inaccessible tissues, such as the substantia nigra (SSN), are available at much smaller sample sizes than those from accessible tissues, such as blood. This severely limits power for identifying tissue-specific expression quantitative trait loci (eQTL). Here we propose Surrogate Phenotype Regression Analysis (Spray) for leveraging information from a correlated surrogate outcome (e.g. expression in blood) to improve inference on a partially missing target outcome (e.g. expression in SSN). Rather than regarding the surrogate outcome as a proxy for the target outcome, Spray jointly models the target and surrogate outcomes within a bivariate regression framework. Unobserved values of either outcome are regarded as missing data. We describe and implement an expectation conditional maximization algorithm for performing estimation in the presence of bilateral outcome missingness. Spray estimates the same association parameter estimated by standard eQTL mapping, and controls the type I error even when the target and surrogate outcomes are truly uncorrelated. We demonstrate analytically and empirically, using simulations and GTEx data, that in comparison with marginally modeling the target outcome, jointly modeling the target and surrogate outcomes increases estimation precision and improves power.

**Keywords**: Asymptotic Relative Efficiency; EM Algorithm; Genetic Association Analysis; Missing Data; Multivariate Analysis.

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115; [2] Department of Biostatistics, MD Anderson Cancer Center, Houston, TX 77030; [3] Department of Statistics, Harvard University, Cambridge, MA 02138.

# 1 Introduction

Tissue-specific expression quantitative trait loci (eQTL) are of substantial biological interest as mechanisms for explaining how the genetic variants identified in genome-wide association studies (GWAS) influence complex traits and diseases (Gamazon et al., 2015; Gusev et al., 2016; Hormozdiari et al., 2016; Zhu et al., 2016; Visscher et al., 2017). Traditional eQTL studies have focused on accessible tissues such as blood (Nica et al., 2011; Westra et al., 2013), while eQTL analyses in inaccessible tissues, such as brain, have been hindered by limited sample sizes. Cross-tissue studies, including the Genotype-Tissue Expression Project (GTEx), have demonstrated that the effect sizes of eQTL are heterogeneous across tissues (Consortium, 2017). Consequently, studying only accessible tissues is insufficient to understand the genetic basis of gene regulation. Larger sample sizes are needed to provide sufficient power for reliable eQTL detection in inaccessible tissues, and there is great interest in borrowing information from accessible tissues to increase the effective sample sizes of inaccessible tissues.

The goal of improving power for eQTL mapping in the substantia nigra (SSN) motivated our work on cross-tissue eQTL mapping. SSN is a region of the midbrain implicated in the development of Parkinson's disease (Poewe et al., 2017). Due to the scarcity of gene expression data, no previous studies have focused on eQTL mapping in this region. At the time of our analysis, only 73 genotyped subjects with expression data in SSN were available from GTEx, in contrast to 369 subjects with expression in whole blood. Among subjects with expression in blood, nearly 90% were missing expression in SSN. The methodology developed here leverages the gene expression measurements from whole blood to improve power for identifying eQTL in SSN.

Existing remedies for eQTL mapping in the presence of missing data are *ad hoc*. Standard eQTL software adjusts for missingness using mean imputation or complete case analysis (Shabalin, 2012). Mean imputation, whether marginal or conditional on covariates, reduces outcome variability, leading to underestimated standard errors and invalid inference (Little and Rubin, 2002). Complete case analysis can introduce bias when missingness is related to covariates, and makes inefficient use of the data even when the missingness occurs completely at random.

In this paper, we propose improving power for eQTL mapping in an inaccessible tissue (e.g. SSN), for which expression measurements are partially missing, by augmenting the sample with expression data from an accessible surrogate tissue (e.g. whole blood), for which the available sample size is substantially larger. Specifically, we propose jointly modeling expression in the target and surrogate tissues while regarding unobserved measurements in either tissue as missing data, an approach we describe as Surrogate Phenotype Regression Analysis (Spray). Spray leverages the correlation in expression levels across tissues to increase

the effective sample size, but maintains eQTL in the target tissue as the focus of inference.

We note that the problem addressed by SPRAY is orthogonal to that considered in Surrogate Variable Analysis (SVA) (Leek and Storey, 2007; Lee et al., 2017), a method developed to identify and adjust for latent factors of variation present in microarray data, such as batch effects and technical artifacts. Importantly, the role of the surrogate phenotype in SPRAY, (e.g. gene expression of blood) is to increase power for inference on the phenotype of interest (e.g. gene expression of SSN) rather than to correct for bias. Moreover, SVA conditions on the surrogate variables (batch effects and artifacts) as covariates in subsequent regression modeling. In contrast, SPRAY models the surrogate phentoype as a component of the bivariate outcome. As discussed in Section 5.2, by regarding the surrogate phenotype an outcome rather than a covariate, we ensure that the genetic effect estimated by SPRAY coincides with that estimated by standard, marginal eQTL mapping.

For estimation, we develop a computationally efficient Expectation Conditional Maximization Either (ECME) algorithm (Meng and Rubin, 1993; Liu and Rubin, 1994), which is adapted to fitting the association model in the presence of bilateral outcome missingness. The algorithm iterates between conditional maximization of the observed data log likelihood with respect to the regression parameters, and conditional maximization of the EM objective function with respect to the covariance parameters. In addition, we have derived the asymptotic covariance of all model parameters, and implemented a flexible Wald test for evaluating hypotheses about the target regression parameters. We show analytically that the asymptotic efficiency of jointly modeling the target and surrogate outcomes, relative to marginally modeling the target outcome only, increases with the proportion of target outcomes missing and with the square of the target-surrogate correlation. We numerically demonstrate the analytical results through extensive simulations evaluating the empirical power of the SPRAY Wald test as a function of the target missingness and the target-surrogate correlation.

Compared to complete data analysis, maximum likelihood estimation as implemented by SPRAY is efficient, making full use of the available data, and provides more precise estimates of the target regression parameters. All estimation and inference procedures described in this article have been implemented in an easy-to-use `R` package (`SurrogateRegression`), which is available on CRAN (McCaw, 2020).

Using data from GTEx, we applied SPRAY to perform cross-tissue eQTL mapping, taking expression in SSN as the target outcome, and expression in whole blood as the surrogate outcome. Compared with marginal eQTL mapping based on expression in SSN only, SPRAY discovered 64.1% more significant eQTL at $\alpha = 5 \times 10^{-8}$. Importantly, while the effect sizes estimated by SPRAY were nearly identical to those obtained via marginal eQTL mapping ($R^2 = 0.998$), the sampling variance was reduced by 19.6%, on

average, suggesting that SPRAY increased power primarily by drawing on the correlated surrogate outcome to improve precision. SPRAY identified 106 more independent genomic loci than the marginal analysis, each of which contained at least one previously reported GWAS variant. We highlight a variant on chromosome 1 (rs483259) that was associated with significantly decreased expression of the LRRIQ3 gene, and is in linkage disequilibrium (LD) with variants recently associated with increased educational attainment (Lee et al., 2018).

The remainder of this paper is organized as follows: Section 2 introduces the setting and model. Sections 3 and 4 detail the estimation and inference procedures. Section 5 addresses the estimand of SPRAY, and the asymptotic relative efficiency of jointly versus marginally modeling the target outcome. Section 6 presents the results of simulation studies, and section 7 the application to GTEx. We conclude with discussions in Section 8.

## 2 Model and Setting

For each of $n$ independent subjects, suppose that two continuous outcomes are potentially observed: the *target outcome* $T_i$ and the *surrogate outcome* $S_i$. Consider the model:

$$\begin{pmatrix} T_i \\ S_i \end{pmatrix} \Big| (\boldsymbol{x}_i, \boldsymbol{z}_i) \sim N\left\{ \begin{pmatrix} \boldsymbol{x}_i' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{z}_i' \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix}, \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix} \right\}, \tag{1}$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates for the target outcome, with regression coefficient $\boldsymbol{\beta}$, and $\boldsymbol{z}_i$ is a $q \times 1$ vector of covariates for the surrogate outcome, with regression coefficient $\boldsymbol{\alpha}$. The $2 \times 2$ residual covariance matrix $\boldsymbol{\Sigma}$ is left unstructured. Let $\boldsymbol{y}_i = \text{vec}(T_i, S_i) \in \mathbb{R}^2$ denote the $2 \times 1$ outcome vector, $\mathcal{X}_i = \text{diag}(\boldsymbol{x}_i', \boldsymbol{z}_i')$ the $2 \times (p+q)$ subject-specific design matrix, and $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ the $(p+q) \times 1$ overall regression coefficient. With this notation, model (1) is succinctly expressible as: $\boldsymbol{y}_i \big| \mathcal{X}_i \sim N\big(\mathcal{X}_i \boldsymbol{\gamma}, \boldsymbol{\Sigma}\big)$.

We suppose that the target and surrogate outcomes are subject to arbitrary patterns of missingness at random (MAR). Define the target $R_{T,i}$ and surrogate $R_{S,i}$ responses indicators:

$$R_{T,i} = \begin{cases} 1, & T_i \text{ is observed,} \\ 0, & T_i \text{ is missing.} \end{cases} \qquad R_{S,i} = \begin{cases} 1, & S_i \text{ is observed,} \\ 0, & S_i \text{ is missing.} \end{cases}$$

These indicators partition the $n$ subjects into the following groups. *Complete cases* are subjects whose target and surrogate outcomes are both observed $R_{T,i} = 1$ and $R_{S,i} = 1$. Subjects with *target missingness* have $R_{T,i} = 0$ and $R_{S,i} = 1$ while subjects with *surrogate missingness* have $R_{T,i} = 1$ and $R_{S,i} = 0$. Subjects with neither outcome observed ($R_{T,i} = 0$ and $R_{S,i} = 0$) make no contribution to the likelihood function and are not considered further. Suppose there are $n_0$ complete cases, $n_1$ subjects with target missingness, and $n_2$ subjects with surrogate missingness, $n = n_0 + n_1 + n_2$.

The MAR assumption requires that observation of the target outcome ($R_{T,i}$) is unrelated to its value ($T_i$), given ($S_i, \boldsymbol{x}_i, \boldsymbol{z}_i$), among subjects with target missingness. Likewise, $R_{S,i}$ is supposed unrelated to $S_i$, given ($T_i, \boldsymbol{x}_i, \boldsymbol{z}_i$), for subjects with surrogate missingness. These assumptions allow for estimation of the model parameters via maximum likelihood as implemented by our EM algorithm. For cross-tissue eQTL mapping in GTEx, the MAR assumption is reasonable because donors were selected to be free of major diseases and the collection of tissue specimen was based on factors such as consent of the donor or next-of-kin, and on the availability of sufficient tissue from the autopsy or surgical procedure (NCI, 2013; Consortium, 2013). Importantly, the ascertainment of a tissue sample was in no way related to gene expression measurements.

# 3 Estimation

## 3.1 Regression Parameters

Define the response indicator matrix $\boldsymbol{R}_i = \text{diag}(R_{T,i}, R_{S,i})$, and note that $\boldsymbol{R}_i$ is a projection matrix. The distribution of the observed data is expressible as:

$$\boldsymbol{y}_i \big| (\boldsymbol{R}_i, \mathcal{X}_i) \sim N\big(\boldsymbol{R}_i \mathcal{X}_i \boldsymbol{\gamma}, \boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i\big),$$

and the observed data log likelihood is:

$$\begin{aligned} \ell_{\text{obs}}(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \propto & -\frac{1}{2}\sum_{i=1}^{n} \ln\det(\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i) \\ & -\frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{R}_i \boldsymbol{y}_i - \boldsymbol{R}_i \mathcal{X}_i \boldsymbol{\gamma})' (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} (\boldsymbol{R}_i \boldsymbol{y}_i - \boldsymbol{R}_i \mathcal{X}_i \boldsymbol{\gamma}). \end{aligned} \tag{2}$$

The observed data score equation for the regression parameters $\boldsymbol{\gamma}$ is:

$$\mathcal{U}_\gamma(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \equiv \frac{\partial \ell_{\text{obs}}}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \mathcal{X}_i' \boldsymbol{R}_i' (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} (\boldsymbol{R}_i \boldsymbol{y}_i - \boldsymbol{R}_i \mathcal{X}_i \boldsymbol{\gamma}).$$

Conditional on $\boldsymbol{\Sigma}$, the maximum likelihood estimator (MLE) of $\boldsymbol{\gamma}$ is the generalized least squares (GLS) estimator:

$$\hat{\boldsymbol{\gamma}}(\boldsymbol{\Sigma}) = \left\{ \sum_{i=1}^{n} \mathcal{X}_i' \boldsymbol{R}_i' (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} \boldsymbol{R}_i \mathcal{X}_i \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathcal{X}_i' \boldsymbol{R}_i' (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} \boldsymbol{y}_i \right\}. \tag{3}$$

## 3.2 Covariance Matrix

Let $\boldsymbol{\epsilon}_i = (\boldsymbol{y}_i - \mathcal{X}_i \boldsymbol{\gamma})$ denote the residual vector. The observed data score equation for $\boldsymbol{\Sigma}$ is:

$$\mathcal{U}_\Sigma(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \equiv \frac{\partial \ell_{\text{obs}}}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{R}_i (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} \boldsymbol{R}_i + \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{R}_i (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i' (\boldsymbol{R}_i \boldsymbol{\Sigma} \boldsymbol{R}_i)^{-1} \boldsymbol{R}_i.$$

However, the score equation for $\boldsymbol{\Sigma}$ does not admit a closed form. To obtain the MLE, we apply the ECME algorithm (Meng and Rubin, 1993; Liu and Rubin, 1994). Define the $2 \times 2$ *residual outer product* matrix:

$$\boldsymbol{V}_i \equiv \boldsymbol{\epsilon}_i \otimes \boldsymbol{\epsilon}_i = \begin{pmatrix} (T_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 & (T_i - \boldsymbol{x}_i'\boldsymbol{\beta})(S_i - \boldsymbol{z}_i'\boldsymbol{\alpha}) \\ (S_i - \boldsymbol{z}_i'\boldsymbol{\alpha})(T_i - \boldsymbol{x}_i'\boldsymbol{\beta}) & (S_i - \boldsymbol{z}_i'\boldsymbol{\alpha})^2 \end{pmatrix}.$$

The *complete data log likelihood* is now expressible as:

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \propto -\frac{n}{2} \ln \det(\boldsymbol{\Sigma}) - \frac{1}{2} \mathrm{tr} \left( \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \boldsymbol{V}_i \right). \tag{4}$$

The *EM objective* is the expectation of the complete data log likelihood in (4) given the observed data $\mathcal{D}_{\text{obs}}$ and the current parameter estimates $(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)})$:

$$Q(\boldsymbol{\gamma}, \boldsymbol{\Sigma} | \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}) \equiv \mathbb{E}\big\{ \ell(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) | \mathcal{D}_{\text{obs}}; \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)} \big\}. \tag{5}$$

To obtain an expression for (5), define the *working outcome vector*:

$$\hat{\boldsymbol{y}}_i^{(r)} \equiv \mathbb{E}\big(\boldsymbol{y}_i | \mathcal{D}_{\text{obs}}; \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}\big) = \begin{cases} (T_i, S_i)', & (R_{T,i} = 1) \wedge (R_{S,i} = 1), \\ (\hat{T}_i^{(r)}, S_i)', & (R_{T,i} = 0) \wedge (R_{S,i} = 1), \\ (T_i, \hat{S}_i^{(r)})', & (R_{T,i} = 1) \wedge (R_{S,i} = 0). \end{cases}$$

For complete cases, the working outcome vector is identically the observed outcome vector. For subjects with target missingness, the unobserved value of $T_i$ is replaced by its conditional expectation given the surrogate outcome and covariates:

$$\hat{T}_i^{(r)} \equiv \mathbb{E}(T_i | S_i, \mathcal{X}_i; \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}) = \boldsymbol{x}_i'\boldsymbol{\beta}^{(r)} + \big(\Sigma_{TS}\Sigma_{SS}^{-1}\big)^{(r)} \big(T_i - \boldsymbol{z}_i'\boldsymbol{\alpha}^{(r)}\big).$$

Similarly, for subjects with surrogate missingness, the unobserved value of $S_i$ is replaced by its conditional expectation give the target outcome and covariates:

$$\hat{S}_i^{(r)} \equiv \mathbb{E}(S_i | T_i, \mathcal{X}_i; \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}) = \boldsymbol{z}_i'\boldsymbol{\alpha}^{(r)} + \big(\Sigma_{ST}\Sigma_{TT}^{-1}\big)^{(r)} (T_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(r)}).$$

Let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ denote the precision matrix. Define the *working residual outer product*:

$$\begin{aligned} \hat{\boldsymbol{V}}_i^{(r)} &\equiv \mathbb{E}(\boldsymbol{V}_i | \mathcal{D}_{\text{obs}}; \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}) \\ &= \big(\hat{\boldsymbol{y}}_i^{(r)} - \mathcal{X}_i\boldsymbol{\gamma}\big) \otimes \big(\hat{\boldsymbol{y}}_i^{(r)} - \mathcal{X}_i\boldsymbol{\gamma}\big) + \begin{cases} \mathrm{diag}(0, 0), & (R_{T,i} = 1) \wedge (R_{S,i} = 1), \\ \mathrm{diag}\big(\Lambda_{TT}^{-1,(r)}, 0\big), & (R_{T,i} = 0) \wedge (R_{S,i} = 1), \\ \mathrm{diag}\big(0, \Lambda_{SS}^{-1,(r)}\big), & (R_{T,i} = 1) \wedge (R_{S,i} = 0). \end{cases} \end{aligned}$$

Note that $\hat{\boldsymbol{V}}_i^{(r)} = \hat{\boldsymbol{V}}_i(\boldsymbol{\gamma} | \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)})$ is a function of the true regression coefficient $\boldsymbol{\gamma}$, and of the current parameter estimates $(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)})$. Expressed in terms of the working residual outer product, the EM objective function is:

$$Q(\boldsymbol{\gamma}, \boldsymbol{\Sigma} | \boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}) = -\frac{n}{2} \ln \det(\boldsymbol{\Sigma}) - \frac{1}{2} \mathrm{tr} \left( \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \hat{\boldsymbol{V}}_i^{(r)} \right).$$

The EM score equation for $\boldsymbol{\Sigma}$ is:

$$\mathcal{U}_{\Sigma}(\boldsymbol{\gamma}, \boldsymbol{\Sigma}|\boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}) \equiv \frac{\partial Q}{\partial \boldsymbol{\Sigma}} = -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n} \hat{\boldsymbol{V}}_i^{(r)}\right)\boldsymbol{\Sigma}^{-1}.$$

Conditional on $\boldsymbol{\gamma}$, the EM update for $\boldsymbol{\Sigma}$ is:

$$\hat{\boldsymbol{\Sigma}}^{(r+1)}(\boldsymbol{\gamma}) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{V}}_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(r)}, \boldsymbol{\Sigma}^{(r)}). \tag{6}$$

### 3.3 Optimization

We propose the following ECME algorithm, in which the regression parameters $\boldsymbol{\gamma}$ are updated via conditional maximization of the observed data log likelihood in (2), and the covariance matrix $\boldsymbol{\Sigma}$ is updated via conditional maximization of the EM objective in (5).

**Algorithm 1** ECME for Bivariate Normal Regression

**Require:** For each subject, observed response and covariate data $(\boldsymbol{R}_i\boldsymbol{y}_i, \mathcal{X}_i)$.
**Require:** Initial estimates of the regression $\boldsymbol{\gamma}^{(0)}$ and covariance $\boldsymbol{\Sigma}^{(0)}$ parameters.
1: **repeat**
2: GLS step: Update $\boldsymbol{\gamma}^{(r+1)} \leftarrow \hat{\boldsymbol{\gamma}}(\boldsymbol{\Sigma}^{(r)})$ via (3).
3: ECM step: Update $\boldsymbol{\Sigma}^{(r+1)} \leftarrow \hat{\boldsymbol{\Sigma}}(\boldsymbol{\gamma}^{(r+1)})$ via (6).
4: Update observed data log likelihood $\ell_{\text{obs}}^{(r+1)} \leftarrow \ell_{\text{obs}}(\boldsymbol{\gamma}^{(r+1)}, \boldsymbol{\Sigma}^{(r+1)})$ via (2).
5: **until** $\ell_{\text{obs}}^{(r+1)} - \ell_{\text{obs}}^{(r)} < \epsilon$, where $\epsilon$ is the tolerance.
6: **return** Final estimates of the regression $\hat{\boldsymbol{\gamma}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ parameters.

The accompanying `R` package initializes $\boldsymbol{\gamma}$ via ordinary least squares using all observed data:

$$\boldsymbol{\gamma}^{(0)} = \left\{\sum_{i=1}^{n} \mathcal{X}_i'\boldsymbol{R}_i\mathcal{X}_i\right\}^{-1}\left\{\sum_{i=1}^{n} \mathcal{X}_i'\boldsymbol{R}_i\boldsymbol{y}_i\right\}.$$

Given $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\Sigma}$ is initialized using the residual outer product of the $n_0$ complete cases:

$$\boldsymbol{\Sigma}^{(0)} = \frac{1}{n_0}\sum_{i=1}^{n_0}(\boldsymbol{y}_i - \mathcal{X}_i\boldsymbol{\gamma}^{(0)}) \otimes (\boldsymbol{y}_i - \mathcal{X}_i\boldsymbol{\gamma}^{(0)}).$$

## 4 Inference

The ECME algorithm presented in the previous section does immediately provide the asymptotic information of the MLEs. The observed-data information matrices were obtained using the following identity:

$$\mathbb{V}\Big[\mathbb{E}\{\mathcal{U}(\boldsymbol{\theta})|\mathcal{D}_{\text{obs}}\}\Big] = \mathbb{V}\{\mathcal{U}(\boldsymbol{\theta})\} - \mathbb{E}\Big[\mathbb{V}\{\mathcal{U}(\boldsymbol{\theta})|\mathcal{D}_{\text{obs}}\}\Big],$$

where $\mathcal{U}(\boldsymbol{\theta})$ is the complete-data score, and $\mathcal{D}_{\text{obs}}$ is the observed data. The observed-data information for the regression parameters $\boldsymbol{\gamma}$ decomposes as:

$$\mathcal{I}_{\gamma\gamma'} \equiv \begin{pmatrix} \mathcal{I}_{\beta\beta'} & \mathcal{I}_{\beta\alpha'} \\ \mathcal{I}_{\alpha\beta'} & \mathcal{I}_{\alpha\alpha'} \end{pmatrix} = \mathcal{I}_{\gamma\gamma',0} + \mathcal{I}_{\gamma\gamma',1} + \mathcal{I}_{\gamma\gamma',2}. \tag{7}$$

$\mathcal{I}_{\gamma\gamma',0}$ is the contribution of complete cases and takes the form:

$$\mathcal{I}_{\gamma\gamma',0} = \sum_{i_0=1}^{n_0} \begin{pmatrix} \boldsymbol{x}_{i_0}'\Lambda_{TT}\boldsymbol{x}_{i_0} & \boldsymbol{x}_{i_0}'\Lambda_{TS}\boldsymbol{z}_{i_0} \\ \boldsymbol{z}_{i_0}'\Lambda_{ST}\boldsymbol{x}_{i_0} & \boldsymbol{z}_{i_0}'\Lambda_{SS}\boldsymbol{z}_{i_0} \end{pmatrix}.$$

$\mathcal{I}_{\gamma\gamma',1}$ is the contribution of subjects with target missingness and $\mathcal{I}_{\gamma\gamma',2}$ is the contribution of subjects with surrogate missingness; these take the following forms respectively:

$$\mathcal{I}_{\gamma\gamma',1} = \sum_{i_1=1}^{n_1} \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{z}_{i_1}'\Sigma_{SS}^{-1}\boldsymbol{z}_{i_1} \end{pmatrix}, \qquad \mathcal{I}_{\gamma\gamma',2} = \sum_{i_2=1}^{n_2} \begin{pmatrix} \boldsymbol{x}_{i_2}'\Sigma_{TT}^{-1}\boldsymbol{x}_{i_2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Complete cases contribute to the information for all regression parameters. Subjects with target missingness contribute to the information for the surrogate regression parameters $\boldsymbol{\alpha}$ only, while subjects with surrogate missingness contribute to the information for the target regression parameters $\boldsymbol{\beta}$ only.

Let $\varsigma = \text{vec}(\Sigma_{TT}, \Sigma_{TS}, \Sigma_{SS})$ vectorize the covariance parameters. The observed-data information for $\boldsymbol{\varsigma}$ likewise decomposes as:

$$\mathcal{I}_{\varsigma\varsigma'} = \begin{pmatrix} \mathcal{I}_{\varsigma_1\varsigma_1} & \mathcal{I}_{\varsigma_1\varsigma_2} & \mathcal{I}_{\varsigma_1\varsigma_3} \\ \mathcal{I}_{\varsigma_2\varsigma_1} & \mathcal{I}_{\varsigma_2\varsigma_2} & \mathcal{I}_{\varsigma_2\varsigma_3} \\ \mathcal{I}_{\varsigma_3\varsigma_1} & \mathcal{I}_{\varsigma_3\varsigma_2} & \mathcal{I}_{\varsigma_3\varsigma_3} \end{pmatrix} = \mathcal{I}_{\varsigma\varsigma',0} + \mathcal{I}_{\varsigma\varsigma',1} + \mathcal{I}_{\varsigma\varsigma',2}. \tag{8}$$

$\mathcal{I}_{\varsigma\varsigma',0}$ is the contribution of complete cases:

$$\mathcal{I}_{\varsigma\varsigma',0} = \frac{n_0}{2} \begin{pmatrix} \Lambda_{TT}^2 & 2\Lambda_{TT}\Lambda_{TS} & \Lambda_{TS}^2 \\ 2\Lambda_{TT}\Lambda_{TS} & 2\Lambda_{TS}^2 + 2\Lambda_{TT}\Lambda_{SS} & 2\Lambda_{TS}\Lambda_{SS} \\ \Lambda_{TS}^2 & 2\Lambda_{TS}\Lambda_{SS} & \Lambda_{SS}^2 \end{pmatrix}.$$

$\mathcal{I}_{\varsigma\varsigma',1}$ is the contribution of subjects with target missingness and $\mathcal{I}_{\varsigma\varsigma',1}$ is the contribution of subjects with surrogate missingness; they take the following forms respectively:

$$\mathcal{I}_{\varsigma\varsigma',1} = \frac{n_1}{2} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Sigma_{SS}^{-2} \end{pmatrix}, \qquad \mathcal{I}_{\varsigma\varsigma',2} = \frac{n_2}{2} \begin{pmatrix} \Sigma_{TT}^{-2} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

These information matrices follow a similar pattern of contributions to that noted for the regression parameters. Complete cases contribute to the information for all covariance parameters, subjects with target missingness contribute only to the surrogate variance, and subjects with surrogate missingness contribute only to the target variance.

The cross information $\mathcal{I}_{\gamma\varsigma'}$ between the regression $\boldsymbol{\gamma}$ and covariance $\boldsymbol{\varsigma}$ parameters is zero. Thus, the MLEs $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Sigma}}$ are asymptotically independent. For eQTL mapping, inference on the target regression parameter $\boldsymbol{\beta} \subseteq \boldsymbol{\gamma}$ is performed using the standard Wald test, the details of which are presented in the supplemental material. Standard errors for all model parameters are provided by the accompanying `R` package, allowing for inference on $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ as well.

# 5 Analytical Considerations

## 5.1 Marginal Interpretation of the Regression Parameter

The choice to jointly model the target and surrogate outcomes, rather than conditioning on the surrogate to predict the target, has important ramifications when interpreting the regression parameters estimated by SPRAY. For exposition, suppose (1) is the generative model, and consider the setting where the target and surrogate means each depend on genotype $g_i$ only:

$$\begin{pmatrix} T_i \\ S_i \end{pmatrix} \Big| g_i \sim N\left\{ \begin{pmatrix} g_i\beta_G \\ g_i\alpha_G \end{pmatrix}, \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix} \right\}. \tag{9}$$

The implied marginal distribution of the target outcome is:

$$T_i | g_i \sim N(g_i\beta_G, \Sigma_{TT}). \tag{10}$$

Observe that the regression parameter for genotype ($\beta_G$) from the joint model (9) is identical to that appearing in the marginal model (10). This equality is unchanged by the presence or absence of an association $\alpha_G$ between genotype $g_i$ the surrogate outcome $S_i$. Importantly, as is confirmed by our simulation studies, this implies that inference on $\beta_G$ under the joint model (1) does not depend on the value of $\alpha_G$. The same is not true of a model that conditions on the surrogate outcome. In particular, when conditioning on the surrogate outcome, the target outcome is distributed as:

$$T_i | (S_i, g_i) \sim N\big\{(\beta_G - \Sigma_{TS}\Sigma_{SS}^{-1}\alpha_G)g_i + \Sigma_{TS}\Sigma_{SS}^{-1}S_i, \Sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}\big\}.$$

Suppose that the target and surrogate outcomes are associated ($\Sigma_{TS} \neq 0$), which is a prerequisite for modeling the surrogate outcome to improve inference on $\beta_G$. Then, in a model that regresses $T_i$ on both $(S_i, g_i)$, the magnitude and direction of the regression coefficient for genotype ($\beta_G - \Sigma_{TS}\Sigma_{SS}^{-1}\alpha_G$) depends on whether and to what extent genotype is associated with the surrogate outcome (i.e. $\alpha_G$).

## 5.2 Efficiency Analysis

Consider again the genotype only model in (9). Suppose initially that all subjects are complete cases, and that the genotypes have been scaled such that: $\sum_{i_0=1}^{n_0} g_{i_0}^2 = n_0$. Under these assumptions, the *efficient information* for $\beta_G$ from (9) is:

$$\mathcal{I}_{\beta_G\beta_G|\alpha_G} = n_0\big(\Lambda_{TT} - \Lambda_{TS}\Lambda_{SS}^{-1}\Lambda_{ST}\big) = n_0\Sigma_{TT}^{-1}.$$

This is identical to the information for $\beta_G$ from the marginal model in (10). Thus, in the absence of missingness, inference on $\beta_G$ under the joint model (9) is asymptotically equivalent to inference on $\beta_G$ under the marginal model (10).

Now suppose there are $n_0$ complete cases and $n_1$ subjects with target missingness. For simplicity, assume no subjects have surrogate missingness, $n_2 = 0$. Genotypes have again been scaled, within outcome missingness groups, such that $\sum_{i_0=1}^{n_0} g_{i_0}^2 = n_0$ for complete cases and $\sum_{i_1=1}^{n_1} g_{i_1}^2 = n_1$ for subjects with target missingness. The efficient information from (9) becomes:

$$\mathcal{I}_{\beta_G\beta_G|\alpha_G} = n_0 \left\{ \Lambda_{TT} - \Lambda_{TS} \frac{n_0}{\left(n_0\Lambda_{SS} + n_1\Sigma_{SS}^{-1}\right)} \Lambda_{ST} \right\},$$

while the information for $\beta_G$ from the marginal model remains $\mathcal{I}_{\beta_G\beta_G} = n_0\Sigma_{TT}^{-1}$. The *asymptotic relative efficiency* (ARE) of inference under the joint model (9) versus inference under marginal model (10) is:

$$\text{ARE} = \frac{\mathcal{I}_{\beta_G\beta_G|\alpha_G}}{\mathcal{I}_{\beta_G\beta_G}} = \Sigma_{TT} \left\{ \Lambda_{TT} - \Lambda_{TS} \frac{n_0}{\left(n_0\Lambda_{SS} + n_1\Sigma_{SS}^{-1}\right)} \Lambda_{ST} \right\}. \tag{11}$$

To better understand (11), suppose the covariance matrix in (9) is a correlation matrix, with $\Sigma_{TT} = \Sigma_{SS} = 1$, and correlation $\Sigma_{TS} = \rho \in (-1, 1)$. The ARE simplifies to:

$$\text{ARE} = \frac{1}{1-\rho^2} \left\{ 1 - \frac{\rho^2}{1-\rho^2} \cdot \frac{n_0}{n_0(1-\rho^2)^{-1} + n_1} \right\} = \frac{1}{1-\pi_T\rho^2},$$

where $\pi_T = n_1/(n_0 + n_1)$ is the proportion of subjects with target missingness. Now, if the target and surrogate outcomes are uncorrelated ($\rho = 0$), or if there is no target missingness ($\pi_T = 0$), then the ARE is 1, and inference based on the marginal model is asymptotically equivalent to inference based on the joint model. For fixed target missingness $\pi_T$, the ARE increases monotonically in the squared target-surrogate correlation $\rho^2$. In the limit as $\rho \to 1$, the ARE is maximized at $(1-\pi_T)^{-1} = 1 + n_1/n_0$. Likewise, for fixed target-surrogate correlation $\rho^2$, the ARE increases monotonically in the target missingness $\pi_T$. In the limit as $\pi_T \to 1$, which occurs when $n_1 \to \infty$, the ARE is maximized at $(1-\rho^2)^{-1}$. Overall, the power gain attributable to jointly modeling the target and surrogate outcomes is expected to increase with the squared target-surrogate correlation $\rho^2$, and with the number of subjects with target missingness $n_1$. This demonstrates an interesting property of the surrogate model: by leveraging the target-surrogate correlation, inference on the target outcome can be improved by incorporating information from subjects whose target outcomes are missing.

# 6 Simulation Studies

## 6.1 Simulation Methods

Additively coded genotypes $g_i \in \{0, 1, 2\}$ were drawn from a binomial distribution with minor allele frequency of 25%. Covariates included age and sex: age was drawn from a gamma distribution with mean 50 and variance 10; sex was drawn from a Bernoulli distribution with proportion 1/2. To emulate population structure, the top 3 PCs of the empirical genetic relatedness matrix (GRM) were included as covariates.

For all simulations, the target $\mu_{T,i}$ and surrogate $\mu_{S,i}$ regressions each included an intercept, age, sex, and 3 genetic PCs. Fixed effect regression coefficients were selected such that the proportion of total outcome-variation explained (PVE) by age and sex was 20%, and the PVE by genetic PCs was 5%. Initially, all simulations were performed in the presence ($\alpha_G \neq 0$) and absence ($\alpha_G = 0$) of an effect for genotype on the *surrogate* outcome. However, since $\alpha_G$ had no effect on the operating characteristics of SPRAY, including both size and power, only results for $\alpha_G = 0$ are reported here. For size simulations, genotype had no effect on the target outcome ($\beta_G = 0$). For power simulations, the effect of genotype was varied to achieve heritabilities between 0.1 and 1.0% ($\beta_G = 1/\sqrt{749}$ to $\beta_G = 1/\sqrt{74}$). Given genotype $g_i$ and covariates $\boldsymbol{x}_i$, the target $T_i$ and surrogate $S_i$ outcomes were generated from the model:

$$\begin{pmatrix} T_i \\ S_i \end{pmatrix} | (g_i, \boldsymbol{x}_i) \sim N \left\{ \begin{pmatrix} g_i\beta_G + \boldsymbol{x}_i'\boldsymbol{\beta}_X \\ g_i\alpha_G + \boldsymbol{x}_i'\boldsymbol{\alpha}_X \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}.$$

The number of complete cases was fixed at $n_0 = 10^3$. The numbers of subjects ($n_1$ and $n_2$) with target and surrogate missingness were varied to achieve specified proportions of target $\pi_T = n_1/(n_0 + n_1 + n_2)$ and surrogate $\pi_S = n_2/(n_0 + n_1 + n_2)$ missingness. For instance, when $\pi_T = 0.25$ and $\pi_S = 0.25$, the strata sizes were $n_0 = 10^3$ complete cases, $n_1 = 500$ subjects with target missingness, and $n_2 = 500$ subjects with surrogate missingness. Seven (target, surrogate) missingness patterns $(\pi_T, \pi_S)$ were considered: no missingness $(0.00, 0.00)$; unilateral target missingness $\{(0.25, 0.00), (0.50, 0.00), (0.75, 0.00)\}$; and bilateral outcome missingness $\{(0.25, 0.25), (0.50, 0.25), (0.25, 0.50)\}$. For each missingness pattern, the target-surrogate correlation $\rho$ spanned $\{0.00, 0.25, 0.50, 0.75\}$. When $\rho = 0$, the target and surrogate outcomes were in fact independent.

## 6.2 Estimation

Estimation simulations were conducted to evaluate the finite sample performance of SPRAY for estimating model parameters, and the calibration of the model-based standard errors (SEs). Table 1 considers estimation of the target genetic effect $\beta_G$, the target variance $\Sigma_{TT}$, and the target-surrogate correlation $\rho$ in the absence of missingness and in the presence of unilateral missingness in the target outcome. The true genetic effect size $\beta_G \approx 0.08$ was selected such that the heritability was 0.5%. The true target variance was $\Sigma_{TT} = 1.00$. The surrogate missingness was fixed at $\pi_S = 0.00$, while the target missing $\pi_T$ and target-surrogate correlation $\rho$ were varied. The point estimate is the average across $R = 5 \times 10^7$ simulation replicates. SEs are presented as the root mean square model-based SE, accompanied by the empirical standard deviation of the simulation point estimates. In all cases, parameter estimation was essentially unbiased, and the model-based SEs, obtained from (7), agreed closely with the empirical standard deviations of the point estimates. Analogous tables for estimation of $(\beta_G, \Sigma_{TT}, \rho)$ in the presence of bilateral missingness

(S1), and for estimation of $(\alpha_G, \Sigma_{SS})$ in the presence of both unilateral and bilateral missingness (S2) are presented in the supplementary materials.

Table 1: **Target parameter estimation and standard error calibration across $R = 5 \times 10^7$ simulations in the presence of unilateral missingness**. The number of complete cases was $n_0 = 10^3$. The true regression coefficient ($\beta_G \approx 0.08$) was chosen such that the heritability of the target outcome was 0.5%. The true variances of the target and surrogate outcomes were $\Sigma_{TT} = \Sigma_{SS} = 1.00$. The surrogate missingness was fixed at $\pi_S = 0.00$. The target missingness $\pi_T$ and target-surrogate correlation $\rho$ were varied. The point estimate (EST) is the average across simulation replicates. The standard error is presented as the root mean square model-based standard error ($\text{SE}_M$), followed by the empirical standard error ($\text{SE}_E$) in parentheses, which is the standard deviation of the simulation point estimates.

| **Settings** | | $\boldsymbol{\beta_G}$ | | $\boldsymbol{\Sigma_{TT}}$ | | $\boldsymbol{\rho}$ | |
|---|---|---|---|---|---|---|---|
| $\rho$ | $\pi_T$ | EST | $\text{SE}_M$ ($\text{SE}_E$) | EST | $\text{SE}_M$ ($\text{SE}_E$) | EST | $\text{SE}_M$ ($\text{SE}_E$) |
| 0.00 | 0.00 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.00 | 0.03 (0.03) |
| 0.25 | 0.00 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.25 | 0.03 (0.03) |
| 0.50 | 0.00 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.50 | 0.04 (0.04) |
| 0.75 | 0.00 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.05) | 0.75 | 0.04 (0.04) |
| 0.00 | 0.25 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.00 | 0.03 (0.03) |
| 0.25 | 0.25 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.25 | 0.03 (0.03) |
| 0.50 | 0.25 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.50 | 0.03 (0.03) |
| 0.75 | 0.25 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.75 | 0.04 (0.04) |
| 0.00 | 0.50 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.00 | 0.03 (0.03) |
| 0.25 | 0.50 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.25 | 0.03 (0.03) |
| 0.50 | 0.50 | 0.08 | 0.05 (0.05) | 1.00 | 0.04 (0.04) | 0.50 | 0.03 (0.03) |
| 0.75 | 0.50 | 0.08 | 0.04 (0.04) | 1.00 | 0.04 (0.04) | 0.75 | 0.03 (0.03) |
| 0.00 | 0.75 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.00 | 0.03 (0.03) |
| 0.25 | 0.75 | 0.08 | 0.05 (0.05) | 0.99 | 0.04 (0.04) | 0.25 | 0.03 (0.03) |
| 0.50 | 0.75 | 0.08 | 0.05 (0.05) | 1.00 | 0.04 (0.04) | 0.50 | 0.03 (0.03) |
| 0.75 | 0.75 | 0.08 | 0.04 (0.04) | 1.00 | 0.04 (0.04) | 0.75 | 0.03 (0.03) |

## 6.3 Type I Error Simulations

Type I error simulations were conducted to assess the validity of the SPRAY Wald test. Table 2 presents the empirical type I error for assessing the $H_0 : \beta_G = 0$ in the presence of unilateral missingness, and type I error estimates in the presence of bilateral missingness are presented in supplementary table S3. For these simulations the genetic effects were set to zero ($\beta_G = 0.00$). The type I error was controlled near the nominal level, and was insensitive to either outcome missingness or target-surrogate correlation.

Supplemental figures S1-S2 demonstrate that, across outcome missingness patterns and target-surrogate correlation levels, the p-values provided by the SPRAY Wald test were uniformly distributed under the null. Thus, SPRAY provides a valid test of association between genotype and the target outcome. Neither the type I error estimates nor the distribution of p-values were affected by the presence ($\alpha_G \neq 0$) or absence ($\alpha_G = 0$) of a genetic effect on the surrogate outcome. It is important to note that here and throughout the simulations, the target-surrogate correlation was estimated even when the generative $\rho = 0$. For a given

Table 2: **Empirical type I error of the Spray Wald test across $R = 5 \times 10^7$ simulation replicates in the presence of unilateral missingness**. The number of complete cases was $n_0 = 10^3$. The surrogate missingness was fixed at $\pi_S = 0$. The genetic effects were set to zero $\beta_G = \alpha_G = 0.00$, while the outcome variances were set to one, $\Sigma_{TT} = \Sigma_{SS} = 1.00$. The target missingness $\pi_T$ and target-surrogate correlation $\rho$ were varied. The point estimate is the average across simulation replicates.

| **Settings** | | **Type I Error** | | |
|---|---|---|---|---|
| $\rho$ | $\pi_T$ | $\alpha = 0.05$ ($\times 10^2$) | $\alpha = 10^{-3}$ ($\times 10^3$) | $\alpha = 10^{-5}$ ($\times 10^5$) |
| 0.00 | 0.00 | 5.10 | 1.08 | 1.09 |
| 0.25 | 0.00 | 5.10 | 1.08 | 1.21 |
| 0.50 | 0.00 | 5.09 | 1.07 | 1.09 |
| 0.75 | 0.00 | 5.08 | 1.07 | 1.20 |
| 0.00 | 0.25 | 5.10 | 1.07 | 1.17 |
| 0.25 | 0.25 | 5.10 | 1.08 | 1.20 |
| 0.50 | 0.25 | 5.09 | 1.06 | 1.06 |
| 0.75 | 0.25 | 5.08 | 1.05 | 1.07 |
| 0.00 | 0.50 | 5.11 | 1.07 | 1.14 |
| 0.25 | 0.50 | 5.11 | 1.07 | 1.18 |
| 0.50 | 0.50 | 5.10 | 1.08 | 1.14 |
| 0.75 | 0.50 | 5.08 | 1.06 | 1.16 |
| 0.00 | 0.75 | 5.12 | 1.09 | 1.22 |
| 0.25 | 0.75 | 5.12 | 1.07 | 1.22 |
| 0.50 | 0.75 | 5.11 | 1.07 | 1.12 |
| 0.75 | 0.75 | 5.09 | 1.06 | 1.06 |

realization of the data, the MLE $\hat{\rho}$ will differ from 0 even when in truth $\rho = 0$. The type I error simulations demonstrate that this spurious estimated correlation ($\hat{\rho} \neq 0$) did not compromise inference on $\beta_G$.

## 6.4 Power Simulations

Power simulations were conducted to assess the probability of correctly rejecting $H_0 : \beta_G = 0$ at $\alpha = 0.05$. Figure 1 presents power curves at $n_0 = 10^3$ in the presence of unilateral missingness. Power curves under bilateral missingness are presented in supplementary Figure S3. The genetic effect on the target outcome $\beta_G$ was varied between 0.037 and 0.14 to achieve heritabilities ranging from 0.1% to 1.0%, while the genetic effect on the surrogate outcome $\alpha_G$ was fixed at zero. In the absence of missingness, no additional power was gained by modeling the surrogate outcome. As the target missingness increased, the power of the Spray Wald test increased in correspondence with the target-surrogate correlation.

## 6.5 Empirical Relative Efficiency

Relative efficiency (RE) simulations were conducted comparing the Spray estimator $\hat{\beta}_G^{\text{Spray}}$ of $\beta_G$, which uses all available data, with the marginal estimator $\hat{\beta}_G^{\text{Marginal}}$ from the model:

$$T_i \big| (g_i, \boldsymbol{x}_i) \sim N\big(g_i \beta_G + \boldsymbol{x}_i' \boldsymbol{\beta}_X, \Sigma_{TT}\big),$$

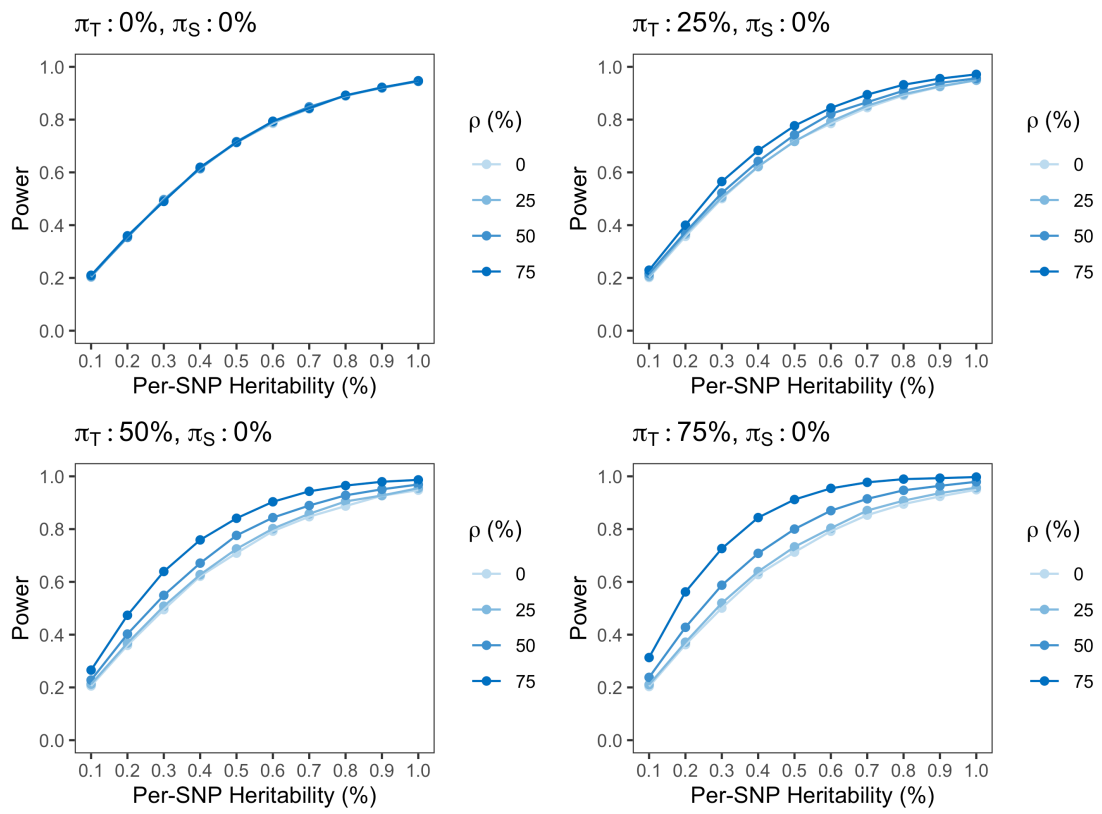


Figure 1: **Power curves for the Spray test of association in the presence of unilateral missingness**. The number of complete cases was $n_0 = 10^3$, and the type I error was $\alpha = 0.05$. Each point on the curve is the average across $R = 5 \times 10^5$ simulation replicates. The standard errors of the point estimates were negligible. The target regression coefficient $\beta_G$ was varied between 0.037 and 0.14 to achieve heritabilities between 0.1% and 1.0%, while the surrogate regression coefficient $\alpha_G$ was fixed at zero. The surrogate missingness was held at $\pi_S = 0$, while the target missingness $\pi_T$ and target-surrogate correlation $\rho$ were varied.

which uses subjects with observed target outcomes only. Note that $\hat{\beta}_G^{\text{Marginal}}$ is the estimator that would conventionally be used in the absence of a surrogate outcome. These simulations quantify the efficiency gain attributable to incorporating information from the surrogate. Table 3 compares the empirical variances of $\hat{\beta}_G^{\text{Spray}}$ and $\hat{\beta}_G^{\text{Marginal}}$ in the presence of unilateral missingness, while supplemental Table S4 compares the empirical variances under bilateral missingness.

In the absence of target missingness ($\pi_T = 0$), or when the target-surrogate correlation was zero ($\rho = 0$), the empirical RE was 1.0, as predicted by (11). Thus, while jointly modeling the target and surrogate outcomes is unnecessary in the absence of missingness, power was not substantially diminished by modeling an uninformative surrogate. In the presence of missingness, modeling an uninformative surrogate ($\rho = 0$) did not spuriously inflate the RE. As the target missingness ($\pi_T$) and target-surrogate correlation ($\rho$) increased, the empirical RE increased as predicted by (11). The precise agreement between the empirical

and theoretical REs suggests that equation (11) could prove useful at the study design stage for assessing the expected efficiency gain provided by SPRAY.

Table 3: **Empirical relative efficiency comparing the Spray estimator to the marginal estimator of $\beta_G$ test across $R = 5 \times 10^7$ simulation replicates in the presence of unilateral missingness**. The number of complete cases was $n_0 = 10^3$. The true regression coefficient ($\beta_G \approx 0.08$) was chosen such that the heritability of the target outcome was 0.5%. The true variances of the target and surrogate outcomes were $\Sigma_{TT} = \Sigma_{SS} = 1.00$. The surrogate missingness was fixed at $\pi_S = 0.00$. The target missingness $\pi_T$ and target-surrogate correlation $\rho$ were varied. Variance refers to the empirical variance of the corresponding estimator across simulation replicates. The empirical RE is the ratio of the variance of $\hat{\beta}_G^{\text{Spray}}$ to that of $\hat{\beta}_G^{\text{Marginal}}$. The theoretical RE was obtained from (11).

| **Settings** | | **Variance** | | **Relative Efficiency** | |
|---|---|---|---|---|---|
| $\rho$ | $\pi_T$ | Marginal | SPRAY | Empirical | Theoretical |
| 0.00 | 0.00 | 0.0027 | 0.0027 | 1.0000 | 1.0000 |
| 0.25 | 0.00 | 0.0027 | 0.0027 | 1.0001 | 1.0000 |
| 0.50 | 0.00 | 0.0027 | 0.0027 | 1.0005 | 1.0000 |
| 0.75 | 0.00 | 0.0027 | 0.0027 | 1.0011 | 1.0000 |
| 0.00 | 0.25 | 0.0027 | 0.0027 | 0.9997 | 1.0000 |
| 0.25 | 0.25 | 0.0027 | 0.0026 | 1.0158 | 1.0159 |
| 0.50 | 0.25 | 0.0027 | 0.0025 | 1.0672 | 1.0667 |
| 0.75 | 0.25 | 0.0027 | 0.0023 | 1.1657 | 1.1636 |
| 0.00 | 0.50 | 0.0027 | 0.0027 | 0.9995 | 1.0000 |
| 0.25 | 0.50 | 0.0027 | 0.0026 | 1.0318 | 1.0323 |
| 0.50 | 0.50 | 0.0027 | 0.0023 | 1.1431 | 1.1429 |
| 0.75 | 0.50 | 0.0027 | 0.0019 | 1.3931 | 1.3913 |
| 0.00 | 0.75 | 0.0027 | 0.0027 | 0.9992 | 1.0000 |
| 0.25 | 0.75 | 0.0027 | 0.0026 | 1.0485 | 1.0492 |
| 0.50 | 0.75 | 0.0027 | 0.0022 | 1.2305 | 1.2308 |
| 0.75 | 0.75 | 0.0027 | 0.0016 | 1.7303 | 1.7297 |

# 7 Application to eQTL Mapping in GTEx

## 7.1 Material and Methods

### 7.1.1 Data Preparation

Genotypes and RNA transcript expression were obtained from the NHGRI Genotype Tissue Expression (GTEx) Project, Version 7 (phs000424.v7.p2). The sample consisted of subjects with expression in at least one of the substantia nigra (SSN) or whole blood. The overall sample consisted of $n = 395$ genotyped subjects. Among these, there were $n_0 = 47$ (12%) complete cases, $n_1 = 322$ (82%) subjects with target missingness, and $n_2 = 26$ (7%) subjects with surrogate missingness. SNPs were excluded from the analysis if the per-SNP missingness exceeded 10% or if the minor allele frequency fell below 5%. The top 3 PCs of the empirical GRM were calculated with `PLINK` (Purcell et al., 2007). To reduce multiple testing, variants in LD ($R^2 > 0.8$) were greedily pruned within a 1 Mb sliding window using PLINK's `clump` command. A transcript was considered *expressed* if the raw read count exceeded 5 for at least 20% of subjects in both

SSN and in blood. For sequences with multiple isoforms, only the isoform with the greatest variation in expression across subjects was retained. Expression levels were normalized within tissues via the rank-based inverse normal transformation (INT) (McCaw et al., 2020), ensuring that the marginal distributions of the target and surrogate outcomes were each normal.

### 7.1.2 Association Testing

Association tests were performed for all variants in *cis* to the transcript's expressed sequence, defined as residing in the region spanning from 1 Mb upstream of the transcription start site to 1 Mb downstream of the transcription end site. Across the 17,007 expressed transcripts, the number of loci in *cis* ranged from 24 to 4,834, with a median of 887. Following GTEx (Consortium, 2017), a separate test of association was conducted for each *cis*-SNP by expressed transcript pairing. Covariates included age, sex, genotyping platform, and the top 3 genetic PCs.

For the marginal analysis, transcript expression in SSN $T_i$ was regressed on genotype, covariates, and genetic PCs, according to the association model:

$$T_i = \beta_0 + \beta_G g_i + \beta_A \text{Age}_i + \beta_S \text{Sex}_i + \beta_P \text{Platform}_i + \boldsymbol{\beta}'_{PC} \text{PCs}_i + \epsilon_{T,i}, \tag{12}$$

where $\epsilon_{T,i}$ is a residual with mean zero and finite variance. The target of inference was $\beta_G$. The null hypothesis $H_0 : \beta_G = 0$ was assessed using the standard Wald test, as implemented by base `R` (Team, 2013). Only the $n_0 + n_2 = 73$ subjects with observed expression in SSN were included in the marginal analysis. Note that expression in blood $S_i$ was not included in the marginal analysis because conditioning on $S_i$ in the regression model would change the interpretation of $\beta_G$ (see Section 5.1).

For the joint analysis, expression in SSN $T_i$ and in blood $S_i$ were jointly regressed on genotype, covariates, and genetic PCs, according to the SPRAY model:

$$\begin{pmatrix} T_i \\ S_i \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_G g_i + \beta_A \text{Age}_i + \beta_S \text{Sex}_i + \beta_P \text{Platform}_i + \boldsymbol{\beta}'_{PC} \text{PCs}_i \\ \alpha_0 + \alpha_G g_i + \alpha_A \text{Age}_i + \alpha_S \text{Sex}_i + \alpha_P \text{Platform}_i + \boldsymbol{\alpha}'_{PC} \text{PCs}_i \end{pmatrix} + \begin{pmatrix} \epsilon_{T,i} \\ \epsilon_{S,i} \end{pmatrix},$$

where $(\epsilon_{T,i}, \epsilon_{S,i}) = \boldsymbol{\epsilon}_i \overset{\text{IID}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$. Estimation and inference were performed using the accompanying `R` package (McCaw, 2020). ECME iterations continued until the improvement in the observed data log likelihood dropped below $10^{-8}$. The null hypothesis $H_0 : \beta_G = 0$ was evaluated using the SPRAY Wald test. All $n = 395$ subjects with expression in at least one of SSN or blood were included in the joint analysis.

## 7.2 Results

### 7.2.1 Discovery and Efficiency Gains

Genome-wide significance was declared at $\alpha = 5 \times 10^{-8}$ (Pe'er et al., 2008). Marginal association analysis, using observed expression in the SSN only, identified 294 significant associations, occurring between 90

unique transcripts and 272 unique genetic variants. Joint association analysis via SPRAY, using expression in both SSN and blood, identified 820 significant associations, occurring between 223 unique transcripts and 745 unique genetic variants.

Figure 2 compares the estimated effect sizes, p-values, and empirical relative efficiency of SPRAY versus the marginal association test at *cis*-eQTL that achieved genome-wide significance in either analysis. All SNPs that reached significance in the marginal analysis were also significant according to SPRAY, but not conversely. Panel A demonstrates that the estimated effect sizes were highly concordant ($R^2 = 0.998$), while panel B demonstrates that the SPRAY test of association was as or more significant than the marginal test. Taken together, these observations verify that SPRAY and the marginal association analysis estimate the same underlying genetic effect, and indicate that the power gain realized by SPRAY is achieved primarily through increasing the estimation precision. Indeed, panel C demonstrates that the empirical relative efficiency, as the ratio of $\mathbb{V}(\hat{\beta}_G^{\text{Marginal}})$ to $\mathbb{V}(\hat{\beta}_G^{\text{Spray}})$, always exceeded 1, and as predicted by (11), increased with the square of the correlation ($r^2$) between expression in SSN and expression in whole blood.

Table 4 presents the discovery and average efficiency gains of SPRAY over the marginal association analysis. The *discovery gain* was defined as the number of significant eQTL uniquely detected by SPRAY relative to the number of significant eQTL detected by either analysis. Among the 820 eQTL identified by either analysis, all 820 were detected by SPRAY, and 526 were detected by SPRAY alone. The discovery gain was thereby 64.1% (= 526/820). The *average efficiency gain* was defined as the ratio of the average sampling variance of the marginal estimated $\overline{\mathbb{V}}(\hat{\beta}_G^{\text{Marginal}})$, across the 820 significant eQTL, to the average sampling variance of the SPRAY estimator, $\overline{\mathbb{V}}(\hat{\beta}_G^{\text{Marginal}})$, minus 1. The average efficiency gain was 19.6% ($= 0.024/0.020 - 1$).

Table 4: **Discovery and average efficiency gains of Spray relative to the marginal association analysis**. Significant associations were declared at ($\alpha = 5 \times 10^{-8}$). The discovery gain is the number of associations uniquely identified by SPRAY relative to the number of associations identified by either analysis. The efficiency gain is one less than the empirical relative efficiency of the SPRAY estimator to the marginal estimate of the genetic effect of averaged over the 820 significant eQTL identified by SPRAY.

| **Significant Associations** | | |
|---|---|---|
| SPRAY | Marginal | Discovery Gain (%) |
| 820 | 294 | 64.1 |
| **Average Estimator Variance** | | |
| SPRAY | Marginal | Efficiency Gain (%) |
| 0.020 | 0.024 | 19.6 |

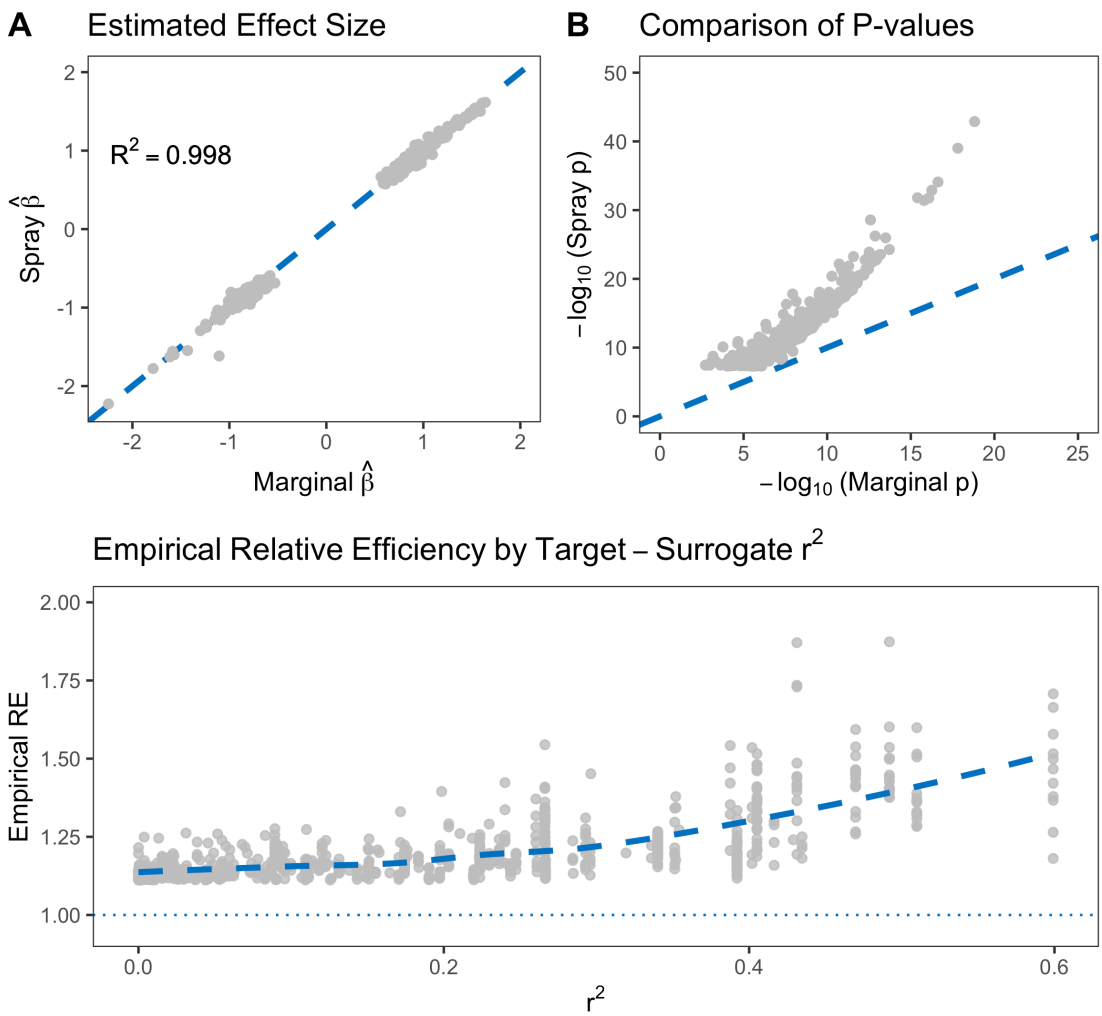


Figure 2: **Comparison of effect sizes, p-values, and empirical relative efficiency of Spray versus the marginal association analyses at genome-wide significant *cis*-eQTL in SSN.** The marginal association test uses observed expression in SSN only, while the joint association test (Spray) uses expression in SSN and in blood. Panel A compares the effect sizes estimated by Spray with those estimated by the marginal analysis, while panel B compares the p-values. Panel C plots the empirical relative efficiency, the ratio of $\mathbb{V}(\hat{\beta}_G^{\text{Marginal}})$ to $\mathbb{V}(\hat{\beta}_G^{\text{Spray}})$, as a function of the square of the correlation between expression in SSN and expression in blood.

#### 7.2.2 Functional Analysis

The biological role of the candidate eQTL identified by Spray was evaluated using FUMA (Watanabe et al., 2017). At the recommended LD threshold of $R^2 \geq 0.6$, the 745 unique genetic variants identified by Spray resolved to 361 independent significant associations, belonging to 183 distinct genomic loci. Meanwhile, the 272 unique genetic variants identified by the marginal analysis collapsed to only 122 independent significant associations, belonging to 77 distinct genomic loci. All 77 genomic loci associated with gene expression in SSN by the marginal analysis were also identified by Spray. Each of the 106 ($= 183 - 77$) loci identified

by SPRAY only harbored at least 1 previously reported GWAS variant (MacArthur et al., 2017). An interesting example is a locus on chromosome 1, spanning base pairs 74,476,321 to 74,710,629. The lead SNP (1_74710629_C_T_b37; alias rs483259) was significantly associated with decreased expression of the protein-coding LRRIQ3 gene ($\hat{\beta}_G = -0.85$; 95% CI -1.12 to -0.59; $p = 2 \times 10^{-10}$) via SPRAY. This SNP is in LD with two nearby SNPs, rs10789387 ($R^2 = 0.66$) and rs11210400 ($R^2 = 0.66$), both of which were recently associated with increased educational attainment, at $p = 7 \times 10^{-16}$ and $p = 2 \times 10^{-17}$, in a meta-analysis of $n > 1.1 \times 10^6$ subjects (Lee et al., 2018). Further work is needed to investigate whether the association of the meta-analysis SNPs with educational attainment is partially mediated by decreased expression of the LRRIQ3 gene.

# 8 Discussion

We have proposed Surrogate Phenotype Regression Analysis (SPRAY) as a novel approach to eQTL mapping in the presence of missing data, and derived a computationally efficient, ECME-type algorithm for fitting the underlying association model. We have demonstrated analytically and empirically, though extensive simulations and in real data, that the SPRAY test of association, which incorporates information from the target and surrogate outcomes, is more powerful than the marginal test of association, which incorporates information from the target outcome only.

We derived and validated a simple formula for assessing the efficiency gain achieved by SPRAY as compared with marginal association analysis. This efficiency gain increases with the target missingness, and with the square of the target-surrogate correlation. We applied SPRAY to eQTL mapping in GTEx, using expression in SSN as the target outcome and expression in blood as the surrogate outcome. SPRAY identified 64.1% more significant *cis*-eQTL than marginal eQTL mapping. Although the joint and marginal effect size estimates were highly concordant ($R^2 = 0.998$), the association statistics estimated by SPRAY had 19.6% greater precision on average. All estimation and inference procedures described in this article have been implemented in the accompanying `R` package (McCaw, 2020).

The problem of cross-tissue eQTL mapping has not previously been studied from the missing data perspective. Established software for eQTL mapping, such as `MatrixEQTL` (Shabalin, 2012), either address missing data via complete case analysis, which is inefficient, or via mean imputation, which invalidates inference. Although sophisticated methods exist for imputing gene expression across tissues (Wang et al., 2016), inference after single imputation leads to inflated type I error (Little and Rubin, 2002), and multiple imputation is computationally infeasible when performing millions of association tests at $\alpha = 5 \times 10^{-8}$. SPRAY fills this gap by providing a valid, powerful, and computationally tractable approach to eQTL

mapping in the presence of missing data.

Our work suggests several areas for further improvement. Although the inverse normal transformation was applied to ensure the marginal distributions of the target and surrogate outcomes were each normal, bivariate normality of the two outcomes is not guaranteed. A future direction is to develop association tests that allow for arbitrary patterns of outcome missingness but do not require specification of joint distribution of the target and surrogate outcomes. Instead of maximum likelihood based estimation, this procedure could use a set of inverse probability weighted estimating equations (Robins et al., 1995). Another avenue for future development is to incorporate multiple surrogate outcomes. One way to achieve this would be to extend the bivariate normal regression framework to a multivariate normal regression framework. However, there are drawbacks to directly modeling multiple surrogate outcomes: the number of nuisance covariance parameters increases quadratically with the number of surrogates, while the number of potential missingness patterns increases exponentially. Finally, although the current work was motivated by eQTL mapping, the idea of leveraging a surrogate outcome to improve inference on a partially missing target outcome is broadly applicable. For example, in large cohort studies such as the UK Biobank (Allen et al., 2014), the target outcome may be an incompletely ascertained phenotype, such as the concentration of a biomarker only measured for a subset of participants, while the surrogate outcome may be a readily ascertained phenotype, such as diagnostic codes from electronic health records.

## Acknowledgments

This work was supported by the National Institutes of Health grants R35 CA197449 and F31 HL140822 (to Z.M.) and R35 CA197449, P01 CA134294, U01 HG009088, U19 CA203654, and R01 HL113338 (to X.L.).

## References

Allen, N. E., Sudlow, C., Peakman, T., Collins, R., et al. (2014). Uk biobank data: come and get it.

Consortium, G. (2013). The genotype-tissue expression (gtex) project. *Nature* page 580–585.

Consortium, G. (2017). Genetic effects on gene expression across human tissues. *Nature* **550,** 204.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47,** 1091.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma,

D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48,** 245.

Hormozdiari, F., van de Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics* **99,** 1245–1260.

Lee, J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* **50,** 1112–1121.

Lee, S., Sun, W., Wright, F., and Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104,** 303–316.

Leek, J. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3,** 1724–1735.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, 2nd edition.

Liu, C. and Rubin, D. (1994). The ecme algorithm: A simple extension of em and ecm with faster monotone convergence. *Biometrika* **81,** 633–648.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Research* **45,** D896–D901.

McCaw, Z. (2020). *SurrogateRegression: Surrogate Outcome Regression Analysis*. `https://CRAN.R-project.org/package=SurrogateRegression`.

McCaw, Z., Lane, J., Saxena, R., Redline, S., and Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76,** 1262–1272.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika* **80,** 267–278.

NCI (2013). Gtex biobank donors. Accessed: 2021-03-15.

Nica, A., Parts, L., Glass, D., Nisbet, J., Barrett, A., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: The muther study. *PLoS Genetics* **7,** e1002003.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology* **32,** 381–385.

Poewe, W., Seppi, K., Tanner, C., Halliday, G., Brundin, P., et al. (2017). Parkinson disease. *Nature Reviews Disease Primers* **3,** 1–21.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81,** 559–575.

Robins, J., Rotnitzky, A., and Zhou, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90,** 106–121.

Shabalin, A. A. (2012). Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358.

Team, R. C. (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics* **101,** 5–22.

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., Cox, N. J., Nicolae, D. L., and Chen, L. S. (2016). Imputing gene expression in uncollected tissues within and beyond gtex. *The American Journal of Human Genetics* **98,** 697–708.

Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with fuma. *Nature Communications* **8,** 1826.

Westra, H., Peters, M., Esko, T., Yaghootkar, H., Schurmann, C., et al. (2013). Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature Genetics* **45,** 1238–1243.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M., et al. (2016). Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature Genetics* **48,** 481–487.

## Supporting Information

Additional simulation results are available in the Support Information published online. `SPRAY` is available at `https://CRAN.R-project.org/package=SurrogateRegression`. Links to this and related software may also be found at `https://content.sph.harvard.edu/xlin/software.html`. Code for reproducing all simulations and summary statistics from the GTEx analysis are available at: `https://github.com/zrmacc/Surrogate-Replication-eQTL`.