**SUPPORTING TEXT: Jeffrey T Leek and John D Storey (2008) A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences, USA*.**

# A Proofs of Theoretical Results

**Proof of Proposition 1.** Theorem 3 from Rüschendorf and de Valk (1993) (1) shows that for a set of random variables $Z_1, Z_2, \ldots, Z_m$ with arbitrary dependence, there exist non-random functions $g_i$ such that $Z_i = g_i(Z_1, Z_2, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_m, V_i)$ a.s., $i = 1, \ldots, m$, where $V_1, V_2, \ldots, V_m$ are jointly independent $Uniform(0,1)$ random variables. We utilize this result to write each $\mathbf{e}_i$ as a function of all other $\mathbf{e}_j$, $j \neq i$, and an independent random variable. We first apply this result to each column of $\mathbf{E}$. That is, by Theorem 3 from Rüschendorf and de Valk (1993) (1) it follows that there exist non-random functions $f_{ij}$ such that

$$e_{ij} = f_{ij}(e_{1j}, e_{2j}, \ldots, e_{i-1,j}, e_{i+1,j}, \ldots, e_{mj}, u^*_{ij}) \text{ a.s.}$$

for $i = 1, \ldots, m$ where $u^*_{1j}, u^*_{2j}, \ldots, u^*_{mj}$ are jointly independent $Uniform(0,1)$ randomly variables. Combining these across the columns of $\mathbf{E}$ there exist non-random functions $f_i$ such that

$$\mathbf{e}_i = f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m, \mathbf{u}^*_i) \text{ a.s.}$$

for $i = 1, \ldots, m$ where $\mathbf{u}^*_1, \mathbf{u}^*_2, \ldots, \mathbf{u}^*_m$ are jointly independent $n$-vectors Uniformly distributed on $[0,1]^n$.

Let $\sigma_{(-i)} = \sigma(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m)$ be the sigma-algebra generated from all $\mathbf{e}_j$, $j \neq i$, and let $F_{(-i)}$ be the probability distribution function on $\sigma_{(-i)}$ with respect to Borel measure. We construct $\mathbf{u}_i$ as follows:

$$\mathbf{u}_i = \int \mathbf{e}_i dF_{(-i)} = \int f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m, \mathbf{u}^*_i) dF_{(-i)}.$$

Since $\mathbf{u}^*_1, \mathbf{u}^*_2, \ldots, \mathbf{u}^*_m$ are jointly independent and $\mathbf{u}_i = t_i(\mathbf{u}^*_i)$, where $t_i(\mathbf{u}^*_i) = \int f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m, \mathbf{u}^*_i) dF_{(-i)}$, it follows that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ are jointly independent. Let $\mathbf{U}$ be the $m \times n$ matrix where row $i$ is $\mathbf{u}_i$, and let $\mathbf{M} = \mathbf{E} - \mathbf{U}$. Then model (a) can be rewritten as $\mathbf{X} = \mathbf{BS} + \mathbf{M} + \mathbf{U}$. Since $\mathbf{M}$ is a matrix of dimension $m \times n$ with $n < m$, there exist an $m \times r$ matrix $\mathbf{\Gamma}$ and an $r \times n$ matrix $\mathbf{G}$ (where $r$ is the column rank of $\mathbf{M}$), such that $\mathbf{M} = \mathbf{\Gamma G}$. Specifically, we take the singular value decomposition $\mathbf{M} = \mathbf{ADV}^T$, and let $\mathbf{\Gamma} = \mathbf{AD}$ and $\mathbf{G} = \mathbf{V}^T$. Thus, $\mathbf{X} = \mathbf{BS} + \mathbf{\Gamma G} + \mathbf{U}$, where the rows of $\mathbf{U}$ are independent from one another.

Since $\mathbf{u}_i = \int \mathbf{e}_i dF_{(-i)}$, it follows that $\mathbf{u}_i = h_i(\mathbf{e}_i)$, where $h_i$ is a non-random function. This shows the existence of the decomposition where $\mathbf{U}$ is a function of $\mathbf{E}$ and only of $\mathbf{E}$. Finally, we will show that $\Pr(\mathbf{u}_i \neq \mathbf{0}) > 0$. If $\mathbf{u}_i = \mathbf{0}$ a.s., then $\mathbf{u}_i = t_i(\mathbf{u}_i^*) = \int f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m, \mathbf{u}_i^*) dF_{(-i)} = \mathbf{0}$ a.s. Since $\mathbf{u}_i^*$ is distributed Uniform on $[0,1]^n$, it follows that $t_i(\cdot) = \mathbf{0}$ a.s. This implies that $\mathbf{e}_i = f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m, \mathbf{u}_i^*) = f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m)$ a.s., which is a contradiction of the assumption that there exists no Borel measurable function $f_i$ such that $\mathbf{e}_i = f_i(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m)$ a.s. Thus, it holds that $\Pr(\mathbf{u}_i \neq \mathbf{0}) > 0$. This shows that the decomposition holds where $\mathbf{U} \neq \mathbf{0}$.

*See the following section* **Remarks on Proposition 1** *for some further remarks on this theorem and proof.*

**Proof of Corollary 1.** Given $\mathbf{Y}$ and $\mathbf{G}$, $\mathbf{B}$ and $\mathbf{\Gamma}$ are fixed. Therefore,

$$\begin{aligned}
\Pr(\mathbf{x}_1, \ldots, \mathbf{x}_m | \mathbf{Y}, \mathbf{G}) &= \Pr(\mathbf{u}_1, \ldots, \mathbf{u}_m | \mathbf{Y}, \mathbf{G}) \\
&= \Pr(\mathbf{u}_1 | \mathbf{Y}, \mathbf{G}) \times \cdots \times \Pr(\mathbf{u}_m | \mathbf{Y}, \mathbf{G}) \\
&= \Pr(\mathbf{x}_1 | \mathbf{Y}, \mathbf{G}) \times \cdots \times \Pr(\mathbf{x}_m | \mathbf{Y}, \mathbf{G}).
\end{aligned}$$

**Proof of Proposition 2.** Let $\mathbf{W} = \begin{pmatrix} \mathbf{S} \\ \mathbf{G} \end{pmatrix}$ and $\mathbf{P}_w = \mathbf{I} - \mathbf{W}^T \left(\mathbf{W}\mathbf{W}^T\right)^{-1} \mathbf{W}$. For the $i$th hypothesis test's data we can write model (b) as:

$$\mathbf{x}_i = (\mathbf{b}_i \;\; \boldsymbol{\gamma}_i)\mathbf{W} + \mathbf{u}_i.$$

The residuals are:

$$\begin{aligned}
\mathbf{r}_i &= \mathbf{x}_i \mathbf{P}_w \\
&= (\mathbf{b}_i \mathbf{S} + \boldsymbol{\gamma}_i \mathbf{G} + \mathbf{u}_i)\mathbf{P}_w \\
&= \mathbf{u}_i \mathbf{P}_w.
\end{aligned}$$

Since the $\mathbf{u}_i$ are independent across rows, the $\mathbf{r}_i$ are as well given $\mathbf{S}$ and $\mathbf{G}$. The estimates for $\mathbf{b}_i$ and $\boldsymbol{\gamma}_i$ are:

$$\begin{aligned}
(\widehat{\mathbf{b}}_i \;\; \widehat{\boldsymbol{\gamma}}_i) &= \left[(\mathbf{b}_i \;\; \boldsymbol{\gamma}_i)\mathbf{W} + \mathbf{u}_i\right] \mathbf{W}^T \left(\mathbf{W}\mathbf{W}^T\right)^{-1} \\
&= (\mathbf{b}_i \;\; \boldsymbol{\gamma}_i) + \mathbf{u}_i \mathbf{W}^T \left(\mathbf{W}\mathbf{W}^T\right)^{-1}
\end{aligned}$$

Since each $\widehat{\mathbf{b}}_i$ is a function of only $\mathbf{b}_i$ and $\mathbf{u}_i$, it follows that the estimates $\widehat{\mathbf{b}}_i$ are independent across tests. To show estimation-level multiple testing independence write:

$$\begin{aligned}
\mathbf{x}_i &= [\mathbf{x}_i - \widehat{\mathbf{x}}_i] + \widehat{\mathbf{x}}_i \\
&= \left[(\mathbf{b}_i \ \boldsymbol{\gamma}_i)\mathbf{W} + \mathbf{u}_i - (\mathbf{b}_i \ \boldsymbol{\gamma}_i)\mathbf{W} - \mathbf{u}_i\mathbf{W}^T\left(\mathbf{W}\mathbf{W}^T\right)^{-1}\mathbf{W}\right] + (\widehat{\mathbf{b}}_i \ \widehat{\boldsymbol{\gamma}}_i)\mathbf{W} \\
&= \mathbf{u}_i\left[\mathbf{I} - \mathbf{W}^T\left(\mathbf{W}\mathbf{W}^T\right)^{-1}\mathbf{W}\right] + (\widehat{\mathbf{b}}_i \ \widehat{\boldsymbol{\gamma}}_i)\mathbf{W} \\
&= \mathbf{g}(\mathbf{u}_i)
\end{aligned}$$

Conditional on $\widehat{\mathbf{b}}_i$, $\mathbf{S}$, $\widehat{\boldsymbol{\gamma}}_i$, and $\mathbf{G}$, the only random component of $\mathbf{g}(\mathbf{u}_i)$ is $\mathbf{u}_i$. It follows that

$$\begin{aligned}
& \Pr\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}\right) \\
= \ & \Pr\left[\mathbf{g}(\mathbf{u}_1), \ldots, \mathbf{g}(\mathbf{u}_m) | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}\right] \\
= \ & \Pr\left[\mathbf{g}(\mathbf{u}_1) | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}\right] \times \cdots \times \Pr\left[\mathbf{g}(\mathbf{u}_m) | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}\right] \\
= \ & \Pr\left(\mathbf{x}_1 | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}\right) \times \cdots \times \Pr\left(\mathbf{x}_m | \widehat{\mathbf{B}}, \mathbf{S}, \widehat{\boldsymbol{\Gamma}}, \mathbf{G}\right).
\end{aligned}$$

The results for the residuals and parameter estimates under the null hypothesis constrained model fits follow analogously. It should be noted that $\mathbf{S}$ can be modified based on $\Omega_0$ so that it restricts $\widehat{\mathbf{b}}_i \in \Omega_0$. This restricted $\mathbf{S}$ is a deterministic adjustment to the original $\mathbf{S}$, so the results follow as above.

**Proof of Corollary 2.** Let $\widehat{\mathbf{b}}_i$ and $\widehat{\mathbf{b}}_i^0$ be the parameter estimates under the unconstrained and null hypothesis constrained model fits, respectively. Let $\mathbf{r}_i$ and $\mathbf{r}_i^0$ be the residuals under the unconstrained and null hypothesis constrained model fits, respectively. By repeating the proof for Proposition 2, one can also show that $\widehat{\boldsymbol{\gamma}}_i$ and $\widehat{\boldsymbol{\gamma}}_i^0$ are independent across tests. For any fixed function, $f\left(\widehat{\mathbf{b}}_i, \widehat{\mathbf{b}}_i^0, \mathbf{r}_i, \mathbf{r}_i^0, \widehat{\boldsymbol{\gamma}}_i, \widehat{\boldsymbol{\gamma}}_i^0\right)$, it follows that these values are independent across tests. The test-statistics and p-values are special cases of such functions.

# B Remarks on Proposition 1

**Remark 1.** If $\mathbf{e}_i$ is independent from all other rows of $\mathbf{E}$, then setting $\mathbf{u}_i = \mathbf{e}_i$ is valid for Proposition 1 to hold.

**Remark 2.** If $\mathbf{e}_i$ has no independent variation, i.e., there exists a Borel measurable $g$ such that $\mathbf{e}_i = g(\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_m)$, then we can set $\mathbf{u}_i = \mathbf{0}$. In this case, the random variation of the data for test $i$ is completely confounded with the variation of the other tests, so test $i$ is stochastically not a separate test, but a convolution of the other tests.

**Remark 3.** Suppose that the dependence in $\mathbf{E}$ is due to unmodeled factors $\mathbf{H}$, where $\mathrm{E}[\mathbf{x}_i|\mathbf{H}] = \mathbf{a}_i\mathbf{T}(\mathbf{H})$, $\mathbf{e}_i = \mathbf{a}_i\mathbf{T}(\mathbf{H}) + \mathbf{u}_i^*$, and the $\mathbf{u}_i^*$ are jointly independent across tests. In this case, setting $\mathbf{u}_i = \mathbf{u}_i^*$ satisfies the properties required for Proposition 1.

**Remark 4.** If we assume that $\mathbf{E}$ is Normal, the result can be derived in a different fashion, which might provide more insight for some readers. First consider just two hypothesis tests, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are dependent Normal vectors. That is, for a fixed sample $j$, $e_{1j}$ and $e_{2j}$ are Normally distributed with $\mathrm{Cov}(e_{1j}, e_{2j}) \neq 0$. It is well known that we can write:

$$\begin{aligned} e_{1j} &= z_j + u_{1j} \\ e_{2j} &= z_j + u_{2j} \end{aligned}$$

where $z_j$, $u_{1j}$, and $u_{2j}$ are all independent Normal random variables. Also, $\mathrm{Cov}(e_{1j}, e_{2j}) = \mathrm{Var}(z_j)$, $\mathrm{Var}(e_{1j}) = \mathrm{Var}(z_j) + \mathrm{Var}(u_{1j})$, $\mathrm{Var}(e_{2j}) = \mathrm{Var}(z_j) + \mathrm{Var}(u_{2j})$, and $\mathrm{Cov}(u_{1j}, u_{2j}) = 0$. In other words, two dependent Normal random variables can be partitioned into a dependent component, $z_j$, and independent components $u_{1j}$ and $u_{2j}$. This can be extended by using standard results on multivariate Normal random variables for any set of $m$ dependent Normal random variables. According to Proposition 3 below, if $\mathbf{e}^j$ is the $j$th column of $\mathbf{E}$, then we can write

$$\mathbf{e}^j = \mathbf{A}\mathbf{z}^j + \mathbf{u}^j,$$

where $\mathbf{A}$ is an $m \times m^*$ matrix ($m^* \leq m$), $\mathbf{z}^j$ is Normally distributed $m^*$ vector, and $\mathbf{u}^j$ is a Normally distributed $m$-vector. The components of $\mathbf{u}^j$ are independent. In the proof of Proposition 1, we showed that $\mathbf{E} = \mathbf{M} + \mathbf{U}$ where $\mathbf{U}$ is independent across its rows. By setting the $j$ column of $\mathbf{U}$ to be $\mathbf{u}^j$ and the $j$th column of $\mathbf{M}$ to be $\mathbf{A}\mathbf{z}^j$, the remainder of the proof follows the same.

**Proposition 3.** Let $\mathbf{e}$ be a random vector of length $m$, where $\mathbf{e} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ is a positive definite matrix, and $\mathbf{C} = \mathrm{diag}\{\sigma_{11}, \ldots, \sigma_{\mathrm{mm}}\}$. Then there exists a matrix $\mathbf{A}$ of constants, a constant $\lambda_0$, and independent random vectors $\mathbf{z} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I})$ and $\mathbf{u} \sim \mathrm{MVN}(\mathbf{0}, \lambda_0\mathbf{C})$ such that:

$$\mathbf{e}^* = \mathbf{A}\mathbf{z} + \mathbf{u},$$

where $\mathbf{e}^*$ and $\mathbf{e}$ have the same distribution.

**Proof of Proposition 3.** Since $\mathbf{\Sigma}$ is positive definite and symmetric, all of the diagonal elements of $\mathbf{\Sigma}$ must be positive, so $\mathbf{C}$ is positive definite and $\mathbf{C}^{-1}$ exists and is positive definite. Then $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^{-1}\mathbf{\Sigma} = \mathbf{C}\mathbf{K}$, where $\mathbf{K}$ is positive definite. Let $\lambda_0 > 0$ be the smallest eigenvalue of

$\mathbf{K}$. Then $\mathbf{K} = \mathbf{K} - \lambda_0\mathbf{I} + \lambda_0\mathbf{I}$ and the matrix $\mathbf{K}^* = \mathbf{K} - \lambda_0\mathbf{I}$ is non-negative definite. Applying the spectral theorem we can write $\mathbf{\Sigma} = \mathbf{C}(\mathbf{K}^* + \lambda_0\mathbf{I}) = \mathbf{L}^T\mathbf{L} + \lambda_0\mathbf{C}$. Setting $\mathbf{A} = \mathbf{L}^T$, we have $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T + \lambda_0\mathbf{C}$. Using properties of the Normal distribution $\mathrm{E}[\mathbf{e}^*] = 0$ and $\mathrm{Var}\,[\mathbf{e}^*] = \mathrm{Var}\,[\mathbf{A}\mathbf{z}] + \mathrm{Var}\,[\mathbf{u}] = \mathbf{A}\mathbf{A}^T + \lambda_0\mathbf{C}$, so $\mathbf{e}^* \sim \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma})$ as required.

# C Estimating G in Practice

We present two scenarios of estimating G in practice, and provide a specific algorithm to estimate $\mathbf{G}$ in one of them.

**Dependence of an Unknown Structure.** The first scenario is when nothing is known about the dependence structure and it is the case that $d + r \ll n$, where $d$ is the row-dimension of $\mathbf{S}(\mathbf{Y})$ and $r$ is the row-dimension of $\mathbf{G}$. This scenario is inspired by the setting where dependence is induced by common latent variables, such as in microarray data analysis (2). Failing to include all relevant factors is a common issue in genomics leading to latent structure (2,3). Consider a study where $n$ subjects are randomly assigned to one of two treatments, and microarrays are utilized to measure genome-wide RNA levels from a tissue of interest in each subject. The goal is to identify genes that show different levels of RNA expression between the two treatment groups. Also suppose that the subjects are of different ages and are composed of both males and females. It is highly likely that there will be a large number of genes that show differential expression with respect to age or sex. If these factors are not included among the primary variables $\mathbf{Y}$, then there will be dependence among those genes differentially expressed with respect to age or sex. Because $\mathbf{\Gamma}\widehat{\mathbf{G}}$ will be included into the model $\mathbf{X} = \mathbf{B}\mathbf{S} + \mathbf{\Gamma}\widehat{\mathbf{G}} + \mathbf{U}$ used to perform the hypothesis tests, it follows that $\widehat{\mathbf{G}}$ acts as a set of surrogate variables for the latent structure or dependence noise. Because of this, we call this approach "surrogate variable analysis" (SVA).

In this scenario, $\mathbf{G}$ can be interpreted as a set of variables that act as surrogates for these unmodeled factors. Let $\mathbf{Z}$ be the set of unmodeled factors whose signature in the data is captured by $\mathrm{E}[\mathbf{x}_i|\mathbf{Z}] = \mathbf{h}_i\mathbf{T}(\mathbf{Z})$, where $\mathbf{h}_i$ is an $n$-vector and $\mathbf{T}(\mathbf{Z})$ is an $r \times n$ matrix. For $\mathbf{G}$ to be a valid dependence kernel, the rows of $\mathbf{G}$ must span the same space as the rows of $\mathbf{T}(\mathbf{Z})$. Therefore, for multiple testing dependence caused by latent structure due to unmodeled factors, $\mathbf{G}$ can be interpreted as a valid linear basis for the effect of the factors on the data. By utilizing techniques similar to factor analysis, $\mathbf{G}$ can be analyzed to scientifically interpret the latent structure and potentially identify the relevant $\mathbf{Z}$.

Below we present an algorithm for estimating $\mathbf{G}$ in this first scenario, called iteratively reweighted surrogate variable analysis (IRW-SVA). The basic idea when estimating $\mathbf{G}$ in this scenario is to identify a subset of tests that show a strong association with $\mathbf{G}$ (i.e., $\boldsymbol{\gamma}_i \neq \mathbf{0}$), but no association

with $\mathbf{S}$ (i.e., $\mathbf{b}_i = \mathbf{0}$). The estimate of $\mathbf{G}$ can then be formed based on the right singular vectors of the data corresponding to this subset of tests. This approach accomplishes two things. First, it does not require the dependence kernel estimate $\widehat{\mathbf{G}}$ to be orthogonal to $\mathbf{S}$. Since it will rarely be the case that $\mathbf{S}$ and $\mathbf{G}$ are orthogonal, even under well-designed randomized studies, forcing $\mathbf{S}$ and $\widehat{\mathbf{G}}$ to be orthogonal will lead to persistent anti-conservative bias. Second, by taking a subset of the data, bias from $\mathbf{S}$ in the estimate $\widehat{\mathbf{G}}$ is reduced. The approach we take is to simultaneously up-weight the tests that show strong association to $\mathbf{G}$ and down-weight tests that show strong association with $\mathbf{S}$. Once the estimate $\widehat{\mathbf{G}}$ is formed, the model $\mathbf{X} = \mathbf{BS} + \mathbf{\Gamma}\widehat{\mathbf{G}} + \mathbf{U}$ is fit and the tests are performed on the $\mathbf{b}_i$ in the usual manner.

**Highly Structured Dependence.** A second scenario where estimating $\mathbf{G}$ is feasible in practice is when there is strong dependence among tests, but enough is known about the dependence structure so that $\mathbf{\Gamma}$ can be characterized *a priori* to a large extent. This is likely in the case of strong spatial dependence where the covariance structure is often specified in the model, for example, in brain imaging problems (4, 5). Here it may be the case that $d + r \approx n$, which would be problematic for our IRW-SVA algorithm below. However, in this second scenario, $\mathbf{\Gamma}$ is largely already characterized because of the covariance constrains, providing the degrees of freedom needed to estimate $\mathbf{G}$ when $d + r \approx n$. Whereas the IRW-SVA algorithm below requires one to fit $\mathbf{\Gamma}$ in addition to $\mathbf{B}$ once $\widehat{\mathbf{G}}$ is formed, this will not be necessary when $\mathbf{\Gamma}$ is highly structured. Indeed, once $\widehat{\mathbf{G}}$ is formed, it can be shown that identifying $\mathbf{\Gamma}$ is straightforward. Thus, we anticipate an effective algorithm in this setting will (i) characterize the structure of $\mathbf{\Gamma}$ based on the well characterized dependence structure, (ii) estimate $\widehat{\mathbf{G}}$ from a properly formed subset of $\mathbf{X}$, (iii) rotate $\mathbf{\Gamma}$ according to $\widehat{\mathbf{G}}$, and (iv) subtract $\mathbf{\Gamma}\widehat{\mathbf{G}}$ from $\mathbf{X}$ before performing any inference. This allows one to utilize what is known about the dependence structure to overcome the fact that $d + r \approx n$ in this case.

In brain imaging, a body of sophisticated theory and methods have been developed that calculate the tail distribution of maximal statistics under this spatial dependence model. Under our framework, instead of the goal being to account for the complex dependence structure among the statistics when calculating their tail probabilities over the population of all studies, the goal would instead be to estimate $\mathbf{G}$ in each specific study. If $\mathbf{G}$ is well estimated, calculating the distribution of maximal statistics now becomes straightforward because they are independent, once we have also conditioned on $\mathbf{G}$ (as detailed in Proposition 2 of the main text). The decomposition $\mathbf{\Gamma G} + \mathbf{U}$ also has a direct scientific interpretation. The matrix $\mathbf{G}$ represents a set of axes in $\mathbb{R}^n$ that fully capture the random realization of the spatial dependence. The vector $\boldsymbol{\gamma}_i$ denotes the position of the $i$th voxel among these axes.

**Iteratively Re-weighted SVA.** The basic idea when estimating $\mathbf{G}$ in this scenario is to identify a subset of tests whose data show a strong association with $\mathbf{G}$, but not a strong association with

**S**. The estimate of **G** can then be formed based on the right singular vectors of this subset. This approach accomplishes two things. First, it does *not* require the dependence kernel estimate $\widehat{\mathbf{G}}$ to be orthogonal to **S**. Since there will generally be a low probability that **S** and **G** are orthogonal, even under randomized studies, forcing **S** and $\widehat{\mathbf{G}}$ to be orthogonal will lead to persistent anti-conservative bias. Second, by taking a subset of the data, bias from **BS** in the estimate $\widehat{\mathbf{G}}$ is reduced. The approach we take is to up-weight the tests that show strong association to **G** and very low association with **S**.

We use the empirical posterior probability estimates, $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}\right)$ obtained from the approach of Storey *et al.* 2005 (6) to weight the tests. This technique is related to other approaches (7,8), but Storey *et al.* specifically demonstrate how to deal with composite hypotheses such as $\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0}$. Briefly, given a our current estimate of $\widehat{\mathbf{G}}_{(b)}$ we break down the probability estimation into two components:

$$\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right) = \widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0} | \boldsymbol{\gamma}_i \neq \mathbf{0}, \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right) \widehat{\Pr}\left(\boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right).$$

Here we sketch the Storey *et al.* (6) approach for estimating $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0} | \boldsymbol{\gamma}_i \neq \mathbf{0}, \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$; the estimation of $\widehat{\Pr}\left(\boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$ is analogous. We first form $F$-statistics $F_1, \ldots, F_m$ using standard linear models for testing the hypotheses:

$$H_{0i} : \mathbf{b}_i = \mathbf{0} \quad \text{vs.} \quad \mathrm{H_{1i}} : \mathbf{b_i} \neq \mathbf{0}.$$

Note that when fitting this model, $\boldsymbol{\gamma}_i$ is a free parameter and $\widehat{\mathbf{G}}_{(b)}$ is also utilized. We then calculate bootstrap null statistics $F_i^{0k}$ for $k = 1, \ldots, K$ by using the standard method (9). Again, by including model fits involving $\widehat{\boldsymbol{\gamma}}_i$ and $\widehat{\mathbf{G}}_{(b)}$ when forming the bootstrap samples, we are able to bootstrap from the proper conditional null distribution (6). Suppose that the null and alternative statistics have probability density functions $g_0$ and $g_1$. Then if $\pi_0$ of the null hypotheses are true, the probability density function of $F_i$ is $g = \pi_0 g_0 + \pi_1 g_1$. From Bayes theorem,

$$\Pr(\mathbf{b}_i = \mathbf{0} | F_i) = \frac{\pi_0 g_0(F_i)}{\pi_0 g_0(F_i) + (1 - \pi_0) g_1(F_i)},$$

where we have replaced $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0} | \boldsymbol{\gamma}_i \neq \mathbf{0}, \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$ with $\Pr(\mathbf{b}_i \neq \mathbf{0} | F_i)$ at a quantifiable loss of information (8). Since the $F_i$ are a sample from $g = \pi_0 g_0 + (1 - \pi_0) g_1$ and the $F_i^{0k}$ are a sample from $g_0$, we can form an estimate of the likelihood ratio $g_0/g$ using a non-parametric logistic regression where we consider the $F_i$ to be "successes" and the $F_{i0}$ to be "failures" (7). Since we only seek an estimate that is proportional to the true probability (because these are being

used as relative weights in the singular value decomposition), we set $\pi_0 = 1$ and calculate the corresponding posterior probability estimate directly from the estimated likelihood ratio. Once $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0} | \boldsymbol{\gamma}_i \neq \mathbf{0}, \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$ and $\widehat{\Pr}\left(\boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$ are formed, they are simply multiplied to obtain $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$.

---

ITERATIVELY RE-WEIGHTED SURROGATE VARIABLE ANALYSIS ALGORITHM

1: Fit the model $\mathbf{X} = \mathbf{BS} + \mathbf{E}$ by least squares, and calculate the residual matrix $\mathbf{R} = \mathbf{X} - \widehat{\mathbf{B}}\mathbf{S}$.

2: Perform a singular value decomposition of $\mathbf{R}$, and let $\mathbf{v}_k$ be the $k$th right eigenvector, $k = 1, \ldots, n$.

3: Let $\widehat{r}$ be the number of statistically significant $\mathbf{v}_k$ according to the algorithm by Buja and Eyuboglu (1992) (10), which is reproduced below.

4: Set $\widehat{\mathbf{G}}_{(0)}$ equal to the $\widehat{r} \times n$ matrix where row $k$ is $\mathbf{v}_k$.

FOR $b = 1, 2, ..., B$ ITERATIONS:

5: Form the empirical Bayes estimates $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$ based on Storey, Akey, and Kruglyak (2005) (6).

6: Perform a weighted singular value decomposition of $\mathbf{X}$ where row $i$ is weighted by $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(b)}\right)$.

7: Set $\widehat{\mathbf{G}}_{(b+1)}$ to be the $\widehat{r} \times n$ matrix of the first $\widehat{r}$ right eigenvectors from STEP 6.

8: Perform a weighted singular value decomposition of $\mathbf{X}$ where row $i$ is weighted by the final weights: $\widehat{\Pr}\left(\mathbf{b}_i = \mathbf{0}, \boldsymbol{\gamma}_i \neq \mathbf{0} | \mathbf{X}, \mathbf{S}, \widehat{\mathbf{G}}_{(B)}\right)$.

9: Set $\widehat{\mathbf{g}}_k$ to be the right eigenvector from STEP 8 that is most correlated with $\mathbf{v}_k$, $k = 1, \ldots, n$. Set $\widehat{\mathbf{G}}$ to be the $\widehat{r} \times n$ matrix where row $k$ is $\widehat{\mathbf{g}}_k$, $k = 1, \ldots, \widehat{r}$.

10: Perform the significance analysis on the $\mathbf{b}_i$ using the model $\mathbf{X} = \mathbf{BS} + \boldsymbol{\Gamma}\widehat{\mathbf{G}} + \mathbf{U}$, where $\widehat{\mathbf{G}}$ is treated as a set of fixed covariates and appropriate adjustments to the degrees of freedom for standard error estimates and hypothesis testing are made.

---

**Remark 1.** When the range of test-specific variances is large, it may improve the surrogate variable estimates to initially scale each test specific variance to one, which could be straightforwardly accomplished at Step 1. This scaling would ensure that each test's data contributes equally in the estimation of $\mathbf{G}$, so estimates are less heavily influenced by those tests with extremely large

variance. However, for the typical range of variances seen in genomic data, we have not observed any improvement in the algorithm when applying such as adjustment.

**Remark 2.** Across studies $\mathbf{G}$ can be viewed as a random matrix so that partitioning $\mathbf{E} = \mathbf{\Gamma G} + \mathbf{U}$ has connections to the familiar partitioning of errors in a mixed model. However, our goal is to perform inference for the variable $\mathbf{B}$ conditional on the observed values of $\mathbf{S}$ and $\mathbf{G}$. When mixed effects model are applied in the usual setting, the dimension along which the sampling occurs is the same as the inference dimension. There is usually a single random sample and a single observation of $\mathbf{G}$, for example, if we were to observe the expression values for a single gene rather than thousands of genes. However, in our scenario, the manifestation of $\mathbf{G}$ can be observed in the data sets corresponding to the multiple tests. This is again due to the fact that the sampling occurs along a different dimension (columns of $\mathbf{X}$) than the inference (rows of $\mathbf{X}$). As compared to traditional studies, the information we have is equivalent to being able to observing the exact same random effect in many studies. This difference has two important implications. First, even when $\mathbf{G}$ and $\mathbf{S}$ are assumed to be independent, such as in a randomized study, by chance $\mathbf{G}$ and $\mathbf{S}$ may be correlated in any study, resulting in confounding for many tests; this would not be the case in the traditional setting where the random effect is modeled over many repeated studies and independent samplings. Second, since the same set of vectors $\mathbf{G}$ are in the data's true model for multiple tests simultaneously, it is possible in our scenario to directly estimate $\mathbf{G}$ by averaging appropriately over the data for all tests. To summarize, over repeated studies $\mathbf{G}$ is a random variable; in any fixed study, $\mathbf{G}$ is a fixed set of vectors that parameterizes the data for the set of multiple hypothesis tests. Because of this, standard random effects estimators, such as the best linear unbiased predictors (BLUPs), are problematic for two reasons. First, they assume $\mathbf{G}$ and $\mathbf{S}$ are independent, which means that at a technical level, the estimates of $\mathbf{G}$ and $\mathbf{S}$ are orthogonal. Standard BLUPs will result in biased estimates of $\mathbf{G}$ and hence $\mathbf{B}$. Second, to estimate BLUPs, the distributions of $\mathbf{G}$ and $\mathbf{U}$ must be specified in advance, which is either difficult or requires substantial assumptions on the part of the analyst.

**Buja and Eyuboglu Algorithm.** For completeness we reproduce the Buja and Eyuboglu (10) algorithm for estimating the dimension of the dependence kernel, $\mathbf{G}$. The algorithm is applied to $\mathbf{R}$ calculated in Step 1 of the iteratively re-weighted SVA algorithm. (It could also be recalculated at each re-weighted iteration, if one chooses to do so.) The algorithm compares the singular values in the observed residual matrix to the corresponding singular values in randomized residual matrices, where each row is permuted individually to break down any structure across rows.

Buja and Eyuboglu Algorithm

1: Calculate the singular value decomposition of the residual matrix $\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Since $\mathbf{R}$ is a residual matrix resulting from a $d$ degrees of freedom model fit, the last $d$ eigen-values are zero.

2: Let $\lambda_\ell$ be the $\ell$th singular value, which is the $\ell$th diagonal element of $\mathbf{D}$, for $\ell = 1, \ldots, n$. For right singular value $k = 1, \ldots, n-d$ set the observed statistic to be:

$$T_k = \frac{\lambda_k^2}{\sum_{\ell=1}^{n-d} \lambda_\ell^2}$$

which is the variance in the residual matrix explained by the $k$th right singular vector.

3: Form a matrix $\mathbf{R}_p$ by permuting each row of $\mathbf{R}$ independently and calculating the residuals $\mathbf{R}_0 = \mathbf{R}_p - \widehat{\mathbf{B}}_p\mathbf{S}$ from fitting the model $\mathbf{R}_p = \mathbf{B}_p\mathbf{S} + \mathbf{E}_p$ to remove any structure across rows of the matrix, and calculate its singular value decomposition $\mathbf{R}_0 = \mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^T$.

4: For right singular value $k$ form a null statistic:

$$T_k^0 = \frac{\lambda_{0k}^2}{\sum_{\ell=1}^{n-d} \lambda_{0\ell}^2}$$

as above, where $\lambda_{0\ell}$ is the $\ell$th diagonal element of $\mathbf{D}_0$.

5: Repeat steps 4-7 a total of $B$ times to obtain null statistics $T_k^{0b}$ for $b = 1, \ldots, B$ and $k = 1, \ldots, n-d$.

6: Compute the p-value for right singular vector $k$ as:

$$p_k = \frac{\#\{T_k^{0b} \geq T_k; b = 1, \ldots, B\}}{B}$$

7: Estimate the number of significant surrogate variables by $\widehat{r}(\alpha) = \sum_{k=1}^{n-d} \mathbf{1}(p_k \leq \alpha)$, for a pre-specified threshold $\alpha$.

# D Evaluation of the IRW-SVA Algorithm

**Simulation Results.** Our approach to estimating $\mathbf{G}$ is based on identifying a subset of tests whose true model includes $\mathbf{G}$ but does not include $\mathbf{S}$ (i.e., the subset of tests $i$ such that $\boldsymbol{\gamma}_i \neq \mathbf{0}$ and $\mathbf{b}_i = \mathbf{0}$). If we could perfectly identify this subset, then it would be possible to form an unbiased estimate of $\mathbf{G}$ regardless of the correlation with $\mathbf{S}$, as long as the subset was large enough to span

**Table S1.** A summary of the parameters for the simulated microarray studies. The simulation scenarios encompass discrete and continuous $\mathbf{G}$ of varying dimension and varying magnitude of the signal associated with $\mathbf{G}$. For each of these different parameterizations, both high and low regression correlation between $\mathbf{S}$ and $\mathbf{G}$ and high and low association overlap in the number of tests with non-zero effects from $\mathbf{S}$ and $\mathbf{G}$ are considered. Full simulation details appear in Table S2.

| Study | Type of $\mathbf{G}$ | Dimension of $\mathbf{G}$ | Magnitude of $\mathbf{b}_i$ | Regression Correlation | Association Overlap |
|---|---|---|---|---|---|
| One | Discrete | $r = 1$ | Moderate | Low | Low |
| Two | Discrete | $r = 1$ | Moderate | Low | High |
| Three | Discrete | $r = 1$ | Moderate | High | Low |
| Four | Discrete | $r = 1$ | Moderate | High | High |
| Five | Continuous | $r = 1$ | Moderate | Low | Low |
| Six | Continuous | $r = 1$ | Moderate | Low | High |
| Seven | Continuous | $r = 1$ | Moderate | High | Low |
| Eight | Continuous | $r = 1$ | Moderate | High | High |
| Nine | Discrete | $r = 1$ | Large | Low | Low |
| Ten | Discrete | $r = 1$ | Large | Low | High |
| Eleven | Discrete | $r = 1$ | Large | High | Low |
| Twelve | Discrete | $r = 1$ | Large | High | High |
| Thirteen | Discrete | $r = 2$ | Moderate | Low | Low |
| Fourteen | Discrete | $r = 2$ | Moderate | Low | High |
| Fifteen | Discrete | $r = 2$ | Moderate | High | Low |
| Sixteen | Discrete | $r = 2$ | Moderate | High | High |

the row space of $\mathbf{G}$. The two parameters that most influence our ability to estimate the relevant subset of tests are (i) the "regression correlation" between $\mathbf{S}$ and $\mathbf{G}$, or the percentage of the column space of $\mathbf{S}$ explained by $\mathbf{G}$ and (ii) the "association overlap", or the percentage of tests whose true model includes $\mathbf{G}$ and $\mathbf{S}$ (via $\boldsymbol{\gamma}_i \neq \mathbf{0}$ and $\mathbf{b}_i \neq \mathbf{0}$, respectively).

We simulated 100 studies with a variety of different forms for $\mathbf{G}$ and the $\boldsymbol{\gamma}_i$. Each simulated study consisted of 1,000 tests and 20 samples divided into two equal treatment groups, parameterized by $\mathbf{S}$. In our notation, $\mathbf{S}$ is a $2 \times 20$ matrix, the first row parameterizing the intercept, the second row the group membership, and we are interested in testing whether $b_{i2} = 0$ for each hypothesis test. For each simulated study, tests 1-300 have non-zero $b_{i2}$ drawn from a common distribution, so that the alternative hypothesis is true for each one. We varied the regression correlation by either randomizing $\mathbf{G}$ with respect to $\mathbf{S}$, or allowing for consistent correlation between $\mathbf{G}$ and $\mathbf{S}$. The first case mimics a randomized study, where we would expect unmodeled latent factors to be orthogonal to $\mathbf{S}$ on average. The second case more closely resembles an observational study, where unmodeled factors are more likely to be correlated with $\mathbf{S}$. We also varied the association overlap by varying the percentage of tests with both nonzero $\mathbf{b}_i$ and $\boldsymbol{\gamma}_i$. The simulation parameters are summarized in Table S1.

For each simulated study we performed an unadjusted significance analysis (i.e., not modeling $\mathbf{G}$ at all), an analysis applying the above Iteratively Re-weighted Surrogate Variable Analysis (IRW SVA) algorithm, and an "ideal scenario" analysis on independent data where we simulate the data

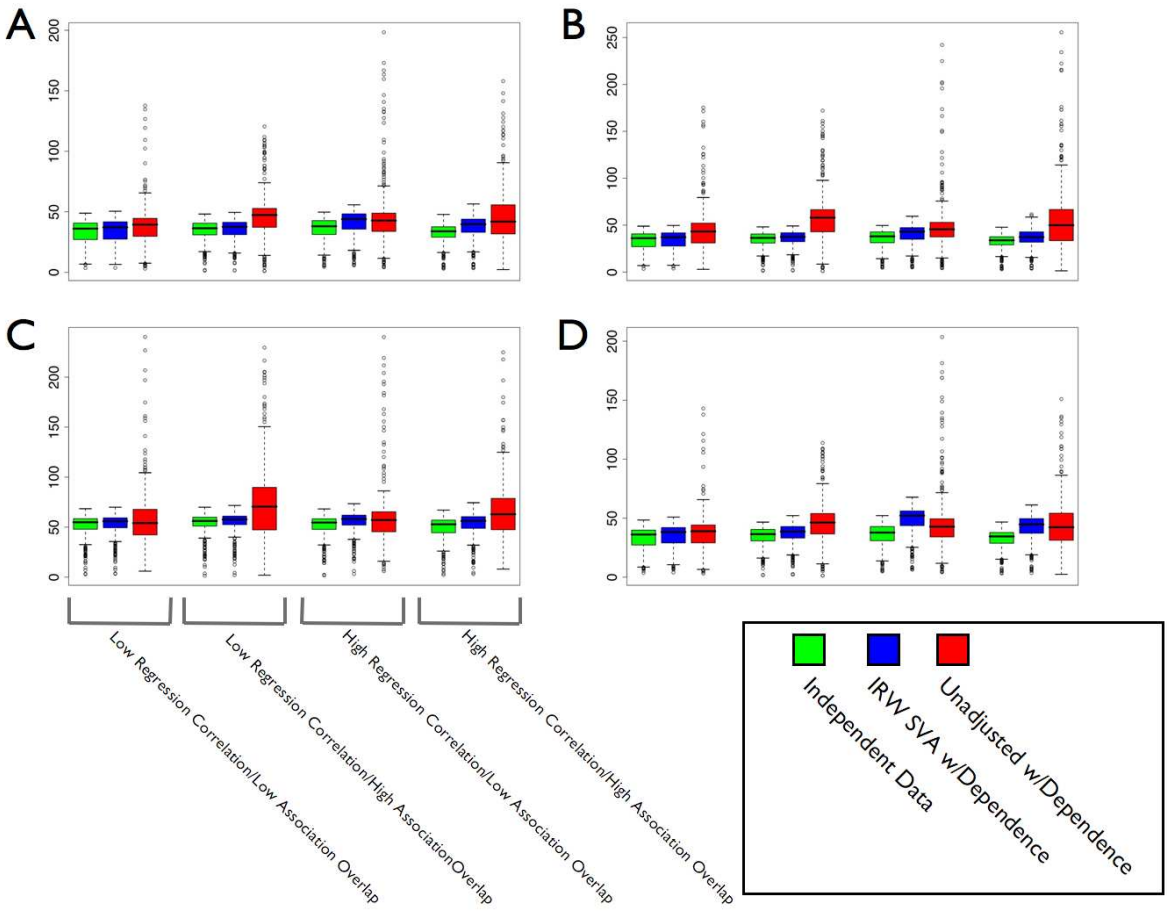


**Fig. S1.** Boxplots of the root mean square error between the true ranking (based on the non-centrality parameter) and the estimated ranking. For each case, the boxplots show the variability in rankings for the ideal scenario with independent tests (green), IRW SVA with dependence (blue), and unadjusted analysis with dependence (red). For each parameterization of $\mathbf{G}$ the four clusters of boxplots correspond to low regression correlation/low association overlap, low regression correlation/high association overlap, high regression correlation/low association overlap, high regression correlation/high association overlap. $\mathbf{G}$ was simulated as $(A)$ discrete with $r = 1$, $(B)$ continuous with $r = 1$, $(C)$ discrete with large non-zero $\boldsymbol{\gamma}_i$, $r = 1$ and $(D)$ discrete with $r = 2$.

exactly as above but set $\mathbf{G} = \mathbf{0}$. We compared the operating characteristics of these three cases across all simulated scenarios on the basis of two important metrics: stability of test rankings and correct null distributions. One of the most important aspects of any multiple testing significance analysis is behavior of the relative significance ranking of tests (i.e., ordering the tests from most to least significant). One way to rank the tests is based on the magnitude of the F-statistics for testing if $b_{i2} \neq 0$. In our two-sample scenario, the ideal ranking is that based on the non-centrality parameter of each true alternative test. In Fig. S1 we show boxplots of the root mean squared error (RMSE) in the rankings for all tests with $\mathbf{b} \neq 0$ with respect to this ideal ranking. The better the ranking, the smaller the mean and standard deviation of the RMSE will be. From Fig. S1 it can be seen that the rankings are variable in the unadjusted analysis with dependence. Applying the IRW SVA algorithm reduces the variability in the rankings nearly to the level of the corresponding study without independence.

A second important component in any significance analysis is that the null statistics should follow the correct null distribution (i.e., the one utilized in forming significance measures). At the level of the p-values, it is well known that most method for estimating multiple testing error rates are accurate only when the distribution of the null p-values is stochastically greater than or equal to the $Uniform(0,1)$ distribution (11–13). For the simulated studies under each set of parameters, we calculated a Kolmogorov-Smirnov (KS) test comparing the distribution of p-values for the null tests (tests 301-1000) to the $Uniform(0,1)$ distribution. If the null p-values are $Uniform(0,1)$ distributed for each study, then across all 100 simulated studies the KS-test p-values should also be $Uniform(0,1)$ distributed. This "double KS-test" is a robust approach for determining if the null p-values have the appropriate distribution across repeated samples from the same population (2). Fig. S2 plots the quantiles of the KS-test p-values across one hundred simulated studies versus the $Uniform(0,1)$ quantiles. If the null distribution is accurate for a particular analysis, the quantiles should lie along the diagonal identity line.

As with the test rankings, the unadjusted analysis under dependence behaves poorly, yielding p-values not following the $Uniform(0,1)$ distribution for the true null hypothesis tests. Again, adjusting for surrogate variables gives nearly identical results to the ideal scenario where the tests are independent. The double KS-test gives strong evidence that this pattern is consistent across hundreds of simulated studies, and we did not just get lucky under a single simulated scenario.

Adjusting for surrogate variables results in a correct null distribution, and this translates into improved FDR estimates. Fig. S3 shows that correcting the null distribution in Experiment 5 reduces error in both q-value estimates and global measures of significance, such as estimates of $\pi_0$, the proportion of true null hypotheses (13). We are able to directly correct dependence at the level of the originally observed data using surrogate variable analysis to obtain corrected error

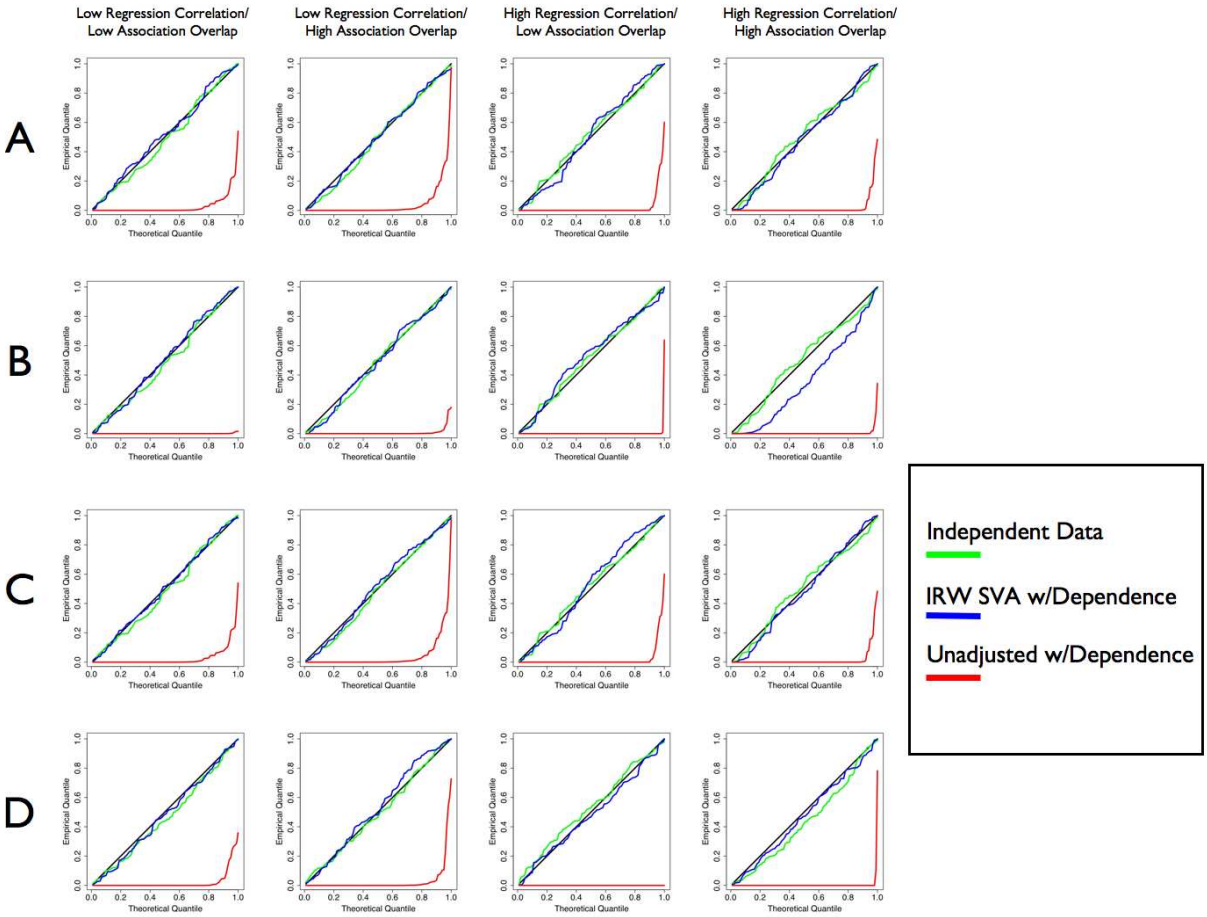


**Fig. S2.** KS-test quantile-quantile plots comparing the distribution of the null p-values for each simulated study to the Uniform distribution. For each case, the plots show the ideal scenario with independent tests (green), IRW SVA with dependence (blue), and unadjusted analysis with dependence (red). For each parameterization of $\mathbf{G}$, the four KS-test plots correspond to low regression correlation/low association overlap, low regression correlation/high association overlap, high regression correlation/low association overlap, high regression correlation/high association overlap. $\mathbf{G}$ was simulated as as $(A)$ discrete with $r = 1$, $(B)$ continuous with $r = 1$, $(C)$ discrete with large non-zero $\boldsymbol{\gamma}_i$, $r = 1$ and $(D)$ discrete with $r = 2$. It can be seen from these plots that p-values in the case of independent data follow the $Uniform(0, 1)$ distribution as expected, but dependence causes deviation from this distribution. Application of IRW-SVA restores the appropriate null distribution and results in correct inference.

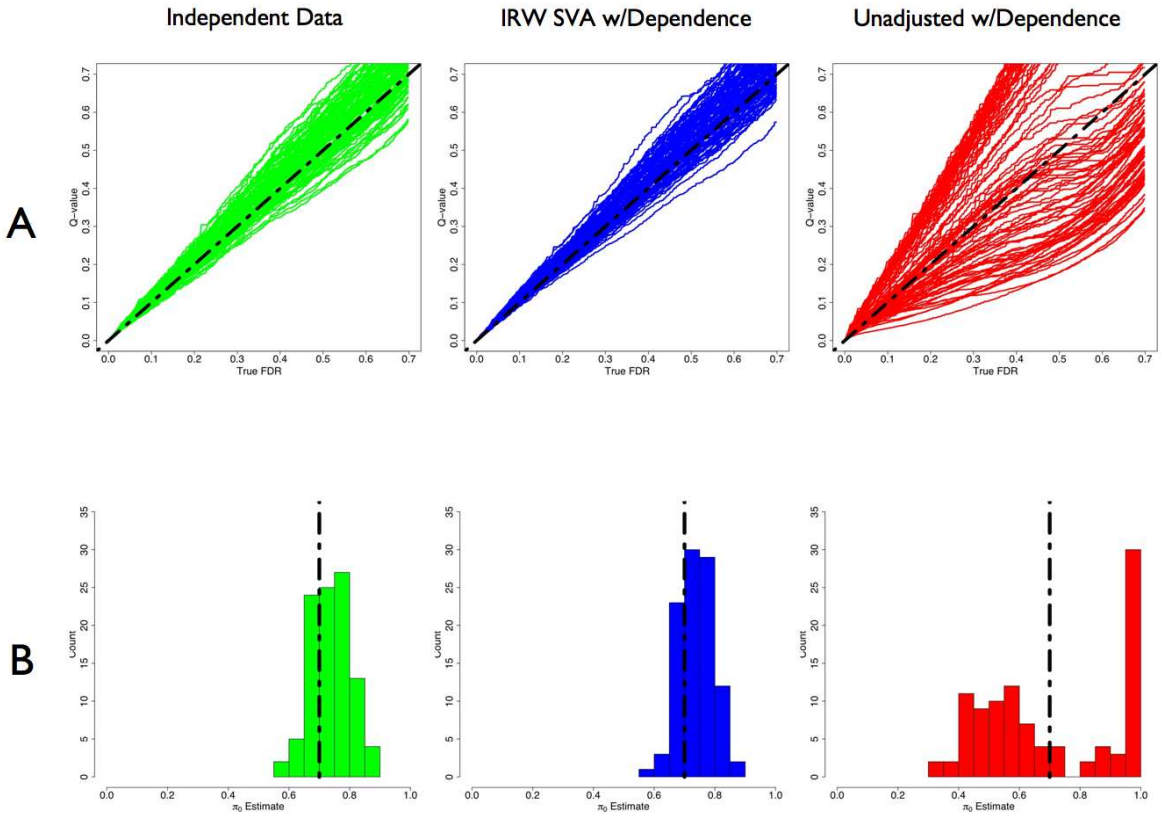


**Fig. S3.** Behavior of FDR estimates for the case of Experiment 5 (see Table S2 for details). For each case, the plots show the ideal scenario with independent tests (green), IRW SVA with dependence (blue), and unadjusted analysis with dependence (red). (*A*) Plots of q-value curves for 100 simulated studies versus the true FDR. Ideally these curves will on average lie slightly above the line of equality, indicating a conservative estimate, and have relatively little variability. This is true for both the independent data and surrogate variable adjusted analyses, but not for the unadjusted analysis with dependence. (*B*) Histograms of the $\pi_0$ estimates ($\pi_0$ = proportion of true null tests) across 100 simulated studies, these estimates should have a slight conservative bias, in other words they should on average be slightly larger than the true value of 0.7, again with small variability. The unadjusted analysis with dependence shows large variation in the $\pi_0$ estimates. This error is eliminated through application of the proposed IRW-SVA algorithm.

rate estimates. By using surrogate variables we eliminate the need for post-hoc adjustments to the null distribution of one-dimensional test-statistics (14) when calculating multiple testing error rates. Our results indicate that valid inference critically depends on a correct distribution of null p-values within any specific study. We have shown that directly estimating and incorporating $\mathbf{G}$ into significance analyses empirically corrects the null distribution across a variety of simulated cases. These results indicate that estimating $\mathbf{G}$ is possible even in difficult scenarios such as observational studies with highly interrelated variables.

**Simulation Details.** For each experiment described in Table S1 we simulated 100 independent studies with $m = 1000$, $n = 20$, and $u_{ij} \overset{i.i.d}{\sim} N(0, \sigma_i)$ where $\sigma_i \overset{i.i.d}{\sim} InvGamma(10, 9)$. For each study $\mathbf{S}$ was a vector of indicator variables equal to one for the first 10 samples and zero for the last 10. $\mathbf{B}$ was simulated as an $m$ vector where the first three hundred elements were distributed as independent $N(0, 2.5)$ random variables in the case of moderate signal and $N(3.5, 1)$ random variables for the large signal. In each case the last 700 elements of $\mathbf{B}$ were set equal to zero, making these the null tests. In all simulated studies $\boldsymbol{\gamma}_{ij} \overset{i.i.d}{\sim} N(0, 2.5)$ or was set equal to zero. When there was low association overlap $\gamma_{i1}$ was non-zero for tests 201-700 and when there was high association overlap $\gamma_{i1}$ was non-zero for tests 101-600. When $r = 2$, $\gamma_{i2}$ was non-zero for tests 401-900. For experiments 1-4 and 9-16, $g_{jk}$ was a discrete random variable. In cases where regression correlation was low $g_{jk} \overset{i.i.d}{\sim} Bernoulli(0.5)$ and in cases where regression correlation was high $g_{jk} \overset{i.i.d}{\sim} Bernoulli(0.7)$ for $j = 1, \ldots, 10$ and $g_{jk} \overset{i.i.d}{\sim} Bernoulli(0.2)$ for $j = 11, \ldots, 20$. For experiments 5-8, $g_{ij}$ was a continuous random variable. In cases where regression correlation was low $g_{jk} \overset{i.i.d}{\sim} N(0, 1)$ for all $j$ and in cases where the regression correlation was high $g_{jk} \overset{i.i.d}{\sim} N(0, 1)$ for $j = 1, \ldots 10$ and $g_{jk} \overset{i.i.d}{\sim} N(1, 1)$ for $j = 11, \ldots 20$. An important point is that the case of low regression correlation mimics a randomized study where on average $\mathbf{G}$ and $\mathbf{S}$ occupy orthogonal linear spaces, but in any fixed study $\mathbf{G}$ and $\mathbf{S}$ may be regression correlated by chance. All p-values were calculated based on a parametric $F$-test based on comparing the full model including $\mathbf{S}$ to the model without $\mathbf{S}$. In the IRW SVA analysis the null and alternative models also included the surrogate variable estimates. All computations were performed in the `R` programming language. Further details on the simulations can be seen in Table S2.

**Remark.** A potentially interesting avenue for future research is to characterize the finite sample and asymptotic properties of estimators of the dependence kernel. In terms of asymptotic results, we expect that the most useful results will be concerned with the scenario where the sample sizes, $n$, stay fixed and the number of tests, $m$, grows large. When asymptotically consistent estimators exist as $m \to \infty$, then in the limit we can extend the result of Proposition 1, replacing the true kernel with the estimate. As a weaker result but potentially equally useful in practice, it may also prove interesting to investigate whether estimates of the dependence kernel are sufficient to induce

**Table S2.** Details on the distributions used in the simulations.

| Experiment | $g_{ij}$ in group 1 | $g_{ij}$ in group 2 | Non-zero $\gamma_{ij}$ | Distribution of $\gamma_{ij}$ |
|---|---|---|---|---|
| One | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 201-700 | $N(0, 2.5)$ |
| Two | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 101-600 | $N(0, 2.5)$ |
| Three | $g_{i1} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 201-700 | $N(0, 2.5)$ |
| Four | $g_{i1} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 101-600 | $N(0, 2.5)$ |
| Five | $g_{i1} \overset{i.i.d}{\sim} N(0, 1)$ | $g_{i1} \overset{i.i.d}{\sim} N(0, 1)$ | Tests 201-700 | $N(0, 2.5)$ |
| Six | $g_{i1} \overset{i.i.d}{\sim} N(0, 1)$ | $g_{i1} \overset{i.i.d}{\sim} N(0, 1)$ | Tests 101-600 | $N(0, 2.5)$ |
| Seven | $g_{i1} \overset{i.i.d}{\sim} N(0, 1)$ | $g_{i1} \overset{i.i.d}{\sim} N(1, 1)$ | Tests 201-700 | $N(0, 2.5)$ |
| Eight | $g_{i1} \overset{i.i.d}{\sim} N(0, 1)$ | $g_{i1} \overset{i.i.d}{\sim} N(1, 1)$ | Tests 101-600 | $N(0, 2.5)$ |
| Nine | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 201-700 | $N(0, 2.5)$ |
| Ten | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 101-600 | $N(0, 2.5)$ |
| Eleven | $g_{i1} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 201-700 | $N(0, 2.5)$ |
| Twelve | $g_{i1} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 101-600 | $N(0, 2.5)$ |
| Thirteen | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 201-700 | $N(0, 2.5)$ |
| | $g_{i2} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i2} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 401-900 | $N(0, 2.5)$ |
| Fourteen | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 101-600 | $N(0, 2.5)$ |
| | $g_{i2} \overset{i.i.d}{\sim} Bern(0.5)$ | $g_{i2} \overset{i.i.d}{\sim} Bern(0.5)$ | Tests 401-900 | $N(0, 2.5)$ |
| Fifteen | $g_{i1} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 201-700 | $N(0, 2.5)$ |
| | $g_{i2} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i2} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 401-900 | $N(0, 2.5)$ |
| Sixteen | $g_{i1} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i1} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 101-600 | $N(0, 2.5)$ |
| | $g_{i2} \overset{i.i.d}{\sim} Bern(0.7)$ | $g_{i2} \overset{i.i.d}{\sim} Bern(0.2)$ | Tests 401-900 | $N(0, 2.5)$ |

so-called weak dependence (15).

**Application to a Population Genomics Study.** Idaghdour *et al.* (16) recently published a study measuring leukocyte gene expression in 46 desert nomadic, mountain agrarian, and coastal urban Moroccan Amazigh individuals. The goal of the study was to identify genes that show differential expression across the three geographically defined populations, which each involve notably different lifestyles and experience very little migration among them. In addition to the population indicator variable, the sex of the patients and batch variable (a technical variable), were also recorded. Our model for the expression of gene $i$ for individual $j$ can then be written as follows:

$$x_{ij} = b_{0i} + b_{1i}\mathbf{1}\{\text{desert}_j\} + b_{2i}\mathbf{1}\{\text{mountain}_j\} + b_{3i}\mathbf{1}\{\text{batch}_j\} + b_{4i}\mathbf{1}\{\text{sex}_j\} + e_{ij}. \qquad \text{(d)}$$

The goal of the study is to test the hypotheses:

$$H_{0i} : b_{1i} = 0 \;\&\; b_{2i} = 0 \qquad \text{vs.} \qquad H_{1i} : b_{1i} \neq 0 \text{ or } b_{2i} \neq 0.$$

We performed an unadjusted analysis (i.e., an analysis ignoring any dependence) using standard $F$-tests of the above hypotheses. At FDR = 1%, 5%, and 10% there are respectively 2,701, 5,111,

and 6,718 genes called significantly differentially expressed with respect to geographical region.

We next applied IRW-SVA algorithm, where the model $\mathbf{S}$ includes both the "batch" and "sex" variables in addition to the geographical population variable. This is an important point for utilizing the proposed algorithm: if measured variables such as batch and sex are known to be possibly in the true model, then they should be included explicitly in the model, even if they are not the focus of the significance analysis. The surrogate variable estimation algorithm can easily incorporate these variables, and conditioning on variables that are known to play a key role in expression can improve the estimates of the remaining surrogate variables. Applying the IRW-SVA algorithm, we can recalculate significance for each gene. The number of genes significant at FDR = 1%, 5% and 10% are respectively 1,940, 4,261, and 5,900. This illustrates another key point: taking into account dependence may actually reduce the observed empirical power in any given study. The reduction in observed power occurs when row spaces spanned by $\mathbf{G}$ and $\mathbf{S}$ overlap. This is seen extensively in the simulated examples above and in the main text. In the expression study, some of the apparent "signal" is in fact due to confounding between the dependence kernel and geographical population variable. Thus, while we have a reduction in the number of genes called significant, the genes that are called significant are more likely to be truly differentially expressed with respect to the geographical population variable. In our analysis, the top surrogate variable (i.e., first row of $\widehat{\mathbf{G}}$) has correlation 0.23 with the "desert" indicator variable. When regressing the geographical population variable on the rows of $\widehat{\mathbf{G}}$, we obtain an $R$-squared value of 0.68. It is likely that some latent factor, partially confounded with being geographically located in the desert, is driving part of the differential expression signal in the unadjusted analysis. This is not surprising given that this study has some observational (as opposed to randomized) sampling characteristics (16).

We can use the estimated posterior weights to determine on a qualitative level the distribution of signal due to geographic location and due to unmodeled factors. Fig. S4 shows the distribution of the scaled weights derived from the $\widehat{\mathbf{Pr}}(b_{1i} \neq 0 \text{ or } b_{2i} \neq 0)$ and $\widehat{\mathbf{Pr}}(\boldsymbol{\gamma}_i \neq \mathbf{0})$ estimates. It is clear that a large percentage of genes are affected by one or more unmodeled factors, much more so than the number of genes that have a high probability of being associated with the geographical population. These relative weights can be used as both a model diagnostic and to identify the most robustly differentially expressed genes between populations.

Finally, as a proof of concept meant for illustrative purposes only, we applied the IRW-SVA algorithm to this study, where we left out the "batch" variable from equation (d). The batch variable is easily verified to be a major source of expression variation. By repeating the analysis without "batch", we are able to see if the $\widehat{\mathbf{G}}$ estimate captures the batch variable. The most significant of the six surrogate variables estimated has correlation 0.71 with the batch variable, so the top surrogate variable is a good estimate of this technical factor. Also, the $R$-squared value of

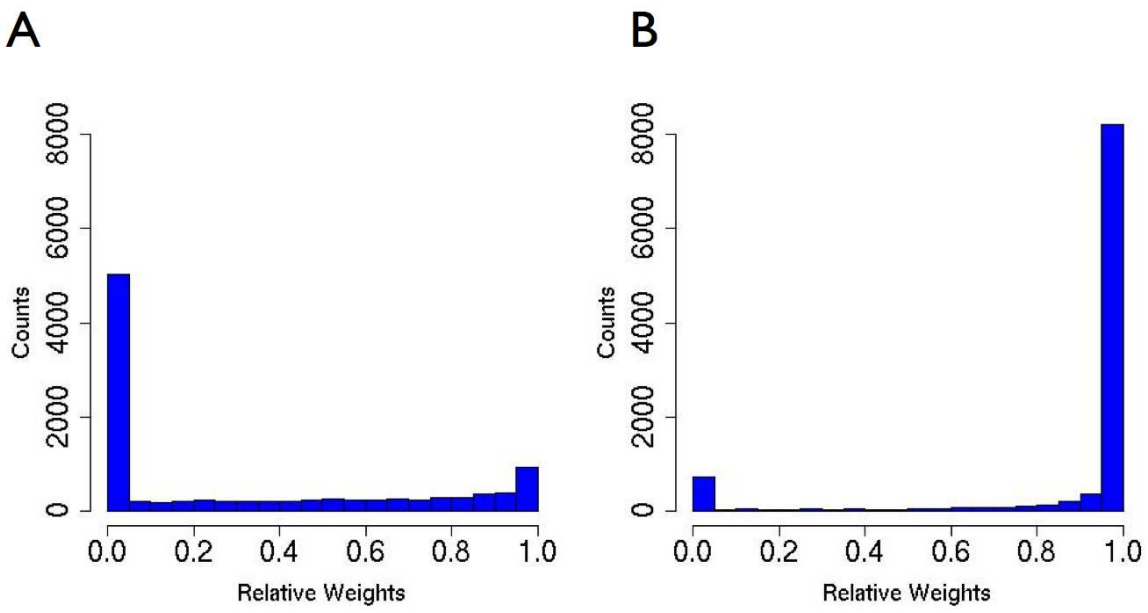


**Fig. S4.** Distribution of relative weights used in the IRW-SVA algorithm. Step 8 of the IRW-SVA algorithm weights the tests' data based on the product of the posterior probability estimates $\widehat{\mathbf{Pr}}(b_{1i} \neq 0 \text{ or } b_{2i} \neq 0)$ and $\widehat{\mathbf{Pr}}(\boldsymbol{\gamma}_i \neq \mathbf{0})$. The distribution of the relative contribution of $(A)$ $\widehat{\mathbf{Pr}}(b_{1i} \neq 0 \text{ or } b_{2i} \neq 0)$ and $(B)$ $\widehat{\mathbf{Pr}}(\boldsymbol{\gamma}_i \neq \mathbf{0})$ to these weights in the Idaghdour *et al.* study is shown.

batch regressed on the rows of $\widehat{\mathbf{G}}$ is 0.54. Therefore, IRW-SVA was able to provide an estimate of this unmodeled technical factor directly from the gene expression data, without using the measured batch information. In practice however, we would include measured variables likely to have a large influence on gene expression such as the batch variable.

## References


[1] Rüschendorf L, de Valk V (1993) On regression representations of stochastic processes. *Stoch Proc and Their Apps*, 4:183–198.

[2] Leek J T, Storey J D (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:e161, 2007.

[3] Pritchard J K, Rosenberg N A (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Amer J Hum Gen*, 65:220–8.

[4] Worsley K J, et al. (1996) A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73.

[5] Geneovese C R, Lazar N A, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:870–8.

[6] Storey J D, Akey J M, Kruglyak L (2005) Multiple locus linkage analysis of geneomewide expression in yeast. *PLoS Biology*, 3:1380–90.

[7] Anderson J A, Blair V (1982) Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrkia*, 69:123–36.

[8] Efron B, Tibshirani R, Storey J D, Tusher V (2001) Empirical bayes analysis of a microarray experiment. *J Comp Bio*, 96:1151–60.

[9] Efron B, Tibshirani R J (1993) *An Introduction to the Bootstrap.* Chapman & Hall: San Francisco, CA.

[10] Buja A, Eyuboglu N (1992) Remarks on parallel analysis. *Multivariate Behav*, 27:509–40.

[11] Lehmann E L (1997) *Testing Statistical Hypotheses.* Springer: New York, NY.

[12] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57:289–300.

[13] Storey J D (2002) A direct approach to false discovery rates. *J Roy Stat Soc B*, 64:479–98, 2002.

[14] Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc*, 99:96–104.

[15] Storey J D, Taylor J E, Siegmund D (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J Roy Stat Soc B*, 66:187–205.

[16] Idaghdour Y, Storey J D, Jadallah S, Gibson G (2008) A genome-wide gene expression signature of lifestlye in peripheral blood of moroccan amazighs. *PLoS Genetics*, 4:e1000052.