

# Statistical Significance of Variables Driving Systematic Variation

Neo Christopher Chung\* and John D. Storey<sup>†‡</sup>

July 2013

## Contents

<b>Introduction</b>	<b>3</b>
<b>Statistical Model and Approach</b>	<b>4</b>
<b>Proposed Algorithms</b>	<b>5</b>
<b>Results</b>	<b>7</b>
<b>Discussion</b>	<b>12</b>
<b>Tables</b>	<b>15</b>
<b>Figures</b>	<b>16</b>
<b>Supporting Information</b>	<b>21</b>
<b>Supplementary Figures</b>	<b>25</b>
<b>References</b>	<b>32</b>

---

\*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.

<sup>†</sup>Lewis-Sigler Institute for Integrative Genomics & Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA.

<sup>‡</sup>To whom correspondence should be sent: John Storey, <http://genomine.org/contact.html>.

## Abstract

There are a number of well-established methods such as principal components analysis (PCA) for automatically capturing systematic variation due to latent variables in large-scale genomic data. PCA and related methods may directly provide a quantitative characterization of a complex biological variable that is otherwise difficult to precisely define or model. An unsolved problem in this context is how to systematically identify the genomic variables that are drivers of systematic variation captured by PCA. Principal components (and other estimates of systematic variation) are directly constructed from the genomic variables themselves, making measures of statistical significance artificially inflated when using conventional methods due to over-fitting. We introduce a new approach called the *jackstraw* that allows one to accurately identify genomic variables that are statistically significantly associated with any subset or linear combination of principal components (PCs). The proposed method can greatly simplify complex significance testing problems encountered in genomics and can be utilized to identify the genomic variables significantly associated with latent variables. Using simulation, we demonstrate that our method attains accurate measures of statistical significance over a range of relevant scenarios. We consider yeast cell-cycle gene expression data, and show that the proposed method can be used to straightforwardly identify genes that are cell-cycle regulated with an accurate measure of statistical significance. We also analyze gene expression data from post-trauma patients, allowing the gene expression data to provide a molecularly-driven phenotype. Using our method, we find a greater enrichment for inflammatory-related gene sets compared to the original analysis that utilizes a clinically defined, although likely imprecise, phenotype. The proposed method provides a useful bridge between large-scale quantifications of systematic variation and gene-level significance analyses.

Abbreviations: KS test, Kolmogorov-Smirnov test; PCA, principal components analysis; PC, principal component

# Introduction

Latent variable models play an important role in understanding variation in genomic data [1, 2]. They are particularly useful for characterizing systematic variation in genomic data whose variable representation is unobserved or imprecisely known (Fig. 1). Principal component analysis (PCA) has proven to be an especially informative method for capturing quantitative signatures of latent variables in genomic data, and it is in widespread use across a range of applications. For example, PCA has been successfully applied to uncover the systematic variation in gene expression [3–5], estimate structure in population genetics [1,6], and account for dependence in multiple hypothesis testing [2, 7]. Generally, principal components (PCs) can be thought of as estimates of unobserved manifestation of latent variables; they are constructed by aggregating variation across thousands or more genomic variables [8]. What is missing from this highly successful system is a method to precisely identify which genomic variables are the statistically significant drivers of the PCs in genomic data, which in turn identifies the genomic variables associated with the unobserved latent variables.

In a typical application of PCA, all genomic variables (e.g. microarray probes, genetic loci, etc.) will have nonzero “loadings”, meaning that they all make some contribution to the construction of PCs. In some cases, when many (or most) of these contributions are forcibly set to zero, similar PCs nevertheless emerge. Methods have been proposed to induce sparsity in the loadings<sup>1</sup>, for example, with a lasso penalized PCA or a Bayesian prior [10–13]. Also, various formulations of statistical significance have been considered previously in the context of PCA. These have usually been focused on scenarios where the number of observations is substantially larger than the number of variables, significance is measured in terms of a completely unstructured data matrix where all variables are mutually independent, or the goal is to only determine the number of significant principal components [14–21]. The problem we consider here differs from those scenarios.

Our goal is not a minimal representation of a PCA; we would like instead to develop a strategy that accurately identifies which genomic variables are truly associated with systematic variation of interest. This can be phrased in statistical terminology as developing a significance test for associations between genomic variables and a given set, subset, or linear combination of PCs estimated from genomic data. We introduce a new resampling approach, which we call the *jackstraw*, to rigorously identify the genomic variables associated with PCs of interest, as well as subsets and rotations of PCs of interest. Our approach is capable of obtaining the empirical null distribution of association statistics (e.g., F-statistics) and applying these to the observed association statistics between genomic features and PCs in order to obtain valid statistical significance measures. Succinctly, new PCs are computed from a dataset with a few independently permuted variables, which become tractable “synthetic” null variables. The association statistics between newly computed PCs and synthetic null variables serve as empirical null statistics, accounting for the measurement error and over-fitting of PCA.

---

<sup>1</sup>The difference between our goal and the goal of sparsity is analogous to the difference between penalized regression and multiple hypothesis testing, both of which have been proven to be useful in genomics and other applications [9].

As an application, we consider the problem of identifying genes whose expression is cell-cycle regulated. In this case, there are infinitely many theoretical curves that would represent “cell-cycle regulation” to the point where a standard statistical analysis involves an unwieldy “composite null hypothesis” [22]. We identify the few realized patterns of cell-cycle regulated gene expression through PCA and we are able to directly test whether each gene is associated with these using the proposed approach. As another application, we analyzed observational gene expression profiles of blunt-force trauma patients [23], whose post-trauma inflammatory responses are difficult to be quantified using conventional means. When the clinical phenotype of interest can not be precisely measured and modeled, we may estimate it directly from genomic data itself. We identify genes driving systematic variation in gene expression of post-trauma patients and demonstrate that our analysis is biologically richer than the original analysis [23].

PCA has direct connections to Independent Component Analysis [9] and  $K$ -means clustering [24, 25]. Therefore, the methods we propose are likely applicable to those models as well. Furthermore, this approach has potential generalizations to a much broader class of clustering and latent variable methods that all seek to capture systematic variation.

## Statistical Model and Approach

Consider a  $m \times n$  row-centered expression data matrix  $\mathbf{Y}$  with  $m$  observed variables measured over  $n$  observations ( $m \gg n$ ).  $\mathbf{Y}$  may contain systematic variation across the variables from an arbitrarily complex function of latent variables  $\mathbf{z}$ . We may calculate the expected influence of the latent variables on  $\mathbf{Y}$  by  $E[\mathbf{Y}|\mathbf{z}]$ , and then write  $\mathbf{Y} = E[\mathbf{Y}|\mathbf{z}] + \mathbf{E}$ , where  $\mathbf{E}$  is defined as  $\mathbf{Y} - E[\mathbf{Y}|\mathbf{z}]$ . There exists a  $r \times n$  matrix, called  $\mathbf{L}(\mathbf{z})$ , that is a row basis for  $E[\mathbf{Y}|\mathbf{z}]$ , where  $r \leq n$  [2, 7]. This low dimensional matrix  $\mathbf{L}(\mathbf{z})$  can be thought of as the manifestation of the latent variables in the genomic data. As illustrated in Fig. 1, this conditional factor model is common for biomedical and genomic data [26]. Since  $\mathbf{z}$  is never directly observed or utilized in the model, we will abbreviate  $\mathbf{L}(\mathbf{z})$  as  $\mathbf{L}$ . This yields the model

$$\mathbf{Y} = \mathbf{BL} + \mathbf{E} \tag{1}$$

where  $\mathbf{B}$  is a  $m \times r$  matrix of unknown parameters of interest. The  $i^{\text{th}}$  row of  $\mathbf{B}$ , which we write as  $\mathbf{b}_i$ , quantifies the relationship between the latent variable basis  $\mathbf{L}$  and genomic variable  $\mathbf{y}_i$ . This model (1) is schematized in Fig. S7.

The principal components (PCs) of  $\mathbf{Y}$  may be calculated by taking the singular value decomposition (SVD) of  $\mathbf{Y}$ . This yields  $\mathbf{Y} = \mathbf{UDV}^T$  where  $\mathbf{U}$  is a  $m \times n$  orthonormal matrix,  $\mathbf{D}$  is a  $n \times n$  diagonal matrix, and  $\mathbf{V}$  is a  $n \times n$  orthonormal matrix. PCs are then the rows of  $\mathbf{DV}^T$ , where the  $i^{\text{th}}$  PC is found in the  $i^{\text{th}}$  row of  $\mathbf{DV}^T$ . The columns of  $\mathbf{U}$  are considered to be the loadings of their respective PCs.

Suppose that the row-space of  $\mathbf{L}$  has dimension  $r$ . The top  $r$  PCs may then be utilized to estimate the row basis for  $\mathbf{L}$  [8]. Specifically, under a mild set of assumptions, it has been shown that as  $m \rightarrow \infty$ , the top  $r$  PCs of  $\mathbf{Y}$  converge with probability 1 to a matrix whose row space is equivalent to that of  $\mathbf{L}$  [26]. For our estimation purposes, we only need

to consider the  $\mathbf{V}^T$  matrix since this captures the row-space. We would therefore estimate  $\mathbf{L}$  by simply obtaining the top  $r$  right singular vectors, which we denote by  $\mathbf{V}_r^T$ .

Let’s now consider a concrete example of  $\mathbf{z}$ ,  $\mathbf{L}$ ,  $\mathbf{V}_r^T$ , and the ultimate inference goal. [27] carried out a gene expression study to identify cell cycle-regulated genes of *Saccharomyces cerevisiae* (Fig. 2). In this experiment,  $m = 5981$  genes’ expression values were measured over  $n = 14$  time points in a culture of yeast cells whose cell cycles had been synchronized. Here,  $\mathbf{z}$  are the latent variables that represent the dynamic gene expression regulatory program over the yeast cell cycle.  $\mathbf{L}$  is the manifested influence of  $\mathbf{z}$  on the observed scale of gene expression measurements (Fig. 1). The ordered time points themselves do not capture the underlying cell cycle regulation, and it is therefore not clear how to *a priori* accurately model  $\mathbf{L}$ . If  $\mathbf{L}$  were directly observed, then we could identify which genes are cell-cycle regulated by performing a significance test of  $H_0 : \mathbf{b}_i = \mathbf{0}$  vs.  $H_1 : \mathbf{b}_i \neq \mathbf{0}$  for each gene  $i$ .

However, since  $\mathbf{L}$  is not observed, we can instead perform the analogous association test using  $\mathbf{V}_r^T$ . Fig. 2(a) shows the first two PCs of  $\mathbf{Y}$ , where it can be seen that these capture systematic variation that resembles cell-cycle regulation. (It should be noted that the remaining PCs, three and higher, do not capture any systematic variation of interest.) Since the row-spaces of  $\mathbf{L}$  and  $\mathbf{V}_r^T$  ( $r = 2$ ) are theoretically close [26], we can instead utilize the model

$$\mathbf{Y} = \mathbf{\Gamma}\mathbf{V}_r^T + \mathbf{E}', \quad (2)$$

where  $\mathbf{\Gamma}$  is a  $m \times r$  matrix of unknown coefficients. We would then perform a significance test of  $H_0 : \gamma_i = \mathbf{0}$  vs.  $H_1 : \gamma_i \neq \mathbf{0}$  for each gene  $i$ .

Note that if  $\mathbf{V}_r^T \rightarrow \mathbf{L}$  in row-space as  $m \rightarrow \infty$ , then these two hypothesis tests would be equivalent. However, for fixed  $m$ , they are not equivalent. There are two main issues: (i)  $\mathbf{V}_r^T$  is a noisy estimate of  $\mathbf{L}$ ; (ii)  $\mathbf{V}_r^T$  is itself a function of  $\mathbf{Y}$ , so hypothesis testing on  $\mathbf{Y} = \mathbf{\Gamma}\mathbf{V}_r^T + \mathbf{E}'$  results in over fitting as shown in the left panel of Fig. 3. Our proposed method deals with problem (ii) by accounting for the over-fitting that is intrinsic to performing hypothesis testing on model (2). The numerical results in this paper are carried out so that we generate the data from model (1) and evaluate the accuracy of the significance based on the truth from model (1). Therefore, we provide direct evidence that our proposed method accounts for both issues (i) and (ii).

## Proposed Algorithms

We have developed a resampling method (Fig. 4) to obtain accurate statistical significance measures of the associations between observed variables and their PCs, accounting for the over-fitting characteristics due to computation of PCs from the same set of observed variables. The proposed algorithm replaces a small number  $s$  ( $s \ll m$ ) of observed variables with independently permuted “synthetic” null variables, while preserving the overall systematic variation in the data. We denote the new matrix with the  $s$  synthetic null variables replacing their original values as  $\mathbf{Y}_{m \times n}^*$ . This is simply the original matrix  $\mathbf{Y}$  with the  $s$  rows of  $\mathbf{Y}$  replaced by independently permuted versions. On each permutation data set  $\mathbf{Y}^*$ , we calculate association statistics for each synthetic null variable, exactly as was done on the

original data. We carry this out  $B$  times, effectively creating  $B$  sets of permutation statistics. The association statistics calculated on  $\mathbf{Y}$  are then compared to the association statistics calculated on *only the  $s$  synthetic null rows of  $\mathbf{Y}^*$*  to obtain statistical significance measures.

Algorithm to Calculate Significance of Variables Associated with PCs

1. Obtain  $r$  PCs of interest,  $\mathbf{V}_r^T$  by applying SVD to the row-centered matrix  $\mathbf{Y}_{m \times n} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ .
2. Calculate  $m$  observed F-statistics  $F_1, \dots, F_m$ , testing  $H_0 : \boldsymbol{\gamma}_i = \mathbf{0}$  vs.  $H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}$  from model (2).
3. Randomly select and permute  $s$  rows of  $\mathbf{Y}_{m \times n}$ , resulting in  $\mathbf{Y}_{m \times n}^*$ .
4. Obtain  $\mathbf{V}_r^{*T}$  from SVD applied to  $\mathbf{Y}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}$ .
5. Calculate null F-statistics  $F_1^{0b}, \dots, F_s^{0b}$  from the  $s$  synthetic null rows of  $\mathbf{Y}^*$  as in Step 2, where  $\mathbf{V}_r^T$  is replaced with  $\mathbf{V}_r^{*T}$ .
6. Repeat steps 3-5  $b = 1, \dots, B$  times to obtain a total  $s \times B$  of null F-statistics.
7. Compute the p-value for variable  $i$  ( $i = 1, \dots, m$ ) by:

$$p_i = \frac{\#\{F_j^{0b} \geq F_i; j = 1, \dots, s, b = 1, \dots, B\}}{s \times B}$$

8. Identify statistically significant tests based on the p-values  $p_1, p_2, \dots, p_m$  (e.g., using false discovery rates).

We call this approach the *jackstraw* for the following reason. By permuting a relatively small amount of observed variables in the original matrix, the underlying systematic variation due to latent variables is preserved as a whole. This makes the PCs of  $\mathbf{Y}^*$  almost identical to the PCs of the original data,  $\mathbf{Y}$ , up to variation due to over-fitting of the noise. Replacing  $s$  variables with null versions is analogous to the game of jackstraws where the goal is to remove one stick at a time from a structured set of sticks without disrupting the overall structure of the sticks. Since the overall structure of  $\mathbf{Y}$  is preserved in  $\mathbf{Y}^*$ , we know that the level of associations between these synthetic null variables and the top  $r$  PCs is purely due to the over-fitting nature of PCA. From these synthetic null statistics, we can therefore capture and adjust for the over-fitting among the original statistics.

A balance between the number of resampling iterations  $B$  and the number of synthetic null variables  $s$  is relevant to the speed of the algorithm and the accuracy of the resulting p-values. One extreme is to set  $s = 1$ , where the accuracy of the p-values is maximized while the algorithm is the least efficient. In each resampling iteration,  $s$  determines the number of estimated null statistics, so to get the same resolution of a particular empirical null distribution ( $s \times B$  total null statistics),  $B$  must increase proportionally with a decreasing

$s$ . For example,  $s = 1$  and  $B = 10000$  yields the same number of null statistics as  $s = 100$  and  $B = 100$ . The number of true null variables in  $\mathbf{Y}^*$  is always greater than or equal to the number of true null variables in the original matrix  $\mathbf{Y}$ . Therefore an increase of  $s$  in the proposed algorithm may lead to a greater over-fitting into the noise of  $\mathbf{Y}^*$  relative to the over-fitting in  $\mathbf{Y}$ , resulting in conservative estimates of significance. Due to this favorable trade-off between  $s$  and  $B$ , the proposed algorithm is always guarded against anti-conservative bias.

The hypothesis test  $H_0 : \boldsymbol{\gamma}_i = \mathbf{0}$  vs.  $H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}$  applied to model (2) may be generalized to performing the test on subspaces spanned by the PCs, shown in the *Supporting Information*. This generalization allows one to perform the association tests on a subset of PCs, while adjusting for other PCs. It also allows for one to consider rotations of  $\mathbf{V}_r^T$  and projections of  $\mathbf{V}_r^T$  onto relevant subspaces. For example, it is possible to rotate the PCs to obtain “independent components” from independent component analysis (ICA) [9] and then perform our algorithm on any desired subset of the independent components.

## Results

We evaluated the proposed method on simulated data so that we could directly assess its accuracy, and we also applied the method to two genomic datasets to demonstrate its utility in practice.

### Simulation Studies

Through a set of simulation studies, we demonstrated that the proposed method is able to accurately estimate the statistical significance of associations between the latent variable basis  $\mathbf{L}$  and observed variables  $\mathbf{y}_i$  (where  $i = 1 \dots m$ ). The data in our simulation studies were generated from model (1)  $\mathbf{Y} = \mathbf{BL} + \mathbf{E}$ , where variables  $\mathbf{y}_i$  corresponding to  $\mathbf{b}_i = \mathbf{0}$  are, by definition, the “null variables” not associated with  $\mathbf{L}$  (Fig. S7). The accuracy of our approach is evaluated by performing  $m$  hypothesis tests using the proposed algorithm (where only  $\mathbf{Y}$  is observed) and assessing whether the joint distribution of p-values corresponding to the null variables is correctly behaved.

#### The joint null criterion.

We utilized the “joint null criterion” of [28] to assess whether the set of p-values corresponding to the null variables follow the desired joint distribution (Fig. S8). When the null hypothesis is true, a valid statistical hypothesis test must generate “null” p-values that are distributed uniformly between 0 and 1. We evaluated the uniformity of null p-values using the Kolmogorov-Smirnov test (KS test) which is designed to detect any deviation from the Uniform(0,1) distribution. The set of null p-values produced by a method satisfies the joint null criterion in a given simulation scenario when their joint distribution is equivalent to a set of i.i.d. observations from the Uniform(0,1) [28].

There are two ways in which we measured deviations from the Uniform(0,1) joint null criterion. The first is via a two-sided KS test, which detects any deviation; the second is a one-sided KS test, which detects anti-conservative deviations where the null p-values are skewed towards zero. Anti-conservative deviations will occur when a method does not properly take into account the fact that the association statistics are formed between the variables and PCs (which have been built from the variables themselves), leading to over-fitting and anti-conservative p-values. Evaluation of the joint null criterion works by simulating many data sets (corresponding to independently repeated studies) from a given data generating process (Fig. S8). The joint behavior of the null p-values is then evaluated among these.

We considered 16 simulation scenarios, described below. For a given scenario, we simulated 500 independent studies and calculated 500 KS test p-values, each of which is based on the set of null p-values from its respective study. In other words, for 500 simulation data sets per scenario, 500 KS test p-values are calculated to measure deviations from the Uniform(0,1); a second application of the KS test (so-called “double” KS test) is then performed on these 500 KS p-values to assess whether any anti-conservative deviation from the Uniform(0,1) among these studies has occurred (Fig. S8). If the statistical method being evaluated provides accurate measures of statistical significance, the collection of double KS test p-values must be distributed Uniform(0,1). This guards against any single simulated data set leading one to an incorrect conclusion by chance.

Overall, we demonstrate that our proposed method provides accurate measures of statistical significance of the associations between variables and the latent variables, when the latent variables themselves are directly estimated from the data via PCA. At the same time, we show that the conventional method does not provide accurate statistical significance measures.

### Simulation scenarios and results.

We constructed 16 simulation scenarios representing a wide range of configurations of signal and noise (Fig. 5), with 500 independent studies simulated from each. Let’s first consider one of the simpler scenarios in detail. Model (1) is used to generate the data. In this particular scenario, we have  $m = 1000$ ,  $n = 20$ ,  $r = 1$  and

$$\mathbf{L} = \frac{1}{\sqrt{n}}(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1, -1, -1),$$

a dichotomous mean shift resembling differential expression between the first 10 observations and the second 10 observations. (The factor  $1/\sqrt{n}$  is to give  $\mathbf{L}$  unit variance.) For 95% of the variables, we set  $b_i = 0$ , implying they are null variables; we parameterize this proportion by  $\pi_0 = 0.95$ . The other 50 non-null variables were simulated such that  $b_i \stackrel{i.i.d}{\sim} \text{Uniform}(0,1)$ . The noise terms are simulated as  $e_{ij} \stackrel{i.i.d}{\sim} \text{Normal}(0,1)$ . The data for variable  $i$  are thus simulated according to  $\mathbf{y}_i = b_i \mathbf{L} + \mathbf{e}_i$ .

For a given simulated data set, we tested for the associations between the observed variables and the latent variables by forming association statistics between the observed  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$  and their collective PC,  $\mathbf{V}_r^T$  ( $r = 1$ ). We calculated p-values using both the



conventional F-test and the proposed method with  $s = 50$  synthetic null variables. Over 500 simulated data sets, the conventional F-test resulted in 500 one-sided KS p-values that exhibit a strong anti-conservative bias with a double KS p-value of  $= 9.71 \times 10^{-196}$  (Fig. S9, black points). On the other hand, the proposed method correctly calculates null p-values, by accounting for the over-fitted measured error in PCA, with a double KS p-value of 0.502 (Fig. S9, orange points). Note that the classification of null p-values is based on the true association status from the population-level data generating distribution from model (1), not based on model (2) or on the observed loadings from the PCA.

We carried out analogous analyses on 15 more simulation scenarios, detailed in Fig. 5. We utilized all possible combinations of the following: (1) either dichotomous or sinusoidal functions for  $\mathbf{L}$ ; (2) the parameters  $\mathbf{B}$  were simulated from either a Bernoulli or Uniform distribution; (3)  $m = 1000$  or  $m = 5000$  variables; and (4) the proportion of true null variables set to either  $\pi_0 = 0.75$  or  $\pi_0 = 0.95$ . The proposed method was applied with  $s = 0.05m, 0.10m,$  and  $0.25m$  to study the impact of the choice of the number of synthetic null variables. For each scenario, we applied the joint null criterion double KS evaluation (Fig. S8), using 500 simulated data sets. The conventional F-test method consistently produced anti-conservative null p-values, while the proposed method yielded accurately distributed null p-values (Fig. 6).

In these simulations, we found that the proposed method tended to produce more conservative null p-values as  $s$  increased (Fig. 6). The explanation for this is that inclusion of a larger number of synthetic null variables leads to a greater over-fitting of PCA to the noise, which in turn yields a conservative empirical null distribution formed by the synthetic null statistics. We therefore identified a trade-off between computational speed and how conservative the calculated p-values are in the choice of  $s$ . We note, however, that the null p-values were never observed to be prohibitively conservative in that the power became unreasonably diminished. In practice, the user has the option to lower the value of  $s$  to minimize this, at the cost of greater computation.

We note that we also investigated a delete- $s$  version of the jackstraw, which draws on ideas from our proposed method, which one could call the permute- $s$  jackstraw. However, this implementation did not produce valid null p-values (*Supporting Information*).

### Testing for associations on subsets of principal components.

We have generalized the proposed method to be able to test for associations on any subset of the top  $r$  PCs, while adjusting for the remaining PCs among the top  $r$ . Here, we demonstrate that the proposed method can identify variables driving a chosen subset of PCs of interest,  $\mathbf{V}_{r_1}^T$ , while adjusting for the remaining of the top  $r$  PCs which are not of interest,  $\mathbf{V}_{r_0}^T$ , where  $r_0 + r_1 = r$ . Based on model (1), we simulated data with  $m = 1000, n = 20, r = 2$  and

$$\mathbf{L}_1 = \frac{1}{\sqrt{n}}(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1),$$

$$\mathbf{L}_2 = \frac{1}{\sqrt{n}}(1, 1, 1, 1, 1, -1, -1, -1, -1, -1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1).$$

$\mathbf{L}_1$  and  $\mathbf{L}_2$  are truly associated with 100 variables and 60 variables, respectively. Among these, 40 variables that are truly associated with both  $\mathbf{L}_1$  and  $\mathbf{L}_2$ . We generated the noise term as  $e_{ij} \stackrel{i.i.d}{\sim} \text{Normal}(0,1)$ . We set  $r = 2$  and tested for associations with the first PC while adjusting for the second PC. Note that the first PC effectively captured the signal from the first latent variable. In this case, the null variables were defined to be 900 variables associated with either only the second latent variable or no latent variable. The conventional F-test resulted in an anti-conservative bias among the null p-values, with a double KS test p-value of  $8.73 \times 10^{-20}$ , while the proposed method produced a correct joint null p-value distribution with a double KS test p-value of 0.352 (Fig. S10).

Furthermore, we generalized the proposed method to be able to test for associations on rotations of the top  $r$  PCs (and on subsets of these rotations). This allows for a general exploration of the row-space spanned by the top  $r$  PCs. The details of these generalizations are given in the *Supporting Information*.

## Application to Gene Expression Studies

Typically, genomic variables are tested for the associations with external variables, which are measured independently of genomic profiling technology, such as disease status, treatment labels, or time points. However, external variables may be imprecise or inaccurate due to poor understanding of the biology or technological limitations; sometimes the external variables of interest may not be capable of being measured at all. For example, in a cancer gene expression study, the cancer types may be based on histological classification of the tumor cells. Then, association tests, such as F-tests, are conducted between the histological classification and transcriptional levels to discover genes of interest. However, the histological classification of cancer tumors may not distinguish important cancer subtypes [29,30]. This lack of information may lead to a spurious signal or reduced power in statistical inference.

When the external variables are unmeasured or imprecise, we are interested in using the latent variable basis,  $\mathbf{L}$ , to discover genes of interest (Fig. 1). Because  $\mathbf{L}$  is never directly measured, we must estimate it from the genomic data, using PCA and related methods. We apply our proposed method to two genomic datasets to demonstrate its utility in practice.

### Cell cycle regulated gene expression in *Saccharomyces cerevisiae*.

It is known that in *Saccharomyces cerevisiae* there is an abundance of genes whose transcription is regulated with respect to the cell cycle [27,31]. Nonetheless, comprehensive identification of the yeast genes whose expression is regulated by the cell cycle is still an active area of research, since it's unclear how the yeast cell cycle regulation should be quantified and modeled [32–35]. The experimental time points after cell population synchronization are readily measured, but this external variable does not directly represent periodic transcriptional regulation with respect to the cell cycle.

Suppose that we want to carry out a hypothesis test on each gene of whether it shows regulation associated with a periodic pattern over the cell cycle. The null hypothesis is then that population mean is not periodic over the cell cycle. This null hypothesis contains

an infinite number of mean time-course trajectories that are non-periodic, making the null hypothesis composite. A composite null hypothesis such as this one is largely intractable because it contains an unwieldy class of potential probability distributions describing gene expression. Indeed, a survey of the literature reveals that this composite null hypothesis is the major challenge when a traditional hypothesis testing approach is taken. However, by using our approach, we can reduce the complexity of this problem by directly estimating the manifested systematic periodic expression variation and applying the proposed method to identify genes associated with this systematic variation due to the latent variables,  $\mathbf{L}$ .

[27] measured transcriptional levels of  $m = 5921$  yeast genes, every 30 min for 390 min after synchronizing the cell cycle among a population of cells by elutriation<sup>2</sup>. The top two PCs capture the manifestation of cell cycle regulation on gene expression [5], explaining 48% of total variance (Fig. 2a,b). By testing for associations between time-course gene expression and the top two PCs, we essentially transform this challenging problem into a tractable association significance testing problem with a simple null hypothesis  $H_0 : \boldsymbol{\gamma}_i = \mathbf{0}$  vs.  $H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}$  (as opposed to a composite null). The hypothesis test is now simply whether gene  $i$  is associated with  $r = 2$  latent variables estimated by the top two PCs.

We applied the proposed method to test this hypothesis and identified a large number of genes associated with yeast cell cycle regulation. We discovered that approximately 84.0% of the 5981 measured genes are associated with the top two PCs ( $\hat{\pi}_0 = 1 - 0.84 = 0.16$ ). At  $\text{FDR} \leq 1\%$ , 2988 genes were found to be statistically significant. Hierarchical clustering of those 2988 genes shows strong, yet diverse, periodic patterns (Fig. 2c). The generalized proposed method allows us to compute statistical significance measures of associations with a subset of PCs. When testing for associations with the first PC while adjusting for the second PC, 1643 genes were called statistically significant at  $\text{FDR} \leq 1\%$ , with the estimated proportion of null variables  $\hat{\pi}_0 = 64.6\%$ . On the other hand, at the same FDR threshold, we found 966 genes were significantly associated with the second PC with  $\hat{\pi}_0 = 59.8\%$ .

### Inflammation associated gene expression in post-trauma patients.

Large-scale clinical genomic studies often lead to unique analytical challenges, including dealing with a large number of clinical variables, unclear clinical endpoints or disease labels, and expression heterogeneity [2]. The ‘‘Inflammation and the Host Response to Injury’’ (IHRI) consortium carried out a longitudinal clinical genomics study on blunt force trauma patients. They collected 393 clinical variables (some longitudinal) and time-course gene expression (total of 797 microarrays) on 168 post-trauma patients [23]. One of the main goals in this study was to elucidate how inflammatory responses after blunt force trauma are manifested on gene expression. To aggregate relevant clinical variables into a manageable daily score, the IHRI consortium used a modified version of the Marshall score to rate the severity of multiple organ dysfunction syndrome [36].

Based on the modified Marshall score trajectories, [23] clustered post-trauma patients into five groups, called ‘‘ordered categorical Multiple Organ Failure’’ (ocMOF) labels. The

---

<sup>2</sup>An inspection of 14 time-series microarray data from [27] reveals an aberrant gene expression profile from 300 min post-elutriation (Fig. S12). Thus, we decided to remove this array in our re-analysis.

time-course gene expression profiles of each patient were summarized by “within patient expression changes” (WPEC) [23]. Then, they tested for correlations between the WPEC genomic variables and the ocMOF score to discover genes associated with inflammatory responses of post-trauma patients. However, the use of the potentially noisy ocMOF clinical variable may impose limitations, as patients with similar Marshall scores may exhibit a wide range of clinical outcomes [37]. Furthermore, five discrete values for the ocMOF scores potentially limits the resolution of the clinical variable.

To investigate this further, we utilized our proposed approach where the gene expression itself was used to construct clinical phenotypes on the patients. We directly used the WPEC data to characterize the molecular signature of inflammatory responses to blunt force trauma. We estimated the manifestation of post-trauma inflammatory responses on gene expression,  $\mathbf{L}$ , with the top 9 PCs (Fig. S13). Then, we applied the proposed method to identify the genomic variables in WPEC associated with the top 9 PCs. The original analysis in [23] estimated 24% of the 54,675 genomic variables (probe sets) to be associated with the ocMOF score. In contrast, our analysis revealed a much larger proportions of the genomic variables to be significantly associated with the major sources of variation, ranging from 62% for 1<sup>st</sup> PC to 39% for 9<sup>th</sup> PC.

The genes identified in the original analysis [23] were largely also identified in our analysis, although our analysis provided many more significant genes. To compare the biological relevance of our re-analysis versus the original analysis, we tested for enrichment of 17 inflammation-related gene sets [38], using one-sided Mann-Whitney-Wilcoxon tests with permutation-based significance. At the  $FDR \leq 1\%$ , none of the inflammatory-related gene sets is enriched for the original analysis using the ocMOF scores [23]. In contrast, a large number of inflammation-related gene sets are significantly enriched when the genomic variables are tested for the associations with the top 9 PCs individually (Table 1). MAPK Signaling is enriched for every PC, except 5<sup>th</sup> PC, whereas Innate Pathogen Detection is enriched for 1<sup>st</sup>, 4<sup>th</sup>, 6<sup>th</sup>, and 9<sup>th</sup> PCs, at the  $FDR \leq 1\%$ . Those two biological pathways were emphasized in the original analysis [23] as indicating down-regulation of innate pathogen detection and up-regulation of MAPK signaling pathway, and they were seen as strong predictors of long-term complications from brute force trauma. Based on enrichment tests, the proposed method appears to provide a biologically richer source of information than the analysis based on the ocMOF scores.

## Discussion

We have developed a method to accurately carry out statistical significance tests of associations between high-dimensional variables and estimated latent variables, which have been estimated through systematic variation present in the observed high-dimensional variables themselves. Our approach is to maintain the overall systematic variation in the high-dimensional data set, while replacing a small number of observed variables with independently permuted synthetic null variables. These synthetic null variables allow us to estimate the null distribution of the association statistics calculated on the original data, that takes

into account the inherent over-fitting that occurs when estimating latent variables through methods such as PCA. We call this approach the *jackstraw* because it draws on the idea of the game of jackstraws, where a player must remove a stick (i.e., a variable) from a pile of tangled sticks without disturbing the overall structure. Through extensive simulations, we demonstrated that the proposed method is capable of accounting for over-fitting and producing accurate statistical significance measures. We also demonstrated that applying conventional association testing methods to this problem artificially inflates the statistical significance of associations.

An input required for the proposed method is the number of PCs,  $r$ , that capture systematic variation from latent variables. Determining the number of “statistically significant” PCs is an active area of research, and defining a number of significant PCs depends on the data structure and the context [15–18, 26]. Setting  $r$  too small results in systematic residual variation in model (2), which hinders the accuracy of the approach. We suggest erring on the side of setting  $r$  larger, and utilizing PCs not of interest as adjustment variables in the more general form of our algorithm. For example, if one would like to identify variables associated with the top three PCs, but is unsure whether the given data has three or four significant PCs, we have found it more robust to input  $r = 4$  which will adjust for potential systematic residual variation captured by the fourth PC.

We demonstrated our approach using PCA. It is well known that individual PCs may not be directly interpretable or may contain multiple signals of interest that the user wishes to distinguish. Our method is capable of considering any subspace spanned by the top  $r$  PCs, while adjusting for the remaining subspace spanned by these PCs (to account for systematic variation to of interest). Our method is also easily adapted to methods such as Independent Component Analysis. Details are given in the *Supporting Information*. This allows the user to intelligently investigate the space spanned by systematic variation and carefully construct estimates of the latent variables of interest. We do not advocate blindly applying our method to the top  $r$  PCs without considering these issues.

The proposed method represents a novel resampling approach operating on variables, whereas established resampling approaches, such as the jackknife and the bootstrap, tend to operate on observations [39–41]. When applying these methods, systematic variation due to latent variables is intentionally perturbed, since their purpose is typically to assess the sampling variation of a single variable. In high-dimensional data, we may need to preserve systematic variation due to latent variables, which is the problem that the jackstraw addresses.

By accurately testing for associations between observed high-dimensional variables and the systematic manifestation of latent variables in the observed variables, our proposed method allows for the automatic discovery of complex sources of variation and the genomic variables that drive them. The proposed method extends PCA and related methods beyond their popular applications in exploring, visualizing, and characterizing the systematic variation to genomic variable level (e.g., gene-level) significance analyses. Given the increasingly important role that non-parametric estimation of systematic variation plays in the analysis of genomic data [1, 2, 5], the proposed method may be useful in many areas of quantitative

biology utilizing high-throughput technologies as well as other areas of high-dimensional data analysis.

### **Acknowledgments**

This research was supported in part by NIH grant HG002913 and Office of Naval Research grant N00014-12-1-0764.

## Tables

Table 1: Q-values from gene enrichment analysis using inflammation-related gene sets

Gene Set	1 <sup>st</sup> PC	2 <sup>nd</sup> PC	3 <sup>th</sup> PC	4 <sup>th</sup> PC	5 <sup>th</sup> PC	6 <sup>th</sup> PC	7 <sup>th</sup> PC	8 <sup>th</sup> PC	9 <sup>th</sup> PC	ocMOF
Adhesion-Extravasation-Migration	0.004	0.034	0.053	0.002	0.144	0.024	0.036	0.003	0.024	0.016
Apoptosis Signaling	0.004	0.018	0.013	0.004	0.036	0.003	0.116	0.006	0.070	0.014
Calcium Signaling	0.021	0.005	0.087	0.100	0.078	0.120	0.046	0.004	0.146	0.078
Complement Cascade	0.116	0.163	0.068	0.012	0.157	0.013	0.167	0.120	0.098	0.196
Cytokine signaling	0.024	0.100	0.033	0.007	0.140	0.003	0.040	0.004	0.066	0.036
Eicosanoid Signaling	0.020	0.031	0.042	0.007	0.163	0.078	0.116	0.122	0.117	0.013
Glucocorticoid/PPAR signaling	0.100	0.034	0.040	0.027	0.182	0.039	0.041	0.005	0.157	0.099
G-Protein Coupled Receptor Signaling	0.133	0.020	0.179	0.046	0.034	0.156	0.026	0.122	0.123	0.039
Innate pathogen detection	0.004	0.077	0.018	0.001	0.087	0.005	0.011	0.011	0.007	0.039
Leukocyte signaling	0.003	0.010	0.001	0.002	0.044	0.001	0.124	0.005	0.014	0.123
MAPK signaling	0.001	0.002	0.007	0.002	0.023	0.002	0.004	0.001	0.002	0.036
Natural Killer Cell Signaling	0.106	0.114	0.015	0.024	0.060	0.039	0.139	0.004	0.036	0.167
NF-kB signaling	0.007	0.020	0.007	0.017	0.120	0.003	0.073	0.025	0.001	0.195
Phagocytosis-Ag presentation	0.025	0.064	0.010	0.011	0.098	0.020	0.013	0.008	0.040	0.205
PI3K/AKT Signaling	0.005	0.001	0.071	0.059	0.163	0.011	0.006	0.024	0.029	0.078
ROS/Glutathione/Cytotoxic granules	0.016	0.007	0.019	0.018	0.158	0.007	0.116	0.058	0.150	0.027
TNF Superfamily Signaling	0.023	0.064	0.070	0.034	0.171	0.007	0.159	0.007	0.078	0.194

Note: Shaded cells indicate q-value  $\leq 0.01$  for a gene set enrichment test.

## Figures

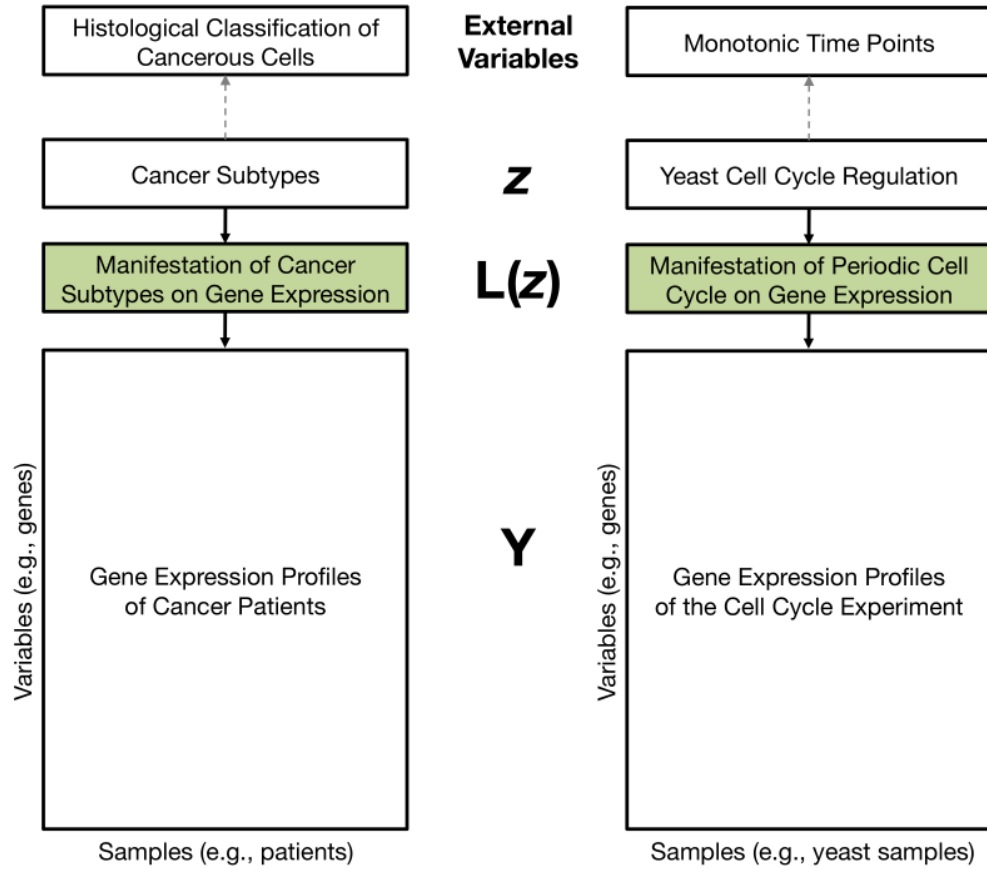


Figure 1: *Illustration of systematic variation genomic data due to latent variables. Complex biological variables, such as cancer subtypes and cell cycle regulation, may be difficult to define, measure, or model. Instead, we can characterize the manifestation of latent variables,  $\mathbf{L}(\mathbf{z})$ , directly from high-dimensional genomic data using PCA and related methods. The proposed method calculates the statistical significance of associations between variables in  $\mathbf{Y}$  and estimates of  $\mathbf{L}$ , while accounting for over-fitting due to the fact that  $\mathbf{L}$  must be estimated from  $\mathbf{Y}$ .*



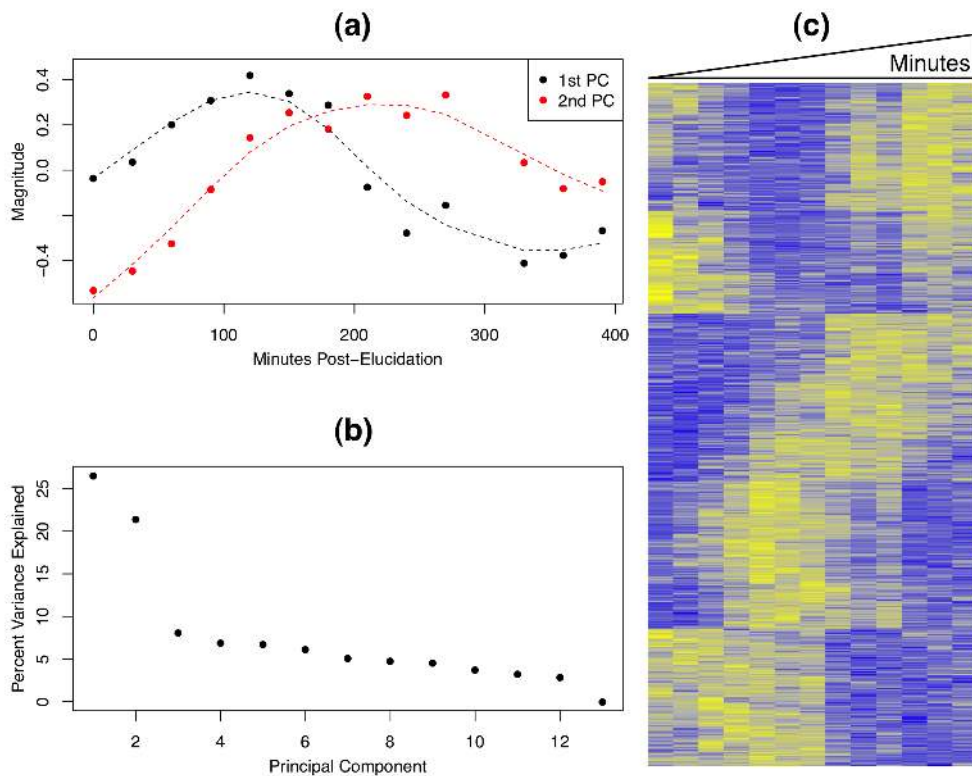


Figure 2: *Identification of yeast genes associated with the cell cycle regulation. (a) The top two PCs of gene expression profiles, measured over time in a population of yeast whose cell cycles have been synchronized by elutriation, capture major transcriptional regulatory patterns [27]. (b) The percent variance explained by PCs shows that the top two PCs capture 48% of the total variance in the data. (c) Hierarchical clustering of transcriptional levels significantly associated with the top two PCs at  $FDR \leq 1\%$  reveals a diverse set of systematic time-course gene expression trajectories.*

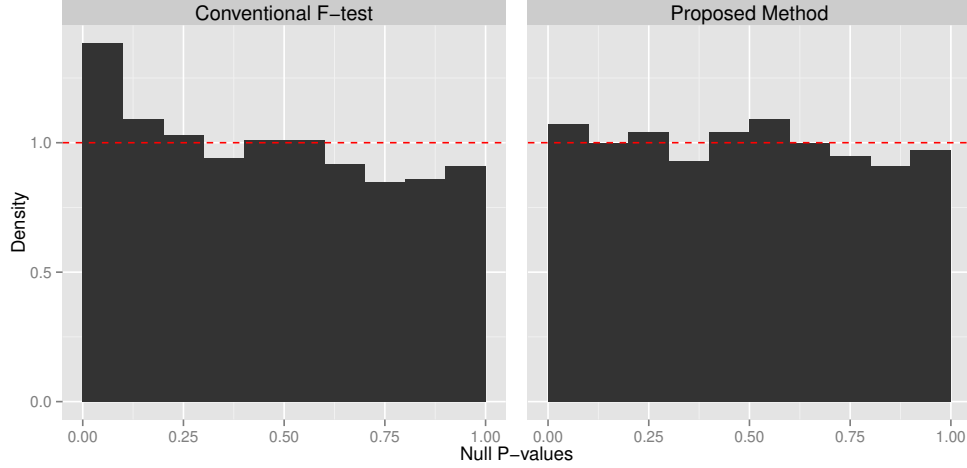


Figure 3: *Evaluation of significance measures of associations between variables and their PCs by comparing true null p-values and the  $Uniform(0,1)$  distribution. (a) The conventional F-test results in anti-conservative p-values, as demonstrated by null p-values being skewed towards 0. (b) The proposed method produces null p-values distributed  $Uniform(0,1)$ . A red dashed line shows the  $Uniform(0,1)$  density function.*

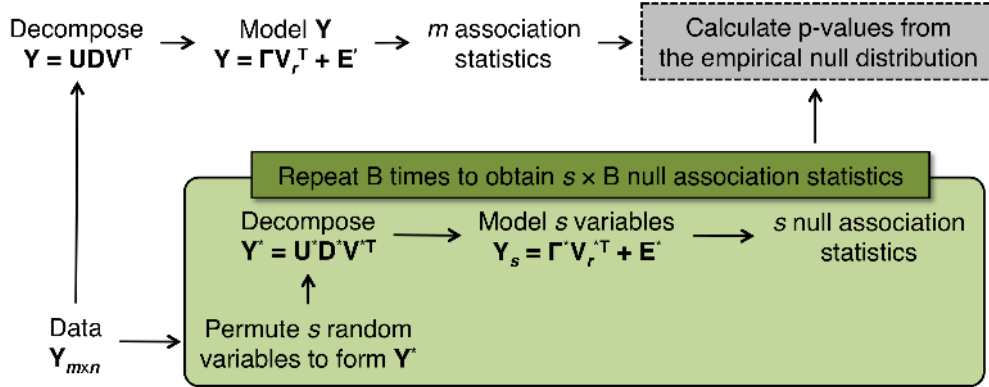


Figure 4: *A schematic of the general steps of the proposed algorithm to calculate the statistical significance of associations between variables (rows in  $Y$ ) and their top  $r$  principal components ( $V_r^T$ ). By independently permuting a small number ( $s$ ) of variables and recalculating the PCs, we generate tractable “synthetic” null variables while preserving the overall systematic variation. Association statistics between the  $s$  synthetic null variables in  $Y^*$  and  $V_r^{*T}$  form the empirical null distribution, automatically taking account over-fitting intrinsic to testing for associations between a set of observed variables and their PCs.*

## Sixteen Simulation Scenarios

All Combinations ( $2^4$ ) of the Following Factors

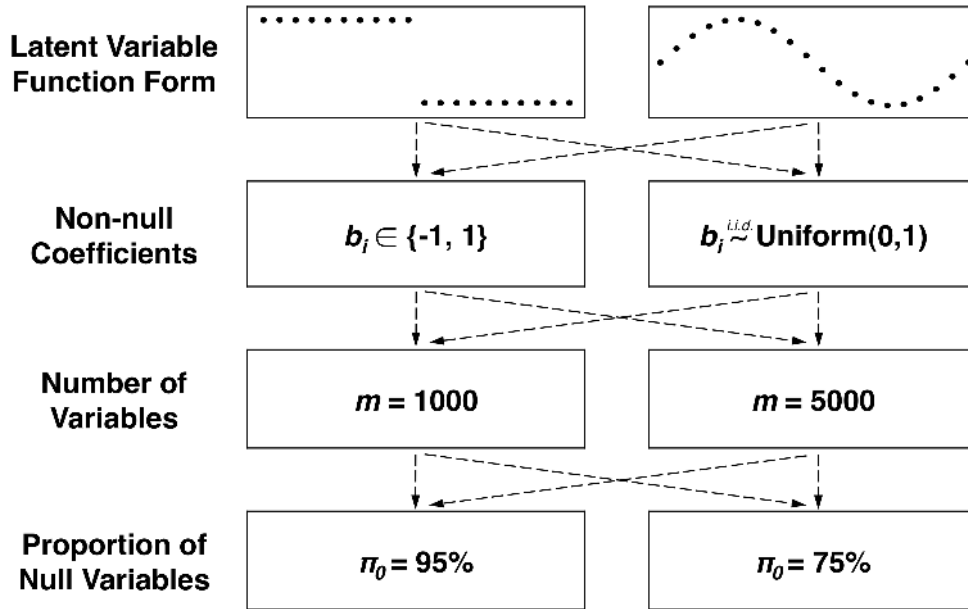


Figure 5: Sixteen simulation scenarios generated by combining four design factors. To assess the statistical accuracy of the conventional  $F$ -test and the proposed method, we simulated 500 independent studies for each scenario, and assessed statistical accuracy according to the “joint null criterion” [28]. For a given simulation study, a valid statistical testing procedure must yield a set of null  $p$ -values that are jointly distributed  $\text{Uniform}(0,1)$ . We use a KS test to identify deviations from the  $\text{Uniform}(0,1)$  distribution. Fig. S8 provides a detailed overview of the evaluation pipeline.

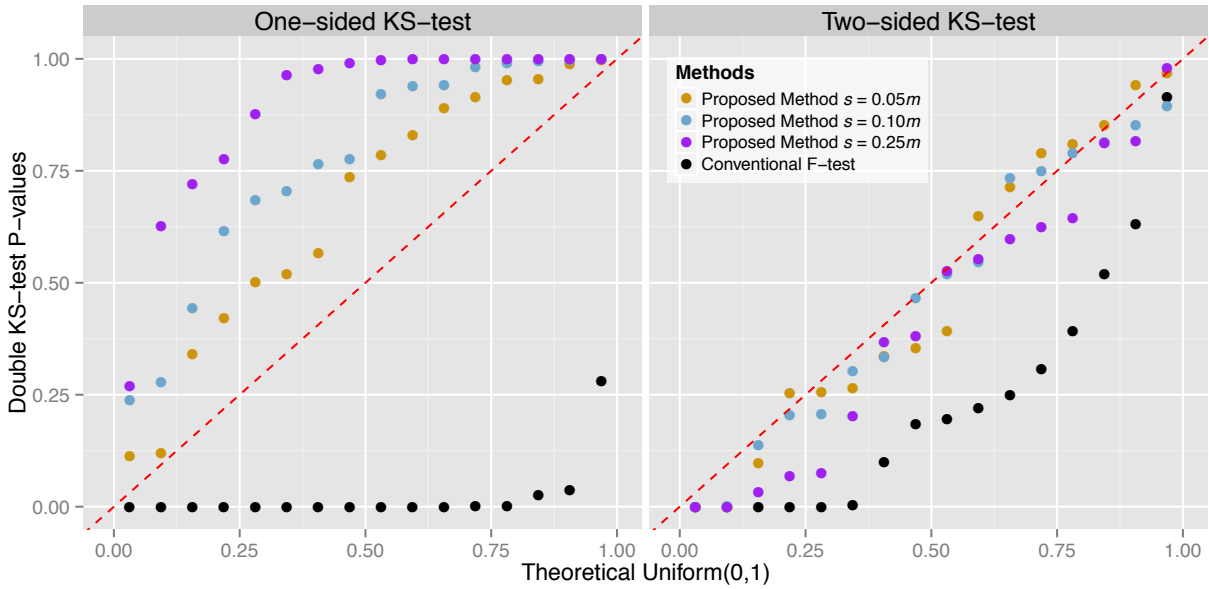


Figure 6: *QQ-plots of double KS test p-values from 16 simulation scenarios versus the  $Uniform(0,1)$  distribution. For each of 500 independent studies per scenario, we tested for deviation of null p-values from  $Uniform(0,1)$ , resulting in 500 KS test p-values for each scenario. An individual point in the QQ-plot represents a double KS test p-value for one scenario, comparing its 500 KS test p-values to  $Uniform(0,1)$ . On the left panel, the systematic downward displacement of 16 black points indicates an anti-conservative bias of the conventional F-test. In contrast, the proposed method produces null p-values that are not anti-conservative. On the right panel, a set of 16 points are below the diagonal red line if the joint null distribution deviates from the  $Uniform(0,1)$  distribution. The proposed method adjusts for over-fitting of PCA and produces accurate estimates of association significance.*

## Supporting Information

### Generalization of the Proposed Algorithm to Subspaces Spanned by Principal Components.

**Statistical Hypothesis Tests.** Here, we expand the hypothesis test,  $H_0 : \boldsymbol{\gamma}_i = \mathbf{0}$  vs.  $H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}$ , of model (2) to a more general test of a linear hypothesis. Let  $\Omega_0$  be the null space of interest and  $\Omega_1$  the alternative space. Suppose that  $\mathbf{C}$  is a  $r \times q$  matrix where  $r \geq q$  and  $\mathbf{a}$  is a  $q$ -vector. Then  $\Omega_0 = \{\boldsymbol{\gamma} \in \mathbb{R}^r : \boldsymbol{\gamma}\mathbf{C} = \mathbf{a}\}$  and  $\Omega_1 = \{\boldsymbol{\gamma} \in \mathbb{R}^r : \boldsymbol{\gamma}\mathbf{C} \neq \mathbf{a}\}$  is a general representation for a null and an alternative space, respectively, of a test of a linear hypothesis. See [42] for a thorough treatment of tests of linear hypotheses.

The hypothesis test  $H_0 : \boldsymbol{\gamma}_i = \mathbf{0}$  vs.  $H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}$  for each gene  $i$  applied to model (2) can be generalized to

$$\begin{aligned} H_0 : \boldsymbol{\gamma}_i &\in \Omega_0 \\ H_1 : \boldsymbol{\gamma}_i &\in \Omega_1. \end{aligned}$$

The proposed algorithm can be modified so that the regression based association test performs this hypothesis test (Steps 2 and 5 in the main text algorithm). This is straightforwardly done by forming F-statistics that are calculated by comparing the unconstrained model fit of model (2) to that under the constraint  $\boldsymbol{\gamma}_i \in \Omega_0$  (see Draper and Smith, 1998).

One important procedure enabled by this generalization is to identify variables associated with a subset of rows of  $\mathbf{V}_r^T$ . Among multiple statistically significant PCs ( $r \geq 2$ ), there may be a subset of PCs of interest, denoted by  $\mathbf{V}_{r_1}^T$ . The complementary subset of PCs that are not of interest is denoted by  $\mathbf{V}_{r_0}^T$ . The proposed method allows one to estimate the significance of associations between observed variables and  $\mathbf{V}_{r_1}^T$ , while adjusting for  $\mathbf{V}_{r_0}^T$ . Operationally,  $\mathbf{V}_{r_0}^T$  has to be included in both the constrained and unconstrained models when calculating the F-statistics, whereas  $\mathbf{V}_{r_1}^T$  is only included in the unconstrained model fit. As illustrated in the simulation study, suppose that  $r = 2$  and we are interested in identifying observed variables associated only with the 1<sup>st</sup> PC. In this case,  $\Omega_0 = \{(0, \gamma_{i,2}) : \gamma_{i,2} \in \mathbb{R}\}$  and  $\Omega_1 = \{(\gamma_{i,1}, \gamma_{i,2}) : \gamma_{i,1} \neq 0 \text{ and } \gamma_{i,1}, \gamma_{i,2} \in \mathbb{R}\}$ . In the proposed algorithm, the 2<sup>nd</sup> PC ( $= \mathbf{V}_{r_0}^T$ ) has to be included in both the constrained and unconstrained models when calculating the F-statistics, whereas the 1<sup>st</sup> PC ( $= \mathbf{V}_{r_1}^T$ ) is only included in the unconstrained model fit. This scenario with two PCs was simulated, and we demonstrated that the proposed algorithm can produce valid p-values, while accounting for 2<sup>nd</sup> PC (right panels in Fig. S10). As expected, the conventional F-test results in artificial inflation of statistical significance (left panels in Fig. S10).

**Rotations of Principal Components.** The top  $r$  PCs,  $\mathbf{V}_r^T$ , collectively capture the systematic variation and estimate the row spanned by  $\mathbf{L}(\mathbf{z})$  (Leek, 2010).  $\mathbf{V}_r^T$  may be rotated while spanning the same row space. The proposed method is capable of estimating the statistical significance of associations between observed variables and any rotation of  $\mathbf{V}_r^T$ .

Consider a  $r \times r$  rotation matrix,  $\mathbf{R}$ . Then, the generalization of model (2) is

$$\begin{aligned} \mathbf{Y} &= \mathbf{\Gamma} \mathbf{R} \mathbf{V}_r^T + \mathbf{E}' \\ &= \mathbf{\Gamma} \mathbf{W}_r + \mathbf{E}'. \end{aligned} \quad (3)$$

To perform a significance test of  $H_0 : \gamma_i \in \Omega_0$  vs.  $H_1 : \gamma_i \in \Omega_1$  for each gene  $i$ , the algorithm needs to adapt the model (3) with  $\mathbf{R} \mathbf{V}_r^T$ , in place of model (2) with  $\mathbf{V}_r^T$ . The rotation matrix  $\mathbf{R}$  must be applied on  $\mathbf{V}_r^{*T}$  in every iteration of the estimation step of null association statistics. This generalization allows one to perform hypothesis tests using various rotations and projections of PCs.

The rotation matrix  $\mathbf{R}$  could reflect biological or clinical measurements, independent of genomic data. By regressing the top  $r$  PCs on a noisy phenotype, we may improve the molecular signature of interest (e.g., separating technical artifacts from biological variation). For example, let's assume a high-dimensional genomic data set contains three statistically significant PCs ( $r = 3$ ) and we are interested in identifying variables associated with a particular linear combination of the top three PCs:  $\mathbf{w}_1 = 0.5\mathbf{v}_1^T - 0.5\mathbf{v}_2^T + \sqrt{0.5}\mathbf{v}_3^T$ . Then, we may construct a  $3 \times 3$  rotation matrix  $\mathbf{R}$ , which must be orthonormal whose determinant equals 1, as required by any proper rotation matrix. In this case, we arrive at

$$\mathbf{R} = \begin{bmatrix} 0.5 & -0.5 & \sqrt{0.5} \\ 0.5 & -0.5 & -\sqrt{0.5} \\ \sqrt{0.5} & \sqrt{0.5} & 0 \end{bmatrix}.$$

After obtaining this particular rotation of the top three PCs,  $\mathbf{W}_r = \mathbf{R} \mathbf{V}_r^T$ , we can continue onto performing a significance test using model (3). The two modifications to the algorithm are:

1. To apply  $\mathbf{R}$  on both  $\mathbf{V}_r^T$  and  $\mathbf{V}_r^{*T}$  to obtain  $\mathbf{W}_r = \mathbf{R} \mathbf{V}_r^T$  and  $\mathbf{W}_r^* = \mathbf{R} \mathbf{V}_r^{*T}$  in the computation of observed association statistics and null association statistics, respectively.
2. To construct a hypothesis test of association between variable  $i$  and  $\mathbf{w}_1$ , while accounting for  $\mathbf{w}_2$  and  $\mathbf{w}_3$ :  $\Omega_0 = \{(0, \gamma_{i,2}, \gamma_{i,3}) : \gamma_{i,2}, \gamma_{i,3} \in \mathbb{R}\}$  and  $\Omega_1 = \{(\gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3}) : \gamma_{i,1} \neq 0 \text{ and } \gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3} \in \mathbb{R}\}$ .

Another interesting example of this generalized jackstraw algorithm arises from transformations of  $\mathbf{V}_r^T$  based on a statistical criterion. Starting with  $\mathbf{V}_r^T$ , Independent Component Analysis (ICA) seeks  $r$  components that are mutually statistically independent (Comon, 1994). In some cases with statistically independent and non-Gaussian sources of variation, independent components may be more sensible, providing an interpretable low-dimensional representation. We may obtain independent components by rotating  $\mathbf{V}_r^T$  to maximize mutual statistical independence (Hastie et al., 2011). Let  $\mathbf{R}$  be a rotation matrix that maps  $\mathbf{V}_r^T$  into the  $r$  independent components. To carry out the association test on the  $r$  independent components, we use the generalized model (3) with  $\mathbf{R} \mathbf{V}_r^T$ , instead of model (2) with  $\mathbf{V}_r^T$ . As above, the analogous substitution occurs at every iteration to ensure the same rotation is applied to synthetic null variables.

Generalized Algorithm to Calculate Significance of Variables Associated with PCs

1. Obtain  $r$  PCs of interest,  $\mathbf{V}_r^T$  by applying SVD to the row-centered matrix  $\mathbf{Y}_{m \times n} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ .
2. Rotate  $\mathbf{V}_r^T$  using a  $r \times r$  rotation matrix  $\mathbf{R}$  to model  $\mathbf{Y} = \mathbf{\Gamma}\mathbf{R}\mathbf{V}_r^T + \mathbf{E}'$ .
3. Calculate  $m$  observed F-statistics  $F_1, \dots, F_m$ , testing  $H_0 : \gamma_i \in \Omega_0$  vs.  $H_1 : \gamma_i \in \Omega_1$  from the Step 2 model.
4. Randomly select and permute  $s$  rows of  $\mathbf{Y}_{m \times n}$ , resulting in  $\mathbf{Y}_{m \times n}^*$ .
5. Obtain  $\mathbf{V}_r^{*T}$  from SVD applied to  $\mathbf{Y}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}$ .
6. Rotate  $\mathbf{V}_r^{*T}$  using a  $r \times r$  rotation matrix  $\mathbf{R}$  to model  $\mathbf{Y}^* = \mathbf{\Gamma}^*\mathbf{R}\mathbf{V}_r^{*T} + \mathbf{E}'^*$ .
7. Calculate null F-statistics  $F_1^{0b}, \dots, F_s^{0b}$  from the  $s$  synthetic null rows of  $\mathbf{Y}^*$ , testing  $H_0 : \gamma_i^* \in \Omega_0$  vs.  $H_1 : \gamma_i^* \in \Omega_1$  from the Step 6 model.
8. Repeat steps 4-7 for  $b = 1, \dots, B$  iterations to obtain a total  $s \times B$  of null F-statistics.
9. Compute the p-value for variable  $i$  ( $i = 1, \dots, m$ ) by:

$$p_i = \frac{\#\{F_j^{0b} \geq F_i; j = 1, \dots, s, b = 1, \dots, B\}}{s \times B}$$

10. Identify statistically significant tests based on the p-values  $p_1, p_2, \dots, p_m$  (e.g., using false discovery rates).

*Remark 1.* If one wants to test linear hypotheses on the unrotated PCs, then Step 2 in the above algorithm is carried out such that  $\mathbf{R}$  is the identity matrix.

*Remark 2.* Step 4 of the generalized algorithm above permutes  $s$  rows of  $\mathbf{Y}$ , thereby breaking all systematic variation spanned by  $\mathbf{V}_r^T$  among these  $s$  variables. If one is testing for associations on only a subspace of  $\mathbf{V}_r^T$ , then it may be desirable to preserve the systematic variation spanned by the subspace of  $\mathbf{V}_r^T$  acting as adjustment variables. (E.g., If one is testing for associations with the first PC where  $r = 2$ , then it may be desirable to preserve the systematic variation spanned by the second PC among these  $s$  variables.) This can be accomplished, for example, by modifying the permutations to act only on the residuals after regressing out the adjustment variables, or by carrying out a bootstrap null procedure whereby residuals resulting from this regression are bootstrapped and added back to the adjustment variable only model fit (see, e.g., [43]).

## An alternative delete- $s$ approach.

We also investigated a delete- $s$  version of the jackstraw, which draws on ideas from our proposed method, the permute- $s$  jackstraw. In the delete- $s$  jackstraw, the set of  $m$  variables is broken up into  $m/s$  disjoint sets of  $s$  variables each, where  $s \ll m$ . The significance of associations between a given set of  $s$  variables and unobserved latent variables is then calculated by testing for the associations between each of the  $s$  variables and the top  $r$  PCs calculated among the remaining  $m - s$  variables. Since the set of  $s$  variables are not used to estimate the latent variables, the association p-values for those  $s$  variables should be marginally valid. However, we found that the delete- $s$  jackstraw did not satisfy the joint null criterion. We applied the delete- $s$  jackstraw to the above 16 simulation scenarios and found that the delete- $s$  jackstraw results in an anti-conservative bias (Fig. S11). As  $s$  increases, the null p-values from the delete- $s$  jackstraw exhibited a greater anti-conservative bias. We hypothesize that this delete- $s$  approach did not produce valid null p-values because of a complex dependence among the  $m/s$  sets of  $s$  p-values (any two sets of  $s$  p-values have an overlap of  $m - 2s$  variables used to construct the PCs used in the association tests), similar to dependence issues encountered in cross-validation.



Supplementary Figures

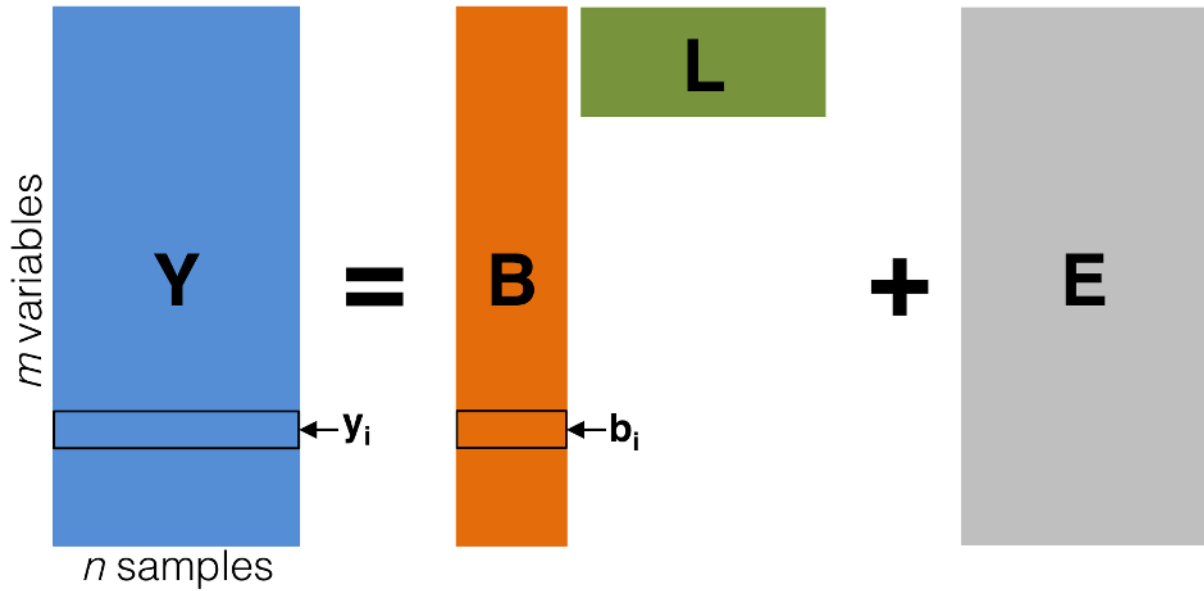


Figure S7: *Diagram of the latent variable model (1). The latent variable basis  $\mathbf{L}$  is not observable, but may be estimated from  $\mathbf{Y}$  using the top  $r$  principal components ( $\mathbf{V}_r^T$ ). The noise term  $\mathbf{E}$  is independent random variation. We are interested in performing statistical hypothesis tests on  $\mathbf{b}_i$  ( $i = 1, \dots, m$ ), which quantifies the relationship between  $\mathbf{L}$  and  $\mathbf{y}_i$  ( $i = 1, \dots, m$ ). Since  $\mathbf{L}$  must be estimated from  $\mathbf{Y}$ , a conventional association test results in anti-conservative  $p$ -values. We have developed the jackstraw method to account for overfitting due to using estimates of  $\mathbf{L}$  to compute the statistical significance of associations between  $\mathbf{L}$  and  $\mathbf{y}_i$ .*

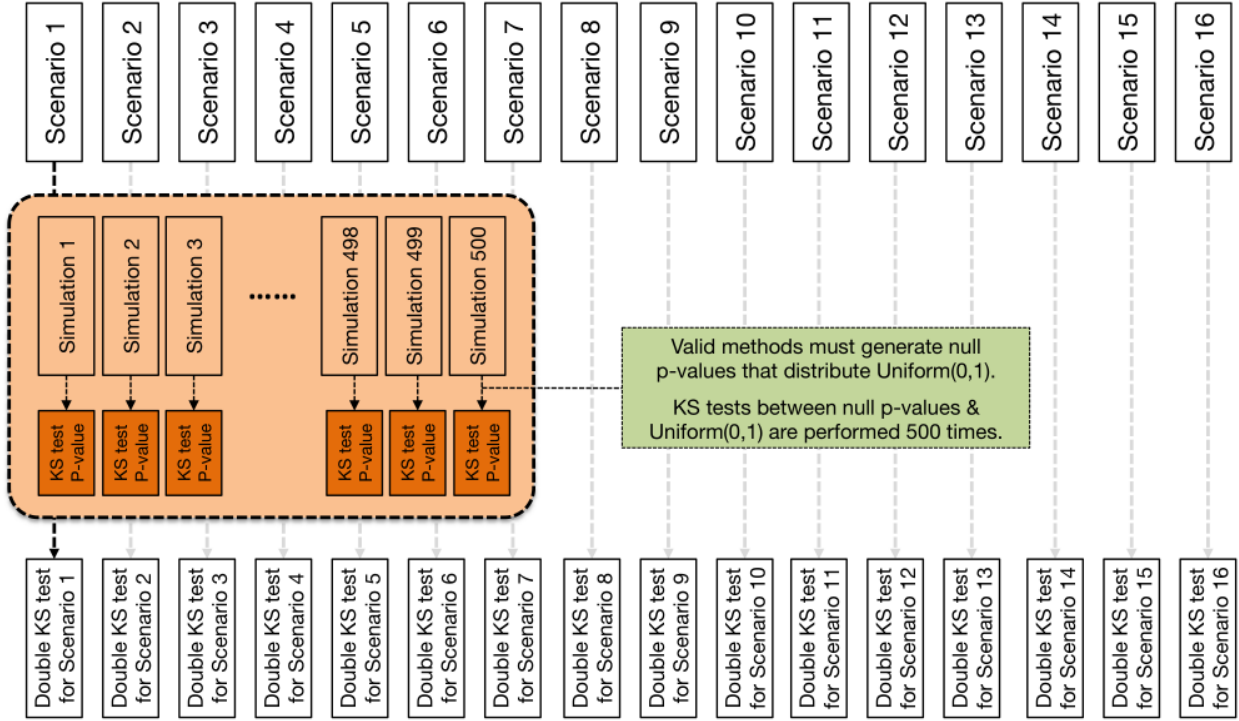


Figure S8: *Evaluation pipeline for 16 simulation scenarios. To assess statistical accuracy in testing the associations between variables and principal components, we generated 16 scenarios from a wide range of configurations. For a given scenario, we simulated 500 independent studies, which resulted in 500 KS test p-values (refer to a green box). For a valid statistical method, this set of 500 KS test p-values should be distributed  $Uniform(0,1)$ ; we evaluate an anti-conservative bias among 500 KS test p-values with another application of the KS test (a “double KS test”). This results in the 16 simulation scenarios being summarized by 16 double KS test p-values (Fig. 6), giving us a comprehensive view of the method’s statistical accuracy.*

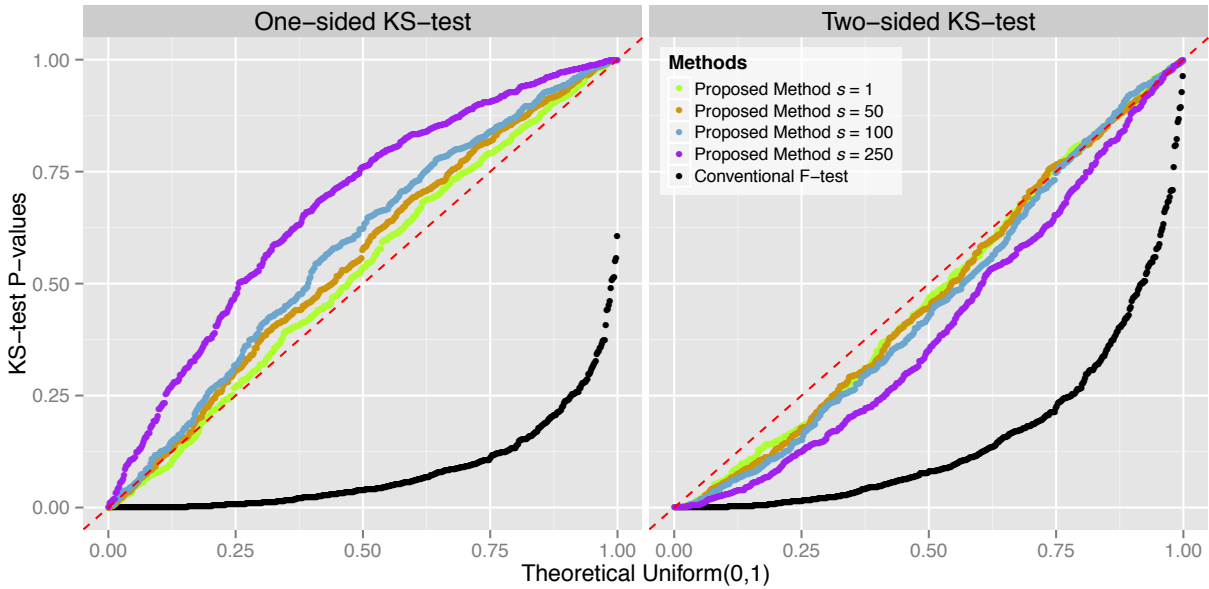


Figure S9: *QQ-plots of KS test p-values from the highlighted simulation scenario in the main text, using the conventional F-test and the proposed method. In this scenario, we generated a dichotomous mean shift between two groups of observations with the true proportion of null variables  $\pi_0 = 0.95$ . We assessed the  $Uniform(0,1)$  property of null p-values using the Kolmogorov-Smirnov (KS) test. In the left QQ-plot comparing one-sided KS test p-values and the  $Uniform(0,1)$  distribution, the downward displacement of the black points below the diagonal red dashed line indicates anti-conservative p-values resulting from the conventional F-test. In contrast, the upward displacement of the colored points demonstrates how the proposed method guards against anti-conservative bias. On the right panel, a two-sided KS test detects any deviation – both conservative and anti-conservative – from the theoretically correct  $Uniform(0,1)$  distribution.*

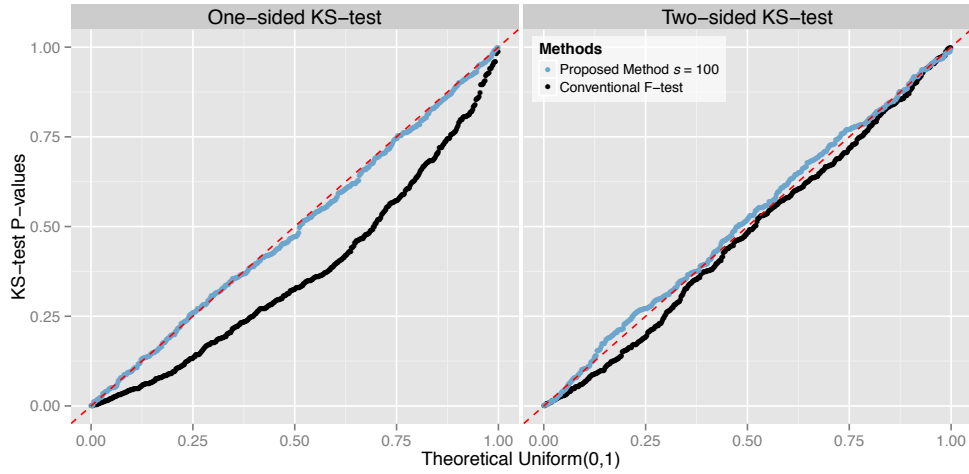


Figure S10: *Evaluation of statistical tests for the associations on subsets of principal components (PCs). We simulated  $m = 1000$  variables of  $n = 20$  observations, with  $r = 2$  significant PCs. We applied the conventional  $F$ -test and the proposed method to test for associations between  $m$  variables and the 1<sup>st</sup> PC. The proposed method is capable of correctly estimating significance measures of the associations between variables and their 1<sup>st</sup> PC, while adjusting for the 2<sup>nd</sup> PC. The conventional  $F$ -test produces anti-conservative  $p$ -values due to over-fitting, where KS test  $p$ -values are skewed towards 0.*

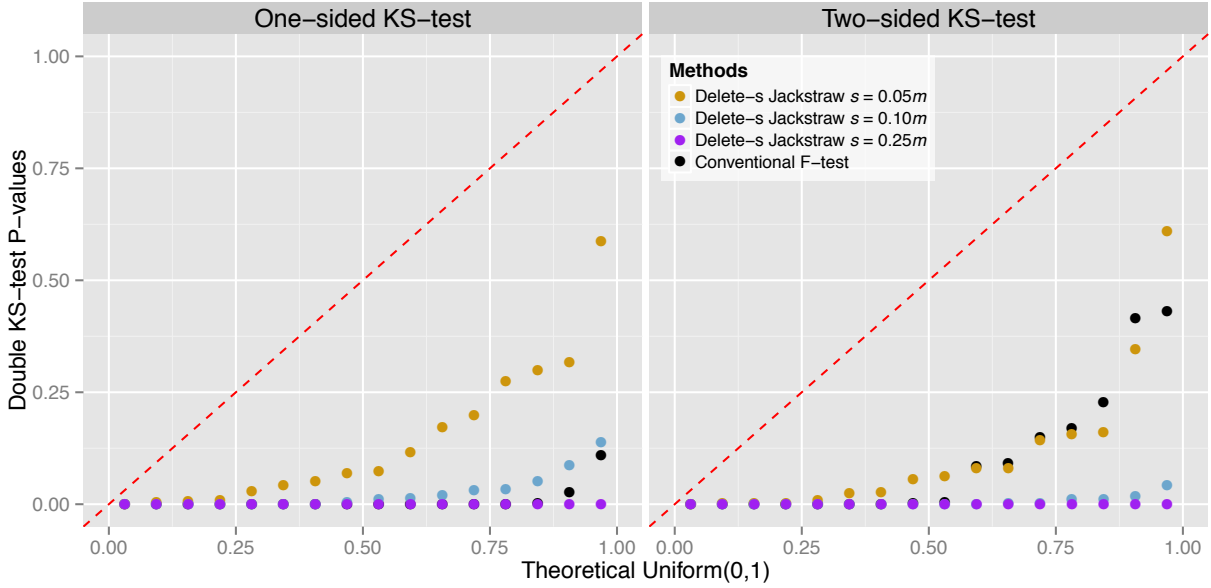


Figure S11: *QQ-plots of double KS-test p-values from applying the delete-s jackstraw on 16 simulation scenarios. In a delete-s version of the jackstraw, the association p-values between the latent variable basis and a given set of  $s$  variables are estimated by testing for the associations between the PCs from  $m - s$  variables and the separate set of  $s$  variables, which were left out in computation of the PCs. A systematic downward displacement of the points below the red diagonal line indicates anti-conservative p-values, present in both the conventional F-test and the delete-s jackstraw method. A larger value of  $s$  in the delete-s jackstraw leads to a greater anti-conservative bias.*

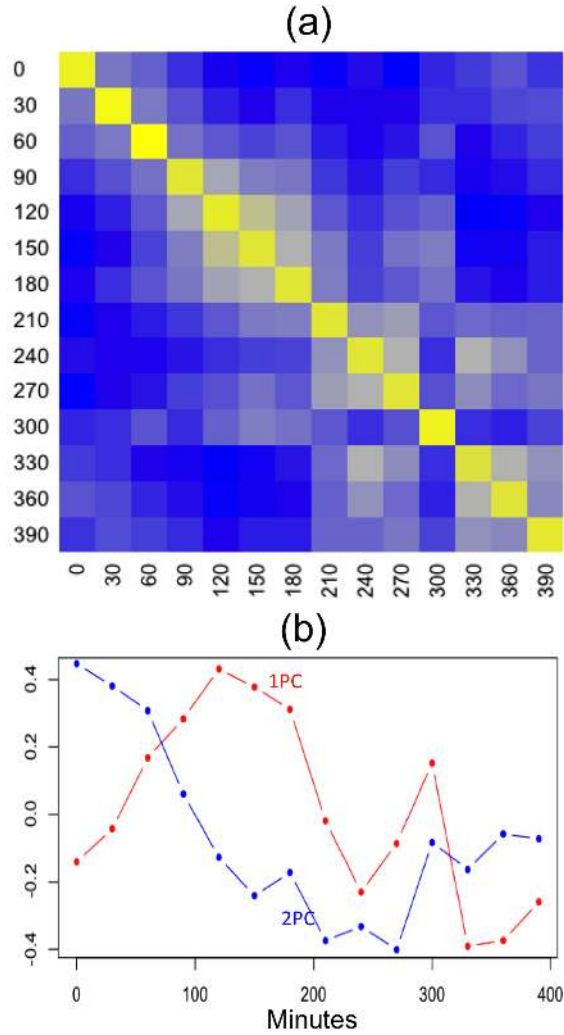


Figure S12: *Original gene expression profiles of the Spellman et al. (1998) yeast cell cycle experiment. (a) A heat map of the covariance matrix of 14 arrays reveals an outlier from the time point at 300 min (low values in blue and high values in yellow). (b) The top two PCs of the original 14 microarrays corroborate an aberrant gene expression profile from 300 min, which is removed from our new analysis.*

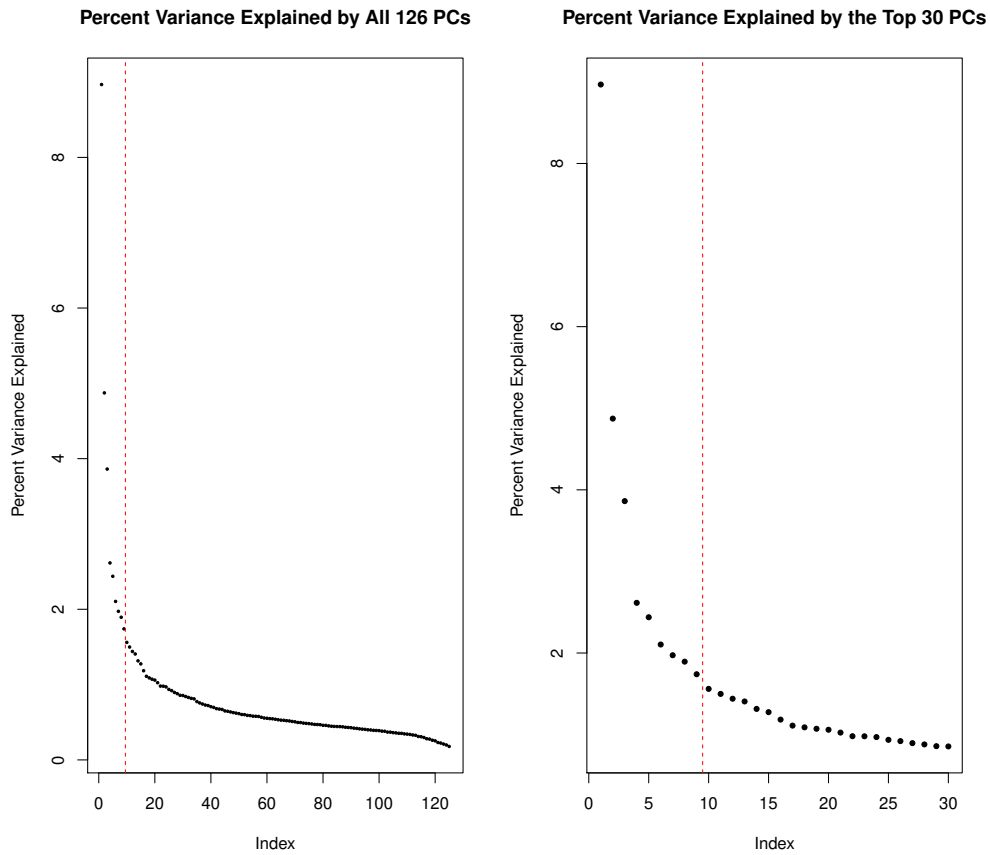


Figure S13: *Percent variance explained by principal components (PCs) of the “Within Patient Expression Change (WPEC)” matrix from Desai et al. (2011). The considerable drop (the “elbow”) in this scree plot helped us to identify the top 9 PCs as significant, likely capturing the molecular signatures of patient responses to blunt force trauma.*

## References

1. Price, A. K., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and D., R. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8), 904–909, Aug (2006).
2. Leek, J. T. and Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**(9), e161 (2007).
3. Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput* **5**, 452–463 (2000).
4. Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., and Fedoroff, N. V. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences of the United States of America* **97**(15), 8409–8414 (2000).
5. Alter, O., Brown, P. O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**(18), 10101–10106 (2000).
6. Zhu, X., Zhang, S., Zhao, H., and Cooper, R. S. Association mapping, using a mixture model for complex traits. *Genet Epidemiol* **23**(2), 181–196, Aug (2002).
7. Leek, J. T. and Storey, J. D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105**(48), 18718–18723 (2008).
8. Jolliffe, I. *Principal Component Analysis*. Springer Verlag, (2002).
9. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, (2011).
10. Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. A modified principal component technique based on the lasso. *JCGS* **12**(3), 531–547 (2003).
11. Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *JCGS* **15**(2), 262–286 (2006).
12. Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**(3), 515–534, Jul (2009).
13. Engelhardt, B. E. and Stephens, M. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**(9), e1001117, Sep (2010).



14. Girshick, M. On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics* **10**(3), 203–224, Sep (1939).
15. Anderson, T. W. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34**(1), 122–148 (1963).
16. Buja, A. and Eyuboglu, N. Remarks on parallel analysis. *Multivariate Behavioral Research* **27**(4), 509–540 (1992).
17. Tracy, C. A. and Widom, H. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177**(3), 727–754 (1996).
18. Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* **29**(2), 295–327 (2001).
19. Peres-Neto, P., Jackson, D., and Somers, K. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology* **84**(9), 2347–2363 (2003).
20. Timmerman, M. E., Kiers, H. A. L., and Smilde, A. K. Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology* **60**(2), 295–314, Dec (2010).
21. Linting, M., van Os, B., and Meulman, J. Statistical significance of the contribution of variables to the PCA solution: An alternative permutation strategy. *Psychometrika* **76**(3), 440–460 (2011).
22. Lehmann, E. *Testing statistical hypotheses*. Springer Verlag, (1997).
23. Desai, K. H., Tan, C. S., Leek, J. T., Maier, R. V., Tompkins, R. G., and Storey, J. D. Dissecting inflammatory complications in critically injured patients by within-patient gene expression changes: A longitudinal clinical genomics study. *PLoS Medicine* **8**(9), e1001093 (2011).
24. Zha, H., He, X., Ding, C., Simon, H., and Gu, M. Spectral relaxation for k-means clustering. In *Neural Information Processing Systems vol.14 (NIPS 2001)*, NIPS 2001, 1057D1064 (, Vancouver, Canada, 2001).
25. Ding, C. and He, X. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine learning, ICML '04*, 225–232 (ACM, New York, NY, USA, 2004).
26. Leek, J. T. Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics* **67**, 344–352 (2010).

27. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9**(12), 3273–3297 (1998).
28. Leek, J. T. and Storey, J. D. The joint null criterion for multiple hypothesis tests. *Statistical Applications in Genetics and Molecular Biology* **10**(1), Article 28 (2011).
29. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* **14**(4), 457–460 (1996).
30. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature* **403**(6769), 503–511, Feb (2000).
31. Cho, R. J., Campbell, M. J., Winzler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**(1), 65–73, Jul (1998).
32. Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* **310**(5751), 1152–1158, Nov (2005).
33. Pramila, T., Wu, W., Miles, S., Noble, W. S., and Breeden, L. L. The forkhead transcription factor *hcm1* regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev* **20**(16), 2266–2278, Aug (2006).
34. Rowicka, M., Kudlicki, A., Tu, B. P., and Otwinowski, Z. High-resolution timing of cell cycle-regulated gene expression. *Proc Natl Acad Sci U S A* **104**(43), 16892–16897, Oct (2007).
35. Wu, W.-S. and Li, W.-H. Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics* **9**, 522 (2008).
36. Marshall, J. C., Cook, D. J., Christou, N. V., Bernard, G. R., Sprung, C. L., and Sibbald, W. J. Multiple organ dysfunction score - a reliable descriptor of a complex clinical outcome. *Critical Care Medicine* **23**(10), 1638–1652 (1995).

37. Cobb, J. P., Mindrinos, M. N., Miller-Graziano, C., Calvano, S. E., Baker, H. V., Xiao, W. Z., Laudanski, K., Brownstein, B. H., Elson, C. M., Hayden, D. L., Herndon, D. N., Lowry, S. F., Maier, R. V., Schoenfeld, D. A., Moldawer, L. L., Davis, R. W., Tompkins, R. G., and Injury, I. H. R. Application of genome-wide expression analysis to human health and disease. *Proceedings of the National Academy of Sciences of the United States of America* **102**(13), 4801–4806 (2005).
38. Loza, M. J., McCall, C. E., Li, L., Isaacs, W. B., Xu, J., and Chang, B. L. Assembly of inflammation-related genes for pathway-focused genetic analysis. *PLoS ONE* **2**(10), e1035 (2007).
39. Quenouille, M. Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B* **11**, 68–84 (1949).
40. Tukey, J. W. Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics* **29**, 614 (1958).
41. Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**(1), 1–26 (1979).
42. Draper, N. R. and Smith, H. *Applied Regression Analysis*. Wiley Series in Probability and Statistics, 3rd edition, (1998).
43. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman & Hall, (1993).