BMC Bioinformatics

METHODOLOGY ARTICLE **Open Access**

Spectral gene set enrichment (SGSE)

H Robert Frost[1,2,3*], Zhigang Li[1,2] and Jason H Moore[1,2,3]

Abstract

Background: Gene set testing is typically performed in a supervised context to quantify the association between groups of genes and a clinical phenotype. In many cases, however, a gene set-based interpretation of genomic data is desired in the absence of a phenotype variable. Although methods exist for unsupervised gene set testing, they predominantly compute enrichment relative to clusters of the genomic variables with performance strongly dependent on the clustering algorithm and number of clusters.

Results: We propose a novel method, spectral gene set enrichment (SGSE), for unsupervised competitive testing of the association between gene sets and empirical data sources. SGSE first computes the statistical association between gene sets and principal components (PCs) using our principal component gene set enrichment (PCGSE) method. The overall statistical association between each gene set and the spectral structure of the data is then computed by combining the PC-level p-values using the weighted Z-method with weights set to the PC variance scaled by Tracy-Widom test p-values. Using simulated data, we show that the SGSE algorithm can accurately recover spectral features from noisy data. To illustrate the utility of our method on real data, we demonstrate the superior performance of the SGSE method relative to standard cluster-based techniques for testing the association between MSigDB gene sets and the variance structure of microarray gene expression data.

Conclusions: Unsupervised gene set testing can provide important information about the biological signal held in high-dimensional genomic data sets. Because it uses the association between gene sets and samples PCs to generate a measure of unsupervised enrichment, the SGSE method is independent of cluster or network creation algorithms and, most importantly, is able to utilize the statistical significance of PC eigenvalues to ignore elements of the data most likely to represent noise.

Background

Gene set testing has become an indispensable tool for the analysis and interpretation of high dimensional genomic data, including measures of DNA sequence variation, DNA methylation, RNA expression and protein abundance [1,2]. By focusing on the collective effect of biologically meaningful groups of genomic variables, rather than just the marginal effect of individual variables, gene set testing methods can significantly improve statistical power, replication of results and biological interpretation. Because of these benefits, significant effort has been devoted over the last decade to building large repositories of functional gene sets [3-5], creating methods for refining and customizing these gene set collections [6-8] and developing effective statistical techniques for gene set testing [9-13].

Gene set testing is normally used to quantify the association between functional groups of genomic variables and a clinical phenotype, e.g., cancer case/control status. Many important use cases exist, however, where a gene set-based interpretation of genomic data is desired in the absence of a phenotype variable, e.g., case-only data collections. For such unsupervised applications, the standard approach for gene set testing involves the computation of the association between gene sets and a categorical variable defined by disjoint clusters of genomic variables. Such methods typically compute the association between each gene set and the variable clustering using either information theoretic measures [14,15] or contingency table-based statistical tests which incorrectly assume independence among the genomic variables [16-18]. Although these techniques provide a measure of

*Correspondence: rob.frost@dartmouth.edu
[1] Institute of Quantitative Biomedical Sciences, Geisel School of Medicine, Lebanon, NH 03756, USA
[2] Section of Biostatistics and Epidemiology, Department of Community and Family Medicine, Geisel School of Medicine, Lebanon, NH 03756, USA
Full list of author information is available at the end of the article

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 2 of 11

gene set enrichment for a given clustering of genomic data, the motivation for most methods is cluster evaluation rather than unsupervised biological interpretation. Cluster-based gene set enrichment results are strongly dependent on the clustering method employed and the number of computed clusters. This sensitivity to the clustering method and number of clusters makes these methods very useful for clustering evaluation but unreliable as general measures of unsupervised gene set enrichment. Specifically, since these methods advocate the use gene set enrichment results to select the clustering method and number of clusters, instead of often unreliable metrics such as the gap statistic [19] or average silhouette width [20], it is unclear what clustering method or number of clusters should be used if the goal is unbiased gene set testing.

An alternative approach for unsupervised gene set testing with many similarities to cluster-based methods is gene set enrichment of gene networks. This approach typically involves the computation of a network from a genomic dataset with network nodes represented by genomic variables, e.g., a co-expression network for gene expression data [21], a community detection algorithm is then used to decompose the network nodes into distinct groups and, finally, gene set testing is performed relative to each community or all communities. If the network communities are treated like gene clusters, the same information theoretic or contingency table-based methods employed for cluster-based enrichment can be used to calculate the association between gene sets and network communities. Approaches have also been developed that directly leverage the network structure to test for the association between gene sets and single network nodes [22] or groups of nodes [23]. Similar to cluster-based approaches, gene set enrichment of networks is highly dependent on the method used to build the network from genomic data and algorithms employed for community detection.

Methods have also been developed to test the association between gene sets and latent variables computed from genomic data sets via techniques such as principal component analysis (PCA) or independent component analysis (ICA). Most of these methods test for the association with just a single latent variable and employ an anti-conservative contingency-table based test on a dichotomized version of the loadings for the latent variable [24-26]. An exception is our recently developed principal component gene set enrichment (PCGSE) method [27] that performs competitive gene set testing relative to each PC using a statistical test that adjusts for correlation among gene set members. Similar to single cluster gene set testing methods, methods that perform gene set testing relative to a single latent variable can only provide an interpretation for a portion of a genomic data

set. To test for the association between gene sets and a collection of latent variables representative of the entire data set, matrix correlation methods [28,29] have been employed, however, such methods are dependent on the number of latent variables included in the test and can only be used for self-contained gene set testing [30] (Q_2 in the terminology of Tian et al. [31]).

Effective methods do not currently exist for unsupervised gene set testing against a competitive null hypothesis that are independent of specific cluster analysis or network analysis approaches. To address this shortcoming, we have developed spectral gene set enrichment (SGSE), an approach for unsupervised competitive testing of the association between gene sets and empirical data sources independent of cluster or network analysis. The SGSE method first computes the statistical association between gene sets and principal components (PCs) using our principal component gene set enrichment (PCGSE) method. The overall statistical association between each gene set and the spectral structure of the data is then computed by combining the PC-level p-values using the weighted Z-method with weights set to the PC variance scaled by lower-tailed p-value computed for the PC variance according to the Tracy-Widom distribution. Although described in the context of gene sets and genomic data, the SGSE method can be used to compute the statistical association between any collection of variable groups and the spectral structure of any empirical data set. To facilitate use of the SGSE method by other researchers, we have included an implementation of the algorithm in the PCGSE R package, which is available from the CRAN repository. Using simulated gene expression data and simulated gene sets, we show that the SGSE method can accurately recover known spectral features from noisy data, features that are undetectable using cluster-based approaches. To illustrate the utility of our method on real genomic data, we compare the performance of the SGSE method and a cluster-based technique on testing the association between MSigDB gene sets and the spectra of two cancer microarray gene expression data sets.

Methods
SGSE inputs
Similar to our principal component gene set enrichment (PCGSE) method [27], the SGSE method takes as input both an $n \times p$ genomic data matrix **X** quantifying p genomic variables under n experimental conditions and an $f \times p$ binary annotation matrix **A** that specifies the association between the p genomic variables and f functional categories.

The genomic data held in **X**, e.g., mRNA expression levels, will be modeled as a sample of n independent observations from a p-dimensional random vector **x**. It is assumed that any desired transformations on

X have been performed and that missing values have been imputed or removed. Although the SGSE method is robust to departures from multivariate normality, as discussed in Section "PC statistical significance" below, it will be assumed that $\mathbf{x} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with correlation matrix **P**. This distributional assumption is usually well justified since sources of genomic data, especially gene expression data, are typically well approximated by a multivariate normal distribution after appropriate transformations.

The rows of **A** represent f biological categories, e.g., KEGG pathways or GO categories, and the elements $a_{i,j}$ hold indicator variables whose value depends on whether an annotation exists between the function i and genomic variable j.

SGSE algorithm

Enrichment of the gene sets defined by **A** relative to the spectra of **X** is performed using the following steps, which are explained in detail in sections "PCA for SGSE" thru "Combined significance of PCGSE p-values" below.

1. Perform PCA on **X**.
2. Determine q, the number of PCs used to represent the spectra of **X**.
3. For all q PCs, use the PCGSE method to compute the statistical significance of the association between each PC and each of the f gene sets defined by **A** according to a competitive null hypothesis.
4. Compute the statistical significance of the association between each of the f gene sets and the spectra of **X** using the weighted Z-method on the q PCGSE p-values with weights based on the PC variances optionally scaled according to PC statistical significance.

PCA for SGSE

Because PCs are not invariant under scaling of the data [28], the PCA solution for SGSE is computed on a mean centered and standardized version of **X**, $\tilde{\mathbf{X}}$. The PC loading vectors and variances of $\tilde{\mathbf{X}}$ are thus the eigenvectors and eigenvalues of:

$$\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} \tag{1}$$

The spectral decomposition of **S** is defined as:

$$\mathbf{S} = \sum_{i=1}^{r_{\tilde{\mathbf{X}}}} \lambda_i v_i v_i^T \tag{2}$$

where $r_{\tilde{\mathbf{X}}}$ is the rank of $\tilde{\mathbf{X}}$, λ_i is the ith eigenvalue of **S** and the variance of the ith PC of $\tilde{\mathbf{X}}$, v_i is the ith eigenvector of **S** and the loadings for the ith PC of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}v_i$ is the ith PC. It is assumed that the eigenvalues are sorted in decreasing

order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r_{\tilde{\mathbf{X}}}}$. Because $\mathbf{x} \sim MVN$, $(n-1)\mathbf{S}$ is approximately Wishart distributed:

$$(n-1)\mathbf{S} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} \sim W(n, \mathbf{P}) \tag{3}$$

Similar to PCGSE, the PCA solution for SGSE is realized via the singular value decomposition (SVD) of a, $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{E}\mathbf{V}^T$, where the columns of **V** represent the PC loading vectors, the entries in the diagonal matrix **E** are proportional to the square roots of the PC variances and the columns of **UE** are the PCs.

PC statistical significance

Random matrix theory (RMT) methods provide useful distributional results for the bulk and extreme eigennvalues of matrices with Wishart distributions [32,33]. As outlined by Johnstone [32], the principal eigenvalue of a sample covariance matrix with a *white* Wishart distribution, where *white* implies that $\boldsymbol{\Sigma} = \mathbf{I}$, tends to a distribution described by a *Tracy-Widom* law of order 1 [34]. Specifically, if $n, p \rightarrow \infty, n/p \rightarrow \eta \geq 1$, then the distribution of the rescaled principal eigenvalue:

$$\frac{\lambda_1 - \mu(p, n)}{\sigma(p, n)} \tag{4}$$

tends to a *Tracy-Widom* law of order 1, where $\mu(p, n) = \frac{(\sqrt{n-1} + \sqrt{p})^2}{n}$ and $\sigma(p, n) = \frac{\sqrt{n-1} + \sqrt{p}}{n}\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{1/3}$.

For $p > n$, the *Tracy-Widom* distribution still holds with p and n simply reversed in the $\mu(p, n)$ and $\sigma(p, n)$ parameter definitions. Although an asymptotic result, this distribution was found to hold well even for p and n values as small as $p = 20$ and $n = 5$ [32]. It also holds well even when the underlying distribution of the elements of $\tilde{\mathbf{X}}$ is not normal [35].

In cases where $\boldsymbol{\Sigma}$ has q variances greater than 1, i.e., a spiked covariance model, Johnstone [32] demonstrated that this distribution approximates the distribution of the $(q+1)^{th}$ eigenvalue but with slightly heavier tails. This result can therefore be used to compute a conservative statistical significance for all of the PCs of $\tilde{\mathbf{X}}$ based on the associated eigenvalues under a null hypothesis of uncorrelated MVN data (e.g., use of PCA for population genetics [36]). Specifically, the statistical significance of PC i can be determined by first computing a *Tracy-Widom* distributed statistic, tw_i, for the eigenvalue, λ_i, associated with the PC using equation 4:

$$tw_i = \frac{\lambda_i - \mu(p-i+1, n)}{\sigma(p-i+1, n)} \tag{5}$$

Under the H_0 that the data is a sample from a MVN distribution with no pair-wise correlation among the individual variables, the p-value for PC i is then computed as

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 4 of 11

the probability of a *Tracy-Widom* law of order 1 statistic more extreme than tw_i:

$$\text{p-value}_{PC_i} = 1 - F_{TW}(tw_i) \tag{6}$$

where $F_{TW}()$ is the cumulative distribution function of a *Tracy-Widom* law of order 1 random variable. This probability can be computed using either numerical lookup tables such as those supported by the RMTstat R package or via the Gamma approximation to the *Tracy-Widom* distribution detailed in Chiani [37]. The SGSE method currently uses the Gamma approximation for more accurate coverage of the tails of the distribution.

Number of PCs used to represent data

The SGSE method supports three options for determining q, the number of PCs used to represent the spectra of $\mathbf{X}$:

1. All PCs with non-zero variance:
   $q = \max_i \lambda_i > 0 = r_{\tilde{\mathbf{X}}}$
2. All statistically significant PCs at a specific α level where statistical significance of a given PC i is determined according (6).
3. A specified number, q^*, with the constraint that q^* cannot be greater than the number of PCs with non-zero variance: $q = q^*$, s.t. $q^* \le r_{\tilde{\mathbf{X}}}$. If specified, q^* will typically be set to a small number, e.g., 1 or 2, to minimize computational cost of the SGSE algorithm.

PCGSE for SGSE

Our PCGSE method [27] is used to compute the statistical significance of the association between each of the f gene sets defined in $\mathbf{A}$ and each of the first q PCs of $\tilde{\mathbf{X}}$ where q is determined using one of the three methods detailed in Section "Number of PCs used to represent data" above. Let the p-value computed via PCGSE for PC i and gene set j be represented using the notation $\text{p-value}_{PC_i,gs_j}$.

Although any supported PCGSE options can be used with SGSE, by default, the SGSE method executes PCGSE using the Fisher-transformed Pearson correlation coefficient between each variable and each PC as the gene-level test statistic and the correlation-adjusted standardized mean difference statistic as the gene set test statistic with statistical significance of the gene set test statistic under a competitive H_0 computed using a two-sided t-test as detailed in Frost et al. [27].

Combined significance of PCGSE p-values

For each of the f gene sets, the p-values computed via PCGSE for the q selected PCs of $\tilde{\mathbf{X}}$ are combined using the weighted Z-method, a generalization of the untransformed Z-transform test [38,39]. The weighted Z-method combines Z-statistics generated for each of multiple independent p-values using weights specific to each p-value. This approach for combining p-values is justified for SGSE

under the assumption of multivariate normality for $\mathbf{x}$ making both the uncorrelated PCs of $\tilde{\mathbf{X}}$, and the p-values generated by PCGSE with respect to those PCs, independent. If $\mathbf{x}$ is significantly non-Gaussian, then the PC-specific p-values will be dependent and techniques such as Kost's method [40] or a generalized version of Fisher's method [41] must be employed instead of the weighted Z-method. In the context of SGSE, a weighed Z-statistic is generated for each of the f gene sets as follows:

$$Z_{gs_j} = \frac{\sum_{i=1}^{q} w_i \Phi^{-1}\left(1 - \text{p-value}_{PC_i,gs_j}\right)}{\sqrt{\sum_{i=1}^{q} w_i^2}} \tag{7}$$

where w_i is a weight specific to PC i of $\tilde{\mathbf{X}}$ and $\Phi^{-1}()$ is the inverse standard normal CDF. Two options are supported for determining the PC-specific weights, w_i:

1. The weight is set to the variance of each PC: $w_i = \lambda_i$
2. The weight is set to the variance of each PC scaled by the lower-tailed p-value computed for the PC variance according to the *Tracy-Widom* distribution as detailed in Section "PC statistical significance": $w_i = (1 - \text{p-value}_{PC_i})\lambda_i = F_{TW}(tw_i)\lambda_i$

The overall p-value representing the statistical significance of the association between gene set j and the spectra of $\mathbf{X}$ is then computed using a one-sided z-test on Z_{gs_j}:

$$\text{p-value}_{gs_j} = 1 - \Phi\left(Z_{gs_j}\right) \tag{8}$$

SGSE evaluation

Benchmark cluster-based gene set testing method

To support comparative evaluation of the SGSE method, we implemented a cluster-based gene set testing method that is representative of a large number of existing cluster and network-based gene set testing methods. Our benchmark cluster-based method computes the statistical significance of the association between gene sets and a data set as follows:

1. Cluster the p genomic variables in $\tilde{\mathbf{X}}$ using k-means clustering with the Hartigan and Wong algorithm [42], 5 restarts and k set according to the global maximum of the gap statistic [19] as computed using the *clusGap()* function in the *cluster* R package [43] with the number of bootstrap resamples defaulting to 100.
2. Compute the statistical significance of the association between each of the f gene sets defined in $\mathbf{A}$ and the k-means clustering using Pearson's χ^2 test of independence on a $2 \times k$ contingency table whose first row holds the counts of gene set members in each of the k clusters and whose second row holds the total size of each of the k clusters.

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 5 of 11

Evaluation using simulated gene sets and simulated data

To explore the type I and type II error rates for the SGSE method and benchmark cluster-based method, a set of simulation studies were performed. In each simulation study, the SGSE method, using both choices for PC weights, and the cluster-based method were used to compute the statistical association between 10 disjoint gene sets, each of size 20, and the spectra of 1000 simulated gene expression datasets each comprised by 50 independent observations of a 200 dimension random vector simulated according to a multivariate normal distribution, $\sim$ MVN($\mathbf{0}, \mathbf{\Sigma}$). The structure of $\mathbf{\Sigma}$ varied between the simulation studies as follows:

1. Test of type I error rate: $\mathbf{\Sigma} = \mathbf{I}$, reflecting a true H_0.
2. Test of power using single factor design: In this case, $\mathbf{\Sigma}$ was generated as a single factor model with $\mathbf{\Sigma} = \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T + \lambda_d \boldsymbol{I}$, where $\lambda_1 = 4$, $\lambda_d = 1$ and $\boldsymbol{\alpha}_1$ is a 200-dimensional vector with all elements equal to 0 except for the first 20 which were set to $\sqrt{.05}$. This population covariance design represents a true association between the first gene set and the first PC.
3. Test of power using two factor design: In this case, $\mathbf{\Sigma}$ was generated as a two-factor model with $\mathbf{\Sigma} = \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T + \lambda_2 \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_2 T + \lambda_d \boldsymbol{I}$, where $\lambda_1 = 4$, $\lambda_2 = 3$, $\lambda_d = 1$, $\boldsymbol{\alpha}_1$ is a 200-dimensional vector with all elements equal to 0 except for the first 10 which were set to $\sqrt{.1}$ and $\boldsymbol{\alpha}_2$ is a 200-dimensional vector with all elements equal to 0 except for the second 10 which were set to $\sqrt{.1}$. This population covariance design represents a true association between the first gene set and the first and second PCs with half of the gene set associated with PC 1 and half associated with PC 2.

For all three simulation studies, the SGSE method was executed on the 1000 simulated datasets using default settings for PCGSE, as specified in Section "PCGSE for SGSE", and both weighting methods outlined in Section "Combined significance of PCGSE p-values". For both the single factor and two-factor simulation designs, the gap statistic method failed to predict any cluster structure, i.e., only a single cluster was predicted. To overcome this issue and prevent a 0 power result for the cluster-based method, the number of clusters used with the benchmark cluster-based method was fixed at k=2 for all simulation cases.

Evaluation using MSigDB C2 v4.0 gene sets and Armstrong et al. leukemia gene expression data

The SGSE method and the benchmark cluster-based method were used to compute the statistical association between the MSigDB C2 v4.0 gene sets and the spectra of the leukemia gene expression data [44] used in the 2005 GSEA paper [9]. The MSigDB C2 v4.0 gene sets and collapsed leukemia gene expression data were both downloaded from the MSigDB repository. With a minimum gene set size of 15 and maximum gene set size of 200, 3,076 gene sets out of the original 4,722 were used in the analysis. The SGSE method was executed on the leukemia gene expression data using all PCs with non-zero eigenvalues, PCGSE was called with default settings as specified in Section "PCGSE for SGSE", and both weighting methods outlined in Section "Combined significance of PCGSE p-values" were employed. The benchmark cluster-based enrichment method was executed as outlined in Section "Benchmark cluster-based gene set testing method" (k=10 was selected as optimal by the gap statistic test). The enrichment of the MSigDB C2 gene sets was also computed relative to the acute myeloid leukemia (AML) versus acute lymphoblastic leukemia (ALL) phenotype using the competitive enrichment method CAMERA [12] with default settings. To quantify how well SGSE and the benchmark cluster-based method were able to capture the known strong association between AML/ALL status and the second PC in the data (see analysis in Frost et al. [27]), the Spearman correlation coefficient was calculated between unsupervised enrichment p-values and phenotype enrichment p-values for all MSigDB C2 gene sets. For gene sets with phenotype enrichment p-values less than 0.05, contingency table statistics were also computed measuring how well SGSE and the cluster-based enrichment method were able to identify MSigDB C2 gene sets significantly associated with the AML/ALL phenotype.

Evaluation using Rosenwald et al. DLBCL gene expression data and MSigDB C2 v4.0 gene sets

The SGSE method and the benchmark cluster-based method were also used to compute the statistical association between the MSigDB C2 v4.0 gene sets and the spectra of the Rosenwald et al. [45] diffuse large B-cell lymphoma (DLBCL) gene expression data. The Rosenwald et al. data set consists of gene expression measurements for 240 patients with DLBCL made using the Lymphochip microarray on 7,399 genes. Microarray data and clinical covariates from the Rosenwald et al. study were both downloaded from the paper's supplemental information web site. To support spectral and phenotype enrichment analysis, the subset of the MSigDB C2 v4.0 gene sets whose members were measured in the Rosenwald et al. data was generated by mapping each of the Lymphochip probes, via Genbank accession numbers, to Entrez gene identifiers and MSigDB C2 v4.0 gene sets. Prior to execution of SGSE and the benchmark cluster-based method, all censored subjects were removed and missing values in the Rosenwald et al. data were imputed using k-nearest neighbor imputation using the impute.knn() function from the

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 6 of 11

R impute package with default settings [46]. With a minimum gene set size of 15 and maximum gene set size of 200, 3,106 gene sets out of the original 4,722 were used in the analysis. The SGSE method was executed on the DLBCL gene expression data using all PCs with non-zero eigenvalues, PCGSE was called with default settings as specified in Section "PCGSE for SGSE", and both weighting methods outlined in Section "Combined significance of PCGSE p-values" were employed. The benchmark cluster-based enrichment method was executed as outlined in Section "Benchmark cluster-based gene set testing method" (k=10 was selected as optimal by the gap statistic test). The enrichment of the MSigDB C2 gene sets was also computed relative to the log of survival time with the competitive enrichment method CAMERA [12] using, as a gene-level test statistic, the z-transformed t-statistic associated with the estimated coefficient from a linear model between gene expression and log survival time. To quantify how well SGSE and the benchmark cluster-based method were able to capture the association between gene set expression and survival time, the Spearman correlation coefficient was computed between unsupervised enrichment p-values and phenotype enrichment p-values for all MSigDB C2 gene sets. For gene sets with phenotype enrichment p-values less than 0.05, contingency table statistics were computed measuring how well SGSE and the cluster-based enrichment method were able to identify MSigDB C2 gene sets significantly associated with log survival time.

Results and discussion
Simulation example
The three simulation studies detailed in Section "Evaluation using simulated gene sets and simulated data" were used to evaluate the type I and type II error rates for the SGSE method (using both weighting options) and the benchmark cluster-based approach. Figure 1 illustrates the results for all three simulation models.

Type I error rate simulation
Because the data was generated according to an identity population covariance matrix, this simulation model is consistent with the H_0 of no association between any of the gene sets and any of the sample PCs. As seen in the quantile-quantile plot of unsupervised gene set enrichment p-values from the simulation study, Figure 1c), the results for all three methods are consistent with this null. At an $\alpha = 0.05$ level, the type I error rate across all 1000 simulated data sets for SGSE with variance weights was 0.02, for SGSE with Tracy-Widom scaled variance weights the type I error rate was 0.034 and for the cluster-based method the type I error rate was 0.031. These results demonstrate that all evaluated methods provide similar, and slightly conservative, control of the type I error rate.

Single-factor power simulation
According to the population covariance matrix used in the single-factor simulation study, only the first gene set should be significantly associated with the first PC. As seen in the Figure 1, plot d), this association is easily detected via the PCGSE method. At an $\alpha = 0.05$ level, the empirical power to detect the association between the first gene set and the spectra of the simulated data for SGSE with variance weights was 0.15, for SGSE with Tracy-Widom scaled variance weights the empirical power was 0.95 and for the cluster-based method the empirical power was 0.92. These results demonstrate that the power of the SGSE method to detect an association in the single factor case is strongly dependent on the choice of weights used to combine the PCGSE-generated p-values in the weighted Z-method, as detailed in Section "Combined significance of PCGSE p-values". The impact of PCGSE p-value weights for the simulation example can be seen in Figure 1 plots b), e) and h). These plots show both the PC variance weights for the simulated datasets as well as weights calculated by scaling the PC variance using the lower-tailed p-value computed using the *Tracy-Widom* distribution for the PC variance. This scaled PC variance weighting results in weights being very close to the standard PC variance weights if the PC variance is highly significant according to the distribution of the principal eigenvalue of a matrix with a white Wishart distribution. As the PC variance becomes less significant, the scaling coefficient decreases lowering the effective weight for the PCGSE-computed p-value associated with that PC. Although the cluster-based method had nearly the same power as the SGSE method with Tracy-Widom scaled variance weights, it is important to note that the gap statistic only identified a single cluster in the data in this case, making the cluster-based χ^2 test for gene set enrichment meaningless if a data-driven approach is taken to determine the number of clusters.

Two-factor power simulation
According to the population covariance matrix used in the two-factor simulation study, the first gene set should be significantly associated with both the first and second PCs. As seen in the Figure 1, plot g), this association is detected via the PCGSE method but the signal is much less strong at the PC level than for the single-factor model. It is only by combining the measured association across all PCs that the SGSE method is able to obtain decent power in such a scenario. At an $\alpha = 0.05$ level, the empirical power to detect the association between the first gene set and the spectra of the simulated data for SGSE with variance weights was 0.31, for SGSE with Tracy-Widom scaled variance weights the empirical power was 0.71 and for the cluster-based method the empirical power was 0.52. These results again demonstrate that the power of
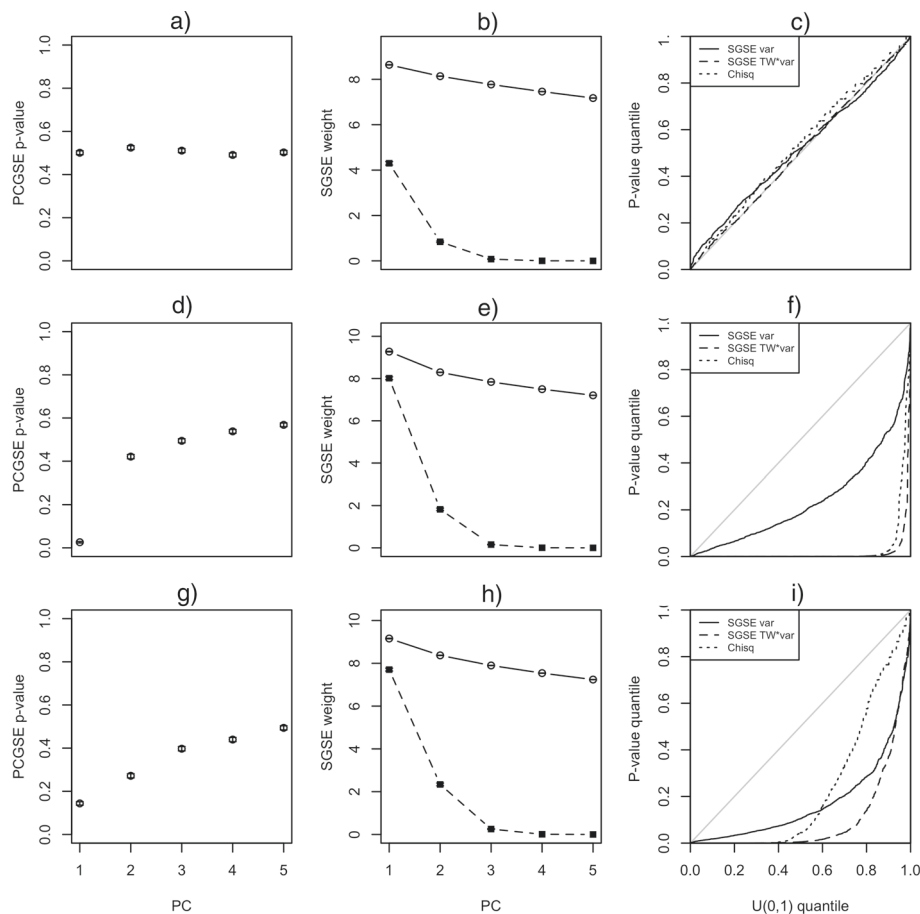
Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 7 of 11

Figure 1 Simulation results. Results for the simulation studies detailed in Section "Evaluation using simulated gene sets and simulated data". For all plots, error bars represent $\pm 1SE$ for the mean value over all 1000 simulated datasets. **a)-c)** Results for the type I error simulation study based on MVN data generated with an identity population covariance matrix. This model is consistent with H_0. **d)-f)** Results for the power simulation study based on MVN data generated according to a single-factor population covariance matrix. Under this model, an association exists between the first gene set and PC 1. **g)-i)** Results for the power simulation study based on MVN data generated according to a two-factor population covariance matrix. Under this model, an association exists between the first gene set and PCs 1 and 2. **a)**, **d)** and **g)** Mean p-values computed using the PCGSE method for the first simulated gene set relative to the first 5 PCs. **b)**, **e)** and **h)** Mean weights used by the SGSE method to combine the PCGSE-computed p-values for each gene set relative to the first 5 PCs. PC variance weights are shown as round points connected by a solid line. PC variance scaled by the lower-tailed p-value computed using the *Tracy-Widom* distribution for the PC variance is shown using square points connected by a dashed line. **c)**, **f)** and **i)** Quantile-quantile plot of the p-values computed using the SGSE method, with both PC variance weights (Var.) or weights defined by the PC variance scaled by the lower-tailed *Tracy-Widom* p-value of the PC variance (TW*Var.), or the benchmark method that uses a Chi-squared test between cluster membership and gene set membership (Chisq).

the SGSE method to detect an unsupervised association is strongly dependent on the choice of weights used to combine the PCGSE-generated p-values in the weighted Z-method. The lower power achieved by the cluster-based method relative to the SGSE method is also noteworthy and is due, in this case, to the fact that portions of the first gene set are associated with both the first and second latent factors. This population covariance design has the effect of generating two separate correlated blocks of variables for the members of this gene set and variable clustering will therefore tend to separate them into distinct clusters, muting the ability of a χ^2 test to identify

an association between cluster membership and gene set membership. Similar to the single-factor simulation, the gap statistic failed to identify more than a single cluster in the datasets simulated according to the two-factor model.

Leukemia gene expression example

The Armstrong et al. [44] leukemia gene expression dataset and MSigDB C2 v4.0 gene sets were selected for SGSE analysis because of the known association between AML/ALL status and the spectra of the gene expression data, as illustrated in Frost et al. [27], the easy accessibility of the data and gene sets from the MSigDB repository

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 8 of 11

and the common use of both the gene expression data and curated gene sets in the gene set enrichment literature (e.g., Subramanian et al. [9]).

Figure 2 shows the association between phenotype and unsupervised gene set enrichment p-values computed using both the benchmark cluster-based method and the SGSE method for the MSigDB C2 v4.0 gene sets, the AML versus ALL phenotype and the Armstrong et al. leukemia gene expression data. Although the true unsupervised enrichment status of the MSigDB C2 v4.0 gene sets relative to the variance structure of the Armstrong et al. [44] gene expression data is unknown, the phenotype enrichment results can be used as a proxy for the true unsupervised gene set enrichment based on the strong association between PC 2 and AML versus ALL status [27] as well as the recent finding by Gorlov et al. [47] that the genes with a large expression variance among cancer cases have a very high likelihood of having a known role in tumor-genesis. As indicated by the correlation between phenotype enrichment and unsupervised gene set enrichment p-values, the SGSE method was able to capture a greater proportion of the AML versus ALL enrichment signal than the benchmark cluster-based method, irrespective of the method used to weight the PC-specific gene set enrichment p-values, with the best performance obtained when PC statistical significance was used to compute the SGSE weights. The benefits of the SGSE method relative to cluster-based enrichment are most clearly visible when considering identification of AML/ALL-associated gene sets via unsupervised enrichment using a phenotype enrichment threshold of $\alpha = 0.1$. In this case, anti-conservative nature of the χ^2 test used in the cluster-based method leads to a high type I error rate

and a very low positive predictive value (PPV) of 0.14 as displayed in plot **(a)**, whereas the SGSE method has a PPV 0.39 when using PC variance weights as displayed in plot **(b)** and a PPV of 0.53 when using as weights the PC variance scaled by the lower-tailed *Tracy-Widom* p-value for the variance as shown in plot **(c)**.

The χ^2 test is anti-conservative in this case because it assumes the contingency table is populated via random sampling from a common distribution. In this case, however, the contingency table is populated using genomic variables whose values are not in fact independent. Thus, the χ^2 test is using a grossly inflated sample size. Similar issues plague other uses of gene sampling in gene set testing, see Goeman et al. [30] for a discussion. Additional file 1 contains the top ten gene sets returned by each of these methods for the Armstrong et al. gene expression data. These lists clearly highlight the performance difference between SGSE and the cluster-based χ^2 test. While SGSE returns a set of cancer related gene sets in the top ten results, the χ^2 test returns gene sets without a clear cancer relationship and with very extreme p-values.

SGSE analysis of the MSigDB C2 v4.0 gene sets and Armstrong et al. [44] leukemia gene expression data illustrates the biological motivation for spectral gene set enrichment, shows the clear superiority of the SGSE approach relative to standard cluster-based gene set tests and demonstrates the importance of PC-specific p-value weights that take into account the statistical significance of each PC.

DLBCL gene expression example

The Rosenwald et al. [45] DLBCL gene expression dataset is another good example of a clear association between the
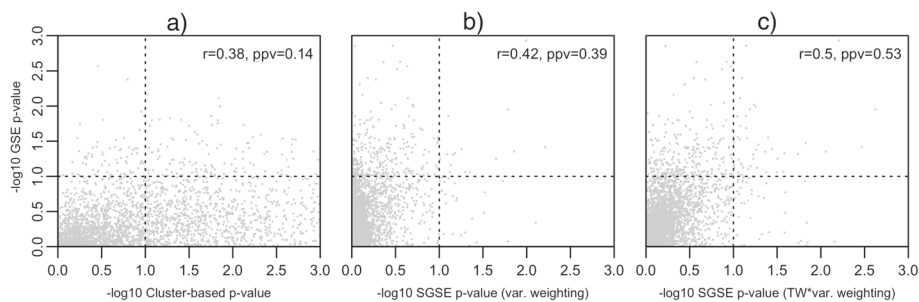

Figure 2 Leukemia gene expression results. Scatter plot showing the association between phenotype gene set enrichment p-values and unsupervised gene set enrichment p-values computed using the benchmark cluster-based method (plot **a)**) and SGSE (plots **b)** and **c)**) for the Armstrong et al. [44] leukemia gene expression data, AML/ALL phenotype, and MSigDB C2 v4.0 gene sets. Phenotype enrichment, unsupervised cluster-based enrichment and spectral gene set enrichment p-values were computed as outlined in Section "Evaluation using MSigDB C2 v4.0 gene sets and Armstrong et al. leukemia gene expression data". Displayed in each plot is the Spearman correlation coefficient between phenotype and unsupervised gene set enrichment p-values and the positive predictive value of unsupervised gene set enrichment for identifying gene sets that are significantly enriched relative to the phenotype at an $\alpha = 0.1$ (shown by dotted lines). The results from the two different SGSE weighting methods outlined in Section "Combined significance of PCGSE p-values" are shown in plots **b)** and **c)** with **b)** plotting SGSE p-values generated using PC variance weighting and **c)** plotting SGSE p-values generated using weights defined by the PC variance scaled by the lower-tailed *Tracy-Widom* p-value for the variance.

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 9 of 11

variance structure of gene expression data and an interesting clinical phenotype, in this case log survival time. Similar to the Armstrong et al. leukemia gene expression data, the Rosenwald et al. DBLCL gene expression data is easily accessible and has been widely reanalyzed in the genomics literature, factors that will support interpretation and replication of the reported SGSE results by other researchers.

Figure 3 shows the association between phenotype and unsupervised gene set enrichment p-values for the MSigDB C2 v4.0 gene sets, log survival time and the spectra of the Rosenwald et al. DLBCL gene expression data. Although the true enrichment status of the MSigDB C2 v4.0 gene sets relative to the variance structure of the Rosenwald et al. gene expression data is unknown, the phenotype enrichment results can again be used as a proxy for the true spectral gene set enrichment based on association between expression variance and cancer-related genes [47]. Although the association between SGSE and cluster-based p-values and phenotype p-values was lower for the Rosenwald et al. DLBCL gene expression data than for the Armstrong et al. leukemia gene expression data, the SGSE method was still able to capture an appreciably greater proportion of the survival time enrichment signal as compared to the benchmark cluster-based method, irrespective of the method used to weight the PC-specific gene set enrichment p-values. Similar to the findings for the leukemia gene expression data, incorporating the PC statistical significance in the SGSE weights improved the Spearman correlation between phenotype enrichment p-values and SGSE p-values for the Rosenwald et al. data. The superior performance of the SGSE method relative to the benchmark cluster-based

method was again most apparent when considering identification of survival time-associated gene sets via unsupervised enrichment using just a phenotype enrichment threshold of $\alpha = 0.1$. In this case, the choice of SGSE weighting method also had a significant impact with a positive predictive value (PPV) of 0.079 for cluster-based enrichment as displayed in plot (a), a PPV of 0.17 for SGSE when using PC variance weights as displayed in plot **(b)** and a PPV of 0.3 when using as weights the PC variance scaled by the lower-tailed *Tracy-Widom* p-value for the variance as shown in plot **(c)**.

Additional file 1 contains the top ten gene sets returned by each of these methods for the Rosenwald et al. gene expression data. Similar to the top gene set lists for the leukemia gene expression data set, these lists highlight the anti-conservative nature of the χ^2 test on this dataset. The fact that the SGSE method with Tracy-Widom scaled variance weights was the only method to includes a gene set directly related to DLBCL in the top ten lends further qualitative support to the efficacy of this approach.

Conclusions

Almost universally, gene set testing is performed in a supervised context to measure the association between functional groups of genes and a clinical phenotype. Many important examples exist, however, where a gene set-based interpretation of genomic data is desired in the absence of a phenotype variable. Although techniques have been developed for unsupervised gene set testing, they predominantly compute enrichment relative to a categorical variable defined by disjoint clusters of the genomic variables. Because such cluster-based methods often use anti-conservative contingency table-based tests
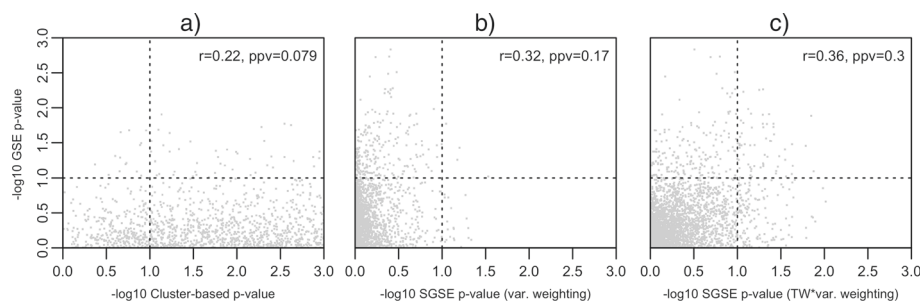

Figure 3 DLBCL gene expression results. Scatter plot showing the association between phenotype gene set enrichment p-values and unsupervised gene set enrichment p-values computed using the benchmark cluster-based method (plot **a)**) and SGSE (plots **b)** and **c)**) for the Rosenwald et al. [45] DLBCL gene expression data, log survival time phenotype, and MSigDB C2 v4.0 gene sets. Phenotype enrichment, unsupervised cluster-based enrichment and spectral gene set enrichment p-values were computed as outlined in Section "Evaluation using Rosenwald et al. DLBCL gene expression data and MSigDB C2 v4.0 gene sets". Displayed in each plot is the Spearman correlation coefficient between phenotype and unsupervised gene set enrichment p-values and the positive predictive value of unsupervised gene set enrichment for identifying gene sets that are significantly enriched relative to the phenotype at an $\alpha = 0.1$ (shown by dotted lines). The results from the two different SGSE weighting methods outlined in Section "Combined significance of PCGSE p-values" are shown in plots **b)** and **c)** with **b)** plotting SGSE p-values generated using PC variance weighting and **c)** plotting SGSE p-values generated using weights defined by the PC variance scaled by the lower-tailed *Tracy-Widom* p-value for the variance.

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 10 of 11

and have performance that is strongly dependent on the clustering algorithm and number of clusters, they are more useful for clustering evaluation than for gene set-based interpretation of genomic data. To address the lack of effective statistical methods for unsupervised competitive gene set testing, we have developed spectral gene set enrichment (SGSE), available in the PCGSE R package from CRAN. The SGSE method first computes the statistical association between gene sets and principal components (PCs) using our principal component gene set enrichment (PCGSE) method. The overall statistical association between each gene set and the spectral structure of the data is then computed by combining the PC-level p-values using the weighted Z-method with weights set to the PC variance scaled by lower-tailed p-values from the Tracy-Widom distribution of the eigenvalue associated with each PC. On both simulated gene sets with simulated data and on curated gene sets with real gene expression data, the SGSE method has been shown to provide superior estimates of unsupervised gene set enrichment relative to standard cluster-based approaches.

Availability of supporting data

The MSigDB C2 v4.0 gene sets can be downloaded from http://www.broadinstitute.org/gsea/msigdb/collections.jsp. The Armstrong et al. [44] leukemia gene expression data can be downloaded from http://www.broadinstitute.org/gsea/datasets.jsp. The Rosenwald et al. [45] DLBCL gene expression data can be downloaded from http://llmpp.nih.gov/DLBCL/. An implementation of the SGSE algorithm is available in the PCGSE R package (version $\geq$ 0.2, http://cran.r-project.org/web/packages/PCGSE/index.html). Due to the dependency on the Bioconductor package safe, it is recommended that PCGSE be installed using the biocLite() function. At the R prompt, enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("PCGSE")
```

Additional file

Additional file 1: Supplemental results for leukemia and DLBCL gene expression datasets. Contains tables listing the top ten most significantly enriched MSigDB C2 v4.0 gene sets returned by SGSE and the cluster-based χ^2 test on the leukemia and DLBCL gene expression datasets.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HRF developed the SGSE methodology, implemented the algorithm, designed the simulation study, performed the real data analysis and drafted the manuscript. ZL participated in the development of the SGSE methodology, assisted with the design of the simulation study and real data analysis and helped draft the manuscript. JHM participated in the development of the SGSE methodology, assisted with the design of the simulation study and real data analysis and helped draft the manuscript. All authors read and approved the final manuscript.

Author details

[1]Institute of Quantitative Biomedical Sciences, Geisel School of Medicine, Lebanon, NH 03756, USA. [2]Section of Biostatistics and Epidemiology, Department of Community and Family Medicine, Geisel School of Medicine, Lebanon, NH 03756, USA. [3]Department of Genetics, Dartmouth College, Hanover, NH 03755, USA.

References

1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2): 1002375. doi:10.1371/journal.pcbi.1002375.
2. Hung J-H, Yang T-H, Hu Z, Weng Z, Delisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinf. 2012;13(3):281–91. doi:10.1093/bib/bbr049.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1): 25–9. doi:10.1038/75556.
4. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
5. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (msigdb) 3.0. Bioinformatics. 2011;27(12):1739–40. doi:10.1093/bioinformatics/btr260.
6. Alterovitz G, Xiang M, Hill DP, Lomax J, Liu J, Cherkassky M, et al. Ontology engineering. Nat Biotechnol. 2010;28(2):128–30. doi:10.1038/nbt0210-128.
7. Davis MJ, Sehgal MSB, Ragan MA. Automatic, context-specific generation of gene ontology slims. BMC Bioinf. 2010;11:498. doi:10.1186/1471-2105-11-498.
8. Frost HR, Moore JH. Optimization of gene set annotations via entropy minimization over variable clusters (emvc). Bioinformatics. 2014;30(12): 1698–706. doi:10.1093/bioinformatics/btu110.
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50. doi:10.1073/pnas.0506580102.
10. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1(1):107–29. doi:10.1214/07-AOAS101.
11. Barry WT, Nobel AB, Wright FA. A statistical framework for testing functional categories in microarray data. Ann Appl Stat. 2008;2:286–315.
12. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40(17):133. doi:10.1093/nar/gks461.
13. Zhou Y-H, Barry WT, Wright FA. Empirical pathway analysis, without permutation. Biostatistics. 2013;14(3):573–85. doi:10.1093/biostatistics/kxt004.
14. Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 2002;12(10): 1574–81. doi:10.1101/gr.397002.
15. Steuer R, Humburg P, Selbig J. Validation and functional annotation of expression-based clusters based on gene ontology. BMC Bioinf. 2006;7: 380. doi:10.1186/1471-2105-7-380.
16. Robinson MD, Grigull J, Mohammad N, Hughes TR. Funspec: a web-based cluster interpreter for yeast. BMC Bioinf. 2002;3:35.
17. Toronen P. Selection of informative clusters from hierarchical cluster tree with gene classes. BMC Bioinf. 2004;5:32. doi:10.1186/1471-2105-5-32.
18. Freudenberg JM, Joshi VK, Hu Z, Medvedovic M. Clean: Clustering enrichment analysis. BMC Bioinf. 2009;10:234. doi:10.1186/1471-2105-10-234.
19. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc. Ser B (Methodological). 2001;63(Part 2):411–23. doi:10.1111/1467-9868.0029.

Frost *et al. BMC Bioinformatics* (2015) 16:70

Page 11 of 11

20. Kaufman L, Rousseeuw PJ. Finding Groups in Data: an Introduction to Cluster Analysis. Hoboken, NJ: Wiley; 2005. http://www.loc.gov/catdir/enhancements/fy0626/2005278659-b.html.
21. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. J Biopharm Stat. 2010;20(2):281–300. doi:10.1080/10543400903572753.
22. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinf. 2005;6:227. doi:10.1186/1471-2105-6-227.
23. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. Enrichnet: network-based gene set enrichment analysis. Bioinformatics. 2012;28(18): 451–7. doi:10.1093/bioinformatics/bts389.
24. Lee S-I, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol. 2003;4(11):76. doi:10.1186/gb-2003-4-11-r76.
25. Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE. Mining gene expression data by interpreting principal components. BMC Bioinf. 2006;7:194. doi:10.1186/1471-2105-7-194.
26. Yao F, Coquery J, Lê Cao K-A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinf. 2012;13:24. doi:10.1186/1471-2105-13-24.
27. Frost HR, Li Z, Moore JH. Principal component gene set enrichment (PCGSE). ArXiv e-prints. 2014arXiv:1403.5148.
28. Jolliffe IT. Principal Component Analysis. Springer Series in Statistics. New York: Springer; 2002. doi:10.1007/b98835.
29. Ramsay JO, Berge J, Styan GPH. Matrix correlation. Psychometrika. 1984;49:403–23. doi:10.1007/BF02306029.
30. Goeman JJ, Buehlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23(8):980–7. doi:10.1093/bioinformatics/btm05.
31. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA. 2005;102(38):13544–9. doi:10.1073/pnas.0506577102.
32. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Ann Stat. 2001;29(2):295–327.
33. Johnstone IM. Approximate null distribution of the largest root in multivariate analysis. Ann Appl Stat. 2009;3(4):1616–33. doi:10.1214/08-AOAS220.
34. Tracy C, Widom H. Level-spacing distributions and the airy kernel. Commun Math Phys. 1994;159(1):151–74. doi:10.1007/BF02100489.
35. Soshnikov A. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. J Statist Phys. 2002;108:1033–56.
36. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLOS Genet. 2006;2(12):190. doi:10.1371/journal.pgen.0020190.
37. Chiani M. Distribution of the largest eigenvalue for real wishart and gaussian random matrices and a simple approximation for the tracy–widom distribution. J Multivariate Anal. 2014;129:69–81. doi:10.1016/j.jmva.2014.04.002.
38. Whitlock MC. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. J Evol Biol. 2005;18(5): 1368–73. doi:10.1111/j.1420-9101.2005.00917.x.
39. Won S, Morris N, Lu Q, Elston RC. Choosing an optimal method to combine p-values. Stat Med. 2009;28(11):1537–53. doi:10.1002/sim.3569.
40. Kost JT, McDermott MP. Combining dependent p-values. Stat Probability Lett. 2002;60(2):183–90. doi:10.1016/S0167-7152(02)00310-3.
41. Dai H, Leeder JS, Cui Y. A modified generalized fisher method for combining probabilities from dependent tests. Front Genet. 2014;5:32. doi:10.3389/fgene.2014.00032.
42. Hartigan JA, Wong MA. A k-means clustering algorithm. Appl Stat. 1979;28(1):100–8. doi:10.2307/2346830.
43. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.1. 2015.
44. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, et al. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet. 2002;30(1):41–7. doi:10.1038/ng765.
45. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. Lymphoma/Leukemia Molecular Profiling Project: The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N Engl J Med. 2002;346(25):1937–47. doi:10.1056/NEJMoa012914.
46. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for dna microarrays. Bioinformatics. 2001;17(6):520–5.
47. Gorlov IP, Yang J-Y, Byun J, Logothetis C, Gorlova OY, Do K-A, et al. How to get the most from microarray data: advice from reverse genomics. BMC Genomics. 2014;15(1):223. doi:10.1186/1471-2164-15-223.