BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1698–1706
doi:10.1093/bioinformatics/btu110

Gene expression

Advance Access publication February 25, 2014

Optimization of gene set annotations via entropy minimization over variable clusters (EMVC)

H. Robert Frost and Jason H. Moore*

Departments of Genetics and Community and Family Medicine, Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH 03755, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Gene set enrichment has become a critical tool for interpreting the results of high-throughput genomic experiments. Inconsistent annotation quality and lack of annotation specificity, however, limit the statistical power of enrichment methods and make it difficult to replicate enrichment results across biologically similar datasets.

Results: We propose a novel algorithm for optimizing gene set annotations to best match the structure of specific empirical data sources. Our proposed method, entropy minimization over variable clusters (EMVC), filters the annotations for each gene set to minimize a measure of entropy across disjoint gene clusters computed for a range of cluster sizes over multiple bootstrap resampled datasets. As shown using simulated gene sets with simulated data and Molecular Signatures Database collections with microarray gene expression data, the EMVC algorithm accurately filters annotations unrelated to the experimental outcome resulting in increased gene set enrichment power and better replication of enrichment results.

Availability and implementation: http://cran.r-project.org/web/packages/EMVC/index.html.

Contact: jason.h.moore@dartmouth.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 31, 2013; revised on January 23, 2014; accepted on February 19, 2014

1 INTRODUCTION

Gene set enrichment is widely used for the analysis and interpretation of the large molecular datasets generated by modern biomedical science (Hung *et al*., 2012; Khatri *et al*., 2012). Despite the development of robust statistical enrichment methods (Efron and Tibshirani, 2007; Subramanian *et al*., 2005; Wu and Smyth, 2012) and extensive functional ontologies such as the Gene Ontology (GO) (Ashburner *et al*., 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and the Molecular Signatures Database (MSigDB) (Liberzon *et al*., 2011) with annotations for many biological molecules across numerous species, the results of enrichment analysis are too often overly general, inaccurate or non-reproducible across experiments (Khatri *et al*., 2012).

Although changes to statistical methods or refinements of functional ontologies can improve enrichment performance, annotation completeness and quality is often a dominant factor driving enrichment accuracy and reproducibility. The annotation of most genes and gene products is incomplete with only a sparse set of annotations to generic high-level categories available (Faria *et al*., 2012). For those annotations that do exist, the overwhelming majority are automatically generated on the basis of sequence or structural similarity without any curatorial review (du Plessis *et al*., 2011; Juncker *et al*., 2009). Such automatically generated annotations have known quality issues relative to manually curated annotations, especially those based on published experimental findings (Bell *et al*., 2012; Dolan *et al*., 2005; Faria *et al*., 2012; Park *et al*., 2011; Schnoes *et al*., 2009; Skunca *et al*., 2012). Electronic annotations are slowly being replaced with higher-quality annotations backed by experimental evidence; however, given the slow pace of experimental validation and manual curation, the preponderance of unreviewed computational annotations and continual generation of new automated annotations, annotation quality will remain a challenge into the foreseeable future.

Current approaches to annotation quality fall into one of several groups: those that create a filtered version of existing annotations, those that subset and/or restructure existing functional ontologies and those that define new, customized, gene sets. In the context of GO, automatic annotation filtering includes methods that use evidence codes, e.g. the MSigDB C5 collection (Liberzon *et al*., 2011), as well as approaches that use the ontology hierarchy to identify and remove redundant annotations (Faria *et al*., 2012). Methods that subset or restructure ontologies include tools for the manual (Binns *et al*., 2009; Carbon *et al*., 2009) or automatic (Davis *et al*., 2010) generation of GO Slims as well as techniques for the information theoretic optimization of the entire GO taxonomy (Alterovitz *et al*., 2010). The process used to generate the MSigDB C4 cancer modules (Segal *et al*., 2004) combines both automatic gene set generation with gene set refinement. In the cancer module process, modules are generated by merging and then refining existing gene sets with gene clusters computed from a large collection of tumor gene expression microarrays.

Although manually customized annotation collections can achieve high specificity, they require domain expertise to create and suffer from ad hoc methods that limit the relevance of any subsequent analysis results. While automatic methods for annotation filtering and ontology sub-setting do not suffer from individual researcher bias, their general purpose nature can prevent them from aligning with the narrow scientific

*To whom correspondence should be addressed.

domain under investigation. An important limitation of many current automatic annotation filtering and ontology subsetting methods is the fact that analysis is only based on the structure of the ontology and the content of the underlying annotation databases. The experimentally observed abundance of the annotated genes and gene products is not used to help identify low-quality annotations or guide ontology restructuring. By focusing on just ontological and annotation data, these methods provide information about the general quality of the annotations and ontology structure, information that is equally relevant to any dataset measuring the annotated genes. Given the large number of proteins with incomplete and therefore coarse-grained functional annotations, a general measure of annotation quality may be a poor predictor of how well annotations will perform within a narrow domain. Even for those approaches that use experimental data, like the process used to create the MSigDB C4 cancer modules, the focus is usually on a broad collection of experimental data (e.g. microarray data for 22 tumor types in the case of the C4 cancer modules), and the output typically combines synthesis of new gene sets with gene set refinement rather than focusing solely on refinement of existing gene sets for a specific experimental context.

Development of high-quality annotations that are specialized to a research domain, yet free from researcher bias, requires techniques that automatically refine annotations using machine learning methods based on representative experimental data. While statistical learning methods are commonly used to predict new annotations from biological data, effective tools are not currently available that apply these techniques for the refinement of existing gene set annotations. To address this gap and enable more accurate and reproducible gene set enrichment analysis, we have developed a novel bioinformatics method, entropy minimization over variable clusters (EMVC), that automatically customizes existing functional annotations for specific sets of biological data. As we demonstrate using simulated gene sets with simulated data and MSigDB collections with microarray gene expression data, the EMVC method accurately filters annotations unrelated to the experimental outcome, resulting in increased gene set enrichment power and better replication of enrichment results.

2 METHODS

Our EMVC algorithm refines gene set annotations to minimize a measure of entropy between each gene set and clusters of genes computed from empirical data. Our method takes as input a collection of functional annotations of genes and gene products (e.g. gene sets from GO, KEGG or MSigDB) and a set of experimental data quantifying the abundance of annotated molecules across multiple experimental conditions. The method outputs the proportion of gene clusterings, averaged over multiple bootstrap resampled datasets, in which each annotation belongs to the minimal entropy solution. Although described in the context of functional gene sets and gene expression data, the EMVC method can be used to optimize any collection of functional annotations given an associated empirical dataset. Mathematical details of the EMVC method, a simple illustrative example and specifics on EMVC evaluation are outlined in the remainder of this section.

2.1 EMVC algorithm

2.1.1 Inputs The EMVC algorithm takes the following data structures as input:

- *Matrix of gene product abundance:* $n \times p$ matrix $\mathbf{X}$ quantifying the abundance of p gene products under n experimental conditions, e.g. mRNA expression levels measured using microarray technology or RNA-seq. These data will be modeled as a sample of n independent observations from a p-dimensional random vector.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \quad (1)$$

  where $x_{i,j}$ represents the abundance of gene product j under condition i. Although the EMVC algorithm does not have specific distributional requirements, sources of genomic data are often well represented by a multivariate normal distribution $\sim \mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$, especially after appropriate transformations. It is assumed that any desired data transformations (e.g. mean centering, standardization, log transformation of mRNA expression ratios) have been performed and that missing values have been imputed or removed for a complete case analysis.

- *Matrix of functional annotations:* $f \times p$ binary annotation matrix $\mathbf{A}$ whose rows represent f different biological functions, e.g. GO categories or KEGG pathways, and whose cells $a_{i,j}$ hold indicator variables whose value depends on whether an annotation exists between the function i and gene product j.

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{f,1} & \cdots & a_{f,p} \end{bmatrix}, a_{i,j} = 1\,[\textit{gene product } j \textit{ has function } i] \quad (2)$$

- *Algorithm parameters:* Required parameters include the variable clustering method (k-means and agglomerative hierarchical clustering using correlation distance are currently supported), the range of cluster sizes (k_{min} to k_{max}) and the number of bootstrap resamples, N.

2.1.2 Entropy measure At the core of our EMVC approach is an entropy measure computed over functional variable groups relative to clusters of variables. In the context of gene sets and genomic data, it is assumed that the p gene products have been divided via a strict partitioning into k clusters with the indicator function $1\,[gene_j \in cluster_l]$ representing the membership state of gene product j within cluster $l, l = 1, \ldots, k$. This clustering can be modeled by f distinct categorical random variables, C_i, one for each function class defined in the annotation matrix A. Each C_i has k categories and a length-k vector of category-specific probabilities Φ^{C_i} with elements $\Phi_l^{C_i}$. The maximum likelihood estimate for the $\Phi_l^{C_i}$ can be computed as the ratio of the number of gene products in cluster l that are annotated to function i over the total number of gene products annotated to function i:

$$\hat{\Phi}_l^{C_i} = \frac{\sum_{j=1}^{p} a_{i,j} 1\,[gene_j \in cluster_l]}{\sum_{j=1}^{p} a_{i,j}} \quad (3)$$

The maximum likelihood estimate for the entropy (Hausser and Strimmer, 2009) of each C_i is therefore as follows:

$$H(C_i) = -\sum_{l=1}^{k} \hat{\Phi}_l^{C_i} log(\hat{\Phi}_l^{C_i}) \quad (4)$$

2.1.3 Annotation optimization Given the data matrix **X**, annotation matrix **A** and required parameters, the EMVC algorithm optimizes **A** using the following core algorithm for a range of k values on each of N bootstrap resampled versions of **X**. The average of all optimized annotation matrices is returned as the final output matrix **O**.

Core algorithm:

- Generate K partitional clusters of the p gene products in **X** using an algorithm such as k-means clustering or a cut of the dendrogram produced by agglomerative hierarchical clustering with correlation distance. Specialized variable clustering methods can also be used, e.g. the principal component analysis-based methods in the R package *ClustOfVar* (Chavent *et al.*, 2012), gene shaving (Hastie *et al.*, 2000), the *varclus* method in the R *Hmisc* package, an R implementation of the SAS VARCLUS procedure.
- For each functional class $i, i = 1, \ldots, f$ whose members are defined by row vector $\mathbf{a}_{i,*}$ of annotation matrix **A**, find the largest subset of annotations that minimizes the entropy measure defined in Equation (4) (i.e. largest minimal entropy subset or LMES). The minimum entropy value of 0 will be achieved when annotations only exist for gene products belonging to a single cluster. Although any cluster with a non-zero number of annotations represents a minimum entropy subset, the EMVC algorithm selects the largest cluster, corresponding to the LMES, to ensure that the fewest annotation changes are made. If multiple clusters are tied for the largest size, a random cluster is selected as the largest. In the case that a functional class has just a single annotation, this annotation will always be a member of the only non-empty cluster and will therefore automatically be retained.
- Create the optimized annotation matrix $\mathbf{A}^*$ by setting $a^*_{i,j} = \mathbb{1}$ [gene$j \in$ LMES for functional class i].

Smoothing across clusterings:

- Generate $\mathbf{A}^*_k$ for $k = k_{min}, \ldots, k_{max}$.
- Average the $\mathbf{A}^*_k$ to create $\mathbf{A}^{sm}$. The elements of $\mathbf{A}^{sm}$ hold the proportion of all variable clusterings in which a particular gene is an element of the LMES.

Bootstrap aggregation:

- Average the $\mathbf{A}^{sm}$ across N bootstrap resampled datasets to form **O** (Breiman, 1996).

2.1.4 Output The EMVC algorithm outputs the $f \times p$ matrix O whose values $o_{i,j}$ reflect the proportion of variable clusterings over all bootstrap resampled datasets in which the annotation of gene product j to function i is kept after entropy minimization.

$$\mathbf{O} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,p} \\ \vdots & \ddots & \vdots \\ o_{f,1} & \cdots & o_{f,p} \end{bmatrix}, o_{i,j} \in [0,1] \quad (5)$$

If an optimized annotation matrix containing binary indicator variables is desired as output, rather than a matrix of proportions, the elements of **O** can be replaced by 0 or 1 according to some desired threshold. For a specific threshold, $\alpha \in [0,1]$, such an $f \times p$ matrix **T** can be generated as follows:

$$\mathbf{T} = \begin{bmatrix} t_{1,1} & \cdots & t_{1,p} \\ \vdots & \ddots & \vdots \\ t_{f,1} & \cdots & t_{f,p} \end{bmatrix}, t_{i,j} = \mathbb{1}[o_{i,j} \geq \alpha] \quad (6)$$

2.2 Simple example

The following simple example illustrates the basic operation of the EMVC method. Assume that just two gene sets are defined over five gene products as specified by the following annotation matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Consider the behavior of the EMVC algorithm for the following two idealized population covariance matrices:

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & \in & \in & 0 & 0 \\ \in & \sigma^2 & \in & 0 & 0 \\ \in & \in & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \in \\ 0 & 0 & 0 & \in & \sigma^2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} \sigma^2 & \in & \in & 0 & 0 \\ \in & \sigma^2 & \in & 0 & 0 \\ \in & \in & \sigma^2 & \in & \in \\ 0 & 0 & \in & \sigma^2 & \in \\ 0 & 0 & \in & \in & \sigma^2 \end{bmatrix}$$

When disjoint variable clusters are generated for experimental data distributed according to Σ_1 with $k = 2$, the cluster assignments will be $\{1, 1, 1, 2, 2\}$ with high likelihood, i.e. two variable clusters corresponding to the block structure in the population covariance matrix. For the gene set corresponding to the first row in **A**, the estimated entropy given by (4) is $H(C_1) = -\frac{3}{3}log(\frac{3}{3}) - \frac{0}{3}log(\frac{0}{3}) = 0$. Because the estimated entropy is already the minimum possible value, the EMVC algorithm will not make any changes to the first row of **A**. For the gene set corresponding to the second row in **A**, the estimated entropy given by (4) is $H(C_2) = -\frac{1}{3}log(\frac{1}{3}) - \frac{2}{3}log(\frac{2}{3}) = .637$. To achieve a minimum entropy of 0 for this gene set with the fewest annotation changes, the EMVC algorithm eliminates all annotations except those belonging to cluster 2, the gene cluster with the most genes annotated to this gene set. Overall, EMVC optimization of **A** for Σ_1 will result in the following optimized annotation matrix:

$$\mathbf{O}_1 = \mathbf{T}_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

When disjoint variable clusters are generated for experimental data distributed according to Σ_2 with $k = 2$, the cluster assignments will alternate between $\{1, 1, 1, 2, 2\}$ and $\{1, 1, 2, 2, 2\}$ with roughly equal likelihood. Because the EVMC algorithm averages optimization results across multiple bootstrap resampled datasets, the optimized matrix **O** will reflect the average of the optimization for these two cluster assignment scenarios:

$$\mathbf{O}_2 = \begin{bmatrix} 1 & 1 & .5 & 0 & 0 \\ 0 & 0 & .5 & 1 & 1 \end{bmatrix}$$

When the $\mathbf{O_2}$ matrix is filtered to generate the binary optimized annotation matrix $\mathbf{T_2}$, either of the following can be generated depending on whether the threshold α is set low or high, respectively:

$$\mathbf{T}_{2,\alpha<.5} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \mathbf{T}_{2,\alpha>.5} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

2.3 EMVC evaluation

To evaluate the effectiveness of our approach, we used the EMVC algorithm to optimize both simulated variable groups for simulated data as well as MSigDB gene set collections for real gene expression data. Variable clusters were generated using both k-means clustering and average-link agglomerative hierarchical clustering with correlation-based distance. Evaluation was based on the following metrics:

(1) Ability of the EMVC algorithm to filter inconsistent gene set annotations and leave valid annotations unchanged. Assuming the validity of each annotation is known, this can be quantified using contingency table statistics for the output matrix **T** and can be represented using a receiver operating characteristic (ROC) curve for the output matrix **O**.

(2) Improvement in gene set enrichment power when using the optimized annotations in **T** versus the unoptimized annotations in **A**. This can be quantified if the identity of gene sets that have a true association with the output for a given dataset is known.

(3) Improved replication of gene set enrichment results across similar datasets when using the annotations in **T** versus **A**. Although knowledge of the true enrichment status of each gene set is not needed to measure replication, multiple datasets are required.

2.3.1 Evaluation using simulated variable groups and simulated data As a straightforward example, the EMVC algorithm was used to optimize 20 disjoint variable groups, each composed of annotations to 15 variables, against twenty-five 100×300 data matrices simulated according to a multivariate normal distribution $\sim \mathrm{MVN}(\mu, \Sigma)$. The population covariance matrix, Σ, was structured such that all variables had a variance of $\sigma^2 = 1$ and a correlation among the first 5 variables within the first 10 variable groups of $\rho = 0.75$. For the first 50 observations, i.e. the *cases*, the mean vector, μ, was set to 0 for all variables except for the first 5 variables within variable groups 1, 2, 11 and 12 (the *enriched* variable groups) for which it was set to 1. For the last 50 observations, i.e. the *controls*, the mean vector was set to zero, $\mu = \mathbf{0}$. According to this design, only the first 5 variables within each of the first 10 variable groups represent valid annotations.

EMVC optimization of the simulated variable groups was performed without bootstrapping and using 50 bootstrap resampled datasets. Variable clusters were created by cutting the dendrogram generated via average-link agglomerative hierarchical clustering with correlation-based distance, $(1 - r)/2$, at $k = 10$ and at k ranging from 5 to 15. Variable group enrichment false discovery rates (FDR) were computed using the Benjamini and Hochberg algorithm (Benjamini and Hochberg, 1995) from two-sided enrichment *P*-values generated by the Correlation Adjusted MEan RAnk (CAMERA) competitive enrichment method (Wu and Smyth, 2012) using the R implementation in the *limma* package (Smyth, 2005) with default settings. Improvement in enrichment replication was quantified using Kendall's coefficient of concordance (Kendall and Smith, 1939), as implemented in the R package *irr*, across the 25 simulated datasets.

EMVC optimization results for additional simulation scenarios involving larger sets of overlapping variable groups and the use of k-means clustering instead of average-link agglomerative hierarchical clustering with correlation distance can be found in Supplementary File S1.

2.3.2 Evaluation using MSigDB C2 v1.0 gene sets and p53 gene expression data The EMVC algorithm was used to optimize the MSigDB C2 v1.0 gene sets for the p53 gene expression data used in the 2005 GSEA paper (Subramanian *et al*., 2005). This classic gene set collection and gene expression dataset were selected principally because of their widespread use in the gene set enrichment literature [e.g. (Efron and Tibshirani, 2007) and (Subramanian *et al*., 2005)] and easy accessibility from the MSigDB repository, factors that will enable other researchers to more easily interpret and replicate the reported EMVC optimization results. As a curated gene set collection with experimentally based annotations, the C2 collection also provides a more meaningful annotation optimization challenge than much larger collections such as GO whose annotations are primarily generated via automated methods and are therefore less likely on average to align with experimental data.

EMVC optimization was performed using the archived MSigDB C2 v1.0 gene sets and collapsed p53 gene expression data downloaded from the MSigDB repository. With a minimum gene set size of 15 and maximum gene set size of 200, 301 gene sets out of the original 522 were used in the analysis. The optimized annotation matrix **O** was generated by executing the EMVC algorithm on 50 bootstrap resampled datasets drawn from the standardized p53 gene expression data, i.e. each column was mean centered and scaled to have a standard deviation of 1, with gene clusters generated by k-means clustering for k in the range of 3–15. An optimized version of the C2 gene sets, representing matrix **T**, was generated by filtering the optimized annotation matrix **O** at a threshold of 0.1. The enrichment of both optimized and unoptimized C2 gene sets was computed for the p53 mutated versus wild-type phenotype using CAMERA (Wu and Smyth, 2012) with default parameters.

Unlike in the simulated data case, where the validity of each annotation was known by design, the consistency of C2 gene set annotations for the p53 data could only be inferred indirectly. For evaluation of the EMVC algorithm via contingency table statistics, the designation of each gene set member by the GSEA algorithm (Subramanian *et al*., 2005) as either a *core* gene or *non-core* gene with respect to enrichment against the p53 mutated phenotype was used as a proxy for annotation validity (e.g. see the detailed results at http://www.broadinstitute.org/gsea/resources/gsea_pnas_results/p53_C2.Gsea/index.html). Although it was not possible to directly quantify the change in gene set enrichment power due to EMVC optimization of the C2 gene sets, the impact was indirectly examined by comparing the change in enrichment FDR values between unoptimized and optimized annotations and the unoptimized enrichment significance. Enrichment replication was analyzed using Kendall's coefficient of concordance on the enrichment results computed using optimized annotations over multiple bootstrap resampled datasets, where these bootstrap datasets used to compute concordance were distinct from the bootstrap datasets used during annotation optimization.

2.3.3 Evaluation using MSigDB C4 v4.0 cancer modules and leukemia gene expression data The EMVC algorithm was also used to optimize the MSigDB C4 v4.0 cancer modules for the leukemia gene expression data (Armstrong *et al*., 2002) used in the 2005 GSEA paper (Subramanian *et al*., 2005). Because the cancer modules (Segal *et al*., 2004) were generated by merging and then refining both existing gene sets drawn from GO, KEGG and the Gene Microarray Pathway Profiler (GenMAPP) (Dahlquist *et al*., 2002) and gene clusters computed from 1975 gene expression microarrays for 22 tumor types, the cancer modules should be well aligned with the structure of tumor gene expression data, making further optimization challenging for a dataset such as the leukemia gene expression data. The automated data-driven process used to create the cancer modules also provides a useful contrast with the curated C2 gene sets for the purpose of evaluating the EMVC algorithm.

Similar to testing on the C2 gene sets and p53 data, optimization was performed using the MSigDB C4 v4.0 cancer modules and collapsed leukemia gene expression data downloaded from the MSigDB repository. With a minimum gene set size of 15 and maximum gene set size of 200, 297 gene sets of the original 431 were used in the analysis. The optimized annotation matrix **O** was generated by executing the EMVC algorithm on 50 bootstrap resampled datasets drawn from the standardized leukemia gene expression data with gene clusters generated by cutting the dendrogran generated via average-link agglomerative hierarchical clustering with correlation distance at k in the range of 3–15. An optimized version of the cancer modules, representing matrix **T**, was generated by filtering the optimized annotation matrix **O** at a threshold of 0.15. The enrichment of both optimized and unoptimized cancer modules was computed for the acute myeloid leukemia (AML) versus acute lymphoblastic leukemia (ALL) phenotypes using CAMERA (Wu and Smyth, 2012).

The computation of contingency table statistics, analysis of enrichment power and quantification of enrichment replication were performed for the cancer modules and leukemia data using the same methods employed for the C2 gene sets and p53 data (see Section 2.3.2 above).

3 RESULTS

3.1 Optimization of simulated variable groups using simulated data

Removal of inconsistent annotations. Optimization results for one of the 25 datasets simulated according the procedure outlined in

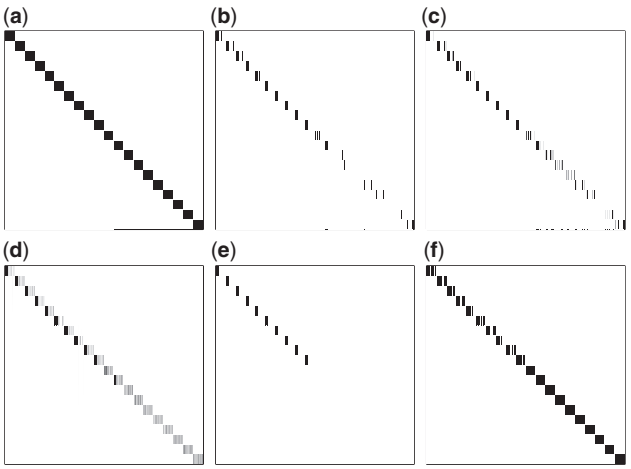

Fig. 1. EMVC optimization results on simulated data. (**a**) Graphical representation of the annotation matrix capturing the non-overlapping assignment of 300 random variables to 20 variable groups. Each row represents a variable group, each column represents a random variable and positive annotation values are indicated by dark cells. (**b**) Annotation matrix after a single execution of the EMVC algorithm on clusters of the 300 variables generated by a cut of the dendrogram generated by single-link agglomerative hierarchical clustering with correlation distance at $k = 10$. Dark cells reflect annotations that were not filtered during optimization. (**c**) Annotation matrix after execution of the EMVC algorithm on clusters of the 300 variables generated by dendrogram cuts at k in the range 5–15. Intensity of the cell shading corresponds to the proportion of the clusterings in which the annotation was kept after optimization. (**d**) Annotation matrix based on the average of 50 executions of the EMVC algorithm on bootstrap resampled datasets. Intensity of cell shading corresponds to the average optimization proportion over all bootstrap resampled datasets. (**e**) Annotation matrix based on sparse filtering of bootstrap results. Only annotations whose average bootstrap optimization proportion is >0.9 are included. (**f**) Annotation matrix based on strict filtering of bootstrap results. Only annotations whose average bootstrap optimization proportion is <0.1 are removed

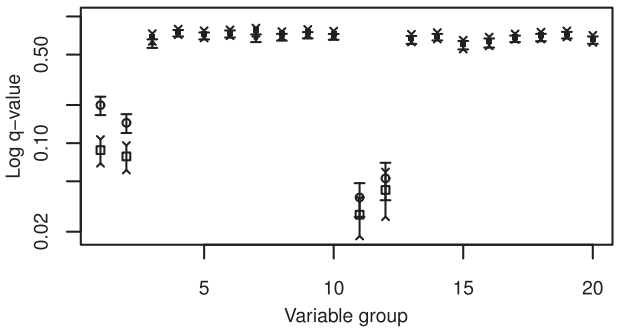

Fig. 2. Distribution of enrichment FDR for simulated variable groups using both unoptimized and optimized annotations. Plotted FDR values were computed using the Benjamini and Hochberg algorithm from two-sided enrichment *P*-values generated by the competitive enrichment method CAMERA (Wu and Smyth, 2012) for each of the 20 variable groups across 25 datasets simulated according to the design outlined in Section 2.3.1. Filled circles and flat error bars represent the average (± one standard error) of the FDR values computed for each of the 20 variable groups using unoptimized annotations on the 25 simulated datasets. Squares and angled error bars represent the FDR values computed using bootstrap optimized annotations with strict filtering. For the four *enriched* variable groups simulated with a true mean difference between cases and controls, open circles and open squares are used. FDR values are plotted on a logarithmic scale

Section 2.3.1 is shown in Figure 1b–f. Figure 1b and c show the EMVC output when results are not averaged over multiple bootstrap resampled datasets. Figure 1b is additionally restricted to just a single number of clusters, in this case 5. Figure 1d illustrates the standard output matrix **O**, which averages results over cluster sizes from 5 to 15 and 50 bootstrap resampled datasets. Figure 1e and f show two versions of the filtered output matrix **T** for thresholds of 0.1 and 0.9, respectively.

For the simulation procedure outlined in Section 2.3.1, the EMVC algorithm filtered inconsistent annotations with high accuracy when applied to a range of cluster sizes and multiple bootstrap resampled datasets. The mean area under curve (AUC) over all 25 simulated datasets was 0.995. When just a single cluster size was used or bootstrapping was not used, EMVC performance declined. The mean AUC for no bootstrapping and $k = 10$ was 0.912, for all cluster sizes and no bootstrapping the mean AUC was 0.941, and for 50 bootstrap datasets and $k = 5$ the mean AUC was 0.993.

Impact on enrichment power. The impact of EMVC optimization on variable group enrichment for all 25 simulated datasets is shown in Figure 2. This figure plots the distribution of variable group enrichment FDR computed using CAMERA (Wu and Smyth, 2012) for each of the 20 variable groups using both unoptimized and optimized annotations. Based on the simulation design, only variable groups 1, 2, 11 and 12 should have significant FDR values because only these variable groups include variables that have a true association with the simulated binary phenotype. Although the EMVC algorithm filters many uncorrelated variables from the first 10 variable groups, enrichment using both unoptimized and optimized annotations results in insignificant FDR values for all truly non-enriched variable groups. The enrichment FDR values for unenriched variable groups are therefore not impacted by EMVC filtering of uncorrelated variables. As shown by the figure, enrichment power for this example is substantially improved after EMVC-based annotation optimization with the mean power to detect the truly enriched variable groups at a q-value of ≤ 1, changing from 0.63 for unoptimized annotations to 0.79 for optimized annotations.

Impact on enrichment replication. EMVC-optimized annotations also improved the replication of enrichment results, as measured by Kendall's coefficient of concordance across the 25 independently simpulated datasets. Using unoptimized annotations, Kendall's W for the enrichment FDR values across the 25 simulated datasets was 0.486. Using optimized annotations, Kendall's W was 0.507.

3.2 Optimization of MSigDB C2 v1.0 using p53 data

Removal of inconsistent annotations. Figure 3 shows the impact of EMVC optimization on the 15 MSigDB C2 v1.0 gene sets with the lowest enrichment *P*-values relative to the p53 mutated versus wild-type phenotype using unoptimized annotations. The contingency table embedded in the lower right corner of this figure holds the results of the overlap between EMVC-filtered genes

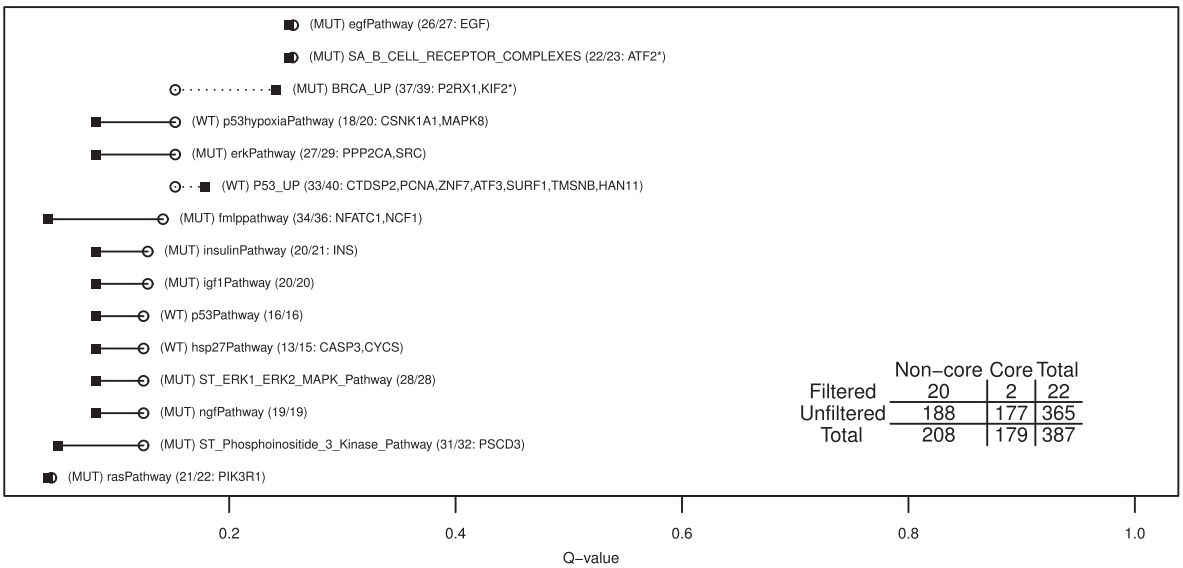

Fig. 3. Enrichment and annotation optimization results for the MSigDB C2 v1.0 gene sets and p53 data used in the 2005 GSEA paper (Subramanian *et al*., 2005). The figure shows the difference between enrichment FDR computed using unoptimized and optimized annotations for the 15 C2 gene sets with the lowest unoptimized enrichment *P*-values. Enrichment FDRs were computed using the Benjamini and Hochberg algorithm on two-sided *P*-values generated by the enrichment method CAMERA (Wu and Smyth, 2012). Open circles represent the FDR values computed using the unoptimized gene sets annotations, and solid squares represent the FDR values computed using optimized annotations. If the optimized FDR value is less than the unoptimized value, a solid line is used, otherwise, a dotted line is used. A WT prefix is used for gene sets enriched using the unoptimized annotations for the wild-type phenotype, and a MUT prefix is used for gene sets enriched for the mutated phenotype. The ratio of optimized to unoptimized annotations for each of the top 15 gene sets is displayed after each gene set name along with the symbols for the filtered genes. An asterisk follows the symbol for filtered annotations that were designated as *core* genes by the GSEA algorithm. The contingency table in the bottom right corner displays the association between EMVC annotation filtering and whether each annotation was designated as a *core* or *non-core* gene by the GSEA algorithm with respect to enrichment against the WT versus MUT phenotype. For the displayed contingency table, filtered annotations were removed by EMVC in more 90% of the cluster results in 50 bootstrap resampled datasets resulting in an odds ratio of 9.38 (95% CI: 2.23–84). When all filtering thresholds are considered, the area under the ROC curve is 0.67

and genes that were designated as *core* or *non-core* by GSEA for these 15 gene sets. In terms of the desired behavior of the EMVC algorithm, *non-core* genes can be viewed as true positives, i.e. annotations that should be removed. As demonstrated by the significant odds ratio of 9.38 (95% CI: 2.23–84) and area under the ROC curve of 0.67 for all annotation filtering thresholds, the EMVC algorithm effectively removed C2 annotations for genes that do not contribute to the mutated versus wild-type phenotype in the p53 data.

Impact on enrichment power. As illustrated in Figure 3, EMVC optimization resulted in an improvement in enrichment FDR values for 13 of the 15 most significant gene sets. As demonstrated by the association between EMVC annotation filtering and the GSEA *core* versus *non-core* designation, this improvement in enrichment FDR values was primarily due to the preferential removal of annotations for genes with either a small association with the outcome or with an association that was the opposite from the overall direction of enrichment of the gene set. Across all 301 tested C2 gene sets, the improvement in the enrichment FDR after EMVC optimization was positively correlated with the original enrichment significance of the gene set, i.e. gene sets with significant enrichment values using the unoptimized annotations were most likely to benefit from optimization. This association was demonstrated by a Spearman correlation between unoptimized enrichment FDR values and the ratio of optimized to unoptimized enrichment FDR values of 0.261 (*P*-value: 4.39e-06). The Spearman correlation between the unoptimized enrichment FDR and the proportion of filtered gene set annotations was −0.0309 (*P*-value: 0.594). The fact that the proportion of gene set annotations filtered during optimization was unassociated with gene set enrichment significance demonstrates that this positive correlation was not the result of preferential annotation filtering for significantly enriched gene sets.

Impact on enrichment replication. EMVC optimization also had a positive impact on gene set enrichment replication, as measured by Kendall's coefficient of concordance on the enrichment *P*-values across multiple bootstrap resampled datasets. Using the unoptimized annotations, Kendall's W for the enrichment *P*-values values of the C2 gene sets relative to the p53 mutated and wild-type phenotypes on 20 bootstrap resampled p53 datasets was 0.372. Using the optimized annotations, Kendall's W was 0.384.

Detailed results. Complete output from both EMVC annotation optimization and CAMERA gene set enrichment can be found in Supplementary File S2.

3.3 Optimization of MSigDB C4 v4.0 cancer modules using leukemia data

Removal of inconsistent annotations. The ability of the EMVC algorithm to successfully remove inconsistent cancer module annotations was verified by examining the overlap between EMVC filtered genes and genes that are designated by GSEA as *core* or

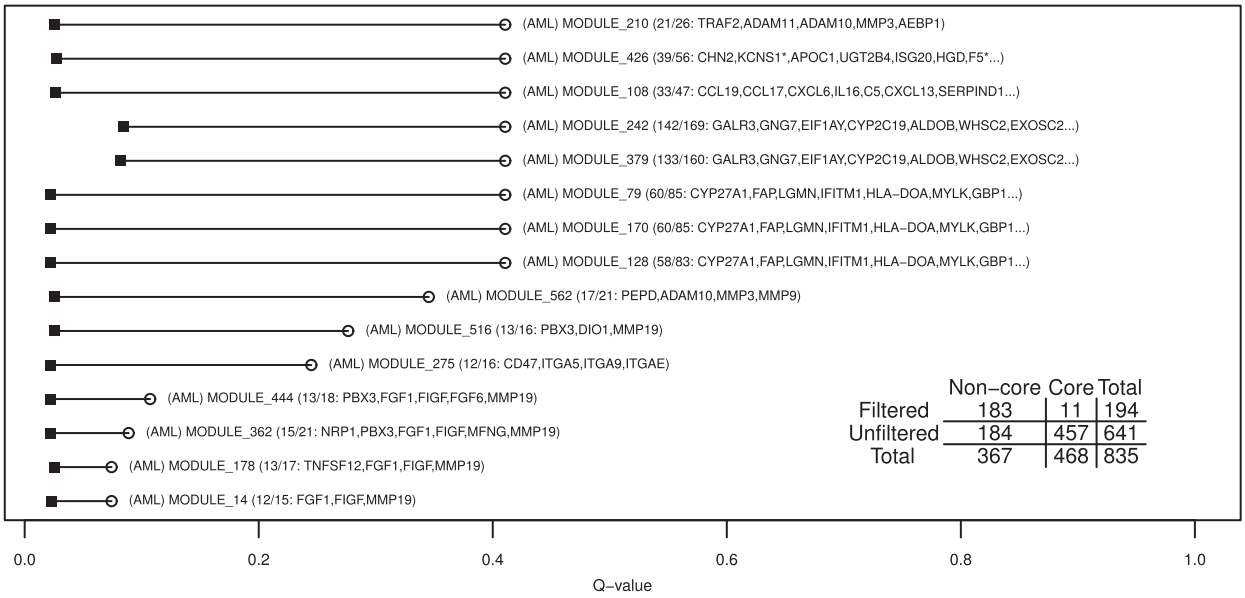

Fig. 4. Enrichment and annotation optimization results for the MSigDB C4 v4.0 cancer modules and leukemia gene expression data (Armstrong *et al.*, 2002). The figure shows the difference between enrichment FDR computed using unoptimized and optimized annotations for the 15 cancer modules with the lowest unoptimized enrichment *P*-values. Enrichment FDRs were computed using the Benjamini and Hochberg algorithm on two-sided *P*-values generated by CAMERA (Wu and Smyth, 2012). Open circles represent the FDR values computed using the unoptimized cancer module annotations, and solid squares represent the FDR values computed using optimized cancer modules. If the optimized FDR value is less than the unoptimized value, a solid line is used, otherwise, a dotted line is used. An AML prefix is used for gene sets enriched using the unoptimized annotations for the acute myeloid leukemia, and an ALL prefix is used for gene sets enriched for the acute lymphoblastic leukemia phenotype. The ratio of optimized to unoptimized annotations for each of the top 15 cancer modules is displayed after each module name along with the symbols for the filtered genes (cropped at 7). An asterisk follows the symbol for filtered annotations that were designated as *core* genes by the GSEA algorithm. For the displayed contingency table, filtered annotations were removed by EMVC in more 85% of the cluster results in 50 bootstrap resampled data sets resulting in an odds ratio of 41.1 (95% CI: 21.8–86). When all filtering thresholds are considered, the area under the ROC curve is 0.92

non-core with respect to enrichment against the AML versus ALL phenotype. Similar to Figure 3, Figure 4 shows the impact of EMVC optimization on the 15 cancer modules with the lowest enrichment *P*-values. The contingency table embedded in the lower right corner of this figure holds the results of the overlap between EMVC filtered genes and GSEA *core* or *non-core* genes for the 15 most enriched cancer modules. As demonstrated by the significant odds ratio of 41.1 (95% CI: 21.8–86) and area under the ROC curve of 0.92 for all annotation filtering thresholds, the EMVC algorithm effectively removed C4 cancer module annotations for genes that do not contribute to the AML versus ALL phenotype.

Impact on enrichment power. As illustrated in Figure 4, EMVC optimization resulted in an improvement in enrichment FDR values for all 15 most significantly enriched cancer modules. The Spearman correlation between the unoptimized enrichment FDR and the ratio of optimized enrichment FDR to unoptimized enrichment FDR was 0.755 (*P*-value: 4.78e-56), while the Spearman correlation between the unoptimized enrichment FDR and the proportion of filtered annotations was 0.277 (*P*-value: 1.2e-06).

Impact on enrichment replication. Using the unoptimized annotations, Kendall's W for the enrichment *P*-values of the cancer modules relative to the AML versus ALL phenotypes on 20 bootstrap resampled leukemia datasets was 0.889. Using the optimized annotations, Kendall's W was 0.934.

Detailed results. Complete output from both EMVC annotation optimization and CAMERA gene set enrichment can be found in Supplementary File S3.

4 DISCUSSION

Gene clusters and gene set enrichment. The EMVC algorithm performs annotation optimization on variable clusters computed using an unsupervised view of experimental data. By minimizing the entropy for each variable group relative to disjoint variable clusters, the annotations for variables that tend to cluster with other variable group members are kept and annotations for variables that cluster apart are filtered. A key advantage of this unsupervised approach is that EMVC-optimized annotations can be used for subsequent variable group enrichment without biasing the computed enrichment statistics. However, the unsupervised EMVC approach can only successfully filter inconsistent annotations, improve gene set enrichment power and improve enrichment replication if the structure of genomic data, as represented by gene clusters, can be used to identify the genomic variables most likely to contribute to gene set enrichment. In other words, the genes that contribute strongly to the enrichment signal for significantly enriched gene sets must be more likely to cluster together than the genes whose expression is not consistent with gene set enrichment.

In the simulation example outlined in Section 3.1, such a relationship between variable group enrichment and inter-variable correlation was explicitly created for the first five variables in the first two variable groups with the predictable result that these variables were not filtered by EMVC and significantly lower enrichment FDR values were obtained using optimized annotations. The results in Sections 3.2 and 3.3 provide important confirmation that this association between gene set enrichment and gene clustering exists in real microarray gene expression data in the context of curated and automatic MSigDB gene sets. This is most clearly demonstrated by the strong relationship between EMVC annotation filtering and the designation of gene set annotations by the GSEA enrichment algorithm as either *core* or *non-core* genes.

Optimal number of gene clusters and use of bootstrap aggregation. Because the true number of clusters for experimental datasets is unknown and cluster size estimation methods such as the gap statistic (Tibshirani *et al*., 2001) or silhouette width (Kaufman and Rousseeuw, 2005) are often unreliable, the EMVC algorithm is executed on multiple variable clusterings where the number of clusters varies over a specified range. One potential enhancement of the EMVC method would be use of results from a method such as the gap statistic to weight the EMVC optimization results for each clustering in the specified range. Bootstrap aggregation is further used to reduce the variance of the annotation optimization estimates (Breiman, 1996; Hastie *et al*., 2009). Averaging over multiple clusterings for multiple bootstrap resampled datasets provides a robust optimization result that is not dependent on a specific estimate of the optimal number of variable clusters. As demonstrated by the simulation example in Section 3.1, this can have a significant impact on optimization performance. The importance of computing information-theoretic measures over a range of cluster sizes has also been highlighted in the paper describing the recently developed maximal information coefficient method (Reshef *et al*., 2011).

Using EMVC to analyze specific genomic datasets. One of the primary applications of the EMVC algorithm involves the optimization of a gene set collection for a specific genomic dataset before enrichment analysis. For this application, it is desirable to perform enrichment analysis against existing gene set categories that have been modified to only contain annotations consistent with the narrow domain under investigation. By using standard gene sets and only allowing the removal of annotations, the computed enrichment results can be directly interpreted in terms of widely known and well understood genomic functions. Such direct and easy interpretation is not possible if annotations are added or if new novel gene sets are derived. The fact that the EMVC algorithm uses an unsupervised view of the data to just filter annotations from existing gene sets therefore makes it well suited for this use case. Additional benefits of the EMVC algorithm in this scenario include the ability to use optimization proportions, rather than filtered annotations based on a threshold, directly with enrichment methods that support annotation weights [e.g. ProbCD (Vêncio and Shmulevich, 2007)], and flexibility regarding the algorithm used to cluster genes.

Using EMVC to refine gene set collections. The EMVC method can also be used for the general refinement of gene set collections, either to create versions of a gene set collection that are customized for a specific domain or to identify and entirely remove annotations that exhibit poor alignment with a broad selection of genomic datasets. For both variants of this use case, the EMVC algorithm would be used to optimize a gene set collection for a large number of individual datasets. For the first variant, the average optimization proportions generated across all target datasets could be used by researchers to create customized versions of the gene set collection at any desired level of confidence. For this application, the ease with which the EMVC algorithm can be parallelized, at the level of different clusterings or different bootstrap resampled datasets, is a major benefit.

EMVC Limitations. Limitations of the EMVC algorithm include the restriction to annotation removal, computational complexity, dependence on gene clustering structure and sensitivity to algorithm parameter settings.

- The EMVC algorithm will only remove potentially inconsistent annotations to a gene set. It will not augment incomplete gene sets or identify new gene sets.
- If gene set members associated with the clinical outcome fail to cluster together, EMVC annotation optimization will not improve gene set enrichment.
- EMVC performance is sensitive to several algorithm parameters. Specifically, the cluster method, k range and filtering threshold must be appropriate for the structure of the experimental data in $\mathbf{X}$ and annotations in $\mathbf{A}$.
- EMVC can be computationally expensive. This is especially true for large genomic datasets and correspondingly large gene sets collections with the k range and number of bootstrap resamples needed to generate stable optimization results.

5 CONCLUSION

Gene set enrichment has become a central element in the analysis and interpretation of genomic data. Although significant progress has been made building gene set collections and developing statistical enrichment methods, annotation quality remains a critical challenge. Because of the broad scope of many gene set collections and the large number of low-quality annotations, enrichment analysis results are frequently inaccurate and non-reproducible. Current approaches to annotation quality are mainly general purpose, largely driven by just the structure and content of the gene set ontology and, when experimental data are considered, focus on gene set synthesis over refinement. To address the annotation quality issue and limitations of current approaches, we have developed a novel annotation optimization method, EMVC, which is available as an R package from CRAN. On both simulated gene sets with simulated data and MSigDB gene sets with real gene expression data, the EMVC algorithm has been shown to effectively filter inconsistent annotations, improve enrichment power and improve enrichment replication.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their insightful and constructive feedback on the original manuscript.

Funding: National Institutes of Health R01 grants (LM010098, LM011360, EY022300, GM103506 and GM103534).

Conflict of Interest: none declared.

REFERENCES

Alterovitz,G. *et al.* (2010) Ontology engineering. *Nat. Biotechnol.*, **28**, 128–130.

Armstrong,S.A. *et al.* (2002) Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bell,M.J. *et al.* (2012) An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB. *Bioinformatics*, **28**, i562–i568.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.*, **57**, 289–300.

Binns,D. *et al.* (2009) Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, **25**, 3045–3046.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Carbon,S. *et al.* (2009) Amigo: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

Chavent,M. *et al.* (2012) ClustOfVar: an R package for the clustering of variables. *J. Stat. Softw.*, **50**, 1–16.

Dahlquist,K.D. *et al.* (2002) Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, **31**, 19–20.

Davis,M.J. *et al.* (2010) Automatic, context-specific generation of gene ontology slims. *BMC Bioinformatics*, **11**, 498.

Dolan,M.E. *et al.* (2005) A procedure for assessing go annotation consistency. *Bioinformatics*, **21** (**Suppl. 1**), i136–i143.

du Plessis,L. *et al.* (2011) The what, where, how and why of gene ontology–a primer for bioinformaticians. *Brief. Bioinform.*, **12**, 723–735.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Faria,D. *et al.* (2012) Mining go annotations for improving annotation consistency. *PLoS One*, **7**, e40519.

Hastie,T. *et al.* (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, [doi:10.1186/gb-2000-1-2-research0003].

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.* 2nd edn. Springer, New York, NY.

Hausser,J. and Strimmer,K. (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, **10**, 1469–1484.

Hung,J.-H. *et al.* (2012) Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.*, **13**, 281–291.

Juncker,A.S. *et al.* (2009) Sequence-based feature prediction and annotation of proteins. *Genome Biol.*, **10**, 206.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kaufman,L. and Rousseeuw,P.J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, Hoboken, NJ.

Kendall,M.G. and Smith,B.B. (1939) The problem of m rankings. *Ann. Math. Stat.*, **10**, 275–287.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Liberzon,A. *et al.* (2011) Molecular Signatures Database (MSigDb) 3.0. *Bioinformatics*, **27**, 1739–1740.

Park,Y.R. *et al.* (2011) Gochase-ii: correcting semantic inconsistencies from gene ontology-based annotations for gene products. *BMC Bioinformatics*, **12** (**Suppl. 1**), S40.

Reshef,D.N. *et al.* (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–1524.

Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.

Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Skunca,N. *et al.* (2012) Quality of computationally inferred gene ontology annotations. *PLoS Comput. Biol.*, **8**, e1002533.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 397–420.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tibshirani,R. *et al.* (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B Methodol.*, **63** (Pt 2), 411–423.

Vêncio,R.Z.N. and Shmulevich,I. (2007) ProbCD: enrichment analysis accounting for categorization uncertainty. *BMC Bioinformatics*, **8**, 383.

Wu,D. and Smyth,G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, **40**, e133.