The limitations of simple gene set enrichment analysis assuming gene independence

Pablo Tamayo,1 George Steinhardt,2 Arthur Liberzon1 and Jill P Mesirov1,2

Abstract
Since its first publication in 2003, the Gene Set Enrichment Analysis method, based on the Kolmogorov-Smirnov statistic, has been heavily used, modified, and also questioned. Recently a simplified approach using a one-sample t-test score to assess enrichment and ignoring gene-gene correlations was proposed by Irizarry et al. 2009 as a serious contender. The argument criticizes Gene Set Enrichment Analysis’s nonparametric nature and its use of an empirical null distribution as unnecessary and hard to compute. We refute these claims by careful consideration of the assumptions of the simplified method and its results, including a comparison with Gene Set Enrichment Analysis’s on a large benchmark set of 50 datasets. Our results provide strong empirical evidence that gene–gene correlations cannot be ignored due to the significant variance inflation they produced on the enrichment scores and should be taken into account when estimating gene set enrichment significance. In addition, we discuss the challenges that the complex correlation structure and multi-modality of gene sets pose more generally for gene set enrichment methods.

Keywords
Gene set enrichment analysis, gene expression

1 Introduction
The availability of global transcription profiling with microarrays in the mid 1990s made possible the analysis and interpretation of the activity of the entire transcriptome to provide insights into biological function and the mechanisms of disease. Early on, it became clear that focusing on long lists of differentially expressed genes gave limited understanding of underlying pathways and processes. Several approaches to consider testing for the over-representation of gene categories rather than genes were introduced.1–6 In Reference [7], we introduced a knowledge-based approach analysis method, Gene Set Enrichment Analysis (GSEA) to address this problem. Briefly, this initial approach ranked genes by their differential expression between two
phenotypes, used the Kolmogorov-Smirnov statistic to score the enrichment of an *a priori* defined sets of genes that share common biological function, chromosomal location or regulation, and evaluated the significance of the score using an empirical permutation test correcting for multiple hypothesis testing. Thus, GSEA provided a prioritized list of sets of genes for hypothesis generation and further study. In this first application, we identified a set of genes involved in oxidative phosphorylation with reduced expression in diabetic patients. None of these genes were down-regulated by more than 20%, but as a group their coordinate down-regulation was significant and, together with subsequent work, lead to a better understanding of the regulation of the oxidative phosphorylation pathway as many of its components turned out to be controlled by the PCG1A transcription factor, which was itself down-regulated in diabetic patients.

As soon as this original version of GSEA appeared, objections were raised to the approach, some of which were immediately refuted in Reference [11], and the rest were met by our subsequent improvement of the GSEA methodology. In Subramanian et al., we introduced a version of GSEA that used a correlation-weighted Kolmogorov-Smirnov statistic, an improved enrichment normalization procedure, and an False Discovery Rate (FDR)-based estimate of significance that collectively made GSEA appreciably more sensitive, more general, and more robust. As a result of these improvements, and the public availability of the software and companion Molecular Signatures Database (MSigDB) [www.broadinstitute.org/gsea], GSEA became a widely used method and was applied to numerous problems across many application domains. Notably, since the original release of the software and database in 2005, the number of GSEA user registrations has grown to over 33,000 and the method used and cited in more than 3100 scientific publications. GSEA and other gene set analysis methods have also motivated the development of general statistical methodologies for large-scale inference for “sets” of variables.

The specific knowledge-based approach pioneered by GSEA is now standard practice in the analysis of gene expression data and inspired the development of a large and growing family of conceptually similar methods. For example, Huang et al. identified at least 68 different gene set enrichment methods in their survey. A family of popular methods estimate the over-representation of Gene Ontology (GO) annotations using a hyper-geometric statistic or Fisher’s exact test (e.g., GoMiner, FatiGO, GoSurfer, EasyGo, David). These methods restrict consideration to the “top” of the list and may miss more subtle signals. They also assume gene independence and thus produce overly optimistic results. In addition, several improvements to gene set enrichment analysis itself have been proposed. These include those used in Reference [26], GSA, SAFE, Catmap, ErmineJ, SAM-GS, and PROPA. They employ alternative ranking metrics, enrichment statistics, and several variations on significance estimation schemes. Notably, Nilsson et al. demonstrates the difficulty of finding a single, optimal statistic due to the complexity, heterogeneity and multi-modal distribution of the expression levels of genes within gene sets. Other somewhat more sophisticated methods (e.g., FunNet, PARADIGM, COFECO) take a network-based approach but restrict the analysis to processes where a deeper understanding of gene–gene interactions is already available. The primary advantages of GSEA are that it only requires gene set membership information to compute enrichment scores, considers the entire ranked list of genes, and maintains the gene–gene dependency that reflects real biology. This yields a good compromise between sensitivity, performance, and applicability.

Recently, Irizarry et al. in their “Gene Set Enrichment Made Simple” article proposed a “simpler” approach to gene set expression analysis assuming gene independence and using a one-sample *t*-test to estimate enrichment. Here, we will refer to their method as Simpler Enrichment Analysis (SEA). The rationale for SEA is based on their perception that gene independence is a reasonable simplifying assumption and thus simpler parametric approaches to gene set analysis have
been ignored. Both of these assumptions disregard a large body of literature where many authors have already introduced “simple” parametric methods for gene set analysis.\(^{38-42}\) Many researchers have demonstrated the unrealistic nature and limitations of the gene independence assumption.\(^{21-25,43}\) In addition, they criticized GSEA for its use of an empirical null distribution, which they argue is unnecessary and hard to compute, and for using a non-parametric weighted statistic that they believe inherits the lack of sensitivity of the original Kolmogorov-Smirnov statistic. The paper concludes by proposing SEA as a serious contender and argues against the use of GSEA in any of its forms.

The aim of this paper is twofold. First, we present the SEA approach and contrast it with GSEA using the statistical framework of Barry et al.\(^ {43}\) We then carefully consider Irizarry et al.’s\(^ {37}\) criticism of GSEA and refute their claims by systematically comparing SEA and GSEA on a large benchmark set of 50 datasets. We show, in agreement with earlier observations, that the gene independence assumption is not realistic because gene correlations are non-trivial and produce a substantial amount of variance inflation in the global statistic that in turn produces a large number of false positive results. Second, we discuss the challenges that the complex correlation structure and multimodality of gene sets pose for gene set enrichment methods in general and propose that future progress in gene set analysis will result from improving the resolution of the gene sets and better ways to model the complex gene set correlation structure.

### 2 Review of SEA and GSEA

Here, we review first the SEA approach proposed in Reference [37] and then the GSEA approach proposed in Reference [12] using the statistical framework of Barry et al.\(^{43}\) Several other valuable and complementary statistical frameworks for gene set analysis have been introduced in recent years.\(^{13,25,44-46}\)

First we will define the quantities relevant for the analysis: the input gene expression dataset \(X\) corresponds to \(N\) genes and \(M\) samples and contains gene expression profiles \(x_{ij}\), where \(i\) indicates a specific gene (row) and \(j\) a specific column (sample). The relevant phenotype of interest is defined as a vector \(Y\) of \(M\) binary values categorizing the sample in two groups \((Y_0, Y_1)\). The gene sets are represented as \(g_k\), where \(k\) runs from 1 to \(K\), the maximum number of gene sets. We define a gene set enrichment method as a two-stage procedure incorporating (a) a local test statistic \(s_i(x_{i*}, Y)\) that measures the association between a gene expression profile \((x_{i*})\) and the phenotype \((Y)\); (ii) a global test statistic \(S_i\) to assess the degree of the gene set’s enrichment, and (iii) a specific null hypothesis and error rate controlling procedure to assess statistical significance and provide a final sorted gene set enrichment result list. The global statistic can be either a parametric or non-parametric function of the local test statistics corresponding to the gene set in question (“self-contained” statistic) and potentially all other genes (the complement or “competitive” statistic).

The SEA method uses as local test statistic a two-sample \(t\)-test statistic \(t_i\) that quantifies the magnitude of differential gene expression for each gene,

\[
s_i(x_{i*}, Y) = t_i = \frac{E(x_{ij}|j \in Y_0) - E(x_{ij}|j \in Y_1)}{\sqrt{\sigma^2(x_{ij}|j \in Y_0) + \sigma^2(x_{ij}|j \in Y_1)}}
\]

where the expected values and standard deviations (\(\sigma\)) are computed in each phenotypic class \((Y_0/Y_1)\). The SEA method can also be used with modified versions of this score\(^ {47}\) or with a
signal to noise ratio as is used in GSEA. SEA’s global test statistic is a one-sample $t$-test (z-score) that is used to estimate the “enrichment” of the entire gene set

$$S^z_k = \sqrt{n_k}E(t_i|i \in g_k) = \sqrt{n_k}\left(\frac{1}{n_k} \sum_{i \in g_k} t_i\right)$$

(2)

The SEA method assumes that these global statistics are independent and identically distributed (gene-independence), proposes a theoretical normal-theory null hypothesis,

$$H^\text{SEA,z}_o : S^1_z, S^2_z, S^3_z, \ldots, S^K_z \text{ are i.i.d and } S^z_k \sim N(0, 1)$$

(3)

and estimates FDR q-values following the methodology of Storey. Realizing that this global statistic only detects changes in location and fails to detect other more complex types of differential gene set behavior, the SEA method adds a second $\chi^2$ global statistic,

$$S^\chi^2_g = \sum_{i \in g_k} (t_i - E(t_i))^2 - (n_k - 1)$$

$$\sqrt{2(n_k - 1)}$$

(4)

and, as in the case of the z-score, an associated normal-theory null hypothesis,

$$H^\text{SEA,} \chi^2_o : S^\chi^2_1, S^\chi^2_2, S^\chi^2_3, \ldots, S^\chi^2_K \text{ are i.i.d and } S^\chi^2_k \sim N(0, 1)$$

(5)

with the FDR computed in the same way as for the z-score. The SEA method is therefore based on the assumption that gene–gene correlations and gene set overlaps have negligible effects and therefore both global statistics (z-score and $\chi^2$) are independent and identically distributed under normal-theory null hypotheses $H^\text{SEA,z}_o$ and $H^\text{SEA,} \chi^2_o$. The SEA null hypotheses are instances of the gene-sampling “Class 1” null hypothesis of Barry et al. We will say more about the applicability of this assumption to real datasets in the next section below.

Before concluding our review of the SEA method, we point out that the second global statistic ironically renders SEA as less “simple,” since the potential user now has to consider two sets of results and their corresponding FDRs. Moreover, no formal procedure is specified by SEA in order to produce a final single list of results. Instead, the authors suggest choosing high scoring gene sets according to either one of the two global statistics.

The GSEA method uses a signal to noise ratio as its local test statistic,

$$s_i(x_{ij}, Y) = \frac{E(x_{ij}|j \in Y_0) - E(x_{ij}|j \in Y_1)}{\sigma'(x_{ij}|j \in Y_0) + \sigma'(x_{ij}|j \in Y_1)}$$

(6)

where $\sigma'$ are the intra-class standard deviations thresholded from below at 20% of the class means,

$$\sigma'(x_{ij}|j \in Y_\ast) = \begin{cases} \sigma(x_{ij}|j \in Y_\ast) \quad & \text{if } \sigma(x_{ij}|j \in Y_\ast) \geq 0.20 \times E(x_{ij}|j \in Y_\ast) \\ 0.20 \times E(x_{ij}|j \in Y_\ast) \quad & \text{if } \sigma(x_{ij}|j \in Y_\ast) < 0.20 \times E(x_{ij}|j \in Y_\ast) \end{cases}$$

(7)
GSEA uses a weighted Kolmogorov-Smirnov global statistic to assess gene set enrichment,

\[
S_{k}^{\text{GSEA}} = \sup_{i=1,\ldots,N} \left( F_{i}^{g_{k}} - F_{i}^{\tilde{g}_{k}} \right);
\]

\[
F_{i}^{g_{k}} = \frac{\sum_{h=1}^{N} |s_{h}|^{\alpha} I_{h}}{\sum_{h=1}^{N} |s_{h}|^{\alpha}};
\]

\[
F_{i}^{\tilde{g}_{k}} = \frac{\sum_{h=1}^{N} (1 - I_{h})}{N - n_{k}};
\]

\[
I_{h} = \begin{cases} 1 & \text{if } h \in g_{k} \\ 0 & \text{if } h \in \tilde{g}_{k} \end{cases}
\]

(8)

An important fact, not often appreciated, is that this global statistic is weighted using a power of the local statistic \(|s_{h}|^{\alpha}\) (typically with \(\alpha = 1\)) and it is therefore much more sensitive to differences at the top and bottom of the gene list than the standard Kolmogorov-Smirnov statistic. Because this statistic cannot be expressed as a simple function of the local statistic, it poses challenges to formal analysis based on parametric modeling.

The global statistics \(S_{k}^{\text{GSEA}}\) depend on size of gene set and therefore are not identically distributed. GSEA addresses this issue by normalizing \(S_{k}^{\text{GSEA}}\) values to factor out the intrinsic gene set size dependence. The relevant normalization is a change of scale using the expected value of the positive (negative) null distribution statistic induced by sample permutation,

\[
\xi_{k}^{\text{GSEA}} = \begin{cases} \frac{S_{k}^{\text{GSEA}}}{E(S_{k}^{\text{GSEA}}|S_{k}^{\text{GSEA}} \geq 0)} & \text{if } S_{k}^{\text{GSEA}} \geq 0 \\ \frac{S_{k}^{\text{GSEA}}}{E(S_{k}^{\text{GSEA}}|S_{k}^{\text{GSEA}} < 0)} & \text{if } S_{k}^{\text{GSEA}} < 0 \end{cases}
\]

(9)

This rescaling is motivated by the asymptotic behavior of the Kolmogorov-Smirnov statistic (for details see supplementary information in Reference [12]). This normalization effectively puts the gene set enrichment scores on the same scale analogous to how the factor \(\sqrt{n_{k}}\) does it for the SEA z-score. This makes it possible to define a null distribution for the GSEA global statistics assuming a null distribution \(F_{0}^{\text{perm}}\) induced by sample permutation,

\[
F_{0}^{\text{GSEA}}, \xi_{1}^{\text{GSEA}}, \xi_{2}^{\text{GSEA}}, \xi_{3}^{\text{GSEA}}, \ldots, \xi_{K}^{\text{GSEA}} \text{ are identically distributed and } \xi_{k}^{\text{GSEA}} \sim F_{0}^{\text{perm}}. \quad (10)
\]

The null distribution \(F_{0}^{\text{perm}}\) corresponds to the Class 2 type subject-sampling null hypothesis of Barry et al.43

Our review of both methods clearly shows that SEA and GSEA differ in two important aspects: the choice of global statistics (z-score combined with \(\chi^{2}\) vs. weighted Kolmogorov-Smirnov) and the specific type of null hypothesis being assumed: gene-sampling/Class 1 vs. subject-sampling/Class 2.

The authors of SEA apply the method to the handful of examples used to introduce GSEA in Reference [12] and make the following claims in favor of the gene independence hypothesis and against GSEA:

(1) Differential gene expression scores can be assumed to be both independent and normally distributed.

(2) As a consequence of (1), the simpler gene set enrichment method (SEA) based on a one-sample \(t\)-test can effectively assess gene set enrichment in generic datasets. When SEA fails to find relevant gene sets it can be “fixed” by applying a second simple statistic (\(\chi^{2}\)).

(3) GSEA is computationally unnecessarily complicated. The complexity of using empirical null distributions in GSEA can be avoided by using theoretical (normal) null distributions. Moreover, the gene–gene independence assumption allows the adjustment for multiple hypotheses by using independent hypotheses FDR \(q\)-values49;
(4) SEA is faster and simpler but equivalent to GSEA and therefore GSEA should not be used. SEA should be the basis of new methodologies for gene enrichment analysis;
(5) GSEA is based on a Kolmogorov-Smirnov statistic, which is known to lack sensitivity and thus is rarely used.

We will refute the authors’ claims below by studying the gene independence assumption behind the SEA method and by performing an empirical study of 50 datasets where we focus on the consequences of assuming gene independence.

3 Empirical analysis of SEA and GSEA

In their SEA approach, Irizarry et al.\textsuperscript{37} justify the gene-independence assumption based on the seemingly “straight” behavior of the gene expression scores in a handful of Q–Q plots (Figure 1, Irizarry et al.\textsuperscript{37}) and conclude that “Barring a few outliers, which are likely associated to differentially expressed genes, the assumption appears appropriate in all datasets.” We believe this is an oversimplification and that the gene independence assumption is not appropriate in general. Here, we present the results of an empirical study to systematically evaluate SEA and GSEA and assess the effect of the gene independence assumption in a representative benchmark set of 50 expression datasets, including some from the GSEA paper,\textsuperscript{12} as well as many more from Gene Expression Omnibus (GEO), the InSilico DB database of datasets (http://insilico.ulb.ac.be) and others from the literature. The complete list of these publicly available datasets is included in Supplementary Table ST1. (Supplementary file published online only).

This benchmark set is much more comprehensive and representative of the universe of datasets that may be used in gene set enrichment analysis. Most of the datasets derive from more recently

![Figure 1. Histograms of z-scores for 1,000 permutations of the samples (grey) and gene identifiers (black), and the simpler enrichment analysis (SEA) null distribution $N(0, 1)$ for the (A) P53 and (B) pancreas datasets. The legend also shows the mean and variance of the distributions.](image)
generated data than the original GSEA examples, contain expression levels for greater numbers of genes, and display a much larger variety of phenotypic distinctions.

In order to perform our comparative analysis we ran implementations of GSEA and SEA on all 50 benchmark sets and computed the corresponding results tables, including enrichment scores, \( p \)- and \( q \)-values for each gene set. In order to make a fair comparison of both methods, we computed the GSEA \( q \)-values using exactly the same procedure as in SEA (i.e., computing \( q \)-values using the nominal \( p \)-values as inputs to the \( q \)-value \( \text{R} \) function/package\(^{49} \)). For SEA, in order to produce a single score per gene set, we generated both proposed scores (\( z \)-score and \( \chi^2 \)) and chose the one with smaller \( p \) value as was suggested in Reference [37].

First we note that SEA uniformly produces many more significant gene sets than GSEA. For example in the Pancreas dataset,\(^{50} \) featured in Figure 1(B), GSEA produces 121 significant gene sets, out of a total of 1368, at the suggested threshold (\( q \)-value < 0.25). In contrast, SEA produces 570 significant gene sets at the most stringent threshold of \( q \)-value < 0.05. This number is almost 5 times more than GSEA and accounts for 42% of the total number of gene sets. Similar remarkably large results sets are produced by SEA in other datasets. This overproduction of significant results is further exacerbated in newer more comprehensive datasets with larger numbers of genes and stronger gene-phenotype correlations. SEA produces large number of significant sets, many of which we suspect are false positives due to the assumption of gene–gene independence. It is, therefore not surprising, that among the SEA sets we find many of the significant results produced by GSEA.\(^{12} \)

It is difficult to assess in more direct terms the specificity/sensitivity of each method and specifically the exact percentage of false positives in SEA results because in general we do not have validated “ground truth” results for any given dataset. However, in our study, we will use the methodology of Gatti et al.\(^{21} \) where for each benchmark dataset, besides the observed global statistics corresponding to the relevant phenotype, we produced results for 1000 randomly permuted phenotypes. Because the phenotypes have been randomized, there is no significant correlation structure between the class labels and the gene profiles, but the gene–gene correlations are preserved. We will use the observed and random-permutation global statistics to perform two groups of analyses: (1) a study of the amount of variance inflation and the (2) inflation of \( p \) values.

### 3.1 Variance inflation

One of the effects of gene–gene correlations and dependency structure in microarray datasets is the increase of variance in the global statistics. For example, the SEA null hypothesis assumes the global statistics are i.i.d. and normally distributed,

\[
H^{\text{SEA},z} : S_1^z, S_2^z, S_3^z, \ldots, S_K^z \text{ are i.i.d and } S_K^z \sim N(0, 1).
\] (11)

Barry et al.\(^{43} \) studied the effects of gene dependency and found that for a difference of means statistic, quite similar to the SEA \( z \)-score, and gene set sizes that are small compared with the length of the gene list, the true variance of the statistic will differ from that under the i.i.d. class 1 gene-sampling null by a “variance inflation” factor \( \Gamma \) that can be approximated by,

\[
\Gamma = 1 + \left( n_k \rho_k^2 \right)
\] (12)

where \( \rho_k^2 \) is the average correlation between the global statistic (\( z \)-score) inside a gene set. For global statistics that are linear functions of the local statistic, \( \rho_k^2 \) can be approximated with \( \rho_k^X \), the average correlation between gene expression profiles of genes in a dataset. From equation (12), it is evident that the variance inflation increases with the product of the average intra-gene set correlation with
the size of the gene set. Therefore, a small number of positive correlations within a gene set can result in substantial variance inflation if the gene set is large enough. For SEA, the assumption of gene independence implies $\rho^X_k = 0$ and therefore a variance inflation of one. Gatti et al.$^{21}$ studied about 200 real datasets and found strong variance inflation effects, roughly in the range from 1 to 6, as a result of rather modest positive gene correlations. Their results demonstrated the importance of this effect and the unrealistic nature of the gene sampling (class 1 and SAE) null hypothesis.

In order to further investigate this issue, we analyzed the variance of the distribution of SEA $z$-scores under 3 null distributions: sample/phenotype-sampling (class 2), gene-sampling (class 1) and SEA (theoretical normal null). In order to produce histograms, we computed $z$-scores generated by 1000 permutations of the samples and 1000 permutations of the gene identifiers. Figure 1 shows the histograms of $z$-scores for P53$^{12}$ and pancreas$^{50}$ datasets from the benchmark set.

The histograms of $z$-scores for 1000 permutations of the samples (shown in grey in Figure 1) show clearly how the gene correlations noticeably increase the width of the distribution and consequently produce significant variance inflation. For example, in the P53 and Pancreas datasets, the variance of the $z$-score distribution for the sample permutations are 2.9 and 4.12, respectively. In contrast, the variances for the gene-sampling distributions are 0.96 and 1.02, which are very close to the SEA null value of 1.

Additional insight can be obtained from the histograms of individual average gene set internal correlations and variance inflation as shown in Figure 2. The gene correlations are on average mostly positive and relatively small, e.g., in the range $[0, 0.30]$ (Figure 2A and 2C), but their effect on variance inflation can be significant and equivalent to a few times the variance of the SEA null distribution as estimated by equation 12 (Figure 2B and 2D).

Because of the strong effects of the gene correlations on the global statistic distribution, the sample/phenotype-sampling class 2 (GSEA) null hypothesis is a much more realistic representative of the situation found in real datasets where there is always a non-negligible amount of gene correlations.

### 3.2 P-value inflation

The variance inflation in the global SEA statistic presented in the previous section is not only of academic interest. It has deleterious consequences in the form of high rates of false positives and inflated $p$ values. In Figure 3, we show histograms of $p$- and $q$-values for SEA and GSEA applied to the P53 (Figure 3(A)) and Pancreas datasets (Figure 3(B)) for 1000 random permutations of their phenotype labels. Recall that these permutations are performed to eliminate, to the extent possible, the biological differences between the two resulting groups. SEA produces a spurious over-population of low $p$ values that can be seen as a spike on the left side of the histogram in panel I of both Figure 3(A) and (B). These spikes include at least about 15% of all gene sets against the P53 dataset and about 20% of all gene sets against the Pancreas dataset.

The over-population of low $p$-values translates into a significant number of gene sets with low $q$-values (panel II of Figure 3(A) and (B)). Using a $q$-value threshold of 0.05, this spurious population includes 11.5% of all gene sets against the P53 dataset and 20.1% of all gene sets against the Pancreas dataset. In contrast, GSEA produces a practically flat histogram of $p$ values and no significant numbers of gene sets with low $q$-values as expected (Figure 3(A) and (B), panels III and IV). In this case, for example, the number of gene sets with $q$-values less than 0.05 is only 0.0248% for the P53 dataset and practically zero for the Pancreas dataset. Thus, we see that SEA produces many significant gene sets in the absence of a biological signal as a consequence of ignoring gene-gene correlations. This is a major drawback of the method.
In Figure 4 and Table ST2, we summarize the randomized phenotype results for the entire benchmark set by showing the percentage of gene sets with $q$-values less than 0.05 and 0.25 by SEA and GSEA. SEA uniformly produces a large number of false positives regardless of the choice of FDR threshold. To further demonstrate that this inflationary effect is indeed produced by the gene–gene correlations, we used the same 1000 randomly permuted phenotypes as before, but we also randomly permuted the gene identifiers effectively destroying the gene–gene correlations. In this case, SEA shows similar behavior to GSEA and neither show inflation of $p$ values (Figure 5(A) and (B)).

The above analysis convincingly demonstrates that ignoring gene–gene correlations and using the class 1 (gene sampling) theoretical null distribution\textsuperscript{13} have very negative consequences on the final results of SEA, and in fact on any enrichment method that assumes gene–gene independence. These results are very similar to those obtained by Gatti et al.\textsuperscript{21} using different datasets and a similar, but different, statistic.

All these results cast doubt on the validity of Irizarry et al.’s claims (i)–(iv) above.
Figure 3. Histograms of p-values and q-values obtained by running simpler enrichment analysis (SEA) and gene set enrichment analysis (GSEA) on 1,000 randomly permuted phenotypes in the P53 dataset (A) and the pancreas dataset (B). The y-axis shows the percentage of gene sets results.

Figure 4. Percentage of gene sets with False Discovery Rate (FDR) less than 0.05 and 0.25 using simpler enrichment analysis (SEA) and gene set enrichment analysis (GSEA) in 1,000 permutations of the phenotype labels for each dataset in the benchmark set.
The complexity of gene score distributions in gene sets

The heterogeneous distribution of the expression of genes in gene sets poses a technical challenge to properly define enrichment scores. This derives from the fact that most gene sets are noisy and imperfect, or too context-dependent to be modeled with a simple assumption or statistic. Sometimes, the genes sets are derived from generic “textbook” descriptions of a biological process, e.g., Biocarta pathways or GO ontologies, and thus have little relevance to coordinately expressed components of these pathways. In other cases, they are defined in a specific cellular context different from the one in which an investigator wishes to assess enrichment or they are a mixture of multiple biological processes that may not occur coherently in any single biological sample. For all these reasons, the distribution of gene scores displays rather complex multi-modal behavior and this in turn makes it difficult to define a single enrichment score that will work well across gene sets and datasets.

Examples of this complex behavior are illustrated in Figure 6, where the behavior of three selected gene sets is shown across three of the benchmark datasets. The different panels illustrate the complex multi-modal behavior of gene sets that may occur. We note that in panels I, III, V and IX, a relatively high enrichment score of the gene set is produced by a subset of the genes, rather than the entire gene set. The genes responsible for the enrichment appear at either the top or bottom of the ranked gene list and are representatives of relevant biological processes that are indeed enriched in the studied phenotype. These examples are typical and show the complexities in modeling that gene set enrichment analyses encounter in practice. They also explain why it is difficult to describe the behavior of gene sets analytically and why overly simplistic assumptions such as those used in Reference [37] are not likely to work. By taking these complexities into account, one can better appreciate the motivation for the weighted Kolmogorov-Smirnov statistic as a good compromise between expecting all or most genes to be coherent on one side and overweighting one or a few genes and allowing them to dominate the score on the other. It provides a reasonable distribution-free, but empirically adaptable, way to deal with the limitations and idiosyncrasies of real-life gene sets.
It may not be the most powerful statistic for any given simplistic circumstance, but the modification we made in Reference [12] of weighting by phenotype correlation is sensitive where it has to be and deals well with the behavior of real gene sets.

5 Discussion

The SEA method justifies the gene independence assumption based on the apparent normality of the local statistics (Q-Q plots) and the assumption that because intra-gene set gene correlations do not impact significantly the distribution of the global statistics. These assumptions are not supported by our empirical results. The strong effect of gene–gene correlations is well known and the need to take into account this dependency structure as part of gene set analysis has been well documented.22–25 Of special note is the study of Gatti et al.21 who performed a large empirical study of 200 datasets in their suggestively titled article “Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets” and also demonstrated the practical impact of strong gene–gene correlation patterns and strongly criticized the use of the gene-independence assumption.

One limitation of the SEA study37 is that they evaluated their method, and obtained their conclusions, in only a handful of datasets, namely the examples from the GSEA paper.12 Those datasets were good examples to illustrate the use of the recently revised GSEA method at that time, but they do not constitute a comprehensive benchmark to systematically study the

Figure 6. Gene set enrichment analysis (GSEA) individual gene set enrichment plots: examples of top scoring gene sets that display complex behavior.
properties of a new method or to perform a comparison between competing methods. The collection of gene sets they used (MSigDB in 2005) is also relatively small (522 gene sets) compared with later releases of the same collection (e.g., 1893 gene sets in MSigDB v2.5) where the overlap between gene sets is more significant.

The authors of SEA also criticized the use of the Kolmogorov-Smirnov statistic in GSEA based on their view that it lacks sensitivity and is rarely used (their claim (v) above). However, they failed to appreciate the fact that the weighted version used in GSEA is not the standard Kolmogorov-Smirnov statistic and was developed specifically to be more sensitive to differences in the tails of the distribution. Non-parametric statistics based on empirical cumulative distribution functions are the basis of new and powerful “goodness of fit” tests.51

Irizarry et al.37 also listed as one of the motivations for SEA the fact that GSEA is slow and hard to compute (claims (iii) and (iv)). In Table 1, we show typical CPU execution times (Mac book Pro 8.3) for SEA and GSEA applied to two benchmark datasets (P53 and Pancreas). As seen in the table, SEA is indeed faster than GSEA because it avoids the generation of an empirical null distribution. However, a typical GSEA run takes only about 7 min and for almost all users this is an acceptable running time. Moreover, SEA may be faster but this speed up comes at the high cost of large numbers of false positives as described above.

6 Conclusions

We have shown strong empirical evidence that gene–gene correlations cannot be ignored and should be taken into account by gene set enrichment methods. Our results agree with the extensive literature providing theoretical or empirical evidence against the gene independence assumption22–25 and counter the chief assumption of SEA.37

We benchmarked SEA against GSEA in a collection of 50 expression datasets. By randomizing phenotypes, we demonstrate that gene–gene correlations produce significant variance inflation in SEA results, which also exhibit very high false positive rates and significant numbers of inflated \(p\) and \(q\) values. Based on our empirical results, we have refuted the claims of Irizarry et al.37 and more broadly we recommend that methods that ignore gene–gene correlations, such as SEA, be avoided.

Finally, we believe that the most important improvements that can be made to gene set enrichment methods are (1) the improvement of gene set databases, like the MSigDB, so that their sets are less redundant and have more coherent behavior in actual biological samples and (2) the development of more sophisticated methodologies that more accurately take into account or model gene correlations and the dependency structure in the data (i.e., the opposite of the SEA approach).

By improving the resolution of gene sets we may overcome many of the limitations we have described above: noisiness, redundancy, multiple-process representation, poor specificity, etc. We are currently experimenting with the creation of a “hallmark” collection of gene sets in MSigDB that will contain more targeted representations of biological processes. In addition, we are investigating

| Method                  | P53 dataset | Pancreas dataset |
|-------------------------|-------------|------------------|
| SEA (R implementation)  | 12.32 s     | 11.35 s          |
| GSEA (desktop application) | 420 s     | 444 s            |

SEA: simpler enrichment analysis; GSEA: gene set enrichment analysis.

Table 1. Typical CPU execution times for SEA and GSEA in the P53 and pancreas datasets (Mac Book Pro 8.3)
the generation of coordinately expressed gene sets derived from the activation or repression of pathways in the laboratory.

Our group, and many others, are currently engaged in efforts to improve gene set analysis through better modeling of the datasets’ correlation structure and by introducing additional information about the behavior of genes (e.g., on a sample per sample basis), as part of the gene set analysis. There are indeed many alternative approaches to pursue these goals and here we conclude by listing just a few: single-sample gene set enrichment, computational and theoretical methods for assessing size and effect of correlation in large-scale testing, eigenvalue-decomposition of covariance matrixes, rotation-based sampling, correlation-adjusted t-scores, modeling dependency among the genes within and across each gene set, and multiple testing procedures using dependency kernels. Given the complexities of genomic data, it is worthwhile to pursue new methodologies such as these; however, we should resist the temptation to over-simplify and remember the admonishments of Pearson and Tukey:

...it is not enough to know that a sample could have come from a normal population; we must be clear that it is at the same time improbable that it has come from a population differing so much from the normal as to invalidate the use of the “normal theory” tests in further handling of the material.—E. S. Pearson

Far better an approximate answer to the right question, than the exact answer to the wrong question, which can always be made precise.—J. Tukey

7 Methods

Our implementation of GSEA is as described in Subramanian et al. The implementation of SEA followed the description given in Irizarry et al. We validated our implementation by exactly reproducing the t-test score and \( \chi^2 \) statistics from the result set of Irizarry et al. in one of the datasets.

We obtained the differential gene-expression scores using the signal to noise ratio as computed in GSEA and in SEA. We also normalized the differential gene expression scores by subtracting the median and dividing by the median absolute deviation as is done in SEA.

To perform the analyses of variance inflation and over-production of false positives and inflated p values, we followed the approach of Barry et al. and Gatti et al. For each dataset in the benchmark, we randomized the phenotype labels 1000 times and ran both algorithms. We also computed enrichment scores by randomizing the gene labels. The p values are computed using the areas under the empirical null histograms from GSEA and areas under the normal distribution for SEA. The q values were computed using the R function qvalue from package qvalue, which implements the method described in Storey and Tibshirani.

Acknowledgements

We thank Helga Thorvaldsdóttir, Aravind Subramanian and other members of the Cancer Program at the Broad Institute for suggestions, and Jon Bistline for help with references and formatting the figures.

Funding

This project was supported in part by a grant from the National Cancer Institute (R01-CA121941).
References

1. Pavlidis P, Lewis DP and Noble WS. Exploring gene expression data with class scores. Pac Symp Biocomput 2002; 474–485.

2. Gerstein M and Jansen R. The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? Cur Opin Struct Biol 2000; 10: 574–584.

3. Hakak Y, Walker JR, Li C, et al. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc Natl Acad Sci USA 2001; 98: 4746–4751.

4. Mirnics K, Middleton FA, Marquez A, et al. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. Neuron 2000; 28: 53–67.

5. Zien A, Kuffner R, Zimmer R, et al. Analysis of gene expression data with pathway scores. Proc Int Conf Intell Syst Mol Biol 2000; 8: 407–417.

6. Ben-Dor A, Bruhn L, Friedman N, et al. Tissue classification with gene expression profiles. J Comput Biol 2000; 7: 559–583.

7. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-lalpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003; 34: 267–273.

8. Mootha VK, Handschin C, Arlow D, et al. Erralpah and Gabpa/b specify PGC-lalpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. Proc Natl Acad Sci USA 2004; 101: 6570–6575.

9. Cunningham JT, Rodgers JT, Arlow DH, et al. mTOR controls mitochondrial oxidative function through a YY1-PGC-1alpha-responsive genes involved in oxidative phosphorylation. Nat Genet 2000; 26: 261–264.

10. Damien D and Gorfine M. Statistical concerns about the GSEA procedure. Nat Genet 2004; 36: 663, author reply 663.

11. Mootha VK, Daly MJ, Patterson N, et al. Reply to “statistical concerns about the GSEA procedure”. Nat Genet 2004; 36: 663.

12. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005; 102: 15545–15550.

13. Efron B. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge, UK: Cambridge Univ Press, 2010, p. 277.

14. Good PI. Analyzing the large number of variables in biomedical and satellite imagery. John Wiley & Sons, Inc., 2011, p. 200.

15. Huang DW, Sherman BT and Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009; 37: 1–13.

16. Zeeberg BR, Feng W, Wang G, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003; 4: R28.

17. Al-Shahrour F, Diaz-Uriarte R and Dopazo J. FatigiO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004; 20: 578–580.

18. Zhong S, Storch KF, Lipan O, et al. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. Appl Bioinform 2004; 3: 261–264.

19. Zhou X and Su Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics 2007; 8: 246.

20. Dennis Jr G, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003; 4: P3.

21. Gatti DM, Barry WT, Nobel AB, et al. Heading down the wrong pathway: on the influence of correlation within gene sets. BMC Genomics 2010; 11: 574.

22. Efron B. Correlation and large-scale simultaneous significance testing. J Am Stat Assoc 2007; 102: 93–103.

23. Lu X and Perkins DL. Re-sampling strategy to improve the estimation of number of null hypotheses in FDR control under strong correlation structures. BMC Bioinformatics 2007; 8: 157.

24. Qiu X, Klebanov L and Yakovlev A. Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. Stat Appl Genet Mol Biol 2005; 4: Article 34.

25. Goeman JJ and Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007; 23: 980–987.

26. Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 2005; 102: 13544–13549.

27. Efron B and Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat 2007; 1: 107–129.

28. Barry WT, Nobel AB and Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005; 21: 1943–1949.

29. Breslin T, Eden P and Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinformatics 2004; 5: 193.

30. Lee HK, Braynen W, Keshav K, et al. ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinformatics 2005; 6: 269.

31. Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 2007; 8: 242.

32. Shen H and West M. Bayesian modelling for biological annotation of gene expression pathway signatures. In: Chen M-H, Dey DK, Müller P, Sun D and Ye K (eds) Frontiers of statistical decision making and Bayesian analysis: in honor of James O Berger, 1st edn. New York, NY: Springer, 2010, pp. 285–302.

33. Nilsson B, Hakansson P, Johansson M, et al. Threshold-free high-power methods for the ontological analysis of genome-wide gene-expression studies. Genome Biol 2007; 8: R74.

34. Prifiti E, Zucker JD, Clement K, et al. FunNet: an integrative tool for exploring transcriptional interactions. Bioinformatics 2008; 24: 2636–2638.

35. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 2010; 26: i237–i245.

36. Sun CH, Kim MS, Han Y, et al. COFECO: composite function annotation enriched by protein complex data. Nucleic Acids Res 2009; 37: W350–W355.

37. Irizarry RA, Wang C, Zhou Y, et al. Gene set enrichment analysis made simple. Stat Meth Med Res 2009; 18: 565–575.

38. Wang L, Zhang B, Wolfinger RD, et al. An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet 2008; 4: e1000115.

39. Kim SY and Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 2005; 6: 144.

40. Smyth G, Gentleman R, Carey VJ, et al. Limma: linear models for microarray data. Bioinformatics and
computational biology solutions using R and bioconductor. New York, NY: Springer, 2005, pp.397–420.
41. Tu K, Yu H and Zhu M. MEGO: gene functional module expression based on gene ontology. *Biotechniques* 2005; **38**: 277–283.
42. Nam D, Kim SB, Kim SK, et al. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics* 2006; **22**: 2249–2253.
43. Barry WT, Nobel AB and Wright FA. A statistical framework for testing functional categories in microarray data. *Ann Appl Stat* 2008; **2**: 296–315.
44. Dudoit S, Keles S and van der Laan MJ. Multiple tests of association with biological annotation metadata. In: Nolan D and Speed T (eds) *Probability and statistics: essays in honor of David A Freedman*. Berkeley, CA: The Berkeley Electronic Press, 2006, pp.153–218.
45. Newton MA, Quintana FA, Boon JAD, et al. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann Appl Stat* 2007; **1**: 85–106.
46. Emmert-Streib F and Glazko GV. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput Biol* 2011; **7**: e1002053.
47. Tusher VG, Tibshirani R and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001; **98**: 5116–5121.
48. Storey JD. A direct approach to false discovery rates. *J Roy Stat Soc: Ser B (Stat Methodol)* 2002; **64**: 479–498.
49. Storey JD and Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100**: 9440–9445.
50. Abdollahi A, Schwager C, Kleeff J, et al. Transcriptional pathway governing the angiogenic switch in human pancreatic cancer. *Proc Natl Acad Sci USA* 2007; **104**: 12890–12895.
51. Zhang J. Powerful goodness-of-fit tests based on the likelihood ratio. *J Roy Stat Soc: Ser B (Stat Methodol)* 2002; **64**: 281–294.
52. Edelman E, Porrello A, Guinney J, et al. Analysis of sample set enrichment scores: assessing the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 2006; **22**: e108–e116.
53. Lee E, Chuang HY, Kim JW, et al. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol* 2008; **4**: e1000217.
54. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 2009; **462**: 108–122.
55. Nam D. De-correlating expression in gene-set analysis. *Bioinformatics* 2010; **26**: i511–i516.
56. Wu D, Lim E, Vaillant F, et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* 2010; **26**: 2176–2182.
57. Zuber V and Strimmer K. Gene ranking and biomarker discovery under correlation. *Bioinformatics* 2009; **25**: 2700–2707.
58. Sohn I, Owzar K, Lim J, et al. Multiple testing for gene sets from microarray experiments. *BMC Bioinform* 2011; **12**: 209.