Assessment Method for a Power Analysis to Identify Differentially Expressed Pathways

Shailesh Tripathi, Frank Emmert-Streib*

Computational Biology and Machine Learning Lab, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

Abstract

Gene expression data can provide a very rich source of information for elucidating the biological function on the pathway level if the experimental design considers the needs of the statistical analysis methods. The purpose of this paper is to provide a comparative analysis of statistical methods for detecting the differentially expression of pathways (DEP). In contrast to many other studies conducted so far, we use three novel simulation types, producing a more realistic correlation structure than previous simulation methods. This includes also the generation of surrogate data from two large-scale microarray experiments from prostate cancer and ALL. As a result from our comprehensive analysis of 41,004 parameter configurations, we find that each method should only be applied if certain conditions of the data from a pathway are met. Further, we provide method-specific estimates for the optimal sample size for microarray experiments aiming to identify DEP in order to avoid an underpowered design. Our study highlights the sensitivity of the studied methods on the parameters of the system.

Citation: Tripathi S, Emmert-Streib F (2012) Assessment Method for a Power Analysis to Identify Differentially Expressed Pathways. PLoS ONE 7(5): e37510. doi:10.1371/journal.pone.0037510

Editor: Zhi Xie, National Institutes of Health, United States of America

Received November 26, 2011; **Accepted** April 22, 2012; **Published** May 18, 2012

Funding: This project is supported by the Center for Cancer Research and Cell Biology (QUB) and the Department for Employment and Learning through its "Strengthening the all-Island Research Base" initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. ST is supported by a studentship from the National Institute of Immunology.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: v@bio-complexity.com

Introduction

The functional analysis of high-throughput data is a challenging but promising direction in the post-genomics era. It is challenging because genome-wide data are high-dimensional and noisy, but promising due to its potential to reveal knowledge about the systemic working mechanisms of biological information processing within cells, which we are currently lacking [1–4].

In the context of expression data the interest shifted in recent years from approaches focusing on the analysis of individual genes, detecting their differentially expression [5–7], toward the analysis of gene sets in order to identify differentially expressed sets of genes [8–11]. The rational behind this is that genes and their products do not work in isolation but interact with each other in a concerted manner in order for a phenotype to emerge [12]. Many univariate, multivariate and nonparametric statistical methods have been either newly developed or existing methodological techniques have been adapted for this problem [13–16]. One important property that allows to distinguish different types of such hypotheses tests was discussed in [17]. There, tests have been distinguished based on the data used for the comparison. A hypothesis test comparing a gene set to all other gene sets available is called *competitive*, whereas a test comparing the same gene set for two different phenotypes is called *self-contained*.

The name '*gene set*' associated with the above methods implies that the choice for defining gene sets by populating them with specific genes is not constraint. However, in the present study, we are considering only gene sets that have been defined by using biological information regarding their association with specific pathways extracted, e.g., from the gene ontology database [18] or KEGG [19]. For this reason we refer to them in the following as *pathway-based methods* [20] to indicate this explicitly.

The major goal of this paper is to compare two self-contained (*sum of t-square* and Hotelling's T^2 [21,22]) and one competitive test (GSEA [23]) with each other, for a variety of simulated and biological expression data, in order to gain insights into the dependence of the power of these methods on the correlation structure among genes. The reason for selecting these tests is their complementary nature, representing univariate (*sum of t-square*), multivariate (Hotelling's T^2) as well as competitive and self-contained tests. In [24] it has been shown that there are currently only about three different null hypotheses effectively tested among all self-contained tests, which include the null hypothesis of the *sum of t-square* test and Hotelling's T^2. Because GSEA is a competitive test, its null hypothesis is conceptually different to the above ones making the three selected methods complementary to each other with respect to the tested null hypotheses. The reason for choosing GSEA over other competitive methods, which are based on methodological extensions [25,26], is the popularity of this method especially among biologists [27], and the vast number of studies it has been already used for.

In contrast to the many studies that have been conducted so far investigating the power of methods for identifying differentially expressed pathways, we are focusing on the *correlation structure* in gene expression data. Other studies conducted in this context either did not consider the correlation among genes [25,28], assumed a constant [13,29], a random [30], an autoregressive [31] or a compound symmetry correlation structure among genes [14,32,33] or studied real microarray data only [34–36], which do not allow to study different configurations by adjusting model parameters. The main problem with the above simulation studies is that they do not provide a realistic correlation structure among genes. The reason for this is that the made assumptions do not lead to a gene network-like correlation structure as observed for gene expression data. Technically, this means that the inverse of the covariance matrix does not reflect the independence relations that can be found in such network structures, as we will discuss in detail in section 'Simulation of network-like correlation structure'. However, due to the fact that we are only considering gene sets that correspond to biological pathways, the strength of the correlation and its structure are important parameters that need to be controlled properly in order to make the transition from gene sets to pathways. In order to overcome this severe limitation, two algorithms have been developed, in a different context, for generating a covariance matrix for a multivariate normal distribution whose inverse is consistent with the independence relations of a network [37,38]. In order to generate such a covariance matrix, both methods need a network as an input for their algorithm. For our simulation study, we employ both algorithms for generating simulated expression data with a gene network-like correlation structure by using a protein interaction network and a transcriptional regulatory network from yeast as input network. This provides a biologically realistic constraint on the resulting correlation structure. In the following we call these simulation types III and IV. Here by '*biologically realistic*' we mean that an experimentally determined protein interaction network and a transcriptional regulatory network are more realistic than artificially generated network structures using a statistical method. In order to conduct a comprehensive analysis of important system parameters, we include in our study also the influence of the sample size, detection call, i.e., the percentage of genes that is differentially expressed within a pathway, and the pathway size on the identification of differentially expressed pathways.

In addition to simulated expression data, we use also two large-scale cancer data sets from DNA microarray experiments [39,40]. By applying a bootstrap approach [41,42], we use these data sets to generate surrogate data of smaller sample sizes. This allows us to study the robustness of the statistical methods over a wide range of realistic sample sizes without the need for making assumptions about the underlying pathology of the pathways in order to declare, e.g., pathways as true positives. Further, we compare the correlation structure of simulated and biological pathways as defined via the gene ontology database [18].

Methods

Pathway-based method

GSEA. This method was introduced by [11,23] in order to identify the differential expression of predefined gene sets. GSEA is considered a competitive test [17] because it compares a test set to a background data set. Let W be the set of genes to be tested and W^c its complement in a way that the union of both sets gives all genes, i.e., $V = W \bigcup W^c$. Briefly, GSEA consists of the following steps, applied to each pathway:

(1) Estimation of gene-wise test statistics.
(2) Rank ordering of the test statistics.
(3) Calculation of an enrichment score (ES) for a pathway.
(4) Permutation of the gene-labels to estimate the significance of the enrichment score (p-value) for the pathway.

The hypotheses tested by GSEA are:

$H_0 : ES = 0$ - vanishing test score
$H_1 : ES \neq 0$ - non-vanishing test score

Hotelling's T^2. The Hotelling T^2 test is a self-contained test that is a multivariate generalization of the univariate t-test. Its null and alternative hypothesis can be formulated as:

$H_0 : \boldsymbol{\mu}^T = \boldsymbol{\mu}^C$ - equality of the p-dimensional population mean vectors
$H_1 : \boldsymbol{\mu}^T \neq \boldsymbol{\mu}^C$ - difference of the p-dimensional population mean vectors

Suppose we have two groups with n_C samples from the control group and n_T samples for the treatment group, each consisting of p genes. Let the expression level of the i^{th} sample of the control group and treatment group be given by $X_i^C = (X_{i1}^C, X_{i2}^C, \ldots X_{ip}^C)^t$ and $X_i^T = (X_{i1}^T, X_{i2}^T, \ldots X_{ip}^T)^t$, respectively. The pooled covariance matrix $\mathbf{S}$ is then defined by

$$\mathbf{S} = \frac{(n_T - 1)\mathbf{S}_T + (n_C - 1)\mathbf{S}_C}{(n_T + n_C - 2)} \tag{1}$$

where $\mathbf{S}_C$ and $\mathbf{S}_T$ are the covariance matrices for the control and treatment group. Hotelling's T^2 is defined as

$$T^2 = \frac{n_T \times n_C}{n_T + n_C} (\boldsymbol{\mu}^T - \boldsymbol{\mu}^C)\mathbf{S}^{-1}(\boldsymbol{\mu}^T - \boldsymbol{\mu}^C)^t. \tag{2}$$

The inverse of the covariance matrix is estimated via the shrinkage estimator [43–46]. The statistical significance of the test statistic T^2 is estimated from sample-label permuted data.

Sum of t-square. The *sum of t-square* test is an univariate test based on t-scores, $\{t_i\}$, obtained for each of the p genes individually for a given set [22]. The test statistic for each pathway is given by

$$\sum_{t,2} = \frac{1}{p}\sqrt{\sum_i^p t_i^2}. \tag{3}$$

Its null and alternative hypothesis can be formulated as:

$H_0 : \sum_{t,2} = 0$ - vanishing test score
$H_1 : \sum_{t,2} \neq 0$ - non-vanishing test score

Again, the significance of $\sum_{t,2}$ is assessed from sample-label permuted data.

Simulation algorithms

In order to assess the performance of the statistical methods we use three principally different algorithms to simulate expression data.

Simulation of uncorrelated data. For this method we, first, define different non-overlapping pathways of varying sizes including a total of p genes. Then we draw iid (Independent and identically distributed) samples from a standard normal distribution, i.e., $X_{ij}^G \sim N(0,1)$, for each gene $i \in \{1, \ldots, p\}$ and sample $j \in \{1, \ldots, n\}$, for the control $(G = c)$ and treatment $(G = t)$ group. In order to make a difference between the control and treatment group we add a constant factor of one to a certain percent of genes of all pathways for the treatment group.

Simulation of correlated data. First, we generate a matrix X with p rows and $2n$ columns with a sample size of n for the control and treatment group, i.e., X_{ij} with $j \in \{1,\ldots,n\}$ corresponds to the control and $j \in \{n+1,\ldots,2n\}$ to the treatment group. Each component of X is independently sampled from a standard normal distribution, i.e., $X_{ij} \sim N(0,1)$. Then we generate a $2n$-dimensional random vector a whose components are also iid drawn from the standard normal distribution. Define

$$Y_{ij} = \sqrt{\rho} a_j + \sqrt{1-\rho} X_{ij} \tag{4}$$

where $i \in \{1,\ldots,p\}$ and $j \in \{1,\ldots,2n\}$, so that the average correlation between the genes (rows of Y) is ρ [29]. To model differential expressed pathways we add a constant factor of one to a certain percent of genes for the treatment group.

Simulation of a network-like correlation structure. For random variables that are from a p-dimensional multivariate normal distribution, i.e., $X_p \sim N(\mu,\Sigma)$, a simple relation between the components of the inverse covariance matrix $\Omega = \Sigma^{-1}$ (also called precision or concentration matrix) and the conditional partial correlation holds [47]

$$\rho_{ij|V\setminus\{ij\}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}. \tag{5}$$

Here $\rho_{ij|V\setminus\{ij\}}$ is the partial correlation coefficient between gene i and j conditioned on all remaining genes and ω_{ij} are the components of Ω. That means if $\rho_{ij|V\setminus\{ij\}} = 0$ then gene i and j are independent from each other,

$$X_i \perp X_j | \{\text{all remaining genes}\}, \tag{6}$$

if and only if $\omega_{ij} = 0$. A multivariate normal distribution that is *Markov* with respect to an undirected network G is called a *Gaussian graphical model*. This means that all conditional independence relations that can be found in Σ^{-1} hold also in G [47]. Hence, such a Σ^{-1} can be considered as consistent with all conditional independence relations in G.

We use the following two algorithms to obtain a covariance matrix for a given network structure G.

(1) The algorithm of [38] is based on successive orthogonal projections constraint by the network structure G, resulting in a consistent covariance matrix Σ.

(2) The algorithm of [37,48] is based on proportional iterative fitting [47] to enforce an average correlation in the data resulting in a covariance matrix Σ consistent with the conditional independence relations in G.

As input network G for these algorithms we use two yeast networks. The first is a protein-protein interaction network provided by the *Biogrid* database [49] and the second is a transcriptional regulatory network [50]. From both networks, we extract the giant connected component. The reason for selecting these networks is that a protein and a transcriptional regulatory network represent *observed* interaction structures among genes and gene product and, hence, provide a more realistic structure than artificially generated networks, e.g., by using the *preferential attachment* model to generate scale-free networks [51]. Once a covariance matrix Σ from one of the above algorithms is obtained, we use Σ to generate iid samples from a multivariate normal distribution, i.e., $X \sim N(\mu,\Sigma)$. In order to simulate the differentially expression of pathways, we use a p-dimensional mean vector of zero, $\mu = 0^p$, for the control group, and a mean vector μ consisting of DC genes with an expression of 1 and $(1-DC)$ genes with an expression of 0 for the treatment group. For both groups, we use the same covariance matrix.

We would like to note that due to the properties of the Gaussian graphical model, as discussed above, this model has been used to infer gene regulatory networks from expression data [46,52–54]. This indicates that the relation between the components of the inverse of the covariance matrix and the independence relations found in a network structure are generally considered to be biologically connected with each other for expression data. Hence, this provides a justification to consider the correlation structure generated from a Gaussian graphical model as biologically plausible.

Simulation types

For our analysis we are using four different simulation types (ST) based on the algorithms described above. Because GSEA is a *competitive test* [17] it requires a background data set against which a pathway is compared. For ST I we simulate such a background data set explicitly. This background data consists of 10000 genes with expression values sampled from a normal distribution $N(0,1)$, for both conditions, and a global correlation structure is imposed by Eqn. 4. For ST II–IV we use, instead, the remaining data *excluding* the pathway under investigation, as background data.

Simulation type I. For this type of simulation we generate simulated expressed data for all $p = 10000$ genes simultaneously, as described in methods section 'Simulation of uncorrelated data' and 'Simulation of correlated data', for which we define non-overlapping pathways of sizes ranging from 5 to 195 (step size 10), $\mathcal{P} = \{5,15,\ldots,195\}$. For each of these $|\mathcal{P}| = 20$ different pathway sizes we generate 5 pathways, resulting in a total of 100 different non-overlapping pathways that contain in total $p = 10000$ genes. Parameters studied: We study the influence of the sample size ($n \in \mathcal{N} = \{5,10,15,\ldots,45\}$, $|\mathcal{N}| = 9$), detection call ($DC \in \mathcal{DC} = \{0\%,10\%,30\%,60\%\}$, $|\mathcal{D}| = 4$) and of the correlation ($\rho \in \mathcal{R} = \{0.0,0.2,0.4,0.6\}$, $|\mathcal{R}| = 4$). Here, the detection call (DC) [24] refers to the percent of differentially expressed genes in a pathway and ρ refers to the correlation between all genes in the overall set. This gives $2880 (= |\mathcal{P}| \cdot |\mathcal{N}| \cdot |\mathcal{D}| \cdot |\mathcal{R}|)$ different parameter configurations. For GSEA, we generate a background data set *without expression difference* between the treatment and control group, i.e., $X^t_{ij}, X^c_{ij} \sim N(0,1)$.

Simulation type II. Here we generate simulated data separately for each pathway, as described in methods section 'Simulation of correlated data'. In contrast to ST I, ST II generates a correlation among the genes within a pathway. We use an overall set of $p = 10000$ genes to define non-overlapping pathways, as for ST I. Parameters studied: We study the influence of the sample size ($n \in \mathcal{N} = \{5,10,15,\ldots,45\}$, $|\mathcal{N}| = 9$), detection call ($DC \in \mathcal{DC} = \{0\%,10\%,30\%,60\%\}$, $|\mathcal{D}| = 4$) and of the correlation ($\rho \in \mathcal{R} = \{0.2,0.4,0.6\}$, $|\mathcal{R}| = 3$). Here, ρ refers to the correlation for the genes within a pathway, whereas the average correlation among all genes is about zero. This gives $2160 (= |\mathcal{P}| \cdot |\mathcal{N}| \cdot |\mathcal{D}| \cdot |\mathcal{R}|)$ parameter configurations.

Simulation type III and IV. For this ST we generate simulated expressed data by sampling from a p-dimensional Gaussian graphical model. In order to obtain a more realistic correlation structure we use two different algorithms, as described in methods section 'Simulation of network-like correlation structure' in combination with a protein interaction network and a transcriptional regulatory network. We used the *gene ontology* database [18] to map the proteins to their corresponding *biological process* for level 4. From this information we selected 76 different but overlapping pathways that consist in total of $p=1588$ genes for algorithm (1) and 41 pathways for 612 genes for algorithm (2). For the transcriptional regulatory network we select 200 different but overlapping pathways that consist in total of $p=1199$ gene for algorithm (1) and (2). Parameters studied for both algorithms: We study the influence of the sample size ($n \in \mathcal{N}=\{5,10,15,\ldots,45\}$, $|\mathcal{N}|=9$) and the detection call ($DC \in \mathcal{DC}=\{0\%,10\%,30\%,60\%\}$, $|\mathcal{DC}|=4$). The overall average correlation between all genes is approximately zero. In addition, for algorithm (2) we study also different values of the correlation, ($\rho \in \mathcal{R}=\{0.2,0.4,0.6\}$, $|\mathcal{R}|=3$). This gives $2736(=\mathcal{N}\cdot\mathcal{D}\cdot 76)$ (ST III) and $4428(=\mathcal{N}\cdot\mathcal{D}\cdot\mathcal{R}\cdot 41)$ (ST IV) different parameter configurations for the protein interaction network and $7200(=\mathcal{N}\cdot\mathcal{D}\cdot 200)$ (ST III) and $21600(=\mathcal{N}\cdot\mathcal{D}\cdot\mathcal{R}\cdot 200)$ (ST IV) different parameter configurations for the transcription regulatory network.

We would like to note that in the results section, the estimates for the *false positive rate* (FPR) have been obtained by setting $DC=0\%$, which corresponds to the case of no differentially expressed pathways [15,24,25].

Surrogate data: ALL and prostate cancer

To assess the power of the three pathway-based methods for microarray data we use two different large-scale data sets based on Affymetrix chips. The first is a prostate cancer data set consisting of 50 control samples and 52 tumor samples [40]. The second data set is from B-cells derived from Acute lymphoblastic leukemia (ALL) [39]. From the entire data set we select 37 samples from the BCR/ABL group and 37 from the NEG group. For the preprocessing and normalization of these data sets we followed [39,40]. After the normalization, we map the genes to the category *biological process* of level four in the gene ontology database [18] in order to obtain information about their association to biological pathways. For prostate cancer we obtain 213 different pathways and for ALL 533.

Our analysis of these data consists of two steps. In the first step, we generate a reference list by testing the significance of pathways for the total number of samples s (prostate: $s=102$, ALL: $s=74$). For the following analysis we use the results from this analysis as reference, because we consider the significant pathways as *true positives* and the nonsignificant pathways as *true negatives*. In the second step, we construct $b(s_i)$ bootstrap data sets for various sample sizes, $s_1>s_2>\cdots>s_{k-1}>s_k$, each data set drawn from the total of s available samples. For each bootstrap data set, each method is applied and a p-value obtained for each pathway. From this, a result is assessed as true positive, true negative, false positive or false negative with respect to the reference list obtained for sample size s (step one). Due to the fact that our reference list may contain false declarations, our results assess the statistical robustness of the methods providing estimates for, e.g., their power, rather than their true value. Further, because we generate bootstrap data sets for each sample size s_i, we consider these as surrogate data for newly generated data from independent experiments, which are not available.

Results

For the following simulations, we use a significance level of $\alpha=0.05$.

Simulation type I and II

In Fig. 1 A-C we show the power for GSEA (red curves), Hotelling's T^2 (green curves) and *sum of t-square* (blue curves) for simulation type I and II and different parameter settings. Here by the power we mean the probability that the statistical hypothesis test is rejected when the null hypothesis is truly false [55]. Practically, we estimate this probability by the population mean over repeated simulations [24]. The different color shadings code for different DC values; DC = 10% (light color), 30% (medium color), 60% (dark color). In these figures, a 'dot' corresponds to a mean value, and the error bars refer to its standard deviation obtained from 50 bootstrap samples. Each figure is indexed by the strength of the mean correlation.

For ST I (Fig. 1 A) the correlation has a much stronger influence on the power of *sum of t-square* and GSEA than on Hotelling's T^2, although Hotelling's T^2 has generally a lower power. Also, the influence of the detection call is for the *sum of t-square* and GSEA strongest resulting in a considerable loss in power for $DC<30\%$. Hotelling's T^2 appears to be relative insensitive against different values of DC. For high correlations and $DC=10\%$ the *sum of t-square* test has by far the worst power. For ST II (Fig. 1 C) GSEA performs significantly worse for all values of DC, compared to simulation type I, showing an almost complete break down. The power of *sum of t-square* and Hotelling's T^2 are comparable to the results for ST I. Regarding the number of significant pathways detected by the three methods, one can see that the *sum of t-square* test declares consistently more pathways as significant than any other method, for all conditions, except for very small sample sizes (≤ 10) for ST I. In this case GSEA declares more pathways as significant. In general, the lower the DC value the lower the number of pathways declared as significant, whereas lower values having a stronger influence. Further, it is interesting to note that for $\rho=0.0$ (Fig. 1 B) the *sum of t-square* and Hotelling's T^2 are different from each other despite that fact that in this case both tests should provide similar result, because the pooled covariance matrix S (see Eqn. 1) becomes diagonal. This indicates a poor behavior of the shrinkage estimator.

For ST II GSEA declares considerably less pathways as significant compared to the other methods, for all conditions. Regarding the false positive rate (FPR), GSEA and *sum of t-square* show a good control of the FPR at a significance level of $\alpha=0.05$. This is in contrast to Hotelling's T^2 which has even for large sample sizes a FPR larger than 0.20. In order to find the cause for this behavior we split the pathways into two categories. In the first category we put all pathways having less than 35 genes, in the second category we place all larger pathways (results not shown). From this analysis we find that also Hotelling's T^2 controls the FPR, but only for pathway sizes less than 35. The reason for this behavior is related to the estimation of the inverse of the covariance matrix, S^{-1}, on which Hotelling's T^2 is based. For smaller pathways, their number of genes, p, is closer to the number of samples, n_C and n_T, and, hence, the estimates for S^{-1} are more accurate than for larger pathways. Hence, for larger pathways one would need to improve the shrinkage estimator. Figure 1 B shows the result for uncorrelated data ($r=0.0$). In this case ST I and II coincide with each other. In general, for the uncorrelated case the power is slightly higher for all methods and also the number of pathways declared significant increases.

Simulation type III and IV

Fig. 2 (A and B) shows the results for simulation type III and IV. Here the *sum of t-square* and Hotelling's T^2 perform much better than GSEA. Interestingly, for ST IV and high correlations ($r=0.6$) and small sample sizes (≤ 25) the power of Hotelling's T^2 is even slightly higher than for the *sum of t-square* test and, more importantly, it is more robust with respect to DC values less than 60%. For the number of significant pathways we find again that GSEA declares less pathways as significant. Hotelling's T^2 does not control well the false positive rate for small sample sizes (≤ 20)

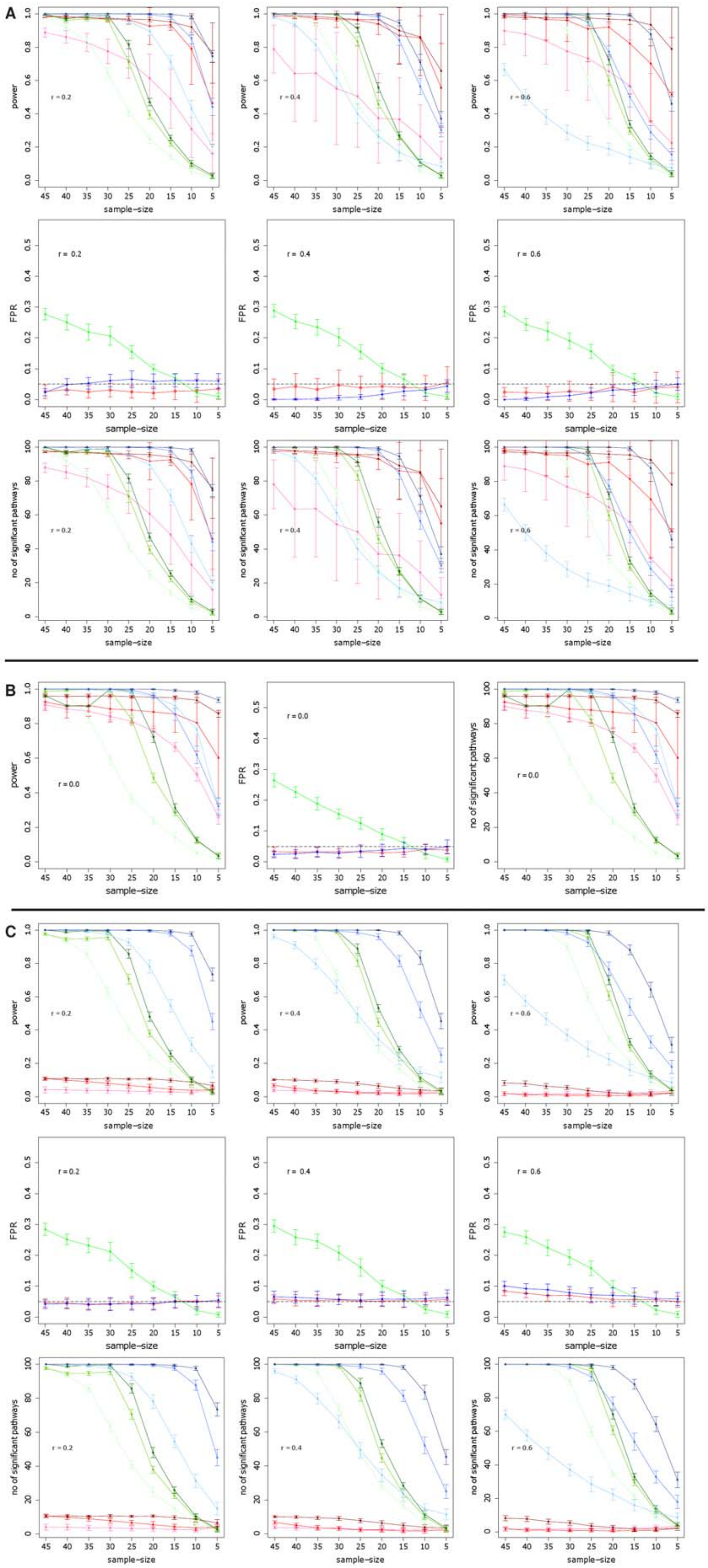
A
power
FPR
no of significant pathways
sample–size
r = 0.2
r = 0.4
r = 0.6
45 40 35 30 25 20 15 10 5
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.1 0.2 0.3 0.4 0.5
0 20 40 60 80 100
B
power
FPR
no of significant pathways
sample–size
r = 0.0
C
power
FPR
no of significant pathways
sample–size
r = 0.2
r = 0.4
r = 0.6

Figure 1. Simulation type I (A, B) and II (C): Power, FPR and number of significant pathways for GSEA (red), *sum of t-square* (blue) and Hotelling's T^2 (green). DC = 10% (light color), 30% (medium color), 60% (dark color).
doi:10.1371/journal.pone.0037510.g001

and has more problems in controlling the FPR with high correlations. For the *sum of t-square* the FPR is in general controlled, except for ST IV and $r = 0.6$. It is of interest to note that GSEA is the only method that controls the FPR for all conditions well.

The significant reduction of the power for GSEA for simulation type I compared to simulation type II to IV can be explained by the background data that have been generated for simulation type I, but not for the other simulation types. Due to the fact that GSEA is a competitive hypothesis test which estimates a test statistics w.r.t the background, the background data have a prominent influence on the power of GSEA whereas a large background dataset without expression changes in the conditions increases the sensitivity of this test.

In Fig. 3, we show similar results as in Fig. 2, however, for the transcriptional regulatory network of yeast [50] instead of the protein network, to generate simulated data. Overall, these results have a large resemblance to the results in Fig. 2. This demonstrates that structural properties of both networks on the pathway level are sufficiently similar to each other to result in similar results for the pathway methods. This corresponds, e.g., to the known similarity of the scale-free degree distribution of these networks [56].

Surrogate data: ALL and prostate cancer

The results for prostate cancer (left) and ALL (right) are shown in Fig. 4. We want to re-emphasize that we used the pathways declared as significant for the total number of samples (prostate: $s = 102$, ALL: $s = 74$) as a reference list. Hence, the power is related to these pathways and not to the *truely* expressed pathways. Similarly, the interpretation for the FPR and the number of significant pathways. On the first sight, all methods seem to perform similarly, although, GSEA has for high sample sizes the lowest power. However, Hotelling's T^2 and *sum of t-square* declare for all studied sample sizes many more pathways as significant than GSEA. Similar to the simulation studies, GSEA controls the FPR well whereas the other two methods assume larger values.

An interesting observation of the power is its rapid decay for an even slightly reduced sample size. More precisely, for prostate cancer the total sample size is $s = 102$ for which we identified the set of significant pathways we consider as 'true positives'. However, when using only 90 bootstrap samples (45 samples per condition) then the power of all three methods is already smaller than 1, see Fig. 4. Even more severe effects are observed for ALL, where the total sample size is $s = 74$, and results for 64 bootstrap samples (32 samples per condition) show a clearly reduced power. This lack of robustness for the initial sample sizes (45 for prostate cancer and 32 for ALL) suggests that the available *total* sample sizes are too small for the employed test statistics, because otherwise we would observe a stable plateau as in Fig. 1, 2 and 3, where a slight reduction of the sample size does not influence the power at all. Hence, from this decay, one can conclude that the total number of samples (prostate: $s = 102$, ALL: $s = 74$) of both cancer data sets (ALL and prostate cancer) is not sufficiently large for the point estimator of the power to converge. This hints to a refinement of the experimental design of studies aiming to detect the DEP to avoid a study that is underpowered.

In order to quantify this observation, we performed a linear regression analysis. For this analysis we use the size of the microarray experiments as predictor variable and the initial step size of the power curves (Fig. 4) as outcome variable, measured by its *distance to convergence*, as found from the comparison with our simulation results in Fig. 2 and 3. That means from Fig. 2 and 3 we obtain the minimal sample sizes for which the power reaches '1', and the *initial step* size of the power corresponds to the power for 45 (prostate cancer) and 32 (ALL) samples from Fig. 4. We are only using the results from ST III and IV for this comparison, because they resemble more closely the correlation structure of real microarray data. We conduct a separate analysis to predict the optimal sample size for each method, see Fig. 5. Here *optimal* refers to the minimal sample size for a method to become invariant against the removal of a small number of samples.

For the regression, we obtain F-statistics (18.83, 7.93 and 25.6) for the three linear regressions, in the order of the figures in 5, which are all significant with p-values of 0.0006, 0.0145 and 0.0071. As a result of this analysis, we predict a sample size of 59 for Hotelling's T^2 and 57 for the *sum of t-square* test (red crosses in Fig. 5). Due to the fact that for GSEA, its power does not converge in our simulation study for ST III and IV, we cannot make a prediction for this method. If we use the results form ST I instead, we obtain an estimated sample size of 83 for GSEA. We would like to emphasize that we consider these estimates as optimistic and, hence, as lower bounds for optimal sample sizes since the simulations constitute only approximations of real data.

A central topic of this paper is the investigation of the influence of the correlation strength and its structure on the identification of differentially expressed pathways. In the introduction we presented arguments supporting the need for such an analysis. Now we add quantitative evidence, directly extracted from the used expression data from prostate cancer and ALL. As discussed in the methods section 'Surrogate data: ALL and prostate cancer', both microarray data sets were normalized. Estimating the average correlation among all genes from the normalized data results in 0.0618 and 0.0138 for ALL and prostate cancer, which are quite small correlation values. However, if we estimate the average correlation among all genes within *each pathway*, we obtain an entirely different result. In Fig. 6 we show these results by ordering the pathways according to their average correlation coefficient. The different number of the pathways results from the fact the we consider 213 pathways for prostate cancer and 533 pathways for ALL, as explained in section 'Surrogate data: ALL and prostate cancer'. Despite the fact that the average correlation among all genes is 0.0618 for ALL (blue) and 0.0138 for prostate cancer (violet), shown as dashed lines in Fig. 6, one can clearly see that within the pathways there is a non-neglectable correlation, which spans a very wide range of different values, as summarized by the two vertical intervals on the left-hand side (violet: prostate cancer; blue: ALL). From these results, we can draw the following conclusions. First, even in normalized expression data there exist quite large correlations within particular pathways, which exhibit much larger values than the average correlation between all genes in the data set. The reason for this is that the purpose of any normalization method is to reduce reduce correlations due to technical artifacts and batch effects in the data but not real biological correlations between genes. These results justify also the selected correlation values for our simulations, which assume correlations up to 0.6. Second, there is a wide dynamic range of observed correlation coefficients that points toward a heterogeneity among the pathways. That means not all pathways possess the same characteristics but they can be quite different from each other.

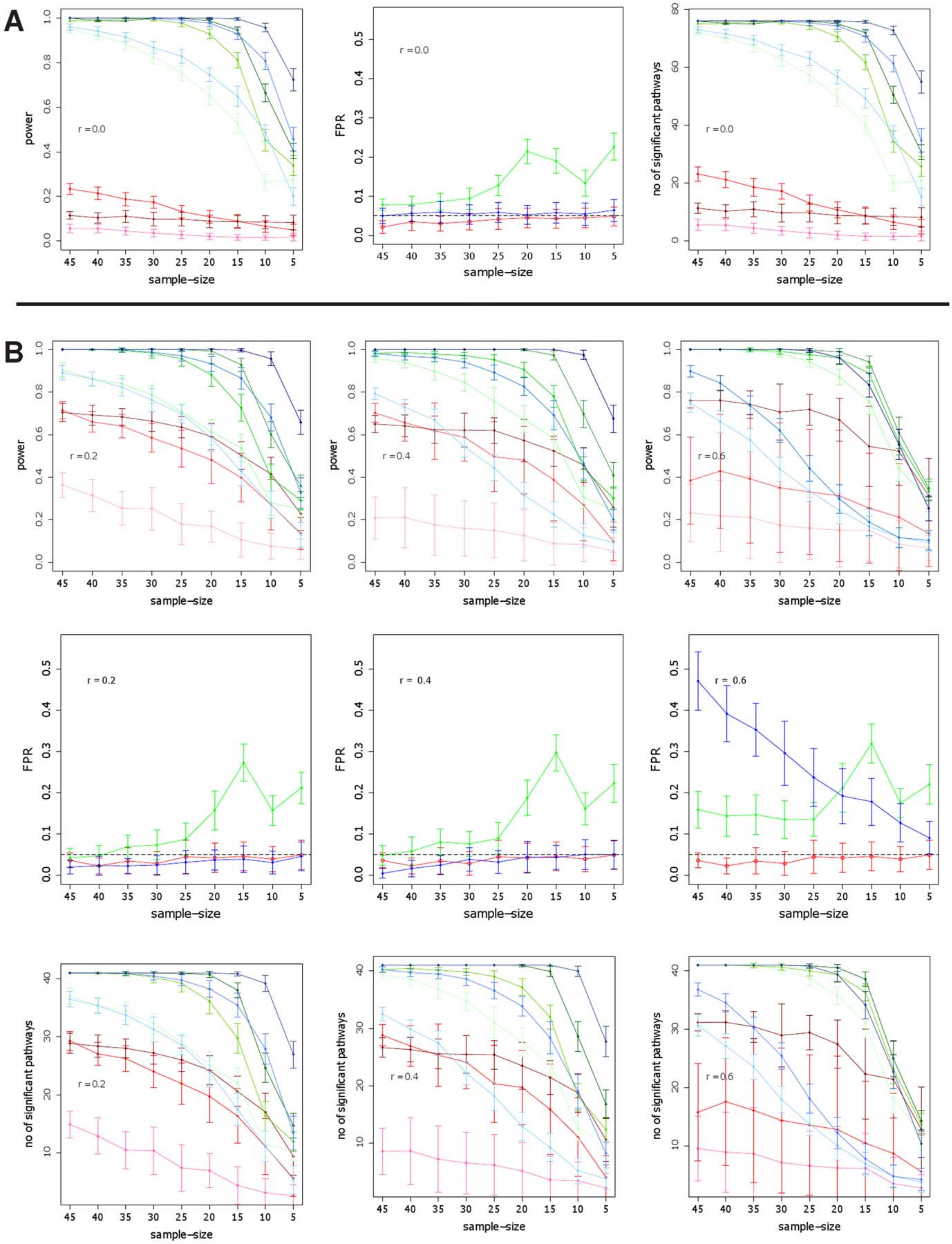

Figure 2. Simulation type III (A) and IV (B): Power, FPR and number of significant pathways for GSEA (red), *sum of t-square* (blue) and Hotelling's T^2 (green). DC = 10% (light color), 30% (medium color), 60% (dark color). Simulated data are from the protein network of yeast [49].
doi:10.1371/journal.pone.0037510.g002

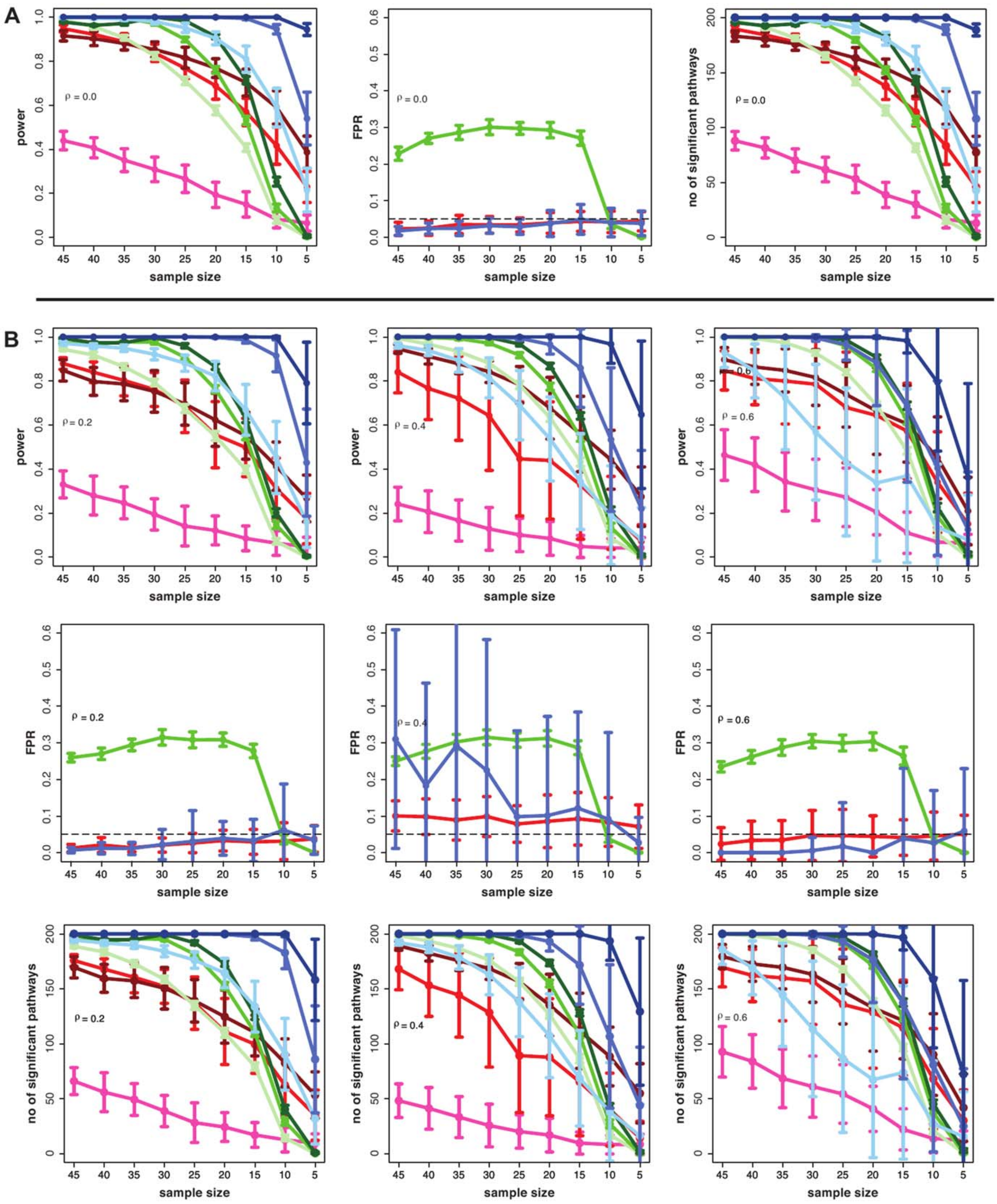

Figure 3. Simulation type III (A) and IV (B) : Power, FPR and *number of significant pathways* for GSEA (red), *sum of t-square* (blue) and Hotelling's T^2 test (green). DC = 10% (light color), 30% (medium color), 60% (dark color). Simulated data are from the transcriptional regulatory network of yeast [50].
doi:10.1371/journal.pone.0037510.g003

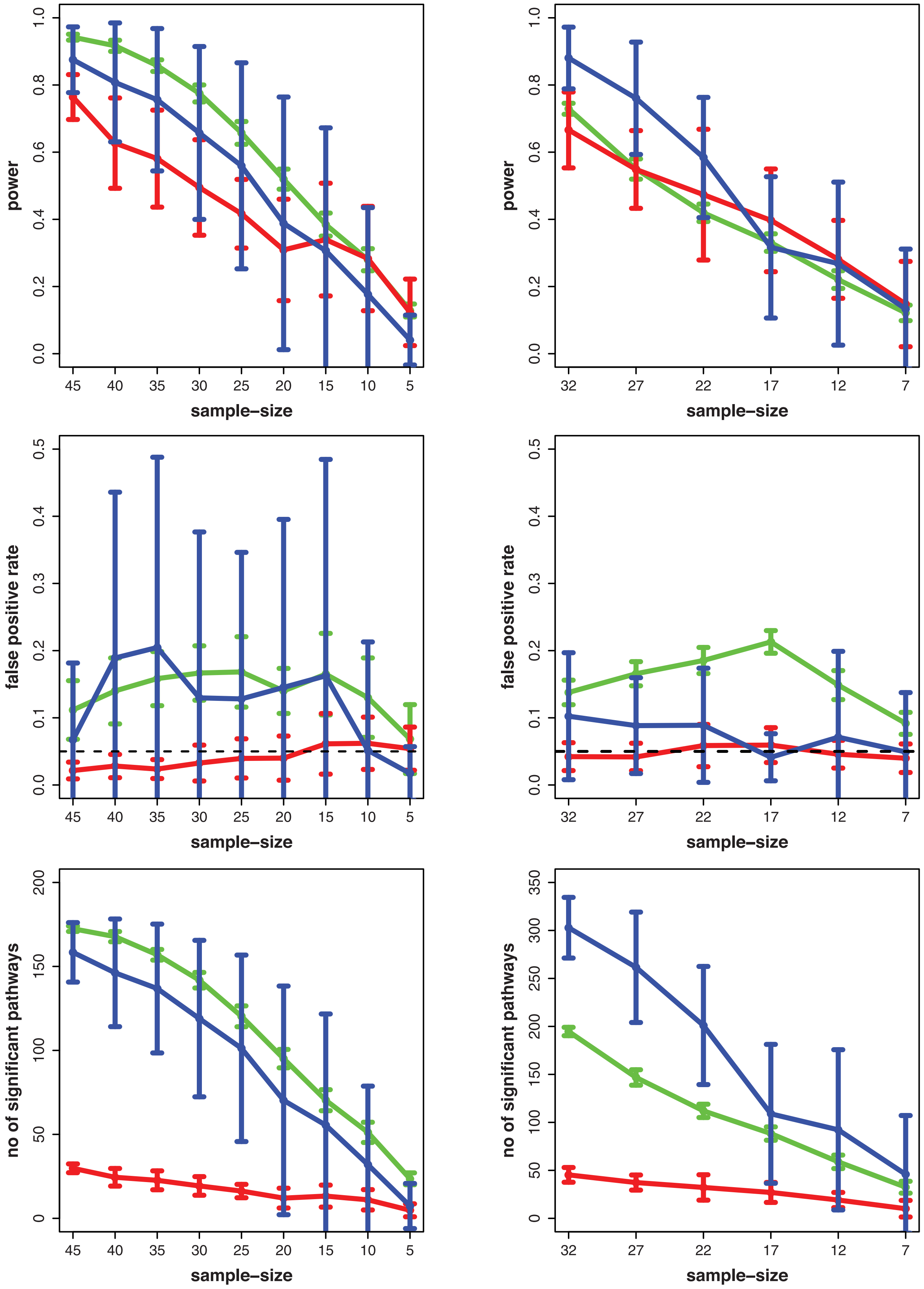
power
1.0
0.8
0.6
0.4
0.2
0.0
45
40
35
30
25
20
15
10
5
32
27
22
17
12
7
sample–size
false positive rate
0.5
0.4
0.3
0.2
0.1
0.0
no of significant pathways
200
150
100
50
0
350
300
250

Figure 4. Left column: prostate cancer. Right column: ALL. Power, false positive rate and number of significant pathways for GSEA (red), *sum of t-square* (blue) and Hotelling's T^2 (green).
doi:10.1371/journal.pone.0037510.g004

For comparison with the simulated data, we include in Fig. 6 the correlation values for ST I to IV for different parameters. The vertical intervals on the right-hand side correspond to the projected correlation range of these four simulation types. Ordered from left to right: ST I (green: $r=\{0.0,0.2,0.4,0.6\}$), ST II (orange: $r=\{0.2,0.4,0.6\}$), ST III (purple: $r=\{0.0\}$) and ST IV (brown: $r=\{0.2,0.4,0.6\}$). Here, we represent only the projected correlation range to simplify the presentation in Fig. 6. However, we would like to note that for each of these individual results the ordering of the correlation values assumes a similar shape as observed for prostate cancer and ALL. From these intervals, two observations are important to emphasize. First, ST I and II result in a shorter range for individual simulations, compared to ST III and IV. Second, the intervals for I and II are non-overlapping. Overall, we observe that if ST III and IV are employed together, the whole range of experimentally observed correlations found within normalized microarray data can be covered without gaps. Further, in comparison to the correlation structures used in previous studies, as discussed in the introduction, we find that ST III and IV provide a more realistic correlation structure compared to studies using a constant [13,29], random [30], autoregressive [31], compound symmetry [14,32,33] or no correlations at all [25,28].

We would like to point out that results about the range of the correlation values is only one indicator that should be met by simulated data. In addition, the structure among the genes is another important characteristics. Due to the fact that the data for ST III and ST IV are generated in a way that the inverse of the covariance matrix does reflect the independence relations that can be found in a protein network, this is another crucial difference to previous studies equipping our approach with a more realistic correlation structure.

Finally, we present results about the biological distribution of DC values in prostate cancer and ALL. Using SAM [57] and a multiple hypotheses correction [58] we identify differentially expressed genes in prostate cancer and ALL for $FDR=0.05$. From this, we estimate a mean DC value of 18% for prostate cancer and 4.5% for ALL; see Fig. 7 for the distributions. Further, we find that only very few pathways have a DC value larger than 30%, namely, 30 out of 213 pathways (corresponding to 14%) in prostate cancer and 2 out of 533 pathways (corresponding to 0.3%) in ALL. This provides evidence that the selected DC values for our simulations correspond to biologically relevant values.

Discussion

For our power analysis of simulated data, we assumed a maximum sample size of 50 because this corresponds well to the number of samples available for the experimental microarray data we used. Further, most other microarray experiments conducted provide usually less than 50 samples per condition making our choice from a biological point of view reasonable. Our results reveal the following. The *sum of t-square* test has for almost all studied cases the highest power if $DC \geq 30\%$, except for ST IV and $r=0.6$. However, if $DC<30\%$ and $r>0.2$ Hotelling's T^2 has a higher power. Due to the fact that this reflects the characteristics of the microarray data better, Hotelling's T^2 seems to be the favorable test. The *sum of t-square* test controls in general the FPR well, except for ST IV and $r=0.6$. The control of the FPR of Hotelling's T^2 depends strongly on the pathway size, and a control is only working for pathway sizes less than 35. GSEA is for almost all studied cases underpowered, except for ST I and $DC=60\%$ which is a condition that has not been found in one pathway in both microarray data sets. In our experience, this problem cannot be solved by increasing the sample size but is caused by *inappropriate* background data, which is out of the control of the experimenter. On the other hand, GSEA has a good control of the FPR for all conditions.

Taking all this into account our findings do not suggest to apply a method unconditionally to all pathways in a given data set, but to *filter* them in order to eliminate conditions for which a method is more likely to cause problems. We suggest to filter the pathways according to the following easy to check criteria: Hotelling's T^2 should only be applied to pathways with less than 35 genes and a sample size larger than 30. The *sum of t-square* test should only be used for pathways with $DC>10\%$ and a sample size of 25 or larger. GSEA should only be used for pathways with $DC>10\%$ and a sample size larger than 25. We want to emphasize that these sample sizes are different to the minimal sample sizes discussed in section 'Surrogate data: ALL and prostate cancer', which consider only the control of the FPR, whereas the optimal sample sizes avoid in addition that a study is underpowered. It is interesting to

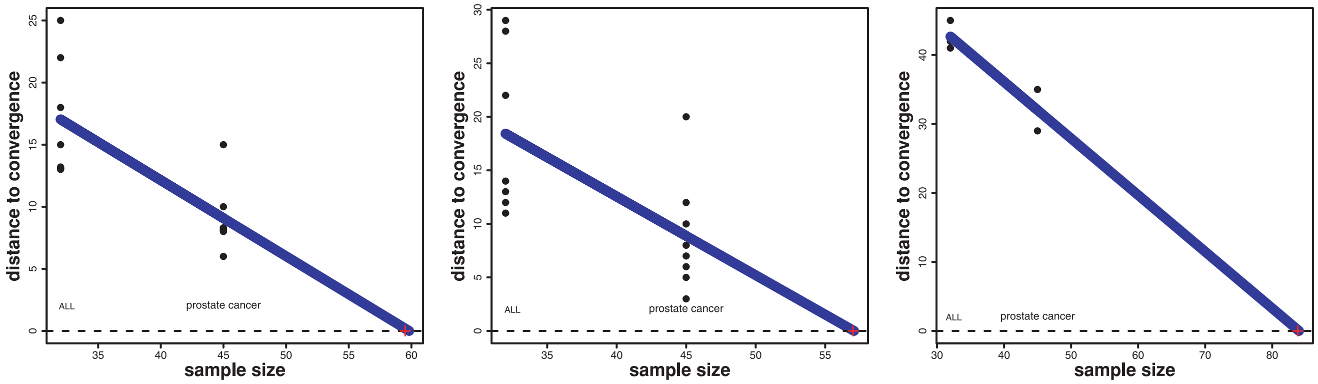

Figure 5. Left: Hotelling's T^2, Middle: *sum of t-square*, Right: GSEA. The regression line is used to predict the *optimal* sample size (red cross) found from the intersection of the regression line with the horizontal dashed line corresponding to a 'zero distance to convergence'.
doi:10.1371/journal.pone.0037510.g005

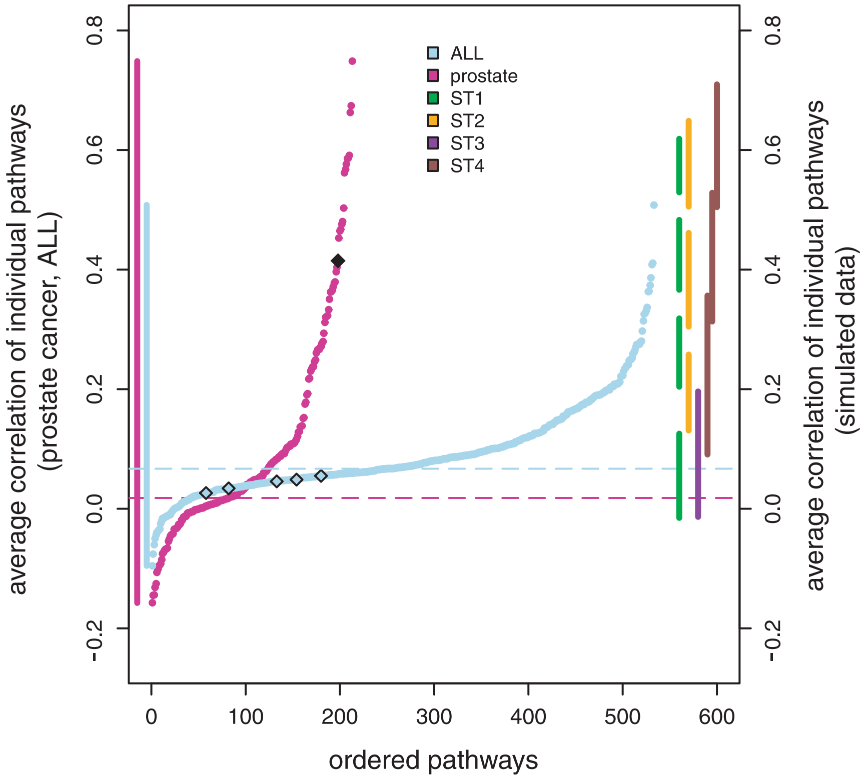

Figure 6. Average correlations for individual pathways for ALL (blue) and prostate cancer (violet) are shown by horizontally dashed lines. The two curves correspond to the rank ordered correlation values for ALL (blue) and prostate cancer (violet). For ST I (green - $\rho \in \{0.0,0.2,0.40.6\}$), ST II (orange - $\rho \in \{0.2,0.4,0.6\}$), ST III (purple, $\rho = 0.0$) and ST IV (brown - $\rho \in \{0.2,0.4,0.6\}$) the projections of the range of correlation values is shown on the right-hand side.
doi:10.1371/journal.pone.0037510.g006

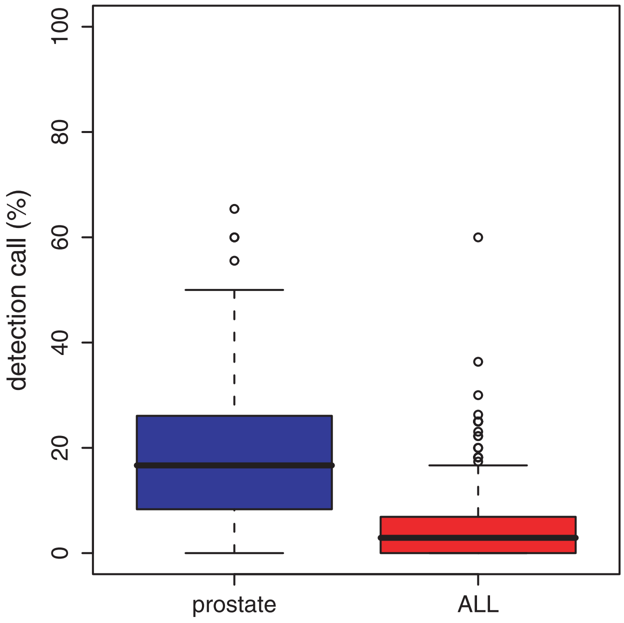

Figure 7. Distribution of the detection call (DC) values for gene expression data from prostate cancer (left) and ALL (right).
doi:10.1371/journal.pone.0037510.g007

note that in [59] a similar sample size recommendation has been given, however, for the stability of clusters obtained from clustering algorithms. Despite the methodological differences among these studies, this correspondence is interesting because it emphasizes that there is a considerable difference between studies for detecting the differentially expression of (single) genes and studies for identifying differentially expressed pathways. For the former it would be plausible to expect lower sample size recommendations than for a clustering analysis trying to estimate the correlation strength among genes. However, due to the fact that our sample size recommendations for DEP methods *coincide* with clustering methods, hints, that DEP analysis methods are not *just* the sum of individual gene test statistics. If one perceives this problem from a biological perspective, this correspondence becomes more plausible because the clustering of genes is frequently used to reveal functional relations between genes corresponding to biological pathways [60–62]. Hence, clustering algorithms and pathway methods respond to similar molecular functional units.

The underlying rationale for our power analysis, which provides statistical estimates of the true positive rates of tests, is to study the unbiased performance of the methods. In contrast, by conducting multiple hypotheses tests one would be obligated to apply a multiple testing procedure to control a selected error measure [58,63]. However, this would introduce a bias in the obtained results because both, the selected error measure and the control procedure, effect the results. In order to minimize this influence (it is probably not possible to completely eliminate this influence) one would need to study which pathway-based method works best together with which error measure and control procedure. However, these technical adjustments would not contribute to a better understanding of the power of a pathway-based method itself.

The optimal experimental design for microarray experiments with respect to the identification of the DEP is an important topic that is currently still under debate. From our comprehensive analysis of simulated and experimental gene expression data of over 3 million pathways, we obtain three major results. First, we find that the heterogeneity of different biological conditions and the sensitivity of the statistical methods suggest a selective application to definite pathways. That means, it is not advisable to apply a method to all accessible pathways but only to selected ones. Second, future gene expression experiments aiming to detect the DEP should be conducted with an increased number of samples in order to avoid non-robust and underpowered studies. From our study, we find method-specific recommendations constituting lower bounds for minimal sample sizes. Specifically, we suggest sample sizes between 60 and 85 to avoid (1) an underpowered study and (2) to allow the control of the FPR. Third, as a more theoretical finding we gained insight into the correlation structure of biological and simulated microarray data. From these results, we suggest the combined usage of ST III and ST IV for simulating gene expression data. Because these simulation types lead to a more realistic correlation structure compared to studies employing a constant, a random or no correlation structure at all. On a side note, we would like to remark that by using simulation methods like *GeneNetWeaver* [64] or *SynTReN* [65], which are aiming to mimic the mechanistic behavior of the transcription regulation of genes, it is also possible to obtain simulated expression data with a realistic correlation structure. However, the generation of data from sampling is simpler and usually less time consuming. Further, the controlled, concerted modification of expression levels of genes in particular pathways may be very challenging for such methods.

For future studies of DEP methods, simulations based on our approach using ST III and ST IV can be very useful to investigate, e.g., the influence of different gene network structures, the effect of overlapping pathways or the influence of heterogeneous effect sizes. For example, one could compare protein networks and transcriptional regulatory networks for different organisms or compare them with gene regulatory networks. Here by gene regulatory networks we mean networks inferred from gene expression data [66]. Also different gene regulatory networks inferred from different inference methods [67–70] could be studied to investigate distinctions on the pathway level. Regarding overlapping pathways and their potential importance for pathway methods, such simulation settings provide ample opportunity to control parameters for testing hypotheses about their influence. Lastly, for our study we used a constant effect size for the differentially expression of genes. That means, we sampled differentially expressed genes for the control group from $N(\mu,1)$ with $\mu = const$. This is similar to all previous studies we are aware of, e.g., [13,25]. However, it could be intricate to identify a distribution from which the mean μ should be sampled.

The studied methods in this paper are expected to be also useful for the analysis of RNA-seq data [71,72]. For this reason, once a sufficiently large data set is available, it would be interesting to repeat the above investigations for this new data type in order to gain a deeper insight into their experimental design in the context of DEP. Another important future direction to explore would be an investigation of the influence that alterations in regulation mechanisms in pathways have on the biological function [8].

Acknowledgments

We would like to thank Galina Glazko and Ricardo de Matos Simoes for fruitful discussions and helpful comments on the manuscript. For our simulations we used [73].

Author Contributions

Conceived and designed the experiments: FES. Performed the experiments: ST FES. Analyzed the data: ST FES. Contributed reagents/materials/analysis tools: ST FES. Wrote the paper: ST FES.

References

1. Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits. Boca Raton, FL: Chapman & Hall/CRC.
2. Emmert-Streib F, Dehmer M, eds. Medical Biostatistics for Complex Diseases. Weinheim: Wiley-Blackwell.
3. Kauffman S (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology 22: 437–467.
4. Niiranen S, Ribeiro A, eds. Information Processing and Biological Systems. Berlin: Springer.
5. Callow M, Dudoit S, Gong E, Speed T, Rubin E (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res 10: 2022–9.
6. Chen Y, Dougherty ER, Bittner ML (1997) Ratio-based decisions and the quantitative analysis of cdna microarray images. Journal Of Biomedical Optics 2: 364–374.
7. Storey J, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–5.
8. Emmert-Streib F (2007) The chronic fatigue syndrome: A comparative pathway analysis. Journal of Computational Biology 14: 961–972.
9. Kim SY, Volsky D (2005) Page: Parametric analysis of gene set enrichment. BMC Bioinformatics 6: 144.
10. Nettleton D, Recknor J, Reecy JM (2008) Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics 24: 192–201.
11. Mootha V, Lindgren C, Eriksson K, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34: 267–273.
12. Emmert-Streib F, Glazko G (2011) Network Biology: A direct approach to study biological function. Wiley Interdiscip Rev Syst Biol Med 3: 379–391.
13. Ackermann M, Strimmer K (2009) A general modular framework for gene set enrichment analysis. BMC Bioinformatics 10: 47.
14. Hummel M, Meister R, Mansmann U (2008) GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics 24: 78–85.
15. Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y (2007) A multivariate extension of the gene set enrichment analysis. J Bioinform Comput Biol 5: 1139–1153.
16. Xiong H (2006) Non-linear tests for identifying differentially expressed genes or genetic networks. Bioinformatics 22: 919–23.
17. Goeman J, Buhlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23: 980–7.
18. Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–29.
19. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopia of Genes and Genomes. Nuclei Acids Res 28: 27–30.
20. Emmert-Streib F, Glazko G (2011) Pathway analysis of expression data: deciphering functional building blocks of complex diseases. PLoS Computational Biology 7: e1002053.
21. Lu Y, Liu P, Xiao P, Deng H (2005) Hotelling's T 2 multivariate profiling for detecting differential expression in microarrays. Bioinformatics 21: 3105–3113.
22. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, et al. (2005) Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Sciences of the United States of America 102: 13544–13549.
23. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–50.
24. Glazko G, Emmert-Streib F (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics 25: 2348–54.
25. Efron B, Tibshiran R (2007) On testing the significance of sets of genes. Annals of Applied Statistics 1: 107–129.
26. Jiang Z, Gentleman R (2007) Extensions to gene set enrichment. Bioinformatics 23: 306–313.

27. Nam D, Kim S (2008) Gene-set approach for expression pattern analysis. Brief Bioinform 9: 189–197.
28. Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza T, et al. (2009) Comparative study of gene set enrichment methods. BMC Bioinformatics 10: 275.
29. Qiu X, Xiao Y, Gordon A, Yakovlev A (2006) Assessing stability of gene selection in microarray data analysis. BMC Bioinformatics 7.
30. Choi Y, Kendziorski C (2009) Statistical methods for gene set co-expression analysis. Bioinformatics 25: 2780–2786.
31. Jung K, Becker B, Brunner E, Beißbarth T (2011) Comparison of global tests for functional gene sets in two-group designs and selection of potentially effect-causing genes. Bioinformatics 27: 1377–1383.
32. Liu Q, Dinu I, Adewale A, Potter J, Yasui Y (2007) Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 8: 431.
33. Tsai C, Chen J (2009) Multivariate analysis of variance test for gene set analysis. Bioinformatics 25: 897–903.
34. Irizarry RA, Wang C, Zhou Y, Speed TP (2009) Gene set enrichment analysis made simple. Statistical Methods in Medical Research 18: 565–575.
35. Luo W, Friedman M, Shedden K, Hankenson K, Woolf P (2009) Gage: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 10: 161.
36. Newton M, Quintana F, den Boon Jea (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Annals of Applied Statistics 1: 85–106.
37. Castelo R (2006) A robust procedure for gaussian graphical model search from microarray data with p larger than n. Journal of Machine Learning Research 7: 2621–2650.
38. Kim KI, van deWiel M (2008) Effects of dependence in high-dimensional multiple testing problems. BMC Bioinformatics 9: 114.
39. Chiaretti S, Li X, Gentleman R, Vitale A, Wang KS, et al. (2005) Gene Expression Profiles of Blineage Adult Acute Lymphocytic Leukemia Reveal Genetic Patterns that Identify Lineage Derivation and Distinct. Mechanisms of Transformation 11: 7209–7219.
40. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1: 203–209.
41. Davison A, Hinkley D (1997) Bootstrap Methods and Their Application Cambridge University Press.
42. Efron B, Tibshirani R (1994) An Introduction to the Bootstrap. New York: Chapman and Hall/CRC.
43. Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Finance 10: 603–621.
44. Ledoit O, Wolf M (2004) A well conditioned estimator for largedimensional covariance matrices. J Multiv Anal 88: 365–411.
45. Ledoit O, Wolf M (2004) Honey, i shrunk the sample covariance matrix. J Portfolio Management 30: 110–119.
46. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4: 32.
47. Whittaker J (1990) Graphical Models in Applied Multivariate Statistics. Chichester: John Wiley & Sons.
48. Castelo R, Roverato A (2009) Reverse engineering molecular regulatory networks from microarray data with qp-graphs. Journal of Computational Biology 16: 213–27.
49. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The BioGRID Interaction Database: 2008 update. Nucl Acids Res 36: D637–640.
50. Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. Journal of Molecular Biology 360: 213–227.
51. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 206: 509–512.
52. Li H, Gui J (2006) Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. Biostatistics 7: 302–317.
53. Werhli A, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22: 2523–31.
54. Wille A, Zimmermann P, Vranova E, Furholz A, Laule O, et al. (2004) Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. Genome Biology 5: R92.
55. Lehman E (2005) Testing Statistical Hypotheses Springer.
56. Albert R (2005) Scale-free networks in cell biology. Journal of Cell Science 118: 4947–4957.
57. Tusher V, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116–21.
58. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57: 125–133.
59. Garge N, Page G, Sprague A, Gorman B, Allison D (2005) Reproducible Clusters from Microarray Research: Whither? BMC Bioinformatics 6: S10.
60. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. PNAS 95: 14863–14868.
61. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC (2006) Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics 22: 2405–2412.
62. Quackenbush J (2006) Microarray analysis and tumor classification. N Engl J Med 345: 2463–72.
63. Dudoit S, van der Laan M (2007) Multiple Testing Procedures with Applications to Genomics. New York; London: Springer.
64. Schaffter T, Marbach D, Floreano D (2011) GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27: 2263–70.
65. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, et al. (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics 7: 43.
66. Emmert-Streib F, Glazko G, Altay G, de Matos Simoes R (2012) Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in Genetics 3: 8.
67. Altay G, Emmert-Streib F (2011) Structural Influence of gene networks on their inference: Analysis of C3NET. Biology Direct 6: 31.
68. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. PLoS Biol 5.
69. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7: S7.
70. Meyer P, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. BMC Bioinformatics 9: 461.
71. Marguerat S, Bähler J (2010) RNA-seq: from technology to biology. Cellular and Molecular Life Sciences 67: 569–579.
72. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.
73. R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.