BIOLOGY DIRECT

REVIEW Open Access

Harnessing the complexity of gene expression data from cancer: from single gene to structural pathway methods

Frank Emmert-Streib*, Shailesh Tripathi and Ricardo de Matos Simoes

Abstract

High-dimensional gene expression data provide a rich source of information because they capture the expression level of genes in dynamic states that reflect the biological functioning of a cell. For this reason, such data are suitable to reveal systems related properties inside a cell, e.g., in order to elucidate molecular mechanisms of complex diseases like breast or prostate cancer. However, this is not only strongly dependent on the sample size and the correlation structure of a data set, but also on the statistical hypotheses tested. Many different approaches have been developed over the years to analyze gene expression data to (I) identify changes in single genes, (II) identify changes in gene sets or pathways, and (III) identify changes in the correlation structure in pathways. In this paper, we review statistical methods for all three types of approaches, including subtypes, in the context of cancer data and provide links to software implementations and tools and address also the general problem of multiple hypotheses testing. Further, we provide recommendations for the selection of such analysis methods.

Reviewers: This article was reviewed by Arcady Mushegian, Byung-Soo Kim and Joel Bader.

Keywords: Gene expression data, Cancer data, Statistical analysis methods, Pathway methods, Correlation structure, Cancer genomics

Review

Background

The early driving forces in biology were reductionist approaches. In general, a reductionist approach tries to break-down a complex system into its parts list and explains its properties as the sum of its individual components. Hence, the individual constituents of a system inform its higher level functions [1-4]. However, the 'one gene, one protein, one function' working hypothesis [5] is not sufficient in order to explain the many emergent properties such as the phenotypic variability of organisms or the heterogeneity of cancer [6]. For this reason, nowadays, it is generally acknowledged that for achieving a functional understanding of biological systems, the genes in a cell need to be studied as a functioning collective [2,3,7]. In such a system, the collective functioning of groups of genes results in, for instance, signaling pathways or protein complexes that regulate cell differentiation, transcription regulation or growth.

A systems integration at the cellular level has the potential to answer many, until now, unsolved questions about biological systems and their collective functioning, regulatory programs for growth, development, phenotypic variability and the causality of many complex diseases [8-10]. Due to the enormous complexity of a cellular system, where many processes and interactions at different levels inside a cell work in harmony to assure the vital functioning of a cell, we need to understand key properties of biological systems like its robustness or modularity [2,8] in order to enhance our understanding of complex diseases. These complex interactions occurring within a cell can be described by networks [11-13], including gene regulatory networks [14,15], protein-protein interaction (PPI) networks [16,17], metabolic networks [18] and transcription regulatory networks [19,20]. The networks are organized at different cellular levels and enable the functionality of the cell. The question now arising is

*Correspondence: v@bio-complexity.com
Computational Biology and Machine Learning Laboratory, Queen's University Belfast, Belfast, UK

how can the complexity inside a cell be understood, and analyzed?

The development of information processing technologies in the post genomic era enabled the generation of huge amounts of data. In this review, we focus on gene expression data from microarray platforms and summarize three major types of analysis strategies: (I) Identification of changes in single genes, (II) identification of changes in gene sets or pathways, and (III) identification of changes in the correlation structure within pathways. We discuss these methods in the context of cancer data sets to emphasize their biological meaning, implications and expressiveness.

Large-scale gene expression data

In the next section, we briefly review high-throughput technologies that enable the generation of large-scale gene expression data [21-23].

Gene expression data from microarray

A microarray experiment measures genome-wide gene expression levels of mRNA in a cell or a tissue sample under a particular condition. A microarray chip quantifies the hybridization of fluorecsent labeled target nucleotide sequences to defined complementary probe sequences that are spotted on a glass or silicon slide. For different microarray platforms the spotted probes are synthetic oligonucleotides ranging from 25 to 80 nucleotides or long cDNA transcripts. Different microarray platforms were designed for a single-channel or a multi-channel experimental setting. For single-channel arrays each condition sample is hybridized separately on individual arrays using a single dye. For multi-channel arrays multiple conditions are hybridized together on individual arrays using multiple dyes. For example Affymetrix is a single-channel platform, where multiple oligonucleotide probes (probeset) of 25 bases are used to measure the concentration of a mRNA transcript. The target mRNAs of expressed genes are extracted from a treatment or a control sample, reverse transcribed to cDNAs, labeled with a fluorescent dye and then hybridized to a microarray. An image of the microarray captures laser induced emitted fluoresent intensities of the probes at each spot. The intensities give a proportional measure of the corresponding mRNA concentration for each gene that was defined on the microarray.

Gene expression data from next generation sequencing (RNA-seq)

The transcriptome of a cell comprises mRNA, tRNA, rRNA, and short regulatory RNAs. RNA-seq is a transcriptome sequencing approach that uses deep sequencing techniques such as 454 (Roche), genome analyzer (Illumina solexa), SOLiD (support oligonucleotide ligation detection), Polonator G.007, HeliScope (Helicos BioSciences) and SMRT (single molecule real time sequencing) [24].

RNA-seq has a wide variety of applications such as the measurement of gene expression levels from transcribed mRNA sequences [25]. In the first step of the procedure RNA is extracted from a given condition sample, fragmented, reverse transcribed to cDNA that is ligated to adapters. In the second step a library of reads is generated from the ligated fragments that are sequenced. In the third step the reads are mapped to known exon sequences of genes. The expression level of a gene is measured from the normalized number of mapped sequences that mapped to the known set of exon sequences of a gene. The RNA-seq transcriptome sequencing approach overcomes several limitations of microarrays for measuring gene expression. For example, RNA-seq measures large ranges of expression levels from very low to highly expressed genes and is able to consider unknown transcribed sequences. Since the novelty of the methodology, gold standard procedures for the management and processing of the data are currently being established.

Gene expression data and cancer

Cancer is a multifactorial disease, i.e., the detection of one mutation in one gene cannot explain the phenotypic plurality of carcino- and pathogenesis by a one-to-one relationship between genotype and phenotype. Instead, cancer can be induced by a multitude of genetic and environmental factors and the accumulation of such events. The intervening of such complex factors makes in general the characterization of complex diseases difficult. For this reason it is astonishing that the seminal work by Weinberg et al. [6,26] presented a relative simple, systematic functional framework for cancer and the role different biological key processes are playing. In this paper, the so called *hallmarks of cancer* have been defined. According to [6], the hallmarks of cancer (see Figure 1) are:

- self sufficiency in growth signals
- insensitivity to anti-growth signals
- evading apoptosis
- limitless replicative potential
- sustained angiogenesis
- tissue invasion and metastasis

Later, this list has been extended by adding two further hallmarks [26]

- deregulating cellular energetics
- avoiding immune destruction

and also two enabling characteristics

- genome instability
- mutation and tumor-promoting inflammation

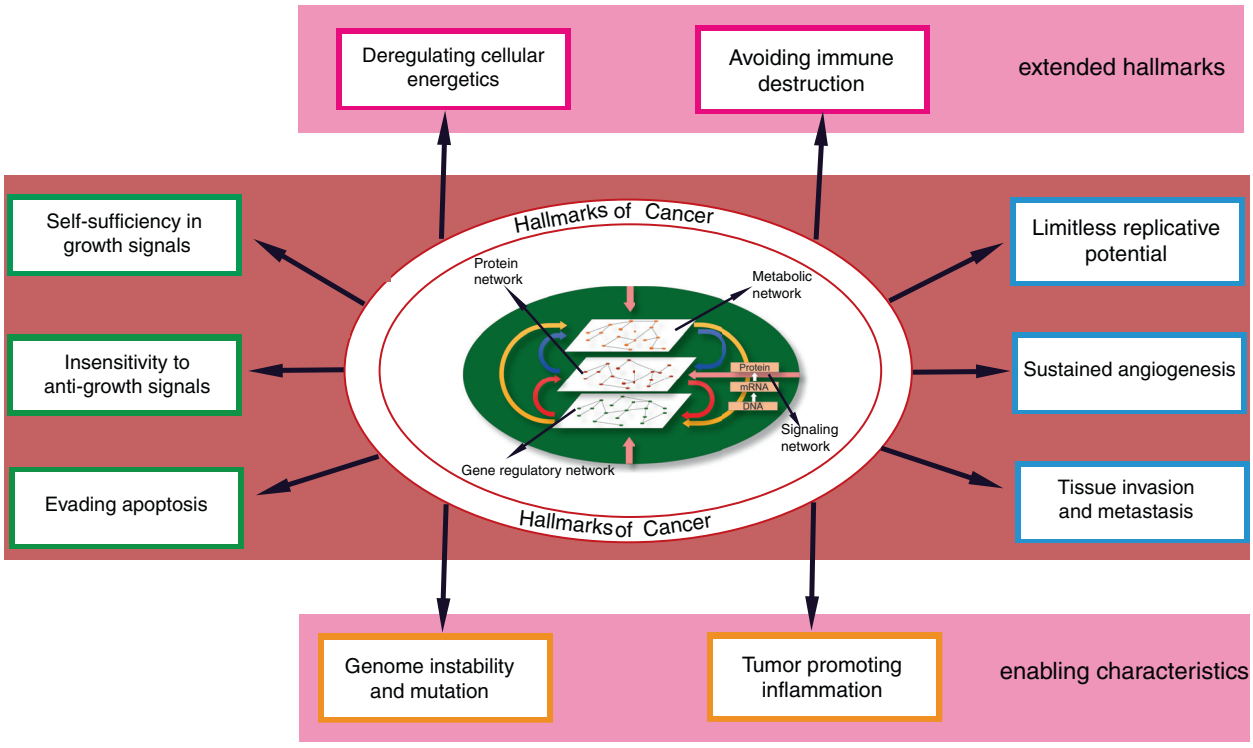

Figure 1 The hallmarks of cancer and enabling characteristics [6,26].

It has been recognized that these hallmarks are gradually acquired by different types of cancers, potentially, in a variable order. This variability in the acquiring of these disease-bearing processes is one of the indicators of the complexity of cancer.

The biological processes in a cell are controlled and regulated by signaling pathways that are activated by internal and external signaling receptors and factors. The signaling pathways governing growth and cell proliferation are likely dysregulated in their functioning in cancer. For example, they become insensitive to anti-growth signals, or they are dys-regulated in growth signaling pathways by gaining autonomy in their growth. It is assumed that interaction changes at various levels (genetic, mRNA or protein) lead to the unlimited growth of cells instead of the up-regulation or down-regulation of a single gene. Further, sometimes, even a moderate change in the expression of a group of genes can lead to a significant change in the biological function of an organism [27].

Currently, the underlying processes that contribute to cancer are being intensively investigated. However, so far, the molecular causes that initiate and maintain cancer are not well understood. For this reason, the understanding of gene expression profiles, which provide signatures of all the active genes and their interconnections in a cell, contain valuable information about the functioning of key pathways, as expressed by the hallmarks of cancer and, hence, enable a practical investigation of functional mechanisms thereof [28-36]. Despite the different focus of many studies of different cancer types, common themes in the form of 'key pathways' can be found throughout. For instance, the NF-κB pathway involved in the cellular responses to external stimuli like cytokines or free radicals, and immune response to infection [29,37-39]; the MAPK signaling pathway responsible for regulating growth factor signaling including the RAF, MEK, and MAPK cascade [34,39,40]; the p53 signaling pathways involved in DNA damage control, apoptosis and inhibition of angiogenesis [37,41,42]; or the Wnt signaling pathway involved in cell differentiation, and cell polarity [31,34,36,38].

Formulating biological hypotheses

A main goal of high-throughput gene expression analysis is to identify differentially expressed genes or gene sets between two or more conditions to enable a functional interpretation of the underlying condition-specific mechanisms. The biological processes at the gene level are complex in nature as they dynamically interact with each other. A single gene can participate in different biological processes and regulate different genes at different time points. The identification of key genes or pathways is a difficult task, because their interactions are unknown. We only observe the phenotypic outcome of test conditions

and the corresponding gene expression patterns measured from a tissue or cell culture. Univariate and multivariate statistical methods can be applied in order to understand such differences from a statistical perspective. The first type of approach that has been used to identify changes in the gene expression is a differential gene expression analysis. This approach is commonly used to compare different conditions of microarray samples to identify differences between them. As a result, a single gene analysis approach gives a list of genes that show a statistically significant difference between two conditions. For cancer, such genes may correspond to oncogenes or tumor suppressor genes.

If we consider the underlying network where different biological functions are being described by groups of interacting genes, a single gene analysis does not resolve the biological functions that are affected primarily in disease conditions and are causal factors of the disease. In order to get a systematic understanding of the disease or phenotypes we have to first understand what biological functions contribute to these changes, and perform a comparison between conditions using *groups of genes* defined by biological pathways. This approach leads to comparing gene expression data at the pathway level where sets of genes are tested for differential expression.

Another interesting property that can be extracted from gene expression data is the correlation structure of gene expression profiles between all genes. This correlation structure shows associations between genes which directly or indirectly interact with each other [8,43-45]. Comparative analyses of gene pathways that consider the correlation structure of expression data can provide a suitable test for the hypothesis of changes in the underlying network.

In summary a gene expression data set can be used to (I) identify differentially altered single genes, (II) identify differentially expressed gene sets or pathways, and (III) identify differentially correlated pathways. In the following sections, we review statistical methods that have been introduced to study the three problems (I-III) above. In Figure 2 we give a graphical overview of the such methods.

Before we procede, we would like to point out that all of these methods test statistical hypotheses [46]. That implies that in order to understand a particular method biologically, i.e., one is capable of providing a biological interpretation, one needs to understand the underlying

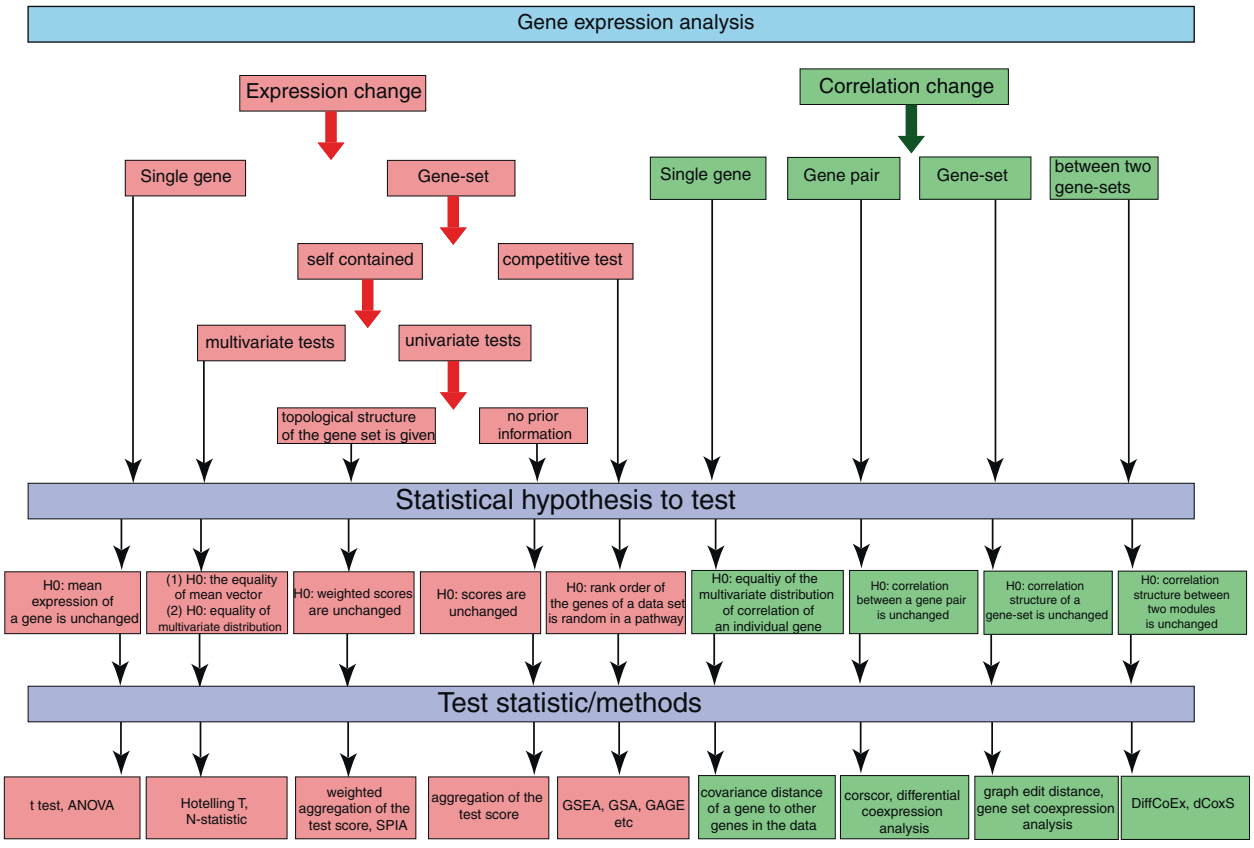

Figure 2 Classification of different univariate and multivariate hypothesis testing based methods for gene expression data analysis. The data are used to test differential expression and co-expression between two conditions for single gene and gene set based approaches.

null hypothesis. In our opinion, it is helpful to approximately categorize all statistical hypotheses into three categories with respect to their biological interpretability, whereas each category represents a different degree of difficulty to find a biological interpretation for a hypothesis. In the following, we provide a brief discussion of these three categories because it enables a better, potentially, more plausible understanding of the methods presented in the next sections.

In category one belong all hypotheses for which it is relatively easy to find a meaningful biological interpretation. An example from this category are tests that compare mean values (μ), e.g., to identify the differential expression of genes (section 'Differential expression of a gene'). That means these tests use the mean as a *test statistic*. Due to the fact that the underlying (probability) distribution of the genes represents, biologically, the activity of the gene expressions, the interpretation of a null hypothesis is directly derived thereof. For this reason the biological interpretation of the rejection of the null hypothesis given in Eqn. 2, is intuitively clear and appealing, because it implies a change in the (mean) expression of genes which may indicate a change in a biological function because the number of available proteins may be altered.

In category two fall tests for which there are several alternative biological interpretations. This makes the interpretation of such tests ambivalent from a biological perspective. As example for such a test, we consider the detection of the differential variance of a gene (section 'Differential variance of a gene'). Despite the fact that the underlying probability distribution of the expression of genes has a clear biological interpretation, the biological interpretation for the rejection of the null hypothesis in Eqn. 4 is not unique. For instance, a gene could have a different variance in two conditions because, e.g., in condition one it is periodically expressed, whereas in condition two it is constantly expressed on an intermediate level. The former condition may be related to the cell cycle or the circadian rhythm, or periodically triggered by an external signaling factor that is released by the administration of a medication that is regularly taken. A second equally plausible interpretation could be that in one condition the cell utilizes parallel pathways to transfer a signal whereas in the other condition only one signaling chain is used. The reason for the utilization of parallel pathways could be triggered by stress factors, e.g., in the presence of an infection, so that the cell is 'running' full power in order to execute all necessary programs that have been initiated by the presence of the intruder.

Lastly, for tests in category three it is very difficult to find sufficiently precise biological interpretations because, statistically, these methods test 'complex' expressions. An example for a test from this category is the N-statistic (section 'N-statistic'). The null hypothesis is based on the comparison of two *distributions* rather than *scalar* test statistics. In order to clarify the crucial difference between the comparison of two *distributions* and *scalar* test statistics we note that, theoretically, every probability distribution can be written as a series expansion in its moments [47]. This means a test for a distribution, compares *implicitly* the moments of this distribution. Here the (k-th) moment is defined as the expectation value of a random variable (to power k), i.e.,

$$m_k = \mathbb{E}[\,x^k\,]\,. \tag{1}$$

An example for a moment is the mean (which is the first moment), other examples of entities that can be expressed as a function of moments are the variance and the kurtosis. This means whenever the null hypothesis in Eqn. 39 is rejected it could be because of a difference in *any* moment of which there are, theoretically, infinite many. Put differently, this kind of unspecificity makes this test very powerful in the sense that it may detect *any possible* difference two distributions can exhibit. On the other hand, if the null hypothesis is rejected it is very difficult to identify a precise reason for its rejection. For instance, this could be related to a difference in the mean, variance, kurtosis or any higher moment or function thereof. These combinatorial factors do usually not allow to find a concise biological interpretation. Nevertheless, such a test can be of valuable use, e.g., for diagnostic purposes.

Single-gene analysis

Single-gene based methods can be subdivided into three major classes. A) Methods for detecting differential gene expression, B) methods for detecting differential correlation, and C) methods for detecting a differential variance.

Differential expression of a gene

The analysis of differential gene expression is based on the mean expression change of individual genes. Suppose for a gene g_i in a microarray data set the mean expression value for two conditions are μ_1 and μ_2 respectively. Then the null hypothesis for the differential expression of the gene g_i is defined as

$$H_0 : \mu_1 = \mu_2 \tag{2}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{3}$$

A gene is called *differentially expressed* when H_0 is rejected. Figure 3A shows an example where the samples for two conditions are drawn from two normal distributions with different mean values, i.e., $N(\mu_1 = 0, \sigma_1 = 1)$ and $N(\mu_2 = 1, \sigma_2 = 1)$.

The first published studies for gene expression analysis selected differentially expressed genes based on a fold-change criteria between a treatment and control condition [48]. For example, an early application of this measure was

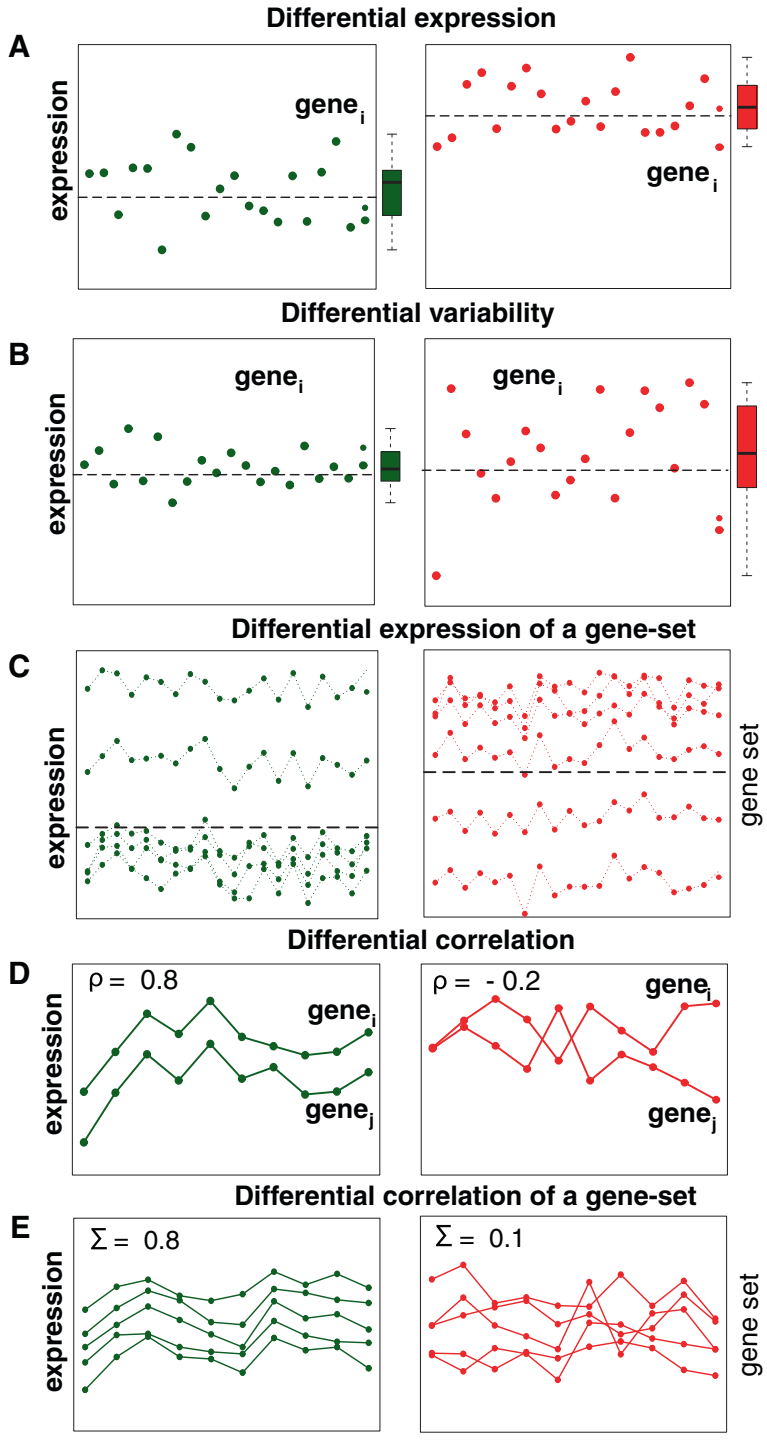

Figure 3 Examples for different biological hypothesis that can be formulated based on different properties of gene expression data (green - control group (condition 1), red - treatment group (condition 2)). E: Σ corresponds to the covariance matrix, corresponding to a constant but different matrix for the two conditions.

used to compare normal colon epithelium and primary colon cancers [49]. Since then, many statistical approaches have been developed to provide more robust measures. Among the most popular methods are, e.g., SAM [50,51], limma [52], and the empirical Bayes approach from Efron et al. [53].

Differential variance of a gene

The analysis of *differential variability* (DV) aims to detect a change in the variance of the gene expression values [54]. Suppose in a microarray expression data set, the mean expression value of a gene i is μ_c and its variance σ_c^2, for condition $c = \{1, 2\}$. Then, the null and alternative hypothesis tested are:

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{4}$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{5}$$

A gene is called *differentially variable* when the null hypothesis H_0 is rejected. The DV analysis in [54] tests H_0 by using a F-test. In Figure 3B we show an example for a gene with a constant mean, but a changed variance in the two conditions. The samples for the two conditions are drawn from a standard normal distribution with the same mean but different variances for the conditions, i.e., $N(\mu_1 = 0, \sigma_1 = 1)$, and $N(\mu_2 = 0, \sigma_2 = 2)$.

Differential correlation of a gene

The analysis of *differential correlation* aims to detect changes in the dependency structure of a single gene [55]. Suppose $r_i = r_{i1}, ..., r_{ip}$ denotes a $p - 1$ dimensional correlation vector, whereas each component corresponds to the correlation between gene i and one of the other $p - 1$ genes in a data set. Then for r_i one obtains distribution functions, denoted by $F_{r_i}^A$ and $F_{r_i}^B$, for condition A and B and the following hypotheses:

$$H_0 : F_{r_i}^A = F_{r_i}^B \tag{6}$$

$$H_1 : F_{r_i}^A \neq F_{r_i}^B \tag{7}$$

A gene i is called *differentially correlated* when H_0 is rejected.

Gene-pair analysis

The functional activities of genes, as measured by gene expression values, reflect the interplay of the genes and their products in the underlying gene network. The objective of a gene-pair analysis is to identify either differential co-correlated or differential co-expressed pairs of genes, instead of individual genes. The reason for looking for pairs of genes is that the concerted changes in genes is due to their common membership in biological pathways.

The principle idea to detect correlation changes in gene-pairs is visualized in Figure 3D. The data are sampled from a multivariate normal distribution with a constant mean vector for both conditions, $\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = (0, 1)$, but a different correlation of $\rho_1 = 0.8$ and $\rho_1 = -0.2$. The point is despite no difference in the mean expression of the gene-pair, there is a difference in their correlation.

In [56] a method (CorScor) has been proposed to identify such gene-pairs. In Figure 4 we show three cases of the joint distribution of expression values of two genes, for two conditions. In this Figure we are showing simulated data for three possible changes in the co-expression of a pair of genes in two conditions. The samples in Figure 4A and B are drawn from a multivariate normal distribution with $\boldsymbol{\mu_1} = \{5, 5\}$ and $\boldsymbol{\mu_2} = \{5, 7\}$. For Figure A the correlation between gene-pairs is $\rho_1 = \rho_2 = 0.9$ and for Figure B it is $\rho_1 = \rho_2 = -0.9$. For Figure 4C the samples are generated from a multivariate normal distribution with $\boldsymbol{\mu_1} = \{5, 5\}$ and $\boldsymbol{\mu_2} = \{5, 5\}$ and the average correlation between gene-pairs is $\rho_1 = 0.9$ and $\rho_2 = -0.3$.

In the first two cases (Figure 4A and B) the correlation of the gene-pairs show a condition specific shift, in [56] denoted as a *gap* and *substitution*. In the third case (Figure 4C), the gene-pairs show a reversed correlation between the two conditions, denoted as *on/off case*. To identify gene-pairs in these two types of conditions, two scoring functions have been suggested in [56] given by:

$$s = \begin{cases} |\rho_A + \rho_B - \alpha\rho| & \text{gap/substitution case} \\ |\rho_A - \rho_B| & \text{on/off case} \end{cases} \tag{8}$$

Here each of the three correlation coefficients are estimated for a gene-pair between gene i and j, i.e., $\rho_A = \rho_A(i, j)$ etc. The value of ρ corresponds to the global correlation coefficient of the gene-pair over the two conditions (A, B) and ρ_A, ρ_B are the correlation coefficients of the gene-pair for condition A and condition B. In Eqn. 8, α is a tuning parameter that governs the balance between separation and parallel alignment. In [56] it was argued to use a value of $\alpha = 1.5$. The null and alternative hypotheses tested are:

$$H_0 : s = 0 \tag{9}$$

$$H_1 : s \neq 0 \tag{10}$$

In [57] the 'expected conditional F-statistic' (ECF-statistic) has been introduced to measure the differential co-expression of gene pairs (X, Y). The method is based on a modified F-statistic, where the variance and the mean parameter of the statistic are estimated from a mixture of two normal distributions.

The R package R/EBcoexpress provides an empirical Bayesian implementation to identify the differential co-expression of gene pairs [58].

Another method called *liquid association* (LA) has been proposed in [59] to identify co-expressed gene pairs. In contrast to pairwise correlation measures, the LA method considers the presence of a mediator gene, Z, for observing a co-expression between two genes at a given cellular

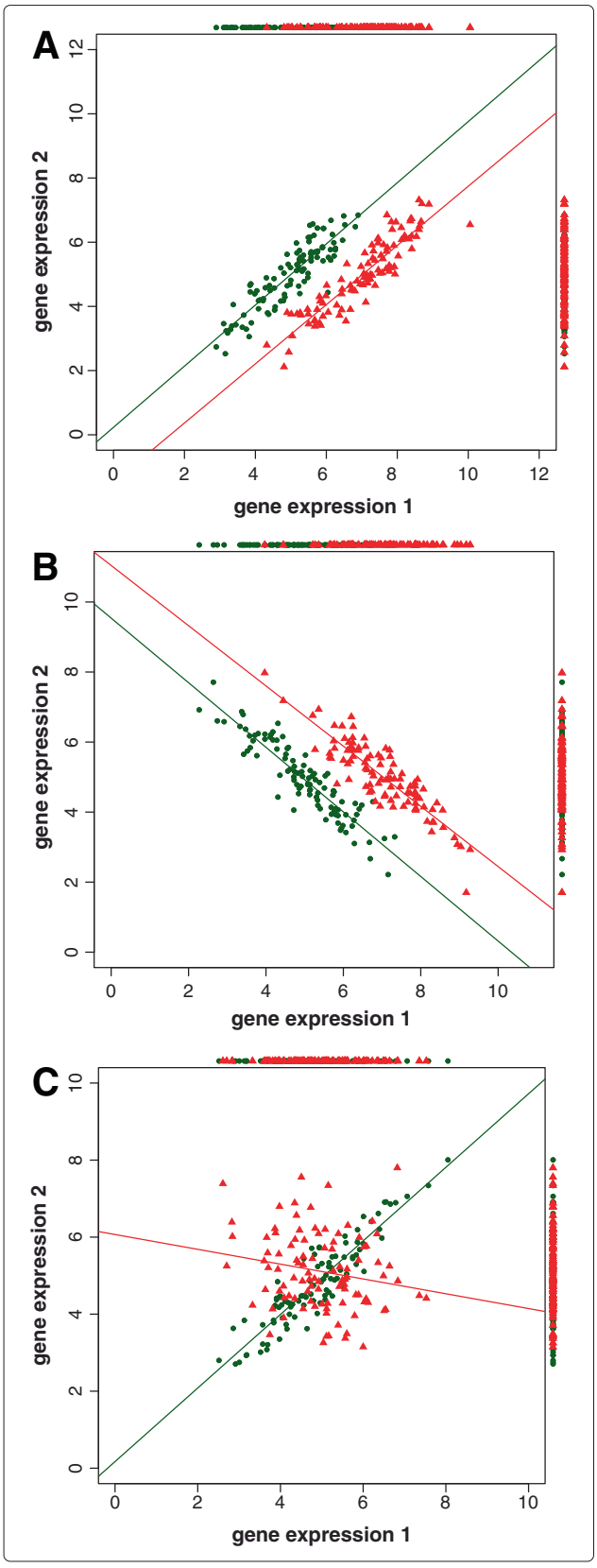

Figure 4 The joint distribution of the expression of two genes. In **A** and **B** the expression values of the two genes show a shift in the two conditions (red and green). In **C** the expression values of a gene pair are anti-correlated between the two conditions. In **A-C** the lines correspond to linear regression fits of the colored expression values and the color dots surrounding the figures provide one-dimensional projections of the two-dimensional distributions.

state. Let X, Y and Z be gene-expression profiles. We say that X and Y form a *liquid association pair* (LAP), if the cellular state of X and Y is correlated with Z. The LA score of X and Y with respect to Z is estimated from rank transformed expression profiles of X, Y, Z given by

$$LA(X, Y|Z) = \frac{1}{m} \sum_{i=1}^{m} X_i Y_i Z_i \tag{11}$$

Here m corresponds to the number of samples. The LA method uses a permutation test for the identification of significant LA gene pair values. Due to the high computational burden of the method that would require N^3 (N is the number of genes) evaluations of Eqn 11 plus additional permutations of the data, which is even for only $N = 10^3$ genes intractable because it requires already more than 10^9 evaluations. For this reason the method is only used to (A) find the gene Z for a given pair of genes or (B) find the LAP, X and Y, for a given gene Z.

Gene set and pathway analysis methods

Generally, a pathway is a group of interacting genes (a gene set) that deploy a cellular function. In a biological system the biological processes are coordinated functions of sets of genes which make the organism work. Some general pathways are, e.g., metabolic pathways, signaling pathways or regulation pathways that represent minimal functioning units of a cellular system. The consideration of pathways or gene sets for a comparative gene expression analysis is an important step toward the exploration of relevant functional mechanisms of a cell.

So far, many multivariate and univariate tests have been proposed for a gene set analysis, see Figure 2. Finding differentially expressed pathways, instead of individual genes, is not straight forward from a statistical and biological perspective and there are several hurdles to this approach. The first is presented by the data themselves, because the number of variables is usually (much) larger than the number of samples, i.e., $n << p$, that leads to many estimation problems. The second hurdle is our incomplete information about the constitution of biological pathways and the potentially high overlap of genes between different pathways. For example, databases like GO [60] provide valuable information about genes for a large variety of different organisms. However,

this information is not static but continuously expanding leaving us at the moment with a snap-shot of knowledge. This makes it difficult to find precise definitions for particular pathways of interest. The third problem comes from the underlying gene network structure that describes the true interactions between genes in a pathway. Here, the problem is that as a result of such interactions among genes it is usually not appropriate to assume their independence, as frequently done for statistical ease.

A motivating example for the general idea underlying gene set methods is shown in Figure 3C. For condition 1 (green), the samples are drawn from a multivariate normal distribution with $\boldsymbol{\mu_1} = \{2, 2.2, 2.4, 2.6, 5, 8\}$ and for condition 2 (red) $\boldsymbol{\mu_2} = \{2, 7, 7.5, 2.6, 5, 8\}$. The covariance matrix, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is for both conditions the same. In this Figure, only 2 of the 6 gene are differentially expressed. This reflects biological situations because, usually, only of fraction of the genes belonging to a pathway is found to be differentially expressed. However, due to the fact that gene set methods are based on the expression of a set of genes such methods borrow strength from the combined analysis of the genes.

Reviews that focus entirely on gene set and pathway analysis methods can be found in [61-65].

Null hypothesis for gene set analysis

Gene set analysis methods can be broadly divided into two major categories, depending on what null hypothesis is tested. The first type of methods are called *competitive methods*, and the second type *self-contained methods* [66]. Briefly, self-contained tests use only the data from a target gene set under investigation, whereas competitive tests use, in addition, also data *outside* the target gene set (background data). In the following we describe popular competitive and self-contained pathway methods.

Competitive gene set and pathway methods

In Table 1 we show an overview of gene set and pathway methods, described in the following.

GSEA

The *gene set enrichment analysis* method (GSEA) [27,68] is one of the most widely used competitive test based method. The test uses a Kolmogorov-Smirnov test statistic to identify differential expressed gene sets. The gene set and background data set are defined in the following. Let W be the target gene set to be tested and W^c its complement in a way that the union of both sets defines all genes, i.e., $V = W \cup W^c$, in the data set. The hypotheses tested by GSEA with respect to an *enrichment score* (ES) are:

$$H_0 : ES = 0(vanishingtestscore)$$
$$H_1 : ES \neq 0(non - vanishingtestscore)$$

Briefly, GSEA consists of the following steps, applied to each pathway:

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

GSEArot

GSEArot (gene set enrichment analysis rotation) [70] is very similar to GSEA, but uses a different approach to randomize data in order to assess the significance of a target pathway. More specifically, a data matrix $\boldsymbol{X}$ is randomized by, first, rotating $\boldsymbol{X}$ around a random angle δ, resulting in a matrix $\boldsymbol{X}(\delta)$. Second, from the matrix $\boldsymbol{X}(\delta)$,

Table 1 Overview of different competitive gene set methods

Principle Method	Reference	Test type	Software
Over-representation analysis (hypergeometric test)	[67]	parametric	GOstats
GSEA	[68]	non-parametric	GSEABase
	[27]	non-parametric	www.broad.mit.edu/gsea
GSA	[69]	non-parametric	http://cran.r-project.org/web/packages/GSA/
GSEArot	[70]	non-parametric	limma
GAGE	[71]	parametric	GAGE
PAGE	[72]	parametric	PGSEA, GAGE
Random Set	[73]	parametric	part of CLEAN
Generalized Random Sets	[74]	parametric	http://GenomicsPortals.org
Gene set enrichment analysis made simple	[75]	parametric	

If available, the name of the software package is provided.

the randomization matrix is obtained by a QR decomposition [76]. In [70] it is argued that this procedure has an advantage for small sample sizes, when only very few permutations are achievable from sample-label permutations. The null hypothesis tested by GSEArot is the same as for GSEA.

Random set

The *random set* method introduced in [73] is a parametric test that is a generalization of Fisher's exact test in the sense that enrichment scores of gene sets are compared with randomly formed sets. The enrichment scores are based on single gene-level test statistics reflecting their differential expression.

1. Estimate the enrichment score of a target gene set W,
$$\bar{s} = \frac{1}{m}\sum_{i\in W} s_i. \tag{12}$$
Here s_i are gene-level scores, e.g., t-scores, and $m = |W|$ is the number of genes in the target pathway.
2. Estimate the enrichment score and its variance of the background gene set $V = W \cup W^c$,
$$\mu = \frac{1}{p}\sum_{i\in V} s_i, \tag{13}$$
$$\sigma^2 = \frac{1}{m}\left(\frac{p-m}{p-1}\right)\left(\frac{\sum_{i\in V} s_i^2}{p} - \left(\frac{\sum_{i\in V} s_i}{p}\right)^2\right), \tag{14}$$
with $p = |W \cup W^c|$.
3. Estimate the standardized enrichment score
$$Z = \frac{\bar{s}-\mu}{\sigma}. \tag{15}$$

The Z score follows a standard normal distribution under the null hypothesis H_0 given by:

$$H_0 : \text{The target set } W \textbf{ is not} \text{ enriched for differentially expressed genescompared with } W \cup W^c \tag{16}$$

$$H_1 : \text{The target set } W \textbf{ is} \text{ enriched for differentially expressed genescompared with } W \cup W^c \tag{17}$$

It is notable that Z can be calculated without a numerical randomization of the data. Further, the background data consist of all genes V, including the ones in the target pathway W. In [77] this method has been applied to *head and neck* and *cervical cancer* for human papillomaviruses-positive and -negative samples.

GAGE

GAGE [71] (generally applicable gene set enrichment) is also a parametric test that, similarly to GSEA, compares the expression in a target gene set with that of the background. But instead of using a Kolmogorov-Smirnov like test [78] it employes a two-sample t-test. The principle steps of the method are as follows:

1. Estimate the mean fold change f and its standard deviation σ_f for the m genes in the target pathway W.
2. Estimate the mean fold change f' and its standard deviation $\sigma_{f'}$ for all p genes in the background gene set $V = W \cup W^c$.
3. Estimate the t-score:
$$t = \frac{f - f'}{\sqrt{\sigma_f^2/m + \sigma_{f'}^2/m}} \tag{18}$$
with
$$df = (m-1)\frac{\left(\sigma_f^2 + \sigma_{f'}^2\right)^2}{\sigma_f^4 + \sigma_{f'}^4} \tag{19}$$
degrees of freedom.

Also GAGE employs all genes V in the background gene set, including the ones in the target set W. The underlying assumption of GAGE is that the (mean) fold changes of genes are independent and identically distributed. The null hypothesis tested by GAGE is:

$$H_0 : \text{The mean fold change of genes (MFG) in set } W \textbf{ is not} \text{ different to the MFG in } W \cup W^c \tag{20}$$

$$H_1 : \text{The mean fold change of genes (MFG) in set } W \textbf{ is} \text{ different to the MFG in } W \cup W^c \tag{21}$$

GSA

Another method is GSA (gene set analysis) [69]. The method, first, calculates z-scores, z_i, with $i \in \{1, \ldots, m\}$, for all m genes in a given target pathway W. Then each z-score is transformed into two scores, assessing the sign of z_i.

$$s^+(z) = \max\{z, 0\} \tag{22}$$

$$s^-(z) = -\min\{z, 0\} \tag{23}$$

This results in two sets of non-zero scores $\mathcal{S}^+ = \{s^+(z_1), \ldots, s^+(z_m)\}$ and $\mathcal{S}^- = \{s^-(z_1), \ldots, s^-(z_m)\}$ from which their mean value is calculated,

$$\bar{s}^+ = mean(\mathcal{S}^+) = \frac{1}{m}\sum_i s^+(z_i) \tag{24}$$

$$\bar{s}^- = mean(\mathcal{S}^-) = \frac{1}{m}\sum_i s^-(z_i) \tag{25}$$

Finally, the *maxmean* test statistic is defined by $s_{mm} = \max\{\bar{s}^+, \bar{s}^-\}$, giving the test statistic for the target pathway.

$$H_0 : s_{mm}(W) = s_{mm}(W^c) \tag{26}$$

$$H_1 : s_{mm}(W) \neq s_{mm}(W^c) \tag{27}$$

The null distribution is assessed by a *restandardization*, combining a sample- and gene-label permutation.

Self-contained gene set and pathway methods

In Table 2 we show an overview of self-contained gene set and pathway methods that are describe in the following in more detail.

Sum of t-square

The *sum of t-square* test is an univariate test that is based on t-scores, $\{t_i\}$, individually obtained for each of the m genes in a given set [95], see also [79]. That means that each t-score assesses the difference of the mean expression between the two conditions,

$$t_i = \frac{\Delta\bar{x}_i - \Delta\mu_i}{\bar{s}_i} \tag{28}$$

Table 2 Overview of self-contained gene set and pathway methods

Principle Method	Reference	Software
Average of single-gene statistics	[79]	sigPathway
Linear Model Toolset for GSEA	[80]	GSEAlm
SAM-GS	[81]	
Globaltest	[82]	globaltest
GlobalANCOVA	[83]	GlobalAncova
Hotelling's T^2	[84-87]	PCOT2
N-statistic	[88]	cramer
RCMAT	[89]	
Non-linear tests for identifying differentially expressed genes or genetic networks	[87]	
Pathway-express	[90]	
Signaling Pathway Impact Analysis	[91]	SPIA (Bioconductor)
SEPEA	[92]	
PARADIGM	[93]	
Gene set analysis exploiting the topology of a pathway	[94]	IPS (available upon request)

with s_i the pooled standard deviation. The test statistic for a pathway is based on the individual t-scores given by

$$\mathrm{TS} = \frac{1}{m}\sqrt{\sum_i^m t_i^2}. \tag{29}$$

Because for each gene $\Delta\mu_i = 0$ should hold if a gene is not differentially expressed, the null and alternative hypothesis can be formulated as:

$$H_0 : \mathrm{TS} = 0 \text{ (vanishing test score)} \tag{30}$$

$$H_1 : \mathrm{TS} \neq 0 \text{ (non-vanishing test score)} \tag{31}$$

The significance of TS is assessed from sample-label permuted data.

SAM-GS

The method *SAM-GS* (Significance Analysis of Microarray for gene sets) [81] uses the test statistics,

$$\text{SAM-GS} = \sum_{k=1}^m d_k^2, \tag{32}$$

with $d_k = \frac{\bar{x}_{1k} - \bar{x}_{2k}}{s_k + s_0}$. Here $\bar{x}_{1k}$ and $\bar{x}_{2k}$ are the sample means of the control and treatment condition of gene k, s_k corresponds to its pooled standard deviation and s_0 is a constant for a sensitivity adjustment. The null and alternative hypothesis can be formulated as:

$$H_0 : \text{SAM-GS} = 0 \text{ (vanishing test score)} \tag{33}$$

$$H_1 : \text{SAM-GS} \neq 0 \text{ (non-vanishing test score)} \tag{34}$$

Statistical significance of *SAM-GS* is again assessed from sample-label permuted data.

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:

$$H_0 : \boldsymbol{\mu}^T = \boldsymbol{\mu}^C \text{ (equality in the m-dimensional population mean vectors)} \tag{35}$$

$$H_1 : \boldsymbol{\mu}^T \neq \boldsymbol{\mu}^C \text{ (difference in the m-dimensional population mean vectors)} \tag{36}$$

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

$$\mathbf{S} = \frac{(n_T - 1)\mathbf{S}_T + (n_C - 1)\mathbf{S}_C}{(n_T + n_C - 2)} \tag{37}$$

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

$$T^2 = \frac{n_T \times n_C}{n_T + n_C} \left(\boldsymbol{\mu}^T - \boldsymbol{\mu}^C\right) \mathbf{S}^{-1} \left(\boldsymbol{\mu}^T - \boldsymbol{\mu}^C\right)^t. \quad (38)$$

The inverse of the covariance matrix is estimated via the shrinkage estimator [96-99]. The statistical significance of the test statistic T^2 is estimated from sample-label permuted data.

N-statistic

The N-statistic is a non-parametric test that is used to test the equality of two distributions. Suppose the expression level of the i^{th} sample of the control group, n_C, and the treatment group, n_T, is given by $X_i^C = \left(X_{i1}^C, X_{i2}^C, \ldots X_{im}^C\right)^t$ and $X_i^T = \left(X_{i1}^T, X_{i2}^T, \ldots X_{im}^T\right)^t$, respectively. Let $i \in \{1 \ldots n_C\}$ correspond to the control data-set and $i \in \{1 \ldots n_T\}$ to the treatment data-set. The null and alternative hypothesis tested by the N-statistic can be formulated as:

$$H_0 : F_C(x) = F_T(x) \quad (39)$$

$$H_1 : F_C(x) \neq F_T(x) \quad (40)$$

whereas $F_C(x)$ and $F_T(x)$ are two multivariate distribution functions from the control and the treatment condition.

The N-statistic itself is defined as follows:

$$\hat{N} = \left[\frac{1}{n_C n_T} \sum_{i=1}^{n_C} \sum_{j=1}^{n_T} 2K\left(x_i^C, x_j^T\right) - \frac{1}{n_C^2} \sum_{i=1}^{n_C} \sum_{j=1}^{n_C} K\left(x_i^C, x_j^C\right) - \frac{1}{n_T^2} \sum_{i=1}^{n_T} \sum_{j=1}^{n_T} K\left(x_i^T, x_j^T\right) \right]^{1/2} \quad (41)$$

Here $K\left(x_i^C, x_j^T\right)$, defined as $K\left(x_i^T, x_j^C\right) = \left\| x_i^T - x_j^C \right\|_2$, is the Euclidean Kernel serving as distance function between the expression values in the two conditions.

Linear model-based pathway methods

There are also several approaches that utilize either a linear or a generalized linear modeling framework for a gene set analysis. Examples for such methods are Global test [82], Extension of GSEA [80] or GlobalAncova [83].

Topological pathway methods based on existing network information

Some recent univariate methods, for instance, Pathway-express [90], SPIA [91] or SEPEA [92], use instead of correlation measures to estimate interactions among genes, predefined topological information as provided, e.g., by the KEGG database [100]. These methods assign each gene in a pathway a score that is based on the position of a gene in the given network structure and, finally, aggregate these individual gene scores to obtain a score for the pathway itself. Yet another approach is provided by PARADIGM [93]. This method uses a factor graph model combining gene copy number variation data with gene expression data for the identification of differentially expressed pathways.

Iterative Proportional Scaling: IPS

IPS (Iterative Proportional Scaling) [94,101] is another method that utilizes the topology of pathways of a given gene set by testing the hypotheses:

$$H_0 : \boldsymbol{\Sigma}_{c_1}^{-1} = \boldsymbol{\Sigma}_{c_2}^{-1} \quad \text{:assuming} \boldsymbol{\Sigma}_{c_1}^{-1}, \boldsymbol{\Sigma}_{c_2}^{-1} \in S^+(G) \quad (42)$$

$$H_1 : \boldsymbol{\Sigma}_{c_1}^{-1} \neq \boldsymbol{\Sigma}_{c_2}^{-1} \quad \text{:assuming} \boldsymbol{\Sigma}_{c_1}^{-1}, \boldsymbol{\Sigma}_{c_2}^{-1} \in S^+(G) \quad (43)$$

In this method, the covariance matrices, $\boldsymbol{\Sigma}_{c_1}, \boldsymbol{\Sigma}_{c_2}$, are estimated from the data, for both conditions, using the Iterative Proportional Scaling (IPS) algorithm. The inverse of the estimated covariance matrices are positive definite (concentration) matrices for which it is assumed that the non-zero elements in $\boldsymbol{\Sigma}_{c_1}^{-1}$ and $\boldsymbol{\Sigma}_{c_2}^{-1}$ are identical; this is the meaning of $\boldsymbol{\Sigma}_{c_1}^{-1}, \boldsymbol{\Sigma}_{c_2}^{-1} \in S^+(G)$ where S^+ indicates the class of all symmetric positive definite matrices with non-zeros elements given by the binary matrix G. This means that the concentration matrices, $\boldsymbol{\Sigma}_{c_1}^{-1}$ and $\boldsymbol{\Sigma}_{c_2}^{-1}$, have identical zero element, but are allowed to have different non-zero entries. In other words, it is assumed that the underlying topology of a pathway is the same for condition c_1 and c_2, given by G, whereas $G_{ij} = 0$ corresponds to an 'absent' interactions among the genes i and j. Since the structure of G is not estimated from the data, it is necessary to obtain it from an independent source, e.g., from the KEGG database or Reactome [100,102]. In [94] it is shown that a log likelihood ($\log(\Lambda)$) ratio test can be used to test for the equality of the concentration matrices for the two conditions and that asymptotically the log likelihood ratio follows a Chi-square distribution with $r + m$ degrees of freedom, i.e., $\log(\Lambda) \sim \chi^2_{r+m}$, whereas m is the number of genes in the pathway and r is the number of non-vanishing edges in G corresponding to the fixed interaction structure of the pathway.

The IPS method has been used in [94] to study acute lymphocytic leukemia with and without BCR/ABL gene rearrangement. As a result, the JUN oncogene with RAS/MAPK/JNK followed by NFAT and NFKB seem to be crucial in distinguishing BCR/ABL positive and negative patients.

Differential correlation/interaction methods

In the previous sections, we discussed different gene set and pathway-based methods for the identification of differentially expressed pathways. These methods focused either only on the expression of genes, or considered an underlying interaction topology among the genes as taken from an independent source. However, even when these

methods considered an underlying network structure, this structure was assumed to be the same for the 'treatment' and 'control' group.

In contrast, in this section we discuss methods that *estimate* the correlation/interaction structure of the genes within pathways, for each experimental condition. The underlying rationale for these approaches is to assume that the expression profiles of genes are dependent on each other [103,104] as the genes in a pathway interact, either directly or indirectly [105]. This assumption results from the observation that genes with similar functions or cellular localization are often co-expressed and cluster together. The methods discussed in this section bear a similarity to the statistical methods for the estimation of differential correlated gene-pairs (see section 'Gene-pair analysis'). However, the extension of such gene-pair measures to the pathway level allows the identification of pathways that show, e.g., a condition specific correlation change.

In Figure 3E we show a simulated example scenario for condition specific correlation changes of the expression profiles for a gene set. In Figure 3E the correlation between all gene-pairs of a gene set is aggregated by a summary statistic. In this example, the mean values between the genes is of a comparable order, whereas the correlation of the gene set in the treatment condition is reduced.

A variety of different pathway methods have been developed that integrate the estimated gene correlations or co-expression structures with gene expression data. A summary of different methods that are used for the identification of differential correlation/interaction changes in pathways is shown in Table 3. In the following these methods are described in more detail.

Graph Edit Distance: GED

Among the first approaches that estimate the interaction structure for a pathway analysis to identify *differentially correlated pathways* (DCP) is a method introduced in [106]. This method uses the *graph edit distance* (GED) score as a test statistic.

More precisely, for a given pathway containing m genes an association graph, G, also called *pseudo-pathway*, is inferred for each condition. That means the resulting network comprises the m genes of this pathway only. The inference method estimates correlation and partial correlation coefficients and tests their statistical significance. That means, if either the correlation or partial correlation coefficient for two genes i and j in this pathway vanishes, then the resulting network will *not* have an interaction between gene i and j, i.e., $E_{ij} = 0$, otherwise there is an interaction, $E_{ij} = 1$. Here E corresponds to the adjacency matrix of the network. Suppose G_{c_1} and G_{c_2} are two networks that have been inferred from gene expression data for condition c_1 and c_2. Further, assume that $M_1, M_2, \ldots, M_n$ are all possible transformations that map G_{c_1} into G_{c_2}, i.e, $M_i(G_{c_1}) = G_{c_2}$. Then the *optimal cost* of the optimal transformation, M', is given by $c(M') = \min\{c(M_i)|1 \leq i \leq n\}$. This value is used to define a dissimilarity measure $d_{\mathrm{GED}}(G_{c_1}, G_{c_2}) = c(M')$ between the two networks G_{c_1} and G_{c_2}, called the *graph edit distance* (GED) score [112]. For arbitrary networks, G_{c_1} and G_{c_2}, the estimation of $d_{\mathrm{GED}}(G_A, G_B)$ is numerically challenging. However, for our specific problem it can be efficiently calculated based on the adjacency matrices, E^{c_1} and E^{c_2}, of the two networks,

$$d_{\mathrm{GED}}(G_{c_1}, G_{c_2}) = \frac{2}{m(m-1)} \sum_{ij}^{m} \left| E_{ij}^{c_1} - E_{ij}^{c_2} \right|. \qquad (44)$$

In [106] the GED score has been used as a test-statistic for the formulation of the hypotheses:

$$H_0 : d_{\mathrm{GED}}(G_{c_1}, G_{c_2}) = 0 \qquad (45)$$

$$H_1 : d_{\mathrm{GED}}(G_{c_1}, G_{c_2}) \neq 0 \qquad (46)$$

In order to assess statistical significance, sample label permutations are performed to obtain the null distribution.

Extensions of this method can be found in [113] where mutual information values have been used to capture nonlinear relations among gene expression values. Further, in [114] a methods based on a *relevance value* (RV) has been defined for integrating different types of genomics data sets which has also a resemblance to the GED.

Table 3 Overview of methods for the identification of differential correlation/interaction changes in pathways

Principle Method	Reference	Software
Graph edit distance	[106]	
Gene-set co-expression analysis (GSCA)	[107]	GSCA (http://www.biostat.wisc.edu/~kendzior/GSCA/)
Differential co-expression (dCoxS) between gene-sets	[108]	dCoxS (http://www.snubi.org/publication/dCoxS/index.html)
DiffCoEx	[109]	R code is provided in the paper
Differential disease network using C3NET	[106,110]	c3net (http://cran.r-project.org/web/packages/c3net/index.html)
Disease associated interactions using Synergy network	[111]	MATLAB code is provided in the paper

Gene set co-expression analysis: GSCA

A method that is based on (zero-order) correlation coefficients is *gene set co-expression analysis* (GSCA) [107]. This method uses as test statistic the *dispersion index*, which is defined as follows:

$$D_s(\rho^{c_1}, \rho^{c_2}) = \sqrt{\frac{1}{P}\sum_{k=1}^{P}\left(\rho_k^{c_1} - \rho_k^{c_2}\right)^2} \tag{47}$$

Here ρ_k^c, with $c \in \{c_1, c_2\}$, is the k-th correlation coefficient for a gene pair, which can be formed among the total number, $P = \binom{m}{2}$, of such pairs for a pathway consisting of m genes. The null and alternative hypotheses tested are:

$$H_0 : D_s(\rho^{c_1}, \rho^{c_2}) = 0 \tag{48}$$

$$H_1 : D_s(\rho^{c_1}, \rho^{c_2}) \neq 0 \tag{49}$$

From the definition of the dispersion index follows that also this method aims at detecting at differential correlation among pathways, despite its name emphasizing co-expression. Interestingly, the *dispersion index* corresponds to the GED score if its components in Eqn. 47 are re-labeled and one defines the components of the adjacency matrices E^{c_1}, E^{c_2} as the correlation coefficients rather than the outcome of the hypotheses tests [106].

A visualization of the underlying idea of GSCA is shown in Figure 3E. The gene expression values are sampled from multivariate normal distribution $N(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu_2}, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$, and the average covariance between gene-pairs is $\boldsymbol{\Sigma}_1 = 0.8$ and $\boldsymbol{\Sigma}_2 = -0.2$. Despite the fact that there is neither a difference in the individual expression of genes nor the the expression of a set of genes, condition 1 and 2 can be distinguished by using a measure based on a correlation change.

Partial least squares based scores: PLS

A statistical framework based on a *partial least squares* score is proposed in [115]. Similar to the above methods, two matrices for the two conditions are inferred. These matrices can be seen as weighted networks, whereas an edge weight corresponds to the strength of the association between two genes. In this paper, three different types of tests are introduced that allow (A) testing for changes in the module structure of the two networks, (B) testing for changed in the connectivity of a particular gene set, and (C) testing for changes in the connectivity of a particular gene.

Differentially co-expressed gene sets: dCoxS

In [108] the *differentially co-expressed gene sets* (dCoxS) algorithm is proposed. This is an entropy-based method that uses the *interaction score* (IS) to measure the difference between two pathways. The IS is estimated by the correlation coefficient between the entropies of the two pathways. The entropies themselves are estimated by using the *Rényi relative entropy*, which is defined by:

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha - 1}\log\left(\int \left(p^\alpha q^{1-\alpha}\right) dp\, dq\right) \approx \log\frac{\hat{f}_h(S_i)}{\hat{f}_h(S_j)} \tag{50}$$

Here α is a parameter and $\hat{f}_h(S_i)$ and $\hat{f}_h(S_j)$ are expression densities of the samples i and j, estimated by using a multiplicative kernel for the density estimation. Further, p and q are the probability density functions of P and Q. From this, the *IS* is estimated by:

$$IS = \frac{\sum_{i<j}(RE^{P_1} - \bar{R}E^{P_2})(RE^{P_1} - \bar{R}E^{P_2})}{\sqrt{\sum_{i<j}(RE^{P_1} - \bar{R}E^{P_1})^2}\sqrt{\sum_{i<j}(RE^{P_2} - \bar{R}E^{P_2})^2}}. \tag{51}$$

In this equation, RE^{P_1} and RE^{P_2} are entropy matrices of two gene sets, P_1 and P_2, for condition c_i, whereas each component of the entropy matrices is proportional to $\sim \log\frac{\hat{f}_h(S_i)}{\hat{f}_h(S_j)}$. That means, strictly $IS = IS(P_1(c_i), P_2(c_i))$. Application of Fisher's Z-transformation to *IS* results in a z-score, $z\Big(IS(P_1(c_i), P_2(c_i))\Big)$, for condition c_i. Combination of both z-scores for condition c_1 and c_2 leads to,

$$z_{comb} = \frac{z\Big(IS\Big(P_1(c_1), P_2(c_1)\Big)\Big) - z\Big(IS\Big(P_1(c_2), P_2(c_2)\Big)\Big)}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}. \tag{52}$$

Here n_1 and n_2 correspond to the number of samples for condition c_1 and c_2. The interpretation of the null hypothesis tested can be stated as:

$$H_0 : z_{comb} = 0 \text{ (equality in entropy changes between gene-pairs in the pathways } P_1 \text{ and } P_2 \text{ between the conditions } c_1 \text{ and } c_2) \tag{53}$$

$$H_1 : z_{comb} \neq 0 \text{ (difference in entropy changes between gene-pairs in the pathways } P_1 \text{ and } P_2 \text{ between the conditions c}_1 \text{ and } c_2) \tag{54}$$

In [108] dCoxS has been applied to gene expression data from lung cancer. Their analysis identified the *Thrombin signaling and protease-activated receptors pathway*, which is known to be involved in the angiogenesis of lung cancer, as the most frequently changed pathway. Another interesting result found is that all significant pathway pairs had a lower interaction score in lung cancer than in the normal control group. This might indicate that the variability in

form of exploited parallel pathways is in cancer lower than in normal cells.

Gene regulatory networks

Finally, we would like to mention that also gene regulatory network inference methods have also been used in this context. More precisely, several attempts have been made to identify disease networks [110,111] that corresponds to particular pathways. For instance, in [110] the C3NET inference method [116,117] has been used to infer pathway specific networks for prostate cancer. A structural comparison between the pathway-specific networks, similar to [106] which is based on testing the hypothesis in Eqn. 45, allowed to identify growth and cell cycle related pathways.

On a side note, we would like to add that *Gaussian graphical models* (GGM), also known as *Markov random fields* [118-120], are also frequently used to infer gene regulatory networks. This model assumes that all variables follow a multivariate normal distribution with a specific structure of the inverse of the covariance matrix, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, whereas $\mathbf{\Omega}$ is called the precision or concentration matrix. Network inference methods based on GGM make use of the relation,

$$\rho_{ij|V\setminus\{ij\}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, \tag{55}$$

connecting the partial correlation coefficient of *full-order* (LHS) with the elements of $\mathbf{\Omega}$, $\omega_{ij} \in \mathbf{\Omega}$. The partial correlation is of full-order (with respect to the number of genes) because $V\setminus\{ij\}$ is the set of all genes excluding i and j, i.e., the largest possible set of genes not considering i and j.

Several methodological improvements have been suggested to infer gene regulatory networks based on GGM [121-123]. These methods differ in the way the inverse of the covariance matrix, $\mathbf{\Sigma}^{-1}$, is estimated and in the statistical tests employed to assess significance. The reason for these technical variants comes from a variety of problems. For instance, if the number of samples is smaller than the number of genes, which is typically the case for a microarray data set, the sample covariance matrix is not positive definite and, hence, not invertible. This means that Eqn. 55 cannot be exploited. In order to overcome such practical estimation problems, recently, several extensions based on the LASSO (least absolute shrinkage and selection operator) have been suggested [124-128].

Importance of multiple hypotheses testing and sample size: An example for differentially expressed genes

Typical microarray experiments measure the concentration of thousands of mRNAs simultaneously. For this reason, usually, one does not just test one statistical hypotheses but dozens, hundreds or even thousands. This makes it mandatory to control the overall error rate for all the tests, because the probability to make *at least* one error, $Pr(V \geq 1|\alpha, t) = (1 - (1 - \alpha)^t)$, for a test with a false positive rate of α, increases rapidly with the number of tests t, as can be seen in Figure 5. Here, V corresponds to the number of false positives. From this one can see that even for a moderate number of tested hypotheses, e.g., 300, this probability is already almost 100%. Hence, each of the three principle types of hypotheses tests discussed in the previous sections are severely effected by this problem.

Classically, a Bonferroni correction is used controlling the Family wise Error Rate (FWER) [129,130],

$$FWER(\alpha) = P(V \geq 1). \tag{56}$$

Unfortunately, this method is often too stringent, which may give no significant results at all. For this reason, alternative error measures and control procedures have been introduced. A recent, very popular measure is the false discovery rate (FDR) [131],

$$FDR = \begin{Bmatrix} E[\,V/R] & \text{if } R > 0 \\ 0 & \text{if } R = 0. \end{Bmatrix} \tag{57}$$

controlled by a procedure introduced by Benjamini & Hochberg (BH) [131]. Subsequently, various related error measures have been proposed like pFDR [132], local FDR [133,134] and a variety of other control procedures [129,135]. Also extensions have been suggested [136] that

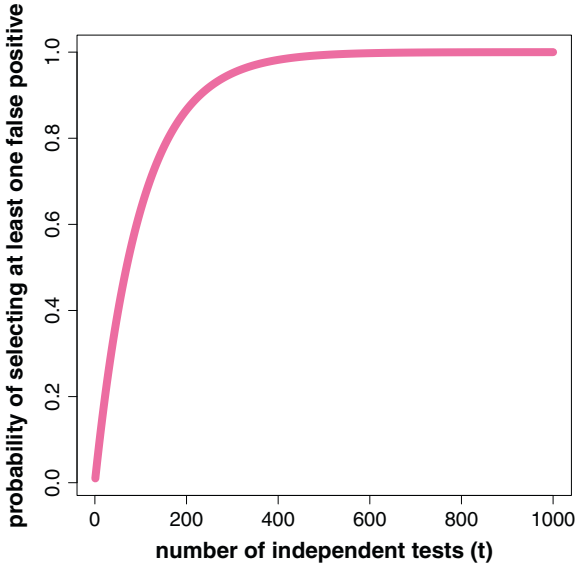

Figure 5 Dependence of $Pr(V \geq 1|\alpha = 0.01, t)$ on the number t of tested hypotheses.

allow the control of an error measure in cases where the underlying tests are not independent from each other. This is particularly important for microarray data that contain a none neglectable correlation structure among the genes.

In order to demonstrate the importance and the influence of different error measures and control procedures, we provide a numerical example, shown in Figure 6. Here, 'BH' means the Benjamini & Hochberg procedure to control the FDR [131], 'pFDR' indicates the positive false positive rate controlled with a method introduced in [132], 'Bonferroni' corresponds to the FWER controlled with a Bonferroni correction and 'localFDR' corresponds to the local FDR [134,137]. The local FDR is defined as,

$$\text{localFDR} = \text{Prob}(\text{null} \,|\text{test statistic}). \tag{58}$$

that means the local FDR is the probability that the null model is true conditioned on the observed test statistic. The data we used for this analysis correspond to simulated gene expression data sampled from a normal distribution

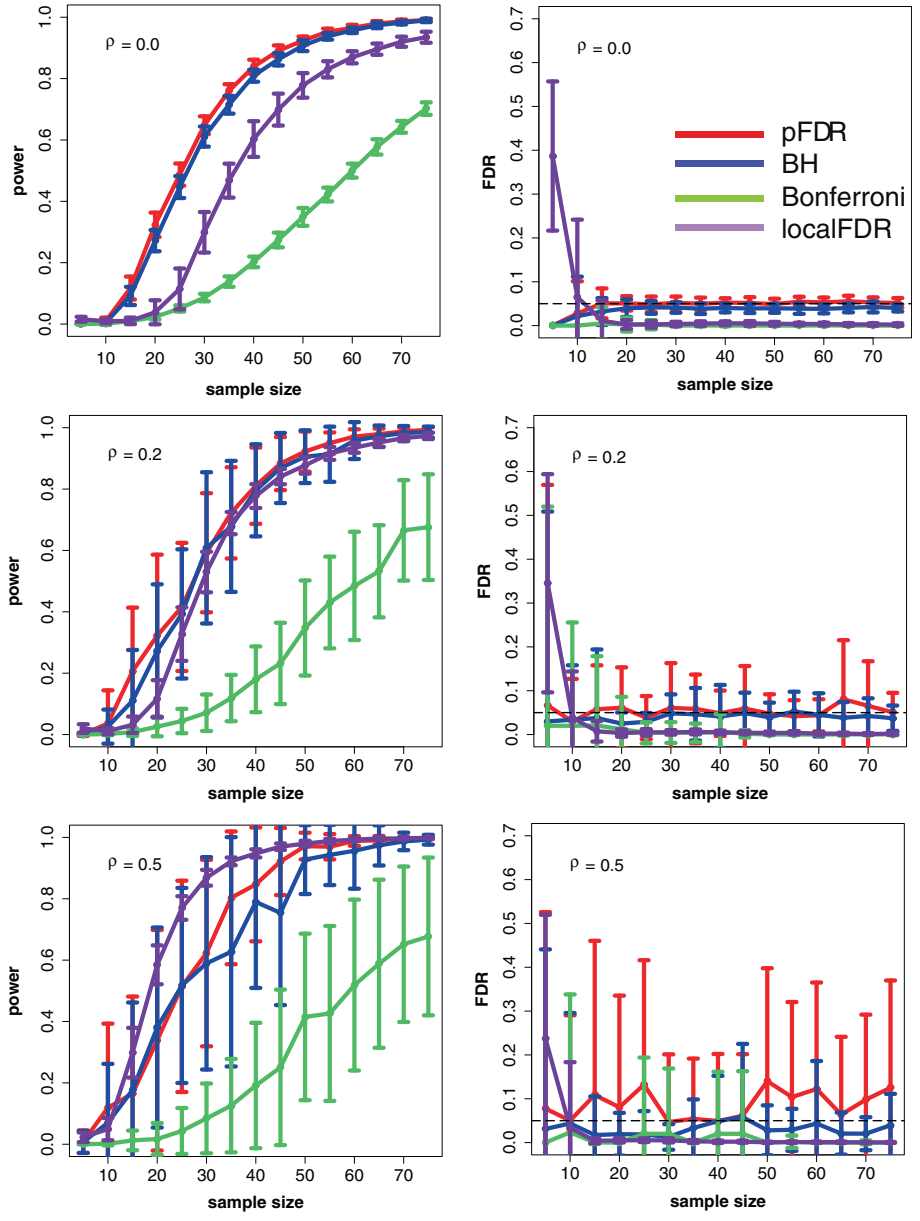

Figure 6 Power and false discovery rate in dependence on the sample size and the correlation in the data. We compare the methods: pFDR (red) [138], BH (blue) [131], Bonferroni (green) and localFDR (purple) [134,137].

with different mean values for the two conditions. More precisely, we simulate 2000 genes of which 400 are differentially expressed (true positives). Further, we study three different (constant) correlation structures with $\rho = \{0.0, 0.2, 0.5\}$. The results shown in Figure 6 are for each sample size averaged over 50 independent runs.

As one can see, 'BH' and 'pFDR' give more significant results and, hence, have a higher power than the Bonferroni correction and the local FDR when there is either no or only a moderate correlation among the genes. However, it is important to note that the utility of these methods depends on the characteristics of the data. For example, if the average correlation in the data is $\rho = 0.0$, then 'pFDR' tends to perform best (see Figure 6A). However, when the average correlation in the data increases ($\rho = 0.5$) then the 'localFDR' [134,137] becomes preferable. We want to note, for a sample size of 5, the power of the methods is usually very low because only a couple of genes test significant. In addition, a large fraction of these can be false positives. This seems to be especially for the local FDR method a problem.

Recommendations

In general, there is a trade-off between a high power of a statistical method on one side, which requires a large number of samples, and low experimental costs on the other. For the identification of differentially expressed genes the results in Figure 6 provide some guidelines. Even for the most favorable condition (for $\rho = 0.0$) a study will usually be underpowered for ≤ 20 samples, however, on the other hand, even for 10 samples the Type I error will be well-controlled.

For gene set and pathway-based methods such recommendation are more delicate. In [105] two self-contained (*sum of t-square* and Hotelling's T^2 [84,95]) and one competitive test (GSEA [27]) have been analyzed. As a results, it is suggested not 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. This can be seen as a reflection of the heterogeneity of cancer, as discussed above in the section 'Gene expression data and cancer'.

In [105] it has been suggested to filter pathways according to the following 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 $DC > 10\%$ (DC is detection call; the percentage of differentially expressed genes in a pathway) 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. That means for the *sum of t-square* test and GSEA, at least 10% of the genes in a pathway should be differentially expressed for the method to work. However, this is not independent of the correlation structure of the data. In general, in the presence of high correlations a larger number of differentially expressed genes is beneficial for these methods.

It is important to emphasize that these sample sizes are different to the minimal sample sizes necessary in order to avoid in addition that a study is underpowered. For the minimal sample sizes [105] predict a sample size of 59 for Hotelling's T^2 and 57 for the *sum of t-square* test and 83 for GSEA. Further, in [95] it was found that using the N-statistic with 40 samples (or more) leads to a good control of the Type I error and a satisfactory power for a variety of differing conditions, including different correlations of the data and DC values in the pathways. Further studies reviewing related methods can found in [61,62,65,139,140].

We would like to emphasize that the above recommendations are data dependent. That means it is not possible to judge solely based on the number of samples which method to use. Instead, one needs to estimate characteristics from a particular data set in order to select an appropriate test. This implies, e.g., to estimate the correlation structure and the detection call. In the context of cancer there is an additional problem that needs to be considered. It is known that a tumor is a heterogeneous collection of cells rather than a homogeneous one [26,141]. This translates into the heterogeneity of gene expression data [142] making it even more dangerous to provide general recommendations without considering a particular data set.

On a general note, we would like to highlight that whenever a given data set allows to (I) identify changes in single genes, (II) identify changes in gene sets or pathways, and (III) identify changes in the correlation structure in pathways, then methods from each of the three categories (I-III) should be applied and there is no need to focus on just one of these. The reason for this is that despite the fact that gene set or pathway methods have more explanatory power than methods to identify changes in single genes [64,95] it does not mean that there are no conditions for which single gene methods reveal interesting biological information that may not be obtained by the other types of methods. For instance, the differential expression of a single gene based on changes in the mean (rather than the variance) may be an indicator for the presence of a single signaling chain rather than of many parallel pathways. Hence, this could provide information about the presence of a Mendelian trait or a complex trait that contains a strong monogenetic component. It appears that for such conditions single gene methods have an advantage over gene set or pathway methods, although, the latter methods may be adaptable to such question as well. However, this may require additional effort. In summary, we recommend to use all different approaches (I-III) side-by side, whenever this is permitted by the data, to interrogate the

data in the broadest way, because this translates into a diverse set of different biological questions.

Our recommendation complements a common line of thought asking for the combination of different types of data. Although it is certainly true that combining different types of high-throughput data, e.g., from DNA microarray and ChIP-chip experiments, is in general more informative, it is also more time and cost intensive to generate such data combinations. For this reason, frequently, only gene expression data are available. Hence, our review provides a survey of method to get the most out of expression data sets.

Finally, we would like to emphasize that all methods require an appropriate filtering and normalization of the data in order to obtain robust and statistically sensible results [143,144].

Conclusions and discussion

In the post genomic era, biology transitioned from a 'gene-centric' to a systems-focused field. This change is also reflected in the transition from methods to identify 'differentially expressed single genes' to approaches for finding 'differentially changed pathways'. Such a transition is natural, because a systems view is required to understand the complex biological functions inside a cell that are responsible for the observable phenotypic outcomes [9,11,145].

As recent findings in cancer research demonstrate, cancer is a heterogeneous disorder, even within a particular cancer type. For example, breast cancer is currently subcategorized into four major tumor subtypes [146]: basal-like, HER2-enriched, luminal (which can be further reduced) and normal-like tumors. Considering the fact that these results have been achieved by using high-throughput data one can expect further refinements when data from different hight-throughput technologies become available and being combined with each other. For this reason, it appears sensible to assume such a heterogeneity not only on the global, phenotype level, but also within the cells, on the pathway-level. This implies that a pathway-based filtering, as suggested in [105], is necessary to apply a method only selectively, and not unconditional, to cancer pathways.

Regarding potential future directions, we expect to see an increase in methods that target changes in the correlation structure in pathways for three reasons. First, genes and their products do interact with each other. This implies that there exists a correlation structure among these entities that represents, potentially, useful biological information that may be missed by co-expression based methods [106]. Second, the costs to generate high-throughput data are declining, which makes it easier for the experimenter to generate a sufficiently large number of samples that enables such an analysis. This is an important point, since the required sample sizes for a pathway analysis is considerably larger than for single gene analyses. Third, biologically, the hallmarks of cancer point to a few pathways as pivotal elements in the molecular elucidation of carcinogenesis, e.g., Wnt/Notch signaling, Hedgehog signaling or DNA damage control [147-149]. Hence, semantically, pathway studies enable the systematic connection of oncogenes, tumor-suppressor genes and stability genes [150] to provide fundamental insights into causal mechanisms underlying cancer. Unfortunately, the temporary literature especially of methodological papers discuss their results rarely in the framework of the *hallmark pathways*. For this reason, we suggest that future studies aim for a conceptual discussion of their results within this enlightening framework. Not because it provides the final answers to understand cancer [151], but due to the fact that it enables a systematic approach to the emperor of all maladies [152].

Reviewers' comments

First of all, we would like to thank all referees for their fruitful suggestions and comments. In the following, we kept our answers to the raised issues short but included our responses in the main text.

Referee 1: Dr. Arcady Mushegian

The manuscript by Emmert-Streib and colleagues is a review of statistical methods for analysis of gene expression data, but it is also much more than that. It is relatively rare for the statisticians to review all classes of such methods and to give an eminently logical classification not only of the techniques on which the methods are based, but also of the kinds of questions that are asked when applying these methods. This, certainly, is a strength of the work and the reason why it should appeal to the biologists that would like to have a deeper insight into which methods are appropriate to which task at hand.

I have, however, several comments that rank somewhere between suggestions and concerns. Most importantly, the authors propose to distinguish three groups of methods: those that identify changes in single genes, those that identify changes in gene sets or pathways, and those that identify changes in the correlation structure in pathways. (By the way, in the Abstract and elsewhere, the description of the groups is almost the same as above, but "changes" are substituted by "differential changes" - is it not a tautology, in particular when there are only two samples?). Then, in discussing the first two classes of methods, the authors almost in every case give a clear formulation of the question that is being asked of the data, in the form of the statistical hypothesis about the data that is being tested. This is an excellent way of explaining things. Unfortunately, it is not consistently applied:

even among these classes of methods, the hypotheses are not mentioned, and then, upon discussing the differential-correlation methods (pp. 15-18), the hypotheses are not explicated at all, except for the IPS method. I think this need to be changed, and the null hypotheses need to be stated for all methods for which this is possible; and if the framework is such that no explicit null hypothesis exist, this needs to be discussed, and the applicable intuitive formulation be given.

Reply:
We appreciate this suggestion and added to all methods the definition of their null hypothesis. In addition, we extended the discussion in section 'Formulating biological hypotheses' explaining why it can be difficult to find a biological interpretation for a null hypothesis and we offer some explanation for this.

Question:
My other concern is abut Figures 3 and 4. The authors never state what the data points there represent. They must be expression values for two genes, but how are these data collected - are they technical replicates? biological replicates? some kind of ordered series? unordered series such as for example different drug treatments? Does it matter what of the above they are?

Reply:
We added an explanation of the data, which are simulated data to visualize the principle idea underlying some methods, to the corresponding methods.

Question:
The third shortcoming of the paper is that there is a significant disconnect between well-covered methodology and the stated goal of discussing the application to cancer biology. In fact, the short discussion about cancer hallmarks is an excellent introduction that points out the way in which analysis of gene expression can lead to the understanding of changes in expression of particular ("hallmark") pathways. This theme, however, is not followed through. Though occasionally we read that such and such method was applied to analysis of a particular type of cancer, there is never any discussion of what was found in gene expression data that allowed an insight into cancer biology. What happens to the hallmark pathways at the level of gene expression programs? Which methods have been used to support (or maybe question?) which aspects of the hallmark hypothesis? Which pathways were predicted or shown to be differentially regulated at the transcription or mRNA concentration level?

Reply:
We agree with the reviewer that 'Which methods have been used to support (or maybe question?) which aspects of the hallmark hypothesis? ' is an important questions. Unfortunately, the methodology oriented literature does rarely touch this topic in a clear manner. That means in order to extend the paper in this direction we could not survey these issues but would need to establish such results. Instead, the concern in our paper is to propagate such an approach in the context of the presented methods. A discussion has been added to 'Discussion and conclusions'.

Question:
Finally, there is the question of, if you will, general biology of transcriptional response. It stands to reason, and indeed has been occasionally shown, that in order for a pathway to be regulated, it may not be necessary to regulate all its components at the same (in this case, mRNA concentration) level. One may find that the gene product amounts are regulated at different levels, or maybe even only one or a few, e.g., rate-limiting, components are regulated at all. This would argue that single gene-based methods may in these cases provide a better clue to the process than pathway-based or gene set enrichment-based methods. It would be interesting to know whether this has been observed in the cancer datasets. A related question is about the rules of thumb in pathway analysis: for example, if a typical pathway (network module?) has a size of N genes, what is the number of genes in this pathway $m < N$ that would still register as an enrichment in some of the tests that the authors discuss?

Reply:
This is an important point. We included a discussion of this to section 'Recommendations'. We added also a discussion of the danger of general suggestions and motivate this by known characteristics of gene expression data from cancer. The problem is twofold. First, each method has its own characteristics under what conditions it works best. Second, data sets from cancer are very heterogeneous so that two data sets containing about the same number of samples can exhibit a very different correlation structure and expression patterns. This holds potentially also for different grades of one cancer type.

Regarding the first question, it appears to us that this is related to the presence or absence of parallel pathways conveying a molecular signal. If for example no parallel pathways exist the detection of differentially expressed genes can provide a robust way to detect functional changes. On the other hand, if there are many alternatives this may not be the case and gene set methods appear to be better suited for such a situation. In general, this kind of cross-method comparisons are not well studied and we are not aware that this has been systematically addressed for cancer or other data sets. One reason for this is that until recently, most data sets

contained less than 20 samples per condition, which usually does not permit a robust analysis of gene set or pathway methods and once larger data sets became available the detection of differentially expressed genes was neglected, potentially, due to the erroneous assumption that differentially expressed gene set methods include the former tests.

In order to emphasize that it is desirable to apply methods from all three different levels simultaneously ((I) identify changes in single genes, (II) identify changes in gene sets or pathways, and (III) identify changes in the correlation structure in pathways) whenever a given data sets allows this, rather than to focus on just one of these levels, we added a discussion to section 'Recommendations'.

Thank you for your suggestions and comments.

Referee 2: Dr. Byung-Soo Kim

General comments This is a well organized review of recent statistical methods of analyzing microarray experiment data sets, particularly on cancers, from single gene analysis to identifying differential changes in pathway, and finally to comparing a given pathway under two different conditions. However, I would like to indicate following four points for the possible improvement. (1) Gaussian graphical model: From the methodological point of view, it is desired to include the sparse Gaussian graphical model (GGM) approach for estimating the gene network under the multivariate normal assumption from a microarray data set. For the recent development of GGM approach one can include glasso (Friedman, Hastie and Tibshirani, 2008; Witten, Friedman and Simon, 2011) [125,127], SCAD penalty of Fan, Feng and Wu (2009) [124], adaptive lasso of Zou (2006) [128] and Kiiveri (2012) [126], among others. (2) Effect of inter-gene correlations on the single gene analysis. A series of Efron's recent work (Efron 2007a, 2007b) [134,137] discussed in detail on how inter-gene correlations could affect the detection of differentially expressed (DE) genes in a single gene analysis? By including Efron's recent work and his R package "locfdr" authors can show how FDR can be used in the real data analysis in their Section on "Importance of multiple hypotheses testing and sample size: An example for differentially expressed genes". (3) Some of the reviews are misleading. These are the few examples. (i) The sentence, at the middle of page 12, "However, in order to use a two-sample t-test with equal size of the two samples it is assumed that the mean fold change f' and its standard deviation σ_f would be the same for a randomly selected background set consisting of only m genes, see Eqn. 10". Actually ([99], Luo et al., 2009) assumes the i.i.d of the fold change of genes to make Eqn 10 have a t distribution. Here the key assumption was the independence, which was missing in the aforementioned sentence. (ii) p. 14. Eqn 16. In ([126], Tian et al., 2005) no t-square statistic was employed. (iii) Eqn 24 of p. 18 does not make sense. Authors of ([20], Cho et al., 2009) didn't make it clear in their equation (3) what Renyi entropy was when the underlying random variables were continuous. (iv) I would suggest authors to allocate more space on the work of ([90], Massa et al, 2010) which was methodologically sound and deserve more coverage than just the IPS algorithm. (4) Inconsistency of notations. In page 11 authors defined p and m to be sizes of the background genes and a target gene set, respectively. However, in line 2 of page 15 "p genes" (which should have been m genes according to page 11 definition) was incorrectly labeled. This inconsistency was repeated in N-statistic section of p.15, and also in Eqn 16 in p. 14 and Eqn 22 of p. 17. The "p-dimensional..." should be "m-dimensional..." at the bottom two lines of p. 14.

Minor Comments

1. p.2 "Gene expression data from next generation sequencing (RNA-seq)". This is an important issue. There is no direct relevance, however, with statistical methods reviewed in this paper.
2. p.4. For detecting differential correlation and differential variance, it would be better to explain why these approaches were taken. For example, in ([54], Ho et al., 2008) it was clearly indicated that changes in expression variability were associated with changes in coexpression pattern, which implied that DV was a signal rather than a noise.
3. Legend of Figure 2. "The data is..." should be "The data are... ".
4. p.7. There is no reference of Figures A, B in the main text. Also indicate in the legend of Figure 3 what Σ is in Figure 3E.
5. p.8. In the legend of Figure 4 what the symbols in the outer-panel represent? What do the lines represent? It is better to use different notation (A, B) to avoid confusion in the main text of the second paragraph of p. 9.
6. p.9 What is "alpha" in Equation (4)?
7. p.9 line -9. You may include two specific patterns of dependence of two genes, namely, type A dependence of Klevanov, Jordan and Yakovlev (2006), and hidden regulator dependence of Lim, Kim and Kim (2011).
8. p.15 line -13. "euclidean Kernel" should be "Euclidean kernel" (9)
9. p.15 line -10. "a either" should be "either".
10. p.15 line -8. Author may want to include Tsai and Chen (2009) for another reference of Hotelling's T-square statistic.
11. p.17. line 15, What are "A" and "B"?
12. p. 18 line 2. Better to include Lauritzen (1996) as a reference of IPS algorithm.

13. p. 22. It would be more beneficial for the read to move the last paragraph of p. 22 (extended to p. 23) to Introduction section.

References Efron B. (2007a). Correlation and large-scale simultaneous signifance testing, J. Amer. Statist. Assoc. 102:93-103. Efron B. (2007b). Size, power and false discovery rates, Annals of Statist. 35:1351-1377. Fan J, Feng Y, Wu Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. Ann. Statist. 3:521-541. Friedman J, Hastie T, Tibshirani R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432-441. Lauritzen SL. (1996), Graphical models, Oxford: Clarendon Press. Lim J, Kim J, Kim BS (2011). An alternative model of type A dependence in a gene set of correlated genews, Statist. Appl. Genet. Mol. Biol. Vol. 9, Article 12. Kiiver H, de Hoog F. (2012). Fitting very large Gaussian graphical models. Comp. Statist. Data Anal. 56:2626-2636. Klebanov L, Jordan C, Yakovlev A. (2006). A new type of stochastic dependence revealed in gene expression data, Statist. Appl. Genet. Mol. Biol. Vol. 5, Article 7.. Tsai C-A, Chen J. (2009). Multivariate analysis of variance test for gene set analysis, Bioinformatics, 25:897-903. Witten DM, Friedman JH, Simon N. (2011). New insights and faster computation for the graphical lasso. J. Comp. Graph. Stat. 20:892-900. Zou H. (2006). The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101:1418-1429.

Reply:

We revised our text correspondingly and addressed all your suggestions. We would like to point out that the major goal of our review is not a full coverage of statistical details but to provide sufficient information for the reader to acquire a basic understanding of major principles and assumptions that underly the methods. The problem is that if too many detail are presented the paper would turn quickly into a formal description which may not be appreciated by a biology oriented readership.

Minor Comments

1. p.12. line 1: What is N?
2. p. 15. line -7 -5: "two i.i.d samples of genes..." is rather confusing. Luo et al. (2005) assumed the i.i.d of the fold change of genes, which was much stronger than just assuming equal mean and variance. It is better to rewrite this sentence to convey the original material.
3. p.17. line -1: "p genes" should be "m genes".
4. p.22. Eqn. (50): What are p and q? What are S_i and S_j ?
5. p.37. Reference 30, p. 38. References 40, 59; The journal title should be consistent with Reference 27 or vise versa.
6. p.40. References 90, 108: Location of the publisher is missing.
7. p.40. References 93,94; The journal title should be consistent with Reference 119.
8. p.40. Reference 109: Author was duplicated at the end. The location and the publisher were missing.
9. p.41. Reference 117. The article title is missing.
10. p.41. Reference 118: The location of the publisher is missing.
11. p.41. Reference 133. The journal title should be consistent with Reference 27.

Reply:

All comments have been addressed and we revised the main text correspondingly.

Thank you for your suggestions and comments.

Referee 3: Dr. Joel Bader

This manuscript reviews methods for analyzing gene expression data with tests of individual genes, gene pairs, gene sets, and networks. The manuscript is strong in covering many methods. It would be more helpful if the authors also provided a point of view or evaluation of methods. Can anything be said about the relative power of different approaches, or which have proven to be more useful in practice? What about the tradeoff between robustness, power, and speed for realistic data? Most of the discussion of method choice is generally about sample size requirements for all methods rather than method choice given sample size. The two parts of the manuscript, gene expression and cancer, don't really mesh. Most of the methods review is not cancer specific. Possibly of greater relevance to cancer are methods that combine different types of data.

The manuscript is generally well written and easy to understand, with ample references to the original work and to previous reviews.

Minor corrections

1. p. 1 'one gene, − > should be ' for open-quote in latex, here and elsewhere
2. p. 2 differnt microarray − > spelling
3. p. 2 comprises, e.g., mRNAs − > 'e.g.' doesn't sound right here. How about providing a full list: mRNA, tRNA, rRNA, and short regulatory RNAs
4. p. 2 'In the third step the reads are mapped to known exon sequences of genes.' Are there also de novo assembly methods that don't require a template? 'allows to overcome' − > overcomes
5. p. 3 allows to measure − > measures. Can also mention other advantages: splice variants, sequence polymorphisms, no need to design and build a custom chip
6. 'correspond to: self sufficiency' − > no colon between preposition and noun phrase. Can the hallmarks be parallel, all start with noun or verb?
7. p. 9 Eq. 4. How is alpha calculated?
8. Eq. 5 need i = 1 underneath the summation

9. p. 14. Eq. 16 Under the null, it seems that $\Sigma_{t,2}$ should approach $1/\sqrt{(p)}$ rather than 0.
10. Eq. 16 How is the significance of SAM-GS calculated?
11. p. 18 Eq. 24 and text after, use log in math mode rather than log.

Reply:

All comments have been addressed and we revised the main text correspondingly.

Thank you for your suggestions and comments.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

ST simulated and visualized the data. ST, RDMS and FES analyzed and interpreted the results. ST, RDMS and FES wrote the paper. FES conceived the paper. All authors read and approved the final manuscript.

Acknowledgements

ST is supported by a studentship from the National Institute of Immunology. FES and RDMS are supported by the Engineering and Physical Sciences Research Council (EPSRC) and DEL.

Received: 30 July 2012 Accepted: 1 October 2012
Published: 10 December 2012

References

1. Bock G, Goode J: *Novartis Foundation Symposium*: John Wiley & Sons; 1998 .
2. Van Regenmortel M: **Reductionism and complexity in molecular biology.** *EMBO reports* 2004, **5**(9):1016–1020.
3. Mazzocchi F: **Complexity in biology.** *EMBO Rep* 2008, **9:**10–14.
4. von Bertalanffy L: **An outline of general systems theory.** *Br J Philosophy Sci* 1950, **1**(2):134–165.
5. Beadle GW, Tatum EL: **Genetic control of biochemical reactions in neurospora.** *Proc Natl Acad Sci USA* 1941, **27**(11):499–506.
6. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:**57–70.
7. Noble D: **Genes and causation.** *Phil Trans R Soc A* 2008, **366:**3001–03015.
8. Kitano H: **Systems biology: a brief overview.** *Science* 2002, **295**(5560):1662–1664.
9. Han JDJ: **Understanding biological functions through molecular networks.** *Cell Res* 2008, **18**(2):224–237.
10. MacDougall-Shackleton SA: **The levels of analysis revisited.** *Phil Trans R Soc B: Biol Sci* 2011, **366**(1574):2076–2085.
11. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev* 2004, **5:**101–113.
12. Brazhnik P, de la Fuente A, Mendes P: **Gene networks: how to put the function in genomics.** *Trends Biotechnol* 2002, **20**(11):467–472.
13. Emmert-Streib F, Glazko G: **Network biology: a direct approach to study biological function.** *Wiley Interdiscip Rev Syst Biol Med* 2011, **3**(4):379–391.
14. Davidson E, Levin M: **Gene regulatory networks.** *Proc Natl Acad Sci USA* 2005, **102**(14):4935.
15. de Matos Simoes R, Tripathi S, Emmert-Streib F: **Organizational structure of the peripheral gene regulatory network in B-cell lymphoma.** *BMC Syst Biol* 2012, **6:**38.
16. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Nat Acad Sci* 1996, **93:**13–20.
17. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296**(5569):910–913.
18. Jeong H, Tombor B, Albert R, Olivai Z, Barabasi A: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651–654.
19. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks.** *Curr Opin Struct Biol* 2004, **14:**283–291.
20. Lee TI, *et al*: **Transcriptional regulatory networks in saccharomyces cerevisiae.** *Science* 2002, **298**(5594):799–804.
21. Allison DB: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7:**55–65.
22. Dehmer M, Emmert-Streib F, Graber A, Salvador A(Eds): *Applied Statistics for Network Biology: Methods for Systems Biology*. Weinheim: Wiley-Blackwell; 2011.
23. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2**(6):418–427.
24. Metzker ML: **Sequencing technologies - the next generation (With NOTES).** *Nat Rev Genet* 2010, **11:**31–46.
25. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57–63.
26. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144**(5):646–674.
27. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, Mesirov J: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102**(43):15545–50.
28. Chuang HY, Lee E, Liu YT, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Syst Biol* 2007, **3:**140.
29. Compagno M, Lim WK, Grunn A, Nandula SV, Brahmachary M, Shen Q, Bertoni F, Ponzoni M, Scandurra M, Califano A, *et al*: **Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma.** *Nature* 2009, **459**(7247):717–721.
30. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, *et al*: **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target.** *Proc Natl Acad Sci USA* 2006, **103**(46):17402–17407.
31. Krivtsov AV, Twomey D, Feng Z, Stubbs MC, Wang Y, Faber J, Levine JE, Wang J, Hahn WC, Gilliland DG, *et al*: **Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9.** *Nature* 2006, **442**(7104):818–822.
32. Oskarsson T, Acharyya S, Zhang XHF, Vanharanta S, Tavazoie SF, Morris PG, Downey RJ, Manova-Todorova K, Brogi E, Massague J: **Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs.** *Nat Med* 2011, **17**(7):867–874.
33. Mavrakis KJ, Wolfe AL, Oricchio E, Palomero T, De Keersmaecker K, McJunkin K, Zuber J, James T, Khan AA, Leslie CS, *et al*: **Genome-wide RNA-mediated interference screen identifies miR-19 targets in Notch-induced T-cell acute lymphoblastic leukaemia.** *Nat Cell Biol* 2010, **12**(4):372–379.
34. Nam S, Park T: **Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with Epithelial-Mesenchymal transition.** *PLoS ONE* 2012, **7**(4):e31685.
35. Guedj M, Marisa L, De Reynies A, Orsetti B, Schiappa R, Bibeau F, Macgrogan G, Lerebours F, Finetti P, Longy M, *et al*: **A refined molecular taxonomy of breast cancer.** *Oncogene* 2011, **31**(July 2011):1196–1206.
36. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA: **Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.** *J Clin Invest* 2011, **121**(7):2750–2767.
37. Fabbri G, Rasi S, Rossi D, Trifonov V, Khiabanian H, Ma J, Grunn A, Fangazio M, Capello D, Monti S, *et al*: **Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation.** *J Exp Med* 2011, **208**(7):1389–1401.
38. Ooi CH, Ivanova T, Wu J, Lee M, Tan IB, Tao J, Ward L, Koo JH, Gopalakrishnan V, Zhu Y, Cheng LL, Lee J, Rha SY, Chung HC, Ganesan K, So J, Soo KC, Lim D, Chan WH, Wong WK, Bowtell D, Yeoh KG, Grabsch H, Boussioutas A, Tan P: **Oncogenic pathway combinations predict clinical prognosis in gastric cancer.** *PLoS Genet* 2009, **5**(10):e1000676.
39. Setlur SR, Royce TE, Sboner A, Mosquera JM, Demichelis F, Hofer MD, Mertz KD, Gerstein M, Rubin MA: **Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer.** *Cancer Res* 2007, **67**(21):10296–10303.

40. Nucera C, Porrello A, Antonello ZA, Mekel M, Nehs MA, Giordano TJ, Gerald D, Benjamin LE, Priolo C, Puxeddu E, *et al*: **B-Raf(V600E) and thrombospondin-1 promote thyroid cancer progression.** *Proc Natl Acad Sci USA* 2010, **107**(23):10649–10654.
41. Shah MA, Khanin R, Tang L, Janjigian YY, Klimstra DS, Gerdes H, Kelsen DP: **Molecular classification of gastric cancer: a new paradigm.** *Clin Cancer Res* 2011, **17**(9):2693–2701.
42. Perroud B, Lee J, Valkova N, Dhirapong A, Lin PY, Fiehn O, Kultz D, Weiss R: **Pathway analysis of kidney cancer using proteomics and metabolic profiling.** *Mol Cancer* 2006, **5:**64.
43. Trewavas A: **A Brief History of Systems Biology: "Every object that biology studies is a system of systems." Francois Jacob (1974).** *Plant Cell* 2006, **18**(10):2420–2430.
44. Emmert-Streib F, Dehmer M: **Networks for systems biology: conceptual connection of data and function.** *IET Syst Biol* 2011, **5**(3):185.
45. Macneil LT, Walhout AJM: **Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression.** *Genome Res* 2011, **21**(5):645–57.
46. Lehman E: *Testing Statistical Hypotheses*. New York: Springer; 2005.
47. DasGupta A: *Probability for Statistics and Machine Learning*. New York: Springer; 2011.
48. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray smages.** *J Biomed Optics* 1997, **2**(4):364–374.
49. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: **Gene expression profiles in normal and cancer cells.** *Science* 1997, **276**(5316):1268–1272.
50. Tusher V, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98**(18):5116–5121.
51. Chu G, Narasimhan B, Tibshirani R, Tusher V: **Significance analysis of microarrays (SAM) software.** *Nature* 2002, **5:**436–442.
52. Smyth GK: **Limma: linear models for microarray data.** In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005:397–420.
53. Efron B, Tibshirani R, JD S, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J Am Stat Assoc* 2001, **96**(456):1151–1160.
54. Ho JWK, Stefani M, Dos Remedios CG, Charleston MA: **Differential variability analysis of gene expression and its application to human diseases.** *Bioinformatics* 2008, **24**(13):i390–i398.
55. Hu R, Qiu X, Glazko G, Klebanov L, Yakovlev A: **Detecting intergene correlation changes in microarray analysis: a new approach to gene selection.** *BMC Bioinformatics* 2009, **10:**20.
56. Dettling M, Gabrielson E, Parmigiani G: **Searching for differentially expressed gene combinations.** *Genome Biol* 2005, **6**(10):R88.
57. Lai Y, Wu B, Chen L, Zhao H: **A statistical method for identifying differential gene-gene co-expression patterns.** *Bioinformatics* 2004, **20**(17):3146–3155.
58. Dawson JA, Ye S, Kendziorski C: **R/EBcoexpress: an empirical Bayesian framework for discovering differential co-expression.** *Bioinformatics* 2012, **28**(14):1939–1940.
59. Li KC: **Genome-wide coexpression dynamics: theory and application.** *Proc Natl Acad Sci USA* 2002, **99:**16875–16880.
60. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The gene ontology consortium.** *Nature Genet* 2000, **25:**25–29.
61. Ackermann M, Strimmer K: **A general modular framework for gene set enrichment analysis.** *BMC Bioinformatics* 2009, **10:**47.
62. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y: **Gene-set analysis and reduction.** *Brief Bioinform* 2009, **10:**24–34.
63. Emmert-Streib F, Glazko G: **Pathway analysis of expression data: deciphering functional building blocks of complex diseases.** *PLoS Comput Biology* 2011, **7**(5):e1002053.
64. Khatri P, Sirota M, Butte A J: **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS Comput Biol* 2012, **8**(2):e1002375.
65. Liu Q, Dinu I, Adewale A, Potter J, Yasui Y: **Comparative evaluation of gene-set analysis methods.** *BMC Bioinformatics* 2007, **8:**431.
66. Goeman J, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics* 2007, **23**(8):980–7.
67. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucl Acids Res* 2009, **37:**1–13.
68. Mootha V, Lindgren C, Eriksson KFea: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nature Genet* 2003, **34:**267 – 273.
69. Efron B, Tibshiran R: **On testing the significance of sets of genes.** *Ann Appl Stat* 2007, **1:**107–129.
70. Dørum G, Snipen L, Solheim M, Sæbø S: **Rotation testing in gene set enrichment analysis for small direct comparison experiments.** *Stat App Genet Mol Biol* 2009, **8:**34.
71. Luo W, Friedman M, Shedden K, Hankenson K, Woolf P: **GAGE: generally applicable gene set enrichment for pathway analysis.** *BMC Bioinformatics* 2009, **10:**161.
72. Kim SY, Volsky D: **PAGE: Parametric Analysis of Gene Set Enrichment.** *BMC Bioinformatics* 2005, **6:**144.
73. Newton M, Quintana F, den Boon Jea: **Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis.** *Ann Appl Stat* 2007, **1:**85–106.
74. Freudenberg JM, Sivaganesan S, Phatak M, Shinde K, Medvedovic M: **Generalized random set framework for functional enrichment analysis using primary genomics datasets.** *Bioinformatics* 2011 Jan 1, **27**(1):70–7.
75. Rafael IA, Chi W, Yun Z, Terence SP: **Gene set enrichment analysis made simple.** *Stat Methods Med Res* 2009, **18**(6):565–575.
76. Lange K: *Numerical Analysis for Statisticians*. Statistics and Computing: Springer; 2010.
77. Pyeon D, Newton MA, Lambert PF, den Boon JA, Sengupta S, Marsit CJ, Woodworth CD, Connor JP, Haugen TH, Smith EM, Kelsey KT, Turek LP, Ahlquist P: **Fundamental differences in cell cycle deregulation in human Papillomavirus-positive and human Papillomavirus-negative head/neck and cervical cancers.** *Cancer Res* 2007, **67**(10):4605–4619.
78. Sheskin DJ: *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd edition. Boca Raton: RC Press; 2004.
79. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102**(38):13544–13549.
80. Jiang Z, Gentleman R: **Extensions to gene set enrichment.** *Bioinformatics* 2007, **23**(3):306–313.
81. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y: **Improving gene set analysis of microarray data by SAM-GS.** *BMC Bioinformatics* 2007, **8:**242.
82. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20:**93–99.
83. Hummel M, Meister R, Mansmann U: **GlobalANCOVA: exploration and assessment of gene group effects.** *Bioinformatics* 2008, **24:**78–85.
84. Lu Y, Liu P, Xiao P, Deng H: **Hotelling's T2 multivariate profiling for detecting differential expression in microarrays.** *Bioinformatics* 2005, **21**(14):3105-3113.
85. Kong S, Pu W, Park P: **A multivariate approach for integrating genome-wide expression data and biological knowledge.** *Bioinformatics* 2006, **22**(19):2373–2380.
86. Tsai C, Chen J: **Multivariate analysis of variance test for gene set analysis.** *Bioinformatics* 2009, **25**(7):897–903.
87. Xiong H: **Non-linear tests for identifying differentially expressed genes or genetic networks.** *Bioinformatics* 2006, **22**(8):919–923.
88. Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y: **A multivariate extension of the gene set enrichment analysis.** *J Bioinform Comput Biol* 2007, **5**(5):1139–1153.
89. Yates P, Reimers M: **RCMAT: a regularized covariance matrix approach to testing gene sets.** *BMC Bioinformatics* 2009, **10:**300.

90. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis.** *Genome Res* 2007, **17**(10):1537–1545.
91. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R: **A novel signaling pathway impact analysis.** *Bioinformatics* 2009, **25:**75–82.
92. Thomas R, Gohlke JM, Stopper GF, Parham FM, Portier CJ: **Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure.** *Genome Biol* 2009, **10**(4):R44.
93. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.** *Bioinformatics* 2010, **26**(12):i237—i245.
94. Massa M, Chiogna M, Romualdi C: **Gene set analysis exploiting the topology of a pathway.** *BMC Syst Biol* 2010, **4:**121.
95. Glazko G, Emmert-Streib F: **Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets.** *Bioinformatics* 2009, **25**(18):2348–2354.
96. Ledoit O, Wolf M: **Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.** *J Empir Finance* 2003, **10:**603–621.
97. Ledoit O, Wolf M: **A well conditioned estimator for largedimensional covariance matrices.** *J Multiv Anal* 2004, **88:**365–411.
98. Ledoit O, Wolf M: **Honey, I shrunk the sample covariance matrix.** *J Portfolio Manage* 2004, **30:**110–119.
99. Schäfer J, Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional Genomics.** *Stat Appl Genet Mol Biol* 2005, **4:**32.
100. Kanehisa M, Goto S: **KEGG: kyoto encyclopia of genes and genomes.** *Nuclei Acids Res* 2000, **28:**27–30.
101. Lauritzen S: *Graphical Models*. New York: Oxford Science Publications, Clarendon Press; 1996.
102. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic Acids Res* 2009, **37**(suppl 1):D619—D622.
103. Klebanov L, Jordan C, Yakovlev A: **A new type of stochastic dependence revealed in gene expression data.** *Stat Appl Genet Mol Biol* 2006, **5**(05/11):Article7.
104. Lim J, Kim J, Kim B: **An alternative model of type A dependence in a gene set of correlated genes.** *Stat Appl in Genet Mol Biol* 2010, **9:**Article 12.
105. Tripathi S, Emmert-Streib F: **Assessment method for a power analysis to identify differentially expressed pathways.** *PLoS ONE* 2012, **7**(5):e37510.
106. Emmert-Streib F: **The chronic fatigue syndrome: a comparative pathway analysis.** *J Comput Biol* 2007, **14**(7):961–972.
107. Choi Y, Kendziorski C: **Statistical methods for gene set co-expression analysis.** *Bioinformatics* 2009, **25**(21):2780–2786.
108. Cho SB, Kim J, Kim JH: **Identifying set-wise differential co-expression in gene expression microarray data.** *BMC Bioinformatics* 2009, **10:**109.
109. Tesson BM, Breitling R, Jansen RC: **DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules.** *BMC Bioinformatics* 2010, **11:**497.
110. Altay G, Asim M, Markowetz F, Neal DE: **Differential C3NET reveals disease networks of direct physical interactions.** *BMC Bioinformatics* 2011, **12:**296.
111. Watkinson J, Wang X, Zheng T, Anastassiou D: **Identification of gene interactions associated with disease from gene expression data using synergy networks.** *BMC Syst Biol* 2008, **2:**10.
112. Bunke H: **What is the distance between graphs?** *Bull EATCS* 1983, **20:**35–39.
113. Fuite J, Vernon S, Broderick G: **Neuroendocrine and immune network re-modeling in chronic fatigue syndrome: an exploratory analysis.** *Genomics* 2008, **92:**393–399.
114. Wang YC, Lan CY, Hsieh WP, Murillo L, Agabian N, Chen BS: **Global screening of potential Candida albicans biofilm-related transcription factors via network comparison.** *BMC Bioinformatics* 2010, **11:**53.
115. Gill R, Datta S, Datta S: **A statistical framework for differential network analysis from microarray data.** *BMC Bioinformatics* 2010, **11:**95.
116. Altay G, Emmert-Streib F: **Inferring the conservative causal core of gene regulatory networks.** *BMC Syst Biol* 2010, **4:**132.
117. Altay G, Emmert-Streib F: **Structural influence of gene networks on their inference: analysis of C3NET.** *Biol Direct* 2011, **6:**31.
118. Dempster A: **Covariance selection.** *Biometrics* 1972, **28:**157–175.
119. Koller D, Friedman N: *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: The MIT Press; 2009.
120. Whittaker J: *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley; 1990.
121. Li H, Gui J: **Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks.** *Biostatistics* 2006, **7**(2):302–317.
122. Schäfer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks.** *Bioinformatics* 2005, **21:**754–764.
123. Wille A, Zimmermann P, Vranova E, Furholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, Zitzler E, Gruissem W, Buhlmann P: **Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana.** *Genome Biol* 2004, **5**(11):R92.
124. Fan J, Feng Y, Wu Y: **Network exploration via the adaptive lasso and SCAD penalties.** *Ann Appl Stat* 2009, **3**(2):521–541.
125. Friedman J, Hastie T, Tibshirani R: **Sparse inverse covariance estimation with the graphical lasso.** *Biostatistics Oxford England* 2008, **9**(3):432–441.
126. Kiiveri H, de Hoog F: **Fitting very large sparse Gaussian graphical models.** *Comput Stat & Data Anal* 2012, **56**(9):2626–2636.
127. Witten DM, Friedman JH, Simon N: **New insights and faster computations for the graphical Lasso.** *J Comput Graphical Stat* 2011, **20**(4):892–900.
128. Zou H: **The adaptive Lasso and its oracle properties.** *J Am Stat Assoc* 2006, **101**(476):1418–1429.
129. Dudoit S, van der Laan, M: *Multiple Testing Procedures with Applications to Genomics*. New York: Springer; 2007.
130. Dudoit S, van der Laan M, Pollard K: **Multiple testing. part I. single-step procedures for control of general type I error rates.** *Stat App Genet Mol Biol* 2004, **3:**13.
131. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc, Ser B (Methodological)* 1995, **57:**125–133.
132. Storey J: **A direct approach to false discovery rates.** *J R Stat Soc, Ser B* 2002, **64:**479–498.
133. Aubert J, Bar-Hen A, Daudin J, Robin S: **Determination of the differentially expressed genes in microarray experiments using local FDR.** *BMC Bioinformatics* 2004, **5:**125.
134. Efron B: **Correlation and large-scale simultaneous signifance testing.** *J Am Stat Assoc* 2007, **102**(477):93–103.
135. Pounds S, Morris SW: **Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values.** *Bioinformatics* 2003, **19**(10):1236–1242.
136. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Stat* 2001, **29**(4):1165–1188.
137. Efron B: **Size, power and false discovery rates.** *Ann Stat* 2007, **35**(4):1351–1377.
138. Storey J, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci USA* 2003, **100**(16):9440–9445.
139. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C: **Gene set enrichment analysis: performance evaluation and usage guidelines.** *Briefings in Bioinformatics* 2012, **13**(3):281–291.
140. Nam D, Kim S: **Gene-set approach for expression pattern analysis.** *Brief Bioinform* 2008, **9**(3):189–197.
141. Weinberg RA: *The Biology of Cancer*. New York: Garland Science; 2007.
142. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS Genet* 2007, **3**(9):e161.
143. McClintick JN, Edenberg HJ: **Effects of filtering by Present call on analysis of microarray experiments.** *BMC Bioinformatics* 2006, **7:**49.
144. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments.** *Proc Natl Acad Sci USA* 2010, **107**(21):9546–9551.

145. Carter GW: **Inferring network interactions within a cell.** *Briefings in Bioinformatics* 2005, **6**(4):380–389.
146. Perou CM, Sorlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**(6797):747–752.
147. Gerstung M, Eriksson N, Lin J, Vogelstein B, Beerenwinkel N: **The temporal order of genetic and pathway alterations in Tumorigenesis.** *PLoS ONE* 2011, **6**(11):e27136.
148. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, *et al*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321**(5897):1801–1806.
149. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, *et al*: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**(5853):1108–1113.
150. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nature Med* 2004, **10**(8):789–799.
151. Lazebnik Y: **What are the hallmarks of cancer?** *Nature Rev Cancer* 2010, **10**(4):232–233.
152. Mukherjee S: *The Emperor of All Maladies: A Biography of Cancer*. London: Fourth Estate; 2011.

doi:10.1186/1745-6150-7-44
Cite this article as: Emmert-Streib *et al.*: Harnessing the complexity of gene expression data from cancer: from single gene to structural pathway methods. *Biology Direct* 2012 **7**:44.