More power via graph-structured tests for differential expression of gene networks

Laurent Jacob, Pierre Neuvial, Sandrine Dudoit

The Annals of Applied Statistics
2012, Vol. 6, No. 2, 561–600
DOI: 10.1214/11-AOAS528

MORE POWER VIA GRAPH-STRUCTURED TESTS FOR DIFFERENTIAL EXPRESSION OF GENE NETWORKS

BY LAURENT JACOB, PIERRE NEUVIAL AND SANDRINE DUDOIT

University of California, Berkeley, University of California, Berkeley, and Université d'Évry Val d'Essonne

We consider multivariate two-sample tests of means, where the location shift between the two populations is expected to be related to a known graph structure. An important application of such tests is the detection of differentially expressed genes between two patient populations, as shifts in expression levels are expected to be coherent with the structure of graphs reflecting gene properties such as biological process, molecular function, regulation or metabolism. For a fixed graph of interest, we demonstrate that accounting for graph structure can yield more powerful tests under the assumption of smooth distribution shift on the graph. We also investigate the identification of non-homogeneous subgraphs of a given large graph, which poses both computational and multiple hypothesis testing problems. The relevance and benefits of the proposed approach are illustrated on synthetic data and on breast and bladder cancer gene expression data analyzed in the context of KEGG and NCI pathways.

1. Introduction. The detection of differentially expressed (DE) genes, that is, genes whose expression levels change between two (or more) experimental conditions, remains a major challenge in biology and medicine, especially in the context of cancer studies. For example, the identification of DE genes between breast cancer patients that are sensitive or resistant to tamoxifen can help understand resistance mechanisms to this drug and eventually improve breast tumor treatment [Loi et al. (2008)]. Similarly, finding DE genes between low-grade, noninvasive or more aggressive bladder tumors may help understand the disease better and ultimately improve its diagnosis and treatment [Stransky et al. (2006)]. The application of the methods developed in this paper will be illustrated on the data sets from the above two papers.

However, the detection of a change in gene expression levels among a large gene list is a difficult problem from a statistical perspective, and lists of differentially expressed genes are generally hard to interpret, as they focus on the level of genes instead of the level of molecular functions. In such a context, expression data from high-throughput microarray and sequencing assays gain much in relevance from their association with graph-structured prior information on the genes, for

Received November 2010; revised October 2011.
Key words and phrases. Differential expression, biological networks, pathways, enrichment analysis, two-sample test, Hotelling T2, spectral graph theory, graph Laplacian, dimensionality reduction.

example, Gene Ontology (GO; http://www.geneontology.org), Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg) or NCI Pathway Integration Database (NCI graphs; http://pid.nci.nih.gov). Most approaches to the joint analysis of gene expression data and gene graph data involve two distinct steps. First, tests of differential expression are performed separately for each gene. Then, these univariate (gene-level) testing results are extended to the level of gene sets, for example, by assessing the over-representation of DE genes in each set based on p-values for Fisher's exact test[1] (or a χ^2 approximation thereof) adjusted for multiple testing [Beissbarth and Speed (2004)] or based on permutation adjusted p-values for weighted Kolmogorov–Smirnov-like statistics [Subramanian et al. (2005)]. Another family of methods directly performs multivariate tests of differential expression for groups of genes, for example, Hotelling's T^2-test [Lu et al. (2005)]. It is known [Goeman and Bühlmann (2007)] that the former family of approaches can lead to incorrect interpretations, as the sampling units for the tests in the second step become the genes (as opposed to the patients) and these are expected to have strongly correlated expression measures. This fact suggests that direct multivariate testing of gene set differential expression is more appropriate than posterior aggregation of individual gene-level tests. On the other hand, while Hotelling's T^2-statistic is known to perform well in small dimensions, it loses power very quickly with increasing dimension [Bai and Saranadasa (1996)], essentially because it is based on the inverse of the empirical covariance matrix which becomes ill-conditioned. Additionally, such direct multivariate tests on unstructured gene sets do not take advantage of information on gene regulation or other relevant biological properties. An increasing number of regulation networks are becoming available, specifying, for example, which genes activate or inhibit the expression of which other genes. If it is known that a particular gene in a tested gene set activates the expression of another, then one expects the two genes to have coherent (differential) expression patterns, for example, higher expression of the first gene in resistant patients should be accompanied by higher expression of the second gene in these patients. Accordingly, the first main contribution of this paper is to propose and validate multivariate test statistics for identifying differential expression patterns (or, more generally, shifts in distribution) that are coherent with a given graph structure.

Next, given a large graph and observations from two data generating distributions on the graph, a more general problem is the identification of smaller nonhomogeneous subgraphs, that is, subgraphs on which the two distributions (restricted to these subgraphs) are significantly different. This is very relevant in the context of tests for gene set differential expression: given a large set of genes, together with their known regulation network, or the concatenation of several such overlapping sets, it is important to discover novel gene sets whose expression changes significantly between two conditions. Currently-available gene sets have often been

[1] Sometimes referred to as a hypergeometric test in the bioinformatics literature.

defined in terms of other phenomena than that under study and physicians may be interested in discovering sets of genes affecting in a concerted manner a specific phenotype. Our second main contribution is therefore to develop algorithms that allow the exhaustive testing of all the subgraphs of a large graph, while avoiding one-by-one enumeration and testing of these subgraphs and accounting for the multiplicity issue arising from the vast number of subgraphs.

As the problem of identifying variables or groups of variables which differ in distribution between two populations is closely related to supervised learning, our proposed approach is similar to several learning methods. Rapaport et al. (2007) use filtering in the Fourier space of a graph to train linear classifiers of gene expression profiles whose weights are smooth on a gene network. However, their classifier enforces global smoothness on the large regularization network of all the genes, whereas we are concerned with the selection of gene sets with locally-smooth expression shift between populations. In Jacob, Obozinski and Vert (2009) and Obozinski, Jacob and Vert (2011), sparse learning methods are used to build a classifier based on a small number of gene sets. While this approach leads in practice to the selection of groups of variables whose distributions differ between the two classes, the objective is to achieve the best classification performance with the smallest possible number of groups. As a result, correlated groups of variables are typically not selected. Other related work includes Fan and Lin (1998), who proposed an adaptive Neyman test in the Fourier space for time series. However, as illustrated below in Section 5, direct translation of the adaptive Neyman statistic to the graph case is problematic, as assumptions on Fourier coefficients which are true for time series do not hold for graphs. In addition, the Neyman statistic converges very slowly toward its asymptotic distribution and the required calibration by bootstrapping renders its application to our subgraph discovery context difficult. By contrast, other methods do not account for shift smoothness and try to address the loss of power caused by the poor conditioning of the T^2-statistic by applying it after dimensionality reduction [Ma and Kosorok (2009)] or by omitting the inverse covariance matrix and adjusting instead by its trace [Bai and Saranadasa (1996), Chen and Qin (2010)] or using a diagonal estimator of the covariance matrix [Srivastava and Du (2008), Srivastava (2009)]. Lopes, Jacob and Wainwright (2011) recently proposed a testing procedure based on random projection of the data in a lower dimension space, and showed that it was asymptotically more powerful than Bai and Saranadasa (1996), Chen and Qin (2010) and Srivastava and Du (2008) in the presence of correlation and when the spectrum of the covariance matrix decays fast enough. Vaske et al. (2010) recently proposed DE tests, where a probabilistic graphical model is built from a gene network. However, this model is used for gene-level DE tests, which then have to be combined to test at the level of gene sets. Several approaches for subgraph discovery, like that of Ideker et al. (2002), are based on a heuristic to identify the most differentially expressed subgraphs and do not amount to testing exactly all possible subgraphs. Concerning the discovery of distribution-shifted subgraphs, Vandin, Upfal and Raphael (2010)

propose a graph Laplacian-based testing procedure to identify groups of interacting proteins whose genes contain a large number of mutations. Their approach does not enforce any smoothness on the detected patterns (smoothness is not necessarily expected in this context) and the graph Laplacian is only used to ensure that very connected genes do not lead to spurious detection. The Gene Expression Network Analysis (GXNA) method of Nacu et al. (2007) detects differentially expressed subgraphs based on a greedy search algorithm and gene set DE scoring functions that do not account for the graph structure.

The rest of this paper is organized as follows. Section 2 explains how to build a lower-dimension basis in which to apply the multivariate test of means. Section 3 presents our graph-structured two-sample test statistic and states results on power gain for smooth-shift alternatives. Section 4 describes procedures for systematically testing (without fully enumerating) all possible subgraphs of a large graph. Section 5 presents results for synthetic data and Section 6 on breast and bladder cancer gene expression data sets analyzed in the light of pathways from the KEGG and NCI databases. Section 7 presents softwares implementing the proposed methods. Finally, Section 8 summarizes our findings and outlines ongoing work.

Although this work is motivated by the specific question of differential expression testing of gene networks, our proposed structured two-sample test of means on a graph and our nonhomogeneous subgraph discovery algorithm can actually be used in any situation where one searches for differences between two populations that are expected to be coherent with a known graph structure. Therefore, our methodological contributions in Sections 3 and 4 are presented in the general context of two-sample tests on graphs.

2. Graph-based dimensionality reduction. As stated in the Introduction, each of the two main paradigms for testing differential expression of a gene set have their limitations. Two-step methods generally do not directly test the existence of a mean shift between two multivariate distributions [Goeman and Bühlmann (2007)]. The second step, which often treats the *genes* as the sampling units, renders the interpretation of p-values problematic and may lead to a large loss of power or Type I error control when sets of genes have correlated expression. Multivariate statistics, on the other hand, allow a direct formulation of and solution to the testing question: the sampling units are vectors of gene expression measures (e.g., corresponding to patients) and the question is whether two such sets of random vectors are likely to have arisen from distributions with equal means. Figure 1 illustrates another classical advantage of multivariate approaches: genes taken individually may have extremely small mean shifts between two populations, although their joint distributions clearly differ between the two populations. Here, again, this phenomenon typically happens for sets of genes whose expression measures are correlated, which is not unlikely for pathways or annotated gene sets.

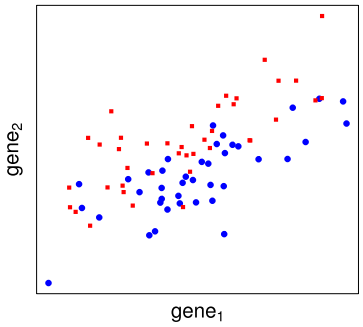

FIG. 1. *Synthetic example of the joint distribution of the expression measures of two genes in two patient populations. The color and shape of the plotting symbols indicate the patient group and the x- and y-axes correspond to the expression measures of the first and second gene, respectively.*

Unfortunately, with moderate sample sizes, multivariate statistics lose power quickly in a high dimension. If some type of side information is available regarding particular properties of the expression shift, a possible approach to get the best of both worlds would be to: (1) project the vectors of covariates in a new space of *lower dimension* that preserves the distribution shift, that is, the distance between the expression measures of the two groups, and (2) apply the multivariate statistic in this new space. One could thus perform the appropriate multivariate test, while avoiding the loss of power caused by the high-dimensionality of the original covariate space.

A possible source of information about the expression shift is the growing number of available gene networks. Indeed, while the difference in mean expression between two groups of patients may not be entirely coherent with an existing network (e.g., because of noise in the data, errors in the annotation, or inappropriateness of the chosen network for the biological question of interest), it is reasonable to expect that this shift will not be entirely contradictory with the given graph structure. For example, repressed genes should be more connected to other repressed genes than to overexpressed genes. Given this assumption, we intend to build a space of lower dimension than the original gene space, but which preserves most of the distribution shift between the two populations.

More precisely, consider a network of p genes, represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = p$ nodes and edge set $\mathcal{E}$. Let $\delta \in \mathbb{R}^p$ denote the mean shift, that is, the vector of differences in mean expression measures for these p genes between the two populations of interest. Suppose we expect the shift δ to be coherent with the graph $\mathcal{G}$, in the sense that it has low energy $E_{\mathcal{G}}(\delta)$ for a particular energy function $E_{\mathcal{G}}$ defined on $\mathcal{G}$. Then, we wish to build a space of lower dimension $k \ll p$ capturing most of the low energy functions. To this end, we start by finding the function that has the lowest possible energy, then the function that has lowest possible energy in the orthogonal space of the first one, up to the kth function with

lowest energy in the orthogonal subspace of the first $k-1$ functions. That is, for each $i \leq k$, we define

$$u_i = \begin{cases} \underset{f \in \mathbb{R}^p}{\arg\min}\, E_{\mathcal{G}}(f) \\ \text{such that } u_i \perp u_j,\ j < i. \end{cases} \tag{2.1}$$

If $E_{\mathcal{G}}$ is a positive semi-definite quadratic form $E_{\mathcal{G}}(\delta) = \delta^\top Q_{\mathcal{G}} \delta$, for some positive semi-definite matrix $Q_{\mathcal{G}} = U \Lambda U^\top$, where U is an orthogonal matrix and Λ a diagonal matrix with elements $\lambda_i, i = 1, \ldots, p$, then the solution to equation (2.1) is given by the k eigenvectors of $Q_{\mathcal{G}}$ corresponding to the smallest k eigenvalues. It is easy to check that these eigenvalues are the energies of the corresponding functions u_i, that is, $E_{\mathcal{G}}(u_i) = \lambda_i$.

Different choices of $Q_{\mathcal{G}}$ lead to different notions of coherence of the expression shift with the network. A classical choice is the *graph Laplacian* $\mathcal{L}$. Suppose $\mathcal{G}$ is an *undirected* graph with adjacency matrix A, with $a_{ij} = 1$ if and only if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise, and degree matrix $D = \mathrm{Diag}(A\mathbf{1})$, where $\mathbf{1}$ is a unit column-vector, $\mathrm{Diag}(x)$ is the diagonal matrix with diagonal x for any vector x, and $D_{ii} = d_i$. The Laplacian matrix of $\mathcal{G}$ is then typically defined as $\mathcal{L} = D - A$ or $\mathcal{L}_{\mathrm{norm}} = I - D^{-1/2} A D^{-1/2}$ for the normalized version, leading to energies $\sum_{i,j \in \mathcal{V}} (\delta_i - \delta_j)^2$ and $\sum_{i,j \in \mathcal{V}} (\frac{\delta_i}{\sqrt{d_i}} - \frac{\delta_j}{\sqrt{d_j}})^2$, respectively. Note that, in this case, the Laplacian matrix $\mathcal{L}$, energy E and basis functions u_i extend the classical Fourier analysis of functions on Euclidean spaces to functions on graphs, by transferring the notions of Laplace operator, Dirichlet energy and Fourier basis, respectively [Evans (1998)].

More generally, any positive semi-definite matrix can be chosen. In the case of gene regulation networks, we do not necessarily expect as strong a coherence as that corresponding to the Dirichlet energy defined by the graph Laplacian, since some of the annotated interactions may not be relevant in the studied context and some antagonist interactions may cancel each other. For example, if a gene is activated by two others, one who is underexpressed and the other overexpressed, we may observe no change in the expression of the gene, but a nonzero Dirichlet energy $\sum_{i,j \in \mathcal{V}} (\delta_i - \delta_j)^2$. Additionally, for applications like structured gene set differential expression detection, one may use negative weights for edges that reflect a negative correlation between two variables, for example, a gene i whose expression inhibits the expression of another gene j. In this case, a small variation of the shift on the edge between i and j should correspond to a small $|\delta_i + \delta_j|$. This can be achieved in the same formalism by simply considering a *signed* version of the adjacency matrix A, that is, $a_{ij} = 1$ if gene i activates gene j and -1 if it inhibits gene j. A signed version of the graph Laplacian is then $\mathcal{L}_{\mathrm{sign}} = D - A$, where $D = \mathrm{Diag}(|A|\mathbf{1})$ is the degree matrix and $|A|$ denotes the entry-wise absolute value of A. Note that such a signed Laplacian was used as a penalty for semi-supervised learning in Goldberg (2007).

In the context of this work, we, moreover, consider *directed* graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the edge set $\mathcal{E}$ consists of ordered pairs of nodes. The adjacency matrix A may be asymmetric, with entries $a_{ij} \neq 0$ if and only if $(i, j) \in \mathcal{E}$, that is, there is an (directed) edge pointing from node v_i to node v_j. We then use the following energy function:

$$E_{\mathcal{G}}(\delta) = \sum_{i:d_i^- \neq 0}^{p} \left(\delta_i - \frac{1}{d_i^-} \sum_{(j,i)\in\mathcal{E}} a_{ji}\delta_j \right)^2, \tag{2.2}$$

where $d_i^- \triangleq \sum_{j=1}^{p} |a_{ji}|$ is the indegree of node v_i, that is, the number of directed edges pointing from any node to v_i. According to this definition, an expression shift will have low energy if the difference in mean expression of any given gene between the two populations is similar to the (signed) average of the differences in mean expression for the genes that either activate or inhibit it.

It is immediate to check that $E_{\mathcal{G}}(\delta) = \delta^\top M_{\mathcal{G}} \delta$, with $M_{\mathcal{G}} \triangleq (\tilde{I} - D_-^{-1}A^\top)^\top(\tilde{I} - D_-^{-1}A^\top)$, where $D_- \triangleq \mathrm{Diag}((d_i^-)_{i=1,\ldots,p})$ is the matrix of indegrees, $\tilde{I} \triangleq \mathrm{Diag}((\mathrm{I}(d_i^- \neq 0))_{i=1,\ldots,p})$ is a modification of the identity matrix where diagonal elements corresponding to nodes with zero indegree are set to zero, and the value of the indicator function I is 1 if its argument is true and zero otherwise. Note that a very similar function was used in the context of regularized supervised learning by Sandler et al. (2009).

Following our principle to build a lower dimension space, we use the first few eigenvectors of $M_{\mathcal{G}}$ to obtain orthonormal functions with low energy. As an example, Figure 2 displays the eigenvectors of $M_{\mathcal{G}}$ for a simple four-node graph with

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \tag{2.3}$$

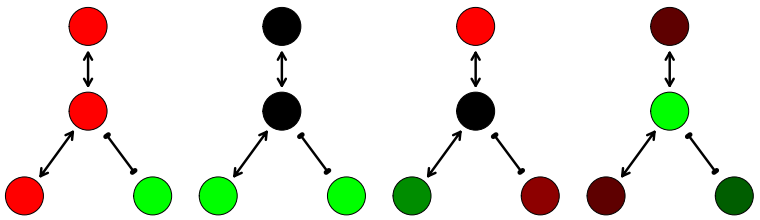

FIG. 2. *Eigenvectors of the signed Laplacian* $\mathcal{L}_{\mathrm{sign}}$ *for the simple undirected four-node graph of example* (2.3). *The eigenvectors of* $M_{\mathcal{G}}$ *for this particular network are the same. The corresponding eigenvalues are* $0, 1, 1, \frac{16}{3}$ *for* $M_{\mathcal{G}}$ *and* $0, 1, 1, 4$ *for* $\mathcal{L}_{\mathrm{sign}}$. *Nodes are colored according to the value of the eigenvector, where green corresponds to high positive values, red to high negative values, and black to* 0. *"T"-shaped edges have negative weights.*

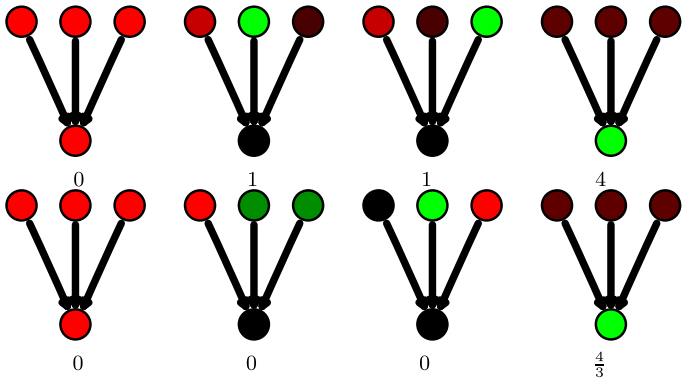

FIG. 3. *Eigenvectors of the signed Laplacian* $\mathcal{L}_{\text{sign}}$ (*top*) *and of* $M_{\mathcal{G}}$ (*bottom*) *for the simple directed four-node graph of example* (2.4). *The corresponding eigenvalues are* $0, 1, 1, 4$ *and* $0, 0, 0, \frac{4}{3}$, *respectively. Nodes are colored according to the value of the eigenvector, where green corresponds to high positive values, red to high negative values, and black to* 0.

where A takes on negative values for negative interactions, such as expression inhibition. The first eigenvector, corresponding to the smallest energy (eigenvalue of zero), can be viewed as a "constant" function on the graph, in the sense that its absolute value is identical for all nodes, but nodes connected by an edge with negative weight take on values of opposite sign. By contrast, the last eigenvector, corresponding to the highest energy, is such that nodes connected by positive edges take on values of opposite sign and nodes connected by negative edges take on values of the same sign. Note that, for this particular example, the adjacency matrix is symmetric, which need not always be the case. Here, the signed Laplacian turns out to have the same eigenvectors as $M_{\mathcal{G}}$, which is not the case generally.

Consider now a slightly different graph, with directed edges, only positive interactions to avoid confusion, and adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{2.4}$$

For this graph, Figure 3 shows that the two notions of energy lead to two different bases. While the signed Laplacian matrix (by definition based on a symmetrized version of A for an undirected graph) has only one (constant) eigenvector of null energy, two of energy 1, and one of 4, $M_{\mathcal{G}}$ has three orthogonal vectors of null energy. Note, however, that the first and last eigenvectors are still the same across the two bases.

More generally, this illustration suggests that projecting on the first eigenvectors of $M_{\mathcal{G}}$ will not preserve the same shifts as projecting on the first eigenvectors of $\mathcal{L}_{\text{sign}}$. It is possible for a shift vector to have low energy (2.2) but larger signed

Dirichlet energy $\sum_{i,j\in\mathcal{V}}(\delta_i - a_{ij}\delta_j)^2$, where we recall that a_{ij} is 1 for an edge indicating a positive interaction between i and j and -1 for an edge indicating a negative interaction. This is, for example, the case of the second eigenvector of $M_{\mathcal{G}}$ on the bottom row of Figure 3. It is therefore conceivable that such a shift essentially lies in the space spanned by the first few eigenvectors of $M_{\mathcal{G}}$, but that its projection in the space formed by the first few eigenvectors of $\mathcal{L}_{\text{signed}}$ is smaller. As a consequence, for a particular shift using one basis or the other for dimensionality reduction will lead to more or less gain in power, which means that the choice of basis should be adapted to the expected type of smoothness of the shift.

While we introduce the idea in the context of gene regulation networks and testing for differential expression, the same dimensionality reduction principle applies to any multivariate testing problem for which the variables have a known structure, as represented by a graph.

As a last remark, we emphasize that our requirement that the shift be coherent with the network is not too strict in practice. It may sound like most pairs of nodes must have shifts whose directions are consistent with the nature of the edge connecting the nodes, but:

- In practice, keeping a few eigenvectors already allows to represent several types of shifts which are not perfectly coherent with the network, as illustrated on Figure 2. The projection only shrinks those shifts which severely contradict the prior given by the network.
- In Section 5.2 we illustrate the fact that this type of projection still leads to gain in power even in case of strong misspecifications in the network, that is, when a lot of edges are missing or wrong.
- Lopes, Jacob and Wainwright (2011) show that in a high dimension, *random projection* of the data in a lower dimension space yields gains in power against the regular Hotelling T^2 in the presence of correlation and if the spectrum of the covariance matrix decays fast enough. This result suggests that there is hope to gain power even in the case where the network doesn't bring much information about the shift.

In the remainder of this paper we denote by $\tilde{f} = U^\top f$ the coefficients of a vector $f \in \mathbb{R}^{|\mathcal{V}|}$ after projection on a basis U (typically the eigenvectors of a $Q_{\mathcal{G}}$ matrix).

3. Graph-structured two-sample test of means under smooth-shift alternatives. For multivariate normal distributions, Hotelling's T^2-test, a classical test of location shift, is known to be a uniformly most powerful invariant against global-shift alternatives. The test statistic is based on the squared *Mahalanobis norm* of the sample mean shift and is given by $T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^\top \hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2)$, where n_i, $\bar{x}_i$ and $\hat{\Sigma}$ denote, respectively, the sample sizes, means and pooled covariance matrix, for random samples drawn from two p-dimensional Gaussian

distributions, $\mathcal{N}(\mu_i, \Sigma)$, $i = 1, 2$. Under the null hypothesis $\mathbf{H}_0 : \mu_1 = \mu_2$ of equal means, NT^2 follows a (central) F-distribution $F_0(p, n_1 + n_2 - p - 1)$, where $N = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p}$. In general, NT^2 follows a noncentral F-distribution $F(\frac{n_1 n_2}{n_1+n_2}\Delta^2(\delta, \Sigma); p, n_1 + n_2 - p - 1)$, where the noncentrality parameter is a function of the Mahalanobis norm of the mean shift δ, $\Delta^2(\delta, \Sigma) = \delta^\top \Sigma^{-1}\delta$, which we refer to as the *distribution shift*. In the remainder of this paper, unless otherwise specified, T^2-statistics are assumed to follow the nominal F-distribution, for example, for critical value and power calculations.

For any orthonormal basis U and, in particular, for our graph-based basis, direct calculation shows that $T^2 = \tilde{T}^2 \triangleq \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^\top U(U^\top \hat{\Sigma} U)^{-1} U^\top (\bar{x}_1 - \bar{x}_2)$, that is, the statistic T^2 in the original space and the statistic $\tilde{T}^2$ in the new graph-based space are identical. More generally, for $k \le p$, the statistic in the original space after filtering out dimensions above k is the same as the statistic $\tilde{T}_k^2$ restricted to the first k coefficients in the new space defined by U:

$$\begin{aligned}\tilde{T}_k^2 &\triangleq \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^\top U_{[k]}(U_{[k]}^\top \hat{\Sigma} U_{[k]})^{-1} U_{[k]}^\top (\bar{x}_1 - \bar{x}_2) \\ &= \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^\top U 1_k U^\top (U 1_k U^\top \hat{\Sigma} U 1_k U^\top)^+ U 1_k U^\top (\bar{x}_1 - \bar{x}_2),\end{aligned}$$

where A^+ denotes the generalized inverse of a matrix A, the $p \times k$ matrix $U_{[k]}$ denotes the restriction of U to its first k columns, and 1_k is a $p \times p$ diagonal matrix, with ith diagonal element equal to one if $i \le k$ and zero otherwise. Note that, as retaining the first k dimensions corresponds to a *noninvertible* transformation, this filtering indeed has an effect on the test statistic, that is, we have $\tilde{T}_k^2 \ne \tilde{T}^2$ in general. As the Mahalanobis norm is invariant to invertible linear transformations, using an invertible filtering (such as weighting each component according to its corresponding eigenvalue) would have no impact on the test statistic.

Hotelling's T^2-test is known to behave poorly in the high dimension; Lemma 1 stated and proved in the supplemental article Supplement A [Jacob, Neuvial and Dudoit (2011a)] shows that gains in power can be achieved by filtering. Specifically, let $\tilde{\delta} = U^\top \delta$ and $\tilde{\Sigma} = U^\top \Sigma U$ denote, respectively, the mean shift and covariance matrix in the new space. Given $k \le p$, let $\Delta_k^2(\delta, \Sigma) = \delta_{[k]}^\top (\Sigma_{[k]})^{-1}\delta_{[k]}$ denote the distribution shift restricted to the first k dimensions of δ and Σ, that is, based on only the first k elements of δ, $(\delta_i : i \le k)$, and the first $k \times k$ diagonal block of Σ, $(\sigma_{ij} : i, j \le k)$. Under the assumption that the distribution shift is smooth, that is, lies mostly in the first few graph-based coefficients, so that $\Delta_k^2(\tilde{\delta}, \tilde{\Sigma})$ is nearly maximal for a small value of k, Lemma 1 states that performing Hotelling's test in the new space restricted to its first k components yields more power than testing in the entire new space. Equivalently, the test is more powerful in the original space after filtering than in the original unfiltered space. The increase in shift $\eta(\alpha, k, l)$

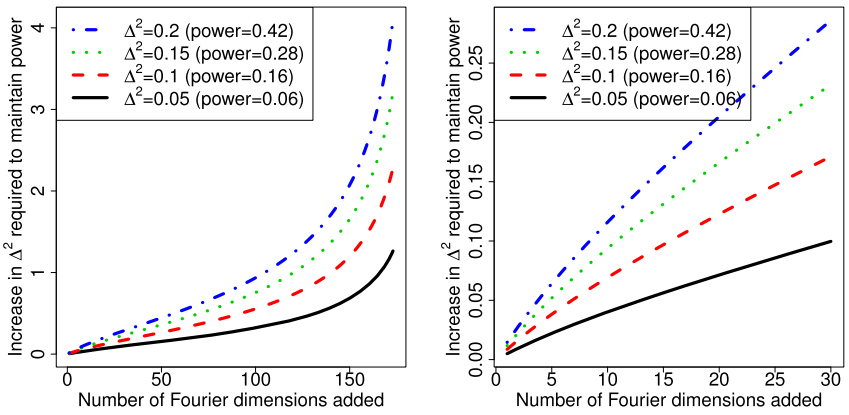

FIG. 4. *Left*: *Increase in distribution shift required for Hotelling's* T^2*-test to maintain a given power when increasing the number of tested new coefficients*: $\Delta^2_{k+l} - \Delta^2_k$ *vs.* l *such that* $\beta_{\alpha,k+l}(\Delta^2_{k+l}) = \beta_{\alpha,k}(\Delta^2_k)$. *Power* $\beta_{\alpha,k+l}(\Delta^2_{k+l})$ *computed under the noncentral* F*-distribution* $F(\frac{n_1 n_2}{n_1+n_2}\Delta^2_{k+l}; k+l, n_1+n_2-(k+l)-1)$, *for* $n_1 = n_2 = 100$ *observations*, $k = 5$, *and* $\alpha = 10^{-2}$. *Each line corresponds to the fixed shift* Δ^2_k *and power* $\beta_{\alpha,k}(\Delta^2_k)$ *pair indicated in the legend. Right*: *Zoom on the first* 30 *dimensions*.

required to maintain power when increasing dimension can be evaluated numerically for any (α, k, l). Note that this result holds because retaining the first k new components is a *noninvertible* transformation.

Corollary 1 in the supplemental article Supplement A [Jacob, Neuvial and Dudoit (2011a)] states that if the distribution shift lies in the first k new coefficients, then testing in this subspace yields strictly more power than using additional coefficients. In particular, if there exists $k < p$ such that $\tilde{\delta}_j = 0\ \forall j > k$ (i.e., the mean shift is smooth) and $\tilde{\Sigma}$ is block-diagonal such that $\tilde{\sigma}_{ij} = 0\ \forall i < k, j > k$, then gains in power are obtained by testing in the first k new components. Although nonnecessary, this condition is plausible when the mean shift lies at the beginning of the spectrum (i.e., has low energy), as the coefficients which do not contain the shift are not expected to be correlated with the ones that do contain it.

Note that the result in Lemma 1 is more general, as testing in the first k new components can increase power even when the distribution shift partially lies in the remaining components, as long as the latter portion is below a certain threshold. Figure 4 illustrates, under different settings, the increase in distribution shift necessary to maintain a given power level against the number of added coefficients.

Under the assumption of block-diagonal covariance, Corollary 2 (in the supplemental article Supplement A [Jacob, Neuvial and Dudoit (2011a)]) directly relates the energy of the mean shift vector to the gain in power. It states that if the energy of the mean shift vector δ is small enough, that is, if the mean shift is coherent enough with the network, then testing in the first k dimensions of the new basis is more powerful than testing in the original space. The corresponding upper

bound on the mean shift energy can be quantified for a given generative setting (μ_1, μ_2, Σ), graph $\mathcal{G}$ and level α. Tighter and looser bounds can be straightforwardly derived using the same principle for the diagonal and general covariance cases, respectively.

4. Nonhomogeneous subgraph discovery. A systematic approach for discovering nonhomogeneous subgraphs, that is, subgraphs of a large graph that exhibit a significant shift in means, is to test all of them one-by-one.

This poses a huge combinatorial problem even for moderately large graphs ($p = 50$, say), as the number of (connected) subgraphs of size k of a graph of size p can be exponential in p and k. Exhaustive search is therefore not feasible in practice, especially for differential expression on gene networks, where p is typically in the dozens or hundreds of genes. Therefore, it is important to rapidly identify sets of subgraphs that all satisfy the null hypothesis of equal means. To this end, we prove an upper bound on the value of the test statistic for any subgraph containing a given set of nodes (Lemma 2 in the supplemental article Supplement A [Jacob, Neuvial and Dudoit (2011a)]). An exact algorithm is derived from this upper bound in Section 4.1, and a quicker, approximate algorithm is proposed in Section 4.2.

4.1. *Exact algorithm.* Given a large graph $\mathcal{G}$ with p nodes, we adopt a branch-and-bound-like approach [Land and Doig (1960)] to test subgraphs of size $q \leq p$ at level α, as described in Algorithm 1. We start by checking, for each node in $\mathcal{G}$, whether the Hotelling T^2-statistic in the first k new components of any subgraph of size q containing this node can be guaranteed (by virtue of Lemma 2) to be below the level-α critical value $T^2_{\alpha,k}$, for example, $(1-\alpha)$-quantile of $F_0(k, n_1 + n_2 - k - 1)$ distribution. If this is the case, the node is pruned, that is, removed from the graph. The algorithm iteratively enriches a list of pruned subgraphs and a list of candidate subgraphs (called `prunedSubgraphs` and `currentSubgraphs` in Algorithm 1, resp.) of increasing number of nodes $s = 1, \ldots, q-1$. Pruned subgraphs are those for which one can guarantee that no supergraph of size q can reach significance level α, and candidate subgraphs are those for which this guarantee cannot be given. The key of the algorithm is that at step s, only those graphs containing a candidate subgraph have to be considered.

This guarantee is obtained by applying Lemma 2 in the supplemental article Supplement A [Jacob, Neuvial and Dudoit (2011a)], which gives an upper bound on the value of the test statistic for any subgraph containing a given set of nodes. For a subgraph g of $\mathcal{G}$ of size $q \leq p$, Hotelling's T^2-statistic in the first $k \leq q$ new components of g is defined as

$$\tilde{T}^2_k(g) = \frac{n_1 n_2}{n_1 + n_2} \big(\bar{x}_1(g) - \bar{x}_2(g)\big)^\top U_{[k]} \big(U_{[k]}^\top \hat{\Sigma}(g) U_{[k]}\big)^{-1} U_{[k]}^\top \big(\bar{x}_1(g) - \bar{x}_2(g)\big),$$

where $U_{[k]}$ is the $q \times k$ restriction of the matrix of q eigenvectors of Q_g to its first k columns [i.e., $U_{[k]}(g)$, where we omit g to ease notation] and $\bar{x}_i(g), i =$

Algorithm 1: Nonhomogeneous subgraph discovery algorithm. The `subgraphBoundary` of a subgraph g of $\mathcal{G}$ is defined as the set of supergraphs of g obtained by adding any one node of $\mathcal{G}$ which is connected to a node of g.

Input: $\mathcal{G}, X_1, X_2, \alpha, q$
Output: selectedSubgraphs

selectedSubgraphs $= \varnothing$;
previousSubgraphs = `nodes` $(\mathcal{G})$;
prunedSubgraphs $= \varnothing$;
foreach $s \in \{1 \ldots q-1\}$ **do**
  checkedSubgraphs $= \varnothing$;
  foreach previousSubgraph **do**
    foreach subgraph $\in$ `subgraphBoundary`(previousSubgraph) **do**
      if subgraph *has been checked or has a pruned subgraph* **then** next;
      if $s < q-1$ **then**
        if `upperBound`(subgraph, $\mathcal{G}, X_1, X_2, q) < T^2_{\alpha,k}$ **then**
          add subgraph to prunedSubgraphs;
        else
          add subgraph to currentSubgraphs;
        end
      else
        foreach q-subgraph $\in$ `subgraphBoundary`(subgraph) **do**
          if q-subgraph *has been checked or has a pruned subgraph* **then**
            next
          else
            if $\tilde{T}^2_k$(q-subgraph, $X_1, X_2) > T^2_{\alpha,k}$ **then**
              add q-subgraph to selectedSubgraphs
            end
            add q-subgraph to checkedSubgraphs
          end
        end
      end
      add subgraph to checkedSubgraphs
    end
  end
  set previousSubgraphs to currentSubgraphs
end

1, 2, and $\hat{\Sigma}(g)$ are, respectively, the empirical means and pooled covariance matrix restricted to the nodes in g. Lemma 2 states that for any number k of retained components, and for any subgraph g' of size q' of g, $\tilde{T}^2_k(g)$ is upper bounded by the T^2 statistic of the subgraph whose nodes are in $\nu(g', q-q')$, that is, the union of the nodes of g' and the nodes of g whose shortest path to a node of g' is less than

or equal to r. The set $v(g', r)$ is called the r-neighborhood of g'. As a corollary of Lemma 2, the subgraphs returned by Algorithm 1 are exactly those who exhibit a significant shift in means at the prescribed level α.

Note that the bound in Lemma 2 takes into account the fact that the T^2-statistic is eventually computed in the first few components of a basis which is not known beforehand: at each step, for each potential subgraph g' which would include the subgraph g which we consider for pruning, the $\tilde{T}_k^2(g')$ that needs to be bounded above depends on $Q_{g'}$.

4.2. *Mean-shift approximation.* For "small-world" graphs above a certain level of connectivity and q large enough, $v(g', q-s)$, the $(q-s)$-neighborhood of g', tends to be large, at least at the beginning of the above exact algorithm, and the number of tests actually performed may not decrease much compared to the total number of possible tests. One can, however, identify much more efficiently the subgraphs whose sample mean shift in the first k components of the new space has Euclidean norm $\|\hat{\tilde{\delta}}_{[k]}(g)\| = \|U_{[k]}^\top(\bar{x}_1(g) - \bar{x}_2(g))\|$ above a certain threshold. Indeed, it is straightforward to see that

$$\begin{aligned}
&\left\|U_{[k]}^\top\left(\bar{x}_1(g) - \bar{x}_2(g)\right)\right\|^2 \\
&\quad \le \left\|U^\top\left(\bar{x}_1(g) - \bar{x}_2(g)\right)\right\|^2 \\
&\quad = \|\bar{x}_1(g) - \bar{x}_2(g)\|^2 \\
&\quad \le \|\bar{x}_1(g') - \bar{x}_2(g')\|^2 \\
&\qquad + \max_{v_1,\ldots,v_{q-s}\in v(g',q-s)} \|\bar{x}_1(v_1,\ldots,v_{q-s}) - \bar{x}_2(v_1,\ldots,v_{q-s})\|^2.
\end{aligned}$$

Using this inequality yields an upper bound on $\tilde{T}_k^2(g)$ that can be used as `upperBound` at line 10 of Algorithm 1. This defines a procedure that identifies all subgraphs for which the Euclidean norm of the sample mean shift exceeds a given threshold: $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2 > \theta$. For any α, if this threshold θ is low enough, all the subgraphs with $\tilde{T}_k^2(g) > T_{\alpha,k}^2$ are included in this set. Performing the actual T^2-test on these preselected subgraphs then yields exactly the set of subgraphs that would have been identified using the exact procedure of Section 4.1. More precisely, Lemma 4, in the supplemental article Supplement A [Jacob, Neuvial and Dudoit (2011a)], states that for any subgraph which would be detected by Hotelling's T^2-statistic $\tilde{T}_k^2(g)$ but not by the Euclidean criterion $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2$, the sample covariance matrix in the restricted new space (after filtering) has an eigenvalue below a certain threshold. This implies that such false negative subgraphs (from the Euclidean approximation to the exact algorithm) have a small mean shift in the new space, but in a direction of small variance. In the context of gene expression, this is related to the well-known issue of the detection of DE genes by

virtue of their small variances. Even though the differences in expression appear to be significant for these genes, they correspond to small effects that may not be interesting from a practical point of view (i.e., biologically nonsignificant). Methods for addressing this problem are proposed in Lönnstedt and Speed (2002). Note that $\lambda_{\min}(\hat{\Sigma}(g)) \leq \lambda_{\min}(\hat{\tilde{\Sigma}}_{[k]}(g))$; thus, the remark on variances holds for both the new and the original spaces. However, if q is large, we expect $\lambda_{\min}(\hat{\Sigma}(g))$ to be very small, while filtering somehow controls the conditioning of the covariance matrix.

4.3. *Multiple hypothesis testing.* Testing for homogeneity over the potentially large number of subgraphs investigated as part of the above algorithms immediately raises the issue of multiple testing. However, because one does not know in advance the total number of tests and which tests will be performed specifically, standard multiple testing procedures, such as those in Dudoit and van der Laan (2008), are not immediately applicable.

In an attempt to address the multiplicity issue, we apply a permutation procedure to control the number of false positive subgraphs under the complete null hypothesis of identical distributions in the two populations. Specifically, one permutes the class/population labels (1 or 2) of the $n_1 + n_2$ observations and applies the nonhomogeneous subgraph discovery algorithm to the permuted data to yield a certain number of false positive subgraphs. Repeating this procedure a sufficiently large number of times produces an estimate of the distribution of the number of Type I errors under the complete null hypothesis of identical distributions.

We evaluate the empirical behavior of the procedures proposed in Sections 3 and 4, first on synthetic data, then on breast cancer microarray data analyzed in the context of KEGG pathways.

5. Results on synthetic data. The performance of the graph-structured test is assessed in cases where the distribution shift Δ^2 satisfies the smoothness assumptions described in Section 3. We first generate a connected random graph $\mathcal{G}$ with $p = 20$ nodes and 20 edges. Next, we generate 10,000 data sets in the space corresponding to the basis U defined by the eigenvectors of the $Q_{\mathcal{G}}$ matrix for the graph $\mathcal{G}$; an inverse transformation is applied to random vectors generated is this new space. Each data set comprises $n_1 = n_2 = 20$ Gaussian random vectors in $\mathbb{R}^p$, with null mean shift δ for 5000 data sets and nonnull mean shift δ for the remaining 5000. For the latter data sets, the nonzero shift is built in the first $k_0 = 3$ graph-based coefficients (the shift being zero for the remaining $p - k_0$ coefficients): $\tilde{\delta}_i \neq 0$ if and only if $i \leq k_0$ and $\Delta^2(\delta, \Sigma) = \Delta^2(\tilde{\delta}, \tilde{\Sigma}) = \tilde{\delta}^\top \tilde{\Sigma}^{-1} \tilde{\delta} = 1$. We consider two covariance settings. In the first one, the covariance matrix in the new space is diagonal, with diagonal elements equal to $\frac{1}{\sqrt{p}}$. In the second, correlation is introduced between the shifted coefficients only. Specifically, for $i, j \leq k_0$, $\tilde{\Sigma}_{ij} = \frac{0.5}{\sqrt{p}}$ if $i \neq j$, $\tilde{\Sigma}_{ii} = \frac{0.9}{\sqrt{p}}$ otherwise.

5.1. *Fixed known network.* Figure 5 displays receiver operator characteristic (ROC) curves for mean shift detection by the standard Hotelling T^2-test, T^2 in the first k_0 graph-based coefficients, T^2 in the first k_0 principal components (PC), the adaptive Neyman test of Fan and Lin (1998), and a modified version of this fourth test where the correct value of k_0 is specified. Note that we do not consider sparse learning approaches [Jacob, Obozinski and Vert (2009), Jenatton, Audibert and Bach (2009), Obozinski, Jacob and Vert (2011)], but it would be straightforward to design a realistic setting where such approaches are outperformed by testing, for example, by adding correlation between some of the functions under $\mathbf{H}_1$.

The first important comparison is between the classical Hotelling T^2-test vs. the T^2-test in the new graph-based space (Figure 5, top row). As expected from Lemma 1, testing in the restricted space where the shift lies performs much better than testing in the full space which includes irrelevant coefficients. The difference can be made arbitrarily large by increasing the dimension p and keeping the shift unchanged. The graph-structured test retains a large advantage even for moderately smooth shifts, for example, when $k_0 = 3$ and $p = 5$. Of course, this corresponds to the optimistic case where the number of shifted coefficients k_0 is known. Figure 6 shows the power of the test in the new space for various choices of k. Even when missing some coefficients ($k < k_0$) or adding a few irrelevant ones ($k > k_0$), the power of the graph-structured test is higher than that of the T^2-test in the full space. The principal component approach is shown in Figure 5 (top row) because it was proposed for the application which motivated our work [Ma and Kosorok (2009)] and because it also illustrates that the improvement in performance originates not only from dimensionality reduction, but also from the fact that this reduction is in a direction that does not decrease the shift. We emphasize that power entirely depends on the nature of the shift and that a PC-based test would outperform our graph-based test when the shift lies in the first principal components rather than graph-based coefficients.

The panels in the middle row show that the statistics of Bai and Saranadasa (1996), Chen and Qin (2010) and Srivastava and Du (2008) are also largely outperformed by our graph-structured statistic. This observation suggests that when such a graph-based prior on the shift is available, working in the new, lower-dimensional space does better at solving the problem of high-dimensionality than methods based on diagonal approximations of the covariance matrix. In addition, as one could expect, the procedures of Bai and Saranadasa (1996), Chen and Qin (2010) and Srivastava and Du (2008) perform very poorly in the presence of correlation. Here again, for a nonsmooth shift, the comparison would be less favorable to our procedure. We also considered the recently-proposed random projection approach of Lopes, Jacob and Wainwright (2011). Random projection was shown to give more power than Bai and Saranadasa (1996), Chen and Qin (2010) and Srivastava and Du (2008) in high-dimensional cases. However, as expected in our simulation setting where the sample size is twice the number of dimensions, it did not improve upon the Hotelling T^2-test (ROC curve not shown for the sake of readability). The

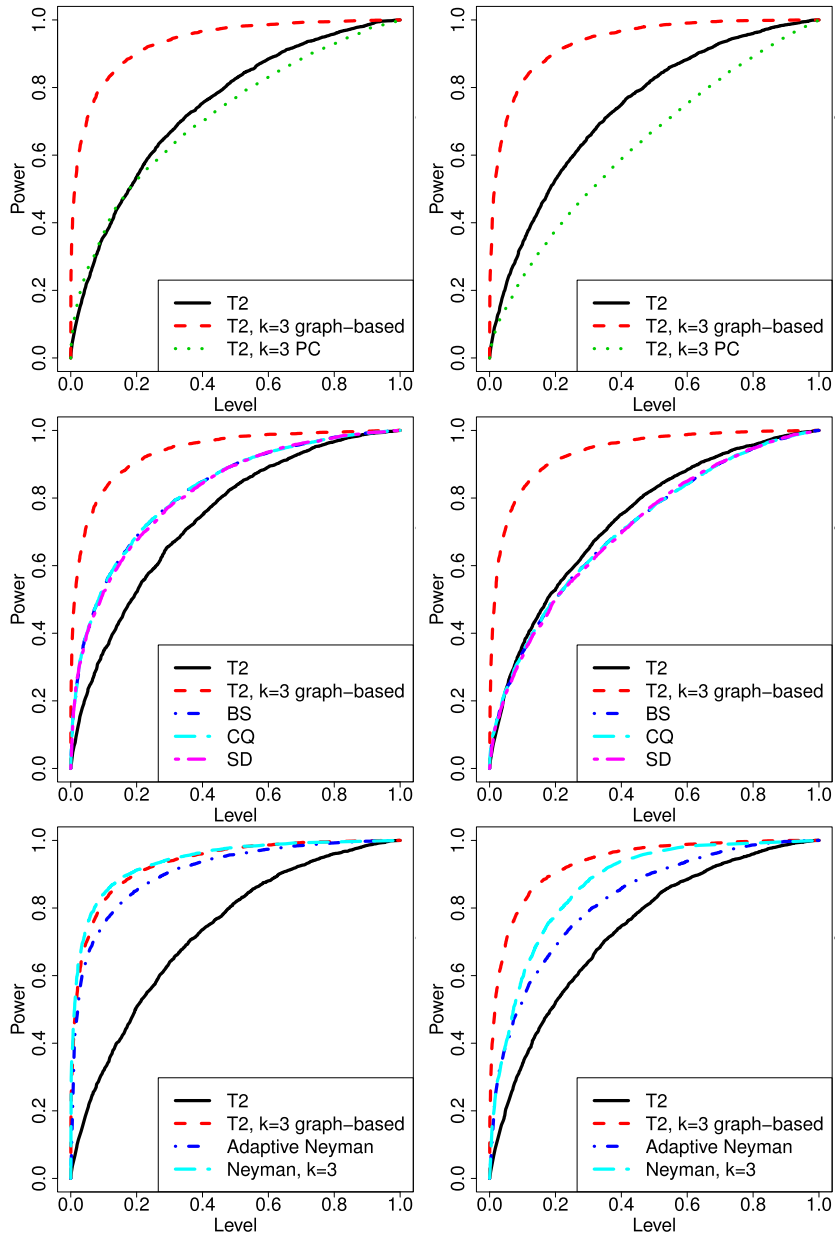

FIG. 5. *Synthetic data*: *ROC curves for the detection of a smooth shift*. *Left*: *Diagonal covariance structure*. *Right*: *Block-diagonal covariance structure*. *Top*: *Comparison of tests based on the standard Hotelling* T^2*-statistic in the original space*, T^2*-statistic in the first* k_0 *graph-based coefficients*, *and* T^2*-statistic in the first* k_0 *principal components*. *Middle*: *Comparison with the statistics of Bai and Saranadasa* (1996) (*BS*), *Chen and Qin* (2010) (*CQ*), *and Srivastava and Du* (2008) (*SD*). *Bottom*: *Comparison with the Neyman statistics of Fan and Lin* (1998).

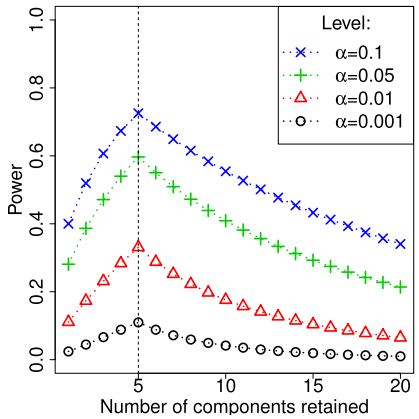

FIG. 6. *Synthetic data*: *Sensitivity to choice of* k. *Power of the* T^2*-test in the first* k *graph-based coefficients for a graph of* 20 *nodes*, *when the actual distribution shift* $\Delta^2 = 1$ *is evenly distributed among the first* $k_0 = 5$ *graph-based coefficients and with* $n_1 = n_2 = 20$.

method of Lopes, Jacob and Wainwright (2011) is more appropriate in a higher dimension and when no prior on the shift direction is available.

Finally, we consider the adaptive Neyman test of Fan and Lin (1998) (bottom two panels of Figure 5), which takes advantage of smoothness assumptions for time series. This test differs from our graph-structured test, as Fourier coefficients for stationary time series are known to be asymptotically independent and Gaussian. For graphs, the asymptotics would be in the number of nodes, which is typically small, and necessary conditions such as stationarity are more difficult to define and unlikely to hold for data such as gene expression measures. In the uncorrelated setting, the modified version of the Fan and Lin (1998) statistic based on the true number of nonzero coefficients performs approximately as well as the graph-structured T^2. However, for correlated data, it loses power and both versions of the Neyman test can have arbitrarily degraded performance. This, together with the need to use the bootstrap to calibrate the test, illustrates that direct transposition of the Fan and Lin (1998) test to the graph context is not optimal.

5.2. *Fixed network with errors*. We now consider the less idealistic case where the network used for testing is not exactly the one which was used to generate the data. More precisely, we follow the same procedure as in the correlated case of Section 5.1, but remove or add some edges to the network between the moment where we use it to generate the two samples and the moment where we use it in our testing procedure. This setting is much closer to what is likely to happen with real data, as available networks may miss several gene interactions which are not known yet and may include some incorrect interactions or some which are irrelevant for the problem under consideration. It is easy to see that in a worst case scenario, removing or adding an edge to the network can arbitrarily shrink the T_k^2

statistic. Take, for example, two disconnected nodes and assume without loss of generality that the empirical covariance matrix is the identity matrix and the empirical mean shifts for the two nodes are 1 and -1. Then, $\delta^\top \delta$ is 2, but adding an edge between the two nodes and projecting on the first eigenvector of the graph Laplacian matrix shrinks the observed shift to 0. A probabilistic analysis over random perturbations would be out of the scope of this paper, but the following simulation study is intended to give insight into what would happen in practice if randomly chosen edges are either wrongly added or omitted.

For the sake of clarity, Figure 7 only shows ROC curves for our graph-structured T^2 with $k = 2, 3, 4$, and the standard Hotelling T^2-statistic. The other competitors [Bai and Saranadasa (1996), Ma and Kosorok (2009), Chen and Qin (2010), Fan and Lin (1998), Srivastava and Du (2008), Lopes, Jacob and Wainwright (2011)] considered above all perform similarly to the Hotelling T^2-statistic. In the case where edges are erroneously removed, we start with a true network having 60

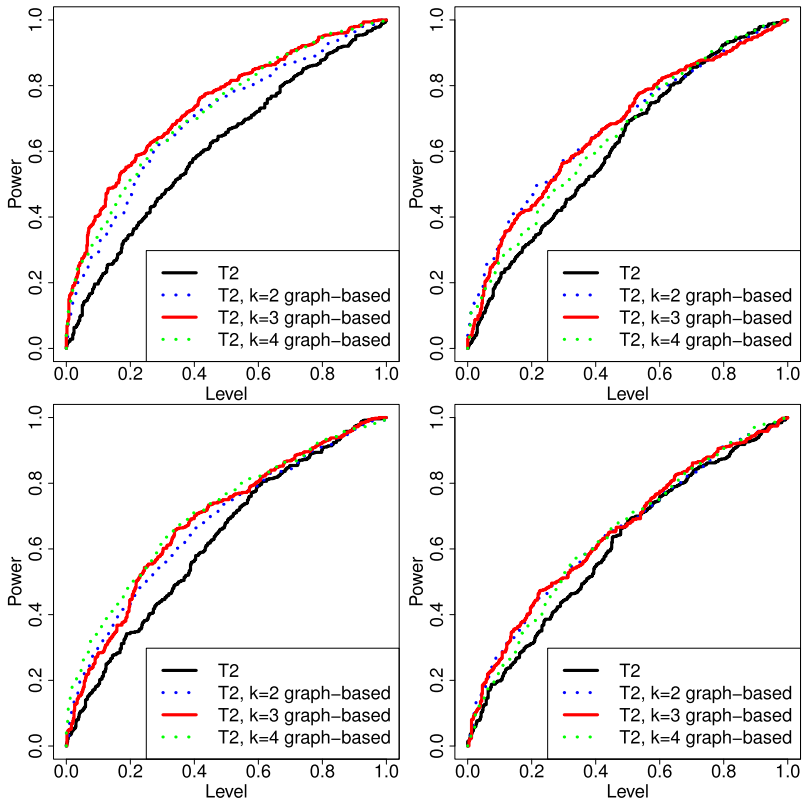

FIG. 7. *Synthetic data*: *ROC curves for the detection of a smooth shift in the presence of errors in the network. Comparison of tests based on the standard Hotelling T^2-statistic in the original space and T^2-statistic in the first k graph-based coefficients. Top*: *After randomly removing* 20 (*left*) *and* 40 (*right*) *edges. Bottom*: *After randomly adding* 20 (*left*) *and* 40 (*right*) *edges.*

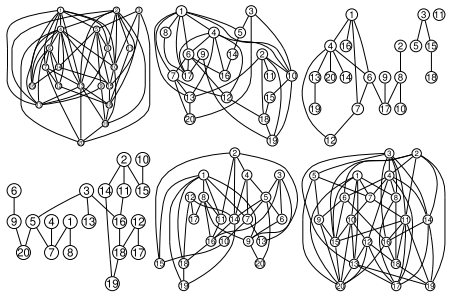

FIG. 8. *Synthetic data*: *Examples of corrupted networks used to generate Figure* 7. *Left column*: *original network used to generate the data of Section* 5.2 *before removing* (*top row*) *and adding* (*bottom row*) *edges*. *Middle column*: *one instance of removing/adding* 20 *edges*. *Right column*: *one instance of removing/adding* 40 *edges*.

edges instead of 20 in Section 5. Figure 7 shows that our graph-based approach can still perform much better than all competing methods, even in cases where the topology of the observed network is very incomplete (1/3 of the true number of edges) or noised by a lot of spurious edges. Figure 8 shows examples of networks corrupted by removing (top row) or adding (bottom row) edges to the original one and used in this experiment. It is visually clear that the information provided to our procedure is very different from the one, that is, used to generate the data. Again, this is an encouraging result, as it is well known that the gene networks available in the literature are missing a lot of interactions and often include incorrect information.

5.3. *Branch-and-bound subgraph discovery.* To evaluate the performance of the subgraph discovery algorithms proposed in Section 4, we generated a graph of 100 nodes formed by tightly-connected hubs of sizes sampled from a Poisson distribution with parameter 10 and only weak connections between these hubs (Figure 9). Such a graph structure is intended to mimic the typical topology of gene regulation networks. We randomly selected one subgraph of 5 nodes to be nonhomogeneous, with smooth shift in the first $k_0 = 3$ coefficients. The mean shift was set to zero on the rest of the graph. We set the norm of the mean shift to 1 and the covariance matrix to identity, so that detecting the shifted subgraph is impossible by just looking at the mean shift on the graph.

We evaluated run-time for full enumeration, the exact branch-and-bound algorithm based on Lemma 2 (Section 4.1), and the approximate algorithm based on the Euclidean norm (Section 4.2). We also examined run-time on data with permuted class labels, as the subgraph discovery procedure is to be run on such data to evaluate the number of false positives and adjust for multiple testing. Averaging over 20 runs, the full enumeration procedure took 732 ± 9 seconds per run and the exact branch-and-bound 627 ± 59 seconds on the nonpermuted data and 578 ± 100 seconds on permuted data. Over 100 runs, the approximation at $\theta = 0.5$

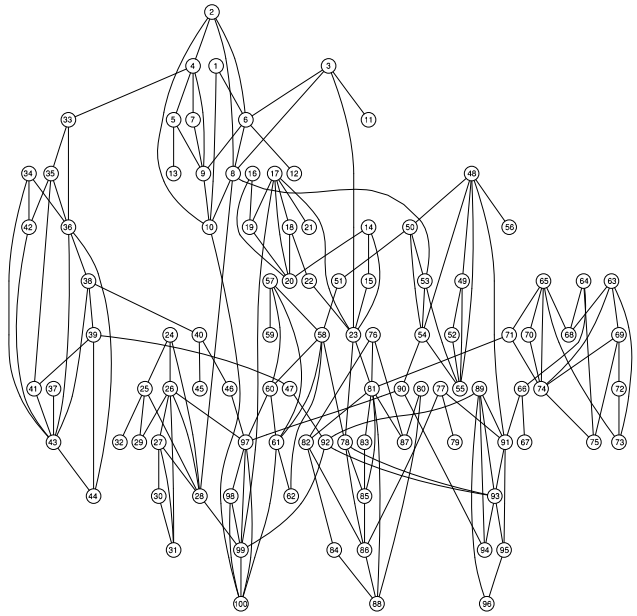

FIG. 9. *Synthetic data*: *Random graph used in the evaluation of the pruning procedure.*

($\lambda_{\min} = 0.52$) took 204 ± 86 seconds (129 ± 91 on permuted data) and the approximation at $\theta = 1$ ($\lambda_{\min} = 1.04$) took 183 ± 106 seconds (40 ± 60 on permuted data). The latter approximation missed the nonhomogeneous subgraph in 5% of the runs.

While neither the exact nor the approximate bounds are efficient enough to allow systematic testing on huge graphs for which the full enumeration approach would be impossible, they allow a significant gain in speed, especially for permuted data, and will thus prove to be very useful for multiple testing adjustment.

6. Results on cancer gene expression data. We also validated our methods using the following two microarray expression data sets: a breast cancer data set [Loi et al. (2008)] and a bladder cancer data set [Stransky et al. (2006)].

Breast cancer data set. The first data set by Loi et al. (2008) comprises the expression measures of 15,737 genes for 255 ER+ breast cancer patients treated with tamoxifen. Breast tumors are generally classified into three main categories [Perou et al. (2000)]: *luminal epithelial/ER+*, *HER2+*, and *triple negative*. ER+ tumors typically express estrogen receptors at a high level and often rely on estrogen for their growth. Tamoxifen is an antagonist of estrogen receptors and therefore prevents its activation by endogenous estrogen. Some ER+ tumors, however, keep growing after being treated with tamoxifen. An important goal is to detect structured groups of genes which are differentially expressed between resistant and sensitive patients, as detecting such groups could help understand resistance mechanisms and eventually improve ER+ breast tumor treatment. Using distant

metastasis-free survival as a primary endpoint, 68 patients from this data set are labeled as resistant to tamoxifen and 187 are labeled as sensitive to tamoxifen.

Bladder cancer data set. The second data set by Stransky et al. (2006) consists of the expression measures of 8323 genes for 57 urothelial tumors. Urothelial tumors are known to be arising and evolving through two distinct pathways, one typically leading to low-grade noninvasive tumors (Ta tumors), the other involving more aggressive tumors [Bakkar et al. (2003), Knowles (2006)]. These two subtypes, however, are not distinguishable from simple markers such as estrogen receptor or HER2 status for breast tumors. Mutation of the FGFR3 gene is sometimes used as a proxy, as about 70% of the noninvasive tumors carry it. As this information was unfortunately not available for this data set, we used the second best proxy which is tumor stage. We defined two groups: 25 tumors either at the Ta or T1 stages (TaT1 group) and 32 tumors at the T2, T3 or T4 stages (T2+ group). The muscle invasive T2+ tumors are aggressive and present a high risk of metastasis, while the Ta tumors have high recurrence level but low chance of progression into muscle invasive tumors. Identifying pathways which differ in expression between the two subtypes could help understand the disease better and improve its treatment.

6.1. *KEGG networks*.

Breast cancer data set. Starting with the breast cancer data set, we tested individually 351 connected components from 100 KEGG pathways corresponding to known gene regulation networks listed in the supplemental article Supplement B [Jacob, Neuvial and Dudoit (2011b)], using the classical Hotelling T^2-test and the T^2-test in the new graph-based space retaining only the first 20% coefficients ($k = 0.2p$). This value was the one chosen in Rapaport et al. (2007) on the same networks, accordingly with an argument based on loss minimization, not on hypothesis testing. The analysis of Lopes, Jacob and Wainwright (2011) suggests that the projection method (on random subspaces in their case) is quite robust to the choice of k. More refined heuristics could be based on eigengaps, that is, on the distances between successive eigenvalues. Indeed, matrix perturbation results suggest that eigenspaces can vary a lot even under small perturbations of the network if the largest discarded eigenvalue is close to the smallest kept eigenvalue [Davis and Kahan (1969), Stewart and Sun (1990), Ipsen (2010)]. Values of k such that $\lambda_k - \lambda_{k+1}$ is as large as possible could therefore be generally preferable.

The networks had 36 nodes in average, with a median of 23. For each of the 351 graphs, (unadjusted) p-values were computed under the nominal F-distributions $F_0(p, n_1 + n_2 - p - 1)$ and $F_0(k, n_1 + n_2 - k - 1)$, respectively. The Benjamini and Hochberg (1995) procedure was then applied to control the false discovery rate (FDR) at level 0.05.

Since there is no gold standard regarding which pathways are actually involved in endocrine resistance, practical validation of the entire set of detected pathways requires advanced biological expertise and further experiments and is the subject

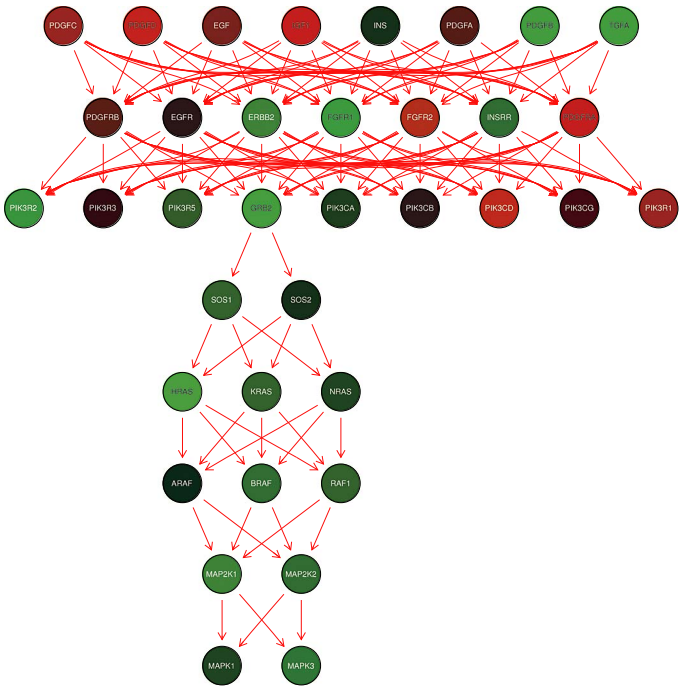

FIG. 10. *Breast cancer data set*: *KEGG prostate cancer pathway. Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the KEGG prostate cancer pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

of ongoing collaborations. Nonetheless, inspection of our list reveals several pathways which would not have been detected (or would have been further down in the list) without accounting for the network structure and which have recently been shown to be central in tamoxifen resistance. Many of these pathways involve the Ras/Raf-1/MAPK cascade [McGlynn et al. (2009)], like one of the connected components of the *prostate cancer* pathway shown in Figure 10 and one connected component of the *GnRH* pathway shown in Figure 11. The former also involves the overexpressed FGFR1, whose amplification was very recently shown to be implicated in endocrine therapy resistance by Turner et al. (2010). The latter pathway involves overexpressed SRC, which is also a well-studied target when trying to prevent tamoxifen resistance [Herynk et al. (2006)]. Both pathways have a much smaller p-value when accounting for their graph structure than when testing in the original gene space: 10^{-4} vs. 0.02 for the prostate cancer pathway and 10^{-3} vs. 0.11 for the GnRH signaling pathway. This is because the differences in expression of individual genes are insufficient to be significant in 36 and 19 dimensions, respectively, while the expression shift projected in the first 8 and 4 graph-based

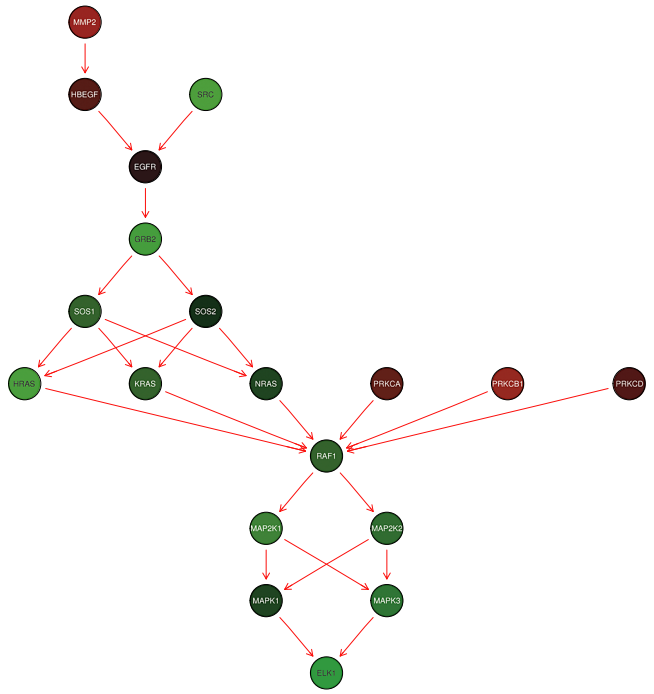

FIG. 11. *Breast cancer data set*: *KEGG GnRH signaling pathway. Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the KEGG GnRH signaling pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

directions, respectively, is significant. Note that the corresponding p-values for the hypergeometric enrichment test are 0.15 and 0.31. The complete gene lists of the two components are reported in Tables 3 and 4, respectively, in the supplemental article Supplement C [Jacob, Neuvial and Dudoit (2011c)]. Using a system-based approach like our proposed graph-based test therefore allows us to recover several known results (which may not have been obvious from the same data when looking at each gene individually) and may give insight regarding other resistance mechanisms by highlighting connections between these results.

Another example of a network selected only when accounting for graph structure is *Leukocyte transendothelial migration*, shown in Figure 12. To the best of our knowledge, this pathway is not specifically known to be involved in tamoxifen resistance. However, its role in resistance is plausible, as leukocyte infiltration was recently found to be involved in breast tumor invasion [Man (2010)]; more generally, the immune system and inflammatory response are closely related to the evolution of cancer. Here again, the p-value of the hypergeometric test is extremely

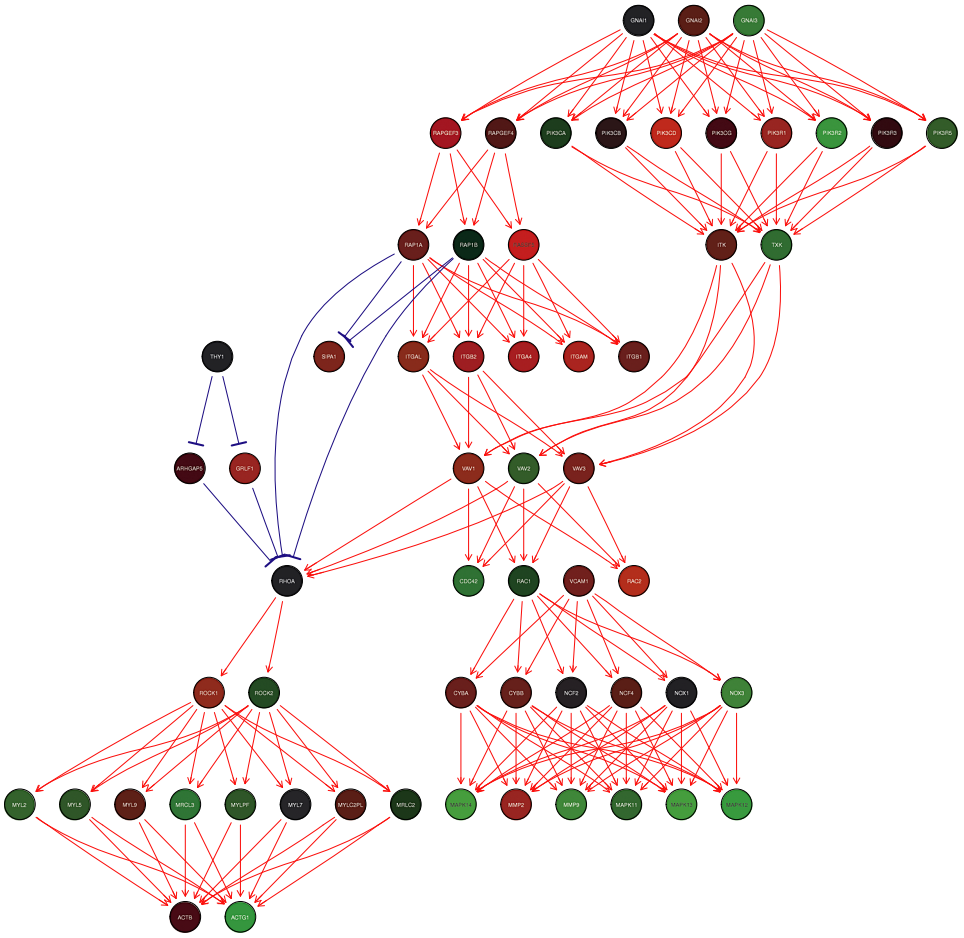

FIG. 12. *Breast cancer data set*: *KEGG leukocyte transendothelial migration pathway. Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the KEGG leukocyte transendothelial migration pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

high (0.31). The entire list of genes in this component is reported in Table 5 in the supplemental article Supplement C [Jacob, Neuvial and Dudoit (2011c)].

Bladder cancer data set. Testing the same KEGG networks on the bladder cancer data set, we immediately notice that several gene sets which are well known to be specific of one of the two bladder cancer progression pathways have much lower p-values under our graph-based approach than using the Hotelling T^2-statistic. This is the case, in particular, for the *p53 signaling pathway* [Spruck et al. (1994), Sanchez-Carbayo et al. (2006)], which is displayed in Figure 13 and for which

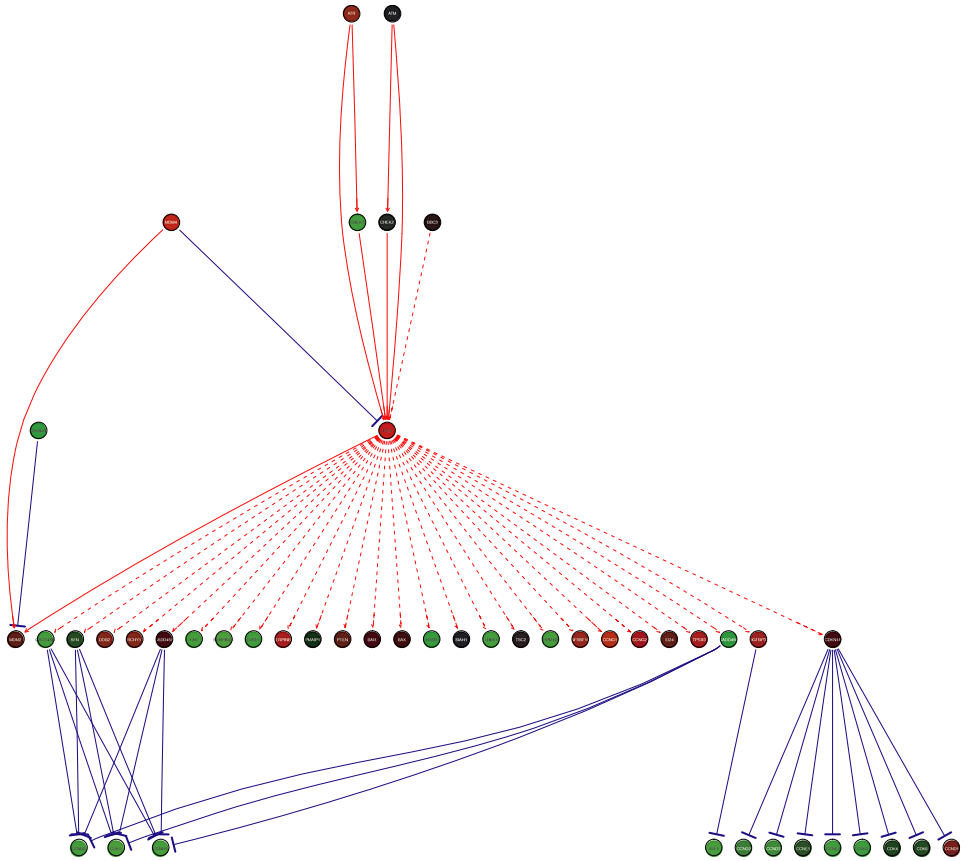

FIG. 13. *Bladder cancer data set*: *KEGG p*53 *signaling pathway. Scaled difference in sample mean expression measures between T*2+ *and TaT*1 *tumors, for genes in one component of the KEGG p*53 *signaling pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

the graph-based procedure outputs a p-value of 8.5×10^{-6} vs. 0.3 for the classical T^2-statistic. The TP53 gene itself is overexpressed in invasive (T2+) tumors. van Rhijn et al. (2004) suggested that FGFR3 and TP53 mutations characterize the two growth pathways and are mutually exclusive. A more recent study [Hernández et al. (2005)] contradicts the exclusion, but the observed underexpression of TP53 in the invasive group could be coherent with its typical mutation in invasive tumors. Genes coding for cyclins, such as CCNB1, CCNB2 and CDC2, are overexpressed. Cyclins are positively involved in cell proliferation, which is coherent with their overexpression in invasive tumors, as it was already observed for other genes of the cyclin family [Levidou et al. (2010)]. IGF1 is also overexpressed in T2+ tu-

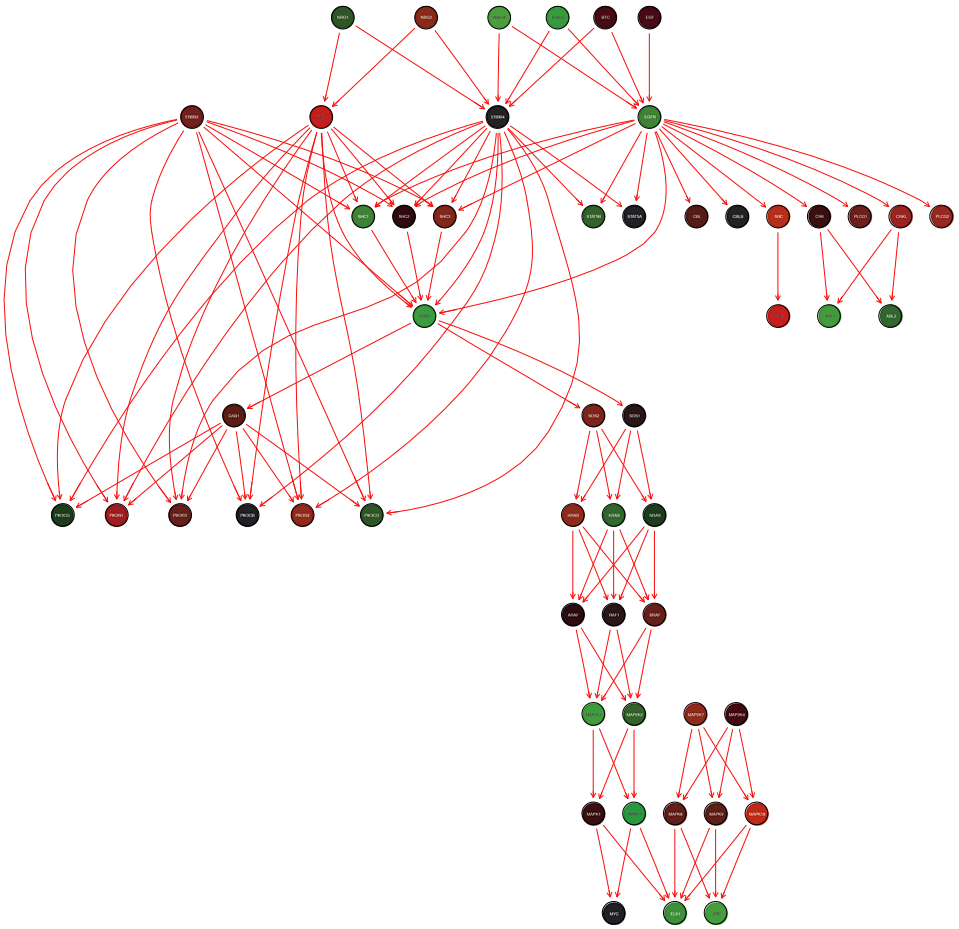

FIG. 14. *Bladder cancer data set*: *KEGG ErbB signaling pathway. Scaled difference in sample mean expression measures between T2+ and TaT1 tumors, for genes in one component of the KEGG ErbB signaling pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

mors, known to induce cell proliferation [Dunn et al. (1997)] and was selected as a prognosis predictor for bladder cancer in Mitra et al. (2009).

We also observe a much lower p-value using our procedure than using the classical T^2-statistic (2.3×10^{-5} vs. 0.066) for the *ErbB signaling pathway*, shown in Figure 14 and known to behave differently in the two bladder cancer growth pathways [Mellon et al. (1996)]. In particular, the network involves the PIK3, RAS and MAPK genes, which are known to be oncogenes specific to one of the growth pathways [Eswarakumar, Lax and Schlessinger (2005)].

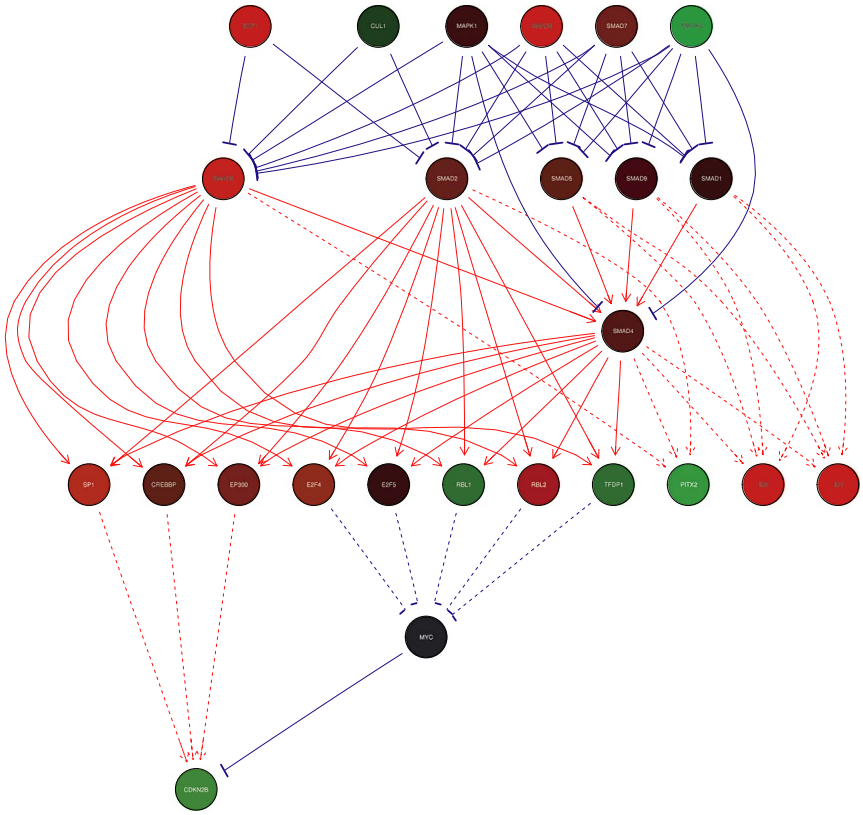

FIG. 15. *Bladder cancer data set*: *KEGG TGF-β signaling pathway. Scaled difference in sample mean expression measures between T2+ and TaT1 tumors, for genes in one component of the KEGG TGF-β signaling pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

Finally, changes in the *TGF-β signaling pathway* are also known to be related to the aggressiveness of bladder cancers [Hung et al. (2008)]. The network is shown in Figure 15 and here again our procedure results in a much lower p-value than the Hotelling test (2.6×10^{-6} vs. 0.049).

Unsurprisingly, these three networks have a relatively large size with respect to the low sample size of this data set and several of their genes show only very moderate differential expression when tested individually.

6.2. *NCI networks*. We also tested 75 connected components coming from gene networks of the NCI Pathway IntegrationDatabase.[2] The NCI networks considered are listed in the supplemental article Supplement B [Jacob, Neuvial and

[2] http://pid.nci.nih.gov.

Dudoit (2011b)]. Unlike KEGG pathways for which the Bioconductor R package `KEGGgraph` had already been developed, NCI pathways were not readily available as R objects. We therefore developed `NCIgraph` [Jacob (2011)], a Bioconductor R package which converts pathways available in BioPAX format to R objects. In addition, instead of importing a gene network as is into R, we provide an option to convert as well as possible the original network, whose nodes can represent proteins, protein complexes or concepts like transport or biochemical reactions, into one whose nodes correspond to *genes* and whose edges represent direct or indirect interactions at the expression level. For instance, if protein A is known to activate protein B, which is a transcription factor for gene C, a relevant network in terms of gene expression should be A and B pointing to C, whereas the BioPAX network will most likely be represented as A pointing to B pointing to C. As discussed in Section 5.2, our method is robust to irrelevant edges in the graph. Such a transformation is nonetheless important, since the method essentially uses biological networks as a prior on the covariance structure of gene expression. After this transformation, however, most networks have much simpler topologies, typically with all genes pointing to one or a few targets. As a result, Laplacian eigenvalues often have high multiplicities, which makes the effect of filtering less drastic.[3] In addition, the networks we consider here have much smaller size than the KEGG networks on average (8.9 vs. 36 for means, 7.5 vs. 23 for medians), which also explain the milder difference between results before and after dimensionality reduction.

For the breast cancer data, the 75 connected components we consider are those which have a nonempty intersection with the genes in this microarray data set. As for the KEGG networks, we compare the classical Hotelling T^2-test and the T^2-test in the new graph-based space retaining only the first 20% coefficients ($k = 0.2p$).

As an example, *NFkB activation by Nontypeable Hemophilus influenzae* shown in Figure 16 includes 21 genes from the breast cancer data set, but keeping the first 20% of the eigenvalues amounts to keeping 16 dimensions because of multiplicities. As a consequence, the p-value obtained after filtering is only slightly lower than that before filtering. Here again, the original context of study for this pathway has nothing to do with breast cancer: the purpose was to uncover the inflammation and mucin overproduction mechanism caused by a particular bacteria. Nevertheless, this network contains several genes which are either known actors of endocrine resistance or whose activity can be directly linked to the resistance phenomenon. Moreover, as one may expect, most of the observed gene-wise differential expression is coherent with the annotated interactions. On the lower part of the figure, IL1B is shown to be overexpressed in sensitive patients. Consistent

[3]If the eigenvalue 0 has a very high multiplicity, for example, then even the most extreme filtering still retains a large number of dimensions.

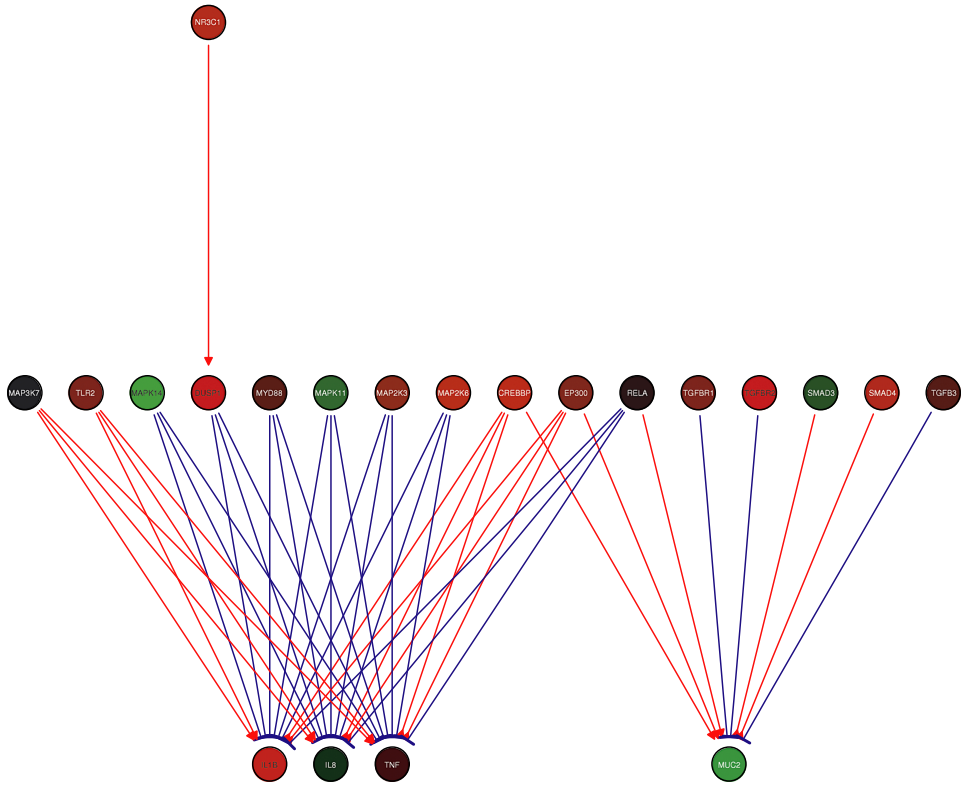

FIG. 16. *Breast cancer data set*: *NCI Nfkb activation by nontypeable hemophilus influenzae pathway. Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the NCI imported BioCarta Nfkb activation by nontypeable hemophilus influenzae pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

with this fact, its negative regulator p38 (MAPK11 and MAPK14) is downregulated in sensitive patients and its positive regulator CREBBP is upregulated. Note that DUSP1 was incorrectly annotated as a negative regulator in the automatic network conversion process of `NCIgraph` but is actually a positive regulator, as it is involved in the inactivation of p38. NR3C1 is involved in the transcription of DUSP1 and is also upregulated in sensitive patients. A few inconsistencies can be observed, like MAP2K3 and MAP2K6 which are negative regulators of IL1B, yet are overexpressed in sensitive patients. Recall, however, that the criterion we use for coherence is based on the difference between the expression of each gene and the (interaction-sign corrected) average expression of its regulators. The second main output of the pathway, MUC2, is downregulated in sensitive patients, which makes sense both in terms of the expression of its negative regulator TGFBR2, which is upregulated, and the already observed fact [Srinivasan et al. (2007)] that

the estrogen receptor upregulates MUC2 and that tamoxifen could block its expression.

The role of MUC2 in resistance to tamoxifen treatment of ductal carcinoma does not seem to be clearly established. Overexpression of MUC2 is sometimes found to be mildly correlated with good prognosis [Walsh et al. (1993), Rakha et al. (2005)], but this may be caused by its correlation with ER+ status. Its overexpression in resistant patients observed in this data set may well be noncausal, but would deserve further investigation. As for TGFBR2, inactivating mutations of the gene have been reported to be associated with recurrence and tamoxifen resistance [Lücke et al. (2001)], which is coherent with underexpression in resistant patients. Regarding IL1B, the main output of the pathway, its overexpression has been shown to be related to inhibition of cancer growth through apoptosis [Roy, Sarkar and Felty (2006)]. DUSP1 is a known negative regulator of cell proliferation and overexpression of p38 is known to be related to tamoxifen resistance [Gutierrez et al. (2005)]. Interestingly, NR3C1 activity has also been described by Wu et al. (2005) as being related to breast cancer cell survival through its induction of MAPK1 expression, which illustrates the interest of studying differential expression patterns at a system level rather than at the single-gene level.

It is also important to note that at least two interpretations can be given for the fact that sensitive patients have several gene expression patterns corresponding to known factors of good prognosis. Some of these patterns may be caused by the treatment, in which case understanding how tamoxifen affects these genes in some patients and not in others may be a proxy to understanding resistance mechanisms. Some of the patterns though may also have been caused by phenotypic traits of the sensitive patients, leading to better prognosis but without any link to the treatment.

Another small but relevant example is the *sonic hedgehog receptor ptc1 regulates cell cycle* pathway shown in Figure 17, which is entirely overexpressed in resistant patients, yielding a 10-fold change between the p-value with and without dimensionality reduction. The genes in this pathway are known to be related to tamoxifen resistance: CCNB1 is related to proliferation and is part of several existing tamoxifen-resistance signatures [Paik et al. (2004)] and inhibition of CDC2 was already proposed as an alternative treatment for endocrine resistant tumors [Johnson et al. (2010)].

6.3. *Branch-and-bound subgraph discovery.* We ran our branch-and-bound nonhomogeneous subgraph discovery procedure on the cell cycle pathway, whose largest connected component, after restriction to edges of known sign (inhibition or activation), has 86 nodes and 442 edges. Specifically, we sought to detect differentially expressed subgraphs of size $q = 5$, after preselecting those for which the squared Euclidean norm of the empirical shift exceeds $\theta = 0.1$; for a test in the first $k = 3$ components at level $\alpha = 10^{-4}$, this corresponds to $\lambda_{\min} < 0.23$ and to an expected removal of 95% of the subgraphs under the approximation that the squared Euclidean norm of the subgraphs follows a χ^2_5-distribution.

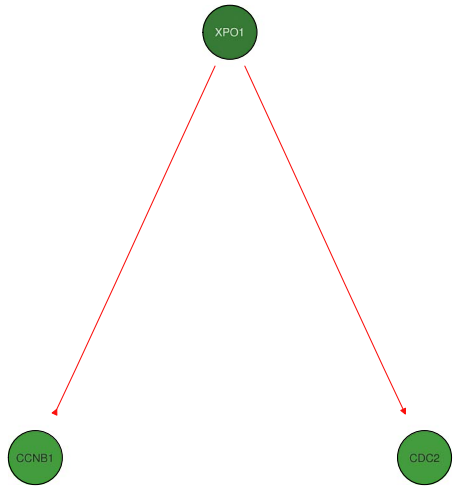

FIG. 17. *Breast cancer data set*: *NCI sonic hedgehog receptor ptc*1 *regulates cell cycle pathway. Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the NCI imported BioCarta sonic hedgehog receptor ptc*1 *regulates cell cycle pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

For $\alpha = 10^{-4}$, out of 100 runs on permuted data, only 9 rejected the null hypothesis for at least one subgraph. More precisely, 4 of these 9 runs detected 1 subgraph and the others detected 3, 6, 6, 21 and 26 subgraphs. In contrast, 41 overlapping subgraphs (Figure 18) were detected on the original data, corresponding to a connected subnetwork of 25 genes. Some of the genes belonging to these networks exhibit large individual differential expression, namely, TP53 whose mutation has been long known to be involved in tamoxifen resistance [Andersson et al. (2005), Fernandez-Cuesta et al. (2010)]. Accordingly, its negative regulator MDM2 is overexpressed and its positive regulator CREBBP is underexpressed. E2F1, whose expression level was recently shown to be involved in tamoxifen resistance [Louie et al. (2010)], is also part of the identified network, as well as CCND1 [Barnes (1997), Musgrove and Sutherland (2009)]. Some other genes in the network have quite low t-statistics and would not have been detected individually. This is the case of CCNE1 and CDK2, which were also described in Louie et al. (2010) as part of the same mechanism as E2F1. Similarly, CDKN1A was recently found to be involved in anti-estrogen treatment resistance [Musgrove and Sutherland (2009)] and in ovarian cancer, which is also a hormone-dependent cancer [Cunningham et al. (2009)]. Interestingly, RBX1, a gene coding for a RING-domain E3 ligase known to be involved in degradation of estrogen receptor α (ERα) [Ohtake et al. (2007)], appears to be overexpressed in resistant patients. This fact may suggest that some of the resistant ER+ patients had fewer receptors and, as a result, their

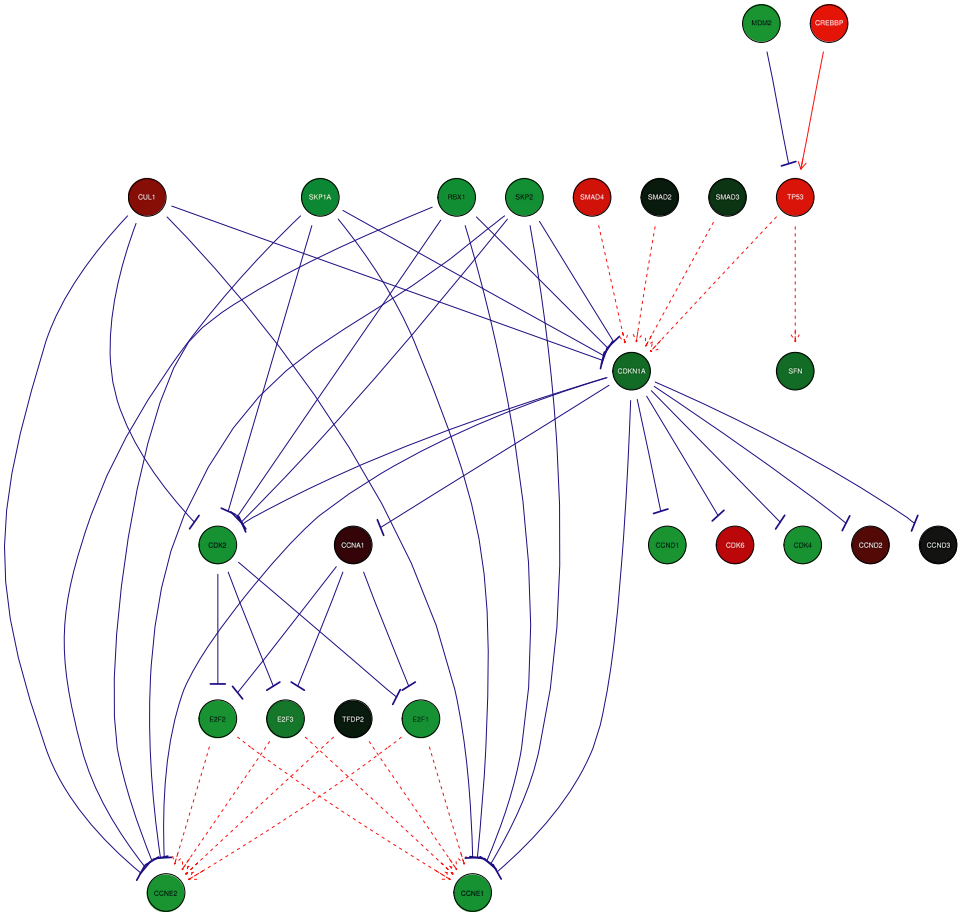

FIG. 18. *Breast cancer data set*: *Subgraph discovery. Difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in the two overlapping subgraphs detected at* $\alpha = 10^{-4}$. *Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to* 0. *Red arrows denote activation, blue arrows inhibition.*

tumors were relying less on estrogen for their growth, hence, the limited effect of a selective estrogen receptor modulator (SERM) like tamoxifen. The networks also contain CDK4, whose inhibition was described in Sutherland and Musgrove (2009) as acting synergistically with tamoxifen or trastuzumab. More generally, a large part of the network displayed in Figure 2A of Musgrove and Sutherland (2009) is included in our network, along with other known actors of tamoxifen resistance. Admittedly, selecting an important regulator like TP53 is not a surprising result, but our system-based approach to pathway discovery directly identifies an important set of interacting genes and may therefore prove to be more efficient than iterative individual identification of single actors.

7. Software implementation. The graph-structured test of Section 3 is implemented in the R software package `DEGraph`, released through the Bioconductor Project (release 2.7). Instructions for download and installation are available at http://www.bioconductor.org. Note that implementations of the branch-and-bound algorithms are not yet included in this package, but are available upon request.

As mentioned in Section 6.2, we also developed `NCIgraph` [Jacob (2011)], a Bioconductor R package which converts pathways available in BioPAX format to R objects with various preprocessing options.

8. Discussion. We developed a graph-structured two-sample test of means, for problems in which the distribution shift is assumed to be smooth on a given graph. We proved quantitative results on power gains for such smooth-shift alternatives and devised branch-and-bound algorithms to systematically apply our test to all the subgraphs of a large graph, without enumerating and testing these subgraphs one-by-one. The first algorithm is exact and reduces the number of explicitly tested subgraphs. The second one is approximate, with no false positives and a quantitative result on the type of false negatives (with respect to the exact algorithm). The nonhomogeneous subgraph discovery method involves performing a large number of tests, with highly-dependent test statistics. However, as the actual number of tested hypotheses is unknown, standard multiple testing procedures are not directly applicable. Instead, we use a permutation procedure to estimate the distribution of the number of false positive subgraphs. Such resampling procedures (bootstrap or permutation) are feasible due to the manageable run-time of the pruning algorithms of Section 4. Results on synthetic data illustrate the good power properties of our graph-structured test under smooth-shift alternatives, as well as the good performance of our branch-and-bound-like algorithms for subgraph discovery. Very promising results are also obtained on the gene expression data sets of Loi et al. (2008) and Stransky et al. (2006).

Future work should investigate the use of other bases, such as graph-wavelets [Hammond, Vandergheynst and Gribonval (2009)], which would allow the detection of shifts with spatially-located nonsmoothness, for example, to take into account errors in existing networks. As for the cutoff selection, more systematic procedures should be considered, for example, the two-step method proposed in Das Gupta and Perlman (1974), adaptive approaches as in Fan and Lin (1998) or heuristics based on the eigengap as mentioned in Section 6. The pruning algorithm would naturally benefit from sharper bounds. Such bounds could be obtained by controlling the condition number of all covariance matrices, using, for example, regularized statistics which still have known nonasymptotic distributions, such as those of Tai and Speed (2009). Concerning multiple testing, procedures should be devised to exploit the dependence structure between the tested subgraphs and to deal with the unknown number of tests. The proposed approach could also be enriched to take into account different types of data, for example, copy number for the detection of DE gene pathways. More subtle notions of smoothness, for

example, "and" (resp., "or") logical relations [Vaske et al. (2010)], could also be included to represent regulation mechanisms where the simultaneous presence of two transcription factors (resp., the presence of one or the other) is necessary to activate the transcription of another gene. Other applications of two-sample tests with smooth-shift on a graph include fMRI and eQTL association studies. For fMRI data, the goal would be to detect whether the brain activity changes between two conditions, using the prior information that parts of the brain which are close up to brain convolutions or known connection patterns should exhibit the same kind of change. One could also want to identify specific areas of the brain whose activity changes between two conditions. In eQTL studies, people are often interested in finding genes whose expression is influenced by single-nucleotide polymorphisms (SNPs), resulting in a large number of individual tests which often need to be aggregated *a posteriori* at the pathway level. Our method could be used to identify pathways whose expression is associated with particular SNPs.

Finally, it would be of interest to compare our testing approach with structured sparse learning (which we briefly described in Section 1) for the purpose of identifying expression signatures that are predictive of drug resistance. Methods should be compared in terms of prediction accuracy and stability of the selected genes across different data sets, a central and difficult problem in the design of such signatures [Ein-Dor et al. (2005), He and Yu (2010), Haury, Jacob and Vert (2010), Haury, Gestraud and Vert (2011)].

Acknowledgments. The authors thank Anne Biton, Noureddine El Karoui, Zaïd Harchaoui, Miles Lopes and Terry Speed for very helpful discussions and suggestions. They also acknowledge the UC Berkeley Center for Computational Biology Genentech Innovation Fellowship, the Stand Up To Cancer Program and The Cancer Genome Atlas Project for funding.

SUPPLEMENTARY MATERIAL

Supplement A: Technical results and proofs (DOI: 10.1214/11-AOAS528SUPPA; .pdf). This section contains our technical results (Lemma and Corollaries) on gain in power along with their proofs. It also contains the upper bound used in the branch and bound algorithm with its proof. Finally, it contains the lemma characterizing the subgraphs that would be missed by the approximated subgraph discovery algorithm presented in Section 4.2 along with its proof.

Supplement B: Pathways considered in the experiments (DOI: 10.1214/11-AOAS528SUPPB; .pdf). This section lists the names of the pathways considered in the experiments.

Supplement C: Gene lists (DOI: 10.1214/11-AOAS528SUPPC; .pdf). This section lists the genes belonging to each of the pathways studied in detail in the experiments along with their t-statistic and corresponding p-value.

REFERENCES

ANDERSSON, J., LARSSON, L., KLAAR, S., HOLMBERG, L., NILSSON, J., INGANÄS, M., CARLSSON, G., OHD, J., RUDENSTAM, C.-M., GUSTAVSSON, B. and BERGH, J. (2005). Worse survival for TP53 (p53)-mutated breast cancer patients receiving adjuvant CMF. *Ann. Oncol.* **16** 743–748.

BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

BAKKAR, A. A., WALLERAND, H., RADVANYI, F., LAHAYE, J.-B., PISSARD, S., LECERF, L., KOUYOUMDJIAN, J. C., ABBOU, C. C., PAIRON, J.-C., JAURAND, M.-C., THIERY, J.-P., CHOPIN, D. K. and DE MEDINA, S. G. D. (2003). FGFR3 and TP53 gene mutations define two distinct pathways in urothelial cell carcinoma of the bladder. *Cancer Res.* **63** 8108–8112.

BARNES, D. M. (1997). Cyclin D1 in mammary carcinoma. *J. Pathol.* **181** 267–269.

BEISSBARTH, T. and SPEED, T. P. (2004). GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20** 1464–1465.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697

CUNNINGHAM, J. M., VIERKANT, R. A., SELLERS, T. A., PHELAN, C., RIDER, D. N., LIEBOW, M., SCHILDKRAUT, J., BERCHUCK, A., COUCH, F. J., WANG, X., FRIDLEY, B. L., OVARIAN CANCER ASSOCIATION CONSORTIUM, GENTRY-MAHARAJ, A., MENON, U., HOGDALL, E., KJAER, S., WHITTEMORE, A., DICIOCCIO, R., SONG, H., GAYTHER, S. A., RAMUS, S. J., PHARAOH, P. D. P. and GOODE, E. L. (2009). Cell cycle genes and ovarian cancer susceptibility: A tagSNP analysis. *Br. J. Cancer* **101** 1461–1468.

DAS GUPTA, S. and PERLMAN, M. D. (1974). Power of the noncentral F test: Effect of additional variates on Hotelling's T^2-test. *J. Amer. Statist. Assoc.* **69** 174–180.

DAVIS, C. and KAHAN, W. M. (1969). Some new bounds on perturbation of subspaces. *Bull. Amer. Math. Soc.* **75** 863–868. MR0246155

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer, New York. MR2373771

DUNN, S. E., KARI, F. W., FRENCH, J., LEININGER, J. R., TRAVLOS, G., WILSON, R. and BARRETT, J. C. (1997). Dietary restriction reduces insulin-like growth factor I levels, which modulates apoptosis, cell proliferation, and tumor progression in p53-deficient mice. *Cancer Res.* **57** 4667–4672.

EIN-DOR, L., KELA, I., GETZ, G., GIVOL, D. and DOMANY, E. (2005). Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* **21** 171–178.

ESWARAKUMAR, V. P., LAX, I. and SCHLESSINGER, J. (2005). Cellular signaling by fibroblast growth factor receptors. *Cytokine Growth Factor Rev.* **16** 139–149.

EVANS, L. C. (1998). *Partial Differential Equations. Graduate Studies in Mathematics* **19**. Amer. Math. Soc., Providence, RI. MR1625845

FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. MR1649196

FERNANDEZ-CUESTA, L., ANAGANTI, S., HAINAUT, P. and OLIVIER, M. (2010). p53 status influences response to tamoxifen but not to fulvestrant in breast cancer cell lines. *Int. J. Cancer* **128** 1813–1821.

GOEMAN, J. J. and BÜHLMANN, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* **23** 980–987.

GOLDBERG, A. B. (2007). Dissimilarity in graph-based semisupervised classification. In *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*.

GUTIERREZ, M. C., DETRE, S., JOHNSTON, S., MOHSIN, S. K., SHOU, J., ALLRED, D. C., SCHIFF, R., OSBORNE, C. K. and DOWSETT, M. (2005). Molecular changes in tamoxifen-resistant breast cancer: Relationship between estrogen receptor, HER-2, and p38 mitogen-activated protein kinase. *J. Clin. Oncol.* **23** 2469–2476.

HAMMOND, D. K., VANDERGHEYNST, P. and GRIBONVAL, R. (2009). Wavelets on graphs via spectral graph theory. Available at arXiv:0912.3848.

HAURY, A. C., GESTRAUD, P. and VERT, J. P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. Preprint. Available at arXiv:1101.5008.

HAURY, A. C., JACOB, L. and VERT, J. P. (2010). Increasing stability and interpretability of gene expression signatures. ArXiv E-prints.

HE, Z. and YU, W. (2010). Stable feature selection for biomarker discovery. Available at arXiv:1001.0887.

HERNÁNDEZ, S., LÓPEZ-KNOWLES, E., LLORETA, J., KOGEVINAS, M., JARAMILLO, R., AMORÓS, A., TARDÓN, A., GARCÍA-CLOSAS, R., SERRA, C., CARRATO, A., MALATS, N. and REAL, F. X. (2005). FGFR3 and Tp53 mutations in T1G3 transitional bladder carcinomas: Independent distribution and lack of association with prognosis. *Clin. Cancer Res.* **11** 5444–5450.

HERYNK, M. H., BEYER, A. R., CUI, Y., WEISS, H., ANDERSON, E., GREEN, T. P. and FUQUA, S. A. W. (2006). Cooperative action of tamoxifen and c-Src inhibition in preventing the growth of estrogen receptor-positive human breast cancer cells. *Mol. Cancer Ther.* **5** 3023–3031.

HUNG, T.-T., WANG, H., KINGSLEY, E. A., RISBRIDGER, G. P. and RUSSELL, P. J. (2008). Molecular profiling of bladder cancer: Involvement of the TGF-beta pathway in bladder cancer progression. *Cancer Lett.* **265** 27–38.

IDEKER, T., OZIER, O., SCHWIKOWSKI, B. and SIEGEL, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. In *ISMB* 233–240.

IPSEN, I. C. F. (2010). The eigenproblem and invariant subspaces: Perturbation theory. In *G. W. Stewart*: *Selected Works with Commentaries* (M. E. Kilmer and D. P. O'Leary, eds.) 71–93. Birkhäuser, Basel. MR2731855

JACOB, L. (2011). NCIgraph: Pathways from the NCI Pathways Database R package version 1.0.0.

JACOB, L., NEUVIAL, P. and DUDOIT, S. (2011a). Supplement A to "More power via graph-structured tests for differential expression of gene networks." DOI:10.1214/11-AOAS528SUPPA.

JACOB, L., NEUVIAL, P. and DUDOIT, S. (2011b). Supplement B to "More power via graph-structured tests for differential expression of gene networks." DOI:10.1214/11-AOAS528SUPPB.

JACOB, L., NEUVIAL, P. and DUDOIT, S. (2011c). Supplement C to "More power via graph-structured tests for differential expression of gene networks." DOI:10.1214/11-AOAS528SUPPC.

JACOB, L., OBOZINSKI, G. and VERT, J.-P. (2009). Group lasso with overlap and graph lasso. In *ICML*'09: *Proceedings of the* 26*th Annual International Conference on Machine Learning* 433–440. ACM, New York.

JENATTON, R., AUDIBERT, J. Y. and BACH, F. (2009). Structured variable selection with sparsity-inducing norms. Research report, WILLOW–INRIA.

JOHNSON, N., BENTLEY, J., WANG, L.-Z., NEWELL, D. R., ROBSON, C. N., SHAPIRO, G. I. and CURTIN, N. J. (2010). Pre-clinical evaluation of cyclin-dependent kinase 2 and 1 inhibition in anti-estrogen-sensitive and resistant breast cancer cells. *Br. J. Cancer* **102** 342–350.

KNOWLES, M. A. (2006). Molecular subtypes of bladder cancer: Jekyll and Hyde or chalk and cheese? *Carcinogenesis* **27** 361–373.

LAND, A. H. and DOIG, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica* **28** 497–520. MR0115825

LEVIDOU, G., SAETTA, A. A., KARLOU, M., THYMARA, I., PRATSINIS, H., PAVLOPOULOS, P., ISAIADIS, D., DIAMANTOPOULOU, K., PATSOURIS, E. and KORKOLOPOULOU, P. (2010).

D-type cyclins in superficial and muscle-invasive bladder urothelial carcinoma: Correlation with clinicopathological data and prognostic significance. *J. Cancer Res. Clin. Oncol.* **136** 1563–1571.

LOI, S., HAIBE-KAINS, B., DESMEDT, C., WIRAPATI, P., LALLEMAND, F., TUTT, A. M., GILLET, C., ELLIS, P., RYDER, K., REID, J. F. et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* **9** 239.

LÖNNSTEDT, I. and SPEED, T. (2002). Replicated microarray data. *Statist. Sinica* **12** 31–46. MR1894187

LOPES, M. E., JACOB, L. and WAINWRIGHT, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. Technical report. Available at arXiv:1108.2401.

LOUIE, M. C., MCCLELLAN, A., SIEWIT, C. and KAWABATA, L. (2010). Estrogen receptor regulates E2F1 expression to mediate tamoxifen resistance. *Mol. Cancer Res.* **8** 343–352.

LU, Y., LIU, P.-Y., XIAO, P. and DENG, H.-W. (2005). Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics* **21** 3105–3113.

LÜCKE, C. D., PHILPOTT, A., METCALFE, J. C., THOMPSON, A. M., HUGHES-DAVIES, L., KEMP, P. R. and HESKETH, R. (2001). Inhibiting mutations in the transforming growth factor beta type 2 receptor in recurrent human breast cancer. *Cancer Res.* **61** 482–485.

MA, S. and KOSOROK, M. R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics* **25** 882–889.

MAN, Y.-G. (2010). Aberrant leukocyte infiltration: A direct trigger for breast tumor invasion and metastasis. *Int. J. Biol. Sci.* **6** 129–132.

MCGLYNN, L. M., KIRKEGAARD, T., EDWARDS, J., TOVEY, S., CAMERON, D., TWELVES, C., BARTLETT, J. M. S. and COOKE, T. G. (2009). Ras/Raf-1/MAPK pathway mediates response to tamoxifen but not chemotherapy in breast cancer patients. *Clin. Cancer Res.* **15** 1487–1495.

MELLON, J. K., LUNEC, J., WRIGHT, C., HORNE, C. H., KELLY, P. and NEAL, D. E. (1996). C-erbB-2 in bladder cancer: Molecular biology, correlation with epidermal growth factor receptors and prognostic value. *J. Urol.* **155** 321–326.

MITRA, A. P., PAGLIARULO, V., YANG, D., WALDMAN, F. M., DATAR, R. H., SKINNER, D. G., GROSHEN, S. and COTE, R. J. (2009). Generation of a concise gene panel for outcome prediction in urinary bladder cancer. *J. Clin. Oncol.* **27** 3929–3937.

MUSGROVE, E. A. and SUTHERLAND, R. L. (2009). Biological determinants of endocrine resistance in breast cancer. *Nat. Rev. Cancer* **9** 631–643.

NACU, S., CRITCHLEY-THORNE, R., LEE, P. and HOLMES, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics* **23** 850.

OBOZINSKI, G., JACOB, L. and VERT, J. P. (2011). Group Lasso with overlaps: The latent group Lasso approach. Technical report. arXiv.

OHTAKE, F., BABA, A., TAKADA, I., OKADA, M., IWASAKI, K., MIKI, H., TAKAHASHI, S., KOUZMENKO, A., NOHARA, K., CHIBA, T., FUJII-KURIYAMA, Y. and KATO, S. (2007). Dioxin receptor is a ligand-dependent E3 ubiquitin ligase. *Nature* **446** 562–566.

PAIK, S., SHAK, S., TANG, G., KIM, C., BAKER, J., CRONIN, M., BAEHNER, F. L., WALKER, M. G., WATSON, D., PARK, T., HILLER, W., FISHER, E. R., WICKERHAM, D. L., BRYANT, J. and WOLMARK, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351** 2817–2826.

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LØNNING, P. E., BØRRESEN-DALE, A. L., BROWN, P. O. and BOTSTEIN, D. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

RAKHA, E. A., BOYCE, R. W. G., EL-REHIM, D. A., KURIEN, T., GREEN, A. R., PAISH, E. C., ROBERTSON, J. F. R. and ELLIS, I. O. (2005). Expression of mucins (MUC1, MUC2, MUC3,

MUC4, MUC5AC and MUC6) and their prognostic significance in human breast cancer. *Mod. Pathol.* **18** 1295–1304.

RAPAPORT, F., ZINOVYEV, A., DUTREIX, M., BARILLOT, E. and VERT, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics* **8** 35.

ROY, D., SARKAR, S. and FELTY, Q. (2006). Levels of IL-1 beta control stimulatory/inhibitory growth of cancer cells. *Front. Biosci.* **11** 889–898.

SANCHEZ-CARBAYO, M., SOCCI, N. D., LOZANO, J., SAINT, F. and CORDON-CARDO, C. (2006). Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays. *J. Clin. Oncol.* **24** 778–789.

SANDLER, T., BLITZER, J., TALUKDAR, P. and PEREIRA, F. (2009). Regularized learning with networks of features. In *Neural Information Processing Systems*. MIT Press, Cambridge, MA.

SPRUCK, C. H., OHNESEIT, P. F., GONZALEZ-ZULUETA, M., ESRIG, D., MIYAO, N., TSAI, Y. C., LERNER, S. P., SCHMÜTTE, C., YANG, A. S. and COTE, R. (1994). Two molecular pathways to transitional cell carcinoma of the bladder. *Cancer Res.* **54** 784–788.

SRINIVASAN, S., ZAFAR, S., NAWAZ, Z. and LOGGIE, B. W. (2007). Transcriptional regulation of MUC2 by estrogen. 2007 Gastrointestinal Cancers Symposium.

SRIVASTAVA, M. S. (2009). A test for the mean vector with fewer observations than the dimension under nonnormality. *J. Multivariate Anal.* **100** 518–532. MR2483435

SRIVASTAVA, M. S. and DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99** 386–402. MR2396970

STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154

STRANSKY, N., VALLOT, C., REYAL, F., BERNARD-PIERROT, I., DIEZ DE MEDINA, S. G., SEGRAVES, R., DE RYCKE, Y., ELVIN, P., CASSIDY, A., SPRAGGON, C., GRAHAM, A., SOUTHGATE, J., ASSELAIN, B., ALLORY, Y., ABBOU, C. C., ALBERTSON, D. G., THIERY, J. P., CHOPIN, D. K., PINKEL, D. and RADVANYI, F. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.* **38** 1386–1396.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. and MESIROV, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

SUTHERLAND, R. L. and MUSGROVE, E. A. (2009). CDK inhibitors as potential breast cancer therapeutics: New evidence for enhanced efficacy in ER+ disease. *Breast Cancer Res.* **11** 112.

TAI, Y. C. and SPEED, T. P. (2009). On gene ranking using replicated microarray time course data. *Biometrics* **65** 40–51. MR2665844

TURNER, N., PEARSON, A., SHARPE, R., LAMBROS, M., GEYER, F., LOPEZ-GARCIA, M. A., NATRAJAN, R., MARCHIO, C., IORNS, E., MACKAY, A., GILLETT, C., GRIGORIADIS, A., TUTT, A., REIS-FILHO, J. S. and ASHWORTH, A. (2010). FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. *Cancer Res.* **70** 2085–2094.

VAN RHIJN, B. W. G., VAN DER KWAST, T. H., VIS, A. N., KIRKELS, W. J., BOEVÉ, E. R., JÖBSIS, A. C. and ZWARTHOFF, E. C. (2004). FGFR3 and P53 characterize alternative genetic pathways in the pathogenesis of urothelial cell carcinoma. *Cancer Res.* **64** 1911–1914.

VANDIN, F., UPFAL, E. and RAPHAEL, B. J. (2010). Algorithms for detecting significantly mutated pathways in cancer. In *RECOMB* (B. Berger, ed.). *Lecture Notes in Computer Science* **6044** 506–521. Springer, Berlin.

VASKE, C., BENZ, S., SANBORN, Z., EARL, D., SZETO, C., ZHU, J., HAUSSLER, D. and STUART, J. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. In *ISMB*.

WALSH, M. D., MCGUCKIN, M. A., DEVINE, P. L., HOHN, B. G. and WRIGHT, R. G. (1993). Expression of MUC2 epithelial mucin in breast carcinoma. *J. Clin. Pathol.* **46** 922–925.

WU, W., PEW, T., ZOU, M., PANG, D. and CONZEN, S. D. (2005). Glucocorticoid receptor-induced MAPK phosphatase-1 (MPK-1) expression inhibits paclitaxel-associated MAPK activation and contributes to breast cancer cell survival. *J. Biol. Chem.* **280** 4117–4124.

L. JACOB
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720-7360
USA
E-MAIL: laurent@stat.berkeley.edu

P. NEUVIAL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720-7360
USA
AND
LABORATOIRE STATISTIQUE ET GÉNOME
UNIVERSITÉ D'ÉVRY VAL D'ESSONNE –
UMR CNRS 8071 – USC INRA
FRANCE
E-MAIL: pierre.neuvial@genopole.cnrs.fr

S. DUDOIT
DEPARTMENT OF STATISTICS
DIVISION OF BIOSTATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720-7360
USA
E-MAIL: sandrine@stat.berkeley.edu