Publication 3

Leo Lahti, Juha E. A. Knuuttila, and Samuel Kaski. 2010. Global modeling of transcriptional responses in interaction networks. Bioinformatics, volume 26, number 21, pages 2713-2720.

BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 21 2010, pages 2713–2720
doi:10.1093/bioinformatics/btq500

Gene expression

Advance Access publication September 2, 2010

Global modeling of transcriptional responses in interaction networks

Leo Lahti[1,*], Juha E. A. Knuuttila[2] and Samuel Kaski[1,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University School of Science and Technology, PO Box 15400, FI-00076 Aalto and [2]Neuroscience Center, University of Helsinki, PO Box 54, FI-00014, Helsinki, Finland

Associate Editor: David Rocke

ABSTRACT

Motivation: Cell-biological processes are regulated through a complex network of interactions between genes and their products. The processes, their activating conditions and the associated transcriptional responses are often unknown. Organism-wide modeling of network activation can reveal unique and shared mechanisms between tissues, and potentially as yet unknown processes. The same method can also be applied to cell-biological conditions in one or more tissues.

Results: We introduce a novel approach for organism-wide discovery and analysis of transcriptional responses in interaction networks. The method searches for local, connected regions in a network that exhibit coordinated transcriptional response in a subset of tissues. Known interactions between genes are used to limit the search space and to guide the analysis. Validation on a human pathway network reveals physiologically coherent responses, functional relatedness between tissues and coordinated, context-specific regulation of the genes.

Availability: Implementation is freely available in R and Matlab at http://www.cis.hut.fi/projects/mi/software/NetResponse

Contact: leo.lahti@iki.fi; samuel.kaski@tkk.fi

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

Received on February 5, 2010; revised on August 24, 2010; accepted on August 26, 2010

1 INTRODUCTION

Coordinated activation and inactivation of genes through molecular interactions determines cell function. Changes in cell-biological conditions induce changes in the expression levels of co-regulated genes in order to produce specific physiological responses. A huge body of information concerning cell-biological processes is available in public repositories, including gene ontologies (Ashburner *et al.*, 2000), pathway models (Schaefer, 2006), regulatory information (Loots and Ovcharenko, 2007) and protein interactions (Kerrien *et al.*, 2007). Less is known about the contexts in which these processes are activated (Rachlin *et al.*, 2006), and how individual processes are reflected in gene expression (Montaner *et al.*, 2009). Although gene expression measurements provide only an indirect view to physiological processes, their wide availability provides a unique resource for investigating gene co-regulation on a genome- and organism-wide scale. This allows the detection of transcriptional responses that are shared by multiple tissues, suggesting shared physiological mechanism with potential biomedical implications, as demonstrated by the *Connectivity map* (Lamb *et al.*, 2006) where a number of chemical perturbations on a cancer cell line were used to reveal shared transcriptional responses between conditions to enhance screening of therapeutic targets. In this work, we study transcriptional responses of different tissues but the same methods can be directly used for modeling sets of cellular conditions within a single or multiple tissues as well.

Transcriptional responses have been modeled using so-called *gene expression signatures* (Hu *et al.*, 2006). A signature describes a co-expression state of the genes, associated with particular physiological states. Well-characterized signatures have proven to be accurate biomarkers in clinical trials, and hence reliable indicators of cell's physiological state. Disease-associated signatures are often coherent across tissues (Dudley *et al.*, 2009) or platforms (Hu *et al.*, 2006). Commercial signatures are available for routine clinical practice (Nuyten and van de Vijver, 2008), and other applications have been suggested recently (Dudley *et al.*, 2009). The established signatures are typically designed to provide optimal classification performance between two particular conditions. The problem with the classification-based signatures is that their associations to the underlying physiological processes are not well understood (Lucas *et al.*, 2009). Our goal is to enhance the understanding by deriving transcriptional signatures that are explicitly connected to well-characterized processes through the network.

We introduce and validate a novel approach for organism-wide discovery and analysis of transcriptional response patterns in interaction networks. Our algorithm has been designed to detect and model local regions in a network, each of which exhibits coordinated transcriptional response in a subset of measurements. In this study, the method is applied to investigate transcriptional responses of the network across a versatile collection of tissues across normal human body. The algorithm is independent of predefined classifications for genes or tissues. Organism-wide analysis can reveal unique and shared mechanisms between disparate tissues (Lage *et al.*, 2008), and potentially as yet unknown processes (Nacu *et al.*, 2007). The proposed NetResponse algorithm provides an efficient model-based tool for simultaneous feature selection and class discovery that utilizes known interactions between genes to guide the analysis. Related approaches include cMonkey (Reiss *et al.*, 2006) and a modified version of SAMBA biclustering (Tanay *et al.*, 2004).

*To whom correspondence should be addressed.

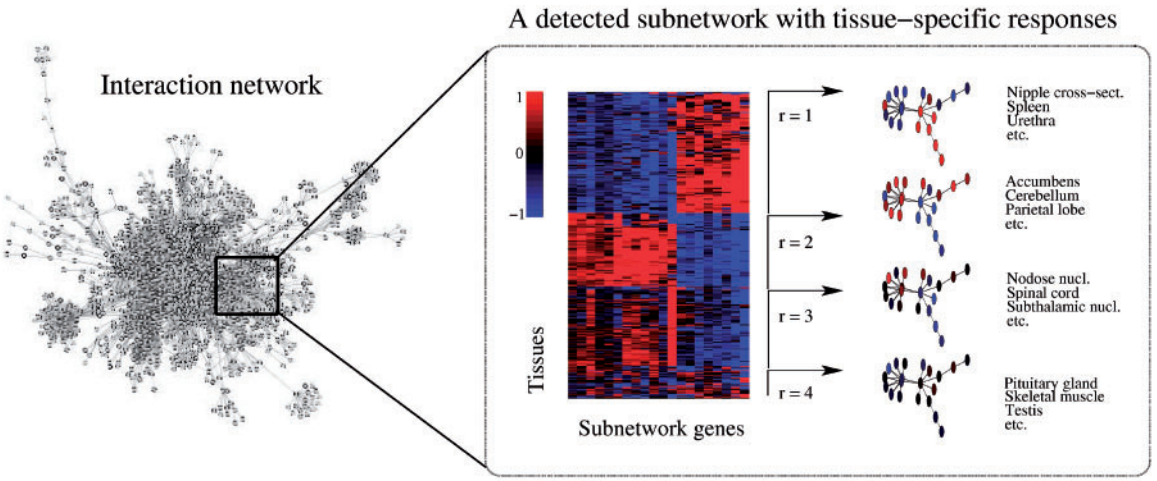

Fig. 1. Organism-wide analysis of transcriptional responses in a human pathway interaction network reveals physiologically coherent activation patterns and tissue-specific regulation. One of the subnetworks and its tissue-specific responses, as detected by the NetResponse algorithm is shown. The expression of each gene is visualized with respect to its mean level of expression across all samples.

However, these are application-oriented tools that rely on additional, organism-specific information, and their implementation is currently not available for most organisms, including human. We provide a general-purpose algorithm whose applicability is not limited to particular organisms.

NetResponse makes it possible to perform data-driven identification of functionally coherent network components and their tissue-specific responses. This is useful since the commonly used alternatives, predefined gene sets or pathways, are collections of intertwined processes rather than coherent functional entities (Nacu *et al.*, 2007). This has complicated their use in gene expression analysis, and methods have consequently been suggested to identify the 'key condition-responsive genes' of predefined gene sets (Lee *et al.*, 2008), or to decompose predefined pathways into smaller functional modules represented by gene expression signatures (Chang *et al.*, 2009). Our network-based search procedure detects the coordinately regulated gene sets in a data-driven manner. Gene expression provides functional information of the network that is missing in purely graph-oriented approaches for studying cell-biological networks (Aittokallio and Schwikowski, 2006). The network brings in prior information of gene function and connects the responses more closely to known processes than purely gene expression-based methods such as biclustering (Madeira and Oliveira, 2004), subspace clustering or other feature selection approaches (Law *et al.*, 2004; Roth and Lange, 2004). A key difference to previous network-based clustering methods, including MATISSE (Ulitsky and Shamir, 2007) and related approaches (Hanisch *et al.*, 2002; Shiga *et al.*, 2007) is that they assume a single correlated response between all genes in a module. NetResponse additionally models tissue-specific responses of the network. This allows a more expressive definition of a functional module, or a signature.

We validate the algorithm by modeling transcriptional responses in a human pathway interaction network across an organism-wide collection of tissues in normal human body. The results highlight functional relatedness between tissues, providing a global view on cell-biological network activation patterns.

2 METHODS

2.1 The NetResponse algorithm

We introduce a new approach for global detection and characterization of transcriptional responses in genome-scale interaction networks. NetResponse searches for local, connected *subnetworks* where joint modeling of gene expression reveals coordinated transcriptional response in particular tissues (Fig. 1). More generally, it is a new algorithm for simultaneous feature selection (for genes) and class discovery (for tissues) that utilizes known interactions between genes to limit the search space and to guide the analysis.

2.1.1 Gene expression signatures Subnetworks are the functional units of the interaction network in our model; transcriptional responses are described in terms of subnetwork activation. Given a physiological state, the underlying assumption is that gene expression in subnetwork n is regulated at particular levels to ensure proper functioning of the relevant processes. This can involve simultaneous activation and repression of the genes: sufficient amounts of mRNA for key proteins has to be available while interfering genes may need to be silenced. This regulation is reflected in a unique expression signature $\boldsymbol{s}^{(n)}$, a vector describing the associated expression levels of the subnetwork genes. The level of regulation varies from gene to gene; expression of some genes is regulated at precise levels whereas other genes fluctuate more freely. Given the physiological state, we assume that the distribution of observed gene expression is Gaussian, $\boldsymbol{x}^{(n)} \sim N(\boldsymbol{s}^{(n)}, \Sigma^{(n)})$.

2.1.2 Modeling tissue-specific transcriptional responses Each subnetwork is potentially associated with alternative transcriptional states, activated in different tissues and corresponding to unique combinations of processes. Since individual processes and their transcriptional responses are in general unknown (Lee *et al.*, 2008), detection of tissue-specific responses provides an efficient proxy for identifying functionally distinct states of the network. Our task is to detect and characterize these signatures. We assume that in a specific observation, the subnetwork n can be in any one of $R^{(n)}$ latent physiological states indexed by r. Each state is associated with a unique expression signature $\boldsymbol{s}_r^{(n)}$ over the subnetwork genes. Associations between the observations and the underlying physiological states are unknown, and treated as latent variables. This leads to a mixture model for gene expression

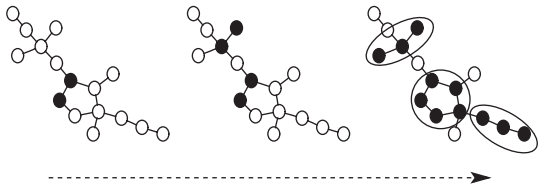

Fig. 2. The agglomerative subnetwork detection procedure. Initially, each gene is assigned in its own singleton subnetwork. Agglomeration proceeds by at each step merging the two neighboring subnetworks that benefit most from joint modeling of their transcriptional responses. This continues until no improvement is obtained by merging the subnetworks.

in the subnetwork n:

$$\boldsymbol{x}^{(n)} \sim \sum_{r=1}^{R^{(n)}} w_r^{(n)} p(\boldsymbol{x}^{(n)} | \boldsymbol{s}_r^{(n)}, \Sigma_r^{(n)}), \quad (1)$$

where each component distribution p is assumed to be Gaussian. In practice, we assume a diagonal covariance matrix $\Sigma_r^{(n)}$.

A particular transcriptional response is characterized by the triple $\{\boldsymbol{s}_r^{(n)}, \Sigma_r^{(n)}, w_r^{(n)}\}$. This defines the shape, fluctuations and frequency of the associated gene expression signature in subnetwork n. The feasibility of the Gaussian modeling assumption is supported by the previous observations of Kong *et al.* (2006), where predefined gene sets were used to investigate differences in gene expression between two predefined sample groups. In our model, the subnetworks, transcriptional responses and the activating tissues are learned from data. In one-channel data such as Affymetrix arrays used in this study, the centroids $\boldsymbol{s}_r^{(n)}$ describe absolute expression signals of the preprocessed array data. Relative differences can be investigated by comparing the detected responses. The model is applicable also on two-channel expression data when a common reference sample is used for all arrays since the relative differences are not altered by the choice of comparison baseline when the same baseline is used for all samples.

Now the model has been specified assuming the subnetworks are given. In practice, they are learned from the data. In order to do this, we make two assumptions. First, we rely on the prior information in the global interaction network, and assume that co-regulated gene groups are connected components in this network. Second, we assume that the subnetworks are independent. This allows a well-defined algorithm, and the subnetworks are then interpretable as independent components of transcriptional regulation. In practice the algorithm, described below, is an agglomerative approximation for searching for locally independent subnetworks.

2.2 Implementation

Efficient implementation is crucial for scalability. For fast computation, we use an agglomerative procedure where interacting genes are gradually merged into larger subnetworks (Fig. 2). Joint modeling of dependent genes reveals coordinated responses and improves the likelihood of the data when compared with independent models, giving the first criterion for merging the subnetworks. However, increasing subnetwork size tends to increase model complexity and the possibility of overfitting since the number of samples remains constant while the dimensionality (subnetwork size) increases. To compensate for this effect, we use a Bayesian information criterion (Gelman *et al.*, 2003) to penalize increasing model complexity and to determine optimal subnetwork size.

The cost function for a subnetwork G is $C(G) = -2L + q\log(N)$, where L is the (marginal) log-likelihood of the data, given the mixture model in Equation (1), q is the number of parameters, and N denotes sample size. NetResponse searches for a joint model for the network genes that maximizes the likelihood of observed gene expression, but avoids increasing model complexity through penalizing an increasing number of model parameters. An optimal model is searched for by at each step merging the subnetwork pair that produces the maximal gain in the cost function. More formally, the algorithm merges at each step the subnetwork pair G_i, G_j that minimizes the cost $\Delta\mathcal{C} = -2(L_{i,j} - (L_i + L_j)) + (q_{i,j} - (q_i + q_j))\log(N)$. The agglomerative scheme is as follows:

Initialize: learn univariate Gaussian mixture for the expression values of each gene, and bivariate joint models for all potential gene pairs with a direct link. Assign each gene into its own singleton subnetwork.

Merge: merge the neighboring subnetworks G_i, G_j that have a direct link in the network and minimize the difference $\mathcal{C}$. Compute new joint models between the newly merged subnetwork and its neighbors.

Terminate: continue merging until no improvement is obtained by merging the subnetworks ($\Delta\mathcal{C} \geq 0$).

The number $R^{(n)}$ of distinct transcriptional responses of the subnetwork is unknown, and is estimated with an infinite mixture model. Learning several multivariate Gaussian mixtures between the neighboring subnetworks at each step is a computationally demanding task, in particular when the number of mixture components is unknown. The Gaussian mixtures, including the number of mixture components, are learned with an efficient variational Dirichlet process implementation (Kurihara *et al.*, 2007). The likelihood L in the model is approximated by the lower bound of the variational approximation. The Gaussian mixture detects a particular type of dependency between the genes. In contrast to MATISSE (Ulitsky and Shamir, 2007) and other studies that use correlation or other methods to measure global co-variation, the mixture model detects coordinated responses that can be activated only in a few tissues. Tissue-specific joint regulation indicates functional dependency between the genes but it may have a minor contribution to the overall correlation between gene expression profiles. In principle, we could also model the dependencies in gene fluctuations within each individual response with covariances of the Gaussian components. However, this would heavily increase model complexity, and therefore we leave dependencies in gene-specific fluctuations within each response unmodeled, and focus on modeling differences between the responses. NetResponse provides a full generative model for gene expression, where each subnetwork is described with an independent joint mixture model. The maximum subnetwork size is limited to 20 genes to avoid numerical instabilities in computation. The infinite Gaussian mixture can automatically adapt model complexity to the sample size. We model subnetworks of 1–20 genes across 353 samples; similar dimensionality per sample size has previously been used with variational mixture models (Honkela *et al.*, 2008).

2.3 Data

2.3.1 Pathway interaction network We investigate the pathway interaction network based on the KEGG database of metabolic pathways (Kanehisa *et al.*, 2008) provided by the signaling pathway impact analysis (SPIA) package (Tarca *et al.*, 2009) of BioConductor (www.bioconductor.org). This implements the pathway impact analysis method originally proposed in Draghici *et al.* (2007), which is currently the only pathway analysis tool that considers pathway topology. SPIA provides the data in a readily suitable form for our analysis. Other pathway datasets, commonly provided in the BioPAX format, are not readily available in a suitable pairwise interaction form. Directionality and types of the interactions were not considered. Genes with no expression measurements were removed from the analysis. We investigate the largest connected component of the network with 1800 unique genes, identified by Entrez GeneIDs.

2.3.2 Gene expression data We analyzed a collection of normal human tissue samples from 10 post-mortem donors (Roth *et al.*, 2006), containing gene expression measurements from 65 normal tissues. To ensure sample quality, RNA degradation was minimized in the original study by flash freezing all samples within 8.5 h post-mortem. Only the samples passing Affymetrix quality measures were included. Each tissue has 3–9 biological replicates measured on the Affymetrix HG-U133plus2.0 platform. The reproducibility of our findings is investigated in an independent human

gene expression atlas (Su *et al.*, 2004), measured on the Affymetrix HG-U133A platform, where two biological replicates are available for each measured tissue. In the comparisons, we use the 25 tissues available in both datasets (adrenal gland cortex, amygdala, bone marrow, cerebellum, dorsal root ganglia, hypothalamus, liver, lung, lymph nodes, occipital lobe, ovary, parietal lobe, pituitary gland, prostate gland, salivary gland, skeletal muscle, spinal cord, subthalamic nucleus, temporal lobe, testes, thalamus, thyroid gland, tonsil, trachea and trigeminal ganglia). Both datasets were preprocessed with RMA (Irizarry *et al.*, 2003). Certain genes have multiple probesets, and a standard approach to summarize information across multiple probesets is to use alternative probeset definitions based on probe-genome remapping (Dai *et al.*, 2005). This would provide a single expression measure for each gene. However, since the HG-U133A array represents a subset of probesets on the HG-U133Plus2.0 array, the redefined probesets are not technically identical between the compared datasets. To minimize technical bias in the comparisons, we use probesets that are available on both platforms. Therefore, we rely on manufacturer annotations of the probesets and use an alternative approach (used e.g. by Nymark *et al.*, 2007), where one of the available probesets is selected at random to represent each unique gene. Random selection is used to avoid selection bias. When available, the 'xxxxxx_at' probesets were used because they are more specific by design than the other probe set types (www.affymetrix.com).

2.4 Validation

The NetResponse algorithm is validated with an application on the pathway interaction network of 1800 genes (Tarca *et al.*, 2009) across 353 gene expression samples from 65 tissues in normal human body (Roth *et al.*, 2006). NetResponse is compared with alternative approaches in terms of physiological coherence and reproducibility of the findings.

2.4.1 Comparison methods NetResponse is designed for organism-wide modeling of transcriptional responses in genome-scale interaction networks. Simultaneous detection of the subnetworks and their tissue-specific responses is a key feature of the model. A straightforward alternative would be a two-step approach where the subnetworks and their tissue-specific responses are detected in separate steps, although this can be suboptimal for detecting tissue-specific responses. Various methods are available for detecting subnetworks based on network and gene expression data (Hanisch *et al.*, 2002; Shiga *et al.*, 2007) in the two-step approach. We use MATISSE, a state-of-the-art algorithm described in Ulitsky and Shamir (2007). MATISSE finds connected subgraphs in the network such that each subgraph consists of highly correlated genes. The output is a list of genes for each detected subnetwork. Since MATISSE only clusters the genes, we model transcriptional responses of the detected subnetworks in a separate step by using a similar mixture model to the NetResponse algorithm. This combination is also new, and called MATISSE+ in this article. The second comparison method is the SAMBA biclustering algorithm (Tanay *et al.*, 2002). The output is a list of associated genes and tissues for each identified bicluster. SAMBA detects gene sets with tissue-specific responses, but, unlike NetResponse and MATISSE+, the algorithm does not utilize the network. Influence of the prior network is additionally investigated by randomly shuffling the gene expression vectors, while keeping the network and the within-gene associations intact. Comparisons between the original and shuffled data help to assess relative influence of the prior network on the results. Comparisons to randomly shuffled genes in SAMBA are not included since SAMBA does utilize network information.

2.4.2 Reproducibility in validation data Reproducibility of the findings is investigated in an independent validation dataset in terms of significance and correlation (for details, see Section 2.3). Each comparison method implies a grouping for the tissues in each subnetwork, corresponding to the detected responses. It is expected that physiologically relevant differences between the groups are reproducible in other datasets. We tested this by estimating differential expression between the corresponding tissues in the validation data for each pairwise comparison of the predicted groups using a standard test for gene set analysis (GlobalTest; Goeman *et al.*, 2004). To ensure that the responses are also qualitatively similar in the validation data, we measured Pearson's correlation between the detected responses and those observed in the corresponding tissues in validation data. The responses were characterized by the centroids provided by the model in NetResponse and MATISSE+. For SAMBA, we used the mean expression level of each gene within each group of tissues since SAMBA groups the tissues but does not characterize the responses. In validation data, the mean expression level of each gene is used to characterize the response within each group of tissues. Probesets were available for 75% of the genes in the detected subnetworks in the validation data; transcriptional responses with less than three probesets in the validation data were not considered. Validation data contained corresponding samples for >79%) of the predicted responses in NetResponse, MATISSE+ and SAMBA (Supplementary Table 1).

3 RESULTS

The validation results reported below demonstrate that the NetResponse algorithm is readily applicable for modeling transcriptional responses in interaction networks on an organism-wide scale. While biomedical implications of the findings require further investigation, NetResponse detects a number of physiologically coherent and reproducible transcriptional responses in the network, and highlights functional relatedness between tissues. It also outperformed the comparison methods in terms of reproducibility of the findings.

3.1 Application to human pathway network

In total, NetResponse identified 106 subnetworks with 3–20 genes (Supplementary Material). For each subnetwork, typically (median) three distinct transcriptional responses were detected across the 65 tissues (Supplementary Fig. 1). One of the subnetworks with four distinct responses is illustrated in Figure 1. Each response is associated with a subset of tissues. Statistically significant differences between the corresponding tissues were observed also in the independent validation data ($P < 0.01$; GlobalTest). Three of the four responses were also qualitatively similar (correlation > 0.8; Supplementary Fig. 2). The first response is associated with immune system-related tissues such as spleen and tonsil. Responses 2–3 are associated with neuronal tissues such as subthalamic or nodose nucleus, or with central nervous system, for example accumbens and cerebellum. The fourth group manifests a 'baseline' signature that fluctuates around the mean expression level of the genes. Testis and pituitary gland are examples of tissues in this group. While most tissues are strongly associated with a particular response, samples from amygdala, bone marrow, cerebral cortex, heart atrium and temporal lobe manifested multiple responses. While alternative responses reveal tissue-specific regulation, detection of physiologically coherent and reproducible responses may indicate shared mechanisms between tissues. Although the responses may reflect previously unknown processes, it is likely that some of them reflect the activation patterns of known pathways. Overlapping pathways can provide a starting point for interpretation. The subnetwork in Figure 1 overlaps with various known pathways, most remarkably with the MAPK pathway with 10 genes (detailed gene–pathway associations are provided in the Supplementary Material; see Subnetwork 12). MAPK is a general signal transduction system that participates in a complex, cross-regulated signaling network that is sensitive to cellular stimuli (Wilkinson and Millar, 2000).

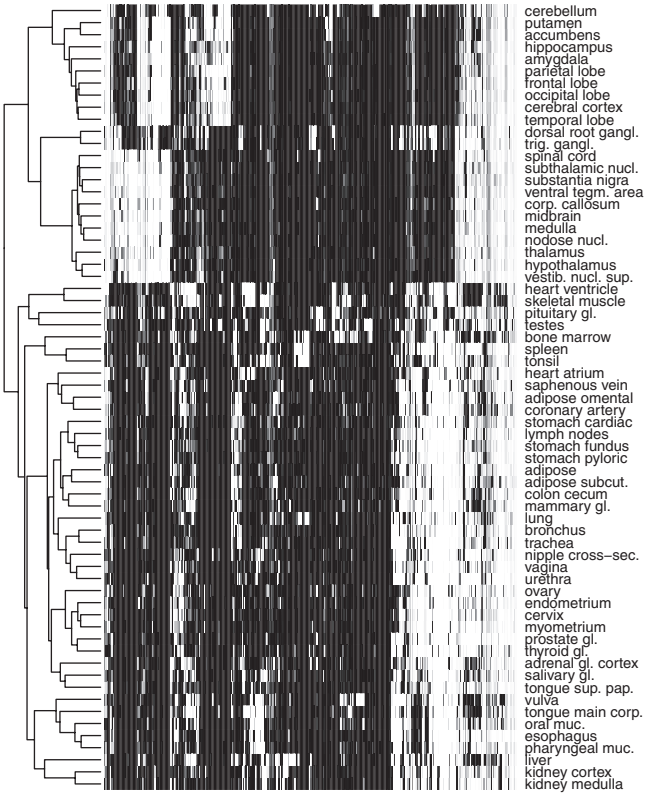

Fig. 3. Associations between 65 tissues (rows) and the detected transcriptional responses of the pathway interaction network of Figure 1. The shade indicates the probability of a particular transcriptional response in each tissue (black: $P=0$; white: $P=1$). Hierarchical clustering based on the signature co-occurrence probabilities between each pair of tissues highlights their relatedness.

Six subnetwork genes participate in the p53 pathway, which is a known regulator of the MAPK signaling pathway. In addition, p53 is known to interact with a number of other pathways, both as an upstream regulator and a downstream target (Wu, 2004). Both MAPK and p53 are associated with processes including cell growth, differentiation and apoptosis, and exhibit diverse cellular responses. Tissue-specific regulation can potentially explain the detection of alternative transcriptional states of the subnetwork.

The detected responses characterize absolute expression signals in our preprocessed one-channel array data. Systematic differences in the expression levels of the individual genes are normalized out in the visualization by showing the relative expression of each gene with respect to its mean expression level across all samples. Note that the choice of a common baseline does not affect the relative differences between the samples.

3.1.1 Tissue-selective network activation Associations between the tissues and the detected transcriptional responses are shown in Figure 3. Some responses are shared by many tissues, while others are more specific to particular contexts such as immune system, muscle or the brain. Related tissues often exhibit similar network activation patterns, which is seen by grouping the tissues according to co-occurrence probabilities of shared transcriptional response.

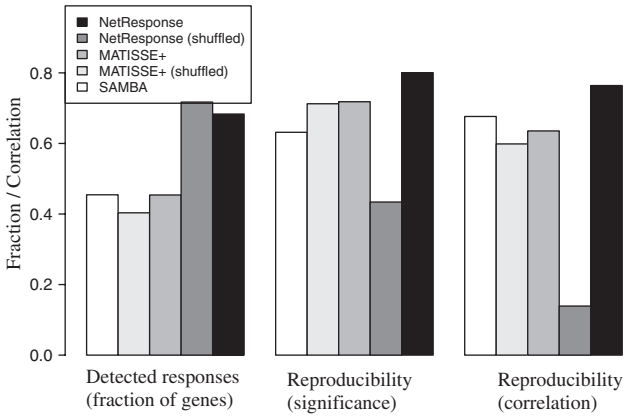

Fig. 4. Comparison between the alternative approaches. *Detected responses:* fraction of genes participating in the detected transcriptional responses. *Reproducibility (significance):* fraction of responses that are reproducible in the validation data in terms of differential expression between the associated tissues ($P<0.05$; GlobalTest). *Reproducibility (correlation):* median correlation between the gene expression levels of the detected responses and the corresponding tissues in the validation data.

This is known as *tissue selectivity* of gene expression (Liang *et al.*, 2006).

3.1.2 Probabilistic tissue connectome Tissue relatedness can be measured in terms of shared transcriptional responses (Supplementary Fig. 3). This is an alternative formulation of the *tissue connectome* map suggested by Greco *et al.* (2008) to highlight functional *connectivity* between tissues based on the number of shared differentially expressed genes at different thresholds. We use shared network responses instead of shared gene count. The use of co-regulated gene groups is expected to be more robust to noise than the use of individual genes. As the overall measure of connectivity between tissues, we use the mean of signature co-occurrence probabilities over the subnetworks, given the model in Equation (1). The analysis reveals functional relatedness between the tissues. In particular, two subcategories of the central nervous system appear distinct from the other tissues. Closer investigation of the observed responses would reveal how the tissues are related at transcriptional level (Supplementary Material).

3.2 Comparison to alternative approaches

NetResponse was compared with the alternative approaches in terms of physiological coherence and reproducibility of the findings (Fig. 4; Supplementary Table 1). NetResponse detected the largest amount of responses; 68% of the network genes were associated with a response, compared with 45% in MATISSE+ and SAMBA. At the same time, NetResponse outperformed the comparison methods in terms of reproducibility of the findings.

3.2.1 Physiological coherence The association between the responses and tissues was measured by normalized mutual information (NMI; Bush *et al.*, 2008) between the sample-response assignments and sample class labels within each subnetwork. The NMI varies from 0 (no association) to 1 (deterministic association). The transcriptional responses detected by NetResponse, MATISSE+ and SAMBA show statistically significant associations to particular

tissues with a significantly higher average NMI (0.46–0.50) than expected based on randomly labeled data (0.26–0.32; $P < 10^{-4}$; Wilcoxon test; Supplementary Table 1). The highest average NMI (0.50) was obtained by NetResponse but differences between NetResponse, MATISSE+ and SAMBA are not significant. NetResponse is significantly physiologically more coherent also when compared with results obtained with shuffled gene expression (NMI 0.22; $P < 10^{-12}$). The observations confirm the potential physiological relevance of the findings in NetResponse, MATISSE+ and SAMBA.

3.2.2 Reproducibility The majority of the detected responses were reproducible both in terms of significance and correlation (Supplementary Fig. 4) as described in Section 2.4. Of the predicted differences between groups of tissues, 80% were significant in validation data with $P < 0.05$ (GlobalTest), compared with 72 and 63% in MATISSE+ and SAMBA, respectively, or 43% obtained for randomly shuffled data with NetResponse (Fig. 4). The changes were also qualitatively similar; in NetResponse the median correlation between the detected responses and corresponding tissues in the validation data is 0.76, which is significantly higher ($P < 0.01$; Wilcoxon test) than in the comparison methods (MATISSE+: 0.64; SAMBA: 0.68), or in randomly shuffled NetResponse data (0.14). NetResponse detected responses for a larger fraction of the genes (68%) than the other methods. This seems an intrinsic property of the algorithm since it detected responses for a similar fraction of the genes also in the network with randomly shuffled genes (72%). However, only the findings from the real data were reproducible.

4 DISCUSSION

Cell-biological networks may cover thousands of genes, but any change in the physiological context typically affects only a small part of the network. While gene function and interactions are often subject to context-specific regulation (Liang *et al.*, 2006), they are typically studied only in particular experimental conditions. Organism-wide analysis could reveal highly specialized functions that are activated only in one or a few tissues. Detection of shared responses between the tissues can reveal previously unknown functional connections and help to formulate novel hypotheses of gene function in previously unexplored contexts. We provide a well-defined algorithm for such analysis.

The results support the validity of the model. NetResponse detected the largest number of responses without compromising physiological coherence or reproducibility of the findings compared with the alternatives. The most highly reproducible results were obtained by NetResponse. Further analysis is needed to establish the physiological role of the findings.

NetResponse is readily applicable for modeling tissue-specific responses in cell-biological networks, including pathways, protein interactions and regulatory networks. The network connects the responses to well-characterized processes, and provides readily interpretable results that are less biased toward known biological phenomena than methods based on predefined gene sets that are routinely used in gene expression studies to bring in prior information of gene function and to increase statistical power. However, these are often collections of intertwined processes rather than coherent functional entities. For example, pathways from KEGG may contain hundreds of genes, while only a small part of a pathway may be affected by changes in physiological conditions (Nacu *et al.*, 2007). This has complicated the use of predefined gene sets in gene expression studies. Draghici *et al.* (2007) demonstrated that taking into account aspects of pathway topology, such as gene and interaction types can improve the estimation of pathway activity. While their SPIA algorithm measures the activity of known pathways between two predefined conditions, our algorithm searches for potentially unknown functional modules, and detects their association to multiple conditions, or tissues, simultaneously. This is useful since biomedical pathways are human-made descriptions of cellular processes, often consisting of smaller, partially independent modules (Chang *et al.*, 2009; Hartwell *et al.*, 1999). Our data-driven search procedure can rigorously identify functionally coherent network modules where the interacting genes show coordinated responses. Joint modeling increases statistical power that is useful since gene expression, and many interaction data types such as protein–protein interactions, have high noise levels. The probabilistic formulation accounts for biological and measurement noise in a principled manner. Certain types of interaction data such as transcription factor binding or protein interactions are directly based on measurements. This can potentially help to discover as yet unknown processes that are not described in the pathway databases (Nacu *et al.*, 2007). False negative interactions form a limitation for the current model because joint responses of co-regulated genes can be modeled only when they form a connected subnetwork.

The need for principled methods for analyzing large-scale collections of gene expression data is increasing with their availability. Versatile gene expression atlases contain valuable information about shared and unique mechanisms between disparate tissues which is not available in smaller and more specific experiments (Lage *et al.*, 2008; Scherf *et al.*, 2000). For example, Lamb *et al.* (2006) demonstrated that large-scale screening of cell lines under diverse conditions can enhance the finding of therapeutic targets. Our model is directly applicable in similar exploratory tasks, providing tools for organism-wide analysis of transcriptional activity in normal human tissues (Roth *et al.*, 2006; Su *et al.*, 2004), cancer and other diseases (Kilpinen *et al.*, 2008; Lukk *et al.*, 2010) in a genome- and organism-wide scale. Similar collections are available for several model organisms including mouse (Su *et al.*, 2004), yeast (Granovskaia *et al.*, 2010) and plants (Schmid *et al.*, 2005). A key advantage of our approach compared with methods that perform targeted comparisons between predefined conditions (Ideker *et al.*, 2002; Sanguinetti *et al.*, 2008) is that it allows systematic organism-wide investigation when the responses and the associated tissues are unknown. The motivation is similar to SAMBA and other biclustering approaches that detect groups of genes that show coordinated response in a subset of tissues (Madeira and Oliveira, 2004), but the network ties the findings more tightly to cell-biological processes in our model. This can focus the analysis and improve interpretability. Since the non-parametric mixture model adjusts model complexity with sample size, our algorithm is potentially applicable also in smaller and more targeted datasets. For example, it could potentially advance disease subtype discovery by revealing differential network activation in subsets of patients.

Many large-scale collections are continuously updated with new measurements. Our algorithm provides no integration technique for new experiments yet; on-line extensions that could directly

integrate data from new experiments provide an interesting topic for further study. Another potential extension would be a fully Bayesian treatment that would provide confidence intervals, removing the need to assess significance of the results in a separate step. While our model provides a model-based criterion for detecting the responses without prior knowledge of the activating tissues, the statistical significance of the findings has to be verified in further experiments. The majority of the responses in our experiments could be verified in an independent dataset. Other potential extensions include adding more structure to address the directionality, relevance and probabilities of the interactions. Not all cell-biological processes have clear manifestations at transcriptome level. Hence, information of transcript and interaction types, as in SPIA, could potentially help to improve the sensitivity of our approach. We could also seek to loosen the constraints imposed by the prior network. However, such extensions would come with an increased computational cost. The simple and efficient implementation is a key advantage.

NetResponse is closely related to subspace clustering methods such as agglomerative independent variable component analysis (AIVGA; Honkela *et al.*, 2008). However, AIVGA and other model-based feature selection techniques (Law *et al.*, 2004; Roth and Lange, 2004) consider all potential connections between the features, which leads to more limited scalability. Finding a global optimum in our model would require exhaustive combinatorial search over all potential subnetworks. Since the complexity depends on the topology of the network, finding a general formulation for the model complexity is problematic. The number of potential solutions grows faster than exponentially with the number of features (genes) and links between them, making exhaustive search in genome-scale interaction networks infeasible. Approximative solutions are needed, and are often sufficient in practice. A combination of techniques is used to achieve an efficient algorithm compared with the model complexity. First, we focus the analysis on those parts of the data that are supported by known interactions. This increases modeling power and considerably limits the search space. Second, the agglomerative scheme finds an approximative solution where at each step the subnetwork pair that leads to the highest improvement in cost function is merged. This finds a solution relatively fast compared with the complexity of the task. Note that the order in which the subnetworks become merged may affect the solution. Finally, the variational implementation considerably speeds up mixture modeling (Kurihara *et al.*, 2007). The running time of our application was 248 min on a standard desktop computer (Intel 2.83GHz; Supplementary Fig. 5).

Investigation of a human pathway interaction network revealed tissue-specific regulation in the network, that is, groups of interacting genes whose joint response differs between tissues. This highlights the context-dependent nature of network activation, and emphasizes an important shortcoming in the current gene set-based testing methods (Nam and Kim, 2008): simply measuring gene set 'activation' is often not sufficient; it is also crucial to characterize *how* the expression changes, and in which conditions. Organism-wide modeling can provide quantitative information about these connections.

5 CONCLUSIONS

We have introduced and validated a general-purpose algorithm for global identification and characterization of transcriptional responses in genome-scale interaction networks across a diverse collection of tissues, applicable also to cell-biological conditions within and between tissues. An organism-wide analysis of a human pathway interaction network validated the model, and provided a global view on cell-biological network activation. The results revealed unique and shared mechanisms between tissues, and potentially help to formulate novel hypotheses of gene function in previously unexplored contexts.

Funding: Academy of Finland (207467); IST Programme of the European Community, under the PASCAL2 Network of Excellence (ICT-216886); Finnish Center of Excellence on Adaptive Informatics Research (AIRC; to L.L. and S.K.).

Conflict of Interest: none declared.

REFERENCES

Aittokallio,T. and Schwikowski,B. (2006) Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.*, **7**, 243–255.

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

Bush,W. *et al.* (2008) Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics*, **9**, 238.

Chang,J.T. *et al.* (2009) A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol. Cell*, **34**, 104–114.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.

Dudley,J.T. *et al.* (2009) Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.*, **5**, 307.

Gelman,A. *et al.* (2003) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

Goeman,J.J. *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Granovskaia,M.V. *et al.* (2010) High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biol.*, **11**, R24

Greco,D. *et al.* (2008) Physiology, pathology and relatedness of human tissues from gene expression meta-analysis. *PLoS ONE*, **3**, e1880.

Hanisch,D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145–154.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Honkela,A. *et al.* (2008) Agglomerative independent variable group analysis. *Neurocomputing*, **71**, 1311–1320.

Hu,Z. *et al.* (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**, 96.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–S240.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36** (Suppl. 1), D480–D484.

Kerrien,S. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35** (Suppl. 1), D561–D565.

Kilpinen,S. *et al.* (2008) Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol.*, **9**, R139.

Kong,S.W. *et al.* (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Kurihara,K. *et al.* (2007) Accelerated variational Dirichlet process mixtures. In Schölkopf,B. *et al.* (eds) *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, pp. 761–768.

Lage,K. *et al.* (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA*, **105**, 20870–20875.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Law,M. *et al.* (2004) Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 1154–1166.
Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.
Liang,S. *et al.* (2006) Detecting and profiling tissue-selective genes. *Physiol. Genomics*, **26**, 158–162.
Loots,G. and Ovcharenko,I. (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics*, **23**, 122–124.
Lucas,J.E. *et al.* (2009) Cross-study projections of genomic biomarkers: an evaluation in cancer genomics. *PLoS ONE*, **4**, e4523.
Lukk,M. *et al.* (2010) A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.
Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.
Montaner,D. *et al.* (2009) Gene set internal coherence in the context of functional profiling. *BMC Genomics*, **10**, 197.
Nacu,S. *et al.* (2007) Gene expression network analysis and applications to immunology. *Bioinformatics*, **23**, 850–858.
Nam,D. and Kim,S.-Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
Nuyten,D. and van de Vijver,M. (2008) Using microarray analysis as a prognostic and predictive tool in oncology: focus on breast cancer and normal tissue toxicity. *Semin. Radiat. Oncol.*, **18**, 105–114.
Nymark,P. *et al.* (2007) Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, **8**, 62.
Rachlin,J. *et al.* (2006) Biological context networks: a mosaic view of the interactome. *Mol. Syst. Biol.*, **2**, 66.
Reiss,D. *et al.* (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **7**, 280.
Roth,R. *et al.* (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, **7**, 67–80.
Roth,V. and Lange,T. (2004) Feature selection in clustering problems. In Thrun,S. *et al.* (eds) *Advances in Neural Information Processing Systems* **16**. MIT Press, Cambridge, MA, pp. 473–480.
Sanguinetti,G. *et al.* (2008) MMG: a probabilistic tool to identify submodules of metabolic pathways. *Bioinformatics*, **24**, 1078–1084.
Schaefer,C.F. (2006) An Introduction to the NCI Pathway Interaction Database. *NCI-Nature Pathway Interaction Database*. [Epub ahead of print, doi:10.1038/PID.2006.001]
Scherf,U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
Schmid,M. *et al.* (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–506.
Shiga,M. *et al.* (2007) Annotating gene function by combining expression data with a modular gene network. *Bioinformatics*, **23**, 468–478.
Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.
Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl Acad. Sci. USA*, **101**, 2981–2986.
Tarca,A.L. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.
Ulitsky,I. and Shamir,R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.*, **1**, 8.
Wilkinson,M.G. and Millar,J.B. (2000) Control of the eukaryotic cell cycle by MAP kinase signaling pathways. *FASEB J.*, **14**, 2147–2157.
Wu,G.S. (2004) The functional interactions between the MAPK and p53 signaling pathways. *Cancer Biol. Therapy*, **3**, 146–151.

Global modeling of transcriptional responses in interaction networks: Supplementary Figures

Leo Lahti, Juha E.A. Knuuttila, and Samuel Kaski
Bioinformatics

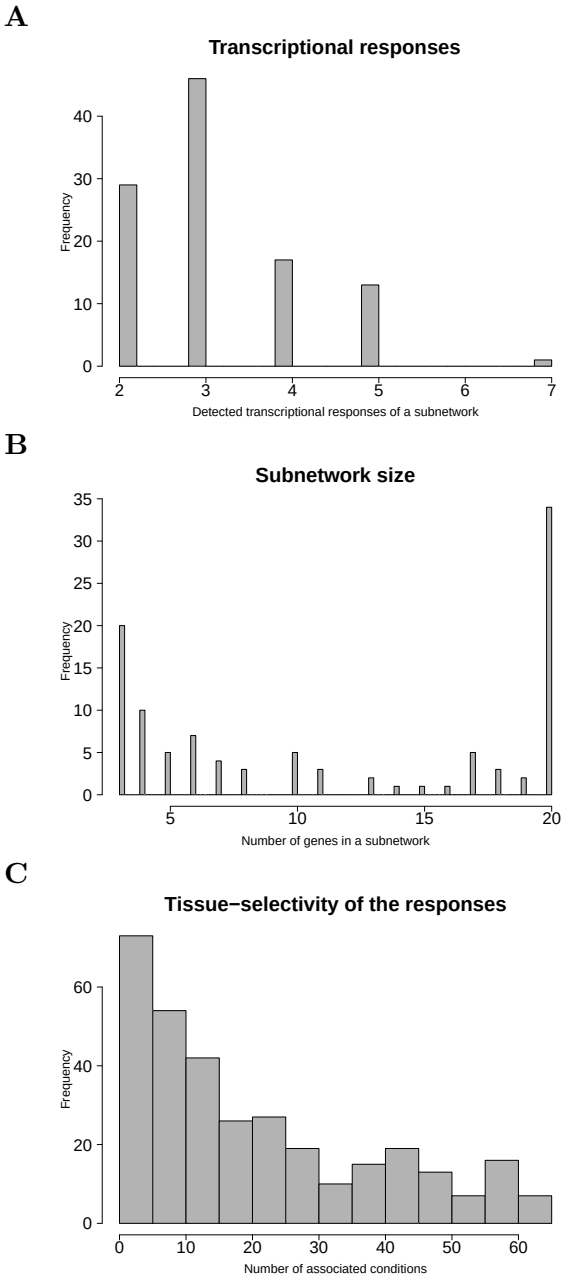

Figure 1: Histograms of model statistics: **A** Number of transcriptional responses in the subnetworks detected by NetResponse. **B** Subnetwork size. **C** Number of tissues associated with each response.

	NetRespose	NetResponse (shuffled)	MATISSE+	MATISSE+ (shuffled)	SAMBA
Reproducibility (corr.)	0.76	0.14	0.64	0.60	0.68
Reproducibility (signif.)	0.80	0.43	0.72	0.71	0.63
Fraction of responses with validation data	0.81	0.34	0.79	0.80	0.89
Physiol. coh. (NMI)	0.50	0.22	0.49	0.41	0.46
Physiol. coh. (signif.)	$< 10^{-4}$	0.65	$< 10^{-4}$	$< 10^{-2}$	$< 10^{-4}$
Fraction of data assigned to subnetworks	0.68	0.72	0.45	0.40	0.45

Table 1: Comparison statistics. *Reproducibility (correlation):* Median correlation between the detected responses and the corresponding tissues in the validation data. *Reproducibility (significance):* Fraction of transcriptional responses that were reproducible in the validation data (GlobalTest $p < 0.05$). The results are shown for the responses where corresponding tissues in the validation data were available. Significance of differential expression was calculated for each pairwise comparison between the associated tissues of the predicted responses in the validation data. *Transcriptional responses with validation data:* Fraction of transcriptional responses for which corresponding samples were available for testing in the validation data. *Physiological coherence (NMI):* Normalized mutual information between the detected transcriptional responses and sample labels (tissues). A higher NMI indicates stronger association between the detected responses and tissues. The differences between NetResponse, MATISSE+, and SAMBA are not significant. *Physiological coherence (significance):* Significance of the physiological coherence (NMI) compared to the expectation based on randomly labeled samples (Wilcoxon test p-value). *Fraction of data assigned to subnetworks:* Fraction of genes participating in the detected responses.

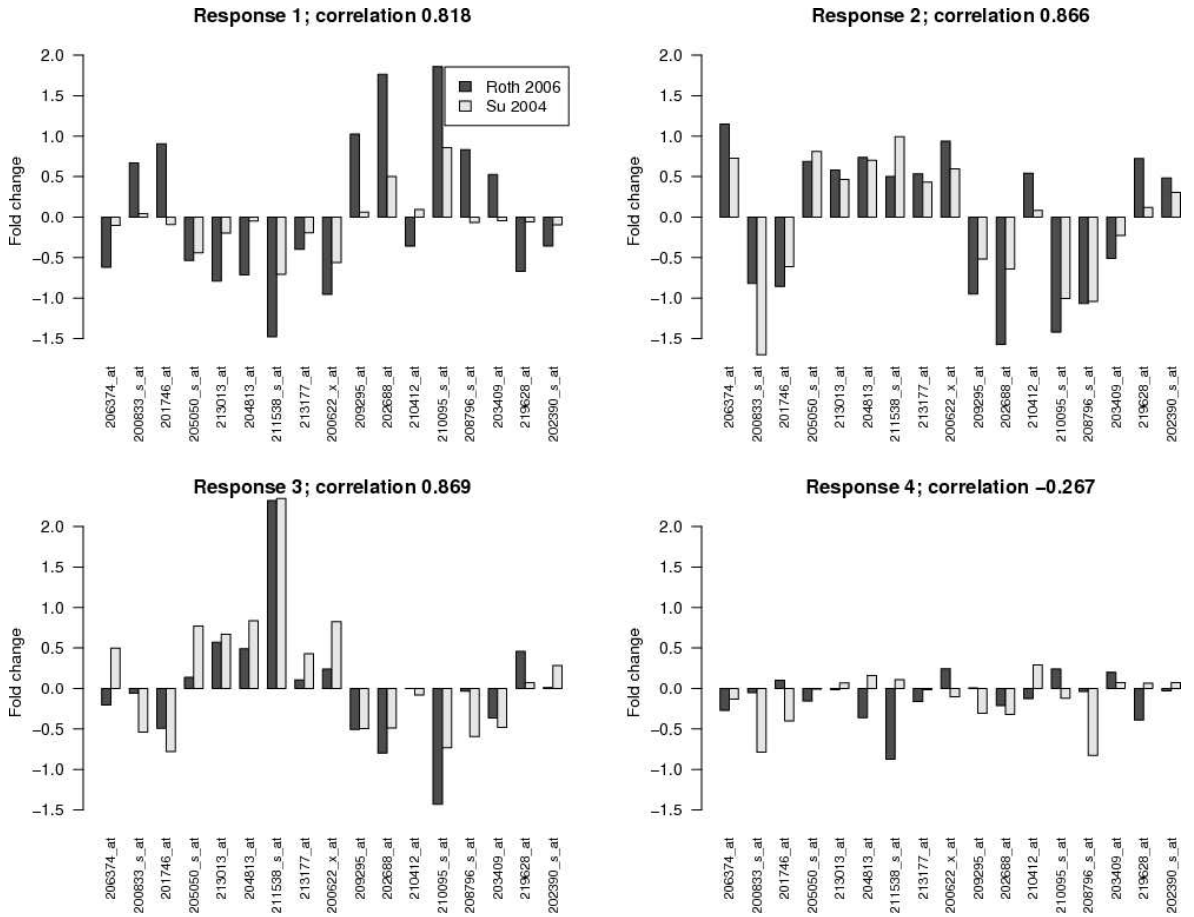

Figure 2: Reproducibility of transcriptional responses of the subnetwork of Figure 1 (main text) in independent validation data. *Correlation:* Qualitatively similar responses are observed in the validation data (Pearson correlation > 0.8), except for the fourth response (correlation -0.27). Differential expression with respect to the mean level of each gene is used in the comparisons. This removes overoptimistic bias in the correlations caused by the systematic differences in the expression levels of the genes. *Significance:* Each response is associated with a subset of tissues. The differences between the corresponding tissues are statistically significant ($p < 0.01$; GlobalTest) in the validation data for each pairwise comparison between the predicted four groups of tissues. The investigated gene expression atlas (Roth et al., 2006) and the validation data (Su et al., 2004) have been measured on different array platforms (HG-U133Plus2 and HG-U133A, respectively). Gene expression levels are here shown for the 17 (out of 20) probesets that are available on both platforms.

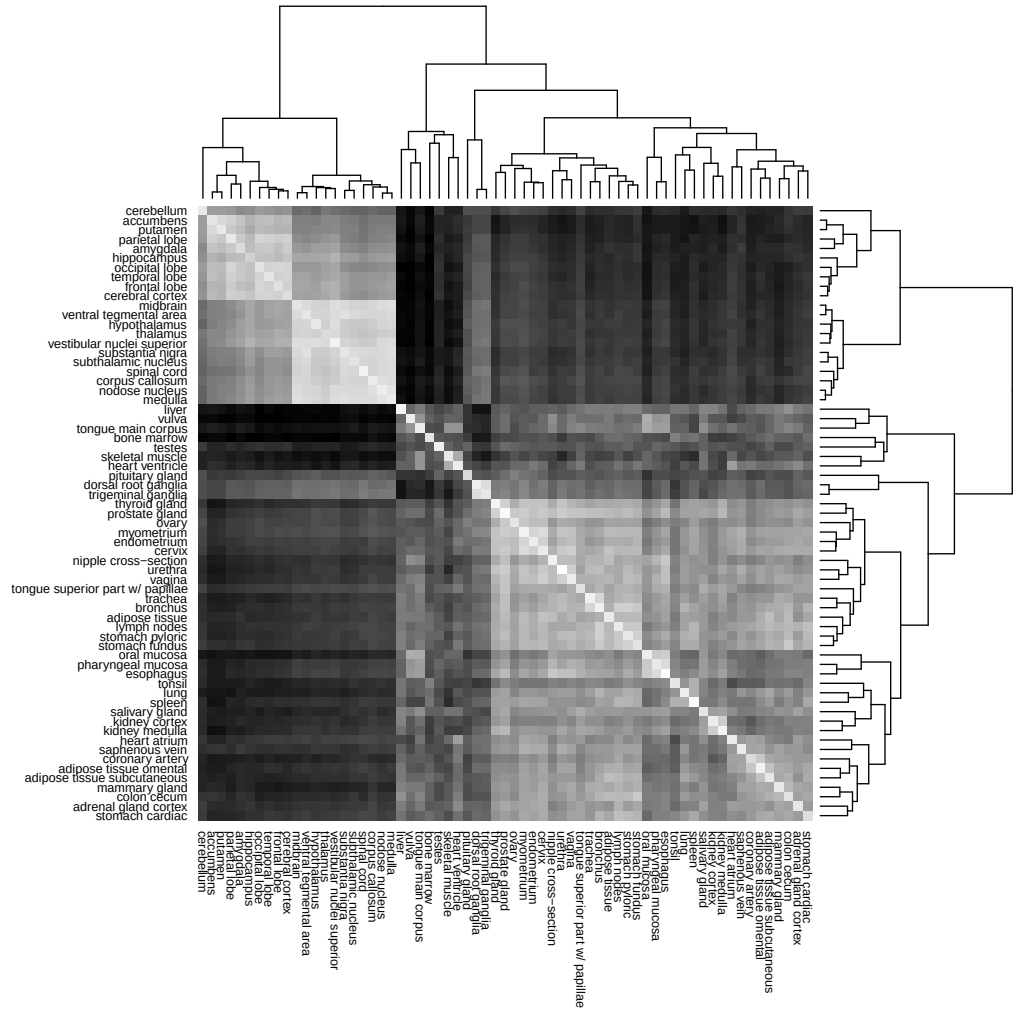

Figure 3: Tissue connectome based on the detected transcriptional responses of the human pathway interaction network. For each pair of tissues the overall probability of shared transcriptional response across the network is shown (black: $P = 0$; white: $P = 1$; see main text for details). This gives a probabilistic measure of tissue similarity based on network activation. The rows and columns are ordered with hierarchical clustering to highlight the relatedness between tissues.

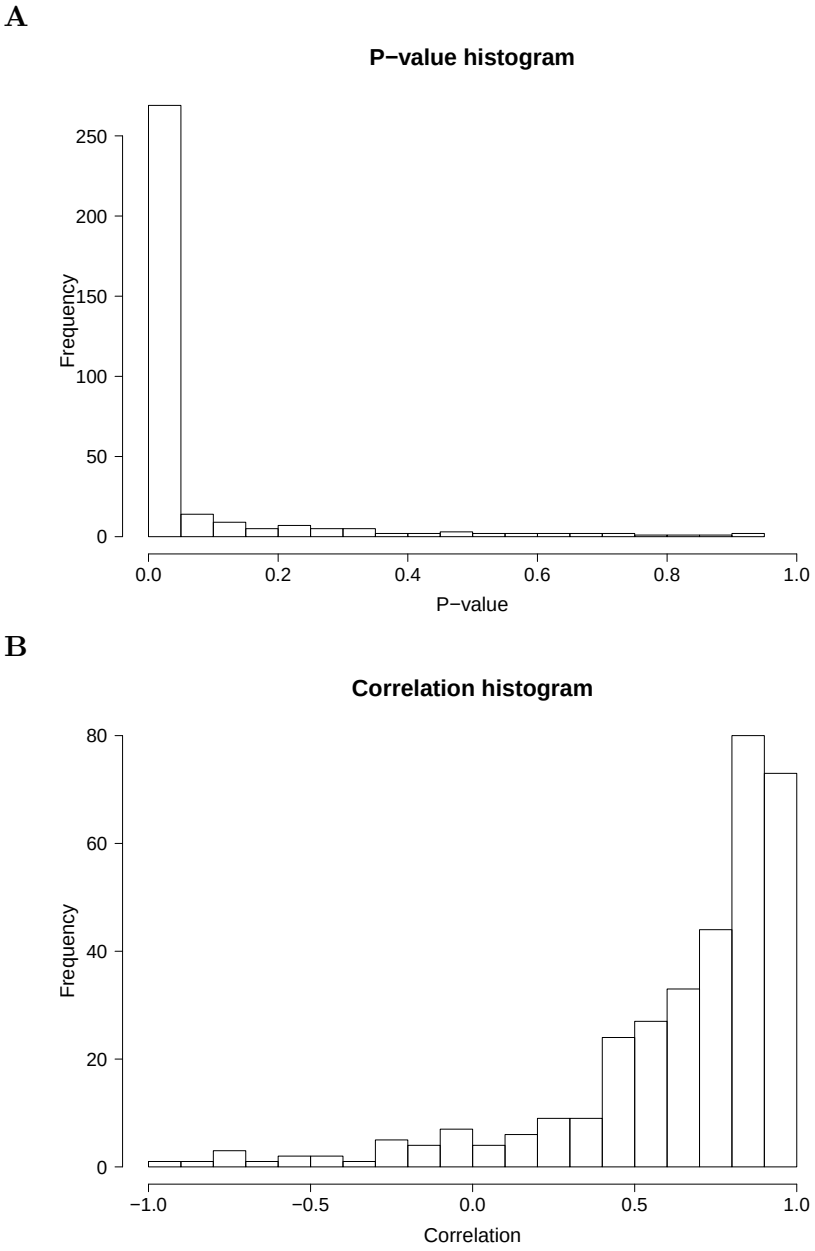

Figure 4: Reproducibility of the detected transcriptional responses in the independent Su et al., 2004 validation data in terms of significance and correlation. **A** Significance of differential expression between each pair of associated tissues for predicted responses in the validation data. 80% of the predicted differences between the tissues were verified in the validation data with $p < 0.05$ (GlobalTest). We tested only the responses where corresponding tissues were available in the validation data (81% of the responses). **B** Correlation between the detected responses in the investigated data set and the corresponding tissues in the validation data. Differential expression with respect to the mean level of each gene was used in the comparisons. This removes the potential bias in the correlations caused by the systematic differences in the expression levels of the genes.

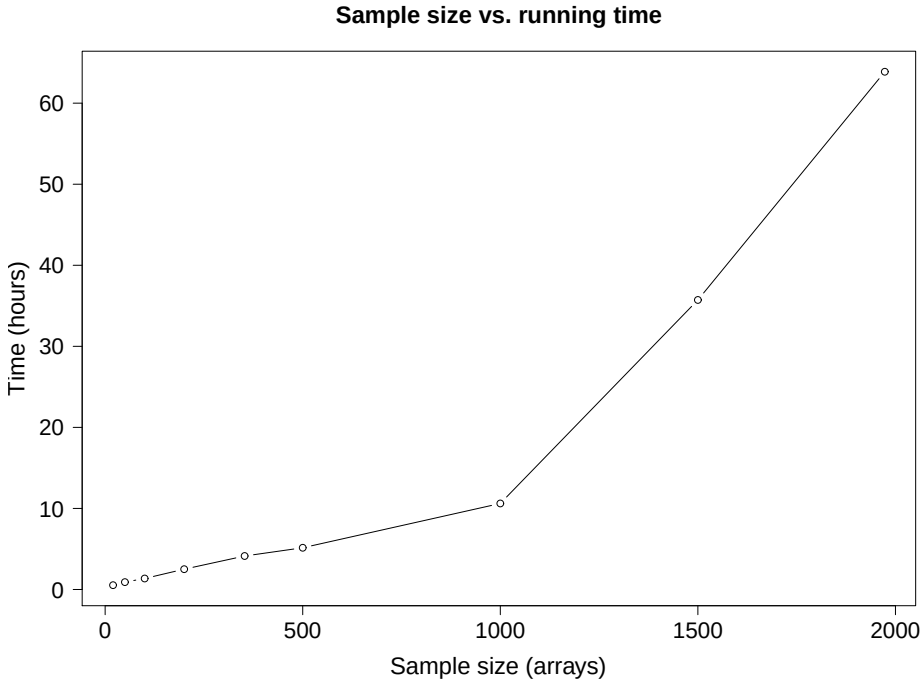

Figure 5: Running time for data sets of different sizes on the pathway network described in the main text. The running time for the GSE3526 data set investigated in the main text was 248 minutes (i.e. 4.1 hours). Computation time increases superlinearly with sample size from 33 minutes with 20 samples to 64 hours with 1977 samples. Model fitting in the algorithm can be parallelized, which will make the model scalable to larger data sets in standard multi-core desktop computers. The running time depends also on the size and connectivity of the network. Our investigated network represents a standard pathway network used in current organism-wide studies. The network has a median of 5 and a maximum of 105 direct interaction partners per gene. This reduces the search space considerably compared to models that would consider all potential interactions between the 1800 network genes. To investigate time consumption we have selected random subsets of various sizes (20, 50, 100, 200, and 353 samples) from the GSE3526 data, described in the main text and having 353 arrays in total. The data sets with 500 and more (1000, 1500, 1973) samples were obtained by picking random subsets from the GSE2109 data set, which has 1973 arrays in total (downloaded 30.5.2008 from http://www.ncbi.nlm.nih.gov/geo/). Both data sets were preprocessed as described in Section 2.3 in the main text.