

A modular approach for integrative analysis of large-scale gene-expression and drug-response data

Zoltán Kutalik^{1,2}, Jacques S Beckmann^{1,3} & Sven Bergmann^{1,2}

High-throughput technologies are now used to generate more than one type of data from the same biological samples. To properly integrate such data, we propose using co-modules, which describe coherent patterns across paired data sets, and conceive several modular methods for their identification. We first test these methods using *in silico* data, demonstrating that the integrative scheme of our Ping-Pong Algorithm uncovers drug-gene associations more accurately when considering noisy or complex data. Second, we provide an extensive comparative study using the gene-expression and drug-response data from the NCI-60 cell lines. Using information from the DrugBank and the Connectivity Map databases we show that the Ping-Pong Algorithm predicts drug-gene associations significantly better than other methods. Co-modules provide insights into possible mechanisms of action for a wide range of drugs and suggest new targets for therapy.

The number of ways to measure the properties and the behavior of biological systems is multiplying rapidly. In particular, microarray technologies have firmly established themselves as a standard tool in biological and biomedical research. They provide cost-efficient but noisy 'snapshots' of the genome-wide transcriptional state of cells. Combining many such snapshots of cells in different conditions has been instrumental for the study of the more global aspects of cellular systems biology¹⁻⁴.

Although many methods⁵⁻¹¹ have been suggested for extracting global properties from massive data, most algorithms only infer information on the structure of one data set at a time. However, with the advent of high-throughput data covering different aspects of gene regulation, as well as other properties of the samples, there is an increasing need for combined analysis of multiple noisy data sets. A particularly challenging application pertains to data sets where the

same cell samples have been studied using assays that probe different aspects of their phenotypes. Pioneering work by Weinstein *et al.*¹² attempted to relate activity profiles of a large collection of compounds across 60 human cancer cell lines with molecular structural features of the tested agents. Their goal was to provide insight into potential target molecules and modulators of activity within these cell lines. Subsequently these cell lines were also characterized by their gene-expression levels using microarray analysis of untreated samples¹³, protein profiles¹⁴ and microRNA expression levels¹⁵. Several studies tried to integrate these data with the drug-response profiles^{14,16,17}. Here we propose integrative analysis methods that take into account the complexity of these kinds of large-scale data and investigate systematically the power of these methods to infer drug-gene interactions and transcriptional response to drugs.

Partitioning similar elements (e.g., genes, drugs, samples) of large data sets into clusters or modules reduces the complexity of the data and facilitates their analysis. First, studying the properties of a limited number of modules is simply more feasible than, for example, looking at the expression levels of thousands of genes over a multitude of conditions. Second, quantitative measures for a module, like the mean expression level of its genes, are much more robust than the measures for single elements, because fluctuations tend to cancel each other out. This is particularly relevant for the noisy data produced by chip-based high-throughput technologies.

The Iterative Signature Algorithm (ISA)¹⁸ is one of the state-of-the-art methods for modular analysis of large-scale data (typically tens of thousands of gene probes tested over hundreds of conditions) according to various criteria^{19,20} and has been used for numerous biological studies^{19,21-23}. Briefly, the ISA identifies from a set of expression data a compendium of transcription modules consisting of co-expressed genes as well as the experimental conditions for which this coherent expression is the most pronounced¹⁸. This algorithm has several advantages. (i) Genes and samples can be assigned to multiple modules (whereas standard clustering produces mutually exclusive units). This is well motivated from the biological point of view, because splice variants may hybridize to the same probe and the same gene can function in several processes, which are induced under different experimental conditions. (ii) Requiring only coherent gene expression over a subset of arrays allows for picking up subtle signals of context-specific and combinatorial co-regulation. Given the experimental noise in microarray

¹Department of Medical Genetics, and ²Swiss Institute of Bioinformatics, University of Lausanne, Rue de Bugnon 27 - DGM 328, CH-1005 Lausanne, Switzerland. ³Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois, Av. de Pierre-Decker 2, CH-1011 Lausanne, Switzerland. Correspondence should be addressed to S.B. (Sven.Bergmann@unil.ch).

Received 23 January; accepted 1 April; published online 8 May 2008;
doi:10.1038/nbt1397

data, these signals may be too weak to be extracted from the correlations over all samples that are used by many clustering algorithms. (iii) Because the ISA does not require the calculation of correlation matrices, it is highly efficient computationally and is thus applicable even to very large data sets.

Most algorithms that aim at integrating massive data from different sources do so only in a sequential manner. For example, groups of genes that have been assigned to a cluster are frequently tested for "significant enrichment"²⁴ with genes of predefined groups (e.g., those having the same functional annotation or belonging to a cluster of a different data set). Yet, such a procedure integrates the information from other data sets only *a posteriori* rather than producing coherent groups of genes by co-analyzing several data sets at the same time. Integrative analysis of the NCI-60 data set was initiated by the

development of the COMPARE software²⁵. Since then several authors have devised methods for such analysis^{14,16,17}. Yet, similarly to COMPARE, they all used simple correlations as a measure of similarity across data sets. Staunton *et al.*¹³ presented promising methodological development by substituting correlations with the more robust *t*-statistics and applying leave-one-out cross-validation to define classifiers. A more advanced method to predict drug response from gene-expression data has been proposed by Ma *et al.*²⁶, who applied classifiers such as Random Forests and Nearest Neighbor methods. However, these classifiers determine predictors for each drug independently, providing no means to structure or simplify the drug-response data. The MA-Networker algorithm²⁷ was used to combine gene expression with chromatin-immunoprecipitation data using regression analysis. The advantage of this approach over correlation analysis is that similarity is not required over the entire profile. However, this method provides no information about the most relevant cell lines and also does not attempt to reduce the complexity of the data. Other methods that integrate different types of data, like Bayesian networks^{28–30}, are less suited for very large data sets, as these algorithms require massive computation power.

Here, we propose extending the modular analysis approach from one to two large-scale data sets that share one common dimension. Specifically, we consider the case of a combined set of gene-expression and drug-response data, as provided by NCI-60 (where the cell lines represent the common dimension). We investigate three distinct extensions of the ISA for the combined modular analysis of these data. The goal of these algorithms is to break down the massive sets of data into smaller building blocks that exhibit similar patterns across certain genes and drugs in some of the cell lines. Throughout the paper we will use the term *co-module* for such a building block generated from two data sets. A co-module is essentially a (weighted) ensemble of certain genes, drugs and cell lines such that its genes are expressed across its cell lines and its drugs induce similar response profiles across the same

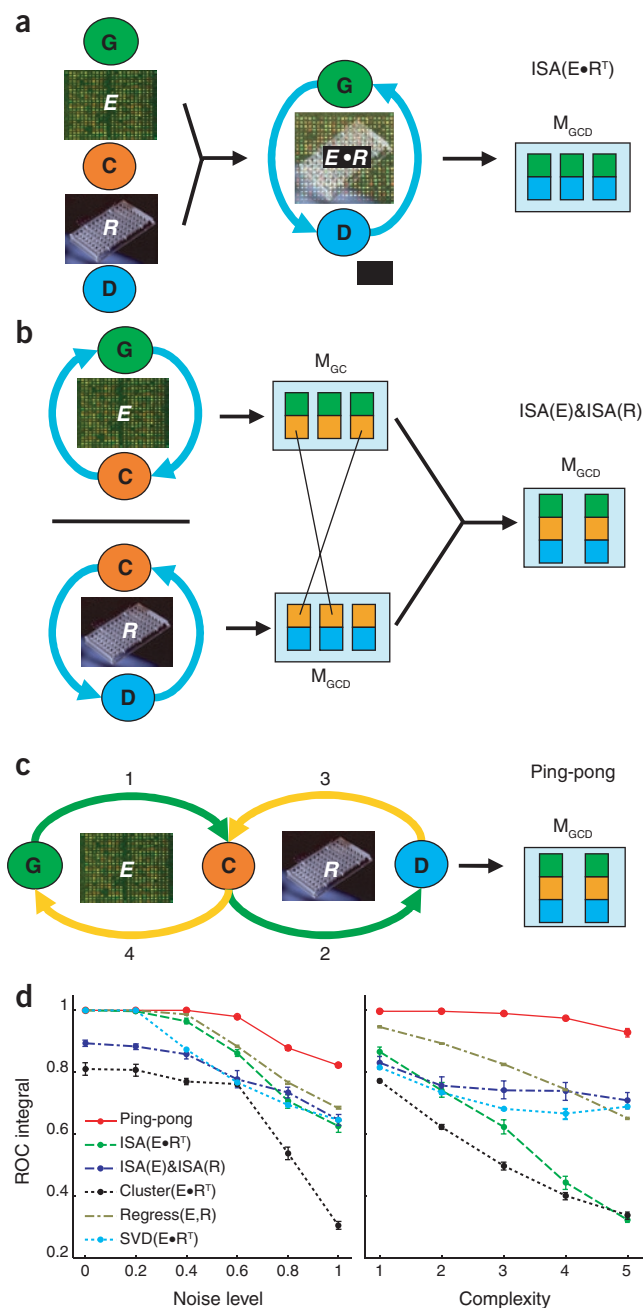


Figure 1 Different modular algorithms and *in silico* performances for paired data sets. The goal of each scheme is to identify subsets of genes (G) and drugs (D) for which (some of) the cell lines (C) exhibit similar profiles in both the gene-expression data E and drug-response data R . These subsets are termed 'co-modules' and denoted M_{GD} (or M_{GCD} , if defined for a subset of cell lines). **(a)** ISA($E \cdot R^T$) performs a modular analysis of the correlation matrix $E \cdot R^T$ between genes and drugs over all cell lines. **(b)** ISA(E)&ISA(R) first extracts transcription modules from E and drug-response modules from R and then matches those that include similar sets of cell lines to construct co-modules. **(c)** The Ping-Pong Algorithm (PPA) iteratively refines coherent patterns across both data sets by alternating between them until convergence to co-modules is reached (see **Supplementary Notes** for details). **(d)** The three modular approaches, as well as hierarchical clustering, regression-based 'association', and SVD of the correlation matrix $E \cdot R^T$, were tested on *in silico* data sets. These data were generated according to a simple model that allows for groups of genes to be co-expressed in certain sets of cell lines and for groups of drugs that affect the growth of certain sets of cell lines. Some of these sets are the same for both genes and drugs whereas others are independent. The goal was to reconstruct from these data the sets of genes and drugs that exhibit coherent expression and response profiles, respectively, in the same sets of cell lines. Performance was evaluated in terms of the area under the receiver operating characteristic (ROC) curve when comparing the underlying structure of the model with the drug-gene associations predicted by the different analysis approaches (**Supplementary Notes**). The graph on the left shows the performance as a function of the noise level. The graph on the right compares the performance as a function of complexity (defined by the 'cross-talk' between the *in silico* modules in terms of common genes, drugs or cell lines). Within this modeling framework the PPA performs best and all modular approaches outperform hierarchical clustering.

cell lines. Representing coherent features across both data sets in terms of such co-modules reduces the complexity of the data. This modular reduction makes it generally easier to study the underlying biology and allows prediction of more robust drug-gene associations.

We first explore and compare these methods, as well as several existing algorithms, using *in silico* data generated according to a model that simulates gene-expression and drug-response data with well-defined and tunable interdependencies. We show that our modular approaches outperform standard algorithms in recovering the implanted drug-gene links when considering noisy or complex data. We then provide an extensive study comparing the output of all methods when applied to the real gene-expression and drug-response data from the NCI-60 cell lines. Using direct information on drug-gene associations from DrugBank³¹ and the Connectivity Map³², we establish that our Ping-Pong Algorithm performs substantially better in predicting such links from the NCI-60 data than nonmodular algorithms. We provide a resource (see Results) of NCI-60 co-modules. This resource allows for a more global and modular outlook at tissue-specific drug-gene interactions, some of which we explore in more detail. Importantly, this co-module resource can be used to predict yet undiscovered interactions using only gene-expression data of untreated cells.

RESULTS

Overview of suggested methods

We considered three modular approaches that follow different strategies. The first method, ISA($E \cdot R^T$), applies the ISA directly to the matrix of pair-wise correlations $E \cdot R^T$ between gene-expression data E and drug-response profiles R across all cell lines (Fig. 1a). The second approach, ISA(E)&ISA(R), first performs independent modular analyses of each data set (giving transcription modules and drug-response modules) and then matches these modules according to the sets of cell lines they contain (Fig. 1b). Finally, the third method, termed Ping-Pong Algorithm (PPA), iteratively refines coherent patterns across both data sets by alternating between them until convergence to co-modules is reached (Fig. 1c).

In silico results

We used receiver operating characteristic (ROC) curves³³ as a measure of performance for benchmarking the different algorithms. Our analysis revealed that all our modular approaches extracted co-modules with higher sensitivity and specificity than standard hierarchical clustering when applied to simulated gene-expression and drug-response data with high levels of noise (Fig. 1d, left panel) or with overlapping modules (Fig. 1d, right panel). Within our benchmarking, the Ping-Pong Algorithm (PPA) performed best in recovering the proper co-modules from *in silico* data with increased noise and complexity. Notably, ISA(E)&ISA(R) performs better than

ISA($E \cdot R^T$) for data of high complexity, indicating that a modularization before linking the two data sets is beneficial in this case. Singular value decomposition, SVD($E \cdot R^T$), performs similarly to ISA(E)&ISA(R) under conditions of high noise and complexity, but not as well as the PPA. It is important to note that these results, of course, hold only within our modeling assumptions (as is true for any *in silico* testing approach). Although we believe that our model captures two major hallmarks of the biological data (that is, complexity and noise), we did not attempt to model other features like modular hierarchies or non-Gaussian noise.

NCI-60 results for drug-gene associations

The PPA was then applied to the NCI-60 data, producing 859 co-modules. Figure 2a illustrates one particular co-module (discussed in more details below). Figure 2b provides basic statistics of the generated output in terms of (i) the distribution of the numbers of co-modules

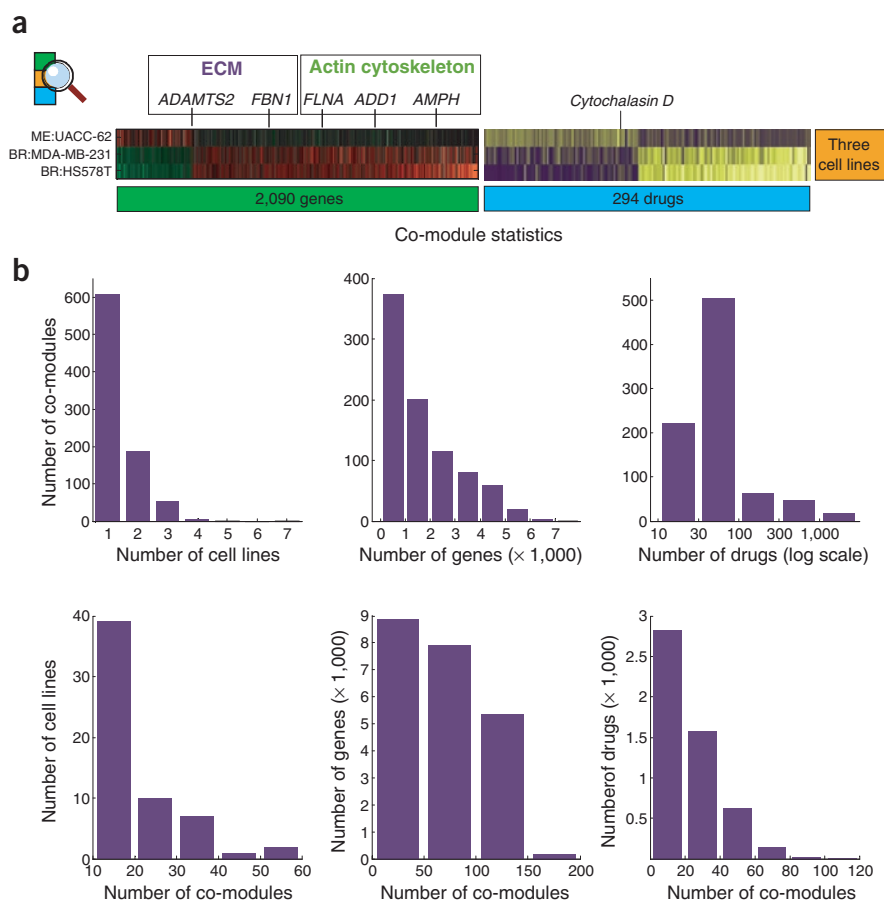
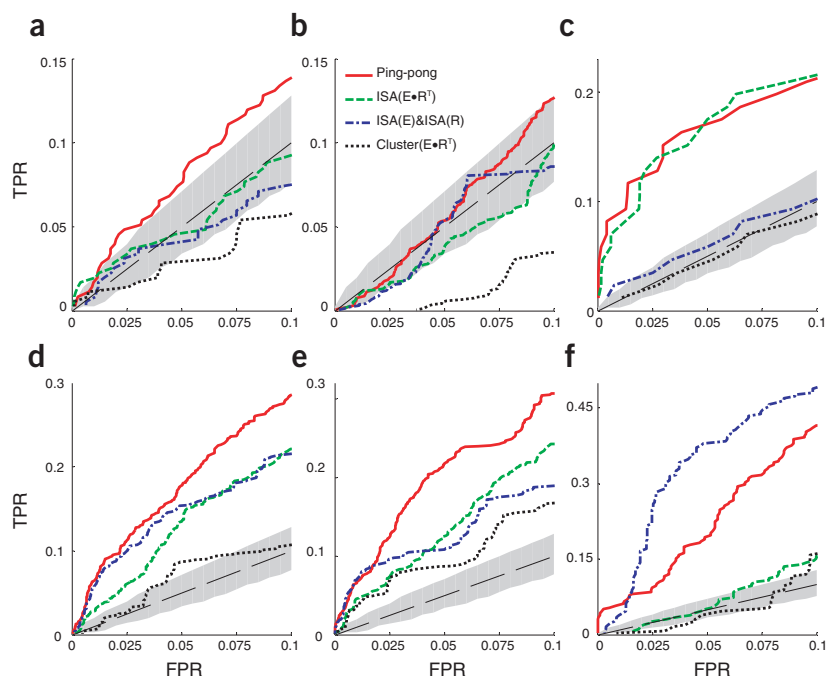


Figure 2 Co-module properties. **(a)** Cytochalasin D co-module. Schematic representation of a particular co-module containing 2,090 genes, 294 drugs and 3 cell lines. The corresponding gene scores are indicated by a color code ranging from green (under-expressed) to red (overexpressed). Note that the two cell lines derived from breast cancers (BR) exhibit similar profiles, whereas the melanoma (ME) cell line has an inverse profile. Similarly, the drug-response profiles have coherent drug scores for breast cancers (blue and yellow indicating sensitivity and resistance, respectively) and reverse scores for the melanoma line. The genes of this co-module are functionally enriched for those encoding actin cytoskeleton and extracellular complex. This enrichment is plausible because the co-module also contains the drug cytochalasin D, which is known to affect these cellular components¹³. **(b)** Co-module statistics. Histograms in the first row illustrate the global distributions for the number of cell lines, genes and drugs attributed to co-modules. More modules were obtained for higher thresholds and contain, therefore, fewer genes, conditions or drugs. The number of drugs is shown on a log scale. Histograms in the second row show in how many co-modules cell lines, genes or drugs occur simultaneously.

Figure 3 Evaluation of different approaches for integrative analysis of high-throughput expression and response data from the NCI-60 study. Drug-gene associations were imputed from the expression and response data gathered for the tumor cell lines using the three modular approaches as well as hierarchical clustering to predict drug-gene associations solely based on high-throughput data (Fig. 1). Note that Regress(E,R) and SVD(E·R^T) do not produce clusters and therefore can be included in the comparison only for drug-gene associations (Supplementary Notes). Each method was applied to provide continuous scores for the likelihood of these interactions. Performance was evaluated by receiver operating characteristic (ROC) curves showing the increase of the true-positive rate (TPR) as a function of the false-positive rate (FPR). The gray area shows the 90% confidence interval for the ROC curve of a random predictor. Predictions were compared with different types of independent experimental evidence. (a,b) Direct (but very sparse) associations between drugs and genes according to the DrugBank (a) and using expression data for four drug-treated cell lines from the Connectivity Map project (b). (c–f) Indirect associations of drugs with predefined groups of genes were predicted using the average of the drug scores over all co-modules, weighted by their enrichment with the genes of these groups (see Supplementary Notes for details). These group scores were then compared with evidences for drugs affecting (c) certain pathways (according to the KEGG database) as well as the GO annotations referring to biological process (d), molecular function (e) and cellular component (f). Each plot shows the region of ROC curves for the critical value of FPR < 10% (see Supplementary Notes for ROC curves in the range of FPR < 20% and comparisons for the area under these curves). All modular approaches outperform hierarchical clustering, and the PPA performs best in most of the cases.



according to the number of cell lines, genes and drugs they contain and, conversely, (ii) the distribution of the numbers of cell lines, genes and drugs according to the number of co-modules they appear in. More than one-third of the genes occur in <50 co-modules, whereas a few genes act as ‘hubs’ by being part of up to 200 different co-modules. Similarly, the majority of co-modules contain <2,000 genes, whereas only a few of them have >5,000 genes.

Next, the PPA was compared to several other algorithms again using ROC curves and two different sources on drug-gene interactions. The superiority of the PPA was even more evident for the analysis of real data than for our *in silico* study (Fig. 3). In fact, only the predictions for drug-gene associations generated by the PPA from the NCI-60 data generated a ROC curve that was consistently above the 90% confidence area for the control curves when using DrugBank as reference (Fig. 3a). The control curves were generated from a random predictor that establishes gene-drug interactions purely by chance. These random predictors generate ROC curves with an area under the curve (AUC) for a false-positive rate <0.1, $AUC(\text{rand}) = 0.005$ (with 90% confidence interval = [0.0033, 0.0067]). For the PPA, $AUC(\text{PPA}) = 0.0079$, which is significantly better ($P < 0.006$) than the AUC of random predictors.

Using the Connectivity Map as an alternative reference, we found that only the predictions produced by the PPA had a significant ($P < 0.022$) true-positive rate (TPR) relative to random predictions for $FPR < 0.1$ (Fig. 3b) and also for $FPR < 0.2$ (Supplementary Notes online). A probable cause is that only 77 (1.5%) of the drugs are represented in the NCI-60 response data, and this only for four common cell lines.

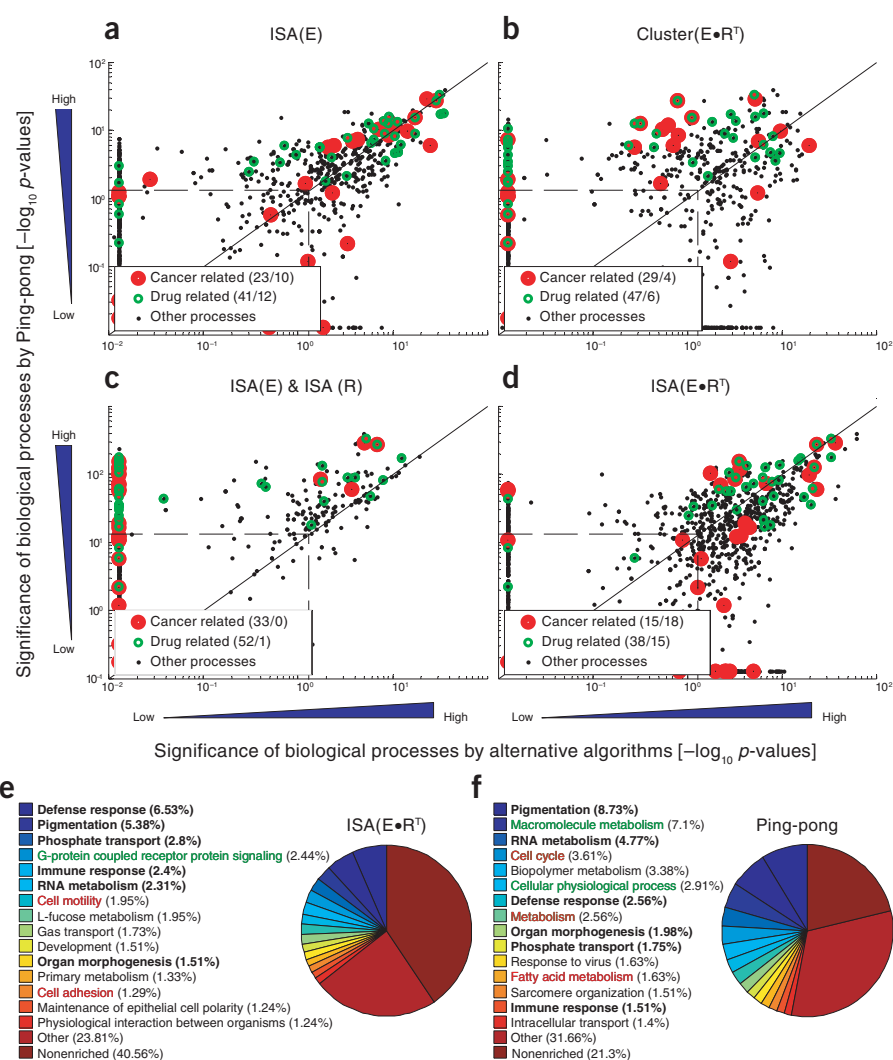
In addition, we applied regression, Regress(E,R)²⁷, to infer drug-gene associations, which yielded ROC curves that were intermediate between PPA and ISA(E·R^T) both for DrugBank and Connectivity

Map. We also evaluated SVD(E·R^T), which performed slightly worse than modular methods, but better than standard clustering (see Supplementary Notes).

NCI-60 results for gene set–drug associations

It is important to note that current annotations for drug-gene associations contain many more ‘false’ than ‘true’ values. In fact, the true values are of a different nature than the false ones, because they have been experimentally confirmed, whereas the false values may just indicate an absence of evidence for a drug-gene association in DrugBank. Thus, predicting associations of drugs with gene sets rather than with single genes provides a much more robust measure for the quality of the predictions based on the NCI-60 data. In addition, our approach may generate drug-sensitivity proxies, which are downstream of the direct targets. The fact that these genes are often in the same functional class as the drug target is another reason for the better accuracy in predicting associations of drugs with gene sets. Indeed, the ROC curves for predicted associations of drugs with KEGG pathways³⁴ (Fig. 3c) using the PPA and ISA(E·R^T) indicate an impressive excess in the TPR with respect to random predictions, whereas ISA(E)&ISA(R) and the non-modular algorithms perform poorly. Thus, a direct coupling of the expression and response profiles appears to be instrumental in relating drugs to pathways. For gene ontology’s (GO²⁴) biological processes and molecular functions the PPA gave superior results (both $P < 10^{-4}$), but predictions made by the other modular approaches were also consistently better than those generated by the nonmodular approaches. Interestingly, for the cellular compound predictions in the region of the ROC curve where $FPR < 0.1$, ISA(E)&ISA(R) worked best, followed by the PPA. We note however that for a $FPR > 0.13$ the TPR of the PPA is higher than that of ISA(E)&ISA(R) (see Supplementary Notes).

Figure 4 Comparison of the different approaches with respect to the coverage of biological processes. For each of the various GO categories for biological processes, we computed significance scores ($-\log_{10}$ transformed Bonferroni-corrected P values) for enrichment with genes attributed to (co-)modules obtained by the different algorithms (Fig. 1). The highest scores across all modules were maintained to indicate how well each biological process was covered. (a–c) The scores for the PPA were then plotted against those of the other algorithms: ISA applied only to the expression data [ISA(E)] (a), Cluster(E·R^T) (b), ISA(E)&ISA(R) (c) and ISA(E·R^T) (d). Note that the significance scores are shown on a log-scale and the blue triangles indicate the direction of increasing significance. Dots appearing along the horizontal and vertical lines correspond to processes exhibiting no significant association with any (co-)module generated by the PPA and the indicated algorithms, respectively. Biological processes reported to be involved in cancer development³⁶ are marked with red dots, whereas drug-related ones (affected by at least five drugs according to DrugBank) are indicated in green. For both types of processes, we indicate in the legends how many were best represented by each algorithm (PPA/ alternative method). The 859 PPA co-modules tend to cover drug-related processes with higher sensitivity than those of the other algorithms. Cancer-related processes are captured with similar sensitivity as by the 2,251 ISA(E·R^T) modules, but with higher significance. Other processes were covered with similar accuracy, except for ISA(E)&ISA(R) which identified only 27 co-modules. (e–f) The fraction of biologically meaningful modules generated by ISA(E·R^T) (e) and PPA (f) were computed to compare the specificity of these algorithms. For each module the best fitting significant biological process (if any) was computed to determine the number of modules most enriched for each biological process. Pie charts show the top 15 biological processes for which ISA(E·R^T) and PPA produced the largest number of matching modules. A much higher fraction of the PPA modules enriched with cancer- and drug-related processes, and a much lower fraction of the PPA modules (21%) is not enriched for any biological process compared to the ISA(E·R^T) modules (41%).



Functional biases in (co-)modules

The functional analysis presented in Figure 4a–d indicates that GO categories pertaining to drug-related processes (according to DrugBank) were matched best in terms of the co-modules generated by the PPA. This effect is most pronounced when comparison is made with the transcription modules generated by the ISA using only expression data (ISA(E)) (Fig. 4a), but is also seen with clusters of the correlation matrix $E \cdot R^T$ (Fig. 4b) or with matched modules (ISA(E)&ISA(R)) (Fig. 4c). For cancer-related processes this bias is somewhat weaker. Interestingly, the 2,251 ISA(E·R^T) modules (Fig. 4d) cover drug-related processes almost as well (whereas cancer-related processes are covered even slightly better on average). Yet, it should be noted that these modules outnumber the 859 PPA co-modules by more than a factor of 2.5, such that they achieve similar performance only at the price of lower global significance.

Our analysis suggests that the perpetual bias for selecting cell lines with similar drug responses used by the PPA is more sensitive in generating functional sets of genes pertaining to drug response. We therefore compared the coverage of the two most sensitive algorithms (that is,

the PPA and ISA(E·R^T)). The pie charts in Figure 4e,f indicate that a much larger fraction of PPA co-modules than of the ISA(E·R^T) modules appears to be biologically meaningful: the fraction of co-modules that are not enriched for any functional category (21.3%) is almost half of that for ISA(E·R^T) modules (40.6%). We also note that >10% of the co-modules were associated with drug-related processes (including 7.1% that appeared to be involved in macromolecule metabolism), whereas this fraction was <3% for the ISA(E·R^T).

Co-modules predict transcriptional response to drug treatment

The Connectivity Map data provide information on differential gene expression upon drug administration for a limited set of drugs in four cell lines. Above, we used these data as an alternative reference to DrugBank to test our drug target predictions. Now, we use it to study to what extent one can predict drug-induced transcriptional response from the NCI-60 data. To this end, we searched for pairs of drugs and cell lines covered by the Connectivity Map that appear in the same co-modules. In fact only three such pairs could be identified. They include the three drugs cytochalasin B, tamoxifen and colforsin, all paired with

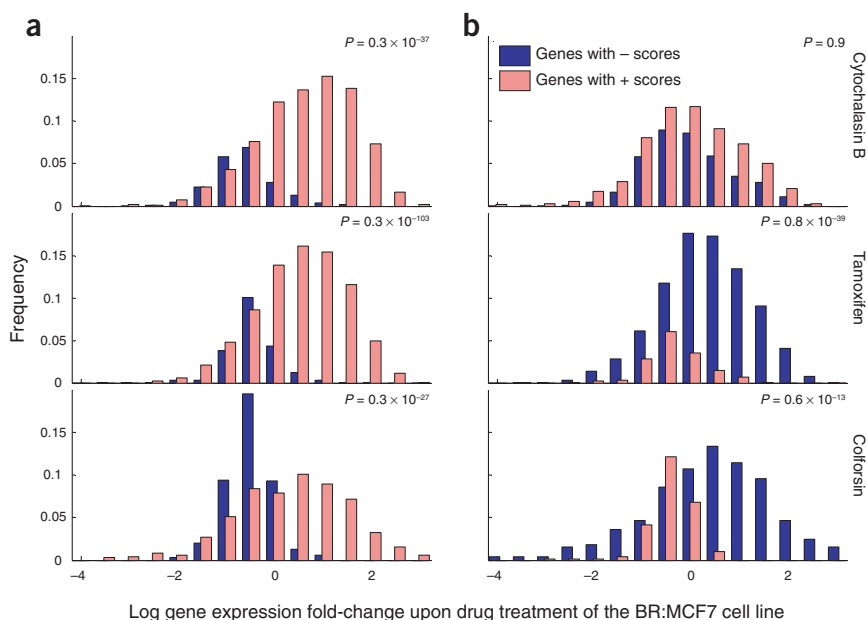


Figure 5 Connection between co-module scores and Connectivity Map data. **(a)** All co-modules were considered that include pairs of a cell line and a drug that were both covered by the Connectivity Map data and had the largest number of cell lines. For example, the pair BR: MCF7–cytochalasin B was found in co-module no. 842. Genes were segregated into a group with positive (2,046) and a group with negative (527) gene scores. The plots show the distribution of the Connectivity Map’s log gene expression fold-change values—observed when the indicated drug was administered to the given cell line relative to the unperturbed expression—separately for these two groups. **(b)** As a more naive, alternative approach, we established how the NCI-60 gene expression profile of different genes correlated with the drug’s sensitivity profile over all cell lines and—to be comparable—selected the same number of most-correlated and anti-correlated genes giving exactly the same number of nonzero gene scores as the PPA co-module. The respective distributions indicate that our PPA approach separates the Connectivity Map expression values much better.

the cell line BR:MCF7. Note that due to this little overlap between the NCI-60 co-modules and the Connectivity Map data the results presented in the subsequent paragraph should be interpreted with care.

The modules providing these links between drugs and cell lines also contain gene sets, which reflect only the NCI-60 expression data of the untreated samples. These genes can be viewed as proxies for the sensitivity of a cell line to the given drug. Remarkably, we observed that their gene scores (indicating their strength of association with a given module, see Methods and **Supplementary Notes** for details) were significantly correlated with the expression fold-changes upon treatment obtained from the Connectivity Map. This is demonstrated in **Figure 5a**, showing separately the distribution of log fold-change values for genes with positive and negative module gene scores for the three aforementioned drugs. In all cases the mean values of the two distributions were significantly different (see indicated P values obtained by t -test). The segregation was much weaker when we grouped the Connectivity Map log fold-change values of the genes based on the sign of correlation between their NCI-60 gene-expression profile and the growth-response profile of each particular drug (**Fig. 5b**). The stronger, co-module-based separation indicates that sets of genes coherently overexpressed in a subset of untreated cell lines tend to be expressed even more when these lines are treated with drugs for which they exhibit similar growth response. A possible explanation is that the genes of certain pathways are already induced in the untreated cell lines (at a level sufficient for attributing the respective genes to co-modules) and that these pathways are then upregulated when the cell lines are exposed to the drug. Thus, in principle, the PPA can be used to predict gene expression change upon drug treatment based on gene expression in untreated samples.

Module database and examples

We established an online database of all co-modules which is accessible at <http://serverdgm.unil.ch/bergmann/PingPong.html>. This database provides a structured resource to explore possible sample-specific drug-gene interactions encoded in the NCI-60 data. Studying our co-module collection, we found many interesting links that were already supported by the literature (**Supplementary Notes**), demonstrating the strength of the PPA for generating hypotheses. In the following we discuss several examples.

It was previously found that a significant fraction of genes successfully predicting sensitivity to cytochalasin D (CAS number 22144-77-0) is related to the cytoskeleton and extracellular complex¹³. Altogether we found 39 co-modules containing cytochalasin D. From these co-modules we singled out the one for which the relative weight of cytochalasin D was the highest. This selected co-module contains three cell lines (ME: UACC-62, BR: MDA-MB-231, BR: HS578T), 292 drugs and 2,090 genes that were significantly enriched for the actin cytoskeleton ($P < 10^{-5}$) and the extracellular matrix (ECM) ($P < 10^{-4}$) among the cellular components, for regulation of actin cytoskeleton ($P < 0.05$) and ECM-receptor interaction ($P < 0.05$) pathways and for actin cytoskeleton organization and biogenesis process ($P < 0.01$). The genes of this co-module were furthermore enriched for cell motility ($P < 0.05$) and blood vessel development ($P < 0.01$), both of which were found to be influenced by cytochalasin D^{35,36}.

Next, we revisited the co-module we used to predict transcriptional response (**Fig. 5a,b**, middle panels) to the estrogen receptor antagonist tamoxifen (CAS number 10540-29-1), a drug to which all co-module breast cancer cell lines were sensitive (scores < -0.72). We found that a highly significant fraction of genes in this co-module were associated with G protein-coupled receptor activity ($P < 10^{-18}$). The co-module also contains the metastasis suppressor gene *KISS1* with a negative gene-score. Indeed, expression levels of *KISS1* and its G protein-coupled receptor in breast cancer are suppressed by estrogens, but can be restored by tamoxifen owing to its ability to block estrogen receptors³⁷. Moreover, low expression levels of *KISS1* and its G-protein-coupled receptor have been shown to be associated with sensitivity to tamoxifen treatment³⁷.

Another interesting co-module contained only leukemia cell lines that were all sensitive (scores < -0.75) to the drug ATRA (all-trans retinoic acid, CAS number 302-79-4). Indeed, ATRA has been used successfully for treating acute promyelocytic leukemia (AML)³⁸. ATRA is an antineoplastic and differentiating agent, which is in accordance with the fact that co-module genes have a significant overlap with numerous cell differentiation processes ($P < 0.05$). Moreover, the drug is thought to target the intracellular membrane-bound organelle, a cellular component that is associated with a highly significant portion of the co-module genes ($P < 10^{-20}$). In addition, the *AKT1* gene and

five phosphoinositide-3-kinases (*PIK3C2B*, *PIK3CA*, *PIK3CB*, *PIK3CD* and *PIK3R3*) had negative gene-scores in this co-module. This result agrees with the finding of a study³⁹ suggesting that lowered *PI3K/AKT1* pathway activity can reverse resistance of leukemia cells to ATRA. In addition, *PIK3CA* is one of the most highly mutated oncogenes recently identified in human cancers⁴⁰. Finally, in AML cells retinoic acid receptor- α (*RARA*) is translocated, and fused to *PLM41*, a gene which occurred in this co-module with positive score.

We also took a drug-class-centered approach and looked into co-modules containing topoisomerase II inhibitors. Topoisomerase II inhibitors are chemotherapy agents designed to interfere with the action of topoisomerase II enzymes, which act on the topology of DNA. Among the most extensively studied topoisomerase II inhibitors are daunorubicin and etoposide. We found 30 co-modules that include at least one topoisomerase II inhibitor drug and at least one DNA topoisomerase II gene (*TOP2A*, or *TOP2B*). Interestingly, in 24 of these co-modules all *TOP2* genes were downregulated and all included cell lines were resistant to topoisomerase II inhibitor drugs. The exact opposite was observed in six remaining other co-modules; that is, they contained upregulated *TOP* genes and topoisomerase inhibitor-sensitive cell lines. Thus, topoisomerase gene expression levels seem to be a good proxy for topoisomerase inhibitor sensitivity. This result is supported by previous findings demonstrating that low levels of *TOP2A* expression is associated with resistance⁴², and we hypothesize that high *TOP2A/TOP2B* expression may be a proxy for sensitivity. Two of the topoisomerase II inhibitor co-modules included membrane transport genes ($P < 0.005$). This seems to be a plausible link because several cell lines have been shown to acquire resistance to etoposide (CAS number 33419-42-0) through membrane transport changes⁴³. In addition, this co-module contained six downregulated ABC transporters, including *ABCC1* and *ABCG2*, and five topoisomerase II inhibitor drugs—all the cell lines in this co-module were moderately sensitive to these inhibitors (drug score < -0.4). This observation is consistent with results by Depeille *et al.*⁴⁴, who showed that lowered levels of *ABCC1* increased the sensitivity to etoposide. Similarly, high *ABCG2* levels result in resistance to mitoxantrone via the effect of *ABCG2* on mitoxantrone transport⁴⁵. Another six topoisomerase II inhibitor co-modules contained only moderately sensitive lung cancer cell lines to all topoisomerase II inhibitors (drug-score < -0.25) and were enriched for glutathione biosynthesis ($P < 10^{-6}$). This hypothetical link is supported by a report suggesting that depletion of glutathione sensitized a small-cell lung cancer cell line to daunorubicin (CAS number 20830-81-3) and etoposide⁴⁶.

DISCUSSION

Our extensive comparative analysis using both *in silico* and NCI-60 data revealed that the Ping-Pong Algorithm (PPA) is the most effective procedure, both with respect to the two other modular algorithms and classical approaches like clustering, regression or SVD. A further advantage is that in contrast to these methods the PPA provides context-dependent links between the two data sets. For the NCI-60 data this implies that drug-gene associations are sample dependent; for example, they might only be relevant for a particular type of tissue. Moreover, the PPA is computationally efficient (see **Supplementary Notes** for more details), making integrated analyses of large-scale data feasible on desktop computers. (More than 850 co-modules were generated from the NCI-60 data in < 100 min on a server-class machine).

Using the NCI-60 drug-response and gene-expression data of untreated cell lines as a paradigm, we have demonstrated on a global scale that our modular approach allows prediction of drug-gene associations from relatively inexpensive high-throughput data and

therefore enables large-scale hypotheses generation that can reduce the cost of studying drug-gene interactions directly. Consistency with the recorded research results in DrugBank is even more striking for functional sets of genes participating in the drug-response indirectly. This is a particularly promising result, because many annotated processes (including those related to drug response) may, of course, not be co-regulated at the transcriptional level within the NCI-60 cell lines.

The modular approach for integrative analysis of large-scale data that we developed in this paper might also be useful for the meta-analyses of other large-scale biological data. For example, co-module analysis has potential for the integration of expression data from two different species. In this case the common dimension of the two sets of expression data are established by the orthologous genes (rather than the samples for the NCI-60 data). Here, co-module analysis could provide a sensitive means to identify not only the sets of genes with conserved co-expression, but also the respective experimental conditions under which this co-expression is induced. Similarly, for data sets covering different types of gene regulation (e.g., post-transcriptional modifications or protein expression), our approach might reveal those sets of genes that are co-regulated at multiple instances. In fact, our method could be particularly useful for these applications because it does not require regulation data for identical sets of conditions but automatically identifies those conditions from both data sets that give the best match for co-regulation.

METHODS

Algorithms. A detailed description of each algorithm is given in the **Supplementary Notes**. The main distinction between the three different modular approaches lies in whether the modularization of the data sets is applied after (ISA(E \cdot R^T)), before (ISA(E)&ISA(R)) or simultaneously with the data integration (PPA). In fact, the PPA can be viewed as a straightforward generalization of the ISA to multiple data sets. Starting with a candidate set of genes (in practice we use hundreds of 'random seeds'), the available expression data are used to identify the cell lines for which these genes exhibit a coherent expression (arrow 1 in **Fig. 1c**). In the next step the response data are used to select drugs that elicit a similar response in these cell lines (arrow 2). This set of drugs is then used to refine the set of cell lines by eliminating those which have an incoherent response profile and adding others that behave similarly across these drugs (arrow 3). Finally, this refined set of cell lines is used to probe for genes that are co-expressed in these lines (arrow 4). This alternating procedure is reiterated until it converges to stable sets of genes, cell lines and drugs. We refer to these sets as co-modules, which generalize the concept of a module from a single to multiple data sets. A mathematical description of the PPA is given in **Box 1**, where the numbered algorithm steps (1–4) correspond to the number of the arrows in **Figure 1c**.

Testing predictions for *in silico* data. Because the features of real data are hardly established (in particular for drug-gene associations), we resorted to an *in silico* approach (similarly to that of refs. 18,20) for the initial benchmarking of our algorithms. We used a model that generates gene-expression and drug-response data containing both independent transcription- and response-modules, as well as some coupled modules. This approach allows for testing our algorithms (as well as other methods) in a well-defined setting with a tunable degree of shared elements between modules. Imposing such modular interdependencies simulates the complexity of real data and is therefore crucial for realistic benchmarking. A second feature of the *in silico* data is the adjustable level of noise. Each algorithm was used to compute putative drug-gene associations based on the modules it produced from the *in silico* data. We used continuous prediction scores, defined as the sum over the products of the gene and drug scores from all (matched) modules: $s_{gd} = \sum_m |s_g^{(m)} s_d^{(m)}|$. A pair of a gene g and a drug d receives a high score s_{gd} either if they are associated through multiple modules or if their respective scores $s_g^{(m)}$ and $s_d^{(m)}$ are high (or both). The evaluation was then performed by computing the ROC curve given the corresponding 'truth' matrix T of the underlying model ($T_{gd} = 1$, if

Box 1 The Ping-Pong Algorithm (PPA)

For a given threshold combination (t_C : condition threshold, t_G : gene threshold, t_D : drug threshold) the Ping-pong algorithm (PPA) is summarized in the following pseudocode:

- $n = 0$; $\mathbf{g}^{(0)} = \text{random}(N_G) \in [0,1]^{N_G}$ (initial random seed)
- **while** ($|\hat{\mathbf{g}}^{(n)} - \hat{\mathbf{g}}^{(n-1)}| + |\hat{\mathbf{d}}^{(n)} - \hat{\mathbf{d}}^{(n-1)}| + |\hat{\mathbf{c}}^{(n)} - \hat{\mathbf{c}}^{(n-1)}| + |\hat{\mathbf{c}}^{(n)}| |\hat{\mathbf{c}}^{(n)}| > \epsilon$)
 1. $\mathbf{c} = \mathbf{E}_G^T \cdot \hat{\mathbf{g}}^{(n)}$; $\mathbf{c}_j^{(n+1)} = \begin{cases} c_j & \text{if } |c_j - \mu(\mathbf{c})| > t_C \sigma(\mathbf{c}) \\ 0 & \text{otherwise} \end{cases}$ ($j = 1, \dots, N_C$)
 2. $\mathbf{d} = \mathbf{R}_C \cdot \hat{\mathbf{c}}^{(n)}$; $\mathbf{d}_k^{(n+1)} = \begin{cases} d_k & \text{if } |d_k - \mu(\mathbf{d})| > t_D \sigma(\mathbf{d}) \\ 0 & \text{otherwise} \end{cases}$ ($k = 1, \dots, N_D$)
 3. $\tilde{\mathbf{c}} = \mathbf{R}_D^T \cdot \hat{\mathbf{d}}^{(n)}$; $\tilde{\mathbf{c}}_l^{(n+1)} = \begin{cases} \tilde{c}_l & \text{if } |\tilde{c}_l - \mu(\tilde{\mathbf{c}})| > \tilde{t}_C \sigma(\tilde{\mathbf{c}}) \\ 0 & \text{otherwise} \end{cases}$ ($l = 1, \dots, N_C$)
 4. $\mathbf{g} = \mathbf{E}_C \cdot \tilde{\mathbf{c}}^{(n)}$; $\mathbf{g}_m^{(n+1)} = \begin{cases} g_m & \text{if } |g_m - \mu(\mathbf{g})| > t_G \sigma(\mathbf{g}) \\ 0 & \text{otherwise} \end{cases}$ ($m = 1, \dots, N_G$)
 5. $n = n+1$
- $\mathbf{g}^* = \mathbf{g}^{(n)}$; $\tilde{\mathbf{c}}^* = \tilde{\mathbf{c}}^{(n)}$; $\mathbf{d}^* = \hat{\mathbf{d}}^{(n)}$

Starting from a random seed, the algorithm identifies weighted lists of genes, conditions and drugs as fixed points of iterating the four linear mappings followed by thresholding (which correspond to the four numbered arrows in **Fig. 1c**). The weights (or scores) are the components x_i of the gene, condition and drug vector, $\mathbf{x} \in \{\mathbf{g}, \mathbf{c}, \mathbf{d}\}$, of length N_G , N_C , N_D , respectively. $|\mathbf{x}|$, $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ denote their norm, mean value and the standard deviation. $\hat{\mathbf{x}} = \mathbf{x}/|\mathbf{x}|$ is a normalized vector. \mathbf{E}_G and \mathbf{E}_C refer to the expression matrix normalized across genes and conditions, respectively. Similarly, \mathbf{R}_D and \mathbf{R}_C denote the response matrix normalized across drugs and conditions, respectively. (See **Supplementary Notes** for more details.)

gene g and drug d are associated, and 0 otherwise). As testing score we used the area under the ROC curve. Scores were computed for all methods using multiple data sets generated for different levels of complexity or noise (**Fig. 1d** and **Supplementary Notes** for more details).

Testing drug-gene association predictions. To test our algorithms in a realistic setting we used the NCI-60 data and direct information on drug-gene associations from DrugBank³¹ and the Connectivity Map³² corresponding to binary truth matrices T_{gd}^{DB} and T_{gd}^{CM} , respectively. DrugBank contains 351 drug-gene associations, involving 251 genes (1.6% of the total 22,283) and 85 drugs (1.6% of the total 5,183) included in the NCI-60 study. Because the DrugBank database does not specify in which tissue a certain drug is reported to interact with a gene (or group of genes), we used again the sum the absolute scores $|s_g^{(m)} s_d^{(m)}|$ over all co-modules (m) to infer a drug-gene association if there is at least one (tissue-specific) co-module that includes both the gene and the drug. The DrugBank associations were generated mostly from literature surveys, few of which might have been influenced by the NCI-60 data. Thus we also considered complementary data from the Connectivity Map, which are derived directly from the transcriptional response of the same 22,283 genes as the NCI-60 study for four cell lines (MCF7, PC3, HL60, SKMEL5) treated with 164 drugs or small molecules, out of which 77 (1.5% of the total 5,183) were also present in the NCI-60 response data. We assumed that genes exhibiting more than a twofold expression change in response to a certain drug to be a plausible target of that drug.

As before, we compared T_{gd}^{DB} and T_{gd}^{CM} with our predictions over the common genes and drugs using ROC curves. Because annotated drug-gene associations are sparse, only predictions with high scores are likely to be relevant. Therefore,

in our analysis, we focused on the region of the ROC curve with FPR < 10% (**Fig. 3a–b**).

Testing drug-gene set association predictions. DrugBank reports association of drugs with gene sets involved in metabolic pathways (according to KEGG³⁴) or annotated in functional categories (according to GO²⁴). We designed another prediction score $s_{Gd} = \sum_m |s_G^{(m)} s_d^{(m)}|$ reflecting the strength of the association of a gene-set G with a single drug d . This score depends on the drug-scores $s_d^{(m)}$ as well as scores $s_G^{(m)}$ describing the significance of overlap between G and the genes G_m of each module m . (We used $s_G^{(m)} = -\log P(G, G_m)$, where $P(G, G_m)$ is computed from the hypergeometric distribution⁴⁷ and Bonferroni-corrected for testing multiple GO categories.) Using ROC curves, we then compared the four different truth matrices T_{Gd}^{DB} (for KEGG, GO's biological processes, molecular functions and cellular components) with our predictions s_{Gd} (**Figure 3c–f**).

Comparing and evaluating the gene content of co-modules. We computed for each biological process in GO²⁴ the significance scores $s_G^{(m)}$ for overrepresentation of the associated gene-set G with that of all modules produced by a particular algorithm. For each biological process we picked the highest score (across all modules from each algorithm) to indicate how well this biological process was represented by the modular decomposition of each algorithm. We then plotted the scores for all biological processes obtained for a reference algorithm (PPA) against those of another algorithm (**Fig. 4a–d**). These scatter plots provide a global picture of which algorithms tend to cover the annotated processes with better accuracy. We highlighted certain processes related to either drug metabolism (according to DrugBank) or cancer (extracted from ref. 48) to give a more refined picture of how well the modules of the different algorithms capture the corresponding genes.

As a complementary approach, we also characterized which biological processes were covered best by the modular decompositions produced by the different algorithms. To this end we determined for each module the process with the highest significance score $s_G^{(m)}$ (> 1.3 , which is equivalent to $P < 0.05$) and then counted the number of modules attributed to each process (or to no process at all). From this we computed the fractional coverage for the 15 processes with the highest number of attributed modules, as well as the fraction of modules that were associated with no process (**Fig. 4e–f**).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We are grateful for financial support from the Giorgi-Cavaglieri Foundation (S.B.), the Swiss National Science Foundation (grants no. 3100AO-116323/1, S.B.; and no. 310000-112552/1, J.S.B.) and the European Framework Project 6 (through the EuroDia and the AnEuploidy projects, S.B.). The authors would like to thank N. Barkai, R. Chrast, L.A. Decosterd, T. Johnson, D. Marek, A. Morton de LaChapelle, B. Peter, C. Rivolta, A. Sewer and O. Spertini for their valuable feedback and comments. We also greatly appreciate the constructive comments of the referees to improve the manuscript.

Published online at <http://www.nature.com/naturebiotechnology/>
Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

1. Benson, M. & Breitling, R. Network theory to understand microarray studies of complex diseases. *Curr. Mol. Med.* **6**, 695–701 (2006).
2. Pereira-Leal, J.B., Enright, A.J. & Ouzounis, C.A. Detection of functional modules from protein interaction networks. *Proteins* **54**, 49–57 (2004).
3. Wang, W., Cherry, J.M., Botstein, D. & Li, H. A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **99**, 16893–16898 (2002).
4. Werner, T. Regulatory networks: linking microarray data to systems biology. *Mech. Ageing Dev.* **128**, 168–172 (2007).
5. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
6. Gasch, A.P. & Eisen, M.B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* **3**, RESEARCH0059 (2002).
7. Getz, G., Gal, H., Kela, I., Notterman, D.A. & Domany, E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics* **19**, 1079–1089 (2003).
8. Holter, N.S. *et al.* Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA* **97**, 8409–8414 (2000).
9. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network.

- Nat. Genet.* **31**, 370–377 (2002).
10. Segal, E., Yelensky, R. & Koller, D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19** Suppl 1, i273–i282 (2003).
 11. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. 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 (2004).
 12. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
 13. Staunton, J.E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* **98**, 10787–10792 (2001).
 14. Shankavaram, U.T. *et al.* Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integrative microarray study. *Mol. Cancer Ther.* **6**, 820–832 (2007).
 15. Gaur, A. *et al.* Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. *Cancer Res.* **67**, 2456–2468 (2007).
 16. Bussey, K.J. *et al.* Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol. Cancer Ther.* **5**, 853–867 (2006).
 17. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **24**, 236–244 (2000).
 18. Bergmann, S., Ihmels, J. & Barkai, N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E* **67**, 031902 (2003).
 19. Ihmels, J., Bergmann, S., Berman, J. & Barkai, N. Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet.* **1**, e39 (2005).
 20. Prelic, A. *et al.* A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22**, 1122–1129 (2006).
 21. Ihmels, J., Bergmann, S. & Barkai, N. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**, 1993–2003 (2004).
 22. Bergmann, S., Ihmels, J. & Barkai, N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* **2**, E9 (2004).
 23. Ihmels, J. *et al.* Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309**, 938–940 (2005).
 24. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
 25. Paull, K.D. *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* **81**, 1088–1092 (1989).
 26. Ma, Y. *et al.* Predicting cancer drug response by proteomic profiling. *Clin. Cancer Res.* **12**, 4583–4589 (2006).
 27. Gao, F., Foat, B.C. & Bussemaker, H.J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**, 31 (2004).
 28. Liu, X., Jessen, W.J., Sivaganesan, S., Aronow, B.J. & Medvedovic, M. Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. *BMC Bioinformatics* **8**, 283 (2007).
 29. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* **36**, 1090–1098 (2004).
 30. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
 31. Wishart, D.S. *et al.* DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
 32. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
 33. Zweig, M.H. & Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577 (1993).
 34. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–D357 (2006).
 35. Hayot, C. *et al.* Characterization of the activities of actin-affecting drugs on tumor cell migration. *Toxicol. Appl. Pharmacol.* **211**, 30–40 (2006).
 36. Melkonian, G. *et al.* Capillary plexus development in the day five to day six chick chorioallantoic membrane is inhibited by cytochalasin D and suramin. *J. Exp. Zool.* **292**, 241–254 (2002).
 37. Marot, D. *et al.* High tumoral levels of Kiss1 and G-protein-coupled receptor 54 expression are correlated with poor prognosis of estrogen receptor-positive breast tumors. *Endocr. Relat. Cancer* **14**, 691–702 (2007).
 38. Hu, J., Zhou, G.B., Wang, Z.Y., Chen, S.J. & Chen, Z. Mutant transcription factors and tyrosine kinases as therapeutic targets for leukemias: from acute promyelocytic leukemia to chronic myeloid leukemia and beyond. *Adv. Cancer Res.* **98**, 191–220 (2007).
 39. Neri, L.M. *et al.* The phosphoinositide 3-kinase/AKT1 pathway involvement in drug and all-trans-retinoic acid resistance of leukemia cells. *Mol. Cancer Res.* **1**, 234–246 (2003).
 40. Samuels, Y. & Velculescu, V.E. Oncogenic mutations of PIK3CA in human cancers. *Cell Cycle* **3**, 1221–1224 (2004).
 41. Chomienne, C. *et al.* All-trans retinoic acid modulates the retinoic acid receptor- α in promyelocytic cells. *J. Clin. Invest.* **88**, 2150–2154 (1991).
 42. Harris, A.L. & Hochhauser, D. Mechanisms of multidrug resistance in cancer treatment. *Acta Oncol.* **31**, 205–213 (1992).
 43. Hande, K.R. Etoposide pharmacology. *Semin. Oncol.* **19**, 3–9 (1992).
 44. Depeille, P., Cuq, P., Passagne, I., Evrard, A. & Vian, L. Combined effects of GSTP1 and MRP1 in melanoma drug resistance. *Br. J. Cancer* **93**, 216–223 (2005).
 45. Breedveld, P. *et al.* The effect of low pH on breast cancer resistance protein (ABCG2)-mediated transport of methotrexate, 7-hydroxymethotrexate, methotrexate diglutamate, folic acid, mitoxantrone, topotecan, and resveratrol in *in vitro* drug transport models. *Mol. Pharmacol.* **71**, 240–249 (2007).
 46. Doyle, L.A. *et al.* An etoposide-resistant lung cancer subline overexpresses the multidrug resistance-associated protein. *Br. J. Cancer* **72**, 535–542 (1995).
 47. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
 48. Gamberoni, G., Storari, S. & Volinia, S. Finding biological process modifications in cancer tissues by mining gene expression correlations. *BMC Bioinformatics* **7**, 6 (2006).