TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

Sackler Faculty of Exact Sciences, Blavatnik School of Computer Science

Novel methods for integrative analysis of heterogeneous large scale biomedical data

THESIS SUBMITTED FOR THE DEGREE OF
"DOCTOR OF PHILOSOPHY"

by

David Amar

The work on this thesis has been carried out
under the supervision of
Prof. Ron Shamir

Submitted to the Senate of Tel-Aviv University
December 2015

Acknowledgments

This thesis summarizes a part of my work in the last three and a half years. This period was great and I was truly lucky to work in such a great environment, travel to fascinating places, and work with interesting and talented researchers.

First, I would like to thank to my advisor Prof. Ron Shamir. During my MSc and PhD Ron was a true mentor for me, professionally and beyond. He taught me what doing science is all about. He gave me freedom to pursuit any goal or any weird idea that I had. Thank you Ron for everything!

I would like to thank many people I got to work with. First, I would like to thank my lab members that were always there for interesting scientific (and non-scientific) discussions. I would like to thank Oren Tzfadia, Erik Alexandersson, and our AllBio team (Itziar, Tatiana, Estelle, Agnieshka, and Sanjeev) for a great collaboration that lead to interesting results in plant research, Adi Maron-Katz for many helpful discussions and helping me understand the fMRI world, Tom Hait for working with me on the Adeptus project, Prof. Daniel Yekutieli for an interesting collaboration and teaching me new things in statistics. I would also like to thank Aloysius Domingo and Prof. Christine Klein for working with me on XDP.

I owe special thanks to the Azrieli foundation for accepting me into the Azrieli program, and for their generous funding. I would like to thank Dr. Naomi Azrieli for her kindness and for spending time with us in person. I would also like to thank the Azrieli foundation team for always being there: Rochelle Avitan, Adi Dagan, Tirza Peleg, and Elinor Tubul.

I would also like to thank the Edmond J. Safra Center for Bioinformatics for the support over the last six years. I would like to thank Gilit Zohar-Oren for always, regardless of the topic at hand, being helpful with a cheerful smile.

Last but definitely not least, I would like to thank my family for their support over these years. Above all, to my second half Anat.

Preface

This thesis is based on the following three articles that were published throughout the PhD period in scientific journals.

1. D. Amar and R. Shamir. **Constructing module maps for integrated analysis of heterogeneous biological networks**. Nucleic Acids Research, doi:10.1093/nar/gku102, 2014 (1).

2. D. Amar, D. Yekutieli, A. Maron-Katz, T. Hendler and R. Shamir. **A hierarchical Bayesian model for flexible module discovery in three-way time-series data**. Bioinformatics, 31 (12): i17-i26, ISMB/ECCB 2015, proceedings paper, doi: 10.1093/bioinformatics/btv228, 2015 (2).

3. D. Amar, T. Hait, S. Izraeli and R. Shamir. **Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets**. Nucleic Acids Research, doi: 10.1093/nar/gkv810, 2015 (3).

Abstract

We live in a unique era in which biomedical research has transformed into an information science. The amount and complexity of the collected data in public databases are huge, and computational methods that can bridge the gap between having information and understanding it are direly needed. In this thesis we describe our studies in which we provided novel computational methods that can help in moving towards this important goal. Our methodologies integrate information from a broad variety of data sources and of diverse types. By utilizing techniques from graph theory, probabilistic modeling, and statistical learning we were able to handle complex large scale data. Our studies present contributions to three main areas of computational biology: network biology, time series data analysis, and big data integration. In each of these fields we showed that our methods for integrative analysis outperform existing methods, provide novel biological insights, and can facilitate suggestion of novel hypotheses that could be used for future research.

Contents

1. Introduction

1.1 The Big Biomedical Data era

Over the past decades technology advances in imaging, molecular cell biology, and genomics have transformed biomedical research into an informatics science. Today, in almost every area of molecular biology and medicine, more and more large scale data are being generated and collected. **Figure I1** below gives an overview of different molecular data types, coupled with their biological context. For each layer of the complex biological machinery different technologies were developed in order to characterize its molecular features. Prominent examples include sequencing DNA or RNA (e.g., across tissues and subjects), measuring metabolite abundance, and even monitoring the activity level of brain regions (not shown).

Many technologies today produce output en masse. That is, they can measure thousands – and sometimes millions - of values in a single experiment. Public databases, such as the ones provided by the National Center for Biotechnology and Information (NCBI), provide freely available datasets from thousands of studies for each type of such high throughput technology. While these data are available for immediate download on demand, they often contain high noise levels, their analysis results are usually hard to interpret, and they may even suffer from low reproducibility. Dealing with such data requires accurate and scalable algorithms for extracting interesting patterns.

One of the main goals of computational biologists is to provide automatic tools that would help researchers integrate data from many heterogeneous studies in order to extract reliable information. In the context of **Figure I1**, the input for integrative analysis is data from many studies that could cover a single layer (for intra-layer analysis) or more (for inter-layer analysis). In both cases, the goal is to improve reproducibility, interpretability, and suggest novel hypotheses.

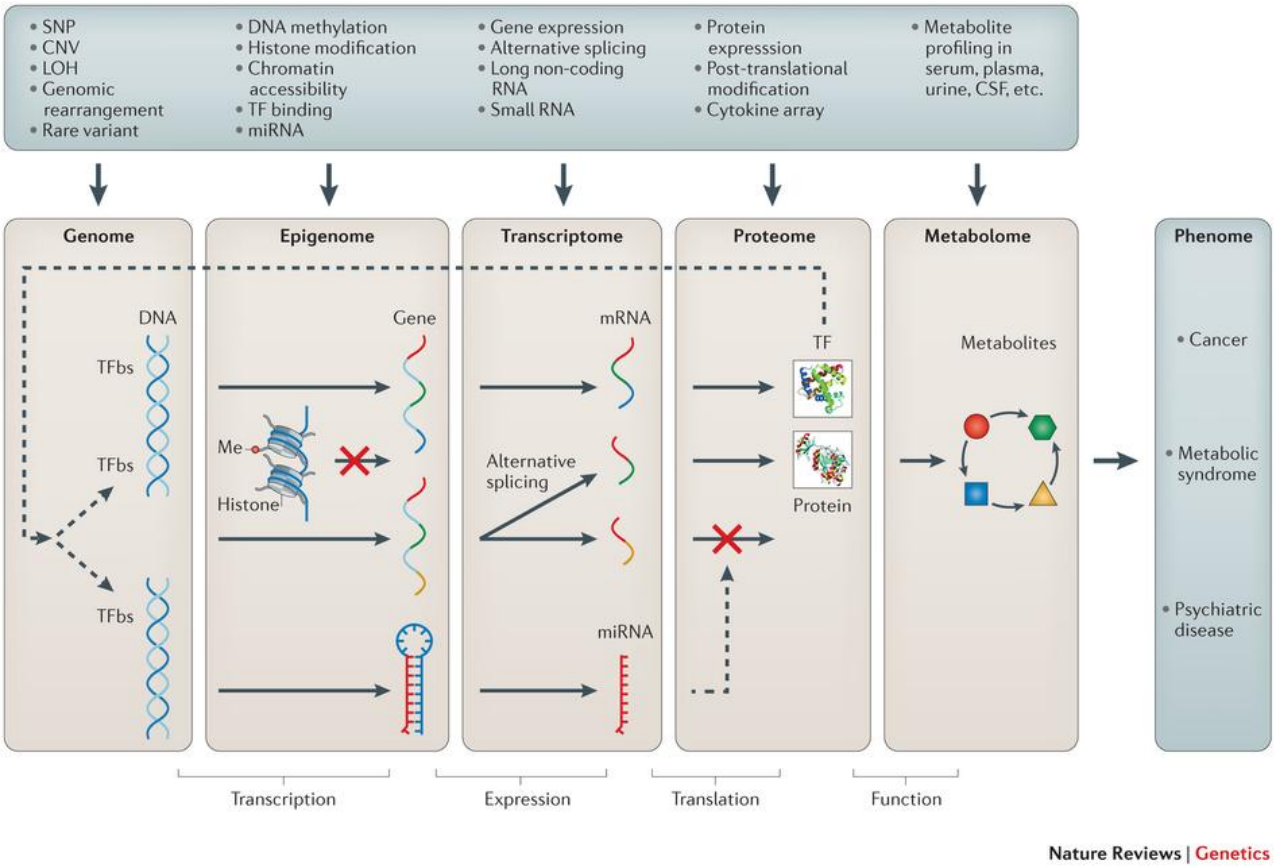


Figure I1. Multi-omics data from the genome to the phenome. The main flow shows a simplified view of the components of cellular systems, organized into different levels. Heterogeneous genomic data exist between and within levels. The top bar lists the specific data types that could be measured within each layer. Red crosses indicate inactivation of transcription or translation. CSF, cerebrospinal fluid; Me, methylation; TFBS, transcription factor-binding site. Source: (4).

In the sections below we lay out the background and terminology required for this thesis. We first introduce the main data types that were studied in order to give a general background (Sections 1.2-1.3). We next discuss advanced experiments that aim to characterize the dynamic features of biological processes (Section 1.4). Finally, we discuss computational aspects of integrative analysis (Section 1.5). As this subject is too broad to fully review it here, we give an overview of pertinent recent advances. We focus on areas covered in this thesis, and highlight the added value of the integrative analysis.

1.2 High throughput profiling

In biomedical research, it is often required to obtain global molecular "snapshots" of the cell at different conditions. These snapshots can later be used, for example, in comparative analysis to reveal the molecular changes between different situations and cell types. See **Figure I2** for an example in cancer research. Ideally, the best snapshot would provide quantification of all biological molecules in the cell, including DNA, RNA, proteins, and metabolites. However, techniques for high quality recording of protein and metabolite quantities on a large scale are still under development. On the other hand, it is possible today to measure the content and concentration of all nucleic acid-based molecules (i.e., DNA and RNA sequences). For example, RNA transcript levels in the cell could be measured using DNA microarrays or sequencing technologies. These measurements are then used as approximation for the current activity of all genes, and indirectly of their protein products.

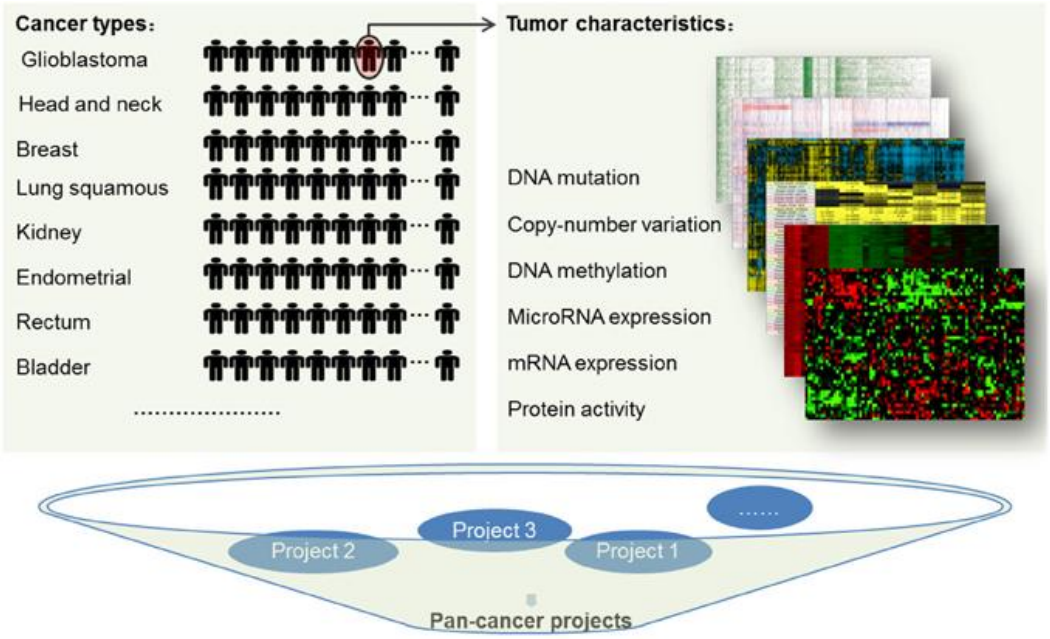


Figure I2. Different molecular profiles from diverse tumor types were collected by the TCGA pan-cancer projects. Each project can include up to six types of profiles (mutation, copy number,

methylation, gene expression, microRNA and reverse phase protein arrays) from tumors occurring in different sites of the body. Source: (5).

Using similar technologies, it is possible to measure the methylation level of all genes in the DNA. The level of DNA methylation is important for many biological processes including gene regulation, cell differentiation, and development (6,7). Profiling of single nucleotide variants (SNVs) can be used for comparing cohorts on the genetic level. Such analyses have been successful in detecting novel mutations in complex disease and cancer, thereby relating genes to diseases (8). In addition, several technologies are available for monitoring the levels of all microRNAs in a given sample. Such transcriptomics and epigenetic experiments were used to detect biomarkers for cancer, but a comparison of different experiments showed low reproducibility (9).

Fortunately, many of these data types are available for exploration and analysis. Databases such as the Gene Expression Omnibus (GEO) (10), ArrayExpress (11), or The Cancer Genome Atlas (TCGA) (12) provide hundreds of thousands of samples that can be utilized to discover changes between patient cohorts. In these databases, a sample is summarized as a vector of scores. In this thesis we shall describe a study in which we utilized data from hundreds of gene expression studies to extract reliable disease biomarkers and relate them to other data sources (e.g., mutation data) in order to reveal therapeutic potential. Such studies can pave the way towards a robust analysis that will improve reproducibility and interpretation.

1.3 Biological Networks

Biological networks provide a comprehensive overview of biological systems. They enable better understanding of biological processes and can shed light on the function of genes and other molecular compounds. Biological networks have been utilized for a wide variety of applications including discovery and prediction of gene interactions, gene functions, and disease-genes associations (13–21). These networks can directly teach us about interactions and dependencies between cell particles. Thus, when analyzed with high throughput profiles, networks can add a complementary view of the cell state.

In biological networks nodes represent molecular entities, and the edges represent interdependencies. For example, in *protein-protein interaction* (PPI) networks nodes represent proteins and edges represent physical interactions. In *genetic interaction* (GI) networks, nodes represent genes and edges represent the organism fitness for double knockout perturbations, yielding two major types of edges: alleviating GIs, and aggravating GIs (22). In alleviating GIs, also called *positive GIs*, the organism fitness after the double-knockout perturbation is better than expected based on the single knockout results. In aggravating or *negative GIs*, the fitness is worse than expected. Positive GIs were shown to be enriched within genes that are expected to work together in the same pathway, whereas negative GIs were shown to be enriched between pathway pairs that can compensate for the loss of each other (23). **Figure I3** shows an example of a genetic interaction network with *suppression interactions*. These are extreme positive GIs that rescue the embryonic development by a deletion of a second gene. Such interactions generally occur between genes with opposing functions in a shared process (24).

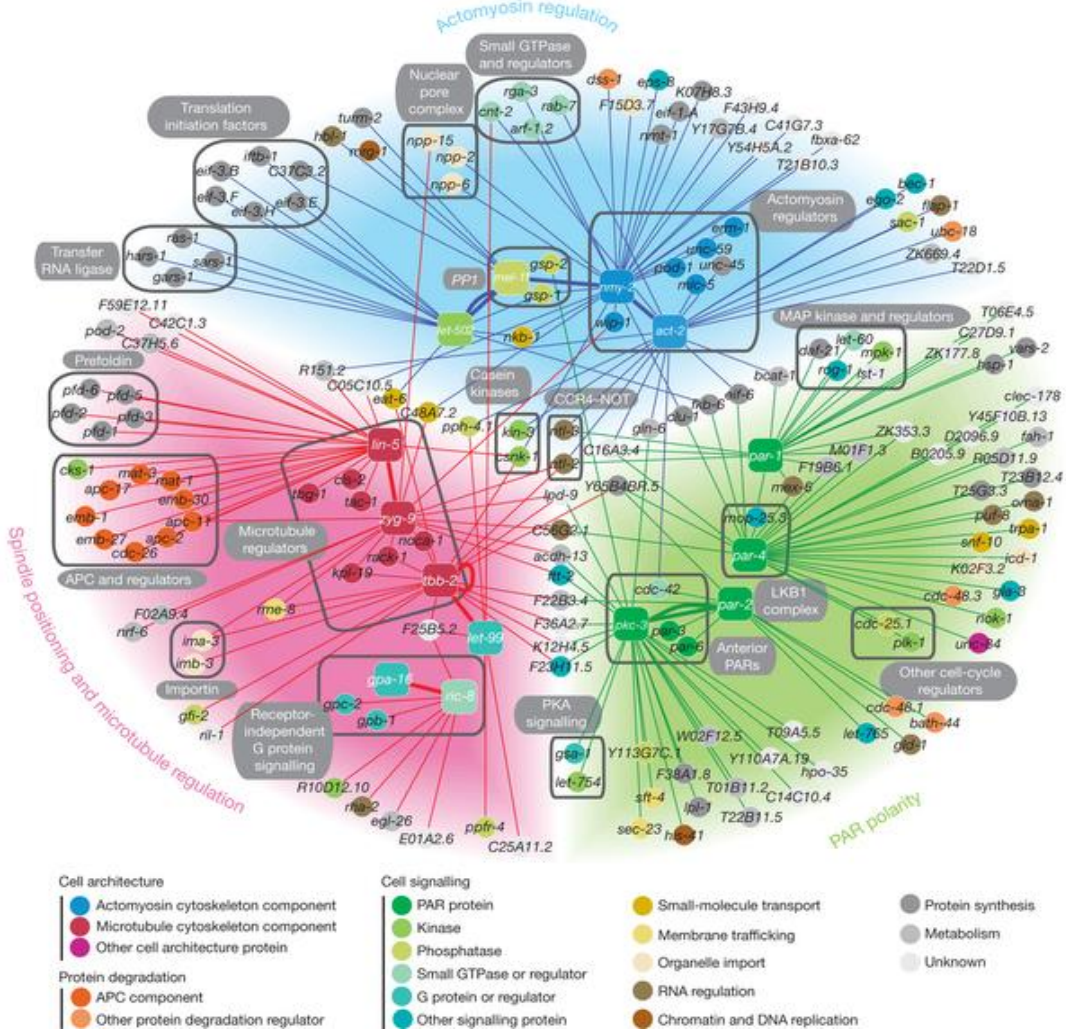


Figure I3. Genetic interaction network found in embryonic Caenorhabditis elegans. Colored areas represent three broad gene functional groups (Actomyosin regulation, blue; PAR polarity, green; Spindle positioning and microtubule regulation, red). Oblong nodes represent temperature-sensitive lethal genes (i.e., genes whose knockdown results in lethality). Circular nodes are their suppressors: genes whose additional knockdown results in embryonic rescue. Lines represent the genetic interactions found in the experiment (colors represent the basis of the mutant functional group). Thick lines represent interactions found between the 14 lethality genes. Suppressors linked to only one seed are positioned in the outskirts of the network. Source: (24).

In gene *co-expression* networks, nodes represent genes and edges score the correlation in expression between the two genes across a set of profiles (25). In gene *differential correlation* (DC) networks, edges score the change in gene pairwise correlation between one set of samples to another (e.g., cases and controls) (26–29). In *metabolic dependency* (MD) networks, nodes represent proteins and an edge is added between a pair of proteins if their associated reactions share a non-common metabolite (30). With the growing use and number of types of biological networks, computational methods that exploit these rich data are of great importance.

The networks explained above are undirected. In contrast, *regulatory pathways* are usually much smaller. In *regulatory pathways* directed edges represent source-target regulation, and in *signaling pathways* they represent signal transduction. More generally, a *biological pathway* is the set of molecular entities involved in a given biological process and the interrelations among those entities. Pathways are usually assembled by expert-based curation of the literature. Therefore, they usually give a simplified view based on the researchers' current knowledge and interpretation. Pathway boundaries are inherently fuzzy and are not always well defined, but they are valuable for understanding biology and for organizing biological knowledge (e.g. as a metabolic or signaling pathway). Pathway databases such as KEGG (31), Biocarta (www.biocarta.com), WikiPathways (32) and Reactome (33) provide manually curated, high quality pathways. A pathway in these databases is essentially a map that tracks the information flow of a biological process. These maps can contain a large variety of molecules, e.g., genes, proteins, metabolites, and their interactions. Using pathways for *enrichment analysis* had emerged as one of the main tools for interpreting results of large scale experiments. Given a set of genes or a ranking of genes (identified in the analysis of an experiment), many statistical methods were developed to assign significance for enrichment of a pathway in the gene set or at the top of the ranking (34–38).

1.4 Advanced dynamic experiments

Using the technologies mentioned above to learn from snapshots of cells has been instrumental in the past decades. As technologies advance and costs decrease, researches are able to design more complex experiments that would allow deciphering the dynamic features of biological systems. This is essential for pinpointing directed and causative relations among genomic components and phenotypes. Integrative analysis is especially crucial in these studies as it can leverage these complex data.

The two main kinds of such experiments are perturbations experiments and time series profiling. In a perturbation experiment, a controlled modification is introduced to the system (e.g., change of nutrients, over-expression of a certain gene, etc.) and measurements are taken before and after applying the modification. In a time series experiment the system is being monitored over k time points, where typically $k \geq 3$ (and in some cases much larger than 3). Both types of experiments can be utilized to learn causal links between components of the biological system. Below we give a short introduction to the main dynamic data sources that were studied in this thesis.

1.4.1 Differential networks

Recent studies in network biology have gone beyond static analysis to detect changes in the networks after the cells had been introduced to either a genetic perturbation or a unique environmental setting (39). One example is the GI network generated in response to DNA damage (40,41). In these studies, GIs were measured both in untreated cells and in cells following perturbation by the DNA-damaging agent methyl methanesulfonate (MMS) (42). These experiments revealed a differential network in response to MMS. This network contains DNA damage-specific interactions within pathways or between different pathways. Such networks could be used to reveal novel pathways that are activated in response to a particular perturbation.

1.4.2 Time series gene expression data

Measuring gene expression over time has been successfully utilized for characterization of developmental processes and response to drugs or other modifications (43–45). While classic experiments measured the expression changes over time in a controlled experiment, or a cell line, recent studies have started producing time series data from multiple different subjects. For example, Parnell et al. (44) measured blood gene expression in 35 patients after septic shock. The expression profiles were taken daily (for up to five days) after the septic shock. The analysis of the data detected major pathways and genes whose expression was significantly differential over time. Such analysis can provide novel candidate regulators that are important during sepsis development and progression.

1.4.3 Brain fMRI

Functional magnetic resonance imaging (fMRI) is a noninvasive neuroimaging method that is typically used to provide a blood-oxygen-level-dependent (BOLD) signal. The signal strength depends on two factors: (1) the systemic coupled increased blood flow (which enhances the signal), and (2) neuronal activity. The former is the main factor that controls the signal and in most analyses it is filtered out during preprocessing. Neuronal activity can enhance the BOLD signal through chemical signaling that causes dilation of blood vessels that increase the blood flow (46). This enhanced flow locally increases the ratio between red blood cells containing oxidized hemoglobin and those that have deactivated form of hemoglobin. As deoxidized hemoglobin has stronger magnetic influence on its surrounding, the result is a detectable change in the BOLD signal. In addition, BOLD is influenced mainly by local neural activity and internal neural processing and not by regional output.

Raw fMRI data requires multiple preprocessing steps including removal of nuisance signals related to head motion (and possibly other physiological variables), slice

timing, spatial smoothing, and temporal bandpass filtering (47). In addition, transformation of coordinates from measurements of different subjects into a common space (e.g., the Montreal Neurological Institute space; MNI) is required to allow group level analysis. As most datasets are generated in grid space of ~3 x 3 x 3 mm, yielding >100,000 brain voxels, dimensionality reduction is often applied. A common approach is to define a whole-brain parcellation based on fMRI signals derived from many subjects. Such approach was shown to better represent brain connectivity than using anatomically-defined atlases such as the anatomic atlas labeling (AAL) or Harvard-Oxford (48,49). Craddock et. al. (48) generated a whole brain functional parcellation by applying correlation-based clustering for fMRI data recorded from 41 healthy subjects at rest. The resulting parcellations, comprised of 200, 500 and 1000 parcels, were validated on an independent dataset. Parcel signals were calculated by averaging BOLD values across all gray matter voxels in it.

The brain function is known to depend on the inter-connectivity between structurally and functionally linked regions (50,51). A set of regions that are known to be active under a specific function is referred to as a "functional brain network". During the past decade, mapping these networks has received a great emphasis. One of the main tools for achieving this goal is resting state fMRI (rs-fMRI) studies, see **Figure I4**. In these studies, subjects are placed in the scanner and are required to stay awake. No particular task is given. The neural activity at rest has been shown to consume large quantities of energy and resources (52). This activity was shown to represent known functional brain networks (53). High correlation between voxels (or parcels) in rs-fMRI data, which is also called functional connectivity (FC), may indicate co-activation. Patterns of FC can be learned by either examining all pairwise correlations, or by examining the correlations of all voxels/parcels with a predefined seed voxel/parcel.

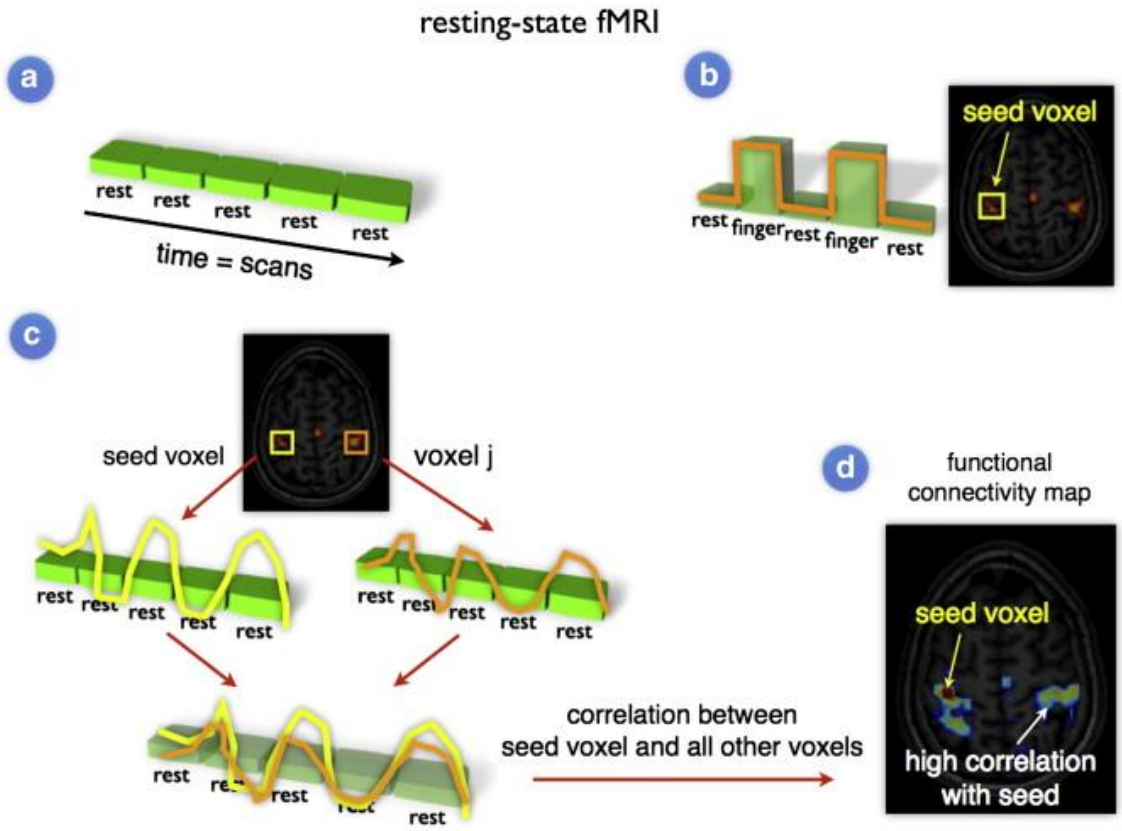


Figure I4. A typical flow of a resting-state fMRI study. (a) Subjects are placed in the MRI scanner and asked to close their eyes without falling asleep. No particular task is given. The BOLD fMRI signal at ~100,000 brain voxels is measured throughout the experiment. Multiple preprocessing steps are taken to reduce undesired effects. The end result is a signal for each voxel over time whose intensity is expected to correlate with neural activity. (b) Additional tasks performed during the study (e.g., moving a finger) between rest periods can be used to distinguish brain activity at rest from that during the task. Voxels (or parcels) with a significant correlation with the task (called "seed voxels") are identified and plotted on the brain map. (c) Functional connectivity is measured by calculating the correlation between the time series signal of a pair of voxels. (d) Highly correlated regions are inserted into the resulting functional connectivity map. An example of a standard visualization is shown. Brain regions that were covered by the selected highly correlated voxels are colored. Additional seed voxel information can be added to the map. Source: (54).

1.5 Integrative analysis

In this thesis we study a large variety of applications that utilize data from biological experiments. The integrative aspect of the studies can refer to two types. First, we tackle problems that emerge when analyzing data from multiple studies that produced the same biological data type. For example, we shall propose methodological ways to learn reliable biomarkers from hundreds of gene expression studies from different technologies, tissues, and diseases. We shall also analyze data of experiments that measured differential networks, where all experiments addressed the same perturbation, but each experiment covered a different gene set. Another example is detection of consistent modules that reappear in multiple subjects for whom time series data was measured (e.g., fMRI or gene expression).

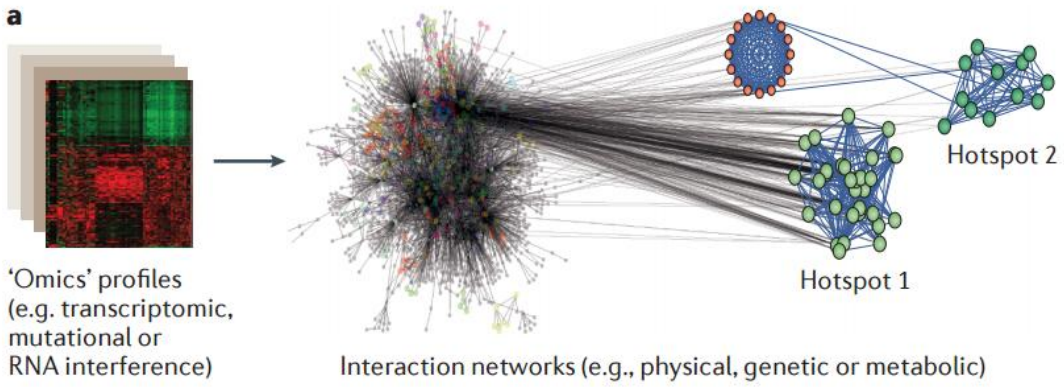
The second type of integrative analysis aims for extracting information from data of two or more different types. For example, we shall propose a way to integrate two different biological networks by learning a module map that summarizes them both. In another analysis we shall integrate gene expression data from hundreds of experiments with multiple gene information sources including cancer somatic mutation data. The result is a gene-based disease overview that can highlight the main disease genes, and can even propose novel candidates for drug repositioning in cancer.

In the sections below we give an introduction to recent integrative research performed on biomedical data similar to those covered in this thesis.

1.5.1 Supervised analysis of gene expression data

In supervised analysis, training samples are provided with labels indicating their classification into different phenotypes, and the goal is to predict the label of a new (unlabeled) sample. A sample is a vector that represents the measurements of an experiment. For example, a gene expression profile of a patient is a real-valued vector where each coordinate is the activity level of a gene. Many supervised analysis methods

that integrate diverse biological data were suggested in recent years. For example, integrative analysis of gene expression profiles and protein-protein interaction data or pathway information was demonstrated to improve patient classification accuracy (55–58). For example, Yang et al. (58) used pathway information to calculate features that summarize the expression level of pathways and used them to predict cancer. A network-based example is the search for *active modules*: connected network subgraphs whose genes exhibit differential signal in a supervised dataset, see **Figure I5** (59,60). However, recent comparisons between gene-based analyses and pathway/network-based methods in breast cancer data showed no significant difference between the approaches in classification performance (57,61,62). Similar ideas were recently used to detect highly connected sub-networks that are enriched with frequently mutated cancer genes (63–65).



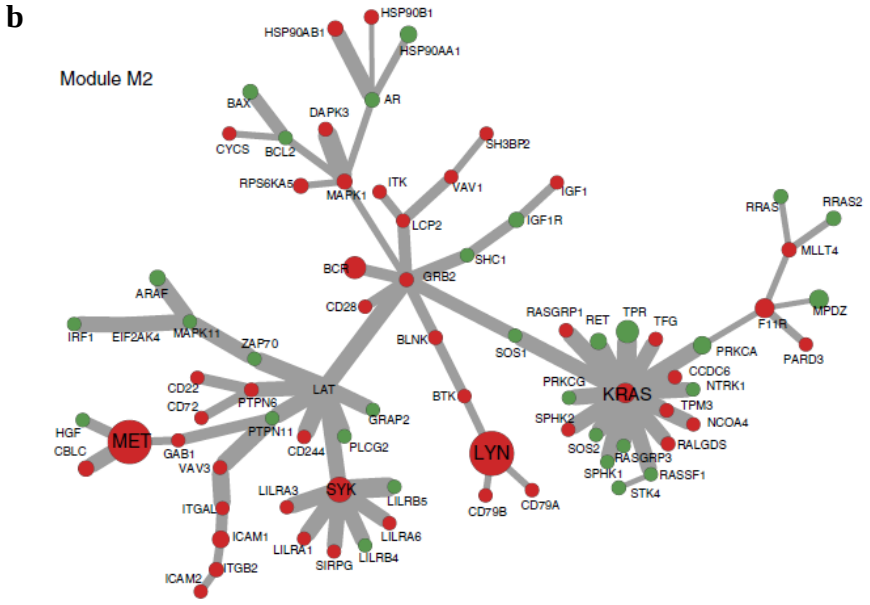


Figure I5. Active modules in biological networks. (a) Detection of active modules. Molecular profiles ("Omics" data) are used to characterize nodes in the network. For example, in gene expression data a score of a node could be a measurement of how differential the gene is when comparing cases and controls. Given the node features, an active module is a well connected subgraph whose genes reflect high scoring nodes. Source: (59). (b) An example of an active module detected in lung cancer. Node color represents up- or down- regulation (red, green, respectively). Node size is proportional to the level of differential expression of the gene. Edges are weighted protein interactions. Source: (66).

In addition, novel integrative methods were developed for feature selection (also known as biomarker or gene signature discovery). For example, the SoFoCles method (67) uses prior knowledge of functional similarity of genes to extract differential genes of similar functionality. Subramanian et al. (38) used pathway information together with gene expression profiles for differential pathway discovery. Mo et al. (68) combined multiple high throughput profiles of the same patients, including mutation, mRNA, and miRNA profiles, to select cancer biomarker genes for unsupervised data analysis (i.e., when the samples have no labels).

A very important integration effort is dealing with the large amounts of gene expression profiles available in public databases. This approach is a promising direction for increasing robustness, a well known problem in gene expression data analysis (69). Moreover, such studies can identify delicate signals that might not be detected when analyzing a single or a few datasets (due to either low quality, high statistical noise levels, or low statistical power). Huang et al. (70) used 9,169 gene expression profiles, each associated with a set of disease terms of the Unified Medical Language System (UMLS). UMLS, and similar databases such as the disease ontology, provide both disease terms and a directed acyclic graph (DAG) that models dependencies among diseases (71,72). The authors presented a *multi-label classifier*: an algorithm that predicts a set of disease terms for each gene expression sample. See **Figure I6** for more details. Their classification algorithm was designed to correct errors in which a sample is predicted to have a disease term but not one of its ancestors. Such predictions violate the path rule of the disease hierarchy, and could not be detected without using data of multiple diseases. The authors showed that such cases can often be corrected by using a method called Bayesian Correction (73). Schmid et al. (74) analyzed 3,030 samples of one platform and predicted their UMLS terms using similarity-based analysis. Lee et al. (75) analyzed >14,000 profiles of one microarray technology. Similarly to Huang et al., a correction for the tissue hierarchy was used, and was shown to significantly improve the results. However, their analysis was limited to prediction of the profile's tissue. Altschuler et al. (76) analyzed 176,971 microarray expression profiles of six different species to detect cross-species, tissue biomarkers that are based on pathways and not single genes.

A **B**

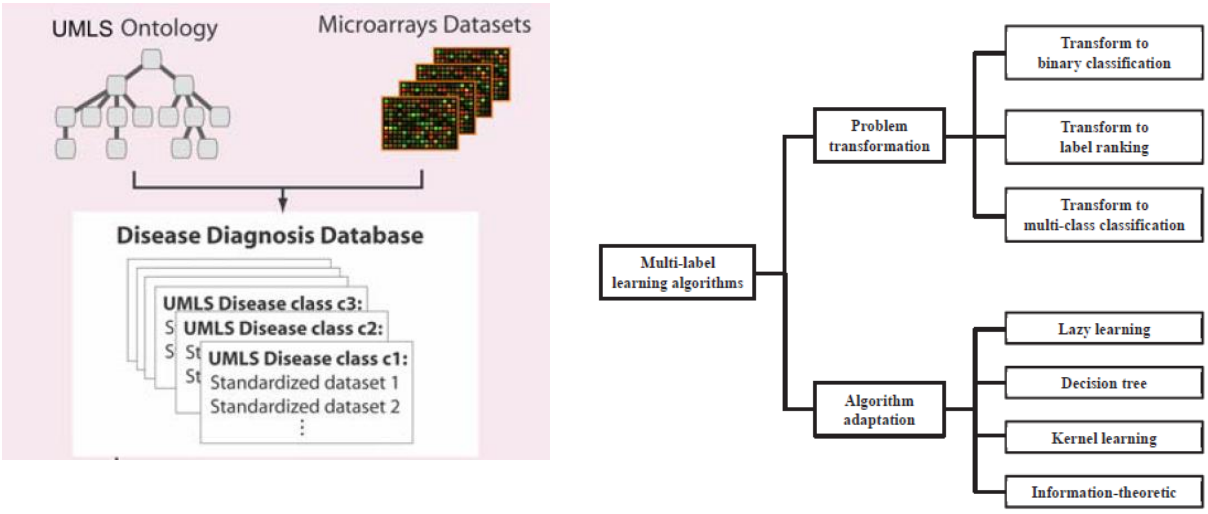


Figure I6. Overview of multi-label learning. (A) The input for the learning problem in our setting. For each patient a molecular profile was measured. In this example each patient has a microarray gene expression profile (in one of the datasets). In addition, each patient is assigned with a set of labels in the UMLS hierarchy that describe his phenotype. Source: (70). (B) Overview of multi-label learning algorithms. In our settings a multi-label classifier is a function that receives as input a molecular profile of a patient x, and for each disease term D, predicts the probability that the x belongs to D. Algorithms for learning multi-label classifiers can be broadly partitioned into two types: *problem transformation* and *algorithm adaptation*. Problem transformation methods transform the original problem into one or more standard classification problems. For example, the label power-set method transforms the problem into a multi-class classification problem. The method defines a new categorical variable (for each sample) whose values are all possible combinations of the original labels, which is then used as the class attribute. This method models the label dependencies implicitly and is usually effective when the number of labels is small (77). Algorithm adaptation methods extend a specific learning algorithm to deal with multi-label classification. For example, predictive clustering tree learns decision trees for the multi-label task (78,79); and Bayesian correction (not shown) uses the known label hierarchy to correct errors introduced when learning an independent single binary classifier for each label (80). Source: (81).

All these studies reported good prediction quality but they have limitations. First, except for (76) only one or two gene expression platforms were considered. Second, in Huang et al. (70) and Schmid et al. (74) the mapping of samples to their disease terms was done automatically using a tool called MetaMap (82) and inevitably mapping errors were present (70). Third, the Huang et al. (70) predictor is only applicable on new independent samples if a set of new control samples is given with them in order to allow calculations of differential expression. Fourth, while the prediction performance of the classifiers was far from random (e.g., 82% precision at 20% recall in Huang et al.) there is room for a significant improvement. Fifth, the low biological interpretability of these multi gene-based predictors poses a problem when a clinician wishes to monitor the decision rules of a specific disease. Finally, the set of studied phenotypes in (76) and (75) were limited as these studies focused on tissue classification.

1.5.2 Integration of interaction networks via module maps

Computational methods that make use of several networks often yield better results than methods that analyze only a single network (17,19,21,59,83–86). For example, combined analysis of PPI networks and gene co-expression networks was utilized to detect gene sets that are co-expressed and are connected in the PPI network. Such analysis outperformed standard clustering algorithms, and was successfully utilized for gene function prediction (15,21,59,86). Alleviating and aggravating GI data were used to find epistasis among and within gene sets. Under the premise that negative GIs tend to occur between compensatory pathways and positive GIs occur within pathways (or complexes), analysis of GIs was used to suggest a map of epistatic relations among functional gene modules (40,83,84,87–89). A marked improvement was reported after adding a connectivity constraint in a PPI network of the modules (84).

Several algorithms were proposed for detecting modules by simultaneous analysis of positive and negative GI data, and PPI networks. Kelley and Ideker (83) proposed a method that is based on local searches in the PPI and GI graphs to find pairs of connected

modules. Ulitsky et al. (84) used a clustering of positive GIs as a starting point and then improved the solution by merging modules that are connected in the PPI network. Other methods make use of the probabilistic scores of each GI edge, and incorporate both positive and negative GIs (87,89). Leiserson et al. (88,89) developed a method called Genecentric, which looks for locally maximum cuts in the GI graph. On the data of Collins et al. (90), this method was reported to outperform other methods, including algorithms that integrate GI and PPI information (40,91). While the methods described above marked a significant advance in the problem of simultaneous network analysis, they were all limited to GI network applications only.

1.5.3 Three-way time series data analysis

Identifying modules of elements acting in concert is a fundamental paradigm in interpreting, visualizing and dissecting complex biomedical data. For two-dimensional data (e.g., genes versus conditions), clustering and biclustering (92,93) have become standard in computational biology (59,94). Recent studies have suggested new methods for more complex input structures beyond the standard row–column data. For example, Meng et al. (95) extended the classic Iterative Signature Algorithm (ISA) for biclustering (96) to analyze a single matrix of time series data together with prior knowledge on gene function to detect temporal transcription modules that are biologically meaningful. Waltman et al. (97) and Dede and Ogul (98) proposed threeway clustering of gene-condition-organism data with or without external information such as sequence information in order to integrate data across species.

Gene expression or fMRI data measured over time and across subjects provide another common data source that calls for three-way analysis. These data are represented by an object x subject x time 3D matrix (i.e. a tensor of order 3) (99,100), where the measured object is either a gene, a parcel, or a voxel. For such matrices, Supper et al. (101) presented EDISA, an extension of ISA that seeks biclusters <G',S'> where G' is a set of genes (objects), and S' is a set of subjects, such that all genes in G' manifest a

similar time response across all subjects in S'. The SMARTS algorithm (102) represents a more flexible approach. This recently suggested algorithm integrates gene expression time series data with regulator-target network data in order to cluster subjects by their regulation patters. For each learned model a set of genes and their regulators are learned for each time point in which a significant change in the expression pattern is observed.

Extant models are limited in their ability to incorporate subject-specific signals into the discovered modules. For example, the set of genes active under one subject in a module may only partially overlap with the gene set of other subjects. Another limitation of some of the methods is the assumption of synchronicity of time points across subjects. Although this assumption is valid for technical repeats or well-tailored experiments, it is less plausible in other situations, e.g. samples taken from patients over time, due to possible heterogeneity in the response of different patients.

1.6 Summary of the articles included in this thesis

1. D. Amar and R. Shamir. **Constructing module maps for integrated analysis of heterogeneous biological networks**. Nucleic Acids Research doi:10.1093/nar/gku102, 2014.

Improved methods for integrated analysis of heterogeneous large-scale omic data are direly needed. Here, we take a network-based approach to this challenge. Given two networks, representing different types of gene interactions, we construct a map of linked modules, where modules are genes strongly connected in the first network and links represent strong inter-module connections in the second. We develop novel algorithms that considerably outperform prior art on simulated and real data from three distinct domains. First, by analyzing protein–protein interactions and negative genetic interactions in yeast, we discover epistatic relations among protein complexes. Second, we analyze protein–protein interactions and DNA damage-specific positive genetic interactions in yeast and reveal functional rewiring among protein complexes, suggesting novel mechanisms of DNA damage response. Finally, using transcriptomes of non–small-cell lung cancer patients, we analyze networks of global co-expression and disease-dependent differential co-expression and identify a sharp drop in correlation between two modules of immune activation processes, with possible microRNA control. Our study demonstrates that module maps are a powerful tool for deeper analysis of heterogeneous high-throughput omic data.

2. D. Amar, D. Yekutieli, A. Maron-Katz, T. Hendler and R. Shamir. **A hierarchical Bayesian model for flexible module discovery in three-way time-series data**. Bioinformatics, 31 (12): i17-i26, ISMB/ECCB 2015, proceedings paper, doi: 10.1093/bioinformatics/btv228, 2015.

Motivation: Detecting modules of co-ordinated activity is fundamental in the analysis of large biological studies. For two-dimensional data (e.g. genes × patients), this is often done via clustering or biclustering. More recently, studies monitoring patients over time have added another dimension. Analysis is much more challenging in this case, especially when time measurements are not synchronized. New methods that can analyze three-way data are thus needed.

Results: We present a new algorithm for finding coherent and flexible modules in three-way data. Our method can identify both core modules that appear in multiple patients and patient-specific augmentations of these core modules that contain additional genes. Our algorithm is based on a hierarchical Bayesian data model and Gibbs sampling. The algorithm outperforms extant methods on simulated and on real data. The method successfully dissected key components of septic shock response from time series measurements of gene expression. Detected patient-specific module augmentations were informative for disease outcome. In analyzing brain functional magnetic resonance imaging time series of subjects at rest, it detected the pertinent brain regions involved.

3. D. Amar, T. Hait, S. Izraeli and R. Shamir. **Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets**. Nucleic Acids Research, doi: 10.1093/nar/gkv810, 2015.

Genome-wide expression profiling has revolutionized biomedical research; vast amounts of expression data from numerous studies of many diseases are now available. Making the best use of this resource in order to better understand disease processes and treatment remains an open challenge. In particular, disease biomarkers detected in case–control studies suffer from low reliability and are only weakly reproducible. Here, we present a systematic integrative analysis methodology to overcome these shortcomings. We assembled and manually curated more than 14 000 expression profiles spanning 48 diseases and 18 expression platforms. We show that when studying a particular disease,

judicious utilization of profiles from other diseases and information on disease hierarchy improves classification quality, avoids overoptimistic evaluation of that quality, and enhances disease-specific biomarker discovery. This approach yielded specific biomarkers for 24 of the analyzed diseases. We demonstrate how to combine these biomarkers with large-scale interaction, mutation and drug target data, forming a highly valuable disease summary that suggests novel directions in disease understanding and drug repurposing. Our analysis also estimates the number of samples required to reach a desired level of biomarker stability. This methodology can greatly improve the exploitation of the mountain of expression profiles for better disease analysis.

1.7 Software tools and websites generated

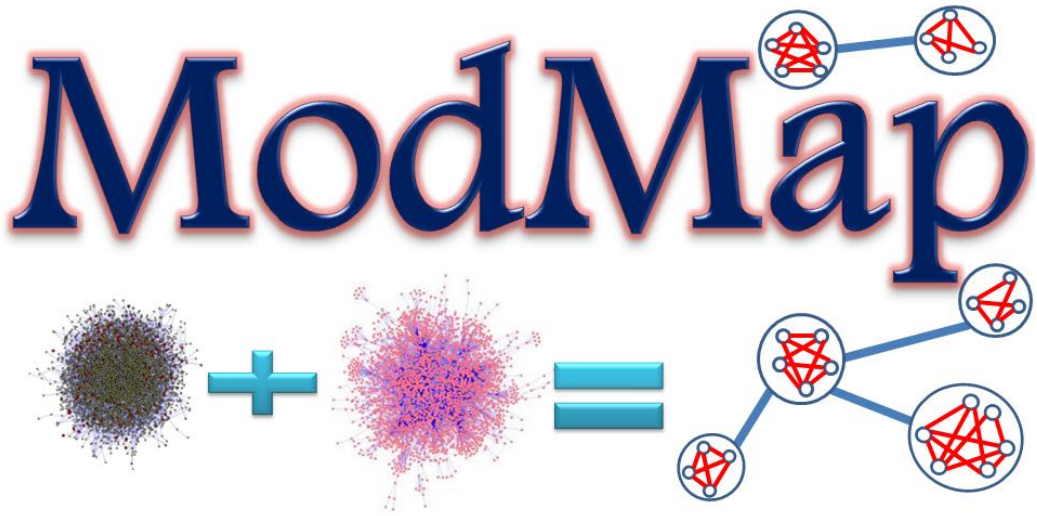
The tools presented in this thesis are available in three websites. All websites provide tutorials and the datasets used in the studies. First, http://acgt.cs.tau.ac.il/modmap/ provides Java runnables of ModMap, and the Java source code. Second, http://acgt.cs.tau.ac.il/twigs/ provides R implementation of our complete algorithm together with auxiliary scripts for running simulations. Researchers can use the scripts to compare algorithms using the same cases shown in the study. Finally, http://acgt.cs.tau.ac.il/adeptus/index.html provides our complete ADEPTUS database, including microarray expression profiles, RNA-Seq expression profiles, somatic mutation data, drug-target interactions, and gene interaction networks. These data are available as RData files that can be easily loaded into any R session. R code implementation of several learning algorithms is also available.

2. **Constructing module maps for integrated analysis of heterogeneous biological networks.**

D. Amar and R. Shamir.

Nucleic Acids Research

doi:10.1093/nar/gku102, 2014



Nucleic Acids Research Advance Access published February 4, 2014

Nucleic Acids Research, 2014, 1–12
doi:10.1093/nar/gku102

Constructing module maps for integrated analysis of heterogeneous biological networks

David Amar and Ron Shamir*

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Received September 24, 2013; Revised December 17, 2013; Accepted January 10, 2014

ABSTRACT

Improved methods for integrated analysis of heterogeneous large-scale omic data are direly needed. Here, we take a network-based approach to this challenge. Given two networks, representing different types of gene interactions, we construct a map of linked modules, where modules are genes strongly connected in the first network and links represent strong inter-module connections in the second. We develop novel algorithms that considerably outperform prior art on simulated and real data from three distinct domains. First, by analyzing protein–protein interactions and negative genetic interactions in yeast, we discover epistatic relations among protein complexes. Second, we analyze protein–protein interactions and DNA damage-specific positive genetic interactions in yeast and reveal functional rewiring among protein complexes, suggesting novel mechanisms of DNA damage response. Finally, using transcriptomes of non–small-cell lung cancer patients, we analyze networks of global co-expression and disease-dependent differential co-expression and identify a sharp drop in correlation between two modules of immune activation processes, with possible microRNA control. Our study demonstrates that module maps are a powerful tool for deeper analysis of heterogeneous high-throughput omic data.

INTRODUCTION

Biological networks provide a comprehensive overview of biological systems. They enable better understanding of the system and can shed light on the function of genes and other molecular compounds. Among other applications, they have been used for discovery and prediction of gene interactions, gene functions and disease–gene associations (1–9).

In these networks, the nodes represent molecular entities and the edges represent interdependencies. For example, in protein–protein interaction (PPI) networks, nodes represent proteins and edges represent physical interactions. In genetic interaction (GI) networks, nodes represent genes and edges represent the organism fitness for double-knockout perturbations, yielding two major types of edges: alleviating GIs and aggravating GIs. In alleviating GIs, also called positive GIs, the organism fitness after the double-knockout perturbation is better than expected based on the single-knockout results. In aggravating or negative GIs, the fitness is worse than expected. In gene co-expression networks, nodes represent genes and edges score the correlation in expression between the two genes (10,11). In gene differential correlation (DC) networks, edges score the change in gene pairwise correlation between one set of samples to another (e.g. cases and controls) (12–14). With the growing use and number of types of biological networks, computational methods that exploit these rich data are of great importance.

Computational methods that make use of several networks are often better than methods that analyze only a single network (4,7,8,15–19). For example, combined analysis of PPI networks and gene co-expression networks was used to detect gene sets that are co-expressed and are connected in the PPI network. Such analysis outperformed standard clustering algorithms and was successfully used for gene function prediction (5,8,16,19). Alleviating and aggravating GI data were used to find epistasis among and within gene sets. Under the premise that negative GIs tend to occur between compensatory pathways and positive GIs occur within pathways (or complexes), analysis of GIs was used to suggest a map of epistatic relations among functional gene modules (15,17,20–23). A marked improvement was reported after adding a connectivity constraint in a PPI network of the modules (15,17). The ability to construct a summary map of several networks allows identifying associations among discovered modules, thus improving the interpretability of the results compared with standard clustering of a single network.

*To whom correspondence should be addressed. Tel: +972 3 640 5383; Fax: +972 3 640 5384; Email: rshamir@tau.ac.il

Building on prior studies of specific pairs of networks, we introduce and study the fundamental problem of constructing a summary map of two biological networks H and G, where the nodes of both are the same genes or proteins, and the edges in each represent a distinct type of relations (see Figure 1D). The map nodes are gene sets that are strongly connected in H, and pairs of sets are connected by links. A link represents strong connection between two gene sets in G. The goal is to find gene modules in H simultaneously with finding module-to-module interactions according to G, by optimizing a specific objective function. We call this computational problem the 'module map problem'.

Most algorithms for the module map problem to date were used to find a summary map of epistatic interactions among pathways (15,17,20–23). Kelley and Ideker (15) proposed a method that is based on local searches in the graphs to find pairs of connected modules. Ulitsky *et al.* (17) used a clustering of H as a starting point and then improved the solution by merging modules. An algorithm akin to (15) has been recently proposed for analyzing gene co-expression and DC networks. The joint analysis of these networks revealed gene groups that are much more (or much less) correlated in one class of individuals (24). Although previous algorithms for the module map problem proved valuable, a thorough analysis of the

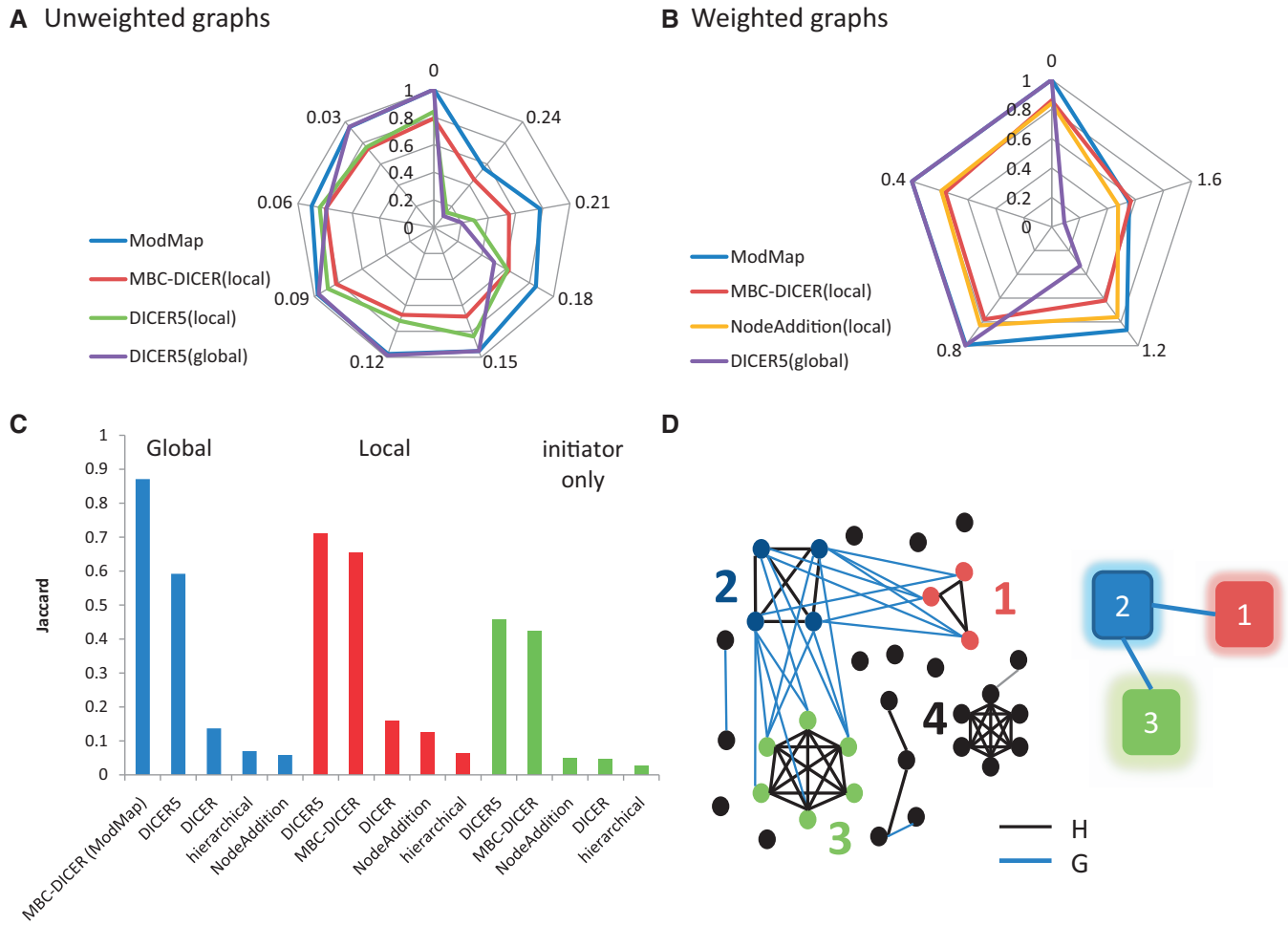


Figure 1. Module map: example and simulation results. (**A** and **B**) Performance of module map algorithms on 500-node graphs. (A) Unweighted graphs. (B) Weighted graphs. Each simulated pair of graphs contained an embedded module map of six modules in a tree structure. In addition, two random cliques and two bicliques were embedded in the graphs as decoys. Module, clique and biclique size was chosen uniformly at random between 10 and 20. In the unweighted model (A) each edge was replaced by a non-edge with probability P and vice versa. In the weighted model (B) edge weights are sampled from the normal distribution $N(1,\sigma)$, and non-edge weights are sampled from the normal distribution $N(-1, \sigma)$. Results are averages of 10 simulations for each data point. The four top performing algorithms for each simulation are presented using radar plots. MBC-DICER with global improvement is denoted as ModMap. The Jaccard coefficient between the modules produced by each algorithm and the true modules is shown as the distance from the center. Consecutive spokes from the top anticlockwise show increasing values of P in A and of σ in B. (**C**) Comparison of module map algorithms on unweighted graphs with 1000 nodes, containing a map of 10 modules and five decoys and $P = 0.15$. (**D**) A toy example of the module map problem; left: the two networks. Nodes are genes, H edges are black and G edges are blue; right: the module map. Nodes are modules and edges are links. Colors and numbers are the same on the left and right. The map contains three modules: module 2 is linked to modules 1 and 3, whereas module 1 and 3 are not linked. Black nodes are not part of the module map. The graph H (black edges) contains a clique that is not linked in G to another module and thus is not a part of the map. The example also demonstrates the difference between the local and global approaches. The local approach identifies modules 1 and 2 as linked, whereas the global approach also identifies module 3 as linked to module 2. See text.

problem and of the merits and weaknesses of these algorithms in different scenarios is required.

The problem of finding an optimal module map is NP hard under most formulations, as it contains the clustering of H as a subproblem. Hence, heuristics are used. These algorithms usually contain two phases. We call the first phase 'initiators': algorithms for finding an initial solution that may contain many small modules. The second phase uses 'improvers': algorithms for improving an initial solution according to a predefined objective function. A variety of algorithms can be formed from different combinations of initiators and improvers.

Here, we study novel and extant initiators and improvers. We show that a new initiator based on maximal bicliques in G together with a statistically formulated global improver strategy performs consistently better or equal to extant methods on synthetic and real networks of several types. We call the resulting algorithm ModMap. We apply ModMap to experimental data in three biological scenarios: (i) using yeast PPIs and negative GIs, we find epistatic relations among protein complexes, (ii) using yeast PPIs and DNA damage-specific positive GIs, we detect emerging connections among protein complexes involved in DNA damage response and (iii) using DC analysis of gene expression profiles of non–small-cell lung cancer (NSCLC) tissues, we identify disease-specific loss of correlation between immune activation processes and detect disease-specific microRNAs.

MATERIALS AND METHODS

Definition of the module map problem

The input to the problem is a pair of networks $H = (V,E_H,W_H)$ and $G = (V,E_G,W_G)$ defined on the same set of vertices. These networks can be weighted or unweighted. The goal is to find a module map that summarizes both networks. A module map is a graph $F = (M,L)$, where M is a collection of disjoint node sets, called modules, $M = \{M_1, \ldots, M_p\}$, $Mi \subseteq V$, $M_i \cap M_j = \emptyset$, and L is a set of module pairs $\{(U_1,V_1), \ldots, (U_p,V_p)\}$, where each U_i and V_i are in M. These pairs are called the map links. In addition, each module must be linked to at least one other module. Roughly speaking, our goal is to find a module map such that each module corresponds to a heavy subgraph of H, and each link represents a heavy bipartite subgraph in G between a pair of modules. A formal notion of heavy subgraphs will be introduced later. Figure 1D shows a toy example of two unweighted networks and their module map.

Previous algorithms for constructing module maps vary in the way they define the objective function and the links. The DICER algorithm (24) seeks one pair of linked modules at a time. A pair of modules is defined as linked if the sum of weights W_G between them is high enough. We call the approach of DICER 'local', as it finds one module pair at a time. The algorithm of Ulitsky *et al.* (17) aims to maximize the 'global score', namely, the total sum of scores within modules in H plus the sum of scores of links in G. In addition to increasing the global score, links between modules are accepted only if they pass a statistical significance test. We call the second approach 'global'. Both methods identify the links and the modules simultaneously.

Figure 1D demonstrates the differences between the local and global approaches. Assume that in both graphs edge weights are 1, non-edge weights are −1 and that the local approach uses a threshold of 0 on the sum of W_G weights between two modules for reporting a link. In both approaches, modules are clusters of nodes with high density in H. According to both approaches, module 1 is linked to module 2: the local score is 4 (8 edges and 4 non-edges), the global analysis *P*-value for linkage is <0.05, and the total score for the module pair is 13 (module score 6 + 3 + link score 4). The sum of W_G weight between modules 2 and 3 is −4 (10 edges and 14 non-edges), and the local method rejects that link. However, the global approach will also link module 2 and 3: the linkage *P*-value is significant ($P = 0.039$), and adding this link will improve the global map score to 24 [13 for the (1,2) pair +15 for module 3–4 for the (2,3) link]. This example illustrates the advantage of the global approach on sparse graphs, in which large modules are not expected to be densely interconnected.

Algorithms

We conducted a systematic study and developed further a family of two-phase algorithms for module map detection that find an initial solution (possibly consisting of many small modules) and then improve it. We call algorithms for the first phase initiators and algorithms for the second phase improvers. For simplicity, we describe the algorithms assuming that edges with positive weight are considered heavy. For unweighted graphs, we assume edge weights to be 1 and non-edge weights to be −1. For weighted graphs, all node pairs (edges) have weights, so there are no non-edges.

Initiators

We tested five different initiators: (i) DICER (24), which finds one pair of linked modules at a time, (ii) hierarchical clustering of the graph H (25), which finds a set of modules, (iii) a greedy node addition algorithm for finding modules in H, (iv) $DICER_k$ a variant of DICER wherein the minimum module size is set to k and (v) an algorithm based on enumeration of maximal bicliques in G using an exhaustive solver (26,27), followed by the cleaning process of DICER. We call the latter algorithm MBC-DICER, see Supplementary Text and Supplementary Figure S1 for a full description of all initiators. Each initiator creates an initial module set, but modules in the map constructed by clustering algorithms are not necessarily linked.

Improvers

The 'local improver' (24) extends module map links by either adding a single node to a module or by merging two module map links. One drawback of this approach is that it cannot create new modules that are not

represented in the initial solution. Another disadvantage is that it cannot merge a module whose two parts are linked to different modules that are unlinked. See Supplementary Figure S2 for examples. Later in the text we introduce the global improver, which can often overcome both problems.

Our 'global improver' is based on the procedure in (17). Let $M = \{M_1, \ldots, M_n\}$ be a collection of disjoint node sets (e.g. a set can be a single gene or not linked to any other set). Given sets (U,V), U, $V \in M$ and $x \in U$, the significance of the linkage of x with V is calculated using Wilcoxon rank-sum test by comparing the edge weights W_G between x and V to the edge weights between x and all nodes not in V. Such *P*-values are calculated for all nodes in U and V, and they are combined using Stoufer's method (28). If the final *P*-value p(U,V) is at most α then U and V are connected by a 'link' in the map. Let $L = \{(U_1,V_1), \ldots, (U_p,V_p)\}$ be the resulting set of links.

The 'global score' of the solution is the sum W_H of edge weights within each M_i plus the sum of W_G edge weights between the linked node sets:

$$S(M,L) = \sum_{i|M_i \in M} \sum_{s,t \in M_i} W_H(s,t) + \sum_{k,l|(M_k,M_l) \in L} \sum_{i \in M_k, j \in M_l} W_G(i,j)$$

$$s.t. \forall_{(M_k,M_l|k \neq l, M_k, \in M, M_l \in M)} (M_k,M_l) \in L \Leftrightarrow p(M_k,M_l) \leq \alpha$$

The improvement stage merges a pair of node sets (two modules or a module and a single gene) if the global score increases and the new link passes the significance test. Considering a merge that creates a new module Y requires recalculating p(Y,Z) for all other modules Z in M, in order to calculate the global score. This process is done greedily: iteratively, the merge that yields the best improvement is performed until no possible merge can improve the global score.

We modified the aforementioned method to allow for fast analysis of large graphs as follows. First, when calculating p(U,V), we consider the links in G' (the unweighted version of G). We use a hypergeometric test to evaluate if a node has significant number of edges in G' to the opposite set (e.g. from a node $v \in V$ to the set U), and then all node *P*-values are merged using Fisher's method (29). The sets U and V are linked if the resulting value $\leq \alpha$. This test is much faster and provides maps of equal quality to using the Wilcoxon test on G (see Supplementary Text). Weighted tests, such as the Wilcoxon test, are not always appropriate for detecting linkage among gene modules. For example, in the DC graphs, a strong link must contain many positive edges, whereas the Wilcoxon test only looks at the ranks of the edge scores.

Second, we set another parameter $\beta >> \alpha$, and if at some point the *P*-value for the possible link between two sets is at least β, we say that the sets are 'anti-linked'. In the original algorithm, when considering merging two sets U and V into W, possible links between W and every other set Y must be calculated. However, if U and Y are anti-linked or V and Y are anti-linked then we mark W and Y as anti-linked, avoiding the need to consider the possible link (W,Y). In practice, we used $\alpha = 0.005$ for the yeast data as suggested in (17) and tested several options in the gene expression data (see Supplementary Text). In all cases we used $\beta = 0.2$. Finally, we perform multiple merge steps simultaneously in a single iteration in a way that guarantees that the global score improves (see Supplementary Text). This provides a speed up of two-fold or more in practice without loss of solution quality.

Simulations

We constructed initially empty 500-node graphs H and G and then added edges creating a perfect module map in which modules are cliques in H and links are bicliques in G. The module map topology (M,L) was a random tree with $|M| = 6$. We then added two H-cliques and two G-bicliques to the graphs to represent additional 'decoy' structures that are not part of the map. Clique, biclique and module sizes were randomly selected in the range 10–20 with uniform distribution and disjoint node sets. Call the resulting edge sets E_H^* and E_G^*. Finally, we modified these graphs by introducing random noise: each edge in G and H was deleted with probability *P*, and each non-edge was replaced by an edge with probability *P*. All reversal steps were done independently. For creating weighted graphs, the same procedure was used, but all possible edges are present in the final H and G: w(u,v) is sampled from $N(1,\sigma)$ if (u,v) is in E_H^* or E_G^*, and from $N(-1,\sigma)$ otherwise. We also generated in this manner 1000 node graphs with 10 or 20 modules and five decoys (cliques and bicliques).

Analysis of negative genetic interactions and protein-protein interactions in yeast

The PPIs and the negative GIs were downloaded from BIOGRID (30). These networks were used to find epistatic relations among protein complexes. The PPI network was used as H, and the GI network was used as G (see Supplementary Table S1).

Analysis of DNA damage-specific genetic interactions data

We used the data of (21), in which all pairwise GIs among 418 genes were tested, and of (31), which tested GIs between 55 query genes and 2022 genes. A 'DNA damage-specific positive GI' was defined as one that had $S < 0$ in the untreated cells, $S > 0.5$ in the treated cells and the *P*-value for differential GI was <0.01. This analysis yielded 840 interactions from (21) and 1677 interactions from (31). We additionally defined a positive GI as 'stable' if it had $S > 1.5$ both in the untreated cells and in the DNA damage cells. This analysis provided 491 interactions in (21) and 3139 interactions in (31). Owing to the different experimental setups most of these GIs are not directly comparable.

Calculating differential correlation scores

Given a training set containing gene expression profiles of subjects, we used the statistical method of (24) to compute for each gene pair its consistent correlation (CC) and DC

scores. First, DC scores are computed using the real labels of the samples. Then, the scores are transformed to log-likelihood ratio (LLR) scores by comparing the original DC scores to scores calculated on the same data with randomly shuffled labels. Thus, positive LLR scores mark gene pairs with significant change in DC. The prior probability of real DC changes was set so that only correlation changes of at least 0.4 will have a positive LLR score. This approach guarantees a similar yet slightly more stringent acceptance threshold compared with (24). See Supplementary Text for additional information.

GO and microRNA enrichment analysis

We used TANGO (32) for Gene Ontology molecular function and biological process enrichment analysis of modules and FAME (33) for microRNA enrichment analysis. Both tools are available as part of the EXPANDER software (34). When a set of modules was analyzed, we corrected for multiple testing using false discovery rate (FDR) with $q = 0.05$. The background set for the enrichment analysis was defined as the set of genes in the networks and not all genes in the organism. This filtering step reduces bias in case of overrepresentation of GO terms in the networks.

Network visualization

Network visualization was done using Cytoscape (35).

Availability

A command line tool for running ModMap is freely available for academic use at http://acgt.cs.tau.ac.il/modmap/.

RESULTS

Simulations

We first tested the different algorithms on synthetic graphs H and G. Starting from a perfect module map, we first added cliques in H and bicliques in G to represent additional structures that are not part of the map and then introduced random noise to the edges. To generate both sparse and denser graphs, we tested a wide range of the noise parameters σ and p in the weighted and the unweighted simulations, respectively (see 'Materials and Methods' section). The results presented here are for graphs with 500 nodes and six modules per map. We also tested larger graphs with similar results (see Supplementary Figures S3 and S4).

We tested 10 combinations of initiator and improver on 10 random data sets for each value of P and σ. We measured the quality of produced solutions using Jaccard coefficient between the reported modules and the known modules. The results of the unweighted and weighted models are shown in Figure 1A and B, respectively. Only the four algorithms that performed best on average in each simulation are shown. Supplementary Table S2 contains the results for all combinations. The local improvement algorithms did not reach perfect scores even on noiseless data. In contrast, MBC-DICER and $DICER_5$ followed by global improver reached perfect Jaccard scores when there was no statistical noise. The high performance of MBC-DICER remained robust even when noise levels were as high as $P = 0.15$ in the unweighted model and $\sigma = 1.2$ in the weighted model. A comparison of all algorithms on unweighted graphs with 1000 nodes and 10 modules for noise level $P = 0.15$ is shown in Figure 1C. Performance remains high although the graphs are much larger. Using the improvers was beneficial compared with using only the initiator solutions, especially for the DICER variants. MBC-DICER with the global improver reached highest performance (0.87). Interestingly, the local improver was better than the global improver for all other algorithms (e.g. 0.71 versus 0.59 for DICER5). This is probably because the MBC-DICER initiator detects robust fully connected modules, which are a better starting point to the global improver at high noise levels. Tests with different values of k for the DICERk algorithm led us to choosing k = 5 (Supplementary Figure S4). In addition, we compared the performance of the global improver with the hypergeometric test and with the Wilcoxon rank-sum test, which was used in previous studies. Our results show that using the hypergeometric test reaches similar quality of results but is much faster (see Supplementary Text). Overall, the results indicate that MBC-DICER followed by the global improver achieved the best performance on both unweighted and weighted data. We call the resulting algorithm ModMap and will use it as the algorithm of choice from now on.

Yeast protein-protein interaction and negative genetic interaction data

We used PPIs and negative GIs from BIOGRID (30) to find epistatic relations among protein complexes. Only genes that had both types of interactions were used. Overall, the networks contained 3979 genes, 45 456 PPIs, and 76 237 negative GIs (the interactions are listed in Supplementary Table S1). This number of genes and edges is larger than in previous studies. For example, (22) covered 1460 genes, and (17) covered 743 genes. Therefore, our networks have the potential to provide a broader overview of the yeast interactome and allow for a comprehensive performance testing of the different algorithms.

As done in previous studies, we evaluated solutions by their statistics and the functional characterization of the modules (17,22). The calculated solution statistics included the number of modules, the number of genes covered and the maximal module size. We used TANGO (34) to measure module functional enrichment, and reported the number of discovered GO terms, the percent of enriched modules and the percent of module map links for which both modules are enriched (with the same or with different functions), which we call 'enriched links'. Enriched links represent dense GIs among known biological terms.

The solution statistics of all algorithms are shown in Supplementary Table S3. One can observe clear superiority of global over local improvers. In contrast to global

improvers, which reported at least 100 modules and covered 800–1000 genes, the local improvers found 2–28 modules covering only 15–192 genes. Except for DICER, the results of all solutions were similar and of high quality. ModMap was the best in terms of the percent of enriched modules (87%) and percent of enriched links (80%). Taken together, the map of ModMap was best in combining functional comprehensiveness and quality. We also compared ModMap with other weighted approaches for GI data analysis (22,36) on the data of Collins *et al.* (37). See Supplementary Text for details. Our results show that ModMap produces high quality maps and improves on extant weighted approaches.

Figure 2 shows a portion of the map constructed by ModMap where links were restricted to $P < 10E\text{-}50$ (for details see Supplementary Tables S4–S6). Each node represents a module, and edges represent map links. All modules in the presented map are enriched at 0.05 FDR with at least one GO term. The node labels show the most significantly enriched term. Three major hubs are marked in green: Rpd3L complex (14 genes, $P = 4.35E\text{-}38$), Swr1 complex (13 genes, $P = 1.08E\text{-}35$) and the mediator complex (17 genes, $P = 4.89E\text{-}43$). The Rpd3L and Swr1 complexes are chromatin related and were previously annotated as hubs of GIs in a gene-based study (38). Bandyopadhyay *et al.* (21) discovered some of the same links; however, module annotation there was manual, whereas our analysis was completely automatic and produced a much larger map. Moreover, our map extends on the previous observations by showing that the three hubs are linked and by providing additional links for the Rpd3L complex. In Figure 3, we focus on the three most significant links in the map ($P < 1E\text{-}70$). Figure 3A shows the connections between the Rpd3L and Set3 complexes and between the Rpd3L and Swr1 complexes. Rpd3L and Set3 are both histone deacetilases, and negative GI between them was reported in (20). The Rpd3L complex was split into two disjoint modules, whereas in our map it is detected as a single module, containing all 14 Rpd3L genes. Figure 3B shows a connection between two well-established subunits of the proteasome complex (39). This example shows how joint analysis of PPIs and GIs correctly detects core functional subunits even when they are connected by many PPIs.

Analysis of DNA damage response networks in yeast

The module map described earlier in the text was obtained by analyzing the entire set of known negative GIs. Recent

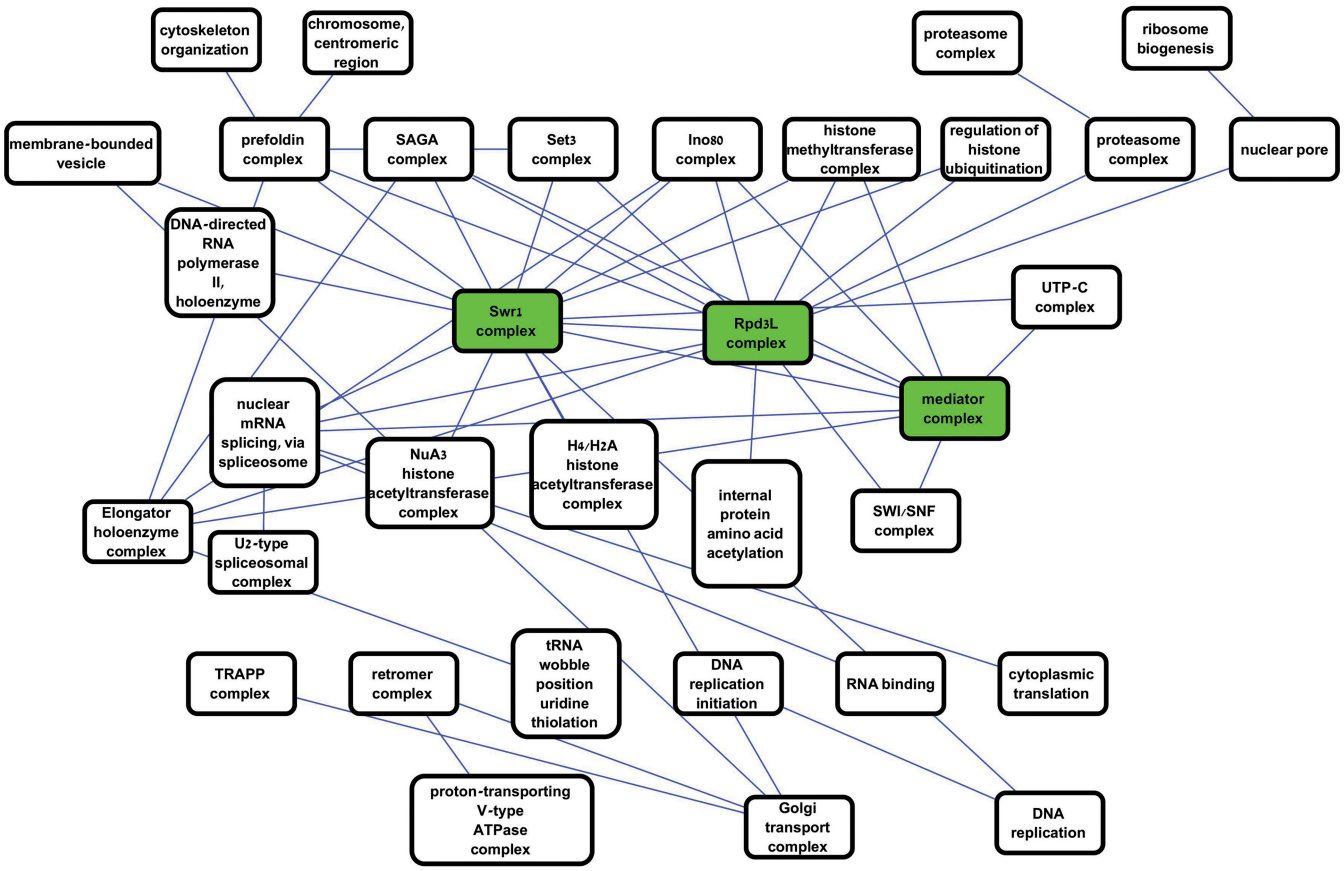


Figure 2. The yeast module map. Each node is a module in the yeast PPI network. The name of a node is the most significantly enriched GO term for that module. Each edge represents a highly significant link between two modules in the negative GI network ($P < 1E\text{-}50$). Modules that were not enriched for any GO term at 0.05 FDR are not shown. Three main chromatin-related hubs are marked in green. Some links connect disjoint modules enriched with similar GO terms (e.g. proteasome–proteasome link, top right), and other links show epistasis between different biological processes (e.g. nuclear pore and ribosome biogenesis, top right).

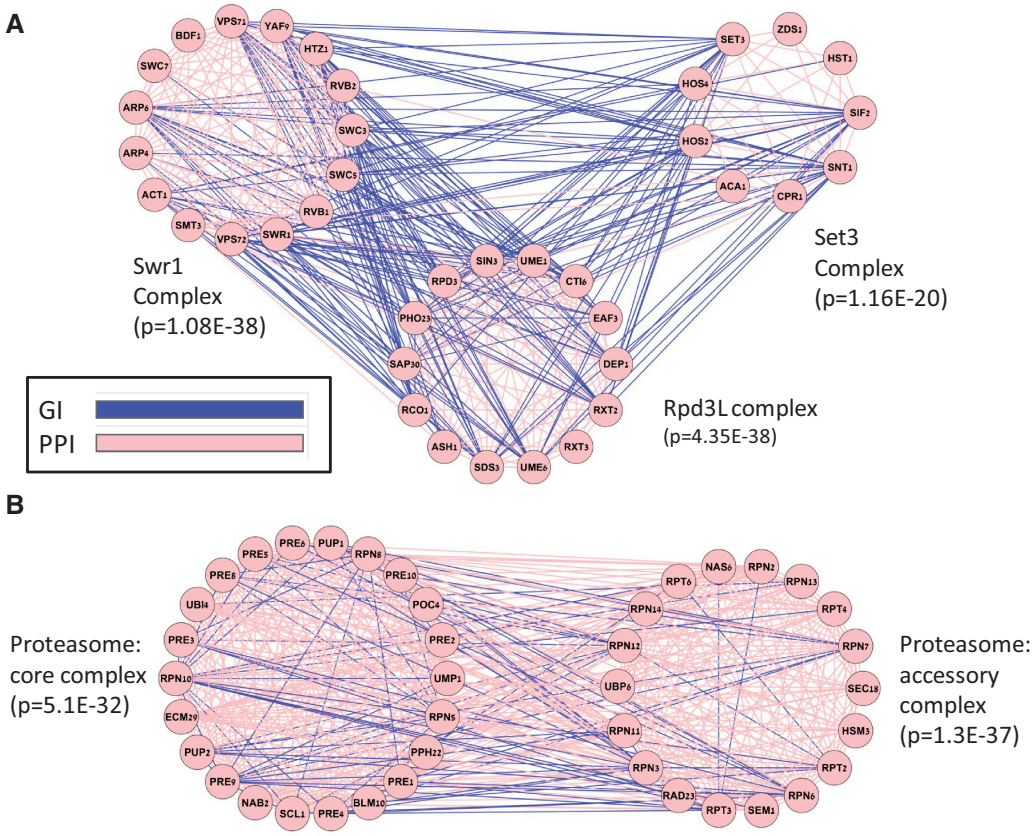


Figure 3. Examples of linked modules in the yeast module map. The genes of each module are arranged in a circle. Blue edges represent negative GIs and pink edges represent PPIs. For each module, the most enriched GO term is shown along with its enrichment *P*-value. (**A**) Linkage among different protein complexes. The significance of the links between Rpd3L and the Set3 complexes and between Swr1 and Rpd3L complexes is <10E-70. The link between Swr1 and Set3 is also highly significant ($P = 4.29E-59$). (**B**) Detection of subcomplexes. The joint analysis of the PPI and GI networks partitions the proteasome complex into its two subcomplexes: the accessory and the core complex.

studies have gone beyond static analysis to detect changes in the GI network in response to DNA damage (21,31). In these studies, GIs were measured in untreated cells and following perturbation by the DNA-damaging agent methyl methanesulfonate (MMS) (40). We combined two such data sets (21,31) to detect 'DNA damage-specific positive GIs', i.e. differential positive GIs that emerge in the treated cells and are not observed in the untreated cells (see 'Materials and Methods' section). Negative GIs are typically observed between genes working in parallel, such as genes that are involved in two compensatory complexes or pathways that backup each other, and thus the loss of one is buffered by the other. Positive GIs are more likely to be observed between genes from the same complex or pathway, where most of the phenotypic effect is already observed in each single-knockout. Hence, DNA damage-specific positive GIs are expected to represent changes of the network in response to MMS, revealing DNA damage-specific interactions within pathways or between different pathways or complexes working in series. In total, 1078 genes were included in both studies, with 2227 DNA damage-specific positive GIs among them (see Supplementary Table S7). There were 6771 PPIs within that gene set.

We applied ModMap with the PPI network as H and the DNA damage-specific positive GI network as G. Because these networks were much smaller than in the previous analysis, we set the minimal module size to three. The small module sizes also affected the attainable *P*-values for links. Here, a pair of modules was defined as linked if its *P*-value was < 0.05 after Bonferonni correction, considering all statistical tests done by the algorithm during the improvement steps.

The generated module map contained 78 genes in 12 modules, with 17 links among them. Module sizes ranged between 3 and 15. A complete description of the map is provided in Supplementary Tables S8–S10. A map of the modules that were significantly enriched with GO terms is shown in Figure 4A. The hub in this map is a module enriched with DNA repair genes, linked to six modules that cover a large variety of functions. In Figure 4B, we focus on the DNA repair-related module and on three of the modules linked to it. The DNA repair module contains four genes: *RAD5*, *RAD18*, *HPR5* and *UBC13*. Interestingly, although *UBC13* is known to physically interact with the three other genes, positive GIs that are consistently stable across experiments (see 'Materials and Methods' section) connect the other three genes, providing further evidence that the four genes are involved in a common process. The *RAD5*, *RAD18* and *UBC13* genes are known to be involved in post-replication repair (41–43) and *HPR5* is involved in checkpoint recovery (44,45).

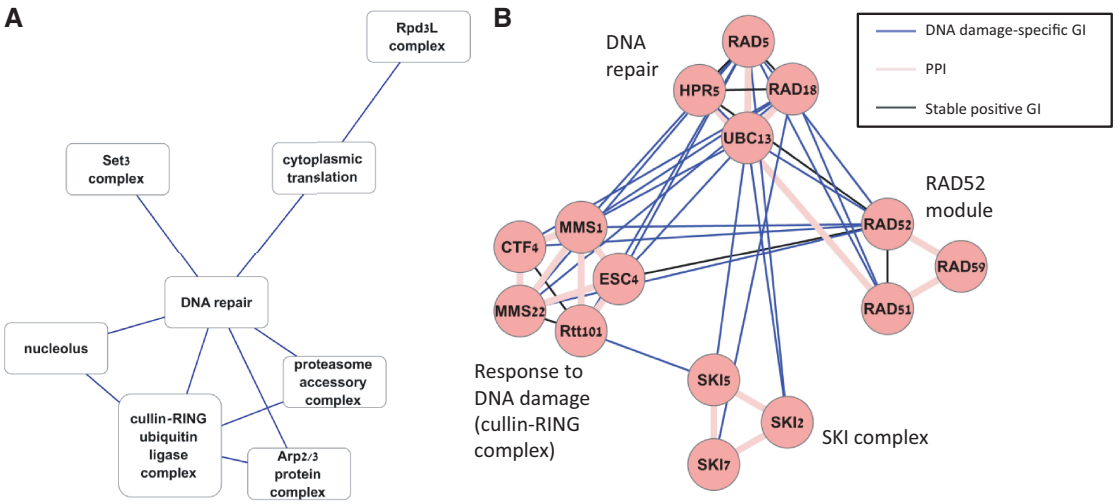


Figure 4. A module map of DNA damage-specific positive GIs. (**A**) A module map of the significantly enriched modules. Nodes represent modules and edges represent significant links (Bonferonni corrected $P < 0.05$). The name of a node is the most significantly enriched GO term. (**B**) A closer look at the DNA repair module and three-linked modules. Nodes represent genes and edges represent interactions: blue—DNA damage-specific positive GIs, pink—PPIs, black—stable positive GIs, which are observed both in the untreated and in the treated cells. This map shows the emerging connections between functional modules on DNA damage response covering DNA repair and checkpoint responses in the DNA repair module, response to damaged replication forks (the DNA damage response module), DNA double-stranded response genes (*RAD52* module) and RNA degradation-related genes (SKI complex module). The *RAD52* and SKI modules do not appear in A, as they reflect functions that do not have established GO terms.

The DNA repair hub module is linked to a module associated with response to DNA damage. It contains five genes: *CTF4*, *ESC4*, *MMS1*, *MMS22* and *Rt101*. The last four genes are part of the cullin-RING ubiquitin ligase complex (GO:0031461). The last three genes were shown to form a complex that stabilizes the replisome during replication stress (46,47). The *CTF4* gene is related to DNA repair and DNA replication initiation according to its GO annotations. The link suggests that this complex might work together with the DNA repair module for coping with damaged replication forks. Interestingly, the two MMS genes were originally detected in MMS sensitivity tests but are not expected to be required for double-stranded repair (47). The *RAD52* module (*RAD51*, *RAD52* and *RAD59*) is related to double-stranded DNA damage repair (48) and is linked both to the DNA damage repair module and to the DNA damage response module, suggesting these modules work together in the same pathway as a result of DNA damage to cope both with damaged replication forks and with double-stranded DNA breaks. The fourth linked module contains three genes of the SuperKiller (SKI) complex (*SKI2*, *SKI5* and *SKI7*). These genes are involved in 3–5 RNA degradation in the cytoplasmatic exosome (49,50). Our analysis suggests that this complex might also be involved in response to DNA damage. Previous studies have shown that RNA degradation cytoplasmatic genes might play a role in DNA damage response separately from their cytoplasmatic activity (51,52). The suggested roles of RNA degradation genes in DNA damage response include DNA stability and telomere stability related functionality (51), mediating the assembly of multiprotein complexes in double-stranded breaks (52) and specific mRNA degradation on DNA damage (53). Hence, our findings match prior studies and strengthen the role of the SKI complex in the response to DNA damage.

Analysis of human co-expression and differential correlation networks

We applied ModMap on case-control gene expression data of NSCLC to reveal DC among highly correlated gene modules. The contribution of this part is two fold. First, we show that DC among gene modules is reproducible in cross-validation tests. Second, we analyze the map of DC patterns between gene modules discovered by ModMap.

Given a data set of gene expression profiles from cases and controls, we used the method of (24) to compute two scores for each gene pair: the CC score, which is positive if the gene pair is consistently correlated across phenotypes, and the DC score, which is positive if the correlation difference between the cases and controls is higher than expected by chance. These scores were then used as edge weights in networks H and G, respectively, on which a module map was learned. The methodology was evaluated using cross-validation: given a module map constructed on a set of profiles (the 'training set') and a disjoint set of samples (the 'test set'), the quality of the predicted map was evaluated on the test set by comparing the DC of links and of non-links using Wilcoxon rank-sum test, where the null hypothesis is that there is no difference in DC between links and non-links. This measure is parameter-free and reflects all DC changes.

We tested several variants of the algorithm using 2-fold cross-validation. The maps produced by the local improver received low *P*-values but suffered from low coverage. For example, for the MBC-DICER initiator, the local improver achieved a *P*-value of 4.43E-4, but

the map covered only 197 genes. In contrast, when applying ModMap (i.e. MBC-DICER with the global improver), the map covered 1289 genes, with *P*-value of 1.54E-10. Supplementary Text contains further results of testing different parameters of the global improver and tests on Alzheimer's disease (54), which got similar cross-validation results. The full results are shown in Supplementary Table S11 for lung cancer and in Supplementary Table S12 for Alzheimer's disease. Taken together, ModMap produces large maps that are robust when tested on independent data sets.

Next, we analyzed the module map obtained by running ModMap on all samples of the NSCLC data. The map covered 1921 genes in 76 modules, connected by 405 links (see Supplementary Tables S13 and S14 for details). To focus on strong changes in correlation between modules, we compared the DC of each link in the map to the DC calculated between random gene sets of the same sizes in 200 repeats and calculated the fold-change between the real link and the best random link as proposed in (24). The link fold-change scores are given in Supplementary Table S14. In all, 150 links had fold-change $\geq$1.5, with the top five links exceeding 2.3. This indicates that the DC of the linked modules is far stronger than expected by chance. We also analyzed the modules of the top links using pathway enrichment analysis and microRNA enrichment analysis (see Supplementary Table S15 for details). One of the links connected two modules related to immune response activation. The linked modules are shown in Figure 5. In Figure 5A, we observe many high co-expression edges between the modules (gene pairs with $r > 0.4$) in the control class. Module 11 is enriched with B-cell receptor signaling pathway genes (6 genes, $P = 3.1E-8$). Module 12 is enriched with T-cell receptor signaling pathway genes (4 genes, $P = 1.37E-4$). Figure 5B shows GeneMANIA analysis of these 10 genes (7,55), which confirms that they are connected by several types of interactions. Figure 5C shows the co-expression of the same modules in the NSCLC class. Within each of the modules a strong level of co-expression is preserved, but the co-expression between the modules is abolished, suggesting that co-regulation of the different immune responses is lost in NSCLC. Finally, module 11 is highly enriched with targets of microRNA 34-a, b, c family (red nodes in Figure 5A), whose members are annotated as causal to NSCLC according to the mir-2-disease database (56). Taken together, these results show the ability of our analysis to detect NSCLC-related functional modules without using any prior knowledge.

DISCUSSION

In this article, we presented a methodology for joint analysis of two gene networks, each representing a different type of omic relation between genes. The method identifies gene sets as modules and the complex structure of relations among them and summarizes the analysis in a module map. Modules correspond to interacting gene sets in the first network, and links in the module map correspond to interacting modules in the second. The map is constructed based on both networks simultaneously and thus can capture and reveal structures that are not identifiable when analyzing each data type separately. Our novel algorithms recovered the planted map structure in simulated data, even when the noise level in the data was high. We tested our methods in three biological applications: (i) yeast PPIs and negative GIs, (ii) yeast PPIs and DNA damage-specific positive GIs and (iii) DC analysis of human disease expression profiles. In all cases, certain parts of our maps are supported by prior biological knowledge, whereas other parts reveal novel structure and suggest new biological findings. The module map paradigm can be applied in principle on any two types of networks with underlying common nodes.

Our analysis of the yeast PPI and negative GI data constructed a large map describing epistatic relations among complexes. Our findings are in agreement with previous studies and show a complex map of interactions among chromatin modification-related complexes but also provide interactions with other functions, such as protein modification-related complexes. The analysis of the yeast PPIs and DNA damage-specific positive GIs produced a smaller map, which contains a DNA repair module as a central hub. The interactions of this module suggest that several mechanisms emerge simultaneously in response to MMS, including double strand repair, damaged replication fork repair and exosome complex activity. In the map constructed based on human NSCLC blood expression profiles, modules represent gene sets that are highly co-expressed both in cases and in healthy controls, whereas the map links correspond to specific rewiring of the co-expression network in NSCLC patients. In particular, we identified two modules enriched with immune activation genes manifesting a sharp drop in correlation in the NSCLC patients, suggesting diminished coordination between the T-cell and the B-cell enriched modules.

The concept of a module map can be viewed as a higher level combination of clustering and biclustering. Each of those problems has been extensively studied and was applied successfully to numerous single-type genomic and proteomic studies (1,57–68). By performing joint analysis on two different data types, we allow some relaxation of the objective function in each of the networks, for the sake of obtaining an overall clearer structure. Therefore, the new analysis can yield results when clustering or biclustering of one data type fails. One of the difficulties in clustering and biclustering is that module (or module-pair) sizes must be large enough to obtain highly significant sets. As our analysis demonstrates, the added power of the module map approach can identify relatively small precise groups that are beyond the detection ability of those prior methods.

Only a handful of studies have addressed the module map problem to date, and most of them focused on joint analysis of yeast PPI and GI networks. Ulitksy *et al.* (17) and Bandyopadhyay *et al.* (69) developed clustering methods that seek a map in which the likelihoods of the edge weights of PPIs and GIs within clusters or of GIs between linked clusters are higher than a given

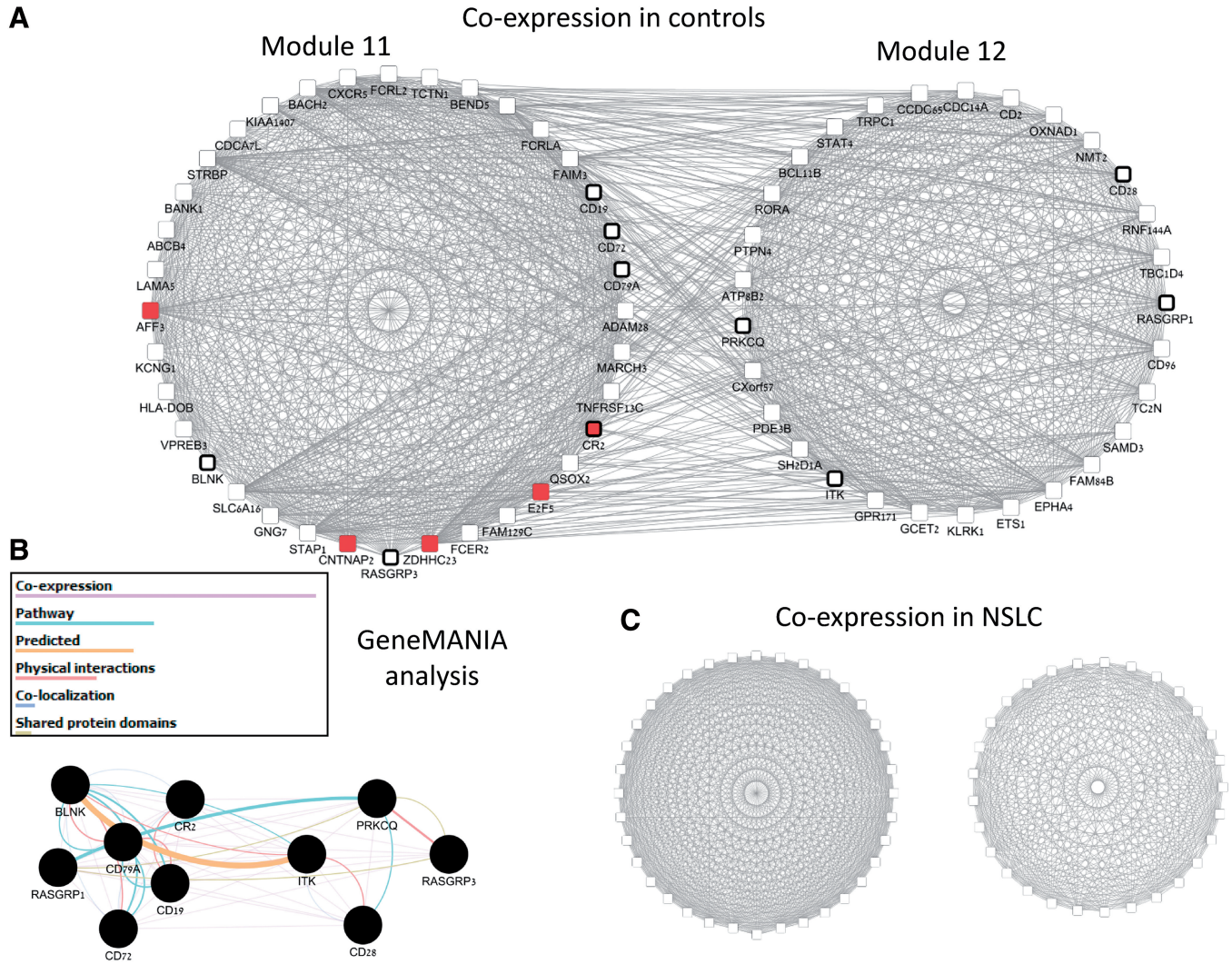


Figure 5. A pair of immune activation-related modules differentially correlated in NSCLC. (**A**) Two-linked modules, which are a part of the constructed module map. Nodes are genes and edges represent correlation >0.4 between the genes in the expression patterns of control class. Edges here correspond to high co-expression between two genes and do not reflect the weights in the CC or DC networks. We observe strong co-expression both within and between the modules. Nodes with black frames are related to immune activation response (six T-cell activation genes in module 11 and four B-cell activation genes in module 12). Red nodes in module 11 are targets of mir-34 family. (**B**) GeneMANIA analysis of the T-cell and B-cell signaling pathway genes shows that the genes of both modules are expected to interact in healthy controls. (**C**) The same two modules and their co-expression network in the NSCLC class. As in A, the genes within each module are highly co-expressed. In contrast to A, co-expression between the modules is completely diminished.

background distribution. Leiserson *et al.* (22,36) sought local maximum cuts in the weighted graph of the GIs by a greedy incremental approach, producing a collection of linked pairs of modules. Kelley and Ideker (20) developed a clustering algorithm that is based on graph compression, where the original GI graph is compressed to a module map. Hence, both (22,36) and (20) look for approximate bicliques that connect gene modules. In contrast, we enumerate the maximal bicliques of GIs, analyze them by taking into consideration the two interaction types to ensure that the initial solution contains dense strongly connected modules and improve the solution using our global improver. Because our approach is generic, it does not exploit the specific probabilistic nature of the GI data as other methods do (22,36). Nevertheless, we show that our method outperforms these and other extant methods in several criteria on GI data. In addition, because our algorithm is not limited by the type of the input data, we are able to combine many heterogeneous data sets (e.g. using all GIs of BioGRID) in our analysis.

When dissecting human expression profiles of disease patients and healthy controls, DC analysis was proposed as a way to discover gene modules whose inter-module correlation levels are altered in disease (12,14,24,70). We previously developed DICER (24), which uses a local approach to detect module pairs. Here, we go beyond it by finding maximal bicliques in the DC graph and by concurrently constructing a global map of modules. As we showed here, in most cases the map links are highly significant. However, we also observed cases where the absolute correlation change of modules

might be mild even though the DC of the module pair is significant. A possible remedy is to give more emphasis to high absolute DC of map links so as to see the DC signal better. Another possible improvement is to enumerate bicliques using established heuristics [e.g. (68)].

A key factor in the performance of the ModMap algorithm is the objective function optimized. Here, we chose to maximize the sum of weights within modules plus the sum of weights of module links and assigned these weights based on a probabilistic model. On unweighted networks, such as the PPI and GI yeast networks, we set the weight of an edge to 1 and the weight of a non-edge to −1, thereby promoting strongly connected modules and links. This setting produced good results and revealed functional interactions among protein complexes. By setting different weights to non-edges in the graphs, future analyses can promote modules that are sparser, thus enabling better detection of interactions among complete pathways.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Yaron Orenstein for his comments on the manuscript. Conceived the method and designed the experiments: D.A. and R.S. Designed and wrote the software used in analysis: D.A. Performed the experiments: D.A. Analyzed the data: D.A. and R.S. Wrote the paper: D.A. and R.S.

FUNDING

Israel Science Foundation [802/08 and 317/13]; Israel Cancer Research Fund; Lee Perlstein Kagan Charitable Trust (in parts). Azrieli Fellowship from Azrieli Foundation, Edmond J. Safra Center for Bioinformatics at Tel Aviv University, Israeli Center of Research Excellence (I-CORE), Gene Regulation in Complex Human Disease, Center No 41/11 (to D.A.); The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Funding for open access charge: Israel Science Foundation and I-CORE.

Conflict of interest statement. None declared.

REFERENCES

1. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
2. Deng,M.H., Zhang,K., Mehta,S., Chen,T. and Sun,F.Z. (2003) Prediction of protein function using protein-protein interaction data. *J. Comput. Biol.*, **10**, 947–960.
3. Kharchenko,P., Chen,L., Freund,Y., Vitkup,D. and Church,G.M. (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, **7**, 177.
4. Pandey,G., Myers,C.L. and Kumar,V. (2009) Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics*, **10**, 142.
5. Sharan,R., Ulitsky,I. and Shamir,R. (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.
6. Kourmpetis,Y.A.I., van Dijk,A.D., Bink,M.C., van Ham,R.C. and ter Braak,C.J. (2010) Bayesian markov random field analysis for protein function prediction based on network data. *PLoS One*, **5**, e9293.
7. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
8. Tzfadia,O., Amar,D., Bradbury,L.M., Wurtzel,E.T. and Shamir,R. (2012) The MORPH algorithm: ranking candidate genes for membership in *Arabidopsis* and tomato pathways. *Plant Cell*, **24**, 4389–4406.
9. Piro,R.M. and Di Cunto,F. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.
10. Boone,C. (2007) Global mapping of the yeast genetic interaction network. *FEBS J.*, **274**, 342–342.
11. Tong,A.H.Y., Lesage,G., Bader,G.D., Ding,H.M., Xu,H., Xin,X.F., Young,J., Berriz,G.F., Brost,R.L., Chang,M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
12. de la Fuente,A. (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet.*, **26**, 326–333.
13. Mentzen,W.I., Floris,M. and de la Fuente,A. (2009) Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumor. *BMC Genomics*, **10**, 601.
14. Tesson,B.M., Breitling,R. and Jansen,R.C. (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, **11**, 497.
15. Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.
16. Ulitsky,I. and Shamir,R. (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, **25**, 1158–1164.
17. Ulitsky,I., Shlomi,T., Kupiec,M. and Shamir,R. (2008) From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. *Mol. Syst. Biol.*, **4**, 209.
18. Ulitsky,I. and Shamir,R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.*, **1**, 8.
19. Narayanan,M., Vetta,A., Schadt,E.E. and Zhu,J. (2010) Simultaneous clustering of multiple gene expression and physical interaction datasets. *PLoS Comput. Biol.*, **6**, e1000742.
20. Kelley,D.R. and Kingsford,C. (2011) Extracting between-pathway models from E-MAP interactions using expected graph compression. *J. Comput. Biol.*, **18**, 379–390.
21. Bandyopadhyay,S., Mehta,M., Kuo,D., Sung,M.K., Chuang,R., Jaehnig,E.J., Bodenmiller,B., Licon,K., Copeland,W., Shales,M. *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science*, **330**, 1385–1389.
22. Leiserson,M.D.M., Tatar,D., Cowen,L.J. and Hescott,B.J. (2011) Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. *J. Comput. Biol.*, **18**, 1399–1409.
23. Ma,X.T., Tarone,A.M. and Li,W.Y. (2008) Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS One*, **3**, e1922.
24. Amar,D., Safer,H. and Shamir,R. (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.*, **9**, e1002955.
25. Defays,D. (1977) Efficient algorithm for a complete link method. *Compu. J.*, **20**, 364–366.
26. Li,J.Y., Li,H.Q., Soh,D. and Wong,L. (2005) A correspondence between maximal complete bipartite subgraphs and closed patterns. *Lect. Notes Artif. Int.*, **3721**, 146–156.
27. Li,J.Y., Liu,G.M., Li,H.Q. and Wong,L. (2007) Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix:

a one-to-one correspondence and mining algorithms. *IEEE Trans. Knowl. Data Eng.*, **19**, 1625–1637.
28. Hedges,L.V. and Olkin,I. (1985) *Statistical Methods for Meta-Analysis*. Academic Press, OR.
29. Schmid,J.E., Koch,G.G. and LaVange,L.M. (1991) An overview of statistical issues and methods of meta-analysis. *J. Biopharm. Stat.*, **1**, 103–120.
30. Chatr-aryamontri,A., Breitkreutz,B.J., Heinicke,S., Boucher,L., Winter,A., Stark,C., Nixon,J., Ramage,L., Kolas,N., O'Donnell,L. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.
31. Guenole,A., Srivas,R., Vreeken,K., Wang,Z.Z., Wang,S.Y., Krogan,N.J., Ideker,T. and van Attikum,H. (2013) Dissection of DNA damage responses using multiconditional genetic interaction maps. *Mol. Cell*, **49**, 346–358.
32. Shamir,R., Maron-Katz,A., Tanay,A., Linhart,C., Steinfeld,I., Sharan,R., Shiloh,Y. and Elkon,R. (2005) EXPANDER - An integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**, 232.
33. Ulitsky,I., Laurent,L.C. and Shamir,R. (2010) Towards computational prediction of microRNA function and activity. *Nucleic Acids Res.*, **38**, e160.
34. Ulitsky,I., Maron-Katz,A., Shavit,S., Sagir,D., Linhart,C., Elkon,R., Tanay,A., Sharan,R., Shiloh,Y. and Shamir,R. (2010) Expander: from expression microarrays to networks and functions. *Nat. Protoc.*, **5**, 303–322.
35. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
36. Gallant,A., Leiserson,M.D., Kachalov,M., Cowen,L.J. and Hescott,B.J. (2013) Genecentric: a package to uncover graph-theoretic structure in high-throughput epistasis data. *BMC Bioinformatics*, **14**, 23.
37. Collins,S.R., Miller,K.M., Maas,N.L., Roguev,A., Fillingham,J., Chu,C.S., Schuldiner,M., Gebbia,M., Recht,J., Shales,M. *et al.* (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**, 806–810.
38. Baryshnikova,A., Costanzo,M., Kim,Y., Ding,H.M., Koh,J., Toufighi,K., Youn,J.Y., Ou,J.W., San Luis,B.J., Bandyopadhyay,S. *et al.* (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods*, **7**, 1017–1024.
39. Dahlmann,B. (2005) Proteasomes. *Essays Biochem.*, **41**, 31–48.
40. Lundin,C., North,M., Erixon,K., Walters,K., Jenssen,D., Goldman,A.S.H. and Helleday,T. (2005) Methyl methanesulfonate (MMS) produces heat-labile DNA damage but no detectable *in vivo* DNA double-strand breaks. *Nucleic Acids Res.*, **33**, 3799–3811.
41. Parker,J.L. and Ulrich,H.D. (2012) A SUMO-interacting motif activates budding yeast ubiquitin ligase Rad18 towards SUMO-modified PCNA. *Nucleic Acids Res.*, **40**, 11380–11388.
42. Blastyak,A., Pinter,L., Unk,I., Prakash,L., Prakash,S. and Haracska,L. (2007) Yeast Rad5 protein required for postreplication repair has a DNA helicase activity specific for replication fork regression. *Mol. Cell*, **28**, 167–175.
43. Brusky,J., Zhu,Y. and Xiao,W. (2000) UBC13, a DNA-damage-inducible gene, is a member of the error-free postreplication repair pathway in *Saccharomyces cerevisiae*. *Curr. Genet.*, **37**, 168–174.
44. Keogh,M.C., Kim,J.A., Downey,M., Fillingham,J., Chowdhury,D., Harrison,J.C., Onishi,M., Datta,N., Galicia,S., Emili,A. *et al.* (2006) A phosphatase complex that dephosphorylates gamma H2AX regulates DNA damage checkpoint recovery. *Nature*, **439**, 497–501.
45. Yeung,M. and Durocher,D. (2011) Srs2 enables checkpoint recovery by promoting disassembly of DNA damage foci from chromatin. *DNA Repair*, **10**, 1213–1222.
46. Mimura,S., Yamaguchi,T., Ishii,S., Noro,E., Katsura,T., Obuse,C. and Kamura,T. (2010) Cul8/Rtt101 forms a variety of protein complexes that regulate DNA damage response and transcriptional silencing. *J. Biol. Chem.*, **285**, 9858–9867.
47. Vaisica,J.A., Baryshnikova,A., Costanzo,M., Boone,C. and Brown,G.W. (2011) Mms1 and Mms22 stabilize the replisome during replication stress. *Mol. Biol. Cell*, **22**, 2396–2408.
48. Mortensen,U.H., Bendixen,C., Sunjevaric,I. and Rothstein,R. (1996) DNA strand annealing is promoted by the yeast Rad52 protein. *Proc. Natl Acad. Sci. USA*, **93**, 10729–10734.
49. Araki,Y., Takahashi,S., Kobayashi,T., Kajiho,H., Hoshino,S. and Katada,T. (2001) Ski7p G protein interacts with the exosome and the Ski complex for 3′-to-5′ mRNA decay in yeast. *EMBO J.*, **20**, 4684–4693.
50. Anderson,J.S.J. and Parker,R. (1998) The 3′ to 5′ degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3′ to 5′ exonucleases of the exosome complex. *EMBO J*, **17**, 1497–1506.
51. Azzalin,C.M. and Lingner,J. (2006) The double life of UPF1 in RNA and DNA stability pathways. *Cell Cycle*, **5**, 1496–1498.
52. Arora,C., Kee,K., Maleki,S. and Keeney,S. (2004) Antiviral protein Ski8 is a direct partner of Spo11 in meiotic DNA break formation, independent of its cytoplasmic role in RNA metabolism. *Mol. Cell*, **13**, 549–559.
53. Hieronymus,H., Yu,M.C. and Silver,P.A. (2004) Genome-wide mRNA surveillance is coupled to mRNA export. *Genes Dev.*, **18**, 2652–2662.
54. Myers,A.J., Webster,J.A., Gibbs,J.R., Clarke,J., Ray,M., Zhang,W.X., Holmans,P., Rohrer,K., Zhao,A., Marlowe,L. *et al.* (2009) Genetic control of human brain transcript expression in Alzheimer Disease. *Am. J. Hum. Genet.*, **84**, 445–458.
55. Montojo,J., Zuberi,K., Rodriguez,H., Kazi,F., Wright,G., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, **26**, 2927–2928.
56. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
57. Ben-Hur,A., Elisseeff,A. and Guyon,I. (2002) A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomput.*, **2002**, 6–17.
58. Chia,B.K. and Karuturi,R.K. (2010) Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms Mol. Biol.*, **5**, 23.
59. Dembele,D. and Kastner,P. (2003) Fuzzy C-means method for clustering microarray data. *Bioinformatics*, **19**, 973–980.
60. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
61. Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, RESEARCH0059.
62. McLachlan,G. (1998) Mathematical classification and clustering. *Psychometrika*, **63**, 93–95.
63. Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.
64. Sharan,R., Maron-Katz,A. and Shamir,R. (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.
65. Van Dongen,S. (2000) Graph clustering by flow simulation. Ph.D Thesis. University of Utrecht.
66. Vlasblom,J. and Wodak,S.J. (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, **10**, 99.
67. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 24–45.
68. Tanay,A., Sharan,R., Kupiec,M. and Shamir,R. (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.
69. Bandyopadhyay,S., Kelley,R., Krogan,N.J. and Ideker,T. (2008) Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput. Biol.*, **4**, e1000065.
70. Watson,M. (2006) CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, **7**, 509.

3. **A hierarchical Bayesian model for flexible module discovery in three-way time-series data**.

D. Amar, D. Yekutieli, A. Maron-Katz, T. Hendler and R. Shamir.

Bioinformatics, 31 (12): i17-i26

ISMB/ECCB 2015, proceedings paper, doi: 10.1093/bioinformatics/btv228, 2015

Bioinformatics, 31, 2015, i17–i26
doi: 10.1093/bioinformatics/btv228
ISMB/ECCB 2015

OXFORD

A hierarchical Bayesian model for flexible module discovery in three-way time-series data

David Amar[1,*], Daniel Yekutieli[2], Adi Maron-Katz[1,3,4], Talma Hendler[3,4] and Ron Shamir[1,*]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel, [2]Department of Statistics and OR, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [3]Functional Brain Center, Wohl Institute for Advanced Imaging, Tel Aviv Sourasky Medical Center, Tel Aviv 64239, Israel and [4]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

*To whom correspondence should be addressed.

Abstract

Motivation: Detecting modules of co-ordinated activity is fundamental in the analysis of large biological studies. For two-dimensional data (e.g. genes × patients), this is often done via clustering or biclustering. More recently, studies monitoring patients over time have added another dimension. Analysis is much more challenging in this case, especially when time measurements are not synchronized. New methods that can analyze three-way data are thus needed.
Results: We present a new algorithm for finding coherent and flexible modules in three-way data. Our method can identify both core modules that appear in multiple patients and patient-specific augmentations of these core modules that contain additional genes. Our algorithm is based on a hierarchical Bayesian data model and Gibbs sampling. The algorithm outperforms extant methods on simulated and on real data. The method successfully dissected key components of septic shock response from time series measurements of gene expression. Detected patient-specific module augmentations were informative for disease outcome. In analyzing brain functional magnetic resonance imaging time series of subjects at rest, it detected the pertinent brain regions involved.
Availability and implementation: R code and data are available at http://acgt.cs.tau.ac.il/twigs/.
Contact: rshamir@tau.ac.il
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identifying modules of elements acting in concert is a fundamental paradigm in interpreting, visualizing and dissecting complex biomedical data. For two-dimensional data (e.g. genes versus conditions), clustering is the simplest way to group the elements of one dimension (Hartigan, 1972). Biclustering seeks row and column subsets that manifest similarity (Cheng and Church, 2000; Hartigan, 1972; Madeira and Oliveira, 2004). Such analysis has become standard in computational biology (Mitra *et al.*, 2013; Oghabian *et al.*, 2014). Algorithms for finding biclusters differ in how they define (and identify) biclusters (Madeira and Oliveira, 2004). For example, biclusters were defined as sub-matrices with constant values (Hartigan, 1972), row or column additive or multiplicative values (Lazzeroni and Owen, 2002) and submatrices with order preserving values (Ben-Dor *et al.*, 2003).

Recent studies have extended the idea of biclustering to more complex input structures beyond the standard row–column data (Mitra *et al.*, 2013). Meng *et al.* (2009) extended the classic Iterative Signature Algorithm (ISA) (Bergmann *et al.*, 2003) to analyze a single matrix of time series data together with prior knowledge on gene function to detect temporal transcription modules that are biologically meaningful. Li and Tuck (2009) introduced an algorithm for joint analysis of ChIP-chip and gene expression data to find biclusters that are likely to be regulated by similar transcription factors. Waltman *et al.* (2010) and Dede and Ogul (2013) proposed three-way clustering of gene-condition-organism data. The algorithm of Waltman *et al.* (2010) uses sequence information to integrate data across species, and a post-processing step allows detection of species-specific information. Gerber *et al.* (2007) cluster tissues hierarchically and then find the representative gene set of each tissue cluster in the hierarchy.

A common data source that calls for three-way analysis is a collection of gene expression profiles measured for a set of subjects over a series of time points. Hence, the data are represented by a

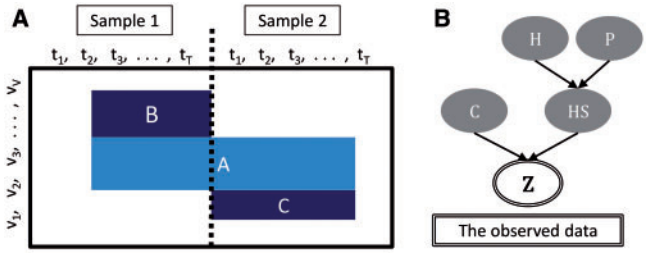


Fig. 1. Overview of the model. (**A**) A toy example of a core module (A) and its private modules (**B**, **C**). (B) An overview of the dependencies in the hierarchical model. P is the vector of subject-specific probabilities P_s

gene × subject × time 3D matrix (i.e. a tensor of order 3) (Mankad and Michailidis, 2014; Zhao and Zaki, 2005). For such matrices, Supper *et al.* (2007) presented EDISA, an extension of ISA that handles a time-course vector for each gene–subject pair instead of a single scalar. Extant models are limited in their ability to detect a signal that is specific to a particular subject. For example, the set of genes active under one subject in a module may only partially overlap with the gene set of other subjects. Another limitation is the assumption of synchronicity of time points across subjects. Although this assumption is valid for technical repeats or well-tailored experiments, it is less plausible in other situations, e.g. samples taken from patients over time, due to possible heterogeneity in the response of different patients.

Here, we introduce a new, flexible definition of a module suitable for three-way data where subjects have entities (e.g. genes) measured over time, but time courses are unsynchronized among the subjects. A *core module* is defined by a subset of the subjects and a subset of the entities, along with subject-specific subset of the time points. In addition, subjects may have *private modules* that only partially coincide with the core set of entities. The assumption is that the resulting submatrices will show values markedly different from the whole matrix. A toy example is shown in Figure 1A.

We developed a statistical framework and algorithm for analyzing such data. Our framework can detect core modules and for each subject in a core module, a private module with relevant time points. We developed a hierarchical Bayesian generative model for the data and a procedure that aims to fit model parameters for a given dataset. Our algorithm uses a regular biclustering solution as a starting point and then performs iterative improvement using a Gibbs sampling procedure. The algorithm is called TWIGS (three-way module inference via Gibbs sampling). In simulations, we show that TWIGS outperforms standard algorithms even when the core modules have no additional subject-specific signal. When subject-specific signals exist, the ability of extant algorithms to detect the core modules declines markedly, whereas the performance of TWIGS remains high.

We demonstrate the advantage of our framework on experimental data from two different domains: gene expression and brain functional magnetic resonance imaging (fMRI) signals. We first analyzed whole blood expression profiles, taken daily for 5 days from 14 patients after septic shock (Parnell *et al.*, 2013). TWIGS detected two core modules of up-regulated genes, showing enrichment for different immune system processes. The first was related to response to bacteria, whereas the second was related to regulation of T-cells. Analysis of the subject-specific private modules revealed multiple enrichments that illustrate patient-specific-activated biological processes. Hence, our analysis produced both shared and subject-specific insights, highlighting biological pathways that repeatedly emerge as up-regulated after septic shock, together with additional biological functions particular to each patient. We also analyzed fMRI readings for 20 subjects at rest (Vaisvaser *et al.*, 2013). The data for each subject are a matrix of 464 brain regions (parcels) measured over 94 time points at 3 s intervals. Each value in the matrix is the parcel's average blood-oxygen-level-dependent (BOLD) contrast. These levels are indicators of the activity at that region. TWIGS revealed several core modules of highly activated bi-lateral brain regions. Reassuringly, the detected modules were enriched with regions that are known to be active during rest. This analysis shows that our framework is able to detect large functional networks that reappear as activated across subjects and also highlight subject-specific activation patterns.

2 Methods

2.1 The probabilistic model

The input for our problem is summarized as a 3D matrix $\mathbf{Z}$ where $Z_{v,t,s}$ is the activity level of the measured object $v \in 1,\ldots,V$, at time $t \in 1,\ldots,T$, for subject $s \in 1,\ldots,S$. We will say that v and t represent the rows and columns of the matrix and s represents layers. In gene expression data, v represents genes, whereas in fMRI, data v represents brain regions (parcels or voxels). For uniformity, from now on we use for v the term row or voxel. Here, we describe a hierarchical probability model for generating a single module from the distribution of $\mathbf{Z}$.

We assume that there is a set of voxels $\mathcal{V}\{1,\ldots,V\}$ that tend to have high values jointly in a subset of the subjects. $\mathcal{V}$ is specified by the indicator vector $\mathbf{H} = (H_1,\ldots,H_V)$, through the relation $\mathcal{V} = \{v : H_v = 1\}$. We assume that $H_v \sim \text{Bernoulli}(\pi^{VC})$. Although H marks the rows of the core module, the signal in each specific subject might change. The subject-specific voxel sets are specified by the matrix $\mathbf{HS} = \{HS_{1,1},\cdots,HS_{V,S}\}$, where $HS_{v,s} = 1$ specifies that voxel v participates in the module of subject s. The relation between $\mathbf{H}$ and $\mathbf{HS}$ is as follows: if $H_v = 1$ then $HS_{v,s} \sim \text{Bernoulli}(p_s)$, otherwise $HS_{v,s} \sim \text{Bernoulli}(p_0)$.

We next model the time-series relations. $\mathbf{C} = C_{t,s} \in \{0,1\}$ indicates whether the voxel set of subject s is active at time t. We assume $\Pr(C_{1,s} = 1) = \pi^{1,1}$. The activity at time $t = 2,\ldots,T$, depends on the time window of size $w \geq 1$ before t. In times $t = 2,\ldots,w$, the time window is $1\cdots t-1$. Let $C'_{t,s} = 1$, if the time window of subject s right before time point t contains at least one active time point and set $C'_{t,s} = 0$ otherwise. We assume that $\Pr(C_{t,s} = 1|C'_{t,s} = 0) = \pi^{1|0}$ and $\Pr(C_{t,s} = 1|C'_{t,s} = 1) = \pi^{1|1}$.

Finally, we assume that for v, t, s for which $C_{t,s} = 1$ and $HS_{v,s} = 1$, $Z_{v,t,s} \sim F_1$, otherwise $Z_{v,t,s} \sim F_0$. An overview of the model hierarchy is shown in Figure 1B. We assume that all hyper-parameters above have Beta prior distributions: $\pi^{VC} \sim \text{Beta}(a_1,b_1)$, $p_s \sim \text{Beta}(a_2,b_2)$, $p_0 \sim \text{Beta}(a_3,b_3)$, $\pi^{1,1} \sim \text{Beta}(a_4,b_4)$, $\pi^{1|1} \sim \text{Beta}(a_5,b_5)$, $\pi^{1|0} \sim \text{Beta}(a_6,b_6)$.

2.2 The Gibbs sampling algorithm

Our algorithm starts from a solution produced using a standard biclustering algorithm and then applies iterative improvement steps. In each step, all parameters are fixed except a single one that is sampled according to its conditional probability. The order of parameters matches the subsections below. This order is repeated cyclically k times. The output of the process is the set of sampled values for each parameter in all iterations. We then extract the core modules and the subject-specific modules from this output.

(1) As H_v are Bernoulli realizations with success probability π^{VC}:

$$\pi^{VC}|\cdots \sim \text{Beta}(a_1 + |\{v : H_v = 1\}|, b_1 + |\{v : H_v = 0\}|).$$

Similarly:

$$\begin{aligned}
p_s|\cdots &\sim \text{Beta}(a_2 + |\{v : HS_{v,s} = 1, H_v = 1\}|,\\
&\quad b_2 + |\{v : HS_{v,s} = 0, H_v = 1\}|).\\
p_0|\cdots &\sim \text{Beta}(a_3 + |\{v, s : HS_{v,s} = 1, H_v = 0\}|,\\
&\quad b_3 + |\{v, s : HS_{v,s} = 0, H_v = 0\}|),\\
\pi^{1,1}|\cdots &\sim \text{Beta}(a_4 + |\{s : C_{1,s} = 1\}|,\\
&\quad b_4 + |\{s : C_{1,s} = 0\}|),\\
\pi^{1|1}|\cdots &\sim \text{Beta}(a_5 + |\{t, s : C'_{t,s} = 1, C_{t,s} = 1\}|,\\
&\quad b_5 + |\{t, s : C'_{t,s} = 1, C_{t,s} = 0\}|),\\
\pi^{1|0}|\cdots &\sim \text{Beta}(a_6 + |\{t, s : C'_{t,s} = 0, C_{t,s} = 1\}|,\\
&\quad b_6 + |\{t, s : C'_{t,s} = 0, C_{t,s} = 0\}|)
\end{aligned}$$

(2) H_v is affected by π^{VC}, and it affects the values of $HS_{v,s}$ for each s:

$$\begin{aligned}
&\Pr(H_v = 1, HS_{v,s} = 1|\cdots) = \pi^{VC} \cdot p_s,\\
&\Pr(H_v = 1, HS_{v,s} = 0|\cdots) = \pi^{VC} \cdot (1 - p_s),\\
&\Pr(H_v = 0, HS_{v,s} = 1|\cdots) = (1 - \pi^{VC}) \cdot p_0,\\
&\Pr(H_v = 0, HS_{v,s} = 0|\cdots) = (1 - \pi^{VC}) \cdot (1 - p_0).
\end{aligned}$$

Thus,

$$\begin{aligned}
\Pr(H_v = 1, HS_{v,*}|\cdots) &= \pi^{VC} \cdot p_s^{\sum_s HS_{v,s}}\\
&\quad \cdot (1 - p_s)^{\sum_s (1 - HS_{v,s})},\\
\Pr(H_v = 0, HS_{v,*}|\cdots) &= (1 - \pi^{VC}) \cdot p_0^{\sum_s HS_{v,s}}\\
&\quad \cdot (1 - p_0)^{\sum_s (1 - HS_{v,s})}.
\end{aligned}$$

Therefore, the conditional posterior of H_v is:

$$\Pr(H_v = 1|\cdots) = \frac{\Pr(H_v = 1, HS_{v,*}|\cdots)}{\Pr(H_v = 1, HS_{v,*}|\cdots) + \Pr(H_v = 0, HS_{v,*}|\cdots)}$$

(3) Given $C_{t,s}$, only the value of $HS_{v,s}$ affects the distribution of $Z_{v,t,s}$. Assume for now, that we condition on $H_v = 1$, then:

$$\begin{aligned}
&\Pr(Z_{v,t,s}, HS_{v,s} = 0|C_{t,s} = 0, \cdots) = (1 - p_s) f_0(Z_{v,t,s}),\\
&\Pr(Z_{v,t,s}, HS_{v,s} = 1|C_{t,s} = 0, \cdots) = p_s \cdot f_0(Z_{v,t,s}),\\
&\Pr(Z_{v,t,s}, HS_{v,s} = 0|C_{t,s} = 1, \cdots) = (1 - p_s) \cdot f_0(Z_{v,t,s}),\\
&\Pr(Z_{v,t,s}, HS_{v,s} = 1|C_{t,s} = 1, \cdots) = p_s \cdot f_1(Z_{v,t,s}).
\end{aligned}$$

From the above, it is clear that when $C_{t,s} = 0$:

$$\begin{aligned}
&\Pr(HS_{v,s} = 1|C_{t,s} = 0, H_v = 1, Z_{v,t,s}, \cdots) = p_s,\\
&\Pr(HS_{v,s} = 0|C_{t,s} = 0, H_v = 1, Z_{v,t,s}, \cdots) = 1 - p_s.
\end{aligned}$$

Therefore, a time point t in which $C_{t,s} = 0$ will not affect the marginal distribution of $HS_{v,s}$. Let $T'_s = \{t : C_{t,s} = 1\}$, then:

$$\begin{aligned}
&\Pr(HS_{v,s} = 1, Z_{v,*,s}|H_v = 1, \cdots)\\
&\quad = p_s \cdot \textstyle\prod_{t \in T'_s} f_1(Z_{v,t,s}),\\
&\Pr(HS_{v,s} = 0, Z_{v,*,s}|H_v = 1, \cdots)\\
&\quad = (1 - p_s) \cdot \textstyle\prod_{t \in T'_s} f_0(Z_{v,t,s}).
\end{aligned}$$

On the basis of the equations above, we can calculate the conditional posterior of $HS_{v,s}$, given that $H_v = 1$, through:

$$\begin{aligned}
&\Pr(HS_{v,s} = 1|H_v = 1, \cdots)\\
&\quad = \frac{\Pr(HS_{v,s} = 1, Z_{v,*,s}|H_v = 1, \cdots)}{\Pr(HS_{v,s} = 0 \vee HS_{v,s} = 1, Z_{v,*,s}|H_v = 1, \cdots)}
\end{aligned}$$

Similarly, the conditional posterior of $HS_{v,s}$ given that $H_v = 0$ can be calculated by replacing every p_s with p_0 in the formulas above.

(4) As the value of $C_{1,s}$ affects the value of the time window $C_{2,s}, \ldots, C_{w+1,s}$ and the values of $Z_{v,1,s}$ with $HS_{v,s} = 1$:

$$\begin{aligned}
&\Pr(C_{1,s} = 1, C_{2,s}, \cdots, C_{w+1,s}, Z_{*,1,s}|\cdots)\\
&\quad = \pi^{1,1} \cdot \pi^{1|1 \sum_{k=2}^{w+1} C_{k,s}} \cdot (1 - \pi^{1|1})^{\sum_{k=2}^{w+1} (1 - C_{k,s})} \cdot \prod_{v:HS_{v,s}=1} f_1(Z_{v,1,s})
\end{aligned}$$

Unlike the equation above, calculating the probability of $C_{1,s} = 0$ requires breaking the window into two parts. Assume that the time window contains at least one active cell. Let l be the first time point of $C_{2,s}, \ldots, C_{w+1,s}$ that changes from 0 to 1. Thus:

$$\begin{aligned}
&\Pr(C_{1,s} = 0, C_{2,s}, \cdots, C_{w+1,s}, Z_{*,1,s}|\cdots)\\
&\quad = (1 - \pi^{1,1}) \cdot \left[\prod_{k=2}^{l-1} (1 - \pi^{1|0})\right] \cdot \pi^{1|0} \cdot \pi^{1|1 \sum_{k=2}^{w+1} C_{k,s}} \cdot\\
&\quad (1 - \pi^{1|1})^{\sum_{k=2}^{w+1} (1 - C_{k,s})} \cdot \prod_{v:HS_{v,s}=1} f_0(Z_{v,1,s})
\end{aligned}$$

If there are no active cells in $C_{2,s}, \cdots, C_{w+1,s}$, then the calculation reduces to:

$$\begin{aligned}
&\Pr(C_{1,s} = 0, C_{2,s}, \cdots, C_{w+1,s}, Z_{*,1,s}|\cdots)\\
&\quad = (1 - \pi^{1,1}) \cdot (1 - \pi^{1|0})^w \cdot \prod_{v:HS_{v,s}=1} f_0(Z_{v,1,s})
\end{aligned}$$

Finally, the conditional of $C_{1,s}$ can be calculated by:

$$\begin{aligned}
&\Pr(C_{1,s} = 1|C_{2,s}, \cdots, C_{w+1,s}, Z_{*,1,s}, \cdots)\\
&\quad = \frac{\Pr(C_{1,s} = 1, C_{2,s}, \cdots, C_{w+1,s}, Z_{*,1,s}|\cdots)}{\Pr(C_{2,s}, \cdots, C_{w+1,s}, Z_{*,1,s}|\cdots)}
\end{aligned}$$

The conditional probability of the event $C_{1,s} = 0$ is computed in the same way.

(5) For $t \in 2, \ldots, T-1$, the value of $C_{t,s}$ is affected by the value of $C'_{t,s}$, and it affects the value of $C_{t,s}, \ldots, C_{\min(w+t,T),s}$, and the values of $Z_{v,t,s}$ with $HS_{v,s} = 1$. Thus:

$$\begin{aligned}
&\Pr(C_{t,s} = 1, C_{*,s}, Z_{*,t,s}|\cdots)\\
&\quad = \pi^{1|1 C'_{t,s}} \cdot \pi^{1|0 1 - C'_{t,s}} \cdot \pi^{1|1 \sum_{k=t+1}^{\min(w+t,T)} C_{k,s}}\\
&\quad \cdot (1 - \pi^{1|1})^{\sum_{k=t+1}^{\min(w+t,T)} (1 - C_{k,s})} \cdot \prod_{v:HS_{v,s}=1} f_1(Z_{v,t,s})
\end{aligned}$$

$$\begin{aligned}\Pr(C_{t,s}=0, C_{*,s}, Z_{*,t,s}|\cdots) &= \prod_{k=t}^{\min(t+w,T)} \pi^{1|1\,C_{k,s}\cdot C'_{k,s}} \cdot \pi^{1|0\,C_{k,s}\cdot(1-C'_{k,s})} \\ &\cdot \prod_{k=t}^{\min(t+w,T)} (1-\pi^{1|1})^{(1-C_{k,s})\cdot C'_{k,s}} \\ &\cdot \prod_{k=t}^{\min(t+w,T)} (1-\pi^{1|0})^{(1-C_{k,s})\cdot(1-C'_{k,s})} \\ &\cdot \prod_{v:HS_{v,s}=1} f_0(Z_{v,t,s})\end{aligned}$$

Thus, $\Pr(C_{t,s}=1|\cdots)$ can be calculated similarly to the calculations in the previous section.

2.3 Setting f_0 and f_1

Here, we discuss two options for setting f_0 and f_1 and their hyperparameters: (i) a *Bernulli-Beta* model for binary data and (ii) a *Normal-Gamma* model for normal distributions. Let A be the cells within the module (including the core and private parts): $A=\{Z_{v,t,s}: C_{t,s}=1 \wedge HS_{v,s}=1\}$. Let B be the cells outside the module: $B=\{Z_{v,t,s}: C_{t,s}=0 \vee HS_{v,s}=0\}$

For binary data, we assume that $Z_{v,t,s}\in\{0,1\}$, $f_0=\text{Bernoulli}(\pi^{C0})$ and $f_1=\text{Bernoulli}(\pi^{C1})$. Thus, our model learns the background probability π^{C0} of observing a value of 1 and the probability π^{C1} of observing 1 within the module. In this model, π^{C0} and π^{C1} follow Beta posterior distributions:

$$\begin{aligned}\pi^{C0}|\cdots &\sim \text{Beta}(a_7+|\{i:B_i=1\}|, b_7+|\{i:B_i=0\}|) \\ \pi^{C1}|\cdots &\sim \text{Beta}(a_7+|\{i:A_i=1\}|, b_7+|\{i:A_i=0\}|)\end{aligned}$$

In the continuous case, we assume that f_0 is $N(\mu_0,\sigma_0)$ and f_1 is $N(\mu_1,\sigma_1)$. Under the *Normal-Gamma* model, the prior distribution for the mean μ and the standard deviation σ of a normal distribution $N(\mu,\sigma)$ is:

$$\begin{aligned}1/\sigma &\sim \text{Gamma}\left(\frac{v_0}{2},\frac{SS_0}{2}\right) \\ \mu &\sim N\left(m_0,\frac{1}{p_0}\right)\end{aligned}$$

The conditional posteriors for $Y=(y_1,\ldots,y_n)$ where for each i, $y_i\sim N(\mu,\sigma)$ are:

$$\begin{aligned}1/\sigma|Y &\sim \text{Gamma}(1/2\cdot(v_0+n), \\ &\quad 1/2\cdot\left(SS_0+\sum_{i=1}^{n}(y_i-\bar{y})^2+\frac{n\cdot p_0}{n+p_0}\cdot(\bar{y}-m_0)^2\right)\right) \\ \mu|Y,\sigma &\sim N\left(\frac{m_0p_0+n\bar{y}}{n+p_0},\frac{\sigma}{p_0+n}\right)\end{aligned}$$

Thus, we apply the model above for A and B, thereby modeling f_0 and f_1 as normal distributions.

2.4 Finding multiple modules

To find a single module, we use a standard biclustering algorithm to produce an initial solution and then use the Gibbs sampler to improve it. The biclustering algorithm is applied on a 2D matrix M obtained by concatenating the layers in Z, i.e. $M_{v,i}=Z_{v,t,s}$, where $i=(s-1)S+t$. In this study, we tested Bimax (Prelic *et al.*, 2006) and ISA (Bergmann *et al.*, 2003) as the base algorithms. To binarize real-valued data Z to run Bimax, we use a threshold τ: we set every value $Z_{v,t,s}\geq\tau$ ($<\tau$) to 1 (0). By default, we set τ to be the 0.9 quantile of the values in Z. After running the Gibbs sampler, we take the mode of H, HS and C as the solution. By default, all hyperparameters of the Gibbs sampler are set to non-informative priors. This means that the algorithm infers these parameters and thus no tunning is required.

To find multiple modules, we tested two previously used heuristics (Serin and Vingron, 2011; Shabalin *et al.*, 2009). In the first (Cheng and Church, 2000), which we call *masker*, we run the algorithm iteratively on the residual matrix of Z. The residual matrix is calculated by going over all cells in the module and updating their values in Z. In the binary model, the update rule is to change all module cells to zero. In the normal model, we subtract the mean of f_1 from the value of each cell.

The second heuristic, called *filter*, takes a set of biclusters U as input and produces a reduced set. It first uses the *overlap reduction* method of Serin and Vingron (2011): initially $U'=\emptyset$, then the largest module in U is added to U' and all remaining modules with a large overlap with it (we used Jaccard index ≥ 0.5) are removed from U. The process is repeated until U is empty. Next, we run the Gibbs sampler on the original matrix Z starting with each module in U'. The result is a set of new modules U''. Finally, as different modules in U' might converge into similar modules in U'', the overlap removal process is used again, taking U'' as input.

For both heuristics, we define when to add a module to the final output. When using *masker*, we add modules until the first time a module is rejected. A module is accepted if it is large enough and the difference $|f_1-f_0|$ for it is large enough. In the binary case, we set $\pi^{C1}>0.5$ and in the normal case we set $\mu>0.3$. Setting the minimal module size depends on the application and on the size of the input data. By default, we set the minimal size of a core module to 5 rows and 5 time points (combining all subjects).

2.5 Performance measures

In the results below, we compare algorithms on simulated data. In each case, we compare the known H, HS, C to the algorithm output H', HS', C' using the Jaccard coefficient. For example, the Jaccard score of H and H' is:

$$J(H,H')=\frac{|\{i;H_i\wedge H'_i\}|}{|\{i;H_i\vee H'_i\}|}$$

When the data contain more than one module we use the running max average of all pairwise Jaccard scores. Given the known solution $H=(H_1,\ldots,H_{k_1})$ and the algorithm output $H'=(H'_1,\ldots,H'_{k_2})$ the running max average score is:

$$\frac{\sum_{i\in1,\ldots,k_1}\left(\max_{j\in1,\ldots,k_2}J(H_i,H'_j)\right)+\sum_{j\in1,\ldots,k_2}\left(\max_{i\in1,\ldots,k_1}J(H_i,H'_j)\right)}{k_1+k_2}$$

The same method is used for HS and C.

3 Results

3.1 Simulations

Our simulations setup was as follows. We set $V=500$, $T=50$, $S=10$ and create an initial matrix Z in which all values are zero. We then add modules to Z in which all values are 1 and later add noise according to the tested model (binary or normal). To define a new module, we first need to randomly select the rows and columns of each subject s. Time points are selected randomly with $w=1$, $\pi^{1,1}=0.05$, $\pi^{1|1}=0.6$ and $\pi^{1|0}=0.1$. Rows are selected randomly as follows. We first select randomly 20 rows $i_1\ldots i_{20}$ for the

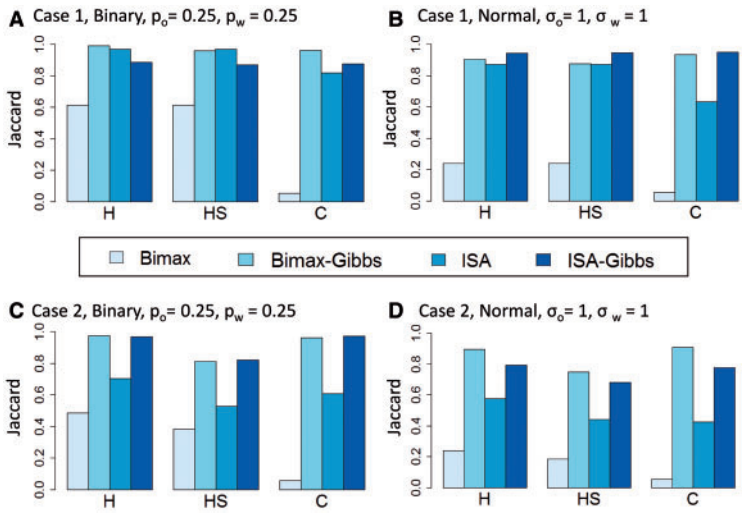


Fig. 2. Simulation results for data with a single module. Each bar represents the average over 10 repeats. (**A**) Case 1: no subject-specific signal. (**B**) Case 2: with subject-specific signals. The Bimax-Gibbs variant was later chosen as the default TWIGS algorithm

core module. Then, we add row r to the private module of subject s with probability p'_s if $r \in \{i_1 \dots i_{20}\}$, otherwise r is added with probability p'_0.

Adding random noise to the data depends on the tested model. In the binary case, we randomly replace $Z_{v,t,s}$ with $1 - Z_{v,t,s}$, with probability p_w if (v, t, s) belongs to the private module of s and with probability p_o otherwise. In the normal model for each (v, t, s) within the private module of s we select $\epsilon_{v,t,s} \sim N(0, \sigma_w)$, otherwise we select $\epsilon_{v,t,s} \sim N(0, \sigma_o)$. We then add the noise by updating $Z_{v,t,s} = Z_{v,t,s} + \epsilon_{v,t,s}$. We tested scenarios of a single module with and without subject-specific signals and of multiple modules.

3.2 Case 1: a single core module

In this test, we set $p'_s = 1$ and $p'_0 = 0$. This case represents the standard biclustering task because there is no subject-specific signal. Thus, biclustering algorithms are expected to achieve high performance.

The results are shown in Figure 2A and B. Each algorithm was tested on 10 instances and the average Jaccard score, which quantifies the agreement between the known solution and the algorithm output, is shown. We set high noise levels both in the binary data (Fig. 2A)—$p'_w = p'_o = 0.25$ and in the normal data (Fig. 2B)—$\sigma'_w = \sigma'_o = 1$. The Bimax algorithm had a low Jaccard score in most cases, since its output covered only a small part of the true bicluster. Although the false-positive rate was very low (<0.01 both for the bicluster rows and columns), the true-positive rate was low as well (<0.25). ISA performed much better, especially in terms of identifying *H* and *HS*. Using TWIGS to improve the solution was beneficial: it was able to keep the high performance of ISA for *H* and *HS* and to considerably improve the score of *C*. It greatly improved the Bimax solution in all criteria. For example, in the normal data, the score of *C* went up from 0.053 to 0.93. The ISA solution improved from 0.63 to 0.95 using TWIGS. Notably, this improvement was achieved with only 50 sampling iterations, which took less than 7 s on average (over simulation repeats). Thus, this boost in performance was achieved at a low cost of running time. We kept this number of iterations also in subsequent analyses. Note, however, that when the data are much larger (e.g. $|T| > 1000$), the running time could increase to several minutes.

We also tested a binary case in which the noise levels were not symmetric: we set $p'_w = 0.5$ and $p'_o = 0.1$. The results are shown in Supplementary Figure 1. In this case, the Bimax–Gibbs combination reached the top performance in all measurements, with very high scores: 0.92 (*H*), 0.86 (*HS*) and 0.93 (*C*). The performance of both ISA and Bimax was low (all scores were <0.7), indicating that standard algorithms have difficulty in such noise levels.

3.3 Case 2: a core module with subject-specific signal

In this test, we set $p'_s = 0.9$ and $p'_0 = 0.01$. Thus, this scenario is different from standard biclustering and triclustering tasks in two ways: (i) not all shared rows are necessarily part of each private module and (ii) each private module is likely to contain additional rows that are not shared among all subjects.

The results (averaged over 10 instances) are shown in Figure 2C and D. The noise levels were $p'_w = p'_o = 0.25$ in the binary data Fig. 2C) and $\sigma'_w = \sigma'_o = 1$ in the normal data (Fig. 2D). Similar to Case 1, Bimax had low scores because it typically covered only a small perfect fraction of the module, whereas ISA reached higher performance. However, the performance of ISA was much lower than in Case 1. For example, in the normal data the score of *H*, which represents the core module rows, dropped from 0.87 in Case 1 to 0.57. This result demonstrates a weakness of standard biclustering algorithms when the data contain subject-specific signal: the algorithms might fail to discover even the shared information. In contrast to ISA and Bimax, TWIGS improved the solution considerably in all measures. For example, the score of *H* and *C* was >0.89 when starting with the Bimax solution.

3.4 Case 3: multiple modules

Here, we tested the performance of TWIGS with *filter* and *masker* on data with five core modules, each with it own subject-specific signals, using as before $p'_s = 0.9$ and $p'_0 = 0.01$. The results are shown in Figure 3. As expected, the results were lower than in the single core module tests. Nonetheless, the results were still high in spite of the high noise levels.

Unlike the previous cases, using *masker* with Bimax as the base algorithm was much better than all other algorithms. For example, in the binary case (Fig. 3A), it reached scores of 0.86 and 0.8 for *H*

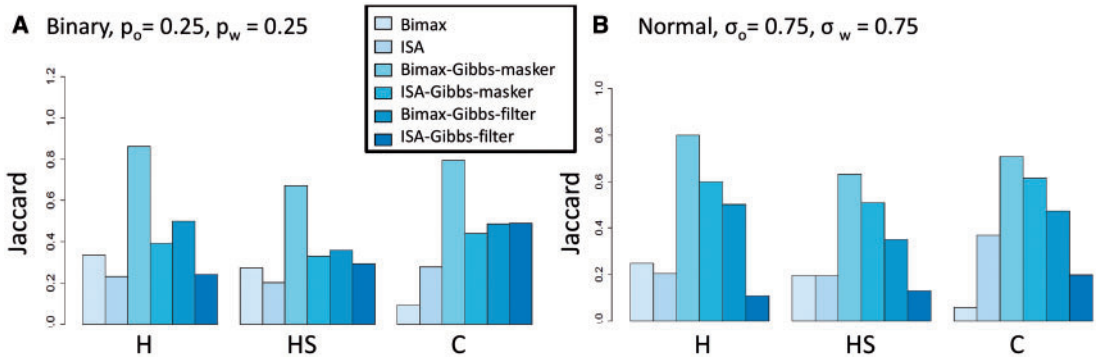


Fig. 3. Simulation results for data with five core modules. Each bar represents the average over 10 repeats. (**A**) Binary data. (**B**) Normal data. The Bimax-Gibbs-masker variant was later chosen as the default TWIGS algorithm

and *C* respectively, where all other algorithms had scores below 0.6. In the normal data, we observed a sharp decrease in performance when setting the noise levels to $\sigma_0 = 1$ as in the previous sections, see Supplementary Figure 2. With a bit lower noise levels of 0.75, the results were similar to the binary case (Fig. 3B). Interestingly, forcing high mean value in f_1 (i.e. by setting $m_0 \geq 1$ and high p_0 constant in the Normal-Gamma model, see Section 2) achieved higher performance scores. For example, setting the mean value to 1.5 improved the score of *H* from 0.8 to 0.9 and the score of *C* from 0.71 to 0.77. We discovered that the non-informative variant had some detrimental instances in which some core modules were grouped together (average number of detected modules was 4.2), whereas enforcing high mean for f_1 detected the correct number of core modules.

On the basis of the above results, from this point on, we used the Bimax-Gibbs-masker as the default variant of the TWIGS algorithm.

3.5 Gene expression data

We tested the performance of TWIGS by analyzing transcriptional response of patients to sepsis. Parnell *et al.* (2013) monitored patients after septic shock. For up to 5 days after sepsis, blood samples were taken daily, and whole blood gene expression was measured using Illumina microarrays. The dataset contained 14 patients for which five profiles, one for each day after sepsis, were available. Our goal was to detect up-regulated biological functions after septic shock. Therefore, for each subject we calculated the log fold change between time points 2, 3, 4, 5 and the first time point. We binarized the data by setting a threshold of 2 for the fold change (i.e. 1 for the log fold change) and ran *masker* with Bimax as the base algorithm. See the Supplementary Text for additional analyses using the non-binarized data and for sensitivity analysis of the binarization threshold. We set the minimal size of the detected core module to 10 rows and 10 columns (number of time points from all patients) and f_1 to have $\pi^{C1} > 0.5$. Using these stop criteria in *masker*, a single small module of 5 genes was detected over 20 repeats in which we independently and randomly shuffled the values of each row in the input matrix.

Two core modules were detected on the real expression matrix. The first covered 11 patients and 53 genes. The second covered seven patients and 62 genes. Four patients were represented in both. Distinct private modules were assigned to each subject in each module. Thus, a total of 20 modules (core or private) were detected in the analysis. GO enrichment analysis [using EXPANDER (Ulitsky *et al.*, 2010)] detected significant enrichment (0.05 FDR) in 19 of the modules. The two detected core modules differed in their enriched biological functions. The first was highly enriched with genes related to killing of cells of other organisms ($P = 2.7E-11$) and response to bacterium ($P = 2.2E-9$).The second core module was enriched with functions that were more specific to T-cell activity (e.g. regulation of T cell activation $P = 4.5E-10$). Thus, TWIGS identified a fuzzy partition of the subjects into two main branches of the immune system and also pointed out the relevant up-regulated genes.

The private modules in the solution were often much larger than the core modules. For example, in the first core module, the private modules of subjects 19 and 24 contained more than 450 genes each. The first core module and the enrichment analysis results of its private modules are shown in Figure 4. See Supplementary Figure 3 for the results of the second core module. Only biological functions that were not significant in the core modules are shown. The figure illustrates how our analysis provides a complementary view to the core modules. That is, although the core modules indicate which biological functions tend to reappear across subjects, the private modules reveal additional enrichments that are sometimes much more specific biologically. For example, the private module of subject 24 was highly enriched with genes related to viral infectious cycle ($P = 1.7E-9$). The network also highlights patients without subject-specific unique enrichments (subjects 30, 46, 49 and 50) and two hubs: subjects 24 and 19. Strikingly, out of the 11 patients covered by this core module, these two patients had much larger private modules and they were the only patients that did not survive the septic shock.

3.6 fMRI data

Vaisvaser *et al.* (2013) collected brain fMRI data from 20 male subjects at rest over 94 time points. In this technique blood flow (BOLD) intensity is measured at every voxel of the brain along time, providing levels of some 100 000 voxels every 2–3 s. The level reflects the activation intensity of the brain voxel. Standard fMRI preprocessing was applied on the raw data as reported in (Vaisvaser *et al.*, 2013). We used a whole brain functional parcellation to transform the data into 517 brain parcels (Craddock *et al.*, 2012). Parcels were masked to include gray matter voxels only using the WFU Pick Atlas Tool (Maldjian *et al.*, 2003; Stamatakis *et al.*, 2010) and 54 parcels that had ≤ 5 gray matter voxels were excluded. For each subject, average BOLD value across all gray matter voxels was calculated within each parcel at each time point. As is standard practice

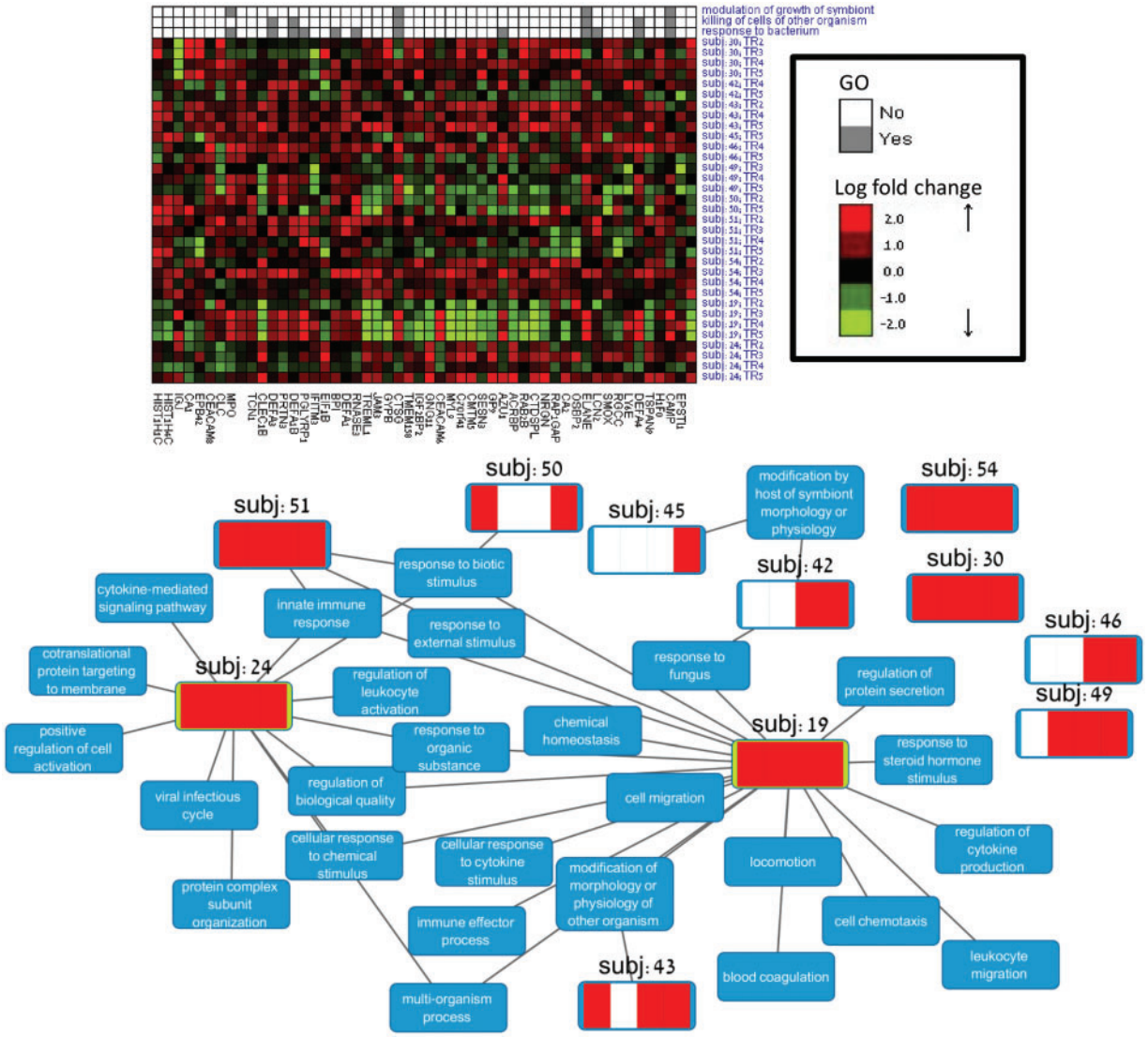


Fig. 4. A module summarizing patient response to sepsis. Top: the first core module heatmap. Bottom: the subject-specific enrichments. The red stripes in each patient's node represent the time points that were covered by its private module. An edge between a subject and a category (blue node) indicates that the subject-specific module was enriched for that category

in fMRI analysis (Birn, 2012), to reduce the effect of physiological artifacts and nuisance variables, the whole-brain mean signal, six motion parameters, cerebrospinal fluid and white matter signals were regressed out of the parcel signals. The result is a matrix M_s for each subject s, in which rows are parcels and columns are time points. We standardized the signal of each row in M_s by subtracting the mean and dividing by the standard deviation. This normalization allows us to find relative changes in the activity of brain regions to highlight temporally activated regions (Rana *et al.*, 2013).

We ran TWIGS with the normal model, Bimax as the initial solution finder and *masker*. With non-informative priors, the algorithm converged to large modules with relatively low mean value (<0.5). As we were interested in highly activated brain regions, we reset the mean of f_1 to a high value: we tested $\mu_1 = 1.5$ and $\mu_1 = 2$. As in the simulations, using such prior improved the results considerably since the non-informative variant tended to merge core modules with high mean value. No module was detected when running the algorithm after randomly and independently shuffling each row of the data matrix (20 repeats).

Unlike in the gene expression analysis, each subject participated in each core module. For $\mu_1 = 2$ (Fig. 5A), four core modules were detected (labeled 1A–4A), with an average of 48.5 parcels. For $\mu_1 = 1.5$ (Fig. 5B), five core modules were detected (labeled 1B–5B), with an average of 66.4 parcels. Out of the five core modules detected using $\mu_1 = 1.5$, four had a parallel core module detected using $\mu_1 = 2$. In addition, modules 1A and 1B maintained similar spatial structure and size and so did 3A and 3B. Modules 2A and 4A were larger than their counterparts.

We evaluated the parcel sets of the identified core modules by comparing them to known functional annotations of the brain (Yeo *et al.*, 2011). The results show that our analysis detected well-known functional modules that are expected to share common activation patterns both during task and at rest. In both solutions, core module 1 was enriched with regions that are involved in visual processing in the occipital lobe of both hemispheres ($q \leq E - 11$) (Belliveau *et al.*, 1991). Core module 2B was enriched with parcels located within the ventral attention network, which is involved in bottom-up orienting of attention ($q \leq 0.02$) (Fox *et al.*, 2006). In both solutions,

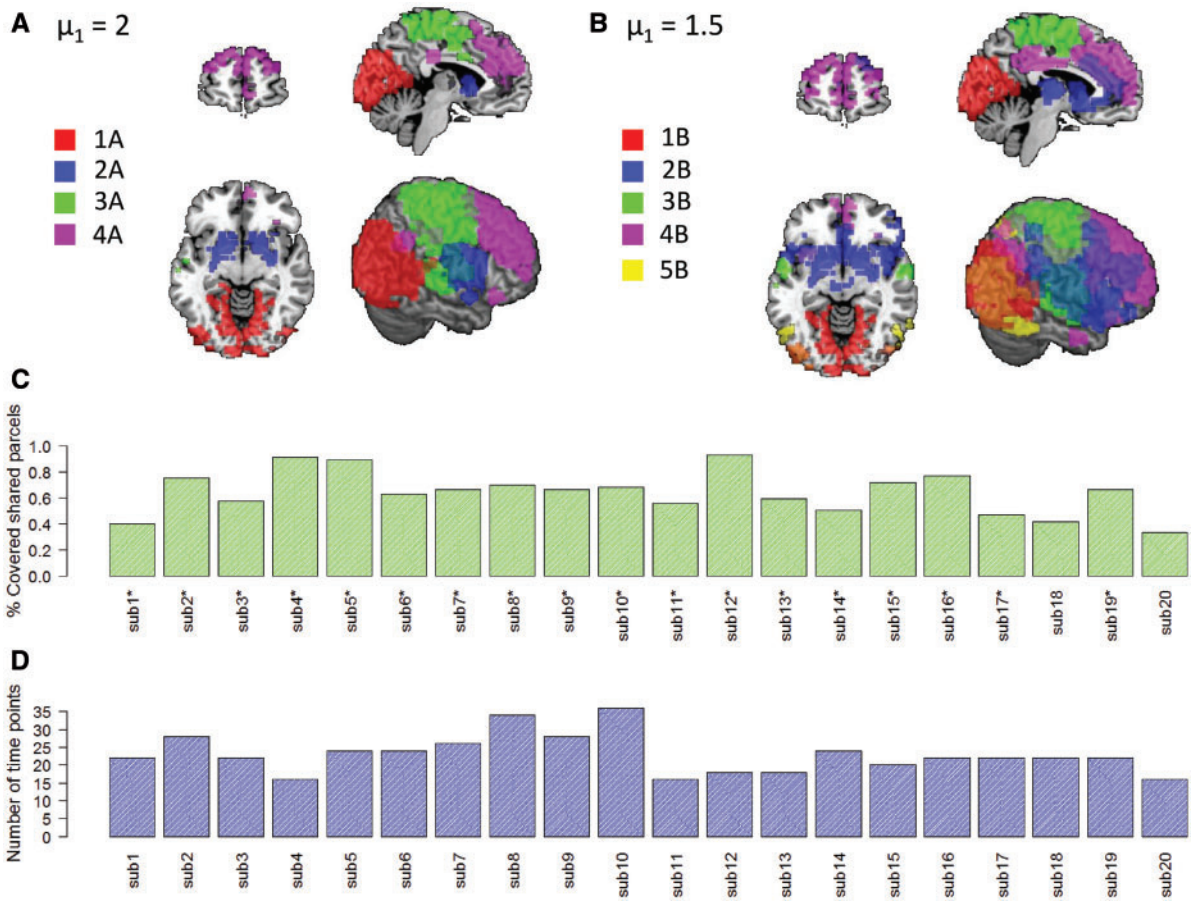


Fig. 5. Results of the fMRI analysis. (**A**) The core module rows of the solution with $\mu_1 = 2$. (**B**) The core module rows of the solution with $\mu_1 = 1.5$. (**C, D**) Examples of subject-specific statistics. This example shows the results for core module 4B. (C) The percent of core module parcels covered by the private modules. Asterisks indicate subjects whose private module had a significant overlap (hyper-geometric $P \leq 0.001$) with the core module. (D) The number of time points in each private module

core module 3 was enriched with parcels located in regions that are involved in sensori-motor processing ($q \leq 1E-11$) and in parcels located within regions of the dorsal attention network, which is involved in top-down orienting of attention ($q \leq 0.03$) (Fox *et al.*, 2006). Modules 4A and 4B were enriched with parcels located in the default mode network ($q \leq 1E-4$), which is composed mainly of midline structures and is involved in self referential functions that include remembering the past as well as planning the future, and the frontoparietal control network, which is responsible for adaptive behavior ($q \leq 1E-7$) (Dosenbach *et al.*, 2007). Finally, core module 5B contained 29 parcels and was enriched with regions that are involved in visual processing ($q \leq 5E-9$) and with parcels that are located within the dorsal attention network ($q \leq 0.001$).

Inspecting the private modules, we observed large heterogeneity in their tendency to overlap with their core module parcels and in the number of time points. Figure 5C and D shows the results for core module 4B. This module is of particular interest as it was enriched with both the default mode network and the frontoparietal control networks. Patterns of co-activation between these two networks have been reported before and suggested to support goal-directed thought processes (Spreng *et al.*, 2010). On average, each private module covered 44.4% of the core module parcels (Fig. 5C) and contained 15.5 time points. In addition, in 18 out of 20 subjects, the overlap between the subject-specific parcels and the core module parcels was significant (hyper-geometric $P < 0.001$). Other modules had much higher coverage. For example, core module 1B had mean coverage of 64.4% and a larger number of time points (mean 23), see Supplementary Figure 4.

When including the private modules in the enrichment analysis, 15 out of 20 private modules of core module 4B were also enriched with the default mode network. The frontoparietal control network was identified in 12 of the 20 subjects. Although to a much lower extent than in the gene expression analysis, we also detected subject-specific signal. For example, ventral attention enrichment was identified in 4 out of 20 subjects but not in the core module. This suggests a tendency of these four subjects to engage in bottom-up processing (e.g. be more attentive to sensory stimuli) during goal-directed thought processes. These results demonstrate the advantage of our multi-subject analysis: it was able to detect large functional networks that reappear as activated across subjects and even highlight subject-specific activation patterns.

3.7 Comparison to related algorithms

Extant algorithms for three-way data analysis were mainly developed for gene expression data. Triclustering (Zhao and Zaki, 2005) assumes that a module is a subcube created by one subset in each of the three dimensions. This setting is too rigid for simultaneous analysis of responses in many patients. Figures 4 and 5 show that our modules are not triclusters since the time points and gene

set of each private module differ under the same core module. Another type of three-way analysis seeks biclusters $< G', S' >$ where G' is a set of genes and S' is a set of subjects, such that all genes in G' manifest a similar time response across all subjects in S'. Two such algorithms are EDISA (Supper *et al.*, 2007), which seeks high correlation between subjects across time points, and the plaid model of Mankad and Michailidis (2014), which extends (Lazzeroni and Owen, 2002) and seeks up- or down-regulated time responses. Finally, Gerber *et al.* (2007) simultaneously cluster tissues and genes to produce biclusters, while accounting for three possible time responses for each tissue when introduced to a drug. However, this analysis answers very different questions than TWIGS as it assumes a hierarchical structure of tissue clusters without overlap, whereas we analyze a single tissue over many time points at rest and allow overlapping core modules.

We compared TWIGS to seven methods: ISA (Bergmann *et al.*, 2003), Bimax (Prelic *et al.*, 2006), SAMBA (Tanay *et al.*, 2004), EDISA (Supper *et al.*, 2007), the plaid model (Mankad and Michailidis, 2014), sliding window analysis of fMRI data (Allen *et al.*, 2014) and modularity analysis of fMRI data (Rubinov and Sporns, 2010, 2011). For each method, we tested a wide range of its internal parameters to fine tune it for the tested dataset. The Supplementary Text provides all details; here we give a brief overview.

Our comparison shows that except for modularity analysis (which enforces using all subjects by the method's definition), extant methods have difficulties in finding modules that cover many subjects. TWIGS provides an almost 2-fold improvement in the ability to find modules that cover many patients. For example, on average, modules identified by EDISA on the sepsis data covered less than five patients compared with nine by TWIGS. The sliding window analysis, which estimates the covariance matrix of each time window and then clusters all windows from all subjects, had an average coverage of less than 10 on the fMRI data, whereas TWIGS covered all 20 subjects. TWIGS was comparable to other methods in enrichment analysis for known biological functions in terms of: (i) the total number of covered functions, (ii) the strength of the detected enrichments and (iii) the fraction of modules with enriched terms. When consolidating scores 1–3 using non-parametric ranking, TWIGS ranked first.

When applying the plaid model to the sepsis data, it tended to find much larger gene sets. However, these modules manifested a very mild up-regulation response compared with the TWIGS modules. The fMRI modularity analysis method of Rubinov and Sporns (2010, 2011) partitioned the brain into clusters, each containing one of our core modules. TWIGS's subject-specific module augmentations provided additional biological results.

4 Discussion

We presented a novel problem formulation and algorithm for flexible three-way clustering of multi-matrix time course data. We defined a core module as (i) a set of rows that are likely to be active together across a set of subjects and (ii) a set of active time points in each covered subject. In addition, each core module has subject-specific private modules that can contain additional genes and have high overlap with the core module. The set of active time points of a module can vary in size and times among subjects.

Our model is much more flexible than existing models. First, it allows different active time points for each subject, thereby accommodating heterogeneity and asynchrony in the response of different subjects. Second, different subjects can differ in their underlying features (rows) and time points (columns). The row set of a particular subject in a module does not necessarily cover all core module rows. This property was crucial in the analysis of fMRI data, where it allowed discovering core modules that better covered active brain regions. In addition, the row set of a private module can contain additional rows that represent subject-specific signal. This property was crucial in the gene expression case as it allowed discovering patient-specific up-regulated immune processes.

We compared TWIGS to seven other methods and showed that extant methods have difficulties in finding modules that cover many subjects, whereas TWIGS easily finds modules that represent a biological function shared by many subjects. In addition, our method outperformed other methods in terms of enrichment analysis. We employed additional metrics for evaluation in each domain. Other comparison criteria can be used in the future, e.g. test-likelihood or perplexity.

Our current analysis has some limitations that can be addressed by future studies. First, we assume that the data originated from two distributions f_0 and f_1. Other approaches could be considered, such as row-based or column-based additive models (Lazzeroni and Owen, 2002). Second, our basic model deals with only a single module at a time. More complex models and algorithms could be proposed to directly model multiple modules. Finally, additional tests are needed to fully exploit the abilities of the model. For example, we focused only on testing a time window of size 1 to find homogenous highly activated private modules.

Funding

This study was supported in part by the Israel Science Foundation (grant 317/13, to R.S.), the Azrieli Foundation Fellowship and the Edmond J. Safra Center for Bioinformatics at Tel Aviv University (to D.A.). Additional support was provided by the Israeli Centers of Research Excellence (I-CORE) Gene Regulation in Complex Human Disease, Center No. 41/11 (R.S.) and Program in the Cognitive Sciences (to T.H.).

Conflict of Interest: none declared.

References

Allen,E.A. *et al.* (2014) Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex*, **24**, 663–676.

Belliveau,J.W. *et al.* (1991) Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, **254**, 716–719.

Ben-Dor,A. *et al.* (2003) Discovering local structure in gene expression data: The order-preserving submatrix problem. *J. Comput. Biol.*, **10**, 373–384.

Bergmann,S. *et al.* (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **67**, 031902.

Birn,R.M. (2012) The role of physiological noise in resting-state functional connectivity. *Neuroimage*, **62**, 864–870.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Craddock,R.C. *et al.* (2012) A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.*, **33**, 1914–1928.

Dede,D. and Ogul,H. (2013) A three-way clustering approach to cross-species gene regulation analysis. In: *2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, IEEE, pp. 1–5.

Dosenbach,N.U.F. *et al.* (2007) Distinct brain networks for adaptive and stable task control in humans. *Proc. Natl Acad. Sci. USA*, **104**, 11073–11078.

Fox,M.D. *et al.* (2006) Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proc. Natl Acad. Sci. USA*, **103**, 10046–10051.

Gerber,G.K. *et al.* (2007) Automated discovery of functional generality of human gene expression programs. *PLoS Comput. Biol.*, **3**, 1426–1440.

Hartigan,J.A. (1972) Direct clustering of a data matrix. *J. Am. Stat. Assoc.,* **67**, 123–129.

Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Stat. Sin.,* **12**, 61–86.

Li,A. and Tuck,D. (2009) An effective tri-clustering algorithm combining expression data with gene regulation information. *Gene Regul. Syst. Biol.,* **3**, 49–64.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE-ACM Trans. Comput. Biol. Bioinform.,* **1**, 24–45.

Maldjian,J.A. *et al.* (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, **19**, 1233–1239.

Mankad,S. and Michailidis,G. (2014) Biclustering three-dimensional data arrays with plaid models. *J. Comput. Graph. Stat.,* **23**, 943–965.

Meng,J. *et al.* (2009) Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics,* **25**, 1521–7.

Mitra,K. *et al.* (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.,* **14**, 719–32.

Oghabian,A. *et al.* (2014) Biclustering methods: biological relevance and application in gene expression analysis. *PLoS One,* **9**, e90801.

Parnell,G.P. *et al.* (2013) Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. *Shock,* **40**, 166–74.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics,* **22**, 1122–1129.

Rana,M. *et al.* (2013) A toolbox for real-time subject-independent and subject-dependent classification of brain states from fMRI signals. *Front. Neurosci.,* **7**, 170.

Rubinov,M. and Sporns,O. (2010) Complex network measures of brain connectivity: uses and interpretations. *NeuroImage,* **52**, 1059–1069.

Rubinov,M. and Sporns,O. (2011) Weight-conserving characterization of complex functional brain networks. *NeuroImage,* **56**, 2068–2079.

Serin,A. and Vingron,M. (2011) Debi: discovering differentially expressed biclusters using a frequent itemset approach. *Algorithms Mol. Biol.,* **6**, 18.

Shabalin,A.A. *et al.* (2009) Finding large average submatrices in high dimensional data. *Ann. Appl. Stat.,* **3**, 985–1012.

Spreng,R.N. *et al.* (2010) Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage,* **53**, 303–317.

Stamatakis,E.A. *et al.* (2010) Changes in resting neural connectivity during propofol sedation. *PLoS One,* **5**, e14224.

Supper,J. *et al.* (2007) EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics,* **8**, 334.

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.

Ulitsky,I. *et al.* (2010) Expander: from expression microarrays to networks and functions. *Nat. Protoc.,* **5**, 303–322.

Vaisvaser,S. *et al.* (2013) Neural traces of stress: cortisol related sustained enhancement of amygdala-hippocampal functional connectivity. *Front. Hum. Neurosci.,* **7**, 313.

Waltman,P. *et al.* (2010) Multi-species integrative biclustering. *Genome Biol.,* **11**, R96.

Yeo,B.T.T. *et al.* (2011) The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.,* **106**, 1125–1165.

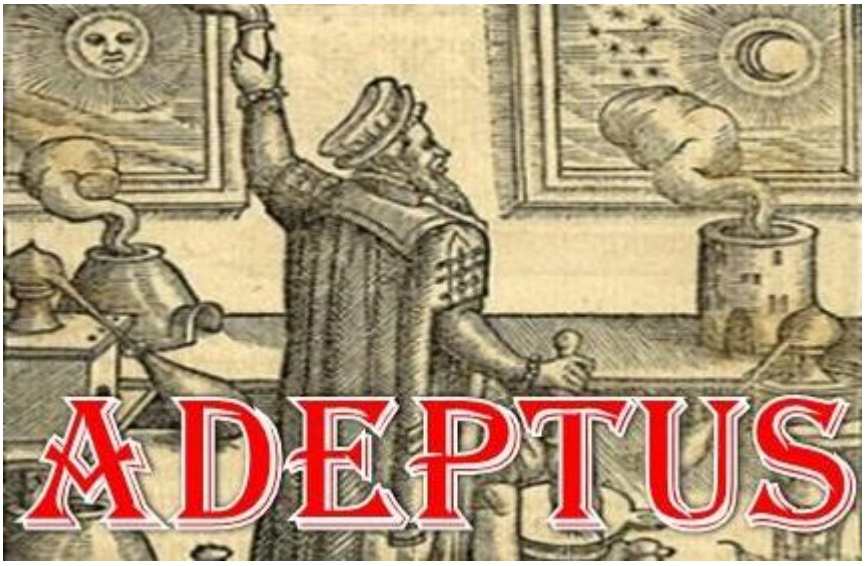
Zhao,L. and Zaki,M.J. (2005) TriCluster: an effective algorithm for mining coherent clusters in 3d microarray data. In: Ozcan,F. (ed.), *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. ACM Press, Baltimore, MD, USA, pp. 694–705.

4. **Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets**.

D. Amar, T. Hait, S. Izraeli and R. Shamir.

Nucleic Acids Research

doi: 10.1093/nar/gkv810, 2015



Published online 10 August 2015

Nucleic Acids Research, 2015, Vol. 43, No. 16 7779–7789
doi: 10.1093/nar/gkv810

Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets

David Amar[1], Tom Hait[1], Shai Izraeli[2,3] and Ron Shamir[1,*]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel, [2]Department of Pediatric Hematology-Oncology, Safra Children's Hospital, Sheba Medical Center, Tel Hashomer, Ramat Gan 52620, Israel and [3]Sackler School of Medicine, Tel-Aviv University, Tel Aviv 69978, Israel

Received March 22, 2015; Revised July 23, 2015; Accepted July 29, 2015

ABSTRACT

Genome-wide expression profiling has revolutionized biomedical research; vast amounts of expression data from numerous studies of many diseases are now available. Making the best use of this resource in order to better understand disease processes and treatment remains an open challenge. In particular, disease biomarkers detected in case–control studies suffer from low reliability and are only weakly reproducible. Here, we present a systematic integrative analysis methodology to overcome these shortcomings. We assembled and manually curated more than 14 000 expression profiles spanning 48 diseases and 18 expression platforms. We show that when studying a particular disease, judicious utilization of profiles from other diseases and information on disease hierarchy improves classification quality, avoids overoptimistic evaluation of that quality, and enhances disease-specific biomarker discovery. This approach yielded specific biomarkers for 24 of the analyzed diseases. We demonstrate how to combine these biomarkers with large-scale interaction, mutation and drug target data, forming a highly valuable disease summary that suggests novel directions in disease understanding and drug repurposing. Our analysis also estimates the number of samples required to reach a desired level of biomarker stability. This methodology can greatly improve the exploitation of the mountain of expression profiles for better disease analysis.

INTRODUCTION

Gene expression studies use expression profiles of cases and controls to understand a disease by identifying genes and pathways that differ in their expression between the two groups. This methodology has become ubiquitous in biomedical research, and is often combined with additional information of either the patients or the genes to interpret the results (1–7). However, these analyses suffer from several limitations: the discovered biomarkers often have low reproducibility, and are difficult to interpret biologically and especially clinically (8,9).

A promising direction for increasing robustness is by integration of many gene expression datasets. The difficulty here is in creating a common denominator of multiple studies, often conducted using different platforms under diverse experimental conditions and tissues. Huang *et al.* (10) used 9169 gene expression samples, each associated with a set of disease terms of the Unified Medical Language System (UMLS). UMLS, and similar databases such as Disease Ontology (DO), provide ontology of disease terms organized in a hierarchy that models dependencies among diseases (11,12). The authors presented an algorithm that predicts a set of disease terms for each gene expression sample (10). Schmid et al. analyzed 3030 samples of one platform and predicted their UMLS terms using similarity-based analysis (13). Lee *et al.* used >14 000 profiles of one microarray technology to predict the tissue of a sample (14). While these studies reported good prediction quality, they have some limitations. First, data of only one or two expression platforms were analyzed, limiting the data used and the applicability of the results. Second, in Huang *et al.* and in Lee *et al.* the mapping of samples to their disease terms was done automatically, inevitably introducing mapping errors (10). Third, the predictor in (10) can be applied on new patient samples only if a set of new control samples accompanies them. Fourth, while the prediction performance of

*To whom correspondence should be addressed. Tel: +972 3 640 5383; Fax: +972 3 640 5384; Email: rshamir@tau.ac.il

the classifiers was far from random there is still substantial room for improvement. Finally, many biomarker sets are hard to interpret biomedically, which hampers their adoption in clinics.

To improve interpretability, several classification methods that integrate different biological data were suggested. For example, combining patient gene expression profiles and protein–protein interaction data or pathway information was demonstrated to improve disease classification accuracy or biological interpretability in some studies (1–7,15,16). However, other studies reported no significant improvement when utilizing network data (1,17). Moreover, in some cases the contribution of the additional gene data was very mild, which questions the benefit from interpreting these models. Other methods for biomarker discovery used prior knowledge on genes to extract differential genes of similar functionality. As another example, Ciriello *et al.*, integrated gene expression profiles, methylation profiles, and single nucleotide polymorphism (SNP) data from 3299 cancer patients to construct a hierarchical structure of the patients, and used it to detect novel biomarkers of cancer subtypes (18).

The main goal of this study is integration of numerous heterogeneous expression profiles to produce reliable results that could be used as a starting point for novel biomedical insights. We focus on identification of the main genes that are specifically differential in a disease of interest and putting them in the context of interactions, mutations, and drugs. To be able to produce such overviews in a meaningful way we developed a four-step procedure (Figure 1). Each step is essential to obtain reliable results. First, we manually annotated more than 14 000 gene expression profiles from 175 datasets to produce a compendium called ADEPTUS (Annotated Disease Expression Profiles Transformed into a Unified Suite). ADEPTUS covers 13 314 microarray samples from GEO and 1526 RNA-Seq samples from TCGA. To overcome study and sample heterogeneity, each sample was normalized using a non-parametric rank-based method. Samples were manually annotated with the most relevant disease terms in DO. Second, as a quality assurance step we tested different multi-label classification algorithms. Two key issues here were: (i) performing leave-dataset-out cross validation to reduce bias of unknown covariates (e.g. batch effects), and (ii) showing that standard performance measures produce over-optimistic results, and rectifying this by introduction of more stringent classification measures. Using our measures, classification performance was very high for 24 diseases, mostly cancer subtypes. Limiting the data to a single platform improved the performance for six additional diseases.

Third, we detect disease-specific differentially expressed genes, by accounting for the diversity of non-disease samples and the relationships among diseases. We demonstrate a shortcoming in the integration of multiple datasets of a single disease: without using alongside it data of other diseases, such analysis might find genes of general disease phenotypes that are not specific to the target disease. Our method was designed to overcome this difficulty. We demonstrate the robustness of our method, and also achieve estimates for the number of datasets and samples required to improve stability of biomarker detection. Functional enrichment analysis shows that the detected gene sets recapitulate known hallmarks of the diseases. Finally, for three cancer types we produce a network that highlights the molecular modification in the disease. This is done by an integrative analysis of the discovered differential genes with information on somatic mutations, drug targets, and gene interactions. We show that our results detect well known disease genes and treatments, and even suggest new indications of several known drugs.

MATERIALS AND METHODS

The expression profile compendium

We constructed a large compendium of expression profiles generated using different technologies, and manually annotated the diseases attributed to each profile (Supplementary Figure S1A). The compendium, called ADEPTUS, contains 174 gene expression studies from GEO (19), each with at least 20 samples. Overall, ADEPTUS covers 13 314 samples from 17 different microarray technologies, and 1526 RNA-Seq samples from TCGA (20). See Supplementary Text regarding using even larger compendia. For each study we used the preprocessed expression matrix given in the database. Each sample was either labeled as 'case' and manually assigned a set of the relevant DO terms based on the its textual description, or labeled as 'control'. To allow cross-validation on whole datasets, we kept only DO terms that were represented by at least five different datasets in our compendium. This resulted in 48 disease terms.

Single sample gene scores

To allow joint analysis across platforms, expression profiles were transformed to rank-based scores (2,21) (Supplementary Figure S1A, see 'Materials and Methods' section). Given a gene expression profile of a single sample S in which k genes were measured, we ranked the genes by their expression levels $g_1, g_2, g_3, \ldots, g_k$ (with g_1 having the highest level), and assigned a score to each gene based on its rank: $W_S(g_i) = ie^{-i/k}$. See Supplementary Text for details.

The final compendium can be summarized as two matrices (Supplementary Figure S1A): A binary (samples × diseases) matrix Y where $Y_{s,d} = 1$ if sample s is annotated with disease d, and a real-valued (samples × genes) matrix X where $X_{s,g} = \mathrm{WS}(g)$.

Multi-label classification

In *multi-label classification* each sample can belong to multiple true classes (e.g. cancer and lung cancer) (22,23). A sample can be predicted to have several labels and the sum over the predicted label probabilities need not be 1. Recent multi-label classification approaches (22,24,25) can be partitioned into two types: *problem transformation* and *algorithm adaptation* (23). See Supplementary Text for details. Here we used the label power-set (LP) transformation method, which defines for each sample a categorical class variable by concatenation of the sample's original labels (26). We also used the Bayesian correction (BC) adaptation method, which uses the known label hierarchy to correct errors after learning an independent single binary clas-

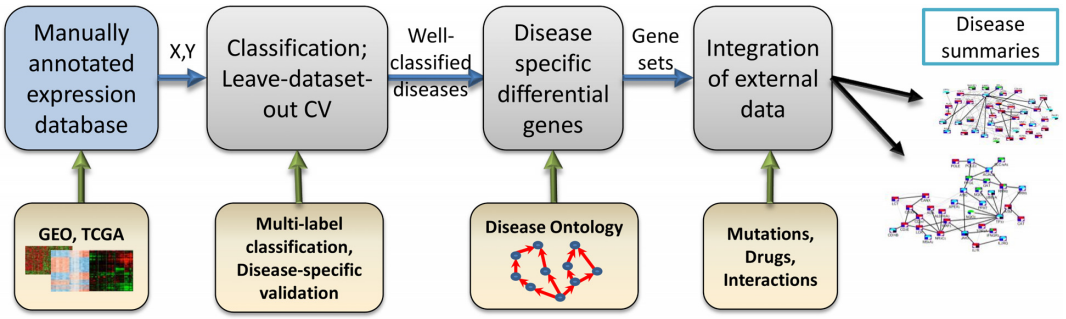


Figure 1. Study workflow overview. Step 1: Assembly of the ADEPTUS database: expression profiles from public sources were normalized and manually annotated. Step 2: Classification methods were used to identify well-classified diseases while avoiding over-optimistic results due to tissue and batch effects. Step 3: Disease-specific biomarker detection using the Disease Ontology structure. Step 4: Integration with other biomedical data produces gene-centric, disease-specific overview with therapeutic potential.

sifier for each label (10,27). Linear SVM (28,29) and random forest (30) were used as the binary classifiers.

Somatic mutation data

We analyzed the raw data of known somatic mutations from COSMIC (31). These data contained associations between genes and tumor samples. We kept only associations to non-silent mutations in coding regions that were also marked as 'confirmed somatic mutations'. The result was 559 727 gene-tumor associations, covering a total of 43 517 tumor samples and 20 332 genes. We then assigned genes to tumor sites by calculating a hyper-geometric (HG) p-value for the overlap between the samples that had a mutation in the gene and the samples from the site. The p-values were FDR corrected for multiple testing and only significant associations were kept ($q \leq 0.05$).

Gene–drug associations

Gene–drug associations were taken from DrugBank (32). Only approved drugs were used.

Network visualization and functional genomics

Network visualization was done using Cytoscape (33) and the Cytoscape application enhancedGraphics (34). Enrichment analysis in Cytoscape was done using BiNGO (35). GeneMania (36) was used to generate networks of a selected gene set. EXPANDER (37) was used for enrichment analysis of all discovered gene sets.

Validation of the multi-label classifier on RNA-Seq data

To test the performance of a multi-label classifier that was trained using the microarray samples, on the RNA-Seq samples, we transformed each RNA-Seq sample to gene weighted ranks. We then performed quantile normalization on all samples together. That is, we created a matrix whose rows are the samples including both the microarray samples and the RNA-Seq samples. The columns were the genes covered by the microarray data and the matrix values were the weighted ranks. Quantile normalization was performed to ensure that rows in the matrix would have similar distributions. This is crucial as any classifier assumes that the tested data and the training data are similarly distributed. Finally, the classifier was tested by computing its predictions on the rows of the RNA-seq samples.

Testing how biomarker stability depends on the amount of data

To test how the stability of our approach depends on the number of datasets used, we focused on DO term 'organ system cancer', which had 46 datasets in the compendium, of which 16 were not assigned to any sub-disease. To measure stability, we (i) randomly selected from these 46 datasets two disjoint subsets A and B of k datasets each, (ii) ran our pipeline and obtained biomarkers on each subset separately and (iii) measured the Jaccard score and the significance of the overlap between the two biomarkers. This process was repeated with k ranging from 5 to 23. As background controls, we added half of the remaining 128 non-'organ system cancer' datasets to A and the rest to B. We rejected sets generated in step (1) if the total numbers of samples in A and B differed by more than 20%.

RESULTS

We collected and curated a large compendium of gene expression profiles from diverse diseases and developed and tested several approaches for classifying patient samples originating from each disease. For those diseases whose classification was validated successfully we developed specific biomarker genes and summarized them in the context of protein interaction, mutation and drug target data. Figure 1 shows an outline of our approach.

A curated gene expression compendium

We constructed a large compendium of >14 000 expression profiles generated using 18 different technologies. Each profile was designated as case or control and cases were manually assigned DO terms. The compendium, called ADEPTUS, contains 174 gene expression microarray studies and 1526 RNA-Seq samples. It covers 48 DO terms, including many cancer subtypes, obesity, neurodegenerative diseases and cardiovascular disease (see Supplementary Figure S2). Each sample was rank-normalized to allow com-

parison of samples from different technologies (see 'Materials and Methods' section). We observed that following rank normalization, the correlation between samples from different platforms (preprocessed using different methods) is high, see Supplementary Text.

Classification

We conducted a systematic analysis of classification methods in order to identify diseases in which the expression signal was consistent across datasets. The main components of the analysis were (i) utilization and comparison of several multi-label classification algorithms, (ii) leave-dataset-out cross validation to overcome technology and batch effects and (iii) a careful examination of the results in each disease separately in order to avoid over-optimistic conclusions due to tissue effects.

The classifiers. Samples in the compendium can have multiple related disease labels (e.g. hematologic cancer and ALL). Classification of the samples can be addressed in such situation as a *multi-label classification problem*. In that problem a sample can be classified into several disease terms, and the sum of its label probabilities need not be 1. See Supplementary Text for full details and references. We first analyzed the 13 314 microarray samples. We tested three multi-label classification approaches: (i) *Single*: learning a classifier for each disease separately, (ii) *LP*: classification using multiclass algorithms on the label power-set of the training data (22,23) and (iii) *BC:* Bayesian correction of single-label classifiers (10,27). All three approaches rely on a binary 'base classifier'. See Supplementary Text for details. We tested support vector machines (SVM) (28,29) and random forest (RF) (30) as the base classifier.

Three sample categories for each disease. The common practice when testing a classifier is to train and evaluate its performance in a binary setting, separating the samples into the cases versus all the rest. However, this is problematic when the data come from many diseases: When classifying one disease, samples that come from other disease studies would typically originate from different tissues, and thus may be easier to separate from the cases based on tissue characteristics, irrespective of the disease. On the other hand, controls in the same study will typically originate from the same tissue, or the same patient, and match the cases in sex and edge distribution. As a result, they would be biologically more similar to the cases and harder to classify. For that reason, for each disease we chose to define three types of samples: (i) ***positives***: patients with the disease; (ii) ***negatives***: control samples originating from the same studies as the positive samples and (iii) ***background controls*** **(BGCs):** all other samples. Thus, a classifier that performs well in separating the positives from the rest may actually provide poor separation of the positives from the negatives (see Supplementary Figure S3).

Cross-validation. We wanted to test the classification quality on samples that are completely unrelated to those used for training, and possibly from different technologies. This would also reduce the risk of batch effects. For this purpose, we used leave-datasets-out cross-validation (LDO-CV): In each cross-validation round 15 complete datasets were put aside, a classifier was learned on the rest of the data and then tested on the left-out datasets. The output of each classifier is a matrix P, where P_{ij} is the probability that sample i has disease j (computed when it was in the left-out test set during the LDO-CV).

Evaluation criteria. For each disease (i.e. taking a specific column of P) we calculated three scores: (i) *PN-ROC:* the area under the ROC curve (AUC-ROC) comparing positives and negatives (ii), *PB-ROC:* AUC-ROC comparing the positives to the BGCs and (iii) a meta-analysis statistical significance score, based on Stouffer's method (38), for separation of positives and negatives within datasets (see Supplementary Text), denoted as *SMQ* (Study-based Meta-analysis *Q*-value). These three scores provide complementary evaluation criteria.

Comparing classifiers. The performance of the classifiers is shown in Figure 2A. We designated a disease *well-classified* if its PB-ROC and PN-ROC scores exceeded 0.7 and its SMQ was significant (<0.05). See Supplementary Text for further explanation on the thresholds. Single-SVM and SVM-BC had the highest average PN-ROC (0.69). Notably, all classifiers had high standard deviation across diseases ($0.175 < \sigma < 0.19$). As expected, the PB-ROC scores were higher than the PN-ROC scores, indicating that obtaining separation between positives and BGCs is an easier task. Overall, Single-SVM performed second in PN-ROC, only slightly below the best algorithm (BC-SVM), and achieved the highest number of well-classified diseases (24). Moreover, when changing the ROC threshold, the SVM-based algorithms consistently outperformed the rest in terms of the number of well-classified diseases. For these reasons, and since the single-SVM classifier is simpler, we used this classifier in all subsequent analyses.

Comparison to extant algorithms. We calculated a global precision-recall curve, also known as a micro-AUC score in learning (22): we treated P and Y as a set of pairs (Y_{ij}, P_{ij}), and used P_{ij} to rank all pairs. This ranking was then used to calculate a precision-recall curve, see Figure 2B. The AUPR was 0.68. Both precision and recall were much higher than in (10): we achieved 93% precision (compared to 82%) at 20% recall, and 44% recall (compared to 20%) at 82% precision (Figure 2B).

Testing classification on samples from a new technology. As an additional validation, we used the 1526 RNA-Seq samples in ADEPTUS. We trained the Single-SVM classifier using all microarray samples and tested its performance on the RNA-Seq samples. The RNA-Seq test data contained 918 breast cancer samples, 102 control breast biopsies, 182 intestinal cancer samples, 173 leukemia samples, and 151 samples from other cancer types. Given the classifier for each disease, we calculated the ROC curve comparing the disease samples to all other RNA-Seq samples, see Figure 2C. All ROC scores were >0.96. Note that in these data only the breast cancer samples had direct negative samples. This can explain why these ROC scores are much higher than those

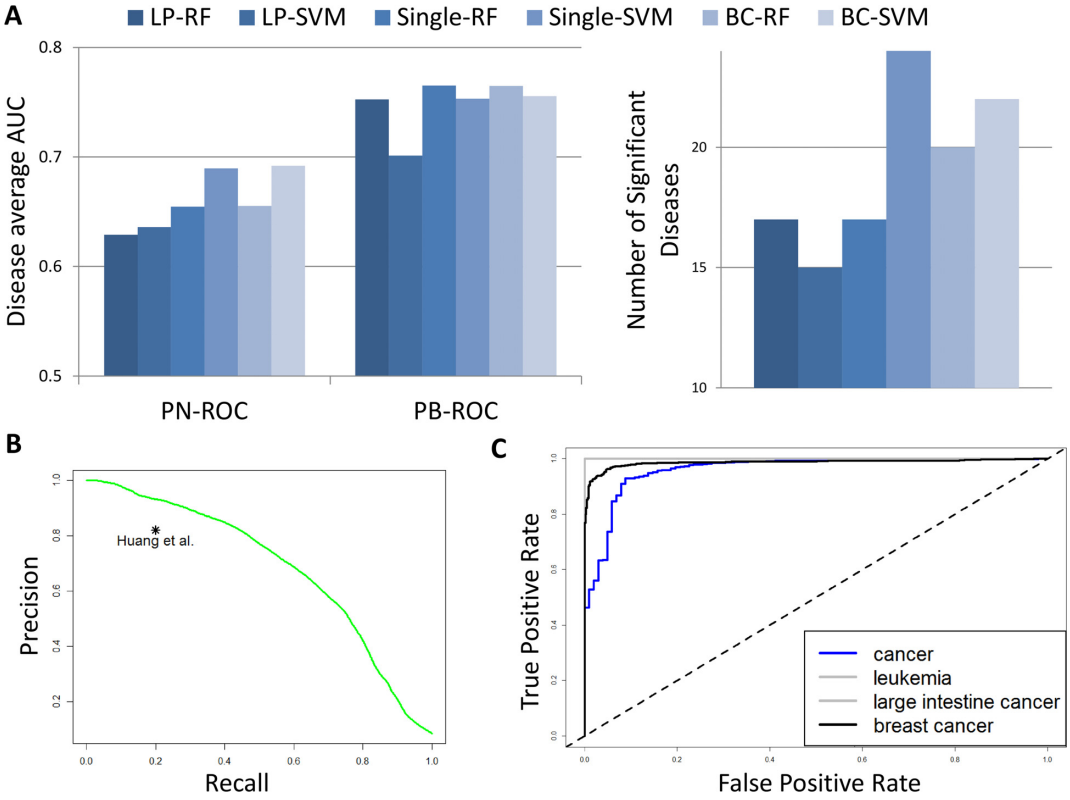


Figure 2. Multi-label classification performance. For each classifier we calculated for each disease the area under the ROC curve comparing positives and negatives (PN-ROC) and the area under the ROC curve comparing positives and background controls (PB-ROC). (**A**) Average performance of the classifiers. Left: Each bar shows the average ROC-AUC over all diseases. LP: label power-set, BC: Bayesian correction, single – single-class classifier; RF: random forest, SVM: support vector machine. Right: The number of disease terms that had both PN-ROC and PB-ROC at least 0.7 and were found significant in the SMQ test. (**B**) The global precision-recall curve of the single-SVM classifier. This analysis measures the overall agreement between the predicted probabilities of sample-disease association and the known labels. The point represents the performance of (18). (**C**) The Single-SVM classifier performance on a test set of 1526 RNA-Seq samples. The ROC score for both leukemia and large intestine cancer is 1.

obtained in the cross validation. Nevertheless, our classifier correctly assigned the cancer to the patients even though the samples were from a technology that was not used at all in the training.

The validation confirms that our classifiers perform well across technologies and platforms. Our analysis produced successful classification for 24 diseases, most of which are cancer subtypes (see Figure 3). It may be possible that the other diseases were less well classified due to loss of information in the rank normalization. To test this, for each of those diseases we reran the LDO analysis process above using only samples from one platform, choosing the platform that had the largest number of the disease datasets. Profiles underwent standard quantile normalization, which retains more of the original signal than the weighted ranking needed when combining data across platforms. The results show that classification performance can be improved for some of these diseases by narrowing down the analysis to one platform, see Supplementary Text. For example, analyzing separately six datasets of 'musculoskeletal system disease' that used the same platform (GPL96, Affymetrix 133A), the classifier achieved a ROC score of 0.84.

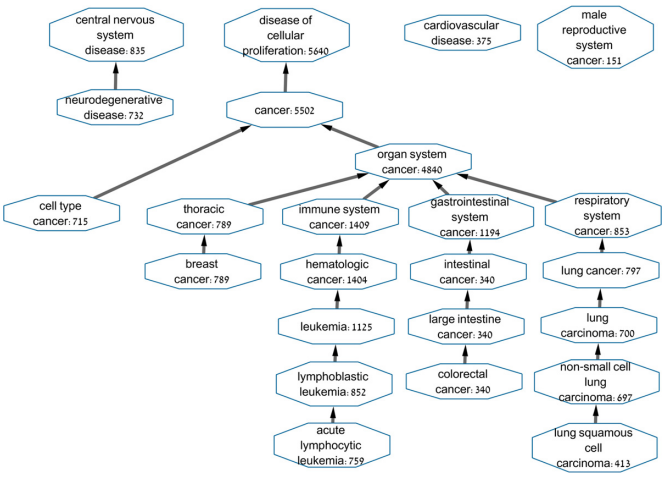


Figure 3. The 24 well-classified diseases. For each node, the Disease Ontology term and the number of positive samples are shown. Edges mark 'is-a' relation in the DO hierarchy.

Detecting disease-specific differential genes

In order to identify genes that are specifically differential in a particular disease, we used the three-way partition of

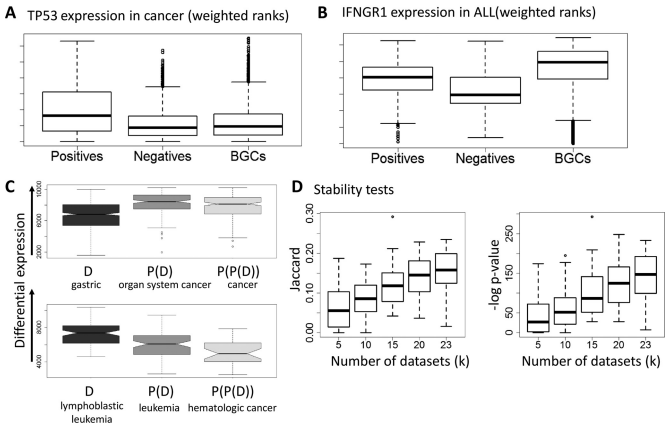


Figure 4. Expression patterns, specificity and robustness. (**A**) Expression of TP53 in cancer. (**B**) Expression of IFNGR1 in ALL. The *y*-axis represents the weighted rank of a gene, where higher ranks have better values. The boxplots show the expression distribution of the three sample cohorts: positives, negatives, and BGCs. TP53 is over-expressed in cancer compared to both negatives and BGCs. IFNGR1 is up-regulated compared to negatives, but down-regulated compared to BGCs. (**C**) Ranking of gastric cancer biomarker genes and of biomarkers for more general cancers on gastric cancer datasets. Plots for biomarkers of more general diseases (parent and grandparent nodes in the Disease Ontology) are further to the right. General cancer genes are ranked higher, indicating that analysis of these datasets alone will not discover the gastric cancer specific genes. (**D**) Testing stability. The plots show the overlap between solutions obtained using two disjoint sets of k disease datasets each, as a function of k. Each boxplot shows the distribution of the overlap scores for a specific k over 50 repeats.

the samples for that disease, and calculated for each gene the PN-ROC, PB-ROC, and SMQ scores. Note that here the distance from 0.5 (in either direction) indicates how informative a gene is. For simplicity, for each ROC score x we report here max (x, 1 – x), and indicate whether the gene is up- or down-regulated. Figure 4A and B shows two differential expression patterns. For TP53 in cancer, the positives are up-regulated compared to both negatives and BGCs (PN- and PB-ROC $\geq$ 0.65, SMQ $\leq$ 2.22E–10). For IFNGR1 in ALL, IFNG1 is up-regulated in positives compared to negatives (PN-ROC = 0.675, SMQ = 0.001) but down-regulated compared to BGCs (PB-ROC = 0.7). Although the PN-ROC and PB-ROC scores are computed by comparing the positives to two disjoint sample groups, we observed high correlation between them across different diseases (0.46 $\pm$ 0.15). For example, in cancer, most differential genes showed the same direction of change in positives versus negatives and in positives versus BGCs, and only a few showed different directions as in Figure 4B.

We designate a gene as specific to a disease D if both of its ROC scores are $\geq$0.65 and it has SMQ score $\leq$0.05. (We chose the ROC threshold more permissively here since some diseases had only few genes with ROC > 0.7). Note that this approach is highly stringent in that we remove genes with a significant q-value whose differential signal is not intense enough. This process produces an initial set of potential genes for D, but can leave high overlap between gene sets of related DO terms. To make sure that a selected gene G is indeed specifically differential in D, D was considered only if it is a leaf or it has at least three datasets whose most specific annotation is D (i.e. samples in them were assigned to D but not to any of its children). In that case we re-calculated the SMQ score using these datasets only. If G was found significant in that test, this indicates that G is differential in D even when we exclude the samples of its sub-diseases. This test markedly reduces the overlap between related DO terms, see Supplementary Text. The resulting disease-specific gene set is designated the disease biomarker. These sets are provided in Supplementary Table S1.

Selection of the biomarker can also be done as part of classifier training. We compared the classification with both gene sets and the results were similar. We preferred determining the biomarker by the procedure described here, as it focuses on genes that are differential and directly addresses the redundancy between related diseases.

Biomarker specificity and robustness

We evaluated the specificity of the disease biomarker sets that we obtained. In each dataset we ranked all the genes by their differential expression score (the difference in the mean rank based score between the cases and the controls) and then computed the median rank of each gene across all the datasets of each disease (see Supplementary Text). Focusing on the datasets of a particular disease, we computed the ranks of its biomarker set, the ranks of the biomarker set of the parent disease, and of the grandparent disease, when available. We expected that a specific biomarker should show higher ranks on its disease data than the biomarker of the more general parent and grandparent disease. The results are summarized in Supplementary Figure S4 and Figure 4C. For most diseases, e.g. lymphoblastic leukemia (Figure 4C), the ranks of the disease gene sets are significantly higher than those of their ancestors. However, in gastrointestinal cancer (Figure 4C), the biomarker sets of the ancestors (organ system cancer and cancer) have much higher ranks (p < 1E–21). Hence, analyzing gastric cancer datasets without expression profiles of BGCs would lead to preferring general cancer genes over genes that are specific to gastric cancer.

A key problem in disease classification has been low overlap between biomarker gene sets obtained in different studies (39). We therefore tested how the stability of our biomarkers depends on the number of datasets used for learning. We focused on the disease term 'organ system cancer' because it had a large number of usable datasets (46, of which 16 were not assigned to any sub-disease). We computed biomarker sets twice based on disjoint data, and measured the overlap between the sets. This was repeated with the number of training datasets ranging from 5 to 23 (see Materials and Methods). The results (Figure 4D and Supplementary Figure S5) show that the overlap is highly significant when k > 10. Importantly, stability increases roughly linearly as a function of the number of datasets in the range we could test. We therefore fit a linear regression model to this trend and estimated the required numbers to achieve higher stability, assuming the linear trend continues. At 46 datasets and 4258 samples (all of the 46 datasets available for that disease) the predicted Jaccard score is 0.29 (expected p < 1E–270). Increasing the numbers to 100 datasets and 10 000 samples is expected to improve the Jaccard score to roughly 0.6.

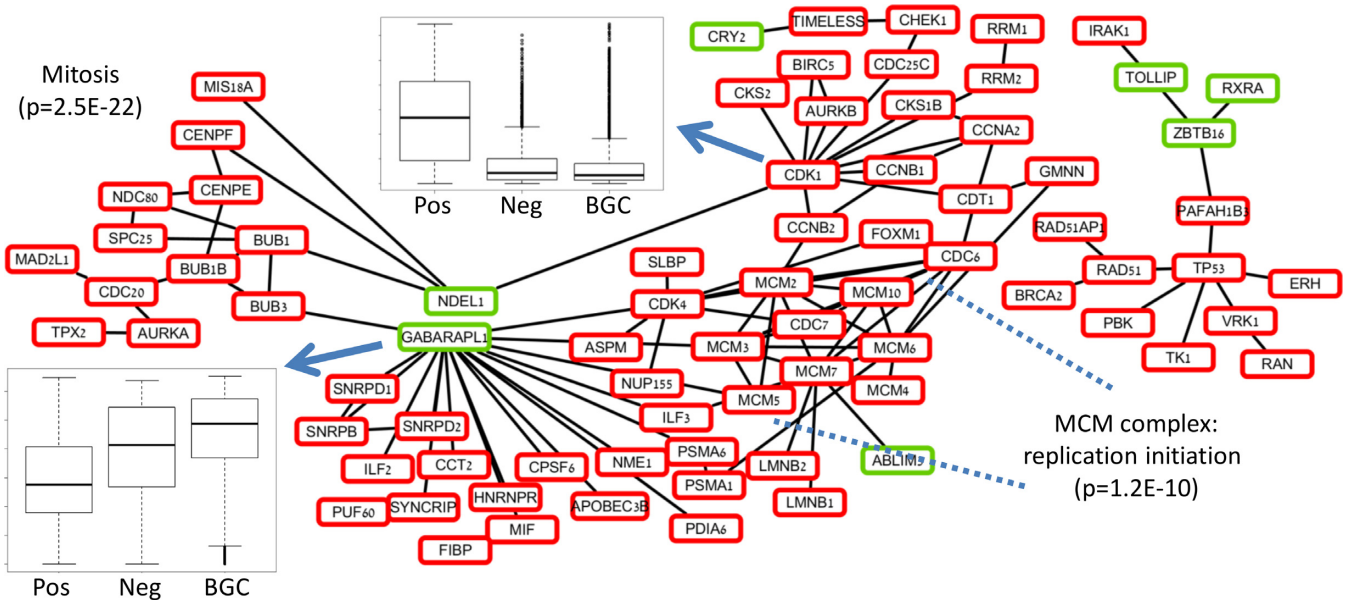


Figure 5. The main connected components of protein-protein interaction network of the cancer-specific differential genes. Up-regulated genes in cases versus negatives and BGCs are in red, down-regulated genes are in green. The large connected component (left) can be separated into two up-regulated sub-modules by removing the down-regulated genes. The down regulated genes are related to cytoskeleton, whereas the sub-modules contain mitosis, replication, and cell cycle genes. The small connected component (right) also contains mainly up-regulated genes, and has TP53 as the main hub.

Functional analysis rediscovers known disease factors and suggests novel ones

For each disease, the set of biomarker genes was partitioned by their differential expression compared to negatives and BGCs (compare Figure 4A and B) and each subgroup was tested for functional and pathway enrichment (see Materials and Methods). The results are summarized in Supplementary Table S2. Overall, the results validated our analysis by rediscovering known disease factors. In cancer the enriched biological processes included well known hallmarks of cancer such as cell cycle regulation, DNA replication, P53 signaling, chromosome organization and cell proliferation (40). In neurodegenerative disorders, the results included oxidative phosphorylation, Alzheimer's disease and Parkinson's disease. In lymphoblastic leukemia, primary immunodeficiency was down-regulated both compared to negatives and BGCs, whereas lymphocyte differentiation and V(D)J recombination were up-regulated. In gastrointestinal cancer, several pathways were down-regulated compared to negative samples, including the calcium signaling pathway and fatty acid metabolism. Interestingly, the latter is *up-regulated* compared to BGCs, indicating that this pathway's expression level in gastrointestinal cancer is reduced but not to the full extent manifested in other unrelated tissue.

We also performed network-based analysis of the identified cancer-specific gene set. This set contained 258 genes, of which 222 were up-regulated in cancer both compared to negatives and to BGCs. Figure 5 shows the two main connected components formed when connecting this gene set with the protein–protein interactions (PPI) from IntAct (41). The first component contains 14 genes including TP53 as the main hub. The second contains 64 genes. Surprisingly, two down-regulated cytoskeleton related genes, NDEL1 and GABARAPL1, connect two up-regulated sub-modules of this connected component. Functional analysis of these two sub-modules revealed that the first is composed of 12 mitosis-related genes ($p = 2.5E{-}22$), whereas the second is related to cell cycle and DNA replication (e.g. the MCM complex, $p = 1.2E{-}10$). Thus, the mitosis related sub-module is up-regulated but its ability to form physical interactions with cytoskeleton related factors is impaired, which suggests differential rewiring of the replication pathway in cancer. Such cellular modifications might cause instability and mitosis defects through impairment of cellular morphogenesis (42).

Integration with information on SNPs and drugs reveals therapeutic potential

In order to interpret our biomarkers, we integrated them with external databases to produce an overview of the molecular changes in a specific cancer and suggest potential consequences to therapy. We used COSMIC (31) for association between genes and cancer types based on occurrence of somatic mutations in coding regions (see Materials and Methods), Drugbank (32) to mark druggable genes, and GeneMania (36) for genetic interactions (GIs) and PPIs between the genes. We tested in detail three examples: lung cancer, ALL, and colorectal cancer. In each case we focused only on genes that (1) were differential in the disease or one of its ancestor DO terms, and (2) either are targets of known drugs or the gene was found associated with the disease in COSMIC.

Lung cancer. Part of the network of lung cancer, containing the two largest connected components in the PPI network, is shown in Figure 6A. The network shows TP53 as a main hub. TP53 and most of its PPI neighbors are

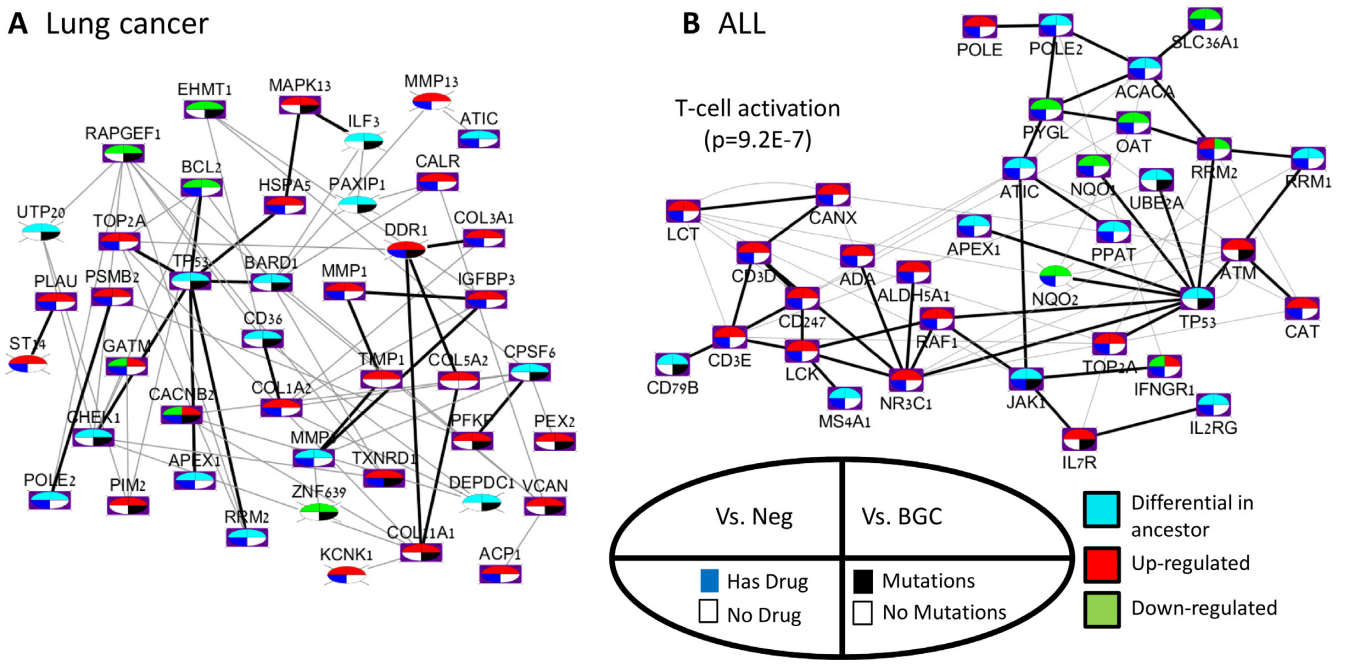


Figure 6. A network overview of the biomarkers in lung cancer and ALL. Each network shows genes that (i) were found differential specifically in the disease or in a more general disease that contains it according to the DO database, and (ii) have a drug targeting them, or were found to be associated with the disease according to the COSMIC database. Black edges are PPIs, and gray edges are GIs. Each node shows four features of a gene: (i) differential pattern compared to negatives, (ii) differential pattern compared to BGCs, (iii) whether a targeting drug exists and (iv) if the gene was associated to lung cancer according to COSMIC. Nodes without a purple background are genes that are not associated with any pathway in KEGG, Reactome, NCI, or Biocarta. (**A**) Lung cancer. The initial network (top left) contained 89 genes. The two largest connected components in the PPI network are shown. The GeneMANIA analysis added COL5A2 and TMP1 to the network. (**B**) ALL. The original network contained 136 genes and 424 edges. The figure focuses on the largest PPI connected component.

differential in cancer but are not specifically differential in lung cancer. Two neighbors of TP53 - TOP2A and HSPA5, however, are up-regulated but are not associated to the disease based on mutations. Interestingly, TOP2A (topoisomerase) is a target of multiple cancer-related inhibitory drugs such as Teniposide, and Valrubicin. In another PPI-based connected component of the network, the hub is DDR1, a key player in communication of cells with their microenvironment (43). It interacts with up-regulated collagen related genes COL5A2, COL11A1 and COL3A1 (44). DDR1, which is not covered by the major pathway databases KEGG (45), Reactome (46), NCI (47) and Biocarta (47), is specifically up-regulated in lung cancer and also associated to lung cancer based on mutations. In addition, this gene is a target of Imatinib, a drug used for treatment of leukemia and gastric cancers (48,49) caused by the bcr-abl1 translocation and by cKit mutations, respectively. In summary, the network highlights two main differential hubs (TP53 and DDR1) and additional connected genes, some of which could be targeted by known cancer drugs.

ALL. In the ALL network (Figure 6B), the largest PPI-based connected component has TP53 as a hub, connected to genes that are specifically up-regulated in ALL such as ATM and TOP2A. An up-regulated sub-module of the network is enriched with to T-cell activation genes ($p = 9.2E{-}7$), which were not found to be associated with leukemia according to COSMIC. However, some of the genes are targets of well known drugs of leukemia sub-diseases, such as ADA (Pentostatin, inhibition - lymphoproliferative malignancies) and LCK (Dasatinib and Ponatinib - chronic myeloid leukemia, ALL) (50–53). Interestingly, NR3C1, a glucocorticoid receptor transcription factor that promotes inflammatory responses, has high degree and is also connected to TP53. This gene is a target of 39 drugs, including both agonists and antagonists (32). In summary, the network reveals two main functional areas in the PPI network: the module surrounding the TP53 hub, and the T-cell sub-module. Both are differential in the disease. In addition, the network captures known related genes and treatments.

Colorectal cancer. As the initial network was large (see Supplementary Table S3) we focused only on up-regulated genes with PN-ROC > 0.8 (Supplementary Figure S6). The result was 27 genes interconnected by 30 GIs, and only one PPI. All GIs were from (54), representing gene pairs that are expected to share similar biological functions (55). The network is enriched with genes related to detection of mechanical stimulus ($p = 2.11E{-}6$). JUN, the main hub, is related to angiogenesis and to positive regulation of endothelial cell development. The network also contains three druggable genes associated with intestinal cancer based on the mutation data: SLC12A2, GABBR1, and CACNA1D. Interestingly, the drugs that target these genes are not known cancer drugs. For example, CACNA1D is a target of 13 inhibitory drugs related mainly to hypertension treatment (e.g. Felodipine, Israpidine) (56). In summary, our results

suggest an up-regulated gene module in colorectal cancer and a possible link between colorectal cancer and other factors related to hypertension and psychological stress.

DISCUSSION

In this study, we present a novel approach for producing reliable disease-specific biomarkers that are readily interpretable, especially in terms of their clinical potential. To be able to do this, we first compiled and manually curated a very large collection of gene expression profiles spanning many studies from multiple diseases, called ADEPTUS. Each sample was normalized separately based on its weighted ranks, in order to allow joint analysis of samples from different technologies and studies, at the expense of some loss of information. Importantly, it also allows the use of a biomarker to classify a single new sample. Future studies could apply other non-parametric approaches that process the raw expression data and do not preserve the measured gene ranking, e.g. Barcode (57) or SCAN (58). ADEPTUS can be readily used to test novel multi-label classification algorithms, and it can be utilized alongside other data (expression or other) in future studies.

We utilized the compendium for improved disease classification. In contrast to previous studies, in our analysis the simple single-classifier approach outperformed more sophisticated methods. A possible explanation is that our analysis used fewer labels compared to other studies (since we only addressed diseases with at least five datasets), and therefore had fewer dependencies among them.

A key insight of our study is the risk of misleadingly optimistic performance when classifying multi-disease data. We showed that one must treat the non-disease samples as two distinct categories: negatives (non-disease samples from studies of the same disease) and background controls (samples from studies of other diseases), and evaluate the performance against each subgroup separately. The good classification results validated the approach and the data quality and allowed us to focus subsequent analyses on well-classified diseases. Our method reached substantially higher classification performance than (10) (e.g. 22% improvement in recall). However, performance is not directly comparable because in (10) fewer samples were used, and samples were limited to just two microarray platforms, the classifier did not predict the control class, and more diseases were tested.

Having identified 24 well-classified diseases, we set out to identify disease-specific genes in each of them using the DO structure, the three-way partition of the samples, and meta-analysis significance. This analysis reduced the overlap between gene sets of related diseases. Reassuringly, the discovered gene sets included established disease factors. While we focused on disease-specific genes, future studies could potentially use our database to search for genes with a similar expression pattern across different cancer types.

The issue of robustness in disease biomarker discovery has been troubling the community for quite some time (59–62). It has two aspects: good predictive power when biomarkers from one study are tested on a different cohort from an independent study, and reproducibility of the same biomarker gene set in independent studies. While the predictive power has been typically high, reproducibility remains low. Domany and colleagues estimated that for breast cancer prognosis prediction, thousands of samples will be needed in order to achieve 50% overlap between two such sets (39). Our study sheds additional light on this issue. It shows that reproducibility of the detected biomarkers improves as the number of disease datasets and samples in the training set grows. When the number of datasets available for a disease is at least 10, our analysis produces biomarker sets that are significantly overlapping on disjoint subsets of the data. Using the whole compendium, the expected Jaccard score for overlap is 0.3 ($p < E-250$) for the most represented disease category. In fact, with over 4200 samples for the organ systems cancer category, robustness is less than predicted by the model of (39). This can be attributed in part to factors that were not taken into account in that model, e.g. batch effects of different studies and technologies. Overall, our results imply that in order to further improve robustness and reproducibility, future studies should aim to increase the number of datasets and samples, while making judicious use of data on other diseases to guarantee specificity.

The final step of our approach involved integration of our results with information from external databases: somatic mutations in cancer, drug–gene associations, and protein interactions. For each tested disease, we summarized all this information and our results in a network. These networks provide a bird's eye view of the disease-specific genes, their relations and properties, and thus point to new therapeutic potential. Such an overview can serve as a starting point for considering novel therapeutics, such as drug repositioning that exploits approved genes for new treatments, or multi-drug treatments, in which several drugs are used to target different aspects of the biological network.

While our approach is effective, it has several limitations that future studies can address. We tested only 48 diseases since we included in the compendium only diseases that had at least five datasets with at least 20 samples each, in order to allow reliable cross validation on whole datasets. In addition, we analyzed only ∼15 000 gene expression profiles, a modest fraction of the human profiles in GEO, since we required manual curation of the disease terms for each profile (automatic curation had unsatisfactory quality). We view our work as a proof of concept: with some more effort of a team of curators, all available large databases can be curated and the same methodology can be applied for their analysis. Second, our multi-platform integration proved beneficial for half of the tested diseases, and most well-classified diseases were related to cancer. Nevertheless, neurodegenerative disorders and cardiovascular disease were well classified as well. In addition, we showed that narrowing down the analysis to a single platform can improve the performance in other disease terms. The low performance in some of the diseases could be due to several reasons: (i) low number of non-cancer datasets, (ii) integration of a large number of platforms, (iii) limitations of using methods that rank genes by their expression levels, (iv) inexistence of gene expression based robust classifier and (v) the tested disease might be too broadly defined (e.g. 'disease of anatomical entity').

AVAILABILITY

The data used in this study and the R code developed are available at http://acgt.cs.tau.ac.il/adeptus, together with a tutorial for using them.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Author Contribution: D.A. and R.S. conceived the study. D.A. and T.H. assembled the database. D.A., T.H., S.I. and R.S. analyzed the data. D.A., S.I. and R.S. wrote the paper.

FUNDING

Israel Science Foundation [317/13]; IDEA grant from the Dotan Center in Hemato-Oncology; D.A. is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship; Edmond J. Safra Center for Bioinformatics at Tel Aviv University (to D.A.); Israeli Center of Research Excellence (I-CORE), Gene Regulation in Complex Human Disease, Center No. 41/11. Funding for open access charge: Israeli Center of Research Excellence (I-CORE), Gene Regulation in Complex Human Disease, Center No. 41/11; IDEA grant from the Dotan Center in Hemato-Oncology.
Conflict of interest statement. None declared.

REFERENCES

1. Lavi,O., Dror,G. and Shamir,R. (2012) Network-induced classification kernels for gene expression profile analysis. *J. Comput. Biol.*, **19**, 694–709.
2. Yang,X., Regan,K., Huang,Y., Zhang,Q., Li,J., Seiwert,T.Y., Cohen,E.E., Xing,H.R. and Lussier,Y.A. (2012) Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Comput. Biol.*, **8**, e1002350.
3. Lee,E., Chuang,H.Y., Kim,J.W., Ideker,T. and Lee,D. (2008) Inferring pathway activity toward precise disease classification. *PLoS Computat. Biol.*, **4**, e1000217.
4. Altschuler,G.M., Hofmann,O., Kalatskaya,I., Payne,R., Sui,S.J.H., Saxena,U., Krivtsov,A.V., Armstrong,S.A., Cai,T.X., Stein,L. *et al.* (2013) Pathprinting: An integrative approach to understand the functional basis of disease. *Genome Med.*, **5**, 68.
5. Xu,M., Kao,M.C.J., Nunez-Iglesias,J., Nevins,J.R., West,M. and Zhou,X.J. (2008) An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics*, **9(Suppl. 1)**, S12.
6. Ulitsky,I., Krishnamurthy,A., Karp,R.M. and Shamir,R. (2010) DEGAS de novo discovery of dysregulated pathways in human diseases. *PLoS One*, **5**, e13367.
7. Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.
8. Rung,J. and Brazma,A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.
9. Ioannidis,J.P., Allison,D.B., Ball,C.A., Coulibaly,I., Cui,X., Culhane,A.C., Falchi,M., Furlanello,C., Game,L., Jurman,G. *et al.* (2009) Repeatability of published microarray gene expression analyses. *Nat. Genet.*, **41**, 149–155.
10. Huang,H., Liu,C.C. and Zhou,X.J. (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6823–6828.
11. Bodenreider,O., Nelson,S.J., Hole,W.T. and Chang,H.F. (1998) Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc. Annu. Symp. AMIA*, 815–819.
12. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
13. Schmid,P.R., Palmer,N.P., Kohane,I.S. and Berger,B. (2012) Making sense out of massive data by going beyond differential expression. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5594–5599.
14. Lee,Y.S., Krishnan,A., Zhu,Q. and Troyanskaya,O.G. (2013) Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, **29**, 3036–3044.
15. Papachristoudis,G., Diplaris,S. and Mitkas,P.A. (2010) SoFoCles: feature filtering for microarray classification based on gene ontology. *J. Biomed. Inform.*, **43**, 1–14.
16. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
17. Staiger,C., Cadot,S., Gyorffy,B., Wessels,L.F. and Klau,G.W. (2013) Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.*, **4**, 289.
18. Ciriello,G., Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127-U1247.
19. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
20. Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
21. Yang,X., Bentink,S., Scheid,S. and Spang,R. (2006) Similarities of ordered gene lists. *J. Bioinform. Computat. Biol.*, **4**, 693–708.
22. Zhang,M.L. and Zhou,Z.H. (2014) A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data En.*, **26**, 1819–1837.
23. Tsoumakas,G., Spyromitros-Xioufis,E., Vilcek,J. and Vlahavas,I. (2011) MULAN: a java library for multi-label learning. *J. Mach. Learn. Res.*, **12**, 2411–2414.
24. Madjarov,G., Kocev,D., Gjorgjevikj,D. and Dzeroski,S. (2012) An extensive experimental comparison of methods for multi-label learning. *Pattern Recognit.*, **45**, 3084–3104.
25. Vens,C., Struyf,J., Schietgat,L., Dzeroski,S. and Blockeel,H. (2008) Decision trees for hierarchical multi-label classification. *Mach. Learn.*, **73**, 185–214.
26. Sucar,L.E., Bielza,C., Morales,E.F., Hernandez-Leal,P., Zaragoza,J.H. and Larranaga,P. (2014) Multi-label classification with Bayesian network-based chain classifiers. *Pattern Recognit. Lett.*, **41**, 14–22.
27. Barutcuoglu,Z., Schapire,R.E. and Troyanskaya,O.G. (2006) Hierarchical multi-label prediction of gene function. *Bioinformatics*, **22**, 830–836.
28. Scholkopf,B.S.P., Smola,A. and Vapnik,V. (1998) Prior knowledge in support vector kernels. *Advances in Neural information processings systems*, **10**, 640–646.
29. Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, NY.
30. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
31. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M.M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
32. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y.F., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
33. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

34. Morris,J.H., Kuchinsky,A., Ferrin,T.E. and Pico,A.R. (2014) enhancedGraphics: a Cytoscape app for enhanced node graphics. *F1000Research*, **3**, 147.
35. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
36. Montojo,J., Zuberi,K., Rodriguez,H., Kazi,F., Wright,G., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, **26**, 2927–2928.
37. Shamir,R., Ulitsky,I., Maron-Katz,A., Shavit,S., Sagir,D., Linhart,C., Elkon,R., Tanay,A., Sharan,R. and Shiloh,Y. (2010) Expander: from expression microarrays to networks and functions. *Nature Protoc.*, **5**, 303–322.
38. Hedges,L.V. and Olkin,I. (1985) *Statistical Methods for Meta-analysis.* Academic Press, Orlando.
39. Ein-Dor,L., Zuk,O. and Domany,E. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 5923–5928.
40. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
41. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
42. Hall,A. (2009) The cytoskeleton and cancer. *Cancer Metast. Rev.*, **28**, 5–14.
43. Hilton,H.N., Stanford,P.M., Harris,J., Oakes,S.R., Kaplan,W., Daly,R.J. and Ormandy,C.J. (2008) KIBRA interacts with discoidin domain receptor 1 to modulate collagen-induced signalling. *BBA-Mol. Cell Res.*, **1783**, 383–393.
44. Xu,H., Raynal,N., Stathopoulos,S., Myllyharju,J., Farndale,R.W. and Leitinger,B. (2011) Collagen binding specificity of the discoidin domain receptors: binding sites on collagens II and III and molecular determinants for collagen IV recognition by DDR1. *Matrix Biol.*, **30**, 16–26.
45. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
46. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
47. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
48. Benjamin,R.S., Blanke,C.D., Blay,J.Y., Bonvalot,S. and Eisenberg,B. (2006) Management of gastrointestinal stromal tumors in the imatinib era: Selected case studies. *Oncologist*, **11**, 9–20.
49. Vigneri,P. and Wang,J.Y.J. (2001) Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. *Nat. Med.*, **7**, 228–234.
50. Chen,X., Ji,Z.L. and Chen,Y.Z. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res.*, **30**, 412–415.
51. Zhu,F., Han,B.C., Kumar,P., Liu,X.H., Ma,X.H., Wei,X.N., Huang,L., Guo,Y.F., Han,L.Y., Zheng,C.J. *et al.* (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.*, **38**, D787–D791.
52. Lindauer,M. and Hochhaus,A. (2010) Dasatinib. *Rec. Results Cancer Res.*, **184**, 83–102.
53. Jackson,R.C., Leopold,W.R. and Ross,D.A. (1986) The biochemical pharmacology of (2′-R)-chloropentostatin, a novel inhibitor of adenosine-deaminase. *Adv. Enzyme Regul.*, **25**, 125–139.
54. Lin,A., Wang,R.T., Ahn,S., Park,C.C. and Smith,D.J. (2010) A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res.*, **20**, 1122–1132.
55. Baryshnikova,A., Costanzo,M., Kim,Y., Ding,H.M., Koh,J., Toufighi,K., Youn,J.Y., Ou,J.W., San Luis,B.J., Bandyopadhyay,S. *et al.* (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods*, **7**, 1017-U1110.
56. Sinnegger-Brauns,M.J., Huber,I.G., Koschak,A., Wild,C., Obermair,G.J., Einzinger,U., Hoda,J.C., Sartori,S.B. and Striessnig,J. (2009) Expression and 1,4-dihydropyridine-binding properties of brain L-type calcium channel isoforms. *Mol. Pharmacol.*, **75**, 407–414.
57. McCall,M.N., Jaffee,H.A., Zelisko,S.J., Sinha,N., Hooiveld,G., Irizarry,R.A. and Zilliox,M.J. (2014) The Gene Expression Barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.*, **42**, D938–D943.
58. Piccolo,S.R., Withers,M.R., Francis,O.E., Bild,A.H. and Johnson,W.E. (2013) Multiplatform single-sample estimates of transcriptional activation. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17778–17783.
59. Chuang,H.Y., Lee,E., Liu,Y.T., Lee,D. and Ideker,T. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.
60. Kim,K., Zakharkin,S.O. and Allison,D.B. (2010) Expectations, validity, and reality in gene expression profiling. *J. Clin. Epidemiol.*, **63**, 950–959.
61. Ein-Dor,L., Kela,I., Getz,G., Givol,D. and Domany,E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.
62. Michiels,S., Koscielny,S. and Hill,C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

5. Discussion

In this thesis we described our contributions to integrative analysis of heterogeneous biological data. We introduced two algorithms for discovery of meaningful modules. The first, ModMap, detects gene modules and links between them in a pair of biological networks. The second algorithm, TWIGS, detects flexible modules that reappear across different subjects for which time series data were measured. In both cases we compared the algorithms to the state of the art and showed a very significant improvement. In addition, we applied each of our algorithms to a wide variety of different biological applications. Next, we described our ADEPTUS study in which we extracted clinically meaningful and reliable disease biomarkers by analyzing more than 14,500 gene expression profiles from more than 180 studies, and somatic mutation data from more than 30,000 patients. All algorithms presented in this thesis were implemented (in R or Java) and made freely available for the community. All analyzed datasets, including the ADEPTUS database, were made available as well.

5.1 Network data analysis and module discovery

5.1.1 The ModMap algorithm

In network biology the key task is to learn interrelations among different components of the biological system. Learning such interactions is essential since any biological molecule (e.g., gene, protein, etc.) is a part of a complex system and does not act alone. The ModMap algorithm, described in Chapter 2, aims for learning a set of gene modules, which represent very basic functional units, and the links between them. This is achieved by simultaneous analysis of two conceptually different networks. The first network contains "positive" edges: these are gene pairs that are expected to work together in the same processes. The second network contains "negative" edges: gene pairs that are expected to work in parallel or in compensatory pathways. The output of ModMap is a

set of gene groups (the modules) and a set of group pairs (links the map). Each module represents a set of genes that are highly connected by positive edges. Each module link represents two gene sets that are highly connected by negative edges.

The ModMap algorithm is based on a global objective function that uses both the within- and between- module scores. A link between two modules is added to the map if and only if they are significantly interconnected in the negative network. This constraint can be a major hurdle when trying to analyze large networks, as testing for significant connectivity between all module pairs could be quite costly in terms of running time. To overcome this difficulty, our algorithm starts with a given initial solution and improves it using an iterative process. It works in a way that quickly updates the solution in each round mainly by merging module pairs. We used two major techniques to allow analysis of large networks in a reasonable running time. First, we keep track of module pairs that cannot induce a link in the map due to insignificant connectivity between them. In subsequent improvements such a module pair (and newly formed modules pairs that contain its members) is never considered as a possible link in the map. We therefore avoid the need to recalculate the connectivity significance among most module pairs. Second, we perform multiple improvement steps in parallel. Our heuristic is guaranteed to improve the global score of a given solution (as long as new merging options are found), which also guarantees convergence (see Supplementary Text in (1) for details and a proof). Using our new algorithm we were able to analyze large networks of more than 5,000 genes. Previous studies analyzed roughly 1,500 genes or less.

We compared ModMap to extant methods on both simulated and real data and showed that it improves upon the other methods. A major advantage of our method is that it uses information on gene pairs and is not dependent on the technology used to generate the network. This allowed us to: (1) use complete interactomes, which contain information from many studies that used different technologies, and (2) analyze three very different data types: (i) PPIs and GIs, (ii) PPIs and dynamic GIs, and (iii) differential coexpression data. We showed that the joint analysis was crucial for detecting patterns that were not detected when each network was analyzed separately. For example, the

proteasome complex genes form a large highly connected subgraph in the PPI network. However, when we jointly analyzed PPIs and negative GIs we found smaller, yet highly functionally relevant gene modules: ModMap correctly (and perfectly) partitioned the proteasome complex into its two sub-complexes: the core and accessory parts (see Figure3 in (1)). While there are many negative GI edges between these two sub-complexes, many genes have no outgoing GI edges that link them to the other sub-complex. Thus, analyzing the GI network separately could not fully recover these two complexes. In summary, ModMap recovered the basic functional subunits of the proteasome complex, which could not be found by analyzing one of the networks separately.

In differential coexpression (DC) analysis we showed that ModMap can markedly improve the number of genes covered by a solution while keeping the detected DC level very high. Recent studies suggested new methods to better quantify the DC level between a pair of genes (103). ModMap can be applied on the output of these methods (i.e., as the negative network) for summarizing the results and detecting gene modules. While these studies looked at the DC of single gene pairs or major gene hubs that showed DC with many other genes, we think that ModMap can give a more systems level interpretation.

In our lung cancer example, we have demonstrated how DC between gene modules can highlight possible dysregulation via miRNA activity. While this analysis was given as an anecdotal example in the ModMap paper, we have demonstrated in a previous work that DC analysis performs much better than other gene expression analyses in detecting modules that are linked to miRNA activity (28). In this work, we focused on algorithmic contributions by showing that ModMap outperformed our previous algorithms both on simulated and real data.

Our approach has two main limitations that are common to most module finding and clustering paradigms: the edge independence assumption, and application-specific manual parameter tuning. In ModMap, we assumed that edges are drawn independently given the degree of each node. While this assumption is naive, modeling edge dependency relies on the assumptions made regarding the null space of the graphs, which

makes any solution application-specific. For example, in a recent work we used an MCMC approach to mimic the expected null space of brain-related networks (104). The downside of this approach is that it is very slow even though it only calculates significance among known modules (i.e., it does not search for patterns in the networks). Thus, it currently cannot be used within ModMap as an alternative for scoring module links. As for manual parameter tuning, we used it both to adjust the parameters when we analyzed relatively small networks (e.g., Figure 4), and to interpret the module maps. For the former we believe that we give reasonable default values from which the user can start. However, for the latter, manual tuning is inevitable, as the goal of ModMap is to give an overview of the analyzed networks in order to help the researcher interpret the data.

5.1.2 The TWIGS algorithm

In Chapter 3 we described TWIGS, an algorithm for detecting flexible modules in three-way time series data. We expect that as technology prices decrease, time series datasets that monitor the state of multiple patients over time will become much more prominent. Thus, computational tools that can exploit these rich data and find information that is shared by different subjects will be needed. In this respect, TWIGS can be viewed as a tool for detecting replicable patterns across patients, which is a basic task that should be solved when moving from analysis of a single patient to analyzing many patients.

In our experiments we observed that available methods for module discovery define patterns that are too rigid. For example, triclustering enforces synchronous response across subjects. While this approach is suitable for data from biological repeats or from well tailored experiments, it does not fit the paradigm of monitoring response in patients over time. As another example, methods that seek biclusters of subjects and genes (or voxels) can suffer from low coverage of the subjects because they require that all genes of the bicluster will be relevant in each of the subjects.

To cope with such problems we suggested a novel definition of a module in three-way datasets. The components of the module are: (1) a set of subjects, (2) a set of genes that constitute the *core* of the module, (3) a gene set for each of the subjects; these include the genes in the core with some high probability, and possibly additional genes, and (4) a set of relevant time points for each subject. We suggested a hierarchical Bayesian generative model for explaining a dataset that contains such a module. Our hierarchical Bayesian model implicitly defines the underlying global scoring function of a candidate module. We detect modules by starting from an initial solution of a biclustering algorithm and improving the results using Gibbs sampling. Multiple modules are detected by rerunning our algorithm on the data after removing the effect of the modules that were detected in previous runs. The algorithm stops when it cannot find an additional module. Finally, we used advanced domain-specific downstream analyses to interpret the results. For example, in the gene expression case we used enrichment analysis and network visualization to stratify the patients based on their detected highly enriched pathways.

The advantages of our approach are: (1) we allow asynchronicity of the biological response across subjects while explicitly modeling the time response, (2) we allow a subject-specific added signal (e.g., an additional activated pathway that is added to the module in a specific subject), (3) some of the core module genes could be missing in any specific subject, and (4) our algorithm for detecting multiple modules can detect a fuzzy solution: a subject can be in more than one module. We applied TWIGS to many different simulated datasets, gene expression data, and fMRI data. In all cases we showed that TWIGS markedly outperforms extant methods.

A detailed analysis of the detected modules indicated that the ability to detect subject-specific information was crucial both for getting better results, and for identifying novel biological insights. Specifically, it led to >2-fold improvement in subject coverage as compared to extant methods, which is a major advantage when analyzing large datasets. In the gene expression data analysis TWIGS outperformed all methods in terms of enrichment analysis. The two detected modules were fuzzy and the core gene sets were

enriched with different immune response pathways. The subject specific information revealed that some of the patients that did not survive the septic shock manifested an increasing number of up-regulated pathways over time. Thus, this module represents a pattern that can be used to detect patients at immediate risk. In the fMRI case TWIGS recovered the main brain networks implicated in rest. Unlike the gene expression case, the extent of the subject specific additional information was mild. Nevertheless, the detected subject-specific information contained interesting patterns. For example, four subjects were pointed out as much more attentive to the situation of staying in the scanner than others. This local signal was not detected by the original studies, nor by the extant methods.

To the best of our knowledge, our definition of a flexible module is the first of its kind. We therefore expect that multiple improvements could be suggested in the future. For example, additional information such as PPI data could be used to improve PPI connectivity of the detected core modules. In addition we used a simple two group model for modeling the likelihood of the data. Future works could use more complex methods. The Markovian window model of time point selection could be replaced by others, e.g., allowing independent subset of time points to accommodate non-sequential profiles of perturbations. Finally, we used a Gibbs sampler to obtain modules, and better algorithms could be suggested. Future studies can tackle the problem in a combinatorial fashion, or suggest a model that allows simultaneous detection of more than a single module.

5.1.3 Characterization of the algorithms

Both problems mentioned above (i.e., the problems studied in Chapters 2 and 3) can be considered as generalizations of biclustering. In the module map problem ignoring the positive network (i.e., by setting all gene pairs as edges with the same positive score) results in a biclustering problem, whereas ignoring the negative network is similar to standard clustering of the nodes. Our flexible module discovery problem for three-way datasets can be reduced to a biclustering problem when there is only a single subject being studied.

Both studies above were designed similarly. We started with a thorough simulation study. We generated both discrete and continuous simulated data in which we planted modules. We then added different noise levels to create different scenarios. We tested a wide range of possible noise levels, some of which are beyond those expected to be in biological data. Using the simulated data we tested the performance of a large set of both extant and novel algorithms. The top selected algorithms, TWIGS and ModMap, were much better than the other alternatives. We then applied all algorithms to real data where TWIGS and ModMap remained the top algorithms in most tests.

These algorithms share some technical similarities, even though they address different problems. First, both algorithms utilize standard biclustering algorithms to obtain an initial solution. These initial solutions are far from the desired output, but they provide a good starting point that is much better than selecting a random point in the search space. Moreover, in both cases the biclustering algorithms that were eventually selected are based on an exhaustive search for perfect bicliques in the underlying graph. Furthermore, several heuristics were added in both cases in order to avoid spending too much time for obtaining the initial solution. Second, both algorithms perform iterative improvement steps that aim to maximize a global score. In ModMap this is done explicitly by selecting only improvement steps that lead to a better score. In TWIGS this is done implicitly by the Gibbs sampler. That is, we try to achieve a solution with a high posterior probability under the hierarchical model.

In summary, we presented a general paradigm for developing algorithms that integrate multiple data sources for finding meaningful modules. Instead of "reinventing the wheel", we first carefully utilize existing algorithms to obtain an initial solution. Thereby, we make use of the rich knowledge accumulated by the community over the recent decades. While extant methods were very beneficial for getting an initial solution, it was essential to develop an additional algorithm that simultaneously analyzes all data sources in order to obtain a final output of the desired quality.

5.2 Integrative analysis of many expression studies

Improving reusability of published biological data is a major effort whose success could lead to a leap in our ability to understand biological information. Although new exciting technologies are being introduced almost every year, it takes time for the community to garner a sufficiently large number of samples using the new technologies. On the other hand, for well established technologies, e.g., microarrays or RNA-Seq for mRNA quantification, the number of samples in the public databases is vast (especially for microarrays). Therefore, we sought a novel methodology for large scale studies that aim to exploit these data. In Chapter 4 we described how researchers can utilize such data for detecting reliable gene expression-based biomarkers, and integrate them with other gene-based information sources. To reach this goal we presented a methodology that contains four major steps.

1. *Collection and standardization*. We collect many gene expression profiles from multiple studies. For each sample public databases provide a vector of probe or gene expression intensities, and a textual description of the phenotype. To transform all expression profiles into a common ground we calculated a rank-based score for each gene (which required transforming probe level data of microarrays into gene level intensity). In addition, for each sample we manually annotated its phenotype description and assigned it a set of disease ontology terms. In our study we also tested an automatic annotation approach. However, although the automatic mapping was far better than random, it was of unsatisfactory quality for our subsequent analyses.

2. *Multi-label learning*. We utilized multi-label learning algorithms in order to validate our database. This step was crucial for selecting well classified diseases for which a reliable gene expression-based, cross-technology, biomarker can be learned. We observed that unlike previous studies using complex multi-label algorithms (e.g., Bayesian Correction (80)) had a negligible advantage over learning an independent binary classifier for each disease term. This probably occurs because the other studies analyzed more disease terms (>100), and thus a more delicate estimation of the dependencies among the disease terms was required there.

A key insight of our study is that when evaluating the performance of a specific disease term, using standard performance scores (e.g., ROC scores) might lead to over-estimation. The reason is that these scores ignore the complexity of the non-positive cases (i.e., samples that do not have the disease). This group can be divided into two: negatives, which are direct controls from the same studies as the cases, and background controls, which are all others. When evaluating a classifier of a disease, one must make sure that it can differentiate between positives and negatives *and* between positives and controls. In fact, more than 20 of the tested disease terms had poor results when we applied this simple methodology, including well studied diseases such as diabetes. There can be several reasons for this low performance. Our rank-based preprocessing probably leads to some loss of information that impairs the ability to learn classifiers for these diseases. To illustrate this we showed that when the analysis is limited to a single technology (so there is no need for our normalization) additional disease terms become well-classified. Still, roughly a third of the tested disease terms remain poorly classified. Additional reasons might explain this low performance: inexistence of a good gene expression-based classifier, and a definition of some disease terms that is too broad (e.g., "disease of anatomical entity").

3. *Finding disease-specific differential genes*. Our next goal was to identify disease-specific differential genes for the selected well-classified diseases. When analyzing if a gene is differential in a disease, we corrected both for biological and disease ontology-related artifacts. Correction for biological covariates was done similarly to the way we tested if a classifier performs well: we tested if the gene is differentially expressed between the positives and the negatives and between the positives and the background controls. Note that the negatives are direct controls of the positives as they probably have similar phenotypic background (e.g., same tissue, age, etc.). In addition, we looked only for genes that exhibit differential expression in both of cases above. While other interesting patterns could be present in the data, we discovered relatively large disease-specific gene sets, which were highly enriched for known relevant pathways. For example, in cancer the discovered gene set induced two highly connected

PPI active modules (see Figure 5 in (3)) that contain the main genes related to the hallmarks of cancer.

Ontology-related artifacts are cases in which a gene is detected as differential in a disease term and is erroneously suggested as differential in one of its ancestors only because the original term has many patients. We avoid these cases by removing the samples of the descendant term and reanalyzing the gene. This simple correction removed most overlaps between child and parent disease terms.

4. *Disease-specific gene-based overview*. In our final step we added multiple non-expression data types in order to improve the interpretability of our discovered gene sets. For a specific disease we created a network whose nodes were the disease-specific differential genes and the gene sets of the ancestral disease terms (e.g., for lung cancer we took the lung cancer genes and the general cancer genes). We used both PPIs and positive GIs as edges that indicate genes that are expected to work together. For each gene we indicate (i) whether it is up- or down- regulated, (ii) whether it is associated to the disease according to the COSMIC database (105), and (iii) whether there are known drugs that could target it. The joint visualization of all these data manifests the major active modules of each of the analyzed diseases (lung cancer, colon cancer, and ALL), pointed out some of the major genes, and even suggested novel candidates for drug repurposing.

In this study, we created an analysis procedure that starts from publicly available data and results in a biologically and clinically interpretable disease summary. We analyzed, in total, 13,314 microarray profiles, 1,526 RNA-Seq profiles, and somatic mutation data from more than 30,000 COSMIC samples. Note that other recent studies that utilized many expression profiles were able to analyze a much larger set of profiles because they were either not limited to human data (e.g., [68], >60,000 microarrays), or they did not require high quality phenotypic annotation (e.g., (106)).

Nevertheless, in this study we analyzed 174 datasets from the Gene Expression Omnibus (GEO) (10). A subsection of GEO contains datasets that were annotated and analyzed by their own team (i.e., datasets with a GDS id). When we searched for annotated datasets that satisfy our criteria for inclusion in the ADEPTUS database (e.g.,

human samples, ≥20 samples in the study, etc.) only 320 studies were detected of which only 254 are of the same platforms covered in our study. Hence, our manually curated collection captures in a single study some 70% of the relevant data collected by the GEO curators over more than a decade. We also note that for seven datasets containing over 1,500 samples that we collected, the GEO description was not detailed enough to assign DO ids, and when we contacted the authors of the publications we got no response. So the numbers of useful GEO samples are effectively even smaller. Finally, note that the previous supervised multi-disease studies covered less samples and platforms, although some of them used automated annotation.

5.3 Future research

The studies in this thesis can be used as a basis for additional research directions. Above, we already outlined some improvements that could be applied to our models and algorithms. Below we list additional research directions.

1. *Module map extensions*

1.1 Module maps with sparse modules. Our current algorithm for module-map discovery outperforms extant methods, but it cannot handle well sparse modules. Sparse modules may represent signaling pathways, in which elements may be connected in a serial fashion and have relatively low degrees. One straightforward option is to use algorithms that analyze weighted graphs by reducing the penalty for non-edges (or low scored edges) in the positive network. For example, consider a PPI network and give a weight of 1 to each edge, and a weight of $-\alpha$ ($0<\alpha<1$) to each node pair that is not an edge. As a result, for small α a module can achieve a positive score even if most pairs within it are not connected. We expect that when applied to PPI and GI networks, such analysis may better detect GI links between large pathways.

1.2 Integrated analysis of miRNA expression profiles and DC networks. In our work we have shown how module maps can detect gene modules that are highly enriched with

miRNA targets. This was achieved without using any auxiliary information on the expression profiles of the miRNA themselves. As these data are becoming more common, a possible extension of ModMap can use the miRNA expression profiles to guide the search process of the DC modules. We expect that such algorithms may be more powerful in detecting the complete landscape of miRNA-mediated dysregulation processes.

2. *Large scale multi-label analysis of additional molecular profiles*.

Our methodology for analyzing many datasets together could be easily extended to non-expression molecular profiles. Two possible data sources are methylation profiles (7) and mutation profiles (8). Creating a database for our multi-label analysis flow requires three components for each sample: (1) a vector of scores for each measured object (i.e., gene, exon, promoter, etc.), (2) a set of DO ids that explain its phenotype, and (3) the original study id. Fortunately, for cancer somatic mutation data public databases provide many well annotated samples. For example, COSMIC (105) provides binary associations between genes and samples (the gene scores), a relatively well defined vocabulary that explains the patient cancer type in terms of site and histology, and additional information from the original studies. Our preliminary results show that mapping the COSMIC phenotypic annotation is relatively simple (in fact, much simpler than for GEO), and that multi-label analysis is successful for roughly 60% of the tested disease terms.

3. *Improving ADEPTUS and adding a new user interface.*

3.1 An extended database. Our ADEPTUS database could be extended to contain many more gene expression profiles. First, the TCGA (12) provides >2,500 gene expression profiles that were not used in our study. In addition, any microarray dataset from the popular Affymetrix and Illumina platforms could be easily added as long as the DO ids of the samples are known. With a sufficient number of additional datasets a larger set of diseases could be studied. If the number of diseases increases markedly (e.g., to >100) it might be worthwhile to rerun the comparison of the multi-label classification algorithms.

Such analysis could reveal whether advanced multi-label algorithms can improve the classification performance.

3.2 Better user interface. In addition, our current website is not interactive and a user cannot instantly analyze a set of genes of interest. Therefore, a we think that implementing a web-server that allows users to analyze either a disease or a gene set to automatically create gene-based overviews (e.g., Figure 6 in Chapter 4) will be a very useful tool for the community. Using such maps for visualization is not limited to our case only. Similar gene-based overviews can also be created to represent the molecular alterations in a single patient.

3.3 Drug repositioning. In our work, we have shown how additional data can be integrated with our results to suggest new possibilities for drug repositioning. This field has received much attention in recent years (107). Standard methods use drug and disease features to predict new repositioning options (see (108) for example). An interesting venue for research is to integrate such machine learning approaches with our analysis in order to seek new hypotheses that cannot be detected by looking only on drug-drug, disease-disease, or drug-disease interactions. A major hurdle to such venture is the shortage of public data on drug repositioning outcomes, which is needed to validate algorithms.

4. *General methods for advanced meta-analysis*.

An interesting computational and statistical problem is dealing with a matrix of p-values. For example, in Chapter 4, we came across such a matrix: each row denoted a gene and each column denoted a dataset. The p-value in a cell i,j indicated whether the gene i was differential in study j. Of course, all p-values were calculated in order to test the same biological question (e.g., test if a gene is differential in neurodegenerative diseases). In our study we used standard meta-analysis that merges all p-values of a gene. However, the underlying null hypothesis of such analyses is that the gene has no effect in each of the studies, and the calculated statistic is based on averaging the log of the p-values. Thus, such methods are very sensitive to outliers, e.g., genes that were assigned

extremely low p-values in only a single study. In addition, these methods do not model the contribution of single datasets (e.g., all datasets get the same weight when the p-values are merged). We think that modeling the gene effect and the dataset effect explicitly can provide better methods for ranking genes by their relevance to the tested hypothesis. Such methods could be used to obtain robust disease biomarkers.

References

1. Amar D, Shamir R. Constructing module maps for integrated analysis of heterogeneous biological networks. Nucleic Acids Res. 2014;42:4208–19.

2. Amar D, Yekutieli D, Maron-Katz a., Hendler T, Shamir R. A hierarchical Bayesian model for flexible module discovery in three-way time-series data. Bioinformatics. 2015;31:i17–26.

3. Amar D, Hait T, Izraeli S, Shamir R. Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. Nucleic Acids Res. 2015;43:7779–89.

4. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16:85–97.

5. Liu Z, Zhang S. Toward a systematic understanding of cancers: a survey of the pan-cancer study. Front Genet. 2014;5:194.

6. Zuo T, Tycko B, Liu T-M, Lin J-JL, Huang TH-M. Methods in DNA methylation profiling. Epigenomics. 2009;1:331–45.

7. Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, Yau P, et al. Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Res. 2006;34:528–42.

8. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.

9. Leidner RS, Li L, Thompson CL. Dampening Enthusiasm for Circulating MicroRNA in Breast Cancer. PLoS One. 2013;8:1–11.

10. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.

11. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, et al. ArrayExpress update - From an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009;37:D868–72.

12. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

13. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. Proc IEEE Comput Soc Bioinform Conf. 2002;1:197–206.

14. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.

15. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3:13.

16. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM. Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics. 2006;7:177.

17. Pandey G, Myers CL, Kumar V. Incorporating functional inter-relationships into protein function prediction algorithms. BMC Bioinformatics. 2009;10:142.

18. Kourmpetis YAI, Van Dijk ADJ, Bink MCAM, Van Ham RCHJ, Ter Braak CJF. Bayesian markov random field analysis for protein function prediction based on network data. PLoS One. 2010;5.

19. Vlasblom J, Zuberi K, Rodriguez H, Arnold R, Gagarinova A, Deineko V, et al. Novel function discovery with GeneMANIA: a new integrated resource for gene function prediction in Escherichia coli. Bioinformatics. 2014;1–5.

20. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, et al. GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop. Bioinformatics. 2010;26:2927–8.

21. Tzfadia O, Amar D, Bradbury LMT, Wurtzel ET, Shamir R. The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways. Plant Cell. 2012;24:4389–406.

22. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nat Rev Genet. 2007;8:437–49.

23. Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, Toufighi K, et al. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods. 2010;7:1017–24.

24. Fievet BT, Rodriguez J, Naganathan S, Lee C, Zeiser E, Ishidate T, et al. Systematic genetic interaction screens uncover cell polarity regulators and functional redundancy. Nat Cell Biol. 2013;15:103–12.

25. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302:249–55.

26. De la Fuente A. From “differential expression” to “differential networking” - identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010;26:326–33.

27. Lai Y, Wu B, Chen L, Zhao H. A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics. 2004;20:3146–55.

28. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol. 2013;9:e1002955.

29. Watson M. CoXpress: differential co-expression in gene expression data. BMC Bioinformatics. 2006;7:509.

30. Orth JD, Palsson BØ. Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng. 2010;107:403–12.

31. Kanehisa M, Goto S. KEGG: Kyoto Encyclopaedia of Genes and Genomes. Nucl Acids Res. 2000;28:27–30.

32. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40:D1301–7.

33. Croft D, O’Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:691–7.

34. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, et al. Expander: from expression microarrays to networks and functions. Nat Protoc. 2010;5:303–22.

35. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Res. 2007;17:1537–45.

36. Huang DW, Lempicki RA, Sherman BT. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.

37. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2006;1:107–29.

38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette M a, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

39. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012;8:1–9.

40. Bandyopadhyay S, Mehta M, Kuo D, Sung M-K, Chuang R, Jaehnig EJ, et al. Rewiring of genetic networks in response to DNA damage. Science. 2010;330:1385–9.

41. Guénolé A, Srivas R, Vreeken K, Wang ZZ, Wang S, Krogan NJ, et al. Dissection of DNA Damage Responses Using Multiconditional Genetic Interaction Maps. Mol Cell. 2013;49:346–58.

42. Lundin C, North M, Erixon K, Walters K, Jenssen D, Goldman ASH, et al. Methyl methanesulfonate (MMS) produces heat-labile DNA damage but no detectable in vivo DNA double-strand breaks. Nucleic Acids Res. 2005;33:3799–811.

43. Schulz MH, Devanny WE, Gitter A, Zhong S, Ernst J, Bar-Joseph Z. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. BMC Syst. Biol. 2012. page 104.

44. Parnell GP, Tang BM, Nalos M, Armstrong NJ, Huang SJ, Booth DR, et al. Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. Shock. 2013;40:166–74.

45. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, et al. Identification of Genes Periodically Expressed in the Human Cell cycle and Their Expression in Tumors. Mol Biol Cell. 2002;13:2001–15.

46. Attwell D, Buchan AM, Charpak S, Lauritzen M, Macvicar B a, Newman E a. Glial and neuronal control of brain blood flow. Nature. 2010;468:232–43.

47. Strother SC. Evaluating fMRI preprocessing pipelines. Eng Med Biol Mag IEEE. 2006;25:27–41.

48. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum Brain Mapp. 2012;33:1914–28.

49. Varoquaux G, Craddock RC. Learning and comparing functional connectomes across subjects. Neuroimage. 2013;80:405–15.

50. Crossley N a, Mechelli A, Vértes PE, Winton-Brown TT, Patel AX, Ginestet CE, et al. Cognitive relevance of the community structure of the human brain functional coactivation network. Proc Natl Acad Sci U S A. 2013;110:11583–8.

51. Bullmore E, Sporns O. The economy of brain network organization. Nat Rev Neurosci. 2012;13:336–49.

52. Raichle ME, Mintun M a. Brain work and brain imaging. Annu Rev Neurosci. 2006;29:449–76.

53. Biswal B, Yetkin FZ, Haughton VM, Hyde JS. Functional Connectivity in the Motor Cortex of Resting. Brain. 1995;34:537–41.

54. Van den Heuvel MP, Pol HEH. Exploring the brain network: A review on resting-state fMRI functional connectivity. Eur Neuropsychopharmacol. 2010;20:519–34.

55. Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008;4:e1000217.

56. Lavi O, Dror G, Shamir R. Network-Induced Classification Kernels for Gene Expression Profile Analysis. J Comput Biol. 2012;19:694–709.

57. Xu M, Kao M-CJ, Nunez-Iglesias J, Nevins JR, West M, Zhou XJ. An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics. 2008;9 Suppl 1:S12.

58. Yang X, Regan K, Huang Y, Zhang Q, Li J, Seiwert TY, et al. Single sample expression-anchored mechanisms predict survival in head and neck cancer. PLoS Comput Biol. 2012;8:e1002350.

59. Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14:719–32.

60. Deshpande R, Sharma S, Verfaillie CM, Hu W-S, Myers CL. A scalable approach for discovering conserved active subnetworks across species. PLoS Comput Biol. 2010;6:e1001028.

61. Staiger C, Cadot S, Györffy B, Wessels LF a, Klau GW. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. Front Genet. 2013;4:1–15.

62. Cun Y, Fröhlich H. Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. BMC Bioinformatics. 2012;13:69.

63. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2014;47:106–14.

64. Vandin F, Clay P, Upfal E, Raphael BJ. Discovery of mutated subnetworks associated with clinical data in cancer. Pac Symp Biocomput. 2012;55–66.

65. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10:1108–15.

66. Balbin OA, Prensner JR, Sahu A, Yocum A, Shankar S, Malik R, et al. Reconstructing targetable pathways in lung cancer by integrating diverse omics data. Nat Commun. 2013;4:2617.

67. Papachristoudis G, Diplaris S, Mitkas P a. SoFoCles: feature filtering for microarray classification based on gene ontology. J Biomed Inform. 2010;43:1–14.

68. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013;110:4245–50.

69. Domany E. Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. Cancer Res. 2014;74:4612–21.

70. Huang H, Liu C-C, Zhou XJ. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. Proc Natl Acad Sci USA. 2010;107:6823–8.

71. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al. Disease ontology: A backbone for disease semantic integration. Nucleic Acids Res. 2012;40.

72. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17–21.

73. Barutcuoglu Z, Schapire RE, Troyanskaya OG, Decoro CR. Bayesian Aggregation for Hierarchical Classification. IEEE Int. Conf. Shape Model. 2007 page 1–8.

74. Schmid PR, Palmer NP, Kohane IS, Berger B. Making sense out of massive data by going beyond differential expression. Proc Natl Acad Sci. 2012;109:5594–9.

75. Lee YS, Krishnan A, Zhu Q, Troyanskaya OG. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. Bioinformatics. 2013;29:3036–44.

76. Altschuler GM, Hofmann O, Kalatskaya I, Payne R, Ho Sui SJ, Saxena U, et al. Pathprinting: An integrative approach to understand the functional basis of disease. Genome Med. 2013;5:68.

77. Zaragoza JH, Sucar LE, Morales EF, Bielza C, Larranãga P. Bayesian chain classifiers for multidimensional classification. IJCAI. 2011;2192–7.

78. Vens C, Struyf J, Schietgat L, Džeroski S, Blockeel H. Decision trees for hierarchical multi-label classification. Mach Learn. 2008;73:185–214.

79. Blockeel H, Schietgat L, Struyf J, Džeroski S, Clare A. Decision Trees for Hierarchical Multilabel Classification: A Case Study in Functional Genomics. Lect Notes Comput Sci Incl Subser Lect Notes Artif Intell Lect Notes Bioinforma. 2006;4213:18–29.

80. Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics. 2006;22:830–6.

81. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. 2014. page 1819–37.

82. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17:229–36.

83. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol. 2005;23:561–6.

84. Ulitsky I, Shlomi T, Kupiec M, Shamir R. From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. Mol Syst Biol. 2008;4 :209.

85. Ulitsky I, Shamir R. Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics. 2009;25:1158–64.

86. Narayanan M, Vetta A, Schadt EE, Zhu J. Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Comput Biol. 2010;6:e1000742.

87. Kelley DR, Kingsford C. Extracting between-pathway models from E-MAP interactions using expected graph compression. J Comput Biol. 2011;18:379–90.

88. Leiserson MDM, Tatar D, Cowen LJ, Hescott BJ. Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. J Comput Biol. 2011;18:1399–409.

89. Gallant A, Leiserson MDM, Kachalov M, Cowen LJ, Hescott BJ. Genecentric: a package to uncover graph-theoretic structure in high-throughput epistasis data. BMC Bioinformatics. 2013;14:23.

90. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature. 2007;446:806–10.

91. Ma X, Tarone AM, Li W. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. PLoS One. 2008;3:e1922.

92. Hartigan J. Direct clustering of a data matrix. J Am Stat Assoc. 1972;67:123–9.

93. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: A survey. IEEE-ACM Trans Comput Biol Bioinforma. 2004;1:24–45.

94. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E. Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis. PLoS One. Public Library of Science; 2014;9:e90801.

95. Meng J, Gao SJ, Huang Y. Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. Bioinformatics. 2009;25:1521–7.

96. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E. 2003;67.

97. Waltman P, Kacmarczyk T, Bate AR, Kearns DB, Reiss DJ, Eichenberger P, et al. Multi-species integrative biclustering. Genome Biol. 2010;11:R96.

98. Dede D, Ogul H. A three-way clustering approach to cross-species gene regulation analysis. Innov Intell Syst Appl (INISTA), 2013 IEEE Int Symp. 2013. page 1–5.

99. Mankad S, Michailidis G. Biclustering Three-Dimensional Data Arrays With Plaid Models. J Comput Graph Stat. Taylor and Francis Ltd; 2014;23:943–65.

100. Zhao L, Zaki MJ. TriCluster: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data. Proc 2005 ACM SIGMOD Int Conf Manag data. ACM Press; 2005. page 694–705.

101. Supper J, Strauch M, Wanke D, Harter K, Zell A. EDISA: extracting biclusters from multiple time-series of gene expression profiles. BMC Bioinformatics. 2007;8:334.

102. Wise A, Bar-Joseph Z. SMARTS: reconstructing disease response networks from multiple individuals using time series gene expression data. Bioinformatics. 2015;31:1250–7.

103. Reznik E, Sander C. Extensive Decoupling of Metabolic Genes in Cancer. PLoS Comput Biol. 2015;11:e1004176.

104. Maron-Katz A, Amar D, Simon E Ben, Hendler T, Shamir R. RichMind: A Tool for Improved Inference from Large-Scale Neuroimaging Results. PLoS One. Public Library of Science; 2016;11:1–14.

105. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014;43:D805–11.

106. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra H-J, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015;47:115–26.

107. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Brief Bioinform. 2015;1–11.

108. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: A machine-learning approach through data integration. J Cheminform. 2013;5.

Appendix

Supplementary Material of Chapter 2 (ModMap)

Here we give the supplementary text and figures. For the supplementary tables please go to the online version of the paper, at http://nar.oxfordjournals.org/content/42/7/4208.

Supplementary Text

Initiators

We tested five different initiators. Three are based on previous methods and two are novel. The three extant initiators contained the DICER algorithm (1) and two clustering algorithms. We developed two additional initiators. The first is a modification of the DICER initiator, and is called $DICER_k$. The second utilizes an exhaustive solver for the maximal biclique problem (2,3) and is called MBC-DICER.

The $DICER_k$ initiator

The DICER initiator (1) starts from a positive edge (u,v) in G, and defines two node sets (U,V), where U is the set of (high weight) neighbors of u in H, and V is the set of neighbors of v in H. The goal is to remove nodes from U and V such that the resulting sets will constitute heavy sub-graphs of H and the weight of edges between U and V will be high in G. A simple example is shown in **Supplementary Figure 1**. Nodes that appear both in U and V are removed. In the next step, nodes in U and V are removed if this improves the score of the module map link in G or the module scores in H.

DICER works greedily, by iteratively removing a "bad" node, that is, a node that either has a negative sum of edge weights in H with its own group, or has a negative sum of weights in G with the other group. The total score of a node is the sum of the two scores. Nodes for which both G and H scores are negative are removed first, followed by other bad nodes, sorted by their total score. The process ends when there are no bad nodes. The resulting node sets U' and V' are accepted as modules only if each of them contains at least k nodes. In that case the nodes of U' and V' are removed from the graphs, and the process is repeated until no new module pair is found. In the original DICER algorithm we used k=2. Here we used k=5, which provided better results on real and simulated data (see Results).

The MBC-DICER initiator

We now describe an alternative method for constructing initial node sets U and V. Define an un-weighted graph G'= (V,E') with the same node set as G, and edge $(u,v) \in E'$ if and only if $W_G(u,v)>0$. Two disjoint node sets (U,V) are called *fully connected* or a *biclique* in G' if every u

∈ U is connected to every v ∈ V. A biclique (U,V) is *maximal* if it is not a proper subset of another biclique. We search for maximal bicliques in G' using an exhaustive solver (3), restricting the search to maximal bicliques (U,V) such that |U|≥k and |V|≥k. Each such pair (U,V) is then subjected to the node removal procedure.

Since the number of maximal bicliques can be exponential we use only the first 50,000 discovered bicliques as candidates for the node removal stage of the $DICER_k$ algorithm. Let S be a heap that contains the current set of candidate bicliques. We select the biclique (U,V) in S of maximal size |U|+|V| as the next candidate. The node removal stage produces from (U,V) a module pair (U',V'). If the latter is accepted, we remove the nodes of U' and V' from G' and from all bicliques in S, and remove bicliques whose new size is less than 2k. When S is empty we try to run the solver again. If the solver fails to find additional bicliques then the process is terminated.

Clustering algorithms

We included in our tests two clustering algorithms. Both look for clusters in H and disregard information from G. The first is complete-linkage hierarchical clustering (4). The second, which we call NodeAddition, starts with all nodes as modules, and repeatedly adds a singleton (a module with a single node) to a module if the sum of edge weights between them is the largest among all singleton-module pairs (5). This process is repeated until no singletons remain or until the best sum is negative.

Proof of the guaranteed improvement during the iterations of the global improver

Notations: The input to the problem is a pair of networks $H=(V,E,W_H)$ and $G=(V,F,W_G)$ defined on the same set of vertices. These networks can be weighted or un-weighted. The goal is to find a module-map that summarizes both networks. A *module-map* is a graph F=(M,L) where M is a collection of disjoint node sets, called *modules*, $M=\{M_1,\ldots, M_p\}$, $M_i \subseteq V$, $M_i \cap M_j = \emptyset$, and L is a set of module pairs $\{(U_1,V_1), \ldots, (U_p,V_p)\}$, where each U_i and V_i are in M. These pairs are called the map *links* an express the set of significant links among modules according to some hypothesis testing function. In addition, each module must be linked to at least one other module.

The *global score* of the solution is the total sum W_H of edge weights within each M_i plus the total sum of W_G edge weights between each linked node set:

$$S(M,L) = \sum_i W_H(M_i) + \sum_{k,l \mid (M_k,M_l)\in L} W_G(M_k,M_l)$$

Where $W_H(M_i)$ is the total sum of weights within M_i in H, and $W_G(M_k, M_l)$ is the total sum of weights between M_k and M_l in G. The improvement stage merges a pair of node sets if the merge improves the global score. This process is done greedily: iteratively, the merge that yields the best improvement is performed until no possible merge can improve the global score.

We perform multiple merge steps simultaneously in a single iteration in a way that guarantees that the global score improves. Let L_i be the group of sets linked to M_i. Denote M_{ij} as the set resulting from merging M_i and M_j. Let L_{ij} be the group of sets linked to M_{ij} after performing the merge. Consider a case where two possible merges can improve the global score of a given solution (M,L): M_i with M_j, and M_a with M_b. If there is no overlap between the union of the sets M_i,M_j, L_i, L_j, L_{ij} and the union of the sets M_a,M_b, L_a, L_b, and L_{ab} then we say that $\{M_i,M_j\}$ and $\{M_a,M_b\}$ are *gain-independent*.

Theorem: When two possible merge steps $\{M_i,M_j\}$ and $\{M_a,M_b\}$ are gain-independent, performing both merge operations will improve the global score. Moreover, if the gain of the first merge is g_{ij} and the gain of the second merge is g_{ab} then the gain of performing both merges is at least $g_{ij} + g_{ab}$.

Proof: Let $M=\{M_1,\ldots, M_n\}$ be the partition of the node set before the merge, and let $L=\{(U_1,V_1), \ldots, (U_p,V_p)\}$ be the links, where each U_i and V_i are in M. The global score after merging M_i and M_j can be written as:

$$S_{new} = S(M,L) + W_H(M_i,M_j) - \sum_{M_k \in L_i} W_G(M_i,M_k) - \sum_{M_k \in L_j} W_G(M_j,M_k) + \sum_{M_k \in L_{ij}} W_G(M_{ij},M_k)$$

Thus, the gain can be written as:

$$g_{ij} = W_H(M_i,M_j) - \sum_{M_k \in L_i} W_G(M_i,M_k) - \sum_{M_k \in L_j} W_G(M_j,M_k) + \sum_{M_k \in L_{ij}} W_G(M_{ij},M_k) > 0$$

Note that under our assumptions of gain-independence this term does not involve any of the sets M_a,M_b, L_a, L_b, and L_{ab}. Therefore after merging M_a and M_b we get:

$$S_{new} = S(M,L) + g_{ij} + g_{ab} + \delta W_G(M_{ij},M_{ab}) \geq S(M,L) + g_{ij} + g_{ab}$$

Where δ=1 if M_{ij} is linked to M_{ab} and δ=0 otherwise. Thus, performing the additional merge between M_a and M_b would add g_{ab} to the new global score. The total gain is at least $g_{ij}+g_{ab}$ since we perform the merge steps without examining the possible link between M_{ij} and M_{ab}. ■

Corollary: A sequence of l merge steps can be performed simultaneously if the k'th merge in the sequence is gain-independent of merges 1 through k-1, for k=1,…,l. . ■

As a result of the theorem, instead of performing a single merge step and estimating the links on the new set we perform several merges, and evaluate the links between the new sets after merging. When we consider the merges in an iteration of the global improver, if many have a positive gain, we select the top B gains (we used B=1000). We then perform the set of merge steps ordered by their gain, skipping a merge if it is not gain-independent with all previous merges. We repeat this process until there is no merge that improves the global score.

While the asymptotic worst-case running time of this procedure is similar to performing a single merge at a time, we discovered that in practice many merge steps are performed per iteration. For example, in the lung cancer differential correlation analysis the maximal number of merges per iteration was 20, and the average was 4.

Comparison of scores for module links in the global improver

Our global improver used a statistical score to determine if two modules are linked. When the graphs are weighted, either the Wilcoxon rank-sum (WRS) test or the simpler hyper-geometric (HG) test can be used. We compared of the results of the global improver with each of the two scores using simulations. We generated weighted and unweighted graphs with 2000 nodes and 20 modules (see the main text for details). In each test, we ran the global improvers with both scores on the initial solution of DICER5. For graphs without any noise (i.e., the graph induces a perfect module map) the running times of the HG and WRS variants were 3 and 240 seconds respectively. Both variants perfectly discovered the planted module-map. On unweighted graphs, when the noise levels were increased to p=0.1, both algorithms reached the same performance of 0.97 but the HG running time was 3.8 seconds and the running time of the WRS variant was 394 seconds. We also applied the same test on weighted graphs with a standard deviation noise level of 0.8. The performance of the HG variant was 1 in 3.85 seconds. The performance of the WRS variant was 0.975 in 603 seconds. Our results show that the HG variant is much faster than the WRS variant, but achieves a similar performance.

Comparison to other weighted approaches

In our analysis in the main text we used un-weighted PPI and GI networks and included algorithms that are akin to previous methods. Other extant methods make use of the probabilistic scores of each GI edge, and incorporate both positive and negative GIs (6,7). Leiserson et al. (7,8) developed a method called Genecentric, which looks for locally maximum cuts in the GI graph. On the data of Collins et al. (9), this method was reported to outperform other methods, including algorithms that integrate GI and PPI information (10,11). We compared the performance of our methods to Genecentric and the graph compression method of Kelley and Kingsford (6), on the Collins data. Note that the other methods use all GIs while our algorithm uses only the negative GIs of the Collins data. Genecentric solution contained 116 modules of average size 10.75. These modules were paired, so that the map contains 58 links. Kelley and Kingsford reported 117 modules of average size 3, and the map contained 403 links. The results are summarized in the table below. Kelley and Kingsford reported many small modules that are not significantly enriched after FDR correction. Thus, the percent of enriched modules and links is not high. The solution of Genecentric covered 1248 genes, whereas the ModMap solution covered only 238 (in 32 modules). The total number of enriched GO terms in our solution was 53, compared to 39 in Genecentric's solution. Finally, 79% of the links in our map were enriched, compared to only 43% in Genecentric. This comparison indicates that ModMap produces comparable or better maps than state of the art methods for analysis of GI data.

Comparison of ModMap to extant methods on the yeast PPI and GI data of Collins et al.

Algorithm	Number of modules	Gene coverage	Maximal module size	Number of enriched GO terms	Percent enriched modules	Number of links	Percent enriched links
ModMap	32	238	20	53	84	67	79
Genecentric	116	1248	25	39	63	58	43
Kelley and Kingsford	117	355	17	32	17	403	6

Differential correlation cross-validation analysis

Our tests on human data utilized expression profiles of lung cancer and Alzheimer's disease and matching controls in each dataset. The first tested dataset, GSE13255 (12), contained 256 peripheral blood mononuclear cells gene expression profiles of patients with non-small cell lung cancer (NSCLC, n=150) and controls (n=106). The second tested dataset, GSE15222 (13), contained 363 post mortem cortex gene expression profiles of Alzheimer's disease (AD) patients (n=176) and controls (n=187). Since the networks used in this analysis were completely different from these used in the yeast studies, we first re-evaluated the different algorithms on them, based on the ability to reveal major changes in co-expression between sick and healthy individuals.

We used the method of Amar et al. (1) to compute two log-likelihood ratio scores for each gene pair: the consistent correlation (CC) score is positive if the gene pair is consistently correlated across phenotypes, and the differential correlation (DC) score is positive if the correlation difference between the cases and controls is higher than expected by chance. These scores were then used as edge weights in networks H and G, respectively, on which a module map was learned.

Given a module map constructed on a set of profiles (the *training set*) and a disjoint set of samples (the *test set*), the quality of the map prediction was evaluated on the test set as follows. For each pair of modules we calculated the absolute average DC between the modules on the test set data, and compared the DC values for links and non-links (i.e., two modules in the map that are not linked) using the Wilcoxon rank-sum test, where the null hypothesis is that there is no difference in DC between links and non-links. This measure is parameter-free and reflects all DC changes. As an additional test, in order to focus on major DC changes, we ignored all links with DC < 0.4, removed unlinked modules and calculated the proportion and number of remaining modules, links and the gene coverage. These parameters reflect the overall predictive quality of each reported map, and its ability to find strong DC signals. We used 2-fold cross-validation, that is, half of the data served as the training set, and the other half served as the test set. The process was repeated with the roles of test and training set switched and results were averaged.

An important parameter in calculating the DC LLR scores is the prior probability of real DC changes. In (1) a parameter K controlled this prior probability. Given a value of K, the prior probability was set such that only DC scores that are distant from the mean of the random distribution by at least K standard deviations (of the random distribution) will get a positive LLR score. Informally, this process guarantees that if the difference between the real and random distributions is minor, all LLR scores will be negative. In (1) a stringent approach was taken and the K parameter was set to 2. In this study we take a different, more direct approach to set the prior probability, using the following simple procedure: given a fixed threshold $\eta>0$ we set the prior to the maximal probability for which the LLR of η is negative. The intuition is that only DC of at least η receives a positive LLR score. Thus, unlike the K parameter, our approach is easily interpretable: we are guaranteed that absolute correlation changes lower than η will be assigned a non-positive LLR score. We used $\eta= 0.4$, which was equivalent to $K \cong 2.3$ on the tested datasets. Thus, our criterion was even slightly more conservative than (1).

An important parameter of the global improver is α, which is used to determine if the link between two modules is significant. We tested several values for α: 1E-4, 1E-6, and 1E-8. For each combination of an initiator and a value of α, we evaluated the map using the Wilcoxon rank sum test as explained above. The performance of the different initiators as a function of α is shown in the table below. A clear advantage for α=1E-6 is observed. For this value, the p-values of all initiators except DICER remain significant after Bonferonni correction over all tests. In addition, a clear advantage for MBC-DICER (i.e., ModMap) is observed, achieving a p-value of 1.54E-10 in the lung cancer data, and 9.06E-6 in the AD data.

Algorithm	$-\log_{10}(\alpha)$	Lung cancer	AD
DICER	4	0.029494	9.02E-07
DICER5	4	0.003697	5.31E-07
hierarchical	4	0.23858	1.48E-07
ModMap	4	0.12515	4.69E-04
NodeAddition	4	0.471779	4.31E-04
DICER	6	1.38E-07	0.026652
DICER5	6	7.23E-07	4.21E-04
hierarchical	6	1.80E-10	1.57E-04
ModMap	6	1.54E-10	9.06E-06
NodeAddition	6	4.49E-05	0.00115
DICER	8	0.067356	0.019777
DICER5	8	0.021281	0.463172
hierarchical	8	0.373946	0.014721
ModMap	8	0.343799	0.110778
NodeAddition	8	0.394908	0.334608

The full cross-validation results for α=1E-6 are shown in **Supplementary Table 11** (NSCLC data) and **Supplementary Table 12** (AD data). The maps produced by the local improver

received a very low p-value in the Wilcoxon rank-sum test between DC of map links and non-links, but suffered from low coverage. For example, for the MBC-DICER initiator, the local improver achieved a p-value of 4.43E-4 in the NSCLC data, and 3.31E-11 in the AD data. However, the map covered 197 genes in the NSCLC data, and 2197 genes in the AD data. In contrast, when applying ModMap (i.e., MBC-DICER with the global improver), the coverage was 1289 and 4955, respectively, with comparable p-values (1.54E-10, and 9.06E-6). Taken together, ModMap produces large maps that are robust when tested on independent datasets.

References

1. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol. 2013;9:e1002955.

2. Li J, Li H, Soh D, Wong L. A correspondence between maximal complete bipartite subgraphs and closed patterns. Knowl Discov Databases PKDD …. 2005;146–56.

3. Li J, Liu G, Li H, Wong L. Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: A one-to-one correspondence and mining algorithms. IEEE Trans Knowl Data Eng. 2007. page 1625–36.

4. Defays D. An efficient algorithm for a complete link method. Comput J. 1977;20:364–6.

5. Ulitsky I, Shlomi T, Kupiec M, Shamir R. From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. Mol Syst Biol. 2008;4 :209.

6. Kelley DR, Kingsford C. Extracting between-pathway models from E-MAP interactions using expected graph compression. J Comput Biol. 2011;18:379–90.

7. Gallant A, Leiserson MDM, Kachalov M, Cowen LJ, Hescott BJ. Genecentric: a package to uncover graph-theoretic structure in high-throughput epistasis data. BMC Bioinformatics. 2013;14:23.

8. Leiserson MDM, Tatar D, Cowen LJ, Hescott BJ. Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. J Comput Biol. 2011;18:1399–409.

9. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature. 2007;446:806–10.

10. Ma X, Tarone AM, Li W. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. PLoS One. 2008;3:e1922.

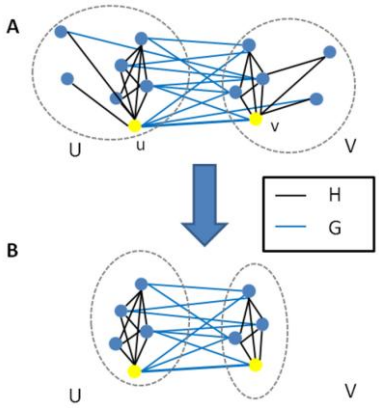
11. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data. PLoS Comput Biol. 2008;4:e1000065.

12. Showe MK, Vachani A, Kossenkov A V., Yousef M, Nichols C, Nikonova E V., et al. Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. Cancer Res. 2009;69:9202–10.

13. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, et al. Genetic Control of Human Brain Transcript Expression in Alzheimer Disease. Am J Hum Genet. 2009;84:445–58.

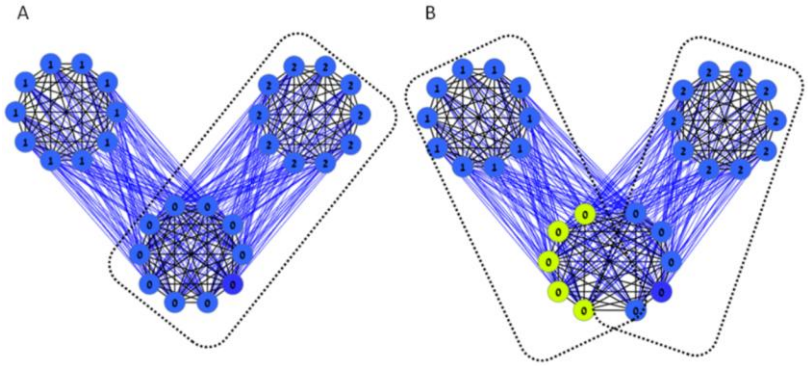
Supplementary Figures

Supplementary Figure 1



Supplementary Figure 1 Illustration of the DICER algorithm local search; A) An edge (u,v) in G is used as a starting point to form two sets U and V, which are the neighbors of u and v in H, respectively. B) Nodes are removed if they are not densely connected to their set in H or not densely connected to the other set in G. The final result after removing these nodes is shown – a pair of modules in H that are strongly linked in G.

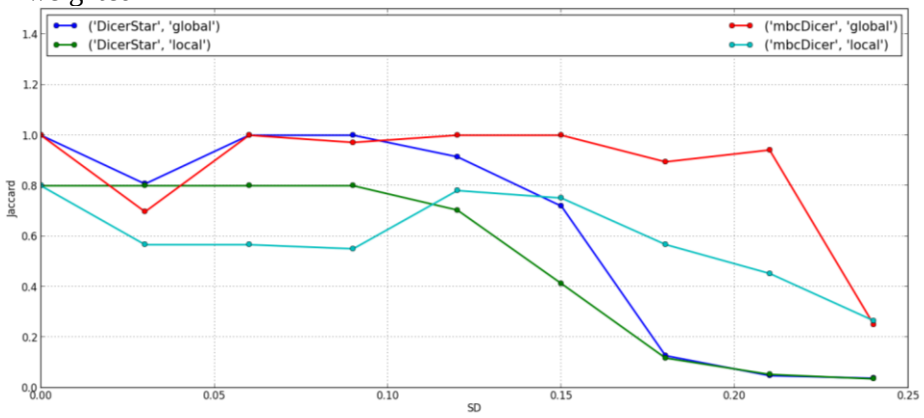
Supplementary Figure 2



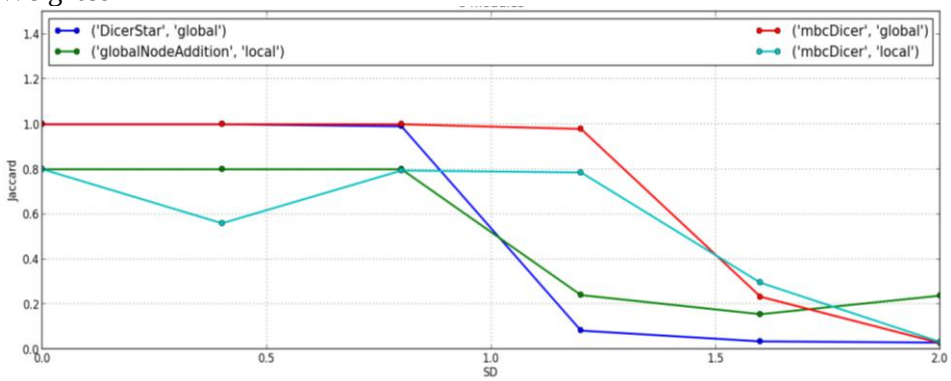
Supplementary Figure 2 Possible pitfalls of local improver that can be solved by the global improver. Edges of H are colored black; edges of G are colored blue. The number of a node is the module it belongs to. The initial solution that is provided to the improver is encircled by a dashed line. A) Given an initial solution that contains modules 0 and 2, a new module cannot be formed by the local improver. Hence, module 1 cannot be detected. B) Given an initial solution that partitions module 0 and links each part to a different module, the local improver cannot merge the two parts of module 0 since module 1 and module 2 are not linked.

Supplementary Figure 3

A) Un-weighted

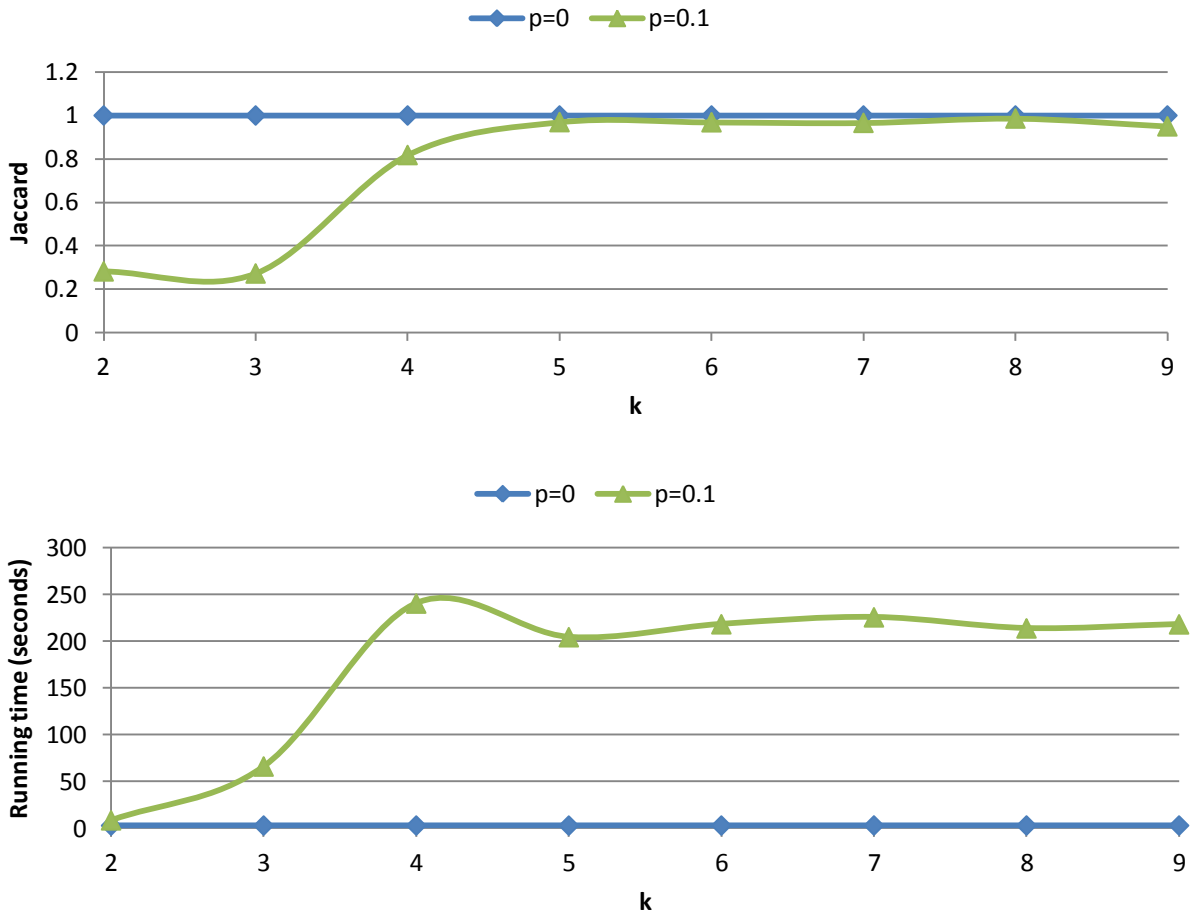


B) Weighted



Supplementary Figure 3 Performance of module map algorithms on simulated data with 1000 nodes. A) Unweighted graphs. B) Weighted graphs. The 1000-node graphs contain an embedded module-map of six modules in a tree structure. In addition, random cliques and bicliques are embedded in the graphs. Module, clique, and biclique size is chosen uniformly at random between 10 and 20. In the un-weighted model each edge is replaced by a non-edge with probability p, and vice versa. In the weighted model edge weights are sampled from the normal distribution N(1,σ), and non-edge weights are sampled from the normal distribution N(-1, σ). A-B) The top four performing algorithms are presented. The y-axis shows the Jaccard coefficient between the output of the algorithms and the known modules.

Supplementary Figure 4



Supplementary Figure 4 Comparison of $DICER_k$ variants for different values of k on simulated unweighted data with 1000 nodes and 20 modules. A) Performance. B) Running times. The 1000-node graphs contain an embedded module-map of 20 modules in a tree structure. In addition, random cliques and bicliques are embedded in the graphs. Module, clique, and biclique size is chosen uniformly at random between 10 and 20. Each edge is replaced by a non-edge with probability p, and vice versa. The results show that using k=5 gives better performance than k<5, and that k>5 does not improve performance. Running times are very similar for k>3. Based on these results, since we expect biological data to contain both large and small modules, we concluded that using k=5 gives a good balance of quality and considering small modules, and used it in subsequent analyses.

Supplementary Material of Chapter 2 (TWIGS)

Supplementary Text

This text is organized as follows. **Section 1** provides additional analyses of the sepsis gene expression data. **Section 2** explains how ISA and BIMAX were used, especially how their parameters were tuned. **Section 3** gives a thorough comparison of TWIGS to five methods: three that were developed for gene expression analysis, and two that were developed for functional connectivity analysis of fMRI data. For each method we tested a wide range of its parameters.

1. Additional results of TWIGS on the sepsis data

1.1 Analysis of the sepsis data without binarization

In this section we analyze the sepsis data using the normal distribution assumption for our model. Under the assumption that the vast majority of the cells (e.g., >90%) in the input matrix represent the background distribution, the overall empirical distribution of the data is expected to be similar to a normal distribution, with possibly heavier tails (Efron, 2009). By analyzing the log fold change values of the sepsis data we observed that our normal distribution assumption did not approximate the empirical distribution well. **Figure ST1** shows the histogram of the log fold change values. The solid curve shows the normal distribution density using the same mean and standard deviation of the empirical distribution. The right plot shows the QQ-norm plot of the data. As can be seen by both plots, the empirical distribution of the data did not fit the normality assumption.

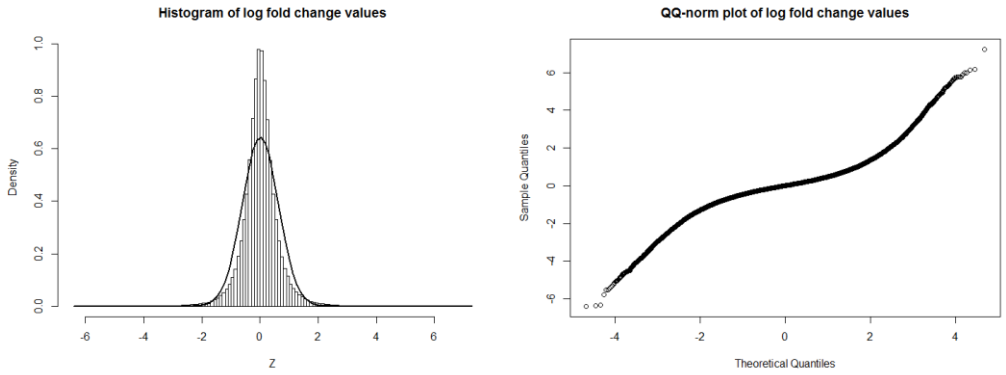


Figure ST1. Distribution of the log fold change values of the sepsis gene expression data. Left: histogram of the values. The solid curve represents a normal distribution with the same mean and standard deviation as the empirical distribution. Right: QQ-norm plot shows that the distribution is not normal as the curve does not fit a straight line.

In contrast, the fMRI data did fit the normal distribution assumption, as can be seen in **Figure ST2.**

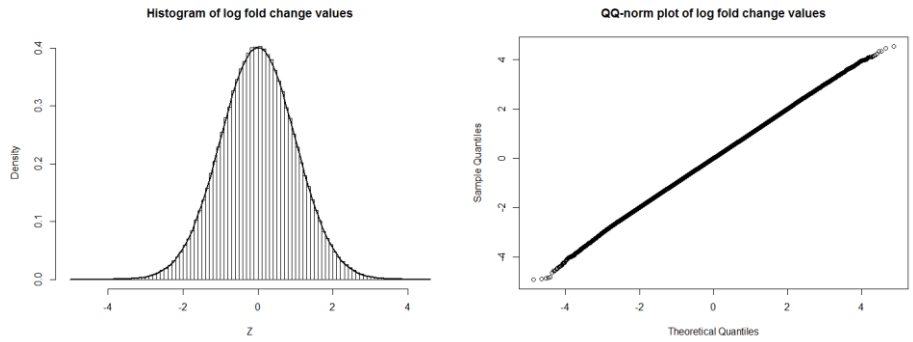


Figure ST2. Distribution of the values of the fMRI data. Left: histogram of the values. The solid curve represents a normal distribution with the same mean and standard deviation as the empirical distribution. Right: QQ-norm plot shows that the distribution is very similar to a normal distribution.

When we tested TWIGS on the sepsis data using the normal variant we observed that the algorithm wrongly inferred extremely high standard deviations (a value > 2 as compared to the empirical value of 0.5). This was detrimental for the algorithm as it now allowed many negative values to be included within biclusters. We therefore conclude that although very simplistic, our binarization based analysis was better for these data. Future studies can use empirical Bayes inference methods as a pre-processing step in order to evaluate the mixture distribution better (Efron, 2009; Dialsingh, 2012).

1.2 Additional results using different binarization thresholds

We also tested different binarization thresholds in the preprocessing step for our binary model. In the main text we report results for a threshold of 1 for the log-fold change (i.e., a fold change threshold of 2). Here we tested changing this threshold to 0.5 or 1.5.

Using a threshold of 0.5 two core modules were identified. These two core modules cover the same subjects as the two core modules detected using a threshold of 1. In addition, the core modules highly overlapped in their gene sets with their counterparts of the threshold 1 solution. For example, the first core module had 19 genes in common with the first core module ($p<1E-33$). However, a noticeable difference is that the gene sets of the subject-specific modules are much larger: the average size increased from 195 to 436. Also, all modules, including the subject-specific modules were enriched with known biological functions. The detected functions were similar to those of the threshold 1 solution. For example, core module 1 was highly enriched with response to bacteria ($q<0.001$). However, it had additional enrichments, such as modification of morphology ($q<0.001$). Using a threshold of 1.5 was too rigid. Only a single core module was identified and it was similar to the first core module of the previous solutions in terms of subjects and biological function.

In summary, this analysis shows that our results remained similar in terms of covered subjects and enriched biological functions when modifying the binarization threshold. Using a threshold of 0.5 more biological terms were detected, but the price was much larger gene sets. Using a threshold of 1.5 was too rigid and a complete core module was lost. We therefore conclude that a threshold of 1 represents a reasonable compromise.

2. ISA and Bimax

In this study we used solutions of Bimax (Prelic *et al.*, 2006) and ISA (Bergmann *et al.*, 2003) to produce starting points for the Gibbs sampling process. On binarized data Bimax searches for complete submatrices that contain only non-zero values. The main parameters are the minimal number of rows *minr*, the minimal number of columns *minc*, and the maximal number of detected biclusters *number*. In our tests, we used number =1000. Setting the other parameters was done as follows. We tested a wide range for both parameters in order to promote detection of biclusters with many columns. We set minr (i.e., minimal number of genes) to 10 and tested a range of minc values, starting from a relatively high value (50 in fMRI data and 10 in gene expression data) to a lower value (10 for both datasets), in decreasing order until at least one bicluster is found. If no bicluster was found we also tested setting both minr and minc to 8, 7, or 6, until we found some biclusters.

The ISA algorithm has two main parameters τ_G and τ_C that represent thresholds for the normalized expression values. We used ISA's common strategy of running the algorithm with different values of these parameters (Bergmann *et al.*, 2003). For both parameters we tested values of 0.5, 1, and 1.5. Since ISA tended to produce many small biclusters, we removed biclusters with less than 5 rows or 5 columns from the final output. For example, this filter removed more than 100 biclusters on the sepsis data.

3. Comparison to other approaches

3.1. Overview

Extant algorithms for three-way data analysis were mainly developed for gene expression data. However, existing models are too rigid for simultaneous analysis of responses in many subjects. For example, Triclustering (Zhao and Zaki, 2005) assumes that a module is a sub-cube created by one subset in each of the three dimensions. Our results show that both in the gene expression data (Figure 4) and in the fMRI data (Figure 5) the discovered modules are not triculsters since the time points and gene/neural position set of each private module differ under the same core module.

Another type of three-way analysis seeks biclusters <G',S'> where G' is a set of genes, and S' is a set of subjects, such that all genes in G' manifest a similar time response across all subjects in S'. Thus, such algorithms do not detect the relevant time points for each subject and they also require that all genes in G' would be relevant in each subject in S'. For the sepsis gene expression data, we compared TWIGS to two such algorithms: the three-way plaid model of (Mankad and Michailidis, 2014), and EDISA (Supper *et al.*, 2007). Of these two, the plaid model searches for a signal that is much more similar to the goal of TWIGS, which was to highlight up- or down- regulated pathways.

In fMRI data analysis we compared TWIGS to two methods. The first (Allen *et al.*, 2014) used a sliding window approach to detect different global correlation patterns in the brain. The output of this method is a clustering solution of all sliding windows from all subjects. We implemented four variants of this method in order to test different ways to learn the covariance matrices. The second approach (Rubinov and Sporns, 2011, 2010) averages the parcel correlation matrices of all subjects (where each correlation score is calculated using all time points), and uses a clustering solution in order to find clusters of brain regions.

In all comparisons we tested a wide range of parameters for each tested method. Below we repost on the tests in detail. We observed that while most methods produced a meaningful

output that is highly enriched with known biological functions of the gene/parcel sets, they lacked the ability to find modules that cover many subjects (see sections 3.2-3.4 below). In addition, in most cases our modules had a comparable and even better enrichment results (see section 3.5). Finally, our subject-specific modules provide an additional dimension of the signal that most methods fail to identify (see sections 3.2-3.4).

3.2. Comparison to three-way data analysis for finding extreme value submatrices

The algorithm of (Mankad and Michailidis, 2014) extends the classic plaid model of (Lazzeroni and Owen, 2002) to find biclusters < G', S' >, such that all genes in G' manifest an up- or down-regulated time response across all subjects in S'. A main threshold of the algorithm is the number of noise layers the algorithm learns internally. We applied the method to the sepsis gene expression data. We tested setting 2-5 noise layers (the default value of the tool is 3), and the number of detected biclusters was 4, 3, 4, and 1, respectively. **Figure ST3** shows the mean and max number of subjects covered by the biclusters in each solution.

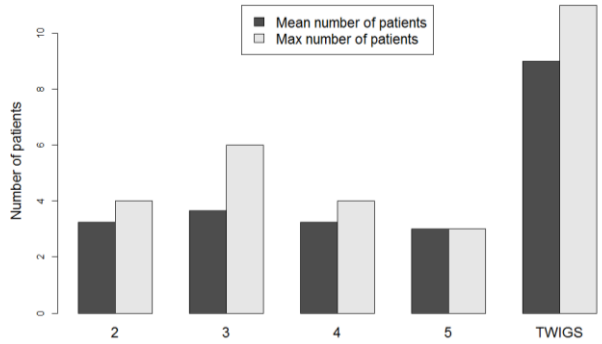


Figure ST3. Comparison of the three-way plaid model to TWIGS in terms of the number of patients covered by each solution on the sepsis gene expression data. For the plaid model different numbers of noise layers were tested: 2 (4 biclusters were detected), 3 (3), 4 (4), and 5 (1). TWIGS detected two core modules.

The figure shows that the plaid model usually covers only a small number of patients in each bicluster. TWIGS, however, can cover many patients due to the flexibility of its statistical model, which identifies subject-specific modules with a different set of genes and time points for each subject.

For each module (a bicluster in the plaid model solution or a module, a core module or a subject-specific module in our solution) we also looked at the induced submatrix in the data and calculated (for each number of noise layers in the plaid model and for the TWIGS solution) the percent of cells with fold-change > 2, and the average number of genes in a bicluster. **Figure ST4** shows that the solution of TWIGS reported far fewer genes (average 183), and a much higher percentage of cells with fold change >2.

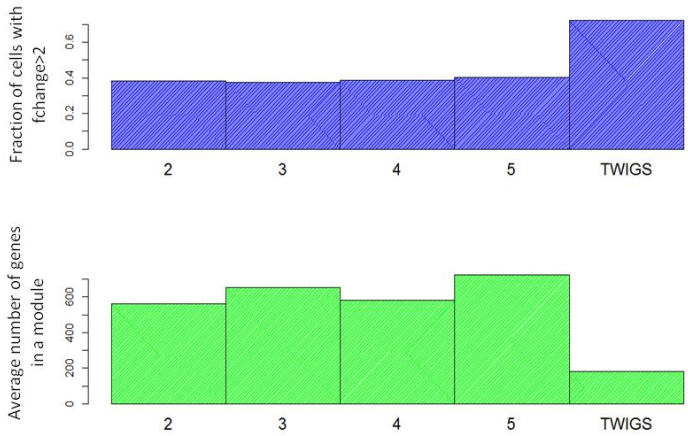


Figure ST4. Additional comparison of the three way plaid model to TWIGS on the sepsis data. Top: the fraction of module cells with high values (fold change > 2). Bottom: the average number of genes in a module. Numbers on the x axis indicate the number of layers used in the plaid model.

In summary, the plaid model identifies biclusters such that each covers a few subjects and has a large gene set that represent a moderate signal of differential expression. In contrast, TWIGS finds core modules with many subjects, the average number of genes in a module is much lower and the differential expression signal is much higher.

3.3. Comparison to correlation-based three-way data analysis methods

3.3.1 ESIDA

The EDISA algorithm (Supper *et al.*, 2007) extended the classic ISA algorithm to deal with three-way data. EDISA analyzes the correlations between genes within each subject and searches for biclusters <G',S'> (G' is a set of genes, and S' is a set of subjects) such that all genes in G' are highly correlated in the subjects in S'. It has three main parameters. First, the type of bicluster sought: (1) biclusters that represent *independent response* in each subject; that is, the genes in G' manifest high correlation within each subject but there is no constraint on correlation between the signals of different subjects; (2) biclusters with *coherent response* across subjects (i.e., a similar signal between subjects), and (3) responses that appear specifically in a single subject, called *single response* modules. In each run of the algorithm we tried all three options. The second and third parameters are τ_G and τ_C that represent thresholds for the correlation between genes within a bicluster (τ_G, default 0.2), and between subjects (τ_C, default 0.1). Low values of τ will increase the required correlation of the biclusters. Hence, increasing these thresholds results in a higher number of genes and higher overlap between biclusters (Supper *et al.*, 2007).

Using the default parameters no module was detected on the binarized sepsis data. Even after increasing the τ parameters up to 0.3 or 0.5 no module was detected. Therefore, the algorithm could not detect modules that represent up-regulation of modules in response to sepsis.

Running the algorithm on the non-binary sepsis data, with default parameters, the algorithm detected seven independent response modules and a single coherent response module. The average number of genes was 108 (σ=62), and the average number of subjects was 4.28 (σ=2.75). Thus, most modules in the EDISA solution cover a much lower number of subjects

compared to TWIGS. Only two out of eight EDISA modules were enriched with GO terms, compared to 19 out of 20 for TWIGS. Forcing higher correlation in the EDISA solution (by decreasing the τ_G parameters up to 0.05) resulted in a single bicluster of 24 genes that covers five subjects. No significant GO enrichment was detected.

Lowering the correlation of the modules by increasing the τ parameters up to 0.3 resulted in 28 independent response modules, two coherent response modules, and six independent response modules. The average number of genes in a bicluster jumped to >400 (for example, an average of 434 for the independent response modules). Surprisingly, the number of subjects decreased to a mean of 3.833 (σ=2) in the independence response modules, and 3 in the coherent response modules (both modules had three subjects). This is a weakness similar to that of the plaid model. GO terms enrichment was detected in 27 out of the 36 modules. That is, 75% of the biclusters were enriched with at least a single GO term, as compared to 95% in the TWIGS solution. An advantage of EDISA is that it searches for general correlation patterns, which results in many genes per module and a larger number of enriched GO terms in modules (70, compared to 55 for TWIGS). However, not all of these can be attributed to up-regulation as in the TWIGS solution.

3.3.2 Sliding window analysis

3.3.2.1 The output of TWIGS in the context of sliding windows

A common approach in fMRI data analysis is to calculate correlation between voxels or parcels in time windows in order to find changes in correlation across time. The goal of such analyses is to find dynamic changes in brain connectivity (Allen *et al.*, 2012; Damaraju *et al.*, 2014).

We used a sliding window approach to analyze the modules of TWIGS. For each subject-specific module we analyzed its submatrix. For each time window of size 15 we calculated the correlation between the module parcels in that window. For a defined subset of windows, we computed the median correlation score. Three subsets were considered: (1) windows containing at least one of the module's time points, (2) windows containing none of the module time points, and (3) all windows. We call the groups W, WO, and ALL, respectively. We expected that W scores would be higher than WO scores as these scores represent correlations around time points that are highly activated. **Figure ST5** compares the W, WO, and ALL distributions from all modules.

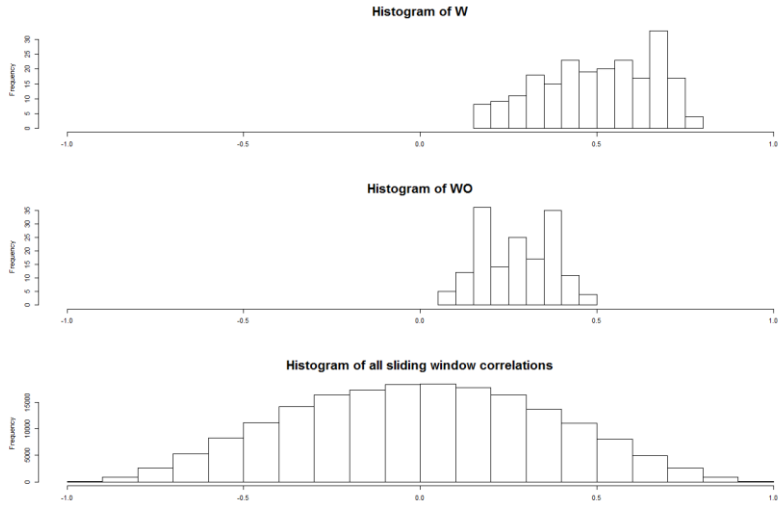


Figure ST5. Distributions of correlation values of time windows of size 15, for different subsets of windows. Top: correlation between the module parcels in windows containing at least one of the module's time points. Center: correlation between the module parcels in windows containing none of the module's time points. Bottom: distribution of the correlation between all pairs of windows across all parcels.

Figure ST5 suggests two main conclusions that are rather obvious due to the characteristics of the signal that TWIGS detects. First, the correlation between parcels from TWIGS modules is much higher than expected by chance (comparing the W and WO scores to the ALL distribution). Second, as expected, the W scores are much higher than the WO scores (Wilcoxon rank-sum test $p = 1.3E-33$).

3.3.2.2 Comparison to sliding window-based analysis

(Allen *et al.*, 2014; Damaraju *et al.*, 2014) clustered the covariance matrices of tapered sliding windows from all subjects in order to detect patterns that represent similar temporal functional connectivity across subjects. To reduce noise levels, they estimated the covariance matrices using the graphical lasso approach (Friedman *et al.*, 2008), which adds a lasso penalty term to the estimation of the inverse covariance matrix. The expected effect is that many low covariance values, which are expected to occur due to noise, would be reduced to zero. The final step of the algorithm is to use k-means to cluster the covariance matrices of all the different windows from all subjects. The value of k is selected by plotting the ratio of the within-cluster sum of squares to the between-cluster sum of squares. Using the plot, k is selected manually using the "elbow" rule: select the k from which little gain is achieved.

The graphical lasso approach depends on a parameter ρ that determines the amount of regularization of the covariance matrix. Higher ρ values will produce more zero values. Estimating ρ is a challenge that requires experimental analysis. When testing the graphical lasso algorithm on our data we encountered two main difficulties: (1) the algorithm had convergence difficulties in low ρ values (e.g., $\rho=0.001$), which led to unreasonable running times (e.g., > 10 minutes for estimating a single covariance matrix), (2) ρ values similar to that used in the original studies either had a little effect (e.g., when a value of 0.05 or 0.1 was used less than 5% of the values were zero) or gave a covariance matrix that was almost diagonal (e.g., using a value of 0.25 or 0.3, >99% of the values were zero).

In our analysis we tested two approaches for learning covariance matrices. First, we used the graphical lasso algorithm with a ρ value of 0.25. Note that as stated above this value removes most of the information from the covariance matrix as typically >99% of the values were set to zero. As an alternative we also tested the shrinkage algorithm of (Schäfer and Strimmer, 2005), which also estimates covariance matrices using a shrinkage approach, but estimates its trade-off parameter internally. Empirically, using this algorithm only a few values were set to exact zero, but the overall shrinkage effect was noticeable. To cluster the covariance matrices we used k-means with k ranging from 2 to 30. For each of the variants above we tested clustering the data with and without standardization.

Figure ST6 shows the performance curve of the clustering solutions as a function of k:

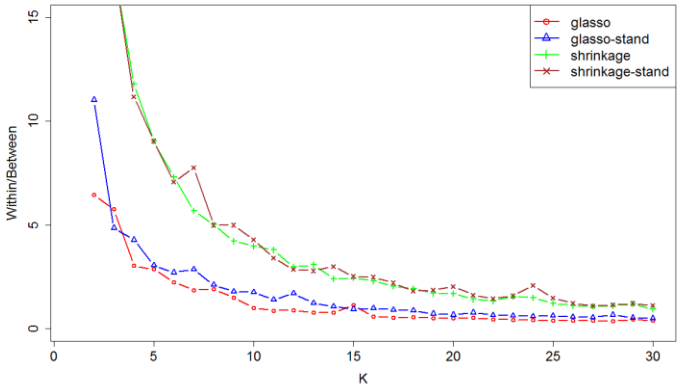


Figure ST6. Performance of the sliding window methods as a function of the number of clusters k. For each value of k between 1 and 30 the ratio of the within-cluster sum of squares to the between-cluster sum of squares is shown.

We observed that using the graphical lasso algorithm the "best" number of clusters was low, roughly 5. These results are in agreement with the original studies. However, as we noted above, this algorithm removes most of the covariance information from the data. Using the shrinkage algorithm we concluded that the “best” k is larger than 5. (Between k=5 and k=10 the within/between value decreases by a factor of 1.8. Similar values are obtained when comparing k to k+5 for k=6-9. For k=10 and beyond the factor drops below 1.4.)

We also observed that all variants had difficulties in covering many subjects in the detected clusters. **Figure ST7** shows the mean number of subjects covered by each clustering solution.

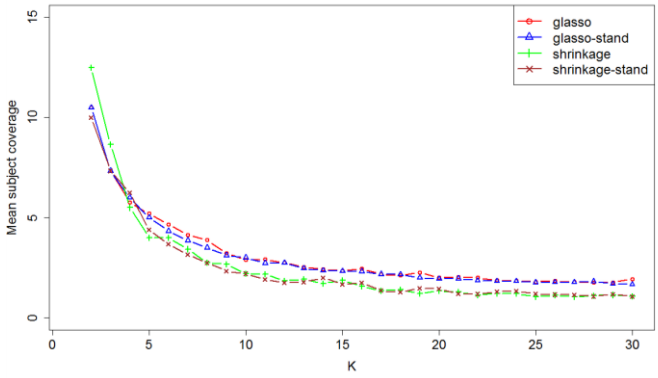


Figure ST7. Mean number of covered subjects (over all clusters) in the sliding window approaches as a function of the number of clusters k.

Figure ST7 shows that as K increases the mean number of covered subjects decreased rapidly. Although this is expected, we note that for values of K that represent the number of selected clusters (i.e., 10 for shrinkage, and 5 for graphical lasso) the mean number of covered subjects is below 5. In contrast, each of TWIG's core modules covered all subjects in the data. The graphical lasso solution with k=5 did have a single cluster that indeed covered all subjects, whereas the shrinkage algorithm with k=10 had a single cluster that covered 11-13 subjects (depending on the algorithm variant).

Our overall conclusion is that sliding window approaches have difficulties in finding several core clusters that cover many subjects whereas TWIGS can easily achieve that. Another important difference is that the output of TWIGS contains the parcel set of each core module, whereas the output of the sliding window approach does not.

3.4. Comparison to modularity analysis approaches of fMRI data

Another standard approach for analyzing multiple fMRI matrices simultaneously is modularity analysis (Rubinov and Sporns, 2011, 2010). In this approach, a preprocessing step is used to create a single graph G=(V,E) in which V represents parcels and E represents edges between parcels. In the weighted version of the algorithm E covers all parcel pairs and the score of each pair is the average correlation across all subjects. In the unweighted version E contains only parcel pairs for which the average correlation across all subjects exceeds a predefined threshold. Given the graph G, a modularity algorithm is used to partition the graph into sub-graphs (i.e. modules) attempting to maximize weights within modules and minimize weights between modules (Rubinov and Sporns, 2011).

We tested both the weighted and unweighted version of the algorithm. In the unweighted case we used a graph-density-based threshold (3%, 5% and 10%). The common practice in modularity analysis is to generate several clustering solutions and compare their Q score, which represents the gain of the solution compared to a random solution in terms of both homogeneity and separation of the clusters (Rubinov and Sporns, 2011, 2010). **Table ST1** shows that the weighted version achieved a much higher Q score.

Table ST1. Statistics of the different variants of the modularity analysis algorithm.

Graph	**Density (%)**	**#Modules**	**Q**
Weighted	100	4	1.658
Binarized	3	6	0.5379
Binarized	5	5	0.4599
Binarized	10	5	0.363

Interestingly we observed that the solution of the weighted version was very similar to the core modules detected by TWIGS. In fact each core module was almost contained in one of the detected modules. **Table ST2** shows for each module in the modularity analysis solution its most similar TWIGS core module, the fraction of parcels in the core module that were in the overlap, and the enriched functional terms in each module (Thomas Yeo *et al.*, 2011). The

results above confirm that TWIGS indeed identified functionally relevant core modules in the brain.

Table ST2. Comparison of the modularity analysis solution to the core modules of TWIGS.

Module	Size	Enriched Terms	Most similar TWIGS core module	Size of parallel TWIGS module (μ1=1.5, μ1=2)	% of TWIGS module covered (μ1=1.5, μ1=2)	% of module covered by TWIGS module (μ1=1.5, μ1=2)	Enriched term
1	93	VAN, AN	2	82,21	1,0.77	0.88,0.17	VAN
2	95	VN	1	57,55	1,1	0.6,0.58	VN
3	120	SMN, DAN, VAN	3	84,72	0.99,0.86	0.69,0.52	DAN, SMN
4	155	FPCN, DMN	4	80,46	0.93,0.86	0.48,0.26	FPCN, DMN

We also evaluated the "strength" of that identification in each solution, calculating its enrichment factor. The *enrichment factor* score (EF) of a module G for a term T is the ratio between the observed number of parcels labeled with term T in G, and the expected number parcels labeled with term T in modules of size |G|. The average EF score of the terms identified by TWIGS was 3.4 with μ=2, and 3.7 with μ=1.5 (see the main text for details), whereas that of the modularity analysis was 2.7. In addition, note that the modularity analysis cannot (by definition) produce a fuzzy clustering of brain parcels, nor can it find subject-specific augmentations of each module. The added value of TWIGS is in the ability to also reveal heterogeneity among the subjects. For example, **Figure 5** in the main text shows that on average each subject covers <50% of the parcels in the core module. The application of such analysis is in identifying subjects with unique patterns. For example, TWIGS identified four subjects whose private modules of core module 4 were enriched with parcels of the ventral attention system as well as parcels of the default-mode and frontoparietal control. This may indicate a tendency of these subjects to engage in bottom-up processing (e.g. be more attentive to sensory stimuli) during goal-directed thought processes.

3.5 Summary of the enrichment analyses on the sepsis gene expression data

Here we summarize the GO enrichment results for the modules and biclusters detected on the sepsis data. For each method we ran the TANGO algorithm (Ulitsky *et al.*, 2010) using 0.05 FDR correction, and calculated three statistics based on the results: (1) the total number of GO terms covered, (2) the average enrichment factor (EF, calculated as explained in Sec. 3.4 above), and (3) the percent of modules in which at least one significant enrichment was found.

The tested methods were: (1) TWIGS, (2) EDISA with default parameters, (3) EDISA with τ=0.3, (4) the plaid model with 2-5 noise layers, (5) SAMBA with default parameters, (6) SAMBA with overlap parameter set to 0.2 (i.e., more biclusters are filtered out compared to the default parameters), denoted as SAMBA-0.2, and (7) ISA with parameters set as described in section 2 above.

In terms of the total number of GO terms covered, EDISA-0.3 achieved the highest number with 89 terms (over all bicluster types, see Section 3.3.1), and TWIGS ranked second (**Figure ST8**). In terms of EF, the top performer was EDISA with default parameters, reaching a mean EF score of 27, followed by SAMBA-0.2 (**Figure ST9**). Both of these algorithms completely failed in terms of the fraction of modules showing enrichment (**Figure ST10**. We counted those modules showing at least one enrichment with a q value $\leq$ 0.05). All plaid solutions reached a perfect score of 1 in this measure, followed by TWIGS with a score of 0.95.

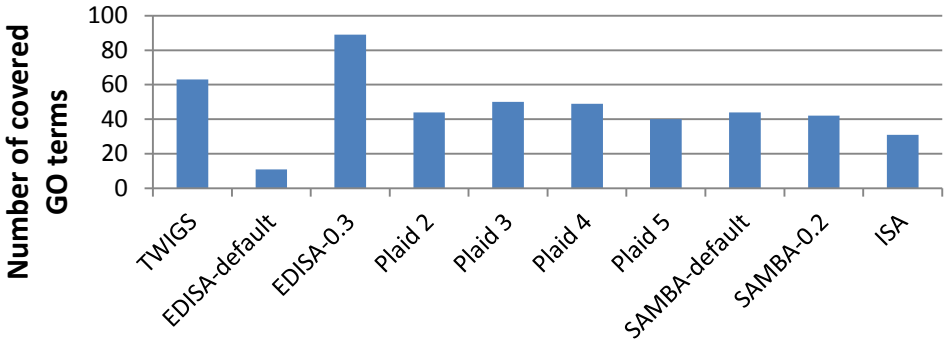


Figure ST8. Number of GO terms in the enrichment analysis results on the sepsis data.

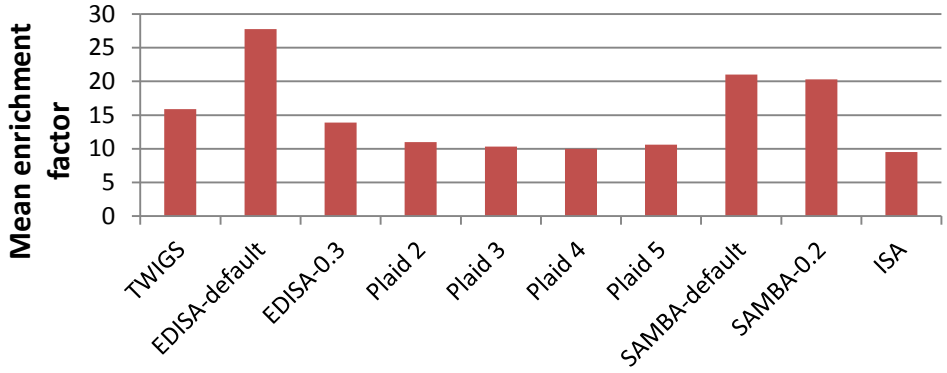


Figure ST9. Enrichment factor scores of the enrichment analysis results on the sepsis data.

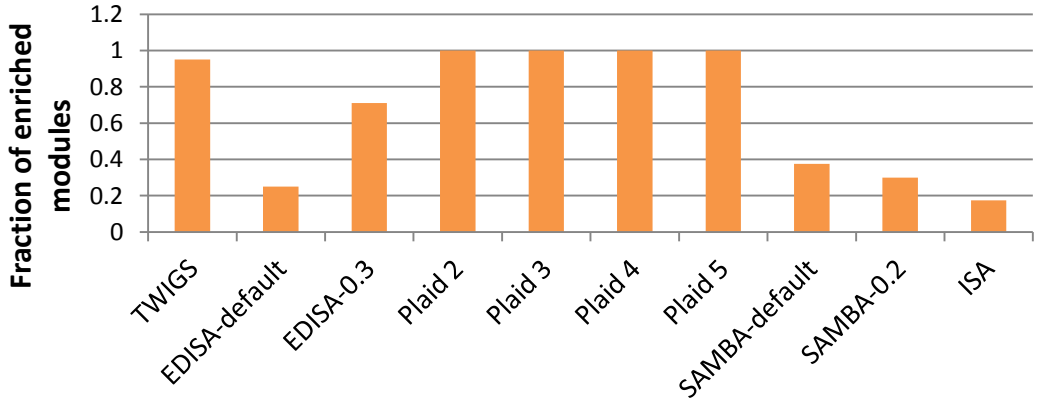


Figure ST10. Fraction of enriched modules in the enrichment analysis results on the sepsis data.

Finally, since no single method reached top performance in all scores, we consolidated the three scores as follows. We ranked the methods according to each measure and calculated the average rank of each method, where 1 means best rank and 10 is worst. **Figure ST11** shows the results. The top method was TWIGS with an average rank of 3.66. The next best methods were EDISA-0.3 and the plaid models with 2-3 noise layers, all with a score of 4.

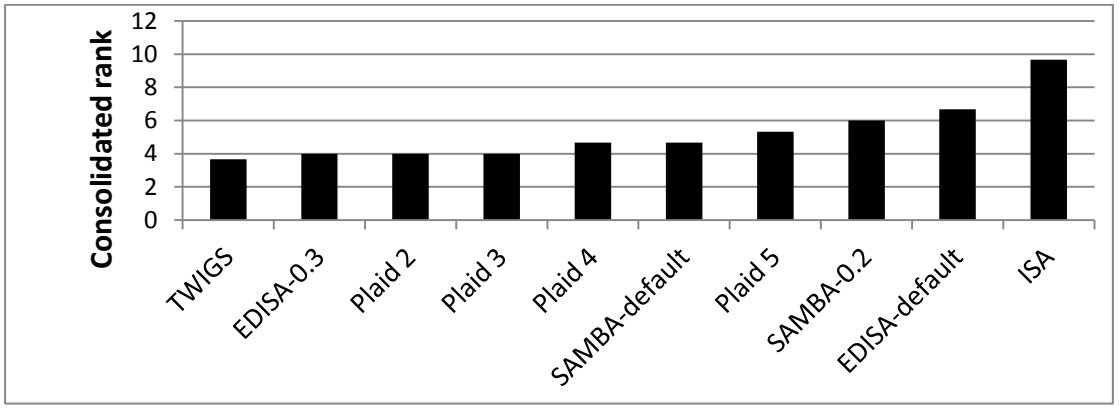


Figure ST11. A consolidated rank of all methods based on all three scores calculated to measure the enrichment results.

References

Allen,E.A. *et al.* (2014) Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex*, **24**, 663–676.

Bergmann,S. *et al.* (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E*, **67**.

Damaraju,E. *et al.* (2014) Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage Clin.*

Dialsingh,I. (2012) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. *J. Appl. Stat.*, **39**, 2305–2305.

Efron,B. (2009) Empirical Bayes Estimates for Large-Scale Prediction Problems. *J. Am. Stat. Assoc.*, **104**, 1015–1028.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Stat. Sin.*, **12**, 61–86.

Mankad,S. and Michailidis,G. (2014) Biclustering Three-Dimensional Data Arrays With Plaid Models. *J. Comput. Graph. Stat.*, **23**, 943–965.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Rubinov,M. and Sporns,O. (2010) Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, **52**, 1059–1069.

Rubinov,M. and Sporns,O. (2011) Weight-conserving characterization of complex functional brain networks. *Neuroimage*, **56**, 2068–2079.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article32.

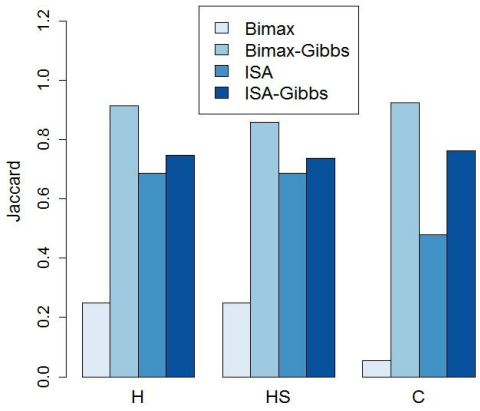
Supper,J. *et al.* (2007) EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics*, **8**, 334.

Thomas Yeo,B.T. *et al.* (2011) The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.*, **106**, 1125–1165.

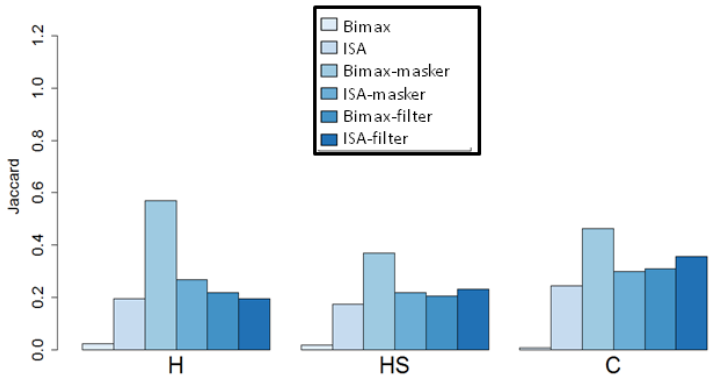
Ulitsky,I. *et al.* (2010) Expander: from expression microarrays to networks and functions. *Nat. Protoc.*, **5**, 303–322.

Zhao,L. and Zaki,M.J. (2005) {TriCluster}: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data. In, *In Proc. of the 2005 ACM SIGMOD international conference on Management of data*. ACM Press, pp. 694–705.

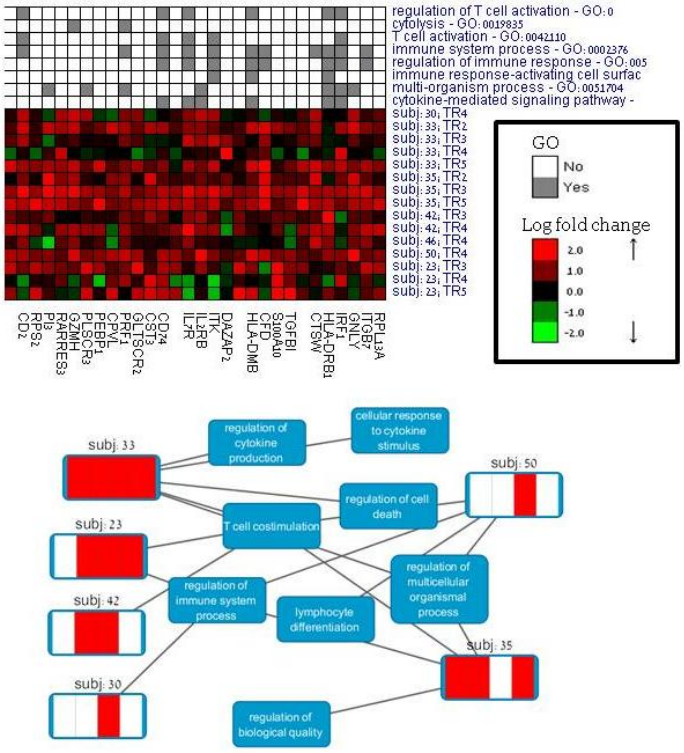
Supplementary Figures



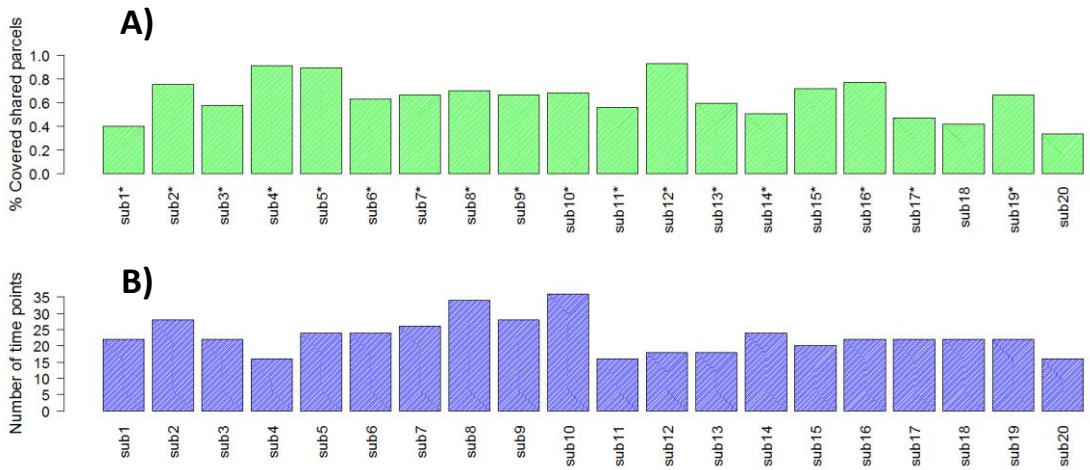
Supplementary Figure 1. Simulation results on data with a single module and non-symmetric noise levels $p'_o = 0.5, p'_w = 0.1$.



Supplementary Figure 2. Simulation results on normal data with five modules and noise levels σ_w=1 and σ_o=1.



Supplementary Figure 3. The second core module and its subject-specific enrichments in the sepsis data. Top: the first core module heatmap. Bottom: the subject-specific enrichments. The red stripes in each patient's node represent the time points that were covered by its private module. An edge between a subject and a category (blue node) indicates that the subject-specific module was enriched for that category.



Supplementary Figure 4. Statistics of the subject-specific module of shared-bicluster 4A. A) The percent of core module parcels covered by the private modules. Asteriks indicate subjects whose private module had a significant overlap (hyper-geometric $p \leq 0.001$) with the core module. B) The number of time points in each private module.

Supplementary Material of Chapter 4 (ADEPTUS)

Here we give the supplementary text and figures. For the supplementary tables please go to the online version of the paper at http://nar.oxfordjournals.org/content/43/16/7779.

Supplementary Text

MetaMap analysis

We tested the MetaMap tool to automatically assign samples to their UMLS or MeSH disease terms. To test the quality of the tool, we used the description of the GEO datasets as an input. Manual examination of the results indicated that using this tool will introduce errors to the labels of the samples. For example, for the GEO dataset GDS4222, the top scored disease was "HIV infections". However, the description of this dataset includes: "Analysis of diagnostic lymph-node biopsies from classic Hodgkins lymphoma HIV- patients before ABVD chemotherapy". Thus, the top prediction is an example of a false positive association between a text and a disease term. In another case, for the GEO dataset GDS1067, MetaMap produced six different disease terms although the dataset contained samples of three diseases (monoclonal gammopathy of undetermined significance, multiple myeloma, and plasma cell leukemia). Although the output was highly relevant for the input text, manual analysis is needed to assign the samples of this dataset to their most specific terms. Due to such examples we decided to manually annotate samples in the current study.

Enlarging the compendium

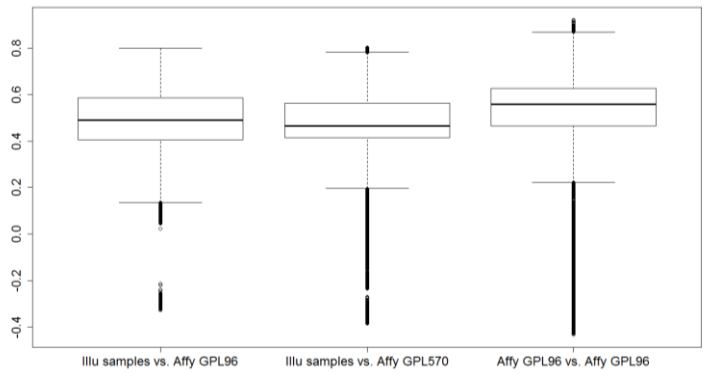
ADEPTUS covers 13,314 samples from 17 different microarray technologies. Note that although the raw number of human datasets in GEO is much larger, annotating datasets is a major hurdle. As of July 2015, the number of annotated human microarray datasets (datasets with a GDS id) with at least 20 samples, a 'disease' flag, and that originate from the platforms covered by ADEPTUS is only 254. Further enlargement of the compendium would require substantially more manual annotation work and is left for future work.

Rank-based scores

Given a gene expression profile of a single sample S in which k genes were measured, we ranked the genes by their expression levels g_1, g_2, g_3,…, g_k (with g_1 having the highest level), and assigned a score to each gene based on its rank: $W_S(g_i) = ie^{-i/k}$. Note that in [1] the weights were $W_S(g_i) = ie^{i/k}$, whereas in [2] the weights were calculated with the minus sign. We preferred the latter as it produces very low differences among genes ranked low (i.e., it keeps a low difference among genes with low expression intensities).

The use of previously preprocessed data followed by our rank-based scoring may raise a concern regarding the ability to compare results of different platforms and preprocessing. The plot below shows the correlation distribution when taking 200 samples at random from each of the three largest platforms. The correlation was calculated between the weighted ranks profiles of the samples. The results show that in general the correlations between platforms are high and are very similar (mean correlation 0.46- 0.49). There is a modest advantage to correlations of samples from the same company and platform. Hence, overall, the correlation

among samples - after our normalization - is very high even between platforms of different companies, preprocessed in completely different ways.



Finding disease-specific differential genes

For each gene G and a disease term D we used G's expression pattern to rank all microarray samples and then utilized the disease labels on this ranking to calculate G's PN-ROC, PB-ROC, and SMQ scores (see details below). Note that the SMQ scores are calculated using the Wilcoxon rank-sum test that assumes independence between samples within a study. Hence, this score could be problematic for datasets with strong batch effects. Nevertheless, we used this analysis only for diseases that were well classified in leave-dataset out cross validation. Hence in these cases the discriminatory signal was consistent across platforms and batches.

We assign a gene G to a disease term D if both of its ROC scores are at least 0.65 and the SMQ (q-value) is at most 0.05. To make sure that a selected gene G is specific to D. D is considered only if it is a leaf or it has at least three datasets with samples whose most specific annotation is D. In the latter case we re-calculate the SMQ score using these datasets only and keep G only if it is found significant again. We call the last constraint the *sub-annotation filter*.

As an example of the effect of the filter, consider the case where a parent DO term and its child term have high overlap in their positive samples. For example, leukemia has 1125 samples, and its child term lymphoblastic leukemia (LL) has 852 samples (see **Figure 3**). Applying the initial gene-disease association process will associate 272 genes to leukemia, and 642 to LL, with an overlap of 176. Hence, 64% of the leukemia genes correspond also to its child LL. Using the sub-annotation filter on 273 leukemia non-LL samples (from 4 studies), the analysis produced 81 leukemia genes, with only four genes in common with the LL gene set.

To better characterize the specificity of biomarkers produced by our method we looked at gene ranks within datasets. For every gene g, we calculated the gene's z-score of differential expression (Wilcoxon rank-sum test) for each pair (x,D) where x is a dataset (out of 174) and D is a disease (out of the 24 well-classified diseases). The dataset x was considered only if it contained both samples of D and other samples (controls). For every pair (x,D) we then ranked all genes by their absolute z-score. For each disease D and gene g, we calculated g's median rank across all of D's datasets. Given a disease D, we computed the distribution of these median ranks for (1) D's biomarker genes (2) the biomarker genes of its parent disease (P(D)) and (3) when available, the biomarker of its grandparent disease (P(P(D))). We emphasize that in all three cases the median ranks are computed on D's datasets.

Multi-label classification algorithms

The input to our learning problem is a training set $\{(x_1,y_1),\ldots,(x_n,y_n)\}$. Each vector $x_i \in \mathbb{R}^p$ contains the p gene scores of sample i, and each vector $y_i \in \{0,1\}^N$, where N is the number of labels (i.e., the number of diseases), and y_{ij} is 1 if and only if sample i originated from a patient with disease D_j. Our goal is to learn a *multi-label classifier*, that is, a function that receives as input a sample $x \in \mathbb{R}^p$, and for each disease D_j predicts the probability that the sample x belongs to D_j. Multi-label classification problems have received considerable attention in recent years [3]. These methods can be broadly partitioned into two types: *problem transformation* and *algorithm adaptation* [4]. Problem transformation methods transform the original problem into one or more standard classification problems. For example, the label power-set (LP) method defines a new categorical variable (for each sample) whose values are all possible combinations of the original labels, which is then used as the class attribute. This method models the label dependencies implicitly and is usually effective when the number of labels is small [5]. Algorithm adaptation methods extend a specific learning algorithm to deal with multi-label classification. For example, predictive clustering tree (PCT) learns decision trees for the multi-label task [6, 7]; and Bayesian correction (BC) uses the known label hierarchy to correct errors introduced when learning an independent single binary classifier for each label [8, 9].
In this study we tested three simple approaches for multi-label classification our database. The first, *Single*, learns a separate classier for each disease. Thus, this approach ignores dependencies among different diseases. We tried two different classifiers: linear support vector machines (R package e1071, with default parameters and logistic models for calculating probabilities), and random forests. When training a classifier for a specific disease we used the top 100 features only (using simple t-test feature selection). Note that when performing leave-dataset out cross validation the feature selection and classifier learning steps are performed using the training set only, in order to avoid over-fitting.

The second approach, *Bayesian Correction* (BC), is a method that adapts single classifiers and uses the DO structure [8, 9]. This method was used in previous studies, see [8-11] for full details. Briefly, single-label classifiers are learned on one portion of the training set. Then, the predictions of these classifiers on a second set of excluded samples are used to learn a Bayesian network in which nodes are diseases. For each disease D_j two connected nodes were added to the network: D_j^1 marks the real labels, and D_j^2 marks the predictions of the single classifiers. The DO network edges are also added between the D^1 nodes. The task is to learn the conditional dependencies. We used the BNLearn R package [12] for inferring the network parameters. Given a new, unlabeled sample x, first a single classifier is used to get a vector of predictions $y' \in \{0,1\}^N$ that is, a binary prediction for each disease term. Then, the network is used to estimate the probability of the unknown true assignments y (D^1 nodes) given y' (the D^2 nodes) and the network structure. Once such probabilities are estimated, the marginal probability of each disease node D_j^1 is calculated. These probabilities are used as the final predictions of the algorithm.

The third approach, *label power-set (*LP*)*, is a simple problem transformation approach [3, 4]. We used the concatenated input label vectors as the classes. For example if the label vector of a sample is y=(0,1,1) then it is assigned to class "011". Note that by construction a sample labeled with a disease is also labeled with all its ancestors in the DO hierarchy. Let $C_1,\ldots,C_M$ be the resulting classes, and let $\delta(D, C_i)$ be 1 if samples of class C_i belong to disease D and 0

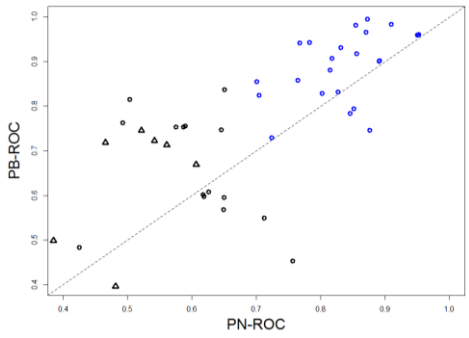
otherwise. After a *multiclass* classifier was learned using the new classes, predicting the probability p_D that a sample x has a disease D is done as follows. First, we use the multiclass classifier to compute p_i, the probability that x belongs to C_i for i= 1,…,M (Note that $\sum_i p_i = 1$). Then, $p_D = \sum_i p_i \delta(D, C_i)$. Intuitively, when considering the DO structure a sample x will be assigned to a disease D if according to the classifier the sample has either D, or one of D's sub-diseases, thus, preserving the is-a property of the DO structure. This strategy is reasonable when the number of the artificial classes generated is not large (in theory it could be an exponential of the number of labels) [5]. We used random forest as the multiclass classifier. The number M of classes in practice was 53, which is only slightly higher than the number of tested diseases (48).

Study-based Meta-analysis q-value (SMQ)

This simple test takes as input three features for each sample originating from a study of a disease D: (1) a score x, (2) a binary value y where y=1 if the sample is positive in D, and y=0 otherwise, and (3) the dataset DS_i that the sample belongs to. The Study-based Meta-analysis Q-value (SMQ) of the score x in disease D is calculated as follows. For each dataset $DS_1,\ldots,DS_k$ that contains positive and negative samples we use the Wilcoxon rank-sum (approximated) test to calculate a Z-score for the separation between the positives and the negatives based on their scores. This approximation is reasonable as we used studies with at least 20 samples each. This calculation produced a set of Z-scores: $Z_1,\ldots,Z_k$. Then, under the assumption that the datasets are independent, and the null hypothesis that the positives and negatives have the same x distribution, the sum of the Z-scores follows a normal distribution with mean zero, and variance k [13]. The SMQ score is the FDR corrected [14] two-tail p-value based on this distribution. Note that we calculated the Wilcoxon rank-sum test p-values for all gene-disease pairs and the SMQ is calculated by correcting all p-values together.

Well-classified diseases

We designated a disease *well classified* if its PB-ROC and PN-ROC scores exceeded 0.7 and its SMQ was significant (< 0.05). Since the number of samples used is very large, even lower ROC scores would have been highly significant although classification quality is low. We therefore chose the threshold of 0.7 as a good bipartition point (see the figure below), and used the SMQ score to account for significance.



In the plot above each point represents a disease. Triangles are diseases that were not significant in their SMQ. Blue points are diseases with both ROC scores above 0.7.

The independence assumption

We assume independence between samples in several parts of the manuscript -- e.g. when assigning genes to tumor site via a hyper-geometric test and when assessing the specificity of a biomarker. This assumption is likely violated when pooling data across multiple studies and multiple platforms and likely even within a single large data set due to batch effects.

As most published literature, our analyses throughout the paper provide reasonable simplified estimations that approximate the complex real underlying distributions (e.g., by using the independence assumption). Nevertheless, in each analysis we took additional precautionary steps to alleviate that effect. For example, in the gene expression analysis the key step was leave-dataset out cross validation. This validation points out diseases for which the discriminatory signal is consistent across studies and batches. In the mutation data analysis we used only "confirmed somatic" mutations, which are the most confident category (reducing false positive rate), and only from whole-genome studies, thereby reducing the false negative rate. We also used a very stringent approach for accepting a mutation: although we used the COSMIC data only for three cancer sites (out of 28 primary cancer sites in COSMIC) we corrected the p-values for association jointly for all sites together (0.05 FDR).

Finally, note that simplifying assumptions, while problematic, are often reasonable for detecting an average global signal. For example, Efron comments that when estimating FDR, dependence increases variance but is not expected to distort the mean FDR [15] . As another example, Cirrielo et al. [16] clustered patients using a statistical modularity score that also assumes independence across samples. Their analysis detected many known factors related to cancer and even suggested novel insights.

Single platform analysis in low performance diseases

In our multi-platform analysis 24 diseases were well classified. For each disease that was not well-classified, we tested leave dataset out cross validation using only samples from the platform with the highest number of datasets. We tested two ways to normalize the datasets: using our rank-based normalization and simple quantile normalization. Our results below show that for six disease terms the classification performance could be improved to achieve a ROC score > 0.7, using the quantile normalization.

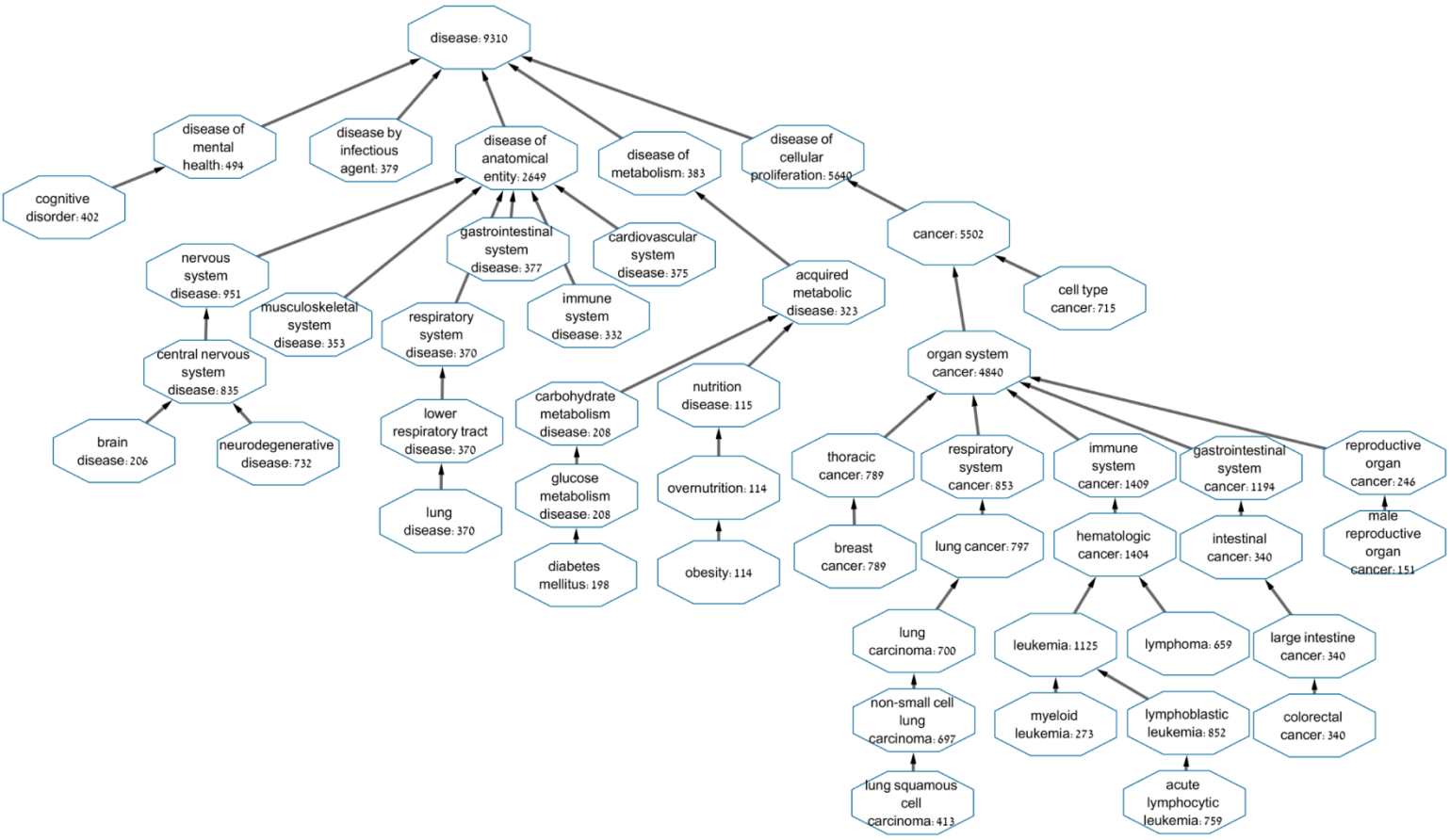
Disease	Platform	#samples	#datasets	Quantile	Ranks
disease of mental health	GPL570	417	6	0.63	0.67
lower respiratory tract disease	GPL570	331	6	0.57	0.52
respiratory system disease	GPL570	331	6	0.57	0.52
disease of anatomical entity	GPL570	1591	28	0.55	0.57
lung disease	GPL570	331	6	0.57	0.52
myeloid leukemia	GPL570	457	4	0.15	0.14
nervous system disease	GPL570	494	8	0.68	0.66
musculoskeletal system disease	GPL96	249	6	0.84	0.83
gastrointestinal system disease	GPL570	165	5	0.59	0.63
brain disease	GPL570	232	3	0.72	0.7
disease by infectious agent	GPL570	127	3	0.63	0.63
acquired metabolic disease	GPL570	210	6	0.5	0.48
diabetes mellitus	GPL570	125	3	0.71	0.6

disease of metabolism	GPL570	256	8	0.26	0.26
glucose metabolism disease	GPL570	125	3	0.71	0.61
carbohydrate metabolism disease	GPL570	125	3	0.71	0.61
Lymphoma	GPL570	412	4	0.82	0.79
reproductive organ cancer	GPL96	255	6	0.37	0.34
immune system disease	GPL570	104	2	0.45	0.44
cognitive disorder	GPL570	271	5	0.63	0.67

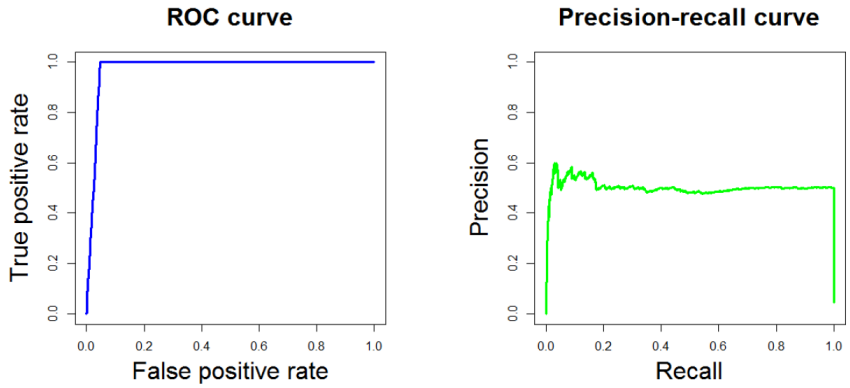
References

1. Yang, X., et al., *Single sample expression-anchored mechanisms predict survival in head and neck cancer.* PLoS Comput Biol, 2012. **8**(1): p. e1002350.
2. Yang, X., et al., *Similarities of ordered gene lists.* J Bioinform Comput Biol, 2006. **4**(3): p. 693-708.
3. Zhang, M.L. and Z.H. Zhou, *A Review on Multi-Label Learning Algorithms.* Ieee Transactions on Knowledge and Data Engineering, 2014. **26**(8): p. 1819-1837.
4. Tsoumakas, G., et al., *MULAN: A Java Library for Multi-Label Learning.* Journal of Machine Learning Research, 2011. **12**: p. 2411-2414.
5. Sucar, L.E., et al., *Multi-label classification with Bayesian network-based chain classifiers.* Pattern Recognition Letters, 2014. **41**: p. 14-22.
6. Vens, C., et al., *Decision trees for hierarchical multi-label classification.* Machine Learning, 2008. **73**(2): p. 185-214.
7. Blockeel, H., et al., *Decision trees for hierarchical multilabel classification: A case study in functional genomics.* Knowledge Discovery in Databases: Pkdd 2006, Proceedings, 2006. **4213**: p. 18-29.
8. Huang, H., C.C. Liu, and X.J. Zhou, *Bayesian approach to transforming public gene expression repositories into disease diagnosis databases.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6823-8.
9. Barutcuoglu, Z., R.E. Schapire, and O.G. Troyanskaya, *Hierarchical multi-label prediction of gene function.* Bioinformatics, 2006. **22**(7): p. 830-836.
10. Kourmpetis, Y.A.I., A.D.J. van Dijk, and C.J.F. ter Braak, *Gene Ontology consistent protein function prediction: the FALCON algorithm applied to six eukaryotic genomes.* Algorithms for Molecular Biology, 2013. **8**.
11. Lee, Y.S., et al., *Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies.* Bioinformatics, 2013. **29**(23): p. 3036-44.
12. Scutari, M., *Learning Bayesian Networks with the bnlearn R Package.* Journal of Statistical Software, 2010. **35**(3): p. 1-22.
13. Hedges, L.V. and I. Olkin, *Statistical methods for meta-analysis*. 1985, Orlando: Academic Press. xxii, 369 p.
14. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
15. Efron, B. and Institute of Mathematical Statistics., *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction*. Institute of Mathematical Statistics monographs. 2010, Cambridge: Cambridge University Press. xii, 263 p.
16. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nature Genetics, 2013. **45**(10): p. 1127-U247.

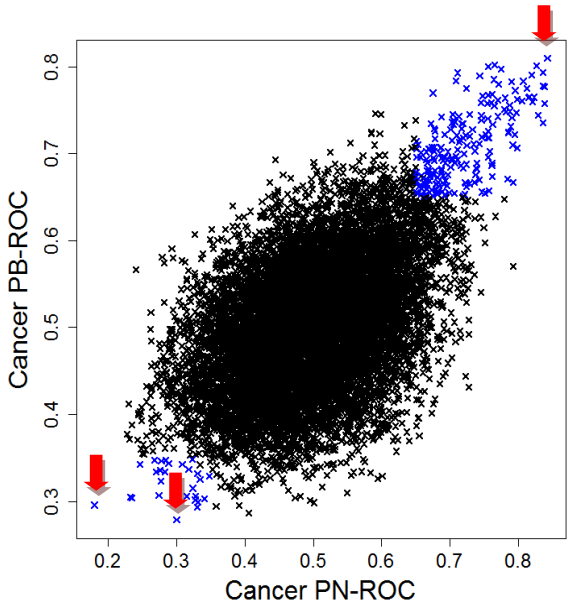
Supplementary Figures



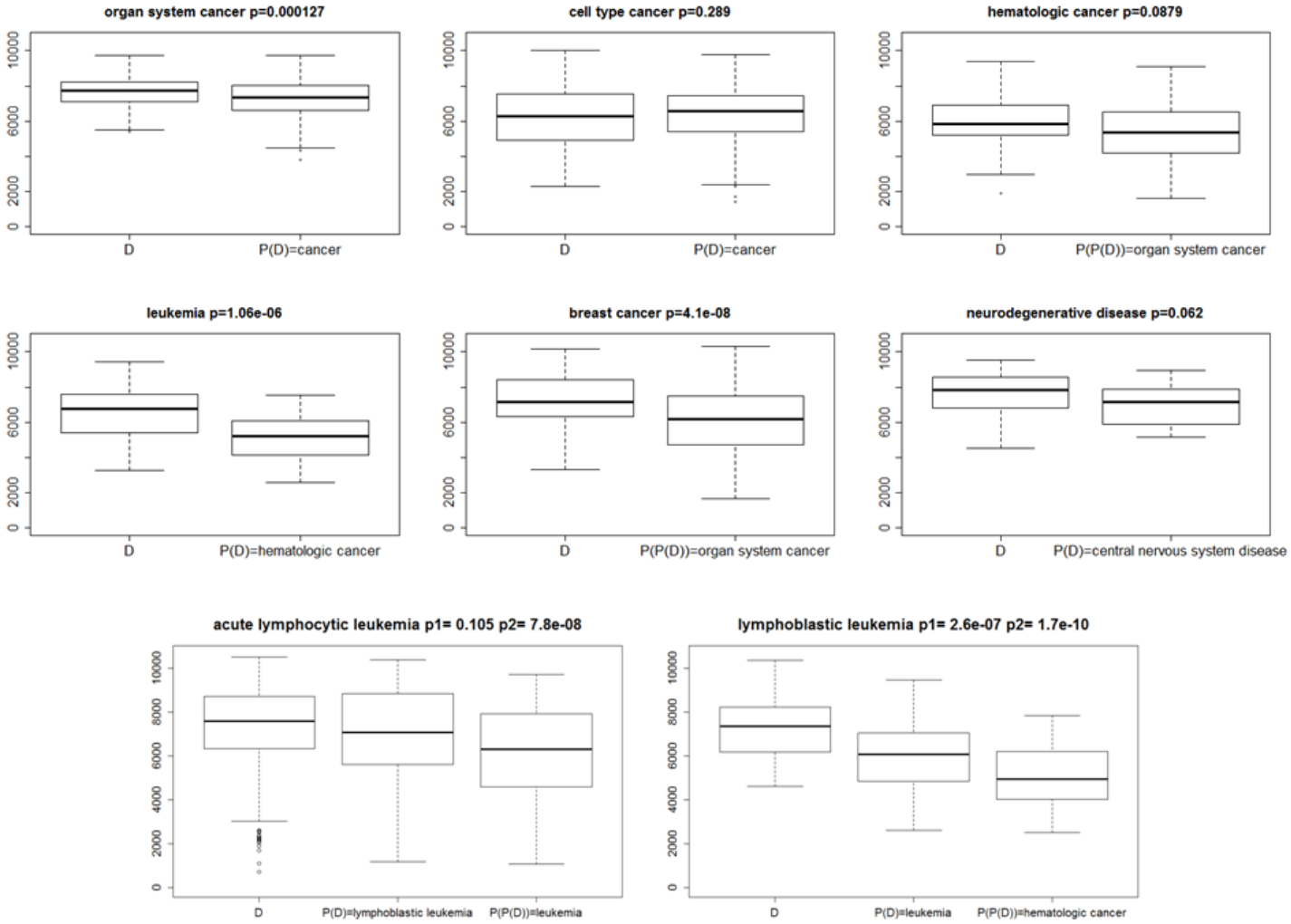
Supplementary Figure S1. The studied disease. The graph shows the 48 Disease terms in ADEPTUS that had at least 100 samples from at least 5 datasets. For each node, the Disease Ontology term and the number of positive samples are shown. Edges mark "is-a" relation in the DO hierarchy.



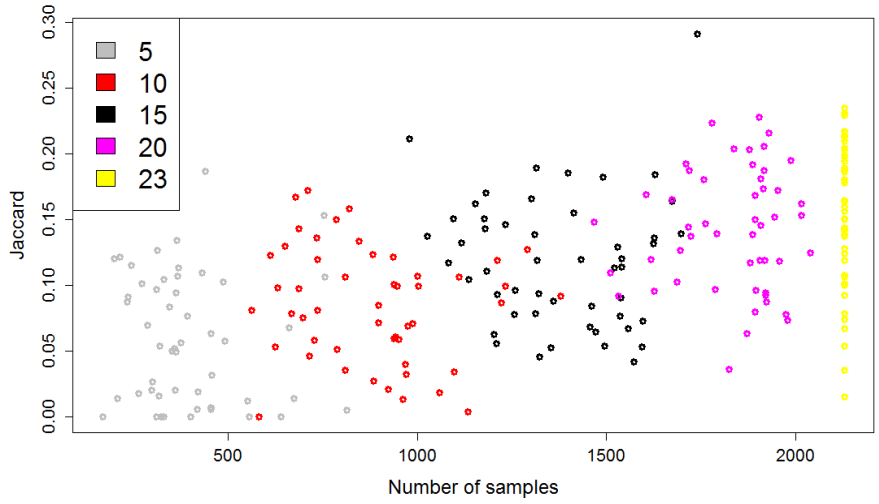
Supplementary Figure S2. Over-optimistic performance when ignoring the partition of non-positive samples into negatives and BGCs. The curves show the result of a simple simulation in which values of positive samples (n=500) and the values of negative samples (n=500) were drawn from the same distribution (U[0,0.5]), whereas the values of BGCs (n=10000) were drawn from a distribution with lower values (U[0.5,1]). The sizes of the classes are extremely skewed as expected in large databases (most samples are of the BGC class). The curves show the separation between the positives and all the rest. ROC and precision-recall were plotted using the sample values to rank the samples. The ROC AUC is 0.97, and the PR AUC is 0.5. These results are misleading, since positives and negatives cannot be separated in this case. Thus, this example illustrates a case in which a classifier can only distinguish BGCs from the rest. In such cases, ignoring the known partition of non-positive samples into negatives and BGCs produces inflated performance scores.



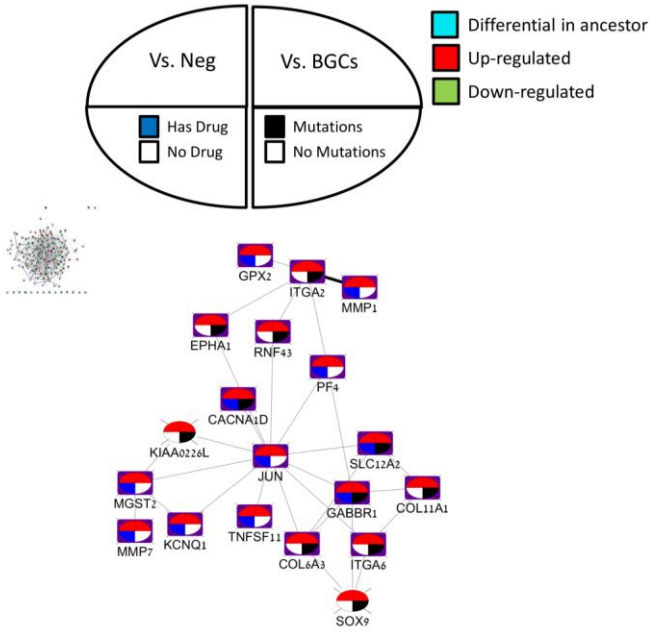
Supplementary Figure S3. High correlation between PN-ROC and PB-ROC scores of genes in cancer. The correlation between the two scores was 0.49. The cancer-specific differential genes are colored in blue. Note that here ROC scores are more significant the further they are from 0.5, in both directions. Also, note that unlike the main text, we plot here the raw ROC scores and not using max(x,1-x) function. This was done in order to better view the correlation between the PB- and PN- ROC scores. The two marked down-regulated genes are CRY2 (PN-ROC 0.3) and CBX7 (PN-ROC 0.18). The marked up-regulated gene is TOP2A.



Supplementary Figure S4. A comparison of disease gene sets with the sets of their ancestral diseases. For every gene g, we calculated the gene's z-score of differential expression for each pair (x,D) where x is a dataset and D is a disease. For every pair (x,D) we then ranked all genes by their absolute z-score (See Supplementary Text). For each disease D, we show (from left to right) (1) the distribution of the genes' median rank across D's datasets. (2) the distribution of the median rank of the biomarker genes of the parent disease (P(D)) across D's datasets, and (3) when available, the distribution of the median rank of the biomarker of the grandparent disease (P(P(D))) across D's datasets. The reported p-values are for the difference between the distributions of D and P(D), and when available, also of D and P(P(D)).



Supplementary Figure S5. Stability tests results by the number of samples. The plots show the Jaccard overlap score between solutions obtained using two disjoint sets of k disease datasets each. The x-axis shows the number of samples in each run, and the color shows k.



Supplementary Figure S6. Networks of molecular modifications in colorectal cancer. The network shows genes that were found specifically differential in the disease or in one of its ancestor DO terms, and for which either a targeting drug exists or the gene was found to be associated with the disease in COSMIC. Black edges are PPIs, and gray edges are GIs. Each node shows four features of a gene: (1) differential pattern compared to negatives, (2) differential pattern compared to BGCs, (3) whether a targeting drug exists, and (4) if the gene was associated to lung cancer according to COSMIC. Nodes without a purple background are genes that are not associated with any pathway in KEGG, Reactome, NCI, or Biocarta. Out of the original network of 454 genes, the figure focuses on 27 extremely up-regulated genes (PN-ROC > 0.8).

Published in *Bioinformatics* [2].

בשנים האחרונות יותר ויותר ניסויים מודדים שינויים במערכות לאורך זמן. שתי דוגמאות נפוצות כוללות מדידת רמות ביטוי גנים לאורך זמן לאחר התפרצות מחלה או לאחר מתן תרופה, ונתוני fMRI אשר מודדים פעילות של אזורים שונים במוח לאורך זמן. נתונים אלו מכילים שלושה מימדים: חולים, אובייקטים נמדדים, וזמן. בעבודה זו אנו מציגים אלגוריתם חדש לניתוח נתונים משלושה מימדים. האלגוריתם שלנו יכול לזהות קבוצות של אובייקטים (למשל גנים) אשר מראים פעילות גבוהה מתואמת לאורך זמן ובצורה עקבית על פני נבדקים שונים. לכל נבדק השיטה גם מזהה את נקודות הזמן הרלוונטיות. בנוסף, אנו מזהים הרחבות המתארות תבניות ייחודיות לנבדקים בודדים. האלגוריתם שלנו מבוסס על מידול הסתברותי בייזיאני של הנתונים ואלגוריתם דגימה גיבס למציאת תבניות באיכות גבוהה. השיטה שלנו נבדקה והראתה שיפור לעומת שיטות קיימות הן על נתונים מלאכותיים והן על נתונים אמיתיים. בפרט, השיטה משפרת בלפחות פי שניים את רמת הכיסוי של הנבדקים על ידי התבניות המזוהות. זהו שיפור משמעותי שכן מטרה עיקרית של הניתוח במקרים אלו היא לזהות תבניות אשר חוזרות על עצמן לאורך נבדקים שונים. בנתונים של ביטוי גנים בחולים לאחר הלם זיהומי (septic shock) זיהינו שתי תבניות המכילות גנים הקשורים למערכת החיסון. המידע הייחודי עבור הנבדקים מראה כיצד בחולים במצב קשה מאוד מצטברים עוד ועוד תהליכים חיסוניים פעילים לאורך זמן. בניתוח נתוני fMRI מנבדקים במצב מנוחה זיהינו את האזורים העיקריים במוח הקשורים לפעילותו במנוחה.

3. **Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets.**
   David Amar, Tom Hait, Shai Izraeli and Ron Shamir.
   Published in *Nucleic Acids Research* [3].

מדידת רמות הביטוי של גנים היא כלי בסיסי במחקר ביו-רפואי עכשווי. מאגרי מידע ציבוריים מכילים תוצאות מאלפי ניסויים שונים. ניתוח משולב של מספר רב של דגימות מניסויים שונים וטכנולוגיות שונות הוא קשה במיוחד, הן מסיבות ביולוגיות כמו ניתוח והבנה של מחלות שונות והן מסיבות טכניות וסטטיסטיות כמו אי תאימות בין טכנולוגיות. בעבודה זו אנו מציגים מסגרת חישובית לניתוח מספר רב של דגימות ושילובן עם תוצאות מניסויים מסוגים נוספים שאינם ביטוי גנים. תחילה, הרכבנו מאגר נתונים אשר מכיל מעל ל-14,500 פרופילים של ביטוי גנים. לכל נבדק מצאנו בעזרת ניתוח ידני את קבוצת המונחים בעץ המחלות המתארת את הפנוטיפ שלו. אנו מראים כיצד יש לבצע ניתוח נכון של הערכת טיב מסווג או בדיקה האם גן משתנה במחלה כאשר מנתחים נתונים מסוג זה (קרי, דגימות מהרבה מחלות ורקמות יחד). עבור מחלות בהן קיימים גנים משתנים באופן ספציפי אנו מציגים ניתוח משולב של התוצאות עם רשתות ביולוגיות, נתונים על מוטציות, וקשרי תרופה-מטרה. ניתוח זה מספק ראיה כללית של הגנים המשתנים במחלה יחד עם ההקשר הטיפולי שלהם. בנוסף, הניתוח מאפשר זיהוי תרופות מועמדות לייעוד מחדש: אלו תרופות שפותחו למחלות מסוימות אך יכולות לעזור במחלות חדשות. בשונה מניתוחים שונים בספרות אנו מספקים שיטה אוטומטית להצעת תרופות מועמדות לייעוד מחדש. לסיכום, השיטה המוצגת בעבודה זו משפרת בצורה משמעותית את היכולת לנצל נתונים קיימים שהצטברו במאגרי המידע הציבוריים לצורך אבחון וניתוח של מחלות.

השונים). לסיום אנו מציגים שיטה חישובית לניתוח משולב של מספר רב של דגימות ביטוי גנים מניסויים שונים. השיטה מאפשרת: (1) איחוי של נתונים מניסויים שונים ומטכנולוגיות שונות, (2) ניתוח של הרבה מחלות יחד, (3) זיהוי מחלות עבורן קיים סיגנל אמין בנתונים, (4) זיהוי גנים אשר משתנים במחלות באופן ספציפי ו-(5) ניתוח משולב של התוצאות עם נתונים ממוטציות, קשרי תרופה-מטרה, ורשתות ביולוגיות. ניתוח זה מזהה גנים מרכזיים במחלות הנבדקות ומציע שימושים חדשים לתרופות קיימות (drug repurposing).

תקציר המאמרים הכלולים בתזה

להלן תקצירי המאמרים עליהם מבוססת עבודה זו:

1. **Constructing module maps for integrated analysis of heterogeneous biological networks.**
   David Amar and Ron Shamir.
   Published in *Nucleic Acids Research* [1].

בעבודה זו, אנו מציגים אלגוריתם חדש לניתוח משולב של רשתות ביולוגיות. הקלט הוא זוג רשתות אשר מתארות סוגים שונים של אינטראקציה בין זוגות גנים. הסוג הראשון מתאר אינטראקציות "חיוביות" - קשתות שנצפה לראות בין גנים מתהלכים ביולוגים דומים. לדוגמה, רשתות של ביטוי משותף. הסוג השני מתאר אינטראקציות "שליליות" - קשתות שנצפה לראות בין גנים מתהליכים שונים. למשל, רשתות אינטראקציה גנטית שלילית מכילות קשתות בין גנים מזוגות תהליכים ביולוגיים שונים אשר יכולים לפצות אחד על האובדן של השני. הפלט של האלגוריתם שלנו הוא מפה המתארת את שתי הרשתות. במפה זו כל צומת מתאר קבוצת גנים עם רמת קישוריות גבוהה ברשת החיובית. כל קשת במפה מתארת רמת קישוריות גבוהה ברשת השלילית. ביצועי האלגוריתם שלנו טובים יותר לעומת שיטות קודמות הן על נתונים מלאכותיים והן על נתונים אמיתיים משלושה סוגים שונים. הסוג הראשון הוא ניתוח משולב של רשתות אינטראקציות חלבון-חלבון כרשת החיובית ורשת אינטראקציה גנטית שלילית. המפה המתקבלת גילתה אינטראקציות אפיסטטיות בין קבוצות גנים אשר פועלות יחד. הסוג השני הוא ניתוח משולב של רשתות אינטראקציות חלבון-חלבון כרשת החיובית ורשת אינטראקציה משתנה כתוצאה מטיפול ב-MMS. במקרה זה המפה שלנו מזהה קישורים חדשים מובהקים הנרקמים בין קבוצות חלבונים כתוצאה מהחשיפה ל-MMS. קישורים אלו יכולים להעיד על היווצרות תהליך חדש כתגובה לחומרים המזיקים לדנ"א. בסוג השלישי והאחרון ניתחנו רשתות ביטוי משותף כרשת החיובית ורשתות ביטוי משותף דיפרנציאלי כרשת השלילית. בנתונים מחולי סרטן, זיהינו מצבים בהם יש ירידה משמעותית ברמת המתאם בין קבוצות גנים הקשורות למערכת החיסון. עבודה זו מראה שסיכום רשתות שונות כמפה של קבוצות הוא כלי לניתוח מדויק ולהבנה טובה יותר של רשתות שונות ומורכבות.

2. **A hierarchical Bayesian model for flexible module discovery in three-way time-series data.**
   David Amar, Daniel Yekutieli, Adi Maron-Katz, Talma Hendler and Ron Shamir.

את רמת השינוי במתאם בין זוג גנים כשעוברים ממצב אחד למצב שני. למשל, הביטוי של זוג גנים יכול להיות במתאם גבוה כשהגנים נמדדים בקבוצת ביקורת המורכבת ממדידות מנבדקים בריאים, אך רמת המתאם יכולה לרדת משמעותית בקבוצת ניסוי המורכבת מחולים במחלה מסוימת.

עד כה תארנו נתונים ביולוגים המאפיינים תמונת מצב רגעית של תא חי. כיום, הודות לירידה משמעותית במחירי הניסויים, חוקרים יכולים לבצע ניסויים מורכבים יותר בהם ניתן לזהות שינויים במערכת הביולוגית לאורך זמן. שני הסוגים הנפוצים ביותר הם ניסויי הסטה (perturbation) ומדידות עיתיות (time-series). בניסויי הסטה בודקים את מצב המערכת לפני ואחרי שינוי מבוקר (למשל, מחיקה של גן או הוספת חומר ספציפי למצע הגידול של התא). מחקרים שמדדו אינטראקציות גנטיות בין זוגות גנים לפני ואחרי הוספת חומר הגורם לנזקים בדנ"א מסוג MMS זיהו רשתות דיפרנציאליות. כלומר, ניסויים אלו זיהו שינוי בקשתות הרשת לפני ואחרי יצירת הנזקים בדנ"א (כמו מחיקה והוספה של קשתות). חשיבות ניסויים אלו היא בזיהוי מנגנונים החשובים לתהליך בו התא מתקן שגיאות בדנ"א, מסלול שהוא בעל חשיבות מכרעת בהתפתחות סרטן. בנתונים ממדידות עיתיות בודקים את מצב המערכת לאורך k נקודות זמן (כאשר k הוא לפחות 3). למשל, ניתן לבדוק את רמות הביטוי של גנים בדם לאורך מספר ימים לאחר זיהום. דוגמה נוספת היא מעולם חקר המוח בו ניתן למדוד בעזרת דימות תהודה מגנטית תפקודית (functional magnetic resonance imaging - fMRI) את רמת הפעילות של איזורים במוח לאורך זמן. שתי הדוגמאות האחרונות מתארות סוג נתונים מעניין במיוחד - אלו הם נתונים בשלושה מימדים: חולים, אובייקטים נמדדים (גנים או אזורים במוח) ונקודות זמן. ניתוח נתונים מסוג זה דורש שיטות חישוביות אשר יכולות לזהות תבניות דומות בנבדקים שונים.

תוצאות

מאגרי מידע ציבוריים, כמו אלה של ה-NCBI, מכילים תוצאות של אלפי ניסויים רחבי היקף. מצד שני, נתונים אלו רחוקים מלהיות מושלמים שכן הם מורכבים מאוד, מכילים רמות גבוהות של רעש סטטיסטי, ולרוב קשה לפרש את תוצאות הניסויים. בנוסף, בהרבה מקרים קיימת סתירה בין תוצאות מניסויים שונים שבדקו השערה ביולוגית דומה. כדי לטפל בבעיות המוזכרות לעיל הדגש העיקרי בתזה זו הוא פיתוח שיטות חישוביות לניתוח משולב של נתונים מהרבה ניסויים יחד. מטרתנו היא לספק לקהילה המדעית כלים חדשים לזיהוי תבניות מעניינות ואמינות תוך שימוש במגוון רחב של ניסויים. השיטות המוצגות בתזה נבחנו והוכחו כעדיפות על פני שיטות קיימות. כל אחת מהשיטות יושמה על פני מגוון רחב של נתונים ביולוגיים. בכל יישום אנו מציגים את הערך המוסף של הניתוח המשולב בעזרת ניתוח מקיף של הנתונים.

בתחום הרשתות אנו מציגים שיטה חדשה לניתוח משולב של שתי רשתות ביולוגיות שונות. הניתוח מייצר מפה אשר מסכמת את שתי הרשתות ובה כל צומת היא קבוצת גנים וכל קשת היא קישור בין זוג קבוצות. המפה מאפשרת זיהוי של יחידות פונקציונאליות בסיסיות של גנים שפועלים יחד וגם קשרים בין קבוצות, פלט אשר לא מתקבל משיטות לזיהוי צבירים או מניתוח של כל רשת בנפרד. בניתוח זמן נתונים עיתיים משלושה מימדים אנו מציגים שיטה חדשה לזיהוי תבניות אשר מופיעות בנבדקים שונים. תבניות אלו מכילות קבוצות אובייקטים אשר פועלים יחד (גנים, או אזורים במוח) באופן עקבי, ואת נקודות הזמן הרלוונטיות בנבדקים. בנוסף השיטה מזהה תבניות ייחודיות בנבדקים בודדים אשר מהוות הרחבה נקודתית של התבנית הכללית (שזוהתה על פני הנבדקים

תקציר

רקע כללי

מערכות ביולוגיות מורכבות ממספר רב של מולקולות אשר פועלות יחד לביצוע מגוון רחב של פעולות. בתא החי קיימות מערכות שונות, ברמות בקרה שונות, שפועלות יחד כדי לאפשר לתא להתקיים בסביבות משתנות. לאורך העשורים האחרונים פיתחה הקהילה המדעית שיטות ביו-טכנולוגיות חדשות אשר יכולות לעקוב אחר פעילותן של אלפי ואף מיליוני מולקולות בבת אחת. בעזרת טכנולוגיות אלה ניתן למדוד ריכוזים של מולקולות ולזהות אינטראקציות בין מולקולות שונות. מדידת ריכוזים או פעילות של מולקולות נעשית כיום גם עבור נבדקים במצבים שונים. הפלט הוא לרוב מערך מספרי, הנקרא גם פרופיל מולקולרי, המתאר את מצב המולקולות בנבדק בודד. למשל, טכנולוגיות שבבי דנ"א וטכנולוגיות ריצוף מודדות באופן כמותי עשרות אלפי מולקולות רנ"א או דנ"א. לרוב מולקולות ה-רנ"א הנמדדות הן של רנ"א שליח ומדידת ריכוזיהן מהווה מדד עקיף לפעילות החלבונים בתא. מדידת התוכן והכמות של מולקולות דנ"א מאפשרת זיהוי של מוטציות סומאטיות המתרחשות בתאים סרטניים. בנוסף, בעזרת טכנולוגיות אלו ניתן למדוד רמות פעילות של מנגנוני בקרה בתא כמו רמות מתילציה בפרומוטורים של גנים, ומדידה כמותית של מולקולות מיקרו-רנ"א.

סוג נוסף של נתונים רחבי-היקף הוא רשתות ביולוגיות. רשתות ביולוגיות מתארות אינטראקציות בין מולקולות בתא. הן מספקות, לרוב בצורה מפשטת, תמונה כללית ומקיפה של המערכת הביולוגית. בעזרת מידול כמותי או איכותי של יחסי הגומלין בין מולקולות שונות, רשתות ביולוגיות משפרות את יכולתנו להבין בצורה מערכתית כיצד מתרחשים תהליכים בתא. רשתות ביולוגיות, ככל המודלים החישוביים, אינן מושלמות ומכילות הן שגיאות סטטיסטיות והן הנחות מלאכותיות שאינן תואמות תמיד לביולוגיה. אף על פי כן, לרשתות אלו שימושים רבים אשר הוכיחו עצמם בשנים האחרונות. למשל, רשתות ביולוגיות מהוות כלי עיקרי לחיזוי פונקציות של גנים וללמידת קישורים בין גנים ומחלות. בנוסף, ניתוח משולב של רשתות עם פרופילי ביטוי גנים יכול לספק תמונה טובה וכוללת יותר של מצב התא מאשר שימוש בכל אחד ממקורות המידע הללו בנפרד.

רשתות ביולוגיות לרוב מוגדרות על ידי צמתים וקשתות. הצמתים מתארים את היחידות הביולוגיות שהרשת מייצגת. לרוב אלו מולקולות כמו חלבון או תרכובת מטבולית, אך הן יכולות לתאר מושגים כלליים יותר כמו גן או קבוצת חלבונים שפועלים יחד (למשל הריבוזום). הקשתות מתארות אינטראקציות בין צמתים. למשל ברשתות אינטראקציות חלבון-חלבון (protein-protein interactions) הקשתות מתארות קשרים פיזיים בין זוגות חלבונים. ברשתות גנטיות הקשתות מתארות קשר מובהק בין זוג גנים לבין הפנוטיפ. אינטראקציה גנטית שלילית (negative genetic interaction) בין זוג גנים מתארת מצב בו מחיקת זוג הגנים גורמת לפנוטיפ רע מהצפוי, למשל יריד משמעותית ביכולת ההישרדות של התא. בהקשר זה, הפנוטיפ הצפוי מחושב על ידי מדידת הפנוטיפ לאחר מחיקת כל אחד משני הגנים בנפרד. באופן דומה אינטראקציה גנטית חיובית (positive genetic interaction) מתארת מצב בו מחיקת זוג גנים גורמת לפנוטיפ טוב מהצפוי. מחקרים קודמים הראו שאינטראקציות גנטיות חיוביות נפוצות בין זוגות גנים אשר פועלים יחד, ואילו אינטראקציות גנטיות שליליות נפוצות בין גנים מתהליכים שונים. ברשתות ביטוי משותף (co-expression) הקשתות מציינות מתאם גבוה ברמות הביטוי של זוג גנים לאורך קבוצת ניסויים. ברשתות ביטוי משותף דיפרנציאלי (differential co-expression) הקשתות מכמתות

תמצית

אנו חיים בתקופה מיוחדת בה המחקר הביו-רפואי נהפך למדע המושתת על עקרונות חישוביים. בעזרת התפתחויות טכנולוגיות פורצות דרך הצליחה הקהילה המדעית לאגור כמויות אדירות של נתונים ממספר רב של נבדקים. למרות התקדמות זו קיים פער גדול בין היכולת שלנו לבצע ניסויים ולאגור את תוצאותיהם לבין היכולת שלנו לדלות מידע שימושי מהם. לכן, יש צורך בשיטות חישוביות חדשות אשר יכולות לנתח בצורה משולבת הרבה נתונים. בתזה זו אנו מתארים מחקרים בהם פיתחנו שיטות חישוביות למטרה זו. הודות לשימוש בעקרונות מתורת הגרפים, מידול הסתברותי, ולמידה סטטיסטית, השיטות שפיתחנו יכולות לנצל בו זמנית נתונים בהיקף גדול וממגוון רחב מאוד. אנו מציגים תרומה משמעותית לשלושה תחומים מרכזיים בביולוגיה חישובית: רשתות ביולוגיות, ניתוח נתונים עיתיים, ואיחוד נתונים ממספר רב של ניסויים. בכל אחד מהתחומים הללו השיטות שלנו משיגות ביצועים טובים יותר משיטות קיימות, ניתן להגיע בעזרתן לתובנות חדשות ואף ניתן להיעזר בהן כדי להציע השערות חדשות שיכולות להוות בסיס למחקר עתידי.

אוניברסיטת תל-אביב TEL AVIV UNIVERSITY

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלבטניק

שיטות חישוביות לניתוח משולב של מידע ביו-רפואי מגוון ורחב-היקף

חיבור לשם קבלת תואר "דוקטור לפילוסופיה"

מאת **דוד עמר**

בהנחייתו של פרופ' **רון שמיר**

הוגש לסנאט של אוניברסיטת ת"א

דצמבר 2015