Identification of Differentially Expressed Gene Modules in Heterogeneous Diseases

Olga Zolotareva

Faculty of Technology
Bielefeld University

This dissertation is submitted for the degree of
Doctor rerum naturalium

April 2021

Olga Zolotareva
German-Canadian DFG International Research Training Group 1906
Computational Methods for the Analysis of the Diversity and Dynamics of Genomes,
Bioinformatics & Medical Informatics Department, Faculty of Technology, Bielefeld University

Referees:

Prof. Dr. Ralf Hofestädt
Bioinformatics & Medical Informatics Department,
Bielefeld University, Germany

Prof. Dr. Martin Ester
School of Computing Science,
Simon Fraser University, Canada

Ph.D., Dr.Sci. Yuriy Lvovich Orlov
Chair of Information and Internet Technologies,
Institute of Digital Medicine,
Sechenov University, Russia, and
Novosibirsk State University, Russia

Printed on permanent, non-aging paper according to DIN ISO 9706.

Acknowledgements

First of all, I would like to thank my supervisors Prof. Dr. Ralf Hofestädt and Dr. Martin Ester for the opportunity to work with them on such exciting projects. I am grateful for their guidance, constructive discussions of the results, teaching me how to write research papers, and for their great attitude and patience.

I would also like to thank all my colleagues, who contributed to the projects described in this thesis for efficient collaboration. Many thanks to Dr. Sahand Khakabimamaghani who proposed to add a probabilistic clustering step to our method and from whom I learned a lot. I am very thankful to Alexey Savchik for the efficient reimplementation of the first step of our method, to Zoe Chervontseva for performing permutation tests and dockerizing the method, and to Olga Isaeva for biological interpretation of biclustering results. Thanks to Cassandra Königs and Maren Kleine for their great contributions to asthma and hypertension project and to the study of gene prioritization methods. Thanks a lot to Cassandra for proofreading the draft of this thesis.

I would also like to acknowledge all members of the International Research Training Group DiDy and all members of Martin Ester group for the great time in Bielefeld and Vancouver. Many thanks to Dr. Roland Wittler for providing a lot of help and advice.

Besides my supervisors and colleagues, I am very grateful to my parents who supported me during the whole time of my education. A big thanks to Dr. Nikolay Zolotarev for encouraging me on the way and for constant willingness to help.

I also acknowledge the International Research Training Group GRK/1906 “Computational Methods for the Analysis of the Diversity and Dynamics of Genomes” (DiDy) for financial support covering the first three years of my work and AG Bioinformatics and Medical Informatics for providing the Doctorate Completion Scholarship for the last six months.

Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Some of the results presented in this thesis have been previously published in:

O. Zolotareva, S. Khakabimamaghani, O. Isaeva, Z. Chervontseva, A. Savchik and M. Ester. Identification of Dysregulated Gene Modules in Heterogeneous Diseases. *Bioinformatics* 2020 Dec 16; Oxford University Press; doi:10.1093/bioinformatics/btaa1038 (subsections 2.6, 4.1-2, 4.5-6, and 5.1-2),

O. Zolotareva, and M. Kleine. "A survey of gene prioritization tools for Mendelian and complex human diseases". *Journal of Integrative Bioinformatics*. 2019 Sep 9;16(4); Walter de Gruyter GmbH; doi:10.1515/jib-2018-0069 (subsections 2.2 and 2.7.1),

O. Zolotareva, O. Saik, C. Königs, E. Bragina, I. Goncharova, M. Freidin, V. Dosenko, V. Ivanisenko, and R. Hofestädt. Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. *Scientific Reports* 2019 Nov 8;9(1):16302; Springer Science and Business Media LLC; doi:10.1038/s41598-019-52762-w (subsection 2.8),

A. Shoshi, R. Hofestädt, **O. Zolotareva**, M. Friedrichs, A. Maier, V. A. Ivanisenko, V. E. Dosenko, E. Yu. Bragina. GenCoNet - A Graph Database for the Analysis of Comorbidities by Gene Networks. *Journal of Integrative Bioinformatics*, 15(4), 2018; Walter de Gruyter GmbH (subsection 2.8).

Olga Zolotareva
April 2021

Abstract

During the last decades, the active development of high-throughput methods led to the discovery of numerous associations between biomolecules and human diseases. This is a great advance for science and medicine since it helps to unravel the mechanisms of the diseases and gives clues for new treatment approaches. At the same time, a tremendous amount of raw experimental data and biomedical knowledge became a great challenge for the researchers. This stimulated the development of automatic solutions for discovery, storage, retrieval, integration, and analysis of biological data.

Associations between genes and diseases have attracted a special interest of researchers. To date, thousands of rare inheritable diseases caused by disruptions of individual genes are described. At the same time, the most widespread disorders are multifactorial, i.e. develop in the result of complex interactions between multiple genetic and environmental factors. Such multifactorial diseases are therefore called complex, and their mechanisms remain far from being understood in our days. To achieve a complete mechanistic picture of a complex disease, it is first necessary to establish a comprehensive list of all pathological changes. This task is complicated by the fact that complex disease may be heterogeneous, i.e. symptomatically similar, but caused by distinct molecular lesions.

The investigation of complex diseases not only brings us closer to the understanding of their mechanisms but also yields a number of useful intermediate results, e.g. the discovery of clinically relevant biomarkers and disease subtypes. This thesis starts from the discussion of the approaches for the discovery and prioritization of gene-diseases associations, and their relevance for complex and heterogeneous diseases. It further focuses on biclustering methods which seem to be very promising in the context of disease heterogeneity. They are capable of identifying genes with a similar expression pattern in a previously unknown subset of samples.

After an overview of existing biclustering methods, this thesis presents a novel biclustering method called DESMOND. Two factors distinguish DESMOND from most of the related works. First, it searches for differentially expressed biclusters, rather than biclusters

with co-expression. Second, it performs a network-constrained search when the majority of biclustering methods are unconstrained.

DESMOND and nine previously published biclustering methods have been applied to simulated data and real breast cancer expression profiles. All the evaluated methods produced very diverse but biologically meaningful biclusters. On the breast cancer datasets DESMOND tended to produce more biologically singnificant gene clusters than the competitors. Compared to baselines, DESMOND and QUBIC identified more similar OS-associated biclusters in two independent breast cancer studies than other methods, possibly owing to their ability to consider gene interactions. Interestingly, these replicated biclusters found by DESMOND and QUBIC were composed of different genes and samples. Such OS-associated biclusters replicated in independent datasets may represent clinically different disease subtypes and are promising biomarker candidates.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AR Androgen Receptor

CC Biclustering method proposed by Cheng and Church

cDNA Complementary DNA

DNA Deoxyribonucleic Acid

EM Expectation-maximization

ER Estrogen Receptor

GO Gene Ontology

GWAS Genome-Wide Association Studies

HER2 Human Epidermal Growth Factor Receptor 2

MAFIA Algorithm for Finding Maximal Frequent Itemsets

mRNA Messenger RNA

MSE Mean Squared Error

MSR Mean Squared Residue

OS Overall Survival

PCR Polymerase Chain Reaction

PPI Protein-Protein Interactions

PR Progesterone Receptor

RNA-seq Sequencing of RNA

RNA Ribonucleic Acid

SMSR Scaling Mean Squared Residue

SSSim Shifting and Scaling Similarity

TCGA The Cancer Genome Atlas Program

VE Virtual Error

Chapter 1

Introduction

Since the development of any disease is the result of the dysregulation of a certain biological process, the detection of molecular entities participating it is necessary for understanding its pathogenesis. During the last decades, the progress in high-throughput technologies led to the aggregation of a large amount of data on the diversity and dynamics of multiple molecular entities. By analogy with genomics, fields of science that study large spectra of these molecular entities were called transcriptomics, epigenomics, proteomics, metabolomics etc. The shared suffix of these terms "omics" began to be used for collective designation of studies, dealing with multidimensional molecular profiles. Investigation of omics profiles in the context of human diseases led to the discovery of many diseases biomarkers – measurable indicators associated with certain features of the disease, e.g. predisposition, disease subtype, severity, prognosis, or treatment response.

The discovery of biomarkers has great importance not only for molecular medicine but also for fundamental research. Since genes encode functional units of cellular machinery, researchers paid special attention to biomarkers reflecting gene functions. Various alterations affecting functions of genes or gene products are shown to trigger diseases, or at least to be involved in their development. Chapter 2.1 of this thesis discusses the approaches for establishing gene-disease associations.

Rapid accumulation of data necessitated the creation of comprehensive databases on gene-disease associations. This, in turn, stimulated the development of computational tools for automatic search and retrieval of the most relevant associations (chapter 2.2, based on [299]). So why are the mechanisms of most diseases still unresolved despite the large number of gene-disease associations discovered? The identification and correct interpretation of key molecular players participating in the diseases are complicated by multiple problems, two of which are in the main focus of this thesis:

- some diseases are complex, which means that multiple genetic and environmental factors contribute to their development (chapter 2.3);
- diseases may be heterogeneous, when the same symptoms manifest in the result of different molecular lesions (chapter 2.4).

Both of these problems motivated the researchers to analyze molecular profiles as a whole in addition to the discovery of isolated gene-disease associations.

The signs of disease heterogeneity and complexity are observed on multiple molecular levels [24, 181, 216]. Recent multi-omics studies demonstrated that gene expression data makes the main contribution to the performance of the models for patient stratification and drug response prediction [61, 99]. This observation makes gene expressions the most promising data type for investigation of disease heterogeneity and complexity, which are in the focus of this thesis.

The methods of gene expression profiling and the ways of downstream analysis are discussed in chapters 2.5 and 2.6 respectively. Briefly, the output of any high-throughput expression profiling method results in a 2D matrix comprising profiles of n genes measured in m samples. The most important steps of the gene expression analysis pipeline include mapping of measured signals to genes or transcripts, and further quantification and normalization [56]. The further analysis depends on the research question posed and on the availability of sample class labels. The most widespread experiment design includes the search of genes significantly differentially expressed between two known groups of samples. Some workflows include an additional step of dimensionality reduction before the detection of differentially expressed genes. The expressions of multiple genes may be collapsed into a single feature using various clustering techniques [148], or based on prior knowledge about their functional relationships [257]. Such dimensionality reduction helps to decrease the risk of overfitting and improve reproducibility of the results.

When class labels are unavailable, one may perform clustering of samples in order to find unknown but biologically relevant subgroups. However, clustering of samples also may not work well in the case when multiple patterns appear in distinct subspaces of expression profiles.

Biclustering methods search for subsets of genes demonstrating similar expression patterns in a subset of samples in gene expression matrix. In contrast with conventional clustering, biclustering methods are able to address disease heterogeneity and therefore are in the focus of this thesis. To date, more than 50 biclustering methods aimed at various expression patterns are published [206, 223]. However, many of them are aimed at the de-

tection of biclusters composed of differentially co-expressed genes rather than differentially expressed. These two patterns of dysregulation represent perturbations with distinct biological significance. Differential expression reflects induction or inhibition of a certain pathway or its downstream. In turn, differential co-expression points to a gain (loss) of co-regulation and highlights regulatory network rewiring in a certain subgroup of samples. Both of these dysregulation types have a great significance for researchers. However, co-expression of two or multiple genes can be only calculated when a group of samples is specified. For a single expression profile, drawing a conclusion about the correlations between gene expressions is impossible, but one can compare gene expressions with the reference values. Therefore, biomarkers based on differential expression appear to be more suitable for clinical use. Given this consideration and the fact that not many biclustering methods are aimed at differentially expressed biclusters, we decided to focus on this specific type of bicluster. Chapter 3 of this thesis provides an overview of biclustering approaches, discusses their applicability for the detection of differentially expressed biclusters, and gives detailed descriptions of nine state-of-the-art biclustering methods.

Since biclustering is a much more complex problem than conventional clustering due to the much larger size of the search space, many biclustering methods put additional constraints on the input data or the biclustering result, e.g. they assume a hidden checkerboard structure [108, 268] of the data, discretize or binarize expression values [227, 242], or incorporate additional data supporting functional relationships of genes [114, 228]. Indeed, taking into account known interactions between the genes may reduce the complexity of the problem. Instead of considering all possible gene subsets, we suggested performing a network-constrained biclustering, i.e., to search for subsets of up- or down-regulated genes that form a connected component in the interaction network [298]. This constrained problem may be also represented as an unsupervised version of active subnetwork detection problem [118], when the groups of samples are unknown.

To solve the formulated problem, a new method for identification of **D**ifferentially **E**xpre**S**sed gene **MO**dules i**N** **D**iseases called DESMOND has been developed [298]. Chapter 4 of this thesis describes in deep detail the method previously published in *Bioinformatics* [298] and also presents its second version. The ability of DESMOND to incorporate prior knowledge about gene interactions directs the search towards biologically meaningful sets of genes and promises to improve the quality of the results. In contrast with the other methods, instead of setting a hard binarization threshold on gene expressions, DESMOND uses flexible thresholds to determine sample groups where genes are differentially expressed. To identify these thresholds, DESMOND analyses expressions of gene pairs connected in the network.

In the second version of DESMOND thresholds are defined for each individual gene using mixture models.

Chapter 5 reports the results of both DESMOND versions and their competitors on synthetic data and on real expression profiles from two large breast cancer datasets comprising more than 3000 tumor samples in total. Breast cancer is one of the most frequent cancer types, more than 2 millions new cases were diagnosed in 2018 worldwide [28]. It is a fatal disease, which course and treatment response is hard to predict. The understanding that cancer mechanisms lie mostly at the molecular and cellular levels motivated the scientist to investigate molecular profiles of tumors.

It has long been known that tumors with distinct molecular features are amenable to treatment in varying degrees [74, 207]. Given that most current anti-cancer treatments cause severe side effects [146, 201], selection of the optimal treatment plan has a great clinical significance. Many recent studies have demonstrated that gene expression profiles of tumors allow stratifying patients according to disease prognosis [209, 216, 226, 267] or predict potential drug response [64, 87, 246, 245, 253].

Current classifications of breast tumors are based on various tumor characteristics and propose different numbers of clinically distinct breast cancer subtypes. In particular, two widely used classifications consider gene expressions: one is based on ER, PR, and Her2 expression [93], and the other distinguishes four [216] (later five [102, 226]) intrinsic subtypes defined by the expressions measured by the PAM50 gene panel [209]. However, these well-established classifications only partially explain molecular heterogeneity of breast tumor and appear to be insufficient for precise prognosis and prediction. The discovery of unknown molecular subtypes may advance optimal therapy selection and give a clue for the design of new treatments, and therefore is of great importance for modern medicine.

Chapter 2

Background

2.1 Associations between genes and diseases

Same as other phenotypic traits, predispositions to many human diseases, from inborn disorders [8] to infectious diseases [38], are inheritable at least partially. Identification of genetic markers, correlated with disease status became one of the primary goals of medical genetics. These markers represented sequence variants, altering gene function, which eventually leads to the development of the disease. Over the past few decades, first genetic mapping and later genome-wide association studies (GWAS) allowed establishing a large number of associations between genes and diseases [5]. Comprehensive databases of genetic associations have been created to facilitate access to this data [6, 59, 147, 173]. These databases are actively replenished and frequently used as starting points for biomedical research projects.

Besides the correlation between the presence of a certain genetic variant and disease-related phenotypes, there are numerous ways to establish a gene-disease association. For example, alterations of amount, localization, structure, or modifications of a transcript or a protein in the disease may be the result of the dysfunction of a corresponding gene. Moreover, novel gene-disease associations may be computationally predicted using known similarities between genes (e.g. functional or sequence similarity) and diseases (e.g. symptom similarity).

The investigation of molecular mechanisms has shown that disease manifestations are accompanied by multiple changes at various molecular levels [53, 58, 86]. However, this does not mean that all of these changes are equally important in the context of a given disease. The identification of essential molecular players is a necessary step towards understanding pathogenesis. To determine genes are most likely to play a central role in a specific disease, the researchers consider orthogonal data which can support or contradict tested associations.

Given the huge number of genetic associations and the volume of prior knowledge about genes and diseases, automatization of this task is of great relevance.

2.2 Methods for prioritization of genes and their limitations

Modern high-throughput experiments produce hundreds or thousands of potential associations between genes and diseases requiring further exploration. In parallel with the simplification of the candidate gene search, the amount of available information about genes, their associations with other biological entities, increased. The emergence of various biological databases and the explosive growth of relevant scientific publications further complicated manual evaluation of candidate genes and stimulated the development of computational methods and tools for gene prioritization. This section reviews and classifies the existing gene prioritization tools and discusses their limitations. It is adapted from [299] published by *Journal of Integrative Bioinformatics (Walter de Gruyter GmbH)*.

Gene prioritization problem could be formulated as follows: rank candidate genes in decreasing order of probability to be truly associated with the disease based on prior knowledge about these genes and the disease. A typical gene prioritization tool is composed of two parts: a collection of evidence sources (i.e. databases of associations between genes, diseases and other biological entities) and a prioritization module (Fig. 2.1). Prioritization module takes two inputs: training data, which is used to define a phenotype of interest and testing data, a set of user-defined candidate genes to prioritize. After that, it extracts information about given genes or terms from evidence sources and calculates a score that reflects the "likelihood" of each gene to be responsible for the phenotype. Training data could be represented either by genes that were previously linked with a phenotype (*seed genes*). Alternatively to seed genes, some tools, e.g. PolySearch2 [165] or PhenoRank [60], accept phenotype or disease terms defining relevant gene-disease associations. The second part of the input is a set of candidate genes to prioritize or in some cases, the whole genome. The output of the program is a list of candidate genes arranged according to calculated scores or p-values. Every gene prioritization tool represents a unique combination of evidence sources, prioritization strategy and input requirements.

At the moment, hundreds of research papers on gene prioritization have been published and about a hundred of them describe computational tools [299]. Gene prioritization tools were extensively applied for prediction of genes involved in human diseases [1, 77, 140, 154,

175, 217, 263, 286], and other polygenic traits [252, 271]. In addition to the evaluation of gene relevance for single diseases, gene prioritization was used for the selection of genes potentially responsible for the comorbidity between two complex diseases – asthma and hypertension [238]. Moreover, taking into account the predicted importance of candidate genes, i.e. score assigned in the result of prioritization, improved the results of pathway enrichment analysis [217, 252, 271], enhanced models for drug response [75], and disease outcome predictions from gene expression profiles [31].

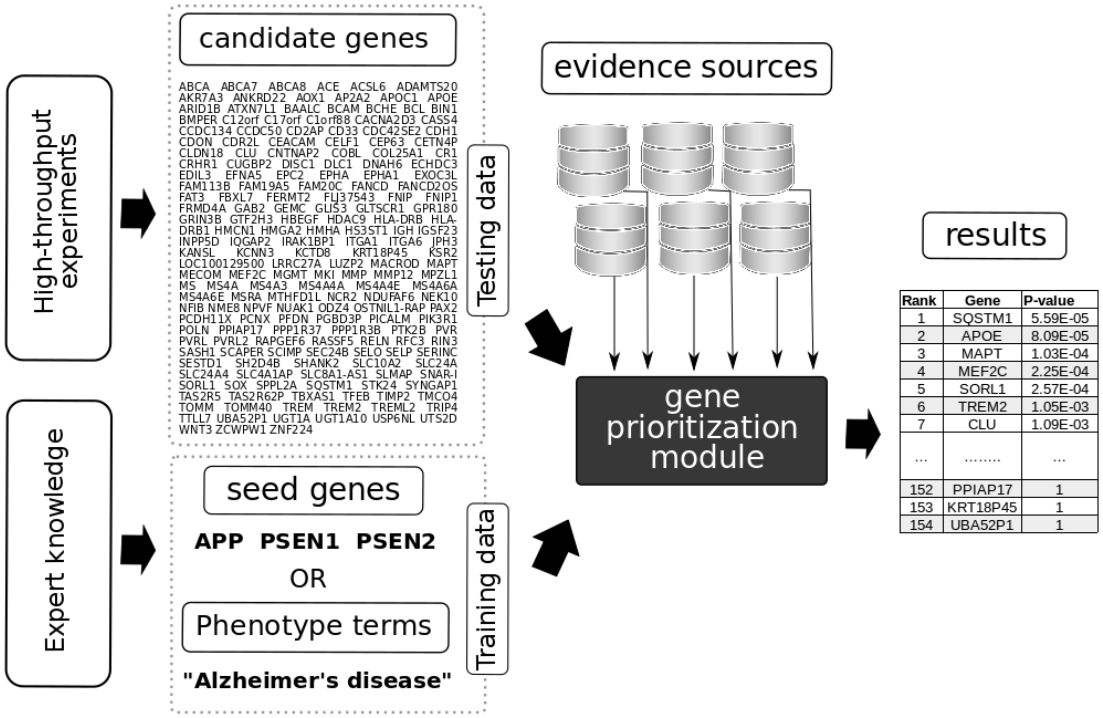

Fig. 2.1 The scheme of a gene prioritization tool. Gene prioritization tool extracts information about specified candidates and seed genes or phenotype terms defining the phenotype from evidence sources and calculates a score that reflects the "likelihood" of each gene to be responsible for the development of a phenotype. In this example, genes which have alleles causing an early-onset autosomal dominant familial form of Alzheimer's disease [20] are used as seeds. Candidate genes were obtained from GWAS Catalog [173]. Each candidate gene has at least one variant associated with Alzheimer's disease. The output of the program is a ranked list of candidate genes arranged according to calculated scores. The figure is reprinted from [299].

In previous works, gene prioritization tools have been classified based on the scope of their application (generic or disease-specific) [220], types of evidence sources used [70],

approaches (filter-based selection or ranking) [192, 220] and method types (network analysis, similarity profiling or text-mining) [90, 192]. We proposed two non-exclusive classifications of gene prioritization tools according to the assumptions they rely and data representation they use.

2.2.1 Assumptions behind prioritization strategies

Gene prioritization strategies rely on two major assumptions. First, genes may be directly associated with a disease, if they are systematically altered in the disease compared to controls (e.g. carry a disease-specific variant). Although various associations may have different strengths and qualities, it is assumed that association, supported by multiple independent studies is more likely to be true. Second, genes can be associated with a disease indirectly, via *guilt-by-association* principle, assuming that the most probable candidates are somehow linked with genes or other biological entities that were previously shown to impact the disease.

Two types of prioritization strategies can be distinguished, depending on the assumption they rely on and, consequently, on the kind of prior knowledge used to solve the prioritization problem. Strategies of the first type integrate all associations of each candidate with the disease of interest and into the overall association score. Such tools require the user to provide keywords or ontology terms specifying the disease and then integrate gene-disease associations of various kinds.

Approaches of the second type reduce the gene prioritization problem to the task of finding genes closely related to known disease genes. They accept a set of seed genes, implicitly defining the disease, instead of specifying the disease explicitly. These tools consider direct and indirect associations between genes, and prioritize candidates by their similarity and/or proximity to a set of seeds.

The majority of tools utilize exclusively one of these two strategies, some tools implement a both of them at different stages. For example, PhenoRank [60] and Phenolyzer [287] accept disease keywords, automatically construct a scored list of seed genes, and rank the rest of genes such that genes associated with high-scored seeds also get higher ranks. Another example is NetworkPrioritizer, which retrieves genes associated with a query disease, builds disease-specific network and identifies the most relevant genes based on the network topology [130].

2.2.2 Data representation

The structure of evidence sources utilized by a gene prioritization tool can be either relational (Figure 2.2A), when data sources are represented by a collection of tables, containing an association of a particular kind, or network (Figure 2.2B), where nodes correspond to genes (or other entities) and edges represent relationships between them. Although these two data representations models are interchangeable, organization of evidence sources is always consistent with the prioritization algorithm. Accordingly, most of the existing approaches can be classified as score aggregation or network analysis methods, or represent their combination [185, 284].

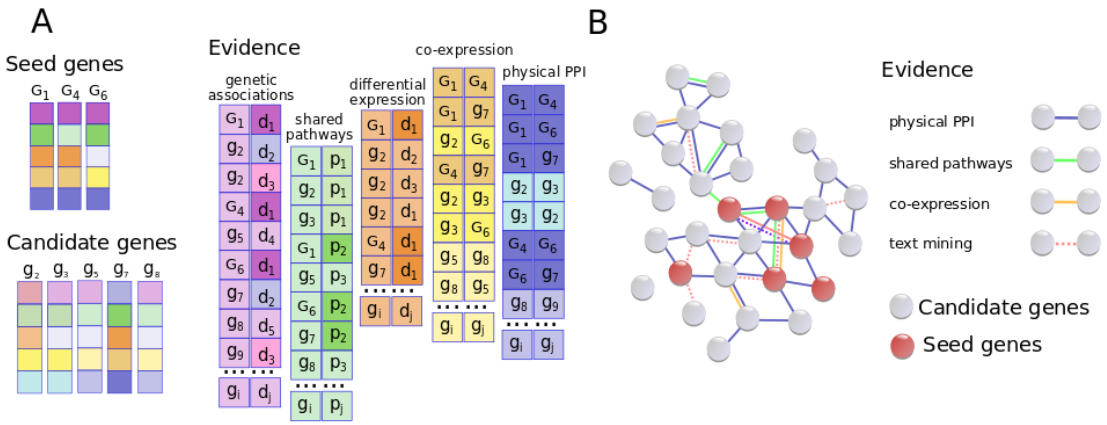

Fig. 2.2 Data representation models utilized by gene prioritization tools. A. Relational data structure. The first and the third evidence sources provide relationships between genes labeled with G (seeds) or g (candidates) and diseases (d), the second source provides gene membership in pathways (p) and the last two evidence sources contain different kinds of interactions between genes. Vector representation of seed and candidate genes are shown on the left. The similarity between colorings of gene g_7 and seed genes shows that g_7 seems to be a promising candidate. B. Network data structure. Nodes depict genes, edges show relationships between genes. Seed genes are shown red. The figure is reprinted from [299].

Network analysis.

Network is a natural representation of pairwise entity relationships, widely used to describe similarities or interactions between biological objects. Many independent studies agree that disease-associated proteins are located close to each other on the PPI network [92, 119, 184]. This observation became the basis for many prioritization approaches assuming that proteins, encoded by unknown disease genes and more tightly connected with known disease proteins

than irrelevant proteins. Moreover, disease proteins demonstrate special network properties (see section 2.8). This observation is used in network-based gene prioritization workflows [27, 130, 205, 237] including also those omitting seed genes [31, 238]. Briefly, these methods retrieve all genes associated with a query disease, build disease-specific interaction network and determine the essentiality of all nodes considering the network topology. Similarly, NetworkPrioritizer [130] ranks candidates according to various network centrality measures, e.g. betweenness and closeness centralities for a set of seed nodes and provides tools for aggregation and comparison of obtained rankings.

The majority of network-based tools require seed genes for input and rank candidates according to their proximity in the network to seeds. The distance from a node to a group of nodes in the network may be defined in numerous ways. MaxLink [94, 204] ranks first neighbours of known genes according to a number of direct links to them. In order to reduce hub bias, it takes into account only candidates which have significantly more connection with seed genes than expected by chance. Similarly with MaxLink, DIAMOnD [89] applies hypergeometric test to detect candidates enriched by seed genes among their first neighbours. In contrast, DIAMOnD ranks genes according to significance of seed overrepresentation among the first neighbours. In every iteration, DIAMOnD includes the most significant candidate into a set of seeds and recalculates p-values for the remaining candidates regarding the updated set of seeds. However, MaxLink considers only first neighbours of seeds and DIAMOnD ignores indirect interactions on every iteration. Gentrepid [88] ranks candidates conforming to the shortest path distance to a seed node. NetShort method implemented in GUILD framework [96] down-weights edges connected to genes with a high score when computing shortest path length. The disadvantage of this measure is that not all the pathways are equally informative, e.g. a path going through promiscuous hub nodes may be short but unspecific to the disease mechanism. In order to address this problem, various network propagation methods, modeling information flow over the network have been developed. ToppNet [42], GeneWanderer [141], PhenoRank [60] and many others [78, 125, 152, 153, 161, 169, 170] apply random walk-based algorithms [281, 139, 283] in order to assess relative importance of a node to a group nodes considering the global network topology. Other methods mathematically related [62] with random walk, modelling diffusion [81, 193, 274] or electric current flow [259] through the network have been used successfully in gene prioritization. GeneMANIA [194] implements Gaussian field label propagation algorithm [297], which redistributes seed gene scores to their neighbors, minimizing differences between both scores of neighboring genes and initial and assigned scores of seed genes. PRINCE [274] uses conceptually similar approach to smooths influence of disease genes

over the network. It simulates the exchange of flows between genes in the network, where every node produces outcoming flows to neighbors, proportional to its score, and computes a new score summarizing incoming flows. The process starts from disease genes and stops after many iterations. In the result, candidates connected with many disease genes gain higher income flow and thus a higher score. eQED [259] represents the network as an electric circuit where seeds are current sources, edges are conductances, candidates are drains, and rank candidates by current flowing through them. Köhler et al., 2008 [141], Navlakha and Kingsford 2009 [199], and Shim et al., 2015 [249] have shown that methods considering global network topology demonstrate higher overall performance than methods based only on local network information. At the same time, methods using local network topology, e.g. direct interactions or shortest path distances, rank true top-ranked candidates higher [95, 249] and therefore are more successful for diseases with few associated genes, tightly connected in the network [249].

Another important feature determining the performance of the network analysis tool is the network type used, its quality and completeness [96]. Some network-based prioritization tools use homogeneous networks modeling only one type of interactions [42, 46, 129, 165]. However, recent studies demonstrate that composite networks, composed of many various kinds of interactions and relationships, outperform any single network, possibly because individual networks contain complementary information [112, 159]. Therefore, many gene prioritization tools use functional protein interaction networks such as GeneMania [280], FunCoup [240], STRING [124] or integrate several networks of different types [89, 287]. Moreover, in previous works gene prioritization is performed on heterogeneous networks including multiple types of biological entities [104, 125, 161, 169, 178, 293].

Score aggregation.

This group includes tools implementing various strategies of aggregating all found associations into a total score. For example, Polysearch [46, 165] recognizes sentences supporting gene-disease associations, weights them according to their reliability and summarizes weights into the total relevancy score. When relevancy scores computed for all genes, Polysearch standardizes them and uses them for prioritization. Similarly, DisGeNET [15, 218, 219] and Open Targets [142] integrate data from multiple evidence sources. For each gene, they compute a weighted sum over all individual gene-disease association scores. Each weighting coefficient is determined by the reliability of association and the type of data source it came from. Thus, strong genetic associations discovered in humans make a bigger impact into

the overall gene score, than less reliable associations inferred from animal models or text mining.

Tools operating with seed genes employ similar ideas to summarize gene-gene associations. Initially, they score each candidate by its similarity with seeds, considering each evidence source independently, and then combine all data source-specific scores into a final score. GPS [185] follows the most straightforward way to integrate multiple rankings: for each gene, it calculates a simple rank average over seven independent rankings. ToppGene [42, 43] and Endeavour [1, 265, 264] realize more sophisticated approaches to obtain the overall ranking. They convert data source-specific scores into p-values and apply meta-analysis-based techniques to compute the overall p-value for each gene.

Score aggregation approaches described above have at least two drawbacks. First, these tools favor genes top-ranked in a maximal number of evidence sources. Meanwhile, they may not consider various reliability and potential dependency of evidence sources. Second, tools from this category do not take into account the fact that the impact of independent rankings into the total score may not be additive.

These deficiencies have been partly overcome with the development of machine learning methods. Similar to ToppGene and Endeavour, machine learning-based methods represent genes as n-dimensional feature vectors, use seed genes as positive training exemplars, genes other than seeds or candidates as negative exemplars, and then classify candidates. Machine learning methods such as multiple linear [44, 144, 284] and logistic [276, 287] regressions, kernel-based approaches [54, 200, 291], neural networks [85] and others [121] were successfully applied for gene prioritization. Recent works have demonstrated that machine learning-based methods tend to outperform other score aggregators [191, 276, 290], possibly owing to their ability to capture unknown or non-linear feature relationships and tuning model parameters.

2.2.3 Limitations

Despite the popularity of gene prioritization tools, almost all of them have two important drawbacks:

- All candidate genes are ranked separately, what suits the case of monogenic diseases, but for polygenic disorders may be disadvantageous.
- Possible disease heterogeneity is not taken into consideration. The disease may comprise several latent subclasses, phenotypically similar but caused by distinct molecular

alterations. Taking into account disease heterogeneity requires the analysis of patient-level experimental data.

The above problems are discussed in the next sections of this thesis.

2.3 Mendelian and Complex diseases

The impact of the individual genetic constitution on the development of the disease varies widely. Some diseases are caused by rare dysfunctional variants of a certain gene and are highly heritable. Therefore, although such diseases are rarely seen in the population, they frequently occur in some families. The distribution of affected individuals in these families follows Mendel's law therefore such monogenic diseases are called Mendelian. This group of diseases includes, for example, sickle-cell anemia (OMIM:#603903) or Duchenne muscular dystrophy (OMIM:#310200). In most cases, the disease is caused by a single mutation leading to the loss of gene function [241] or its modification [172].

To date, more than six thousand human diseases caused by the dysregulation of a single gene are known [7]. Because unrelated individuals may carry different variants affecting the same gene, precise diagnostics usually require whole exome or even whole genome sequencing of the affected individuals and their parents. Disease variants are searched among rare variants with a strong effect on gene function, e.g. complete loss of function, or modifying crucial gene regions.

Unlike Mendelian diseases, complex diseases e.g. cancers, Alzheimer's disease, or asthma are multifactorial. These diseases cannot be explained by a single mutation with a strong effect but thought to be the result of interactions between multiple genetic and environmental factors. Genome-wide association studies aimed to discover single nucleotide polymorphisms (SNPs), which allelic states significantly correlate with disease status. Compared to variants causing Mendelian diseases, SNPs are much more frequent in the population. NHGRI-EBI GWAS Catalog [173] provides a curated and regularly updated lists of published GWAS, and contains over sixty thousand of SNP-trait associations. The effects of such risk variants identified in GWAS are much weaker than the effects of Mendelian variants. Moreover, linking an associated variant with its effect on a certain gene may be not straightforward. Unlike Mendelian disease variants, only a small fraction of GWAS variants has an obvious effect on the protein, e.g. missense substitution, premature stop gain, or frameshift. Most of the significant GWAS hits are located in intronic or intergenic regions [106]. Mapping them to the closest gene may not always be correct, and considering regulatory annotations for

	Mendelian diseases	Complex diseases
causes	abnormal function of certain gene due to a pathogenic mutation	complex interactions between genetic, epigenetic and environmental factors
penetrance	high	low
heritability	high	low
number of genetic associations	one or several variants affecting the same gene	many in multiple loci
types of associated variants	non-silent mutations with a strong effect on protein	all kinds of variants, including intronic and intergenic
distribution of associated variants in populations	extremely rare variants	frequent, infrequent, and rare variants

Table 2.1 Comparison of Mendelian and complex diseases.

mapping of silent GWAS variants to genes appears to be a better strategy [73]. Joehanes et al. 2017 [126] have shown that about half of GWAS variants lay in expression quantitative trait loci (eQTL) – genome regions with markers correlated with expression levels of one or several genes. Recently, it has been shown by Li et al., 2018 [158] and independently by our group [300], that comorbidity of some complex diseases is more likely to be explained by the overlap of eQTL-controlled genes, rather than a direct genetic overlap.

Summarizing the above, complex and Mendelian diseases are intrinsically different (2.1) and therefore the investigation of their causes and underlying mechanisms requires different approaches. The primary cause of a Mendelian disorder is relatively easy to find, and after that, the mechanism of disease development becomes more or less clear. Having sequencing data of an individual, the development of Mendelian disease can be predicted with a high degree of confidence. In contrast, the primary causes of complex diseases remain hypothetical to date. The attempts to predict complex traits including disease predisposition given genomic data had limited success [17, 123, 128], because the individual effect of each associated variant is very small.

2.4 Disease Heterogeneity

Molecular heterogeneity of the disease implies that similar clinical manifestations may be caused by different molecular lesions. Sometimes genetically distinct disease subtypes may be hardly clinically distinguishable [136]. In other cases, molecular heterogeneity may result in wide phenotypic heterogeneity. A group of mechanistically distinct diseases,

but characterized by similar symptoms might be historically united under a single name [145, 197].

Molecular heterogeneity was shown for many human diseases, complex and monogenic [181, 211]. The most famous examples of such diseases are cancers, demonstrating a tremendous molecular and phenotype heterogeneity [107]. For a long time, cancer was classified only by the tissue of origin. The choice of anti-cancer treatment depended only on the cancer type and general health condition of the patient. With the accumulation of data on incredibly high heterogeneity within each type of cancer, the paradigm shifted towards tailoring the treatment to target weaknesses of each specific tumor [37, 207]. At the same time, the investigation of cancer subtypes leads to a more detailed understanding of some of the mechanisms of its development [107]. During the last years, the search for new subtypes and subtype-specific biomarkers with prognostic or predictive power became one of the hottest areas of modern biology.

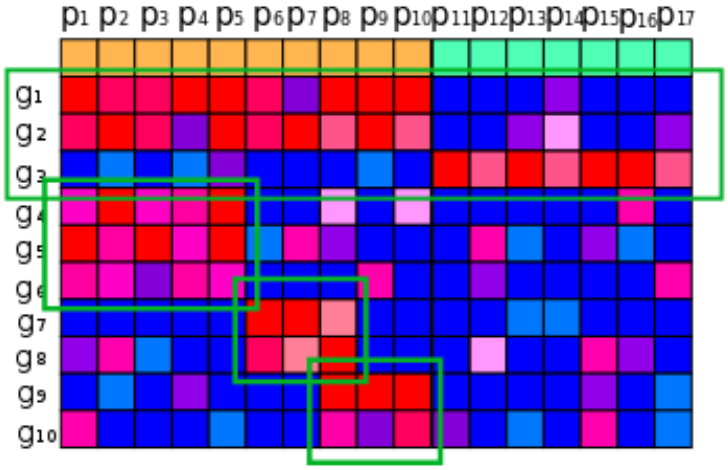

Fig. 2.3 A toy example of the expression matrix, genes are shown in rows and patients are in columns. The top row depicts class labels, e.g. disease (orange) and controls (green). Some genes are altered only in a specific subgroup of disease samples. These genes can be missed in a case-control study, if the corresponding group is not big enough or expression fold change is small.

Evidence of disease heterogeneity is observed at different molecular levels, e.g. genomic [11], transcriptomic [216], epigenomic [179], and metabolic [212]. The presence of heterogeneity complicates the discovery of biomarkers and further development of predictive models. In the case of heterogeneity, a biomarker may be relevant only for a subgroup of patients representing a certain subtype (Figure 2.3). For example, if a gene or a group of genes is differentially expressed only by a small subgroup of samples, it may be missed if

a standard case-control design is chosen. Moreover, the discovery of rare disease subtypes requires larger sample sizes [278]. Besides that, if a specific disease subtype is characterized by differential expression of a small group of genes, it might not be visible after dmentionaly reduction, e.g. at PCA plot.

Breast cancer was chosen for this thesis, because

- it is well known to be molecularly heterogeneous and has many characterized subtypes;
- it is a frequent cancer type diagnosed in more than 2 million women worldwide annually [28];
- several large breast cancer expression datasets with detailed annotations are publicly available.

In 2001 Perou et al. [216], proposed the first molecular classification of breast tumors based on their expression profiles distinguishing four molecular subtypes:

- **Luminal-like** tumors highly express ER and keratins 8 and 18. The latter are known to be makres of luminal epithelial cells, what points to their cell type of origin. Luminal-like tumors are split into Luminal A and Luminal B subtypes, which differ in the level of tissue differentiation and survival prognosis [254].
- **Basal-like subtype** got its name owing to high express keratines 5,6, and 17, intrinsic of the basal layer of epithelium. Most of basal-like tumors actively proliferate and demonstrate low expression of ER, PR, and HER2, i.e. are triple-negative.
- **Her2 subtype** is characterized by elevated expression and frequent amplification of Her2 (also known as ERBB2), which encodes human epidermal growth factor receptor 2. Signaling through this receptor with tyrosine kinase domain suppresses apoptosis and induces cell proliferation. Patients with Her2/ERBB2-positive tumors have the bad survival prognosis, but better respond to trastuzumab, an antibody binding Her2 [215].
- **Normal breast-like group** includes tumors with expression profiles more similar to normal breast tissues samples, rather than tumors and therefore not falling to any of the above subtypes.

Later in 2007, Herschkowitz et al. [102], defined yet another subtype called **Claudin-low** and characterized by suppressed expression of claudines, E-cadherine, and other genes, responsible for cell adhesion. Another feature of the claudin-low subtype is high lymphocyte infiltration [226].

The above molecular, however, takes into account far from all expressed genes. It does not reflect the whole spectrum breast tumor heterogeneity and seems to be insufficient for precise prognosis and prediction. The discovery of unknown molecular subtypes remains the direction of further research.

2.5 Gene expression profiling

Gene expression is a multistage process of realization of information encoded by this gene in the form of its product, transcript, or protein. Expression of genes is necessary for their functioning, therefore it is strictly regulated by the cell. The increase (up-regulation) and decrease (down-regulation) of gene expression mean an increase and decrease in the amount of gene product respectively.

Various changes in gene expressions point to alterations of gene functions. Although the final result of protein-coding gene expression is the creation of protein molecules, the expression profile most commonly implies a transcriptome profile. In humans and other eukaryotes, a gene can encode several variants of a transcript (isoforms) owing to the excision of certain transcript subsequences (splicing). To identify genes which functions are altered under a specific condition, e.g. in disease compared to control, researchers search for alterations of

- expression levels of genes, transcripts, isoforms [149]
- splicing sites [176]
- localizations [113]

Two types of expression patterns related to the level of expression should be distinguished: differentially expressed and differentially co-expressed groups of genes (Figure 2.4). These two patterns represent perturbations with distinct biological significance. The differential expression reflects the induction or inhibition of a certain pathway or its downstream. Differential co-expression points at gain or loss of co-regulation and highlights regulatory network rewiring.

To date, several methods for investigation of gene expression exist [198]. The abundances of selected transcripts can be measured by various quantitative PCR-based protocols or Northern blotting. However, if the aim is to find new candidates, rather than test a small number of previously known, the researchers use high-throughput methods, such as microarrays or RNA sequencing (RNA-seq). Both methods allow simultaneous profiling of thousands of

transcripts in a single experiment. High-throughput gene expression profiling provides a detailed snapshot of gene activities and therefore is widely used in the studies of disease heterogeneity [216].

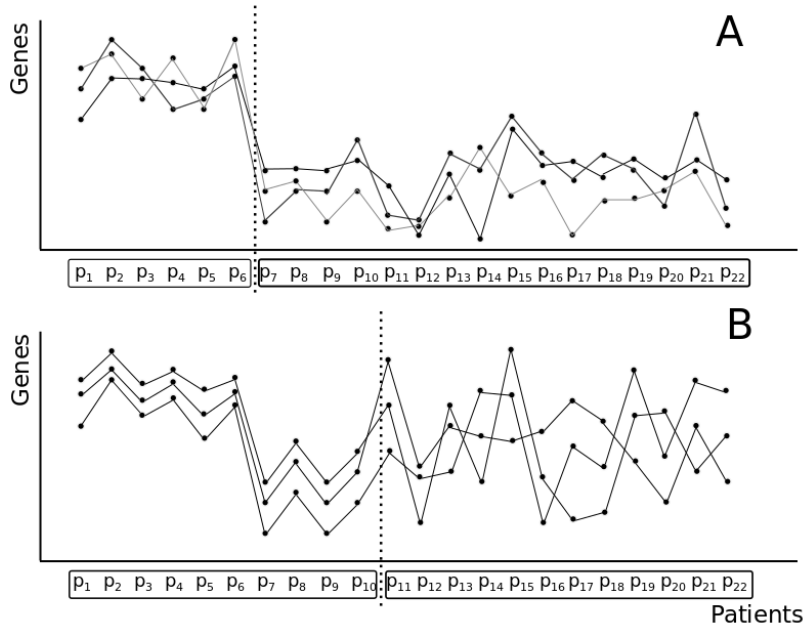

Fig. 2.4 A toy example of two gene expression patterns. A. Differential expression. Three genes are up-regulated in patients 1-6 compared to the other ones. B. Differential co-expression. Expression levels of three genes are correlated in patients 1-10 but not in patients 11-22.

Microarrays

Microarray is a solid chip with attached oligonucleotide probes complementary to sequences of known transcripts. Modern microarrays developed for human transcriptome contain up to several hundreds of thousands of probes complementary to the regions of most human transcripts. The process of expression profiling with one-color microarray includes the following steps [229]:

1. total RNA is extracted from the samples and purified. If necessary, the desired sequences may be enriched (e.g. poly-A pooling of mRNA) and undesired depleted (e.g. depletion of ribosomal RNA)

2. reverse transcriptase produces cDNA which is more stable on the matrix of RNA

3. cDNA is amplified via PCR

4. cDNA fragments are labeled with a fluorescent dye
5. the mixture is added to the microarray. cDNA fragments hybridize with probes on the chip
6. the abundance of cDNA complementary to each of the probes is defined according to the intensity of fluorescence, recorded by the machine

The procedure is slightly different for two-color microarrays, which simultaneously hybridize cDNA from two samples labeled by different markers. It shows relative abundances of expressions in case and control samples [243].

Sequencing of RNA

With the advent of high-throughput sequencing methods, the popularity of expression microarrays is declining. The main advantage of RNA-seq over microarray technology is that the former allows not only quantification of known transcripts, but also detection and sequencing of the unknown ones. Besides that, RNA-seq more correctly measures weakly expressed transcripts than microarrays [295]. The beginning of the sample preparation protocol for RNA-seq is the same as for microarrays: RNA is extracted, purified, converted into cDNA, and amplified. This is followed by a library preparation step which depends on the sequencing platform chosen by researchers. To date, several sequencing platforms present on the market, of which Illumina is the most commonly used [177].

Sequencing results in a large number of short *reads* – fragments of transcript subsequences determined, (i.e. read) by the machine. Every base in a read is accompanied by a quality score reflecting the probability of a wrong base call. To obtain gene expression reads are mapped to the reference transcriptome sequences and quantified [56]. To make gene/transcript expression abundances comparable within and between samples, they are subject to normalization procedures [79, 230, 233].

2.6 Identification of differentially expressed genes and gene sets

Correlation of gene expression level with any disease features may be a sign of the possible involvement of this gene in the disease mechanism. This consideration stimulated the development of approaches for the identification of genes differentially expressed in disease and

control groups [167, 180, 231]. However, the reproducibility of differentially expressed gene lists obtained in independent studies was rather low [260, 292]. To improve reproducibility and consistency of differential expression analysis results, genes demonstrating a similar pattern of dysregulation may be grouped together. Working with gene sets instead of single genes is beneficial because it reduces the dimensionality of the data and makes aggregated expression less prone to noise [118]. Moreover, the resulting smaller groups of coordinately altered genes are easier to interpret than the whole list of dysregulated genes. Such groups of functionally related disease-associated genes also called *gene modules* [188, 236].

Similar to individual genes participating in the development of the disease, the genes of a disease-associated module must demonstrate a specific pattern of dysregulation in the disease samples compared to controls. Consequently, some clustering methods group genes into modules based on the dependency of their expression profiles without considering any prior knowledge about these genes [148, 236]. Alternatively, genes can be grouped using predefined gene sets representing known pathways or functional groups of genes either before [14] or after [257] testing for dysregulation. The reference gene sets can be obtained from a specialized gene set [162, 288] or pathway [133, 134] databases.

However, expert-curated pathways and gene sets are limited by the current knowledge and cover only a small part of the whole interactome [184]. On the other hand, computationally predicted pathways and gene interactions provide a more complete view of interactome but they might be inaccurate. Moreover, due to cell type specificity and the dynamic nature of cellular circuits, even valid pathways or gene sets may be irrelevant to the object of research and absent in a given dataset. Therefore, instead of using *a priori* defined gene sets, some methods predict novel gene modules through integrative analysis of gene expression profiles and protein-protein interactions (PPI). These methods require dysregulated genes to cluster in gene networks derived from PPI networks, forming so-called active subnetworks. This network constraint is valid because functionally related genes are likely to be co-regulated, to interact and act together. jActiveModules [118], BioNet [16] and other methods [50, 65] map differentially expressed genes to a PPI network and search for minimal connected components containing as many dysregulated genes as possible. These methods accept differentially expressed genes scored by effect size or p-value or define them before the network search [52, 269]. Some more sophisticated methods such as OptDis [66] or CoSINE [171], take normalized expression profiles as input and determine dysregulated genes while discovering active subnetworks.

All the above methods search for differentially expressed gene modules in a supervised manner, when class labels, e.g. disease and control or disease subtypes, are known. However,

in many real-world scenarios, class labels may be unavailable. Moreover, even when class labels are provided, the compared sample groups may demonstrate high internal heterogeneity and consist of several unknown molecular subtypes [25, 40, 181]. Identification of genes, under- or overexpressed only in a certain unknown group of samples, is especially challenging when such a group is small. This problem of the detection of between-sample heterogeneity may be addressed by clustering and biclustering methods. Biclustering methods [206, 223, 285] are searching for subsets of genes demonstrating similar expression patterns in a subset of samples, given a matrix of genes profiled in these samples. The third chapter of this thesis is entirely devoted to biclustering methods and explains their advantage over conventional clustering.

2.7 Biological networks

A network is a natural representation of pairwise entity relationships, widely used to describe similarities or interactions between biological objects. The methods of network biology advanced the understanding of biological phenomena including complex diseases [13, 117, 296]. Disease genes were shown to possess special network properties. For example, it has been shown that disease-associated proteins tend to cluster on the PPI networks [92, 119, 184]. The researchers pay special attention to two types of nodes, which are likely to play important roles in biological networks [92, 289]:

- *hubs* – nodes with the highest degree;
- *bottlenecks* – nodes with the highest betweenness centrality, i.e. those through which the passes the maximal number of shortest paths.

"Importance" here means that these nodes are crucial for many biological processes and too important to be dysfunctional. Therefore, disease genes are less likely to be found among hubs or bottleneck genes, although some examples of essential and disease-specific genes are known [13]. At the same time, the most influential disease genes tend to be central in the disease-specific networks [155, 205]. Identification of these and other relationships between functional properties of genes and network topology motivated the invention of network-based algorithms for the discovery of genome-wide associations [279], gene prioritization [42, 60, 130, 141, 205], and determination of drivers in cancer [251, 273].

The researchers created and analyzed numerous kinds of biological networks, from homogeneous, including only one type of objects and interactions, to composite heterogeneous networks [105, 124, 250], demonstrating relationships between various biological entities.

Types of biological networks widely used in molecular biology and medicine and some of their properties are reviewed in the next section and were previously published in [299].

2.7.1 Types of biological networks

Physical protein-protein interactions

Physical protein-protein interactions (PPI) point to a potential functional interaction between these proteins and subsequently, to the association between corresponding genes. Physical PPI can be experimentally identified using high-throughput methods, such as yeast two-hybrid assay, affinity purification with mass spectrometry or confirmed in single experiments, e.g. X-ray crystallography. Primary PPI databases obtain data from curation of published literature, e.g. DIP [239], HPRD [225], BioGRID [39], InnateDB [29] or MatrixDB [150] or from single large-scale experiments [115, 116]. Other PPI databases, such as IntAct [203], MINT [163], MENTHA [33], HitPredict [166], integrate protein interaction data from multiple primary databases and assign interaction reliability scores according to the level of supporting evidence. In order to facilitate access to a large number of redundant PPI databases, a standardized query interface PSIQUIC was created.

In addition to direct physical contacts, proteins can also interact indirectly, collectively performing their function. For example, since a protein complex functions as a whole, all its members, including those non-interacting directly, are strongly functionally related. CORUM [235] and Complex Portal [183] provide curated human and animal protein complexes, their subunit composition, structure, and functions.

Pathways and regulation

Proteins participating in consequent steps of a biological pathway are also considered to be functionally related. In a broad sense, a biological pathway is a chain of molecular events, such as chemical reactions, conformational changes, binding or dissociation, etc., which leads to certain changes in the cell. Pathguide [10] is a comprehensive catalog comprising of 702 resources related to pathways and molecular interactions in human and other organisms. Pathways are classified according to prevailing interaction type as metabolic, signaling, and regulatory. Metabolic pathways, representing chains of chemical reactions catalyzed by enzymes, can be found in MetaCyc [35], which is a part of BioCyc, including pathway-related information for more than 13000 species. Signalling databases, such as OmniPath [266], Signor [214], SignaLink [82], PhosphoSite [109], contain literature-curated information on cellular signal transduction via post-translational modifications, relocation,

binding or conformational changes. Genetic regulation databases contain manually curated and computationally inferred relationships between genes and transcriptional factors (TFs), e.g. JASPAR [138] and TRANSFAC [282], or miRNA, e.g. miRTarBase [49]. Large pathway databases, such as KEGG [135], Reactome [80] and ConsensusPathDB [132] are not specialized on a particular type of pathway or process and provide biological interaction of multiple types for human and other organisms, while the other resources have a certain focus, e.g. innate immunity [29] or a specific disease [131, 189].

Predicted interactions

Since biological pathways are mediated by gene products, proteins or RNAs, pathway data is the invaluable source of functional relationships between genes. However, known pathways cover only a small part of all the existing interactions and not all human genes are fully functionally annotated. Unknown gene functions and interactions can be computationally predicted on the basis of gene co-expression [270], sequence similarity [55] or interactions [156, 275] with well-annotated genes. Genes or proteins with expression levels correlated across different conditions are likely to be co-regulated and may share functions [270]. Sequence similarity and domain composition can also give a clue about the function of an unannotated protein and help to identify its interaction partners. Recent paralogs may have the same function [160], but later their functions tend to diverge. Orthologs are more functionally conservative [262] and therefore functional annotations of genes from related species and PPI [277] may be transferred on their human orthologs.

Functional similarity

The amount of knowledge regarding gene and protein roles in the cell is diverse, enormous and continuously growing. The unification and formalization of this knowledge are crucial to ensure its computational processing and analysis. Gene Ontology (GO) consortium [9] created in 1999, develops and maintains a controlled vocabulary of concepts describing gene functions, localizations and participation in biological processes. GO consortium provides regularly updating [57] whole-genome annotations, either supported by experimental evidence or computationally predicted, for multiple species, from human to bacteria, which allows within and between-species comparisons of gene functions. GO term enrichment analysis became a community standard for functional annotation of gene sets and interpretation of the experiment results. Since genes sharing GO terms are considered to be functionally related, many gene prioritization tools utilize GO as an additional source of evidence.

Text mining

Yet another way of establishing putative associations between genes, diseases and other biological entities is text mining of biomedical literature. Many gene prioritization tools utilize the results of co-occurrence based text mining, assuming that frequent colocalization of two entities in biomedical texts points to their possible interaction. More sophisticated pattern-based text-mining methods use advanced weighting schemes to assign qualities to predicted associations [46, 165]. Other text-mining systems, e.g. ANDsystem [122], apply natural language processing (NLP) algorithms allowing to differentiate between various kinds of biological entities and associations between them. The major drawbacks of the networks built on the results of text mining are the high rate of false positives and the lack of accuracy in the detection of associations and determination of their types. Despite that, text mining remains the only way to absorb the whole volume of relevant scientific literature, impossible to handle manually.

2.7.2 Properties of biological networks

Biological networks are not random and possess several characteristic properties [13], of which two are in the focus of this thesis: scale-free property and modularity.

The first was proposed in 1999 by Barabasi and Albert [12] who demonstrated that degrees of nodes in natural networks are distributed according to the power law, i.e. the probability of a node with degree n is proportional to $n^{-\gamma}$. They have also shown that values of γ lay between 2 and 3 for most biological networks. The authors proposed that a scale-free network is formed in the result of the process with a preferential attachment when the probability of a new node to attach to one of the network nodes is proportional to the node degree [4].

Currently, it is debated at which extent the scale-free property is fulfilled in real-world networks [30]. Broido and Clauset analyzed almost a thousand various natural networks, including biological, sociological, and information networks and demonstrated that in many cases node degree distribution is better fitted by log-normal, than power law.

Another feature inherent of biological networks, and other aspects of living organisms organization, was formulated by Hartwell et al., 1999 [98]. Structural modularity has long been known, but Hartwell et al. emphasized functional modularity. They proposed to study *functional modules* – groups of biological molecules united by a common function and separated from the other such groups. This idea stimulated the research in the field of network biology and led to the emergence of many works aimed at the identification and

interpretation of modules in biological networks [3, 48, 91, 188, 236]. Speaking of modules in biological networks, three interrelated concepts should be distinguished (Figure 2.5):

- *Topological communities* represent a set of nodes [91] or edges [3] more connected with each other, than with nodes (edges) outside it. Densely connected communities found in PPI networks represent protein complexes [255]. In co-expression networks communities correspond to clusters of co-expressed genes, representing downstream of a certain pathway [236]. Many methods for community detection have been developed [84], for example, Markov Cluster algorithm [72] or Louvain method [22]. densely connected regions of the networks. Community is defined as

- *Functional modules* are topological communities of interacting biomolecules united by a common function. Functional modules are not necessarily so densely connected as protein complexes or clusters of co-expressed genes. Depending on the network, functional modules may represent metabolic pathways, signaling cascades, epigenetic regulation, etc.

- *Disease modules* may be understood as a subset of functional modules. Disease modules are fully composed of genes involved in the process of disease development or enriched by such genes [92]. In 2016, Choobdar et al. organized a community competition on the identification of disease modules at dreamchallenges.org. They attracted more than 400 teams and published the benchmark of 75 methods on PPI, gene co-expression, signaling, homology, and cancer gene essentiality networks [48].

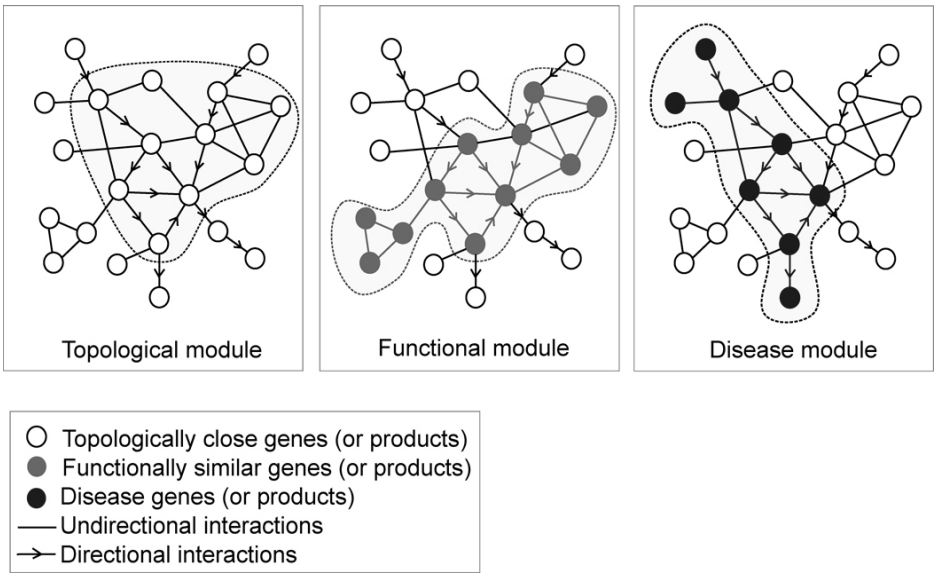

Fig. 2.5 Three types of network modules. The figure is reprinted from [13].

2.8 Network analysis for medical bioinformatics

Besides homogeneous networks, including only one type of objects and interactions, the researchers created and analyzed numerous kinds of composite heterogeneous networks [105, 124, 250], modeling various relationships between multiple kinds of biological entities. An example of such study is our recent work aimed at unraveling molecular-genetic reasons of frequent co-occurence of asthma and hypertension [300].

In many patient cohorts, asthma and hypertension coincide more frequently than would be expected by chance [51, 69, 83, 101, 127]. Such correlation of two or several diagnoses is called comorbidity. It is shown for many human disorders, complex and Mendelian [21, 103, 111, 182]. Comorbidity and may point to causal relationships between two diseases, e.g. shared susceptibility loci. However, this does not seem to be the case of asthma and hypertension [202]. At the same time, isolated asthma and hypertension have inheritable components, which means the presence of some shared molecular-genetic mechanisms. This observation motivated us to search for genetic overlap considering multiple kinds of molecular evidence in addition to genetic associations. From public databases we extracted genes which:

- carried at least one non-silent variant associated with any of diseases in GWAS;
- had a variant or variants causing familial forms of hypertension or asthma, or a mendelian disease characterized by hypertension or asthma among other symptoms;
- were regulated by eQTL variants matching with GWAS hits for asthma or hypertension;
- demonstrated differential expression in tissues of patients with asthma or hypertension, compared to healthy controls;
- were targeted by drugs used to treat asthma or hypertension;
- were targeted by drugs worsening or asthma or hypertension;
- frequently co-occurred in texts with asthma or hypertension;

In total, 330 genes were associated with both diseases through the associations of various kinds. These shared genes were further projected on the PPI network obtained via Cytoscape [244] version 3.6.1 stringApp [71] version 1.3.0. Genes not connected with any other shared genes were excluded from consideration. Applying the EAGLE algorithm [248] on the resulting network of 257 nodes revealed six modules which were further characterized

according to overrepresented pathways, GO terms, and tissue-specific gene sets. The scheme of the study workflow is shown on Figure 2.6.

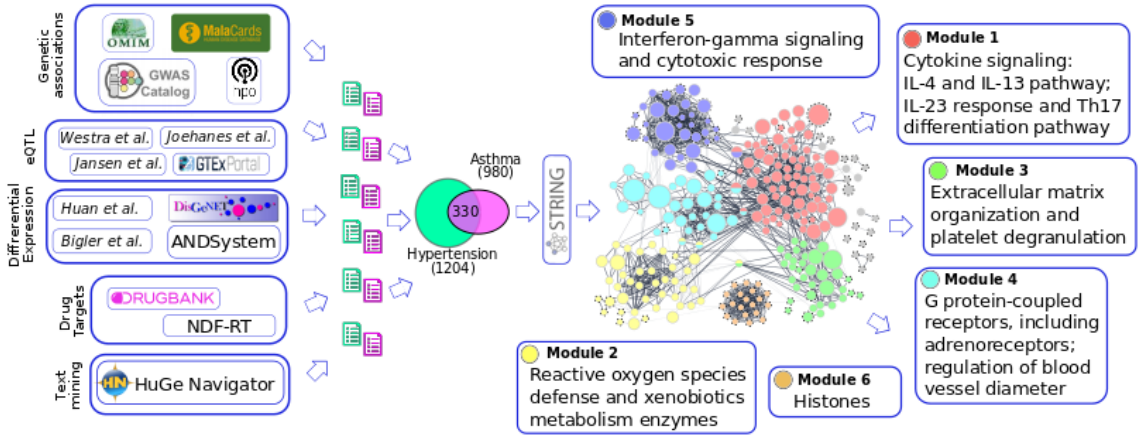

Fig. 2.6 Identification and characterization of gene modules associated with asthma and hypertension. Network nodes represent genes and are colored according to membership in a module. Nodes not assigned to clusters are shown in grey. Size of each node is proportional to the number of evidence sources supporting the association of corresponding gene with asthma or hypertension. The figure is adapted from [300].

The description of each of the modules and the discussion of their possible role in asthma and hypertension is given in [300]. Interestingly, testing of the whole set of shared genes would give a smoothed picture of overrepresented GO, pathway, and tissue-specific labels, similar to the largest module. This result demonstrates that considering gene interactions enables a more detailed view of biological processes implicated in asthma and hypertension and potentially responsible for their comorbidity.

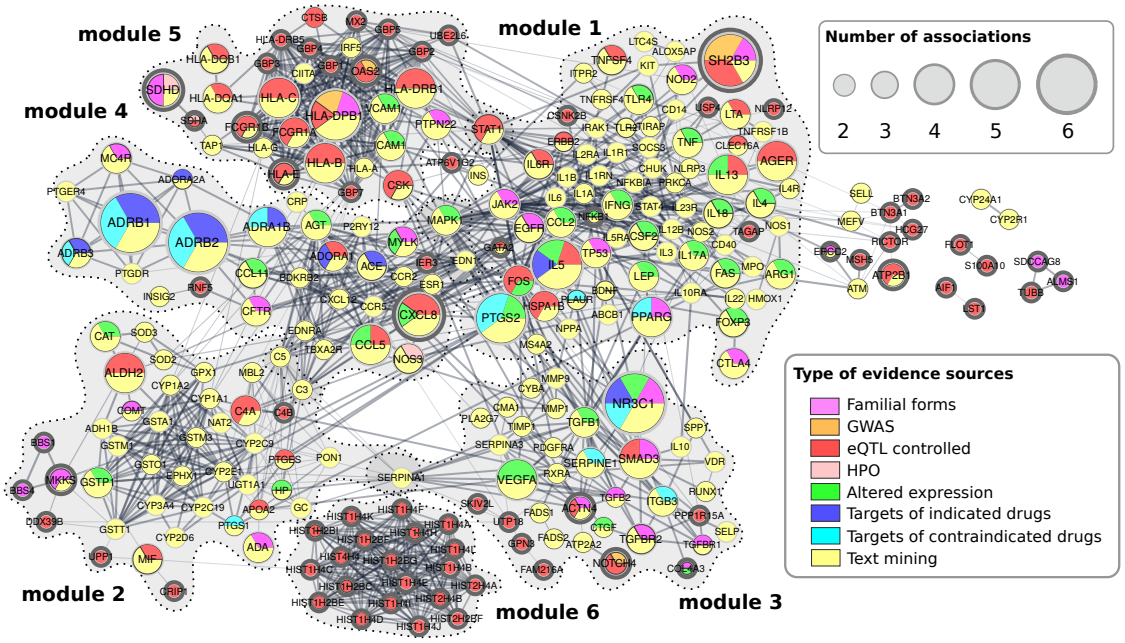

Fig. 2.7 Evidence sources supporting gene associations with asthma and hypertension. In this figure, a node style similar to Figure 1B in [247] was used. Here, nodes represent genes associated with both asthma and hypertension, edges correspond to gene interactions. Genes are colored according to evidence sources (see figure legend) from which associations came from. The size of each node is proportional to the number of evidence sources supporting its association with asthma and hypertension. The figure is reprinted from [300].

Module 4 contained several genes targeted by drugs indicated or contraindicated in asthma and hypertension. Since several drugs had opposite effects on asthma and hypertension, we hypothesized that drug side effects may also contribute to the developemt of comorbidity. For example, drugs used against one disease could make patients more prone to another disease and thus increasing risks of comorbidity. To evaluate this hypothesis, drugs that influence asthma or hypertension were classified into four groups:

- drugs used to treat asthma or relieve its symptoms
- drugs decreasing blood pressure and used against hypertension
- drugs contraindicated for patients with asthma or worsening or inducing its symptoms
- drugs elevating blood pressure and/or contraindicated for patients with hypertension

Eight non-selective beta-blockers used to treat hypertension (timolol, nadolol, sotalol, pindolol, carvedilol, labetalol, propranolol) were not recommended for asthma patients due to the risk of asthma exacerbations [190]. At the same time, seven anti-asthmatic drugs

classified as beta-agonists or corticosteroids were in the list of drugs that may elevate blood pressure: triamcinolone, prednisolone, methylprednisolone, dexamethasone, hydrocortisone, and epinephrine, ephedra, ephedrine.

Since drugs from the same class affected the same target genes, to identify all targets that may potentially mediate drug effects on asthma and hypertension, target overrepresentation analysis has been carried out for each of four drug groups. It revealed 96 genes significantly overrepresented among targets of at least one of four drug groups (summarized in Figure 2.8), but only 16 of them were in the asthma-hypertension network. As expected, *ADRB1* and *ADRB2* were targeted by drugs from all four groups, since activation and inhibition of beta-adrenoreceptors had opposite effects on asthma and hypertension. *NR3C1* which encodes glucocorticoid receptor, was activated by drugs indicated in asthma but potentially harmful for hypertension. Yet another target of corticosteroid drugs, *ANXA1* mediates the anti-inflammatory effect via inhibition of phospholipase A2 [208]. *PTGS1* was overrepresented among targets of drugs contraindicated in asthma, while inhibition of its paralog *PTGS2* potentially promoted both diseases.

Taken together, our findings suggest that genes targeted by prescribed drugs may contribute to pathophysiologic mechanisms of comorbidities. Moreover, this particular case of asthma and hypertension leads to the conclusion that drug side effects may be used to connect genes and diseases and advance the understanding of disease mechanisms.

To facilitate further analysis of the data on associations relevant for asthma and hypertension comorbidity, a Neo4j database called GenCoNet has been created [250]. GenCoNet describes relationships between four types of biological entities: genes, diseases, drugs, and gene variants (Fig. 2.9).

GenCoNet allows the user to answer research questions using Cypher query language, for example:

- find drugs indicated for one disease, but contraindicated for the other;
- find all genes, targeted by drugs used to treat the disease and differentially expressed in this disease;
- find all genes, which are controlled by eQTL variants associated with a disease;

In contrast with other databases of gene and disease relationships such as DisGeNET [15, 218, 219], Open Targets [142], or TargetMine [45], GenCoNet contains more reliable manually curated data on asthma and hypertension. One important limitation of this work is the absence of a gold standard for direct validation of the resulting gene sets. Experimental

validation of gene roles in comorbidity is strongly desired but remains beyond the scope for this thesis.

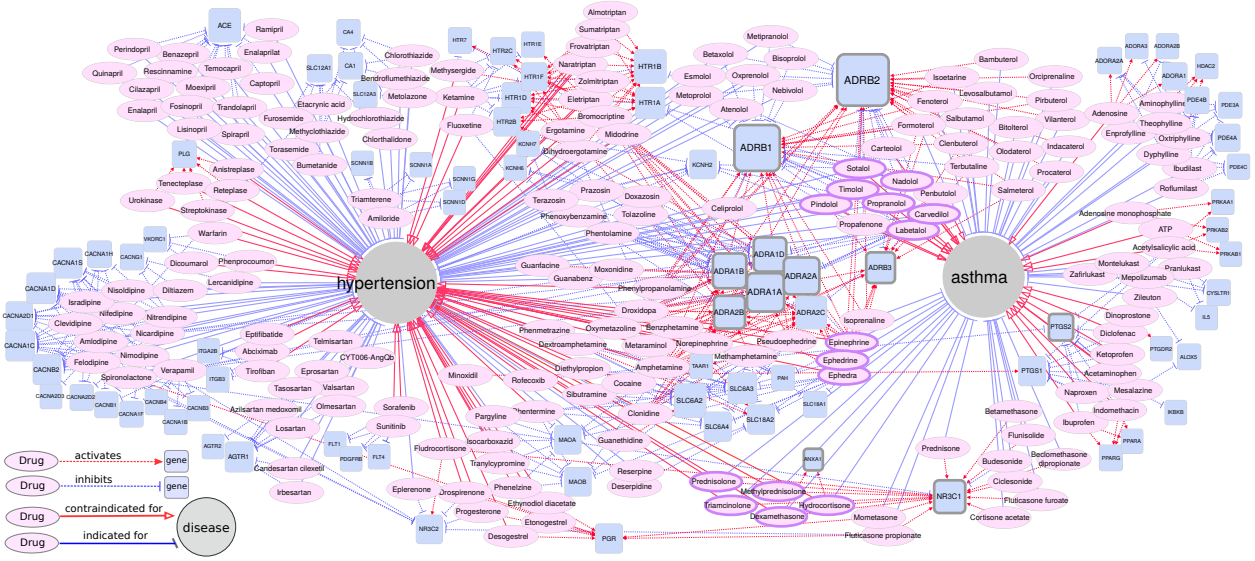

Fig. 2.8 Relationships between genes and drugs indicated and contraindicated in asthma and hypertension. All target genes significantly overrepresented in one of four drug groups are shown. Drugs influencing both diseases and target genes overrepresented in more than one group are shown with bold frames. The figure is reprinted from [300].

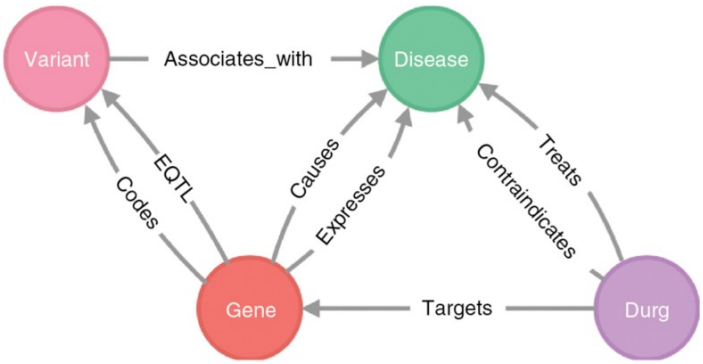

Fig. 2.9 The scheme of the GenCoNet database reprinted from [250].

Chapter 3

Overview of Biclustering Methods

3.1 Introduction to Biclustering

The term *biclustering* introduced by B. Mirkin in 1996 [186] designates the methods performing simultaneous clustering of rows and columns of a 2-dimensional data matrix. The result of biclustering is a set of submatrices demonstrating a specific pattern and called *biclusters*. Fig. 3.1 illustrates the general concept of biclustering on the example of an artificial real-valued matrix of 50 rows and 25 columns with five implanted biclusters.

In order to understand the concept of biclustering, one may compare it with conventional clustering. Let us consider a 2-D matrix, which columns represent the objects, (e.g. expression profiles of samples) and rows correspond to the features of these objects (e.g. genes). Conventional clustering methods group the objects in a way such that the objects from one group are more similar to each other than to the objects from the other groups. To solve this problem, clustering methods operate on similarities computed in the space of all features.

In contrast with conventional clustering, biclustering methods search for multiple independent groupings of objects, such that each grouping is supported by a local pattern. Locality here means that these patterns manifest only in certain subspaces of features, e.g. subsets of genes. These relevant subspaces may overlap and vary in size. The ideas behind biclustering agree with the understanding of phenotype heterogeneity and complexity discussed in chapter 2. Complexity implies that multiple genes or even functional groups of genes are involved in the development of phenotype of interest, e.g. a disease. Heterogeneity implies that the desired pattern may present only in a small subgroup of genes and samples. Moreover, there may present multiple such subgroups and they can overlap in genes and samples.

Biclustering is particularly useful when the data contain local patterns independent from each other and relatively small compared to the total number of features. Fig. 3.2 A shows a

data matrix with 5 biclusters overlapping in rows or columns. Rows of the bottom half of this matrix do not belong to any bicluster and contain random values drawn from the standard normal distribution. For example, hierarchical clustering applied independently on rows and columns of this matrix splits almost all the biclusters. Fig. 3.2B demonstrates a similar matrix, but its bottom rows are correlated. In this case, the clustering of columns will be driven by the largest pattern in this matrix. The appropriate biclustering approach would also detect the largest pattern, but along with smaller ones.

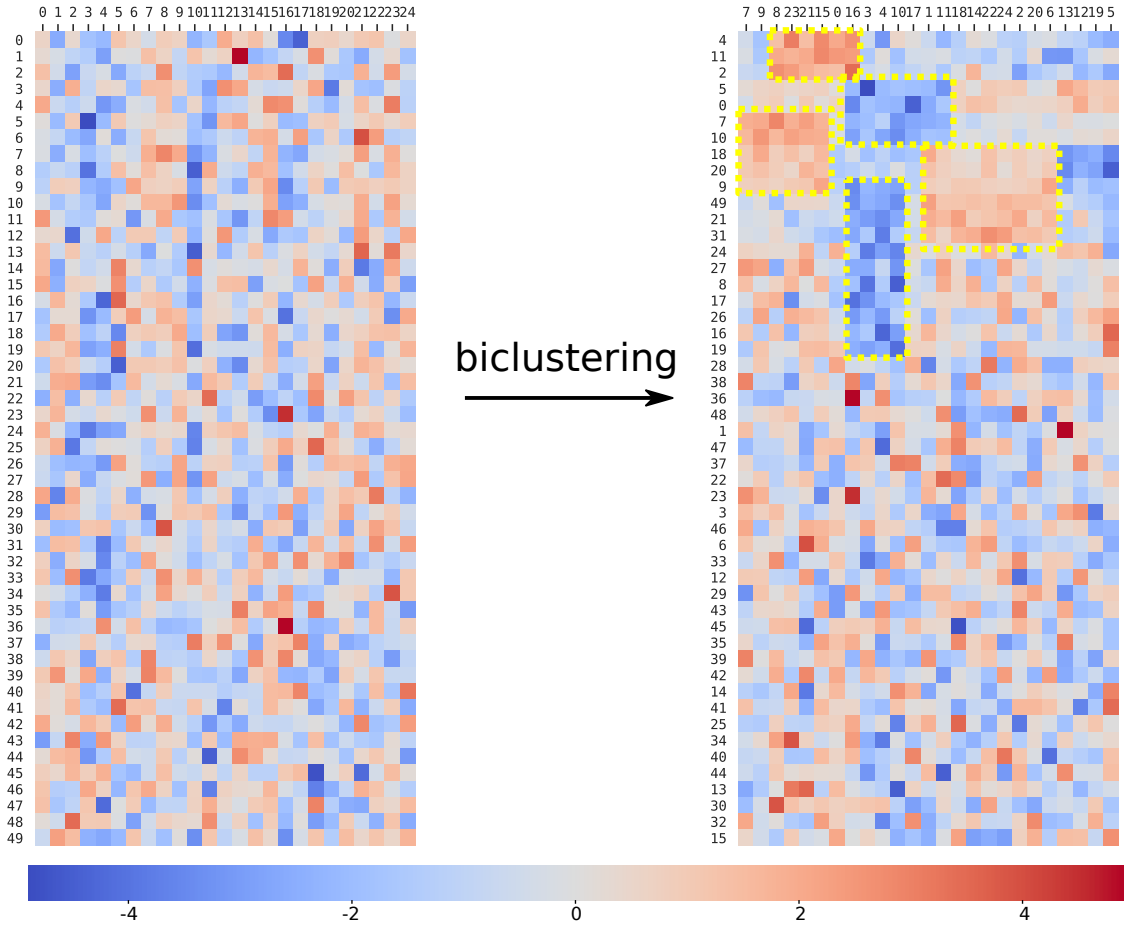

Fig. 3.1 The concept of biclustering. Dashed frames highlight biclusters which became visible in the matrix after the rearrangement of columns and rows.

The main challenge of biclustering is the size of search space which is much larger than for conventional clustering on columns and rows. Moreover, it is often the case that neither the number of biclusters (i.e. relevant groupings), nor their sizes are known, which makes biclustering problems extremely computationally complex and cannot be solved by a simple

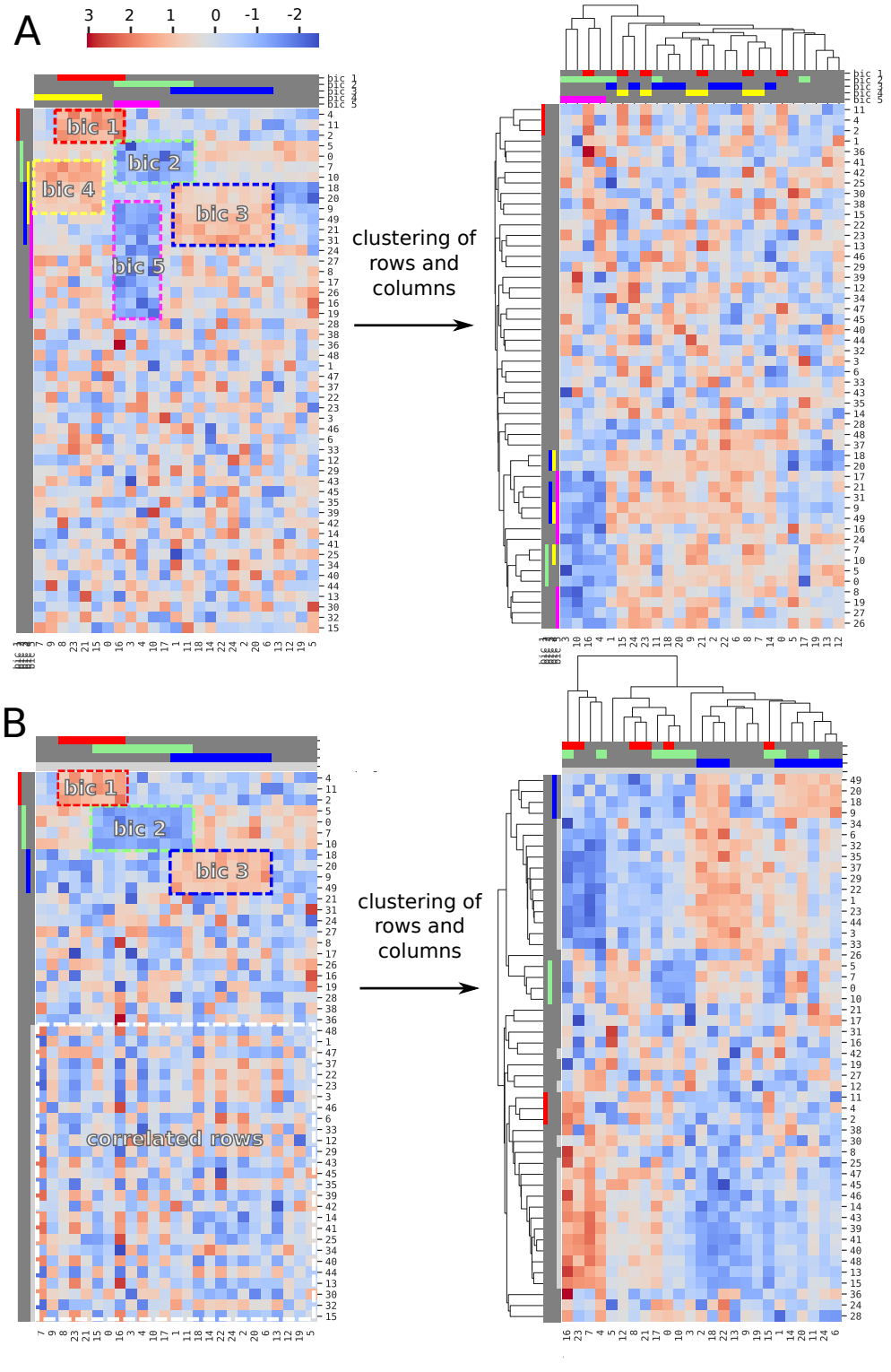

Fig. 3.2 The examples demonstrate the advantage of biclustering over conventional clustering. Both panels show matrices before and after hierarchical clustering with average linkage in a space of euclidean distances. Colored bars and dashed frames highlight membership in biclusters. A. The input matrix contains five implanted biclusters highlighted by dashed frames. Because these biclusters overlap in rows or columns, hierarchical clustering mixes them up. B. The input matrix contains three biclusters overlapping in columns. Approximately half of all rows outside biclusters are correlated (marked by grey bars and the frame). This group of rows is larger than any of the biclusters and therefore makes the major impact on the clustering of columns.

brute force in a reasonable time. Some biclustering problems were proven to be NP-complete [261, 174]. Briefly, the simplest problem definition of binary biclustering can be reformulated in terms of graph theory as the problem of finding a maximum edge biclique in a bipartite graph, which is NP-complete [210]. Other problem definitions consider non-binary data and/or searching for more complex patterns must have higher complexity. To reduce the complexity and find suboptimal solutions in a shorter time, many methods apply various heuristics or solve restricted versions of the problem.

Since the pioneering work by Hartigan published in 1972 [97], dozens of various biclustering problem definitions and approaches to solve them have been proposed [223]. Because the ideas behind biclustering agree with the understanding of phenotype heterogeneity and complexity, it has been widely used in bioinformatics for the analysis of data of various kinds: genomic, transcriptomic, epigenomic, etc. [285].

Same as clustering problems, the exact definitions of the problems solved by biclustering methods may greatly differ in details, in particular:

- in patterns characterizing the desired biclusters
- input data type: binary, discrete or real
- whether the number of biclusters is known
- whether the overlap of biclusters is allowed
- the exhaustiveness towards columns and rows
- whether any additional data about rows and columns are considered

Characterization of the existing methods according to the above characteristics is provided below in this chapter. A special emphasis is placed on the applicability of the methods for the purpose of gene expression data analysis.

3.2 Classification of biclustering methods

3.2.1 Input data type

The exact definitions of biclustering problems vary in assumptions about the input data. The majority of computational tools of biclustering are capable of handling any real-valued matrix, e.g. a matrix of gene expressions. However, before biclustering, some tools can automatically

perform all necessary data transformations in order to satisfy method requirements. Basically, input data types required by the methods reviewed in this thesis may be split into three groups:

- **Real.** In most cases, expression data represent a matrix of floats, each corresponding expression levels of genes in samples. Some methods are suitable to handle non-normalized inputs (although normalization is almost always recommended), while the others perform some mandatory data transformations before biclustering. For example, NMF-based methods, such as nsNMF [34], require all values to be positive. The real-valued matrix provides a more realistic expression data representation that allows more flexibility in pattern discovery. However, this data presentation does not simplify the task in any way, as compared to binarization or discretization of the data.

- **Discrete.** Some methods, e.g. xMOTIFs [196] and QUBIC [157], approximate expression values by a small set of integers. For example, in the simplest case, gene expression may be split into three bins, e.g. 1 if it is up-regulated, -1 is down-regulated and 0 otherwise. The number of such bins is usually determined by the user and may vary, although it must remain less than the number of samples. The boundaries of bins may be determined based on the quantiles of the distribution of the expression levels for each gene. The results of Eren et al. have demonstrated the choice of discretization level affects the results [76]. On the one hand, discretization simplifies and unifies input data thus helping to reduce the complexity of the problem. On the other hand, such data transformation may lead to the smoothing or even the loss of dependencies between genes, especially when the number of bins is small. This can make the method less appropriate for some patterns, e.g. those which are related to gene co-expression (see subsection 3.2.2).

- **Binary.** There are different types of biological data for which, in contrast to expression, the binary format is naturally most suitable. These include, for instance, mutation data, e.g. profiles of single nucleotide alterations, which are frequently presented as binary vectors, where every component contains 1 is the gene is mutated or 0 otherwise (see our recent works for the examples [246, 253]). Therefore, some biclustering methods, such as e.g. BicBin [272], were initially developed for handling sparse mutational data.

  Since binarization significantly simplifies the problem, binary biclustering was also applied to the expression data preceded by the binarization procedure [227, 234, 242].

For binarization of expression data, BiMAX [227] applies a cutoff equal to the mean between the minimum and maximum values of a whole dataset. DeBi [242] assigned ones to all the samples where a certain gene is above (below) of a user-defined fold-change cutoff, set the same for all genes. All the above binarization methods have important drawbacks: they all apply a threshold, same for all genes, regardless of the shape of the distribution, which may vary. Moreover, after the binarization, one has no clue about how well separated the groups labeled with 0 and 1. Finally, some relationships between gene expressions, e.g. correlation, may be lost after the binarization procedure, which makes binary biclustering methods not suitable for the detection of differential co-expression.

3.2.2 Patterns

The definition of a pattern characterizing the desired biclusters is a key part of a biclustering problem. Most methods are aimed at the detection of a specific pattern and may not perform well when the actual data contains biclusters of a different kind [76, 206]. This must be taken into account when choosing the most appropriate methods for a given research question. If the type of the desired pattern is unknown, it is worth trying several methods designed for the discovery of different bicluster types.

In agreement with the previous reviews, the patterns aimed by biclustering methods can be divided into three classes: constant values, coherent values, and coherent evolutions [223]. The examples of these patterns, as well as functions used to measure their quality, are described below in this chapter. The relationships of these patterns to biologically meaningful phenomena of differential expression and co-expression are also discussed below. Here and below $B(G',S')$ denotes a bicluster B in the matrix with expressions of G genes and S samples which include genes $G' \in G$ and samples $S' \in S$.

Constant values on rows and/or columns. Constant values on rows and/or columns. This class includes biclusters with constant values on rows, columns, or both (Fig. 3.3). As quality function for biclusters with constant values on rows and columns. Hartigan [97] proposed to use bicluster variance defined as:

$$Var(B(G',S')) = \sum_{g \in G'} \sum_{s \in S'} (b_{gs} - b_{G'S'})^2, \tag{3.1}$$

where $b_{G'S'}$ is the average over all elements of $B(G',S')$:

$$b_{G'S'} = \frac{\sum_{g\in G'}\sum_{s\in S'} b_{gs}}{|S'||G'|}. \tag{3.2}$$

If all elements b_{gs} of $B(G',S')$ are close to a constant value C, bicluster variance will be close to 0.

The definition of a constant value bicluster says nothing about the difference between values within the bicluster and outside it. If values within the bicluster are higher or lower than background values, this pattern corresponds to differential expression.

Coherent values on rows or columns. This type of biclusters implies that each of its elements b_{gs} combines the effects of g-th gene and s-th sample in an additive or multiplicative way. Addition and multiplications give rise to shifting and scaling patterns, which may present separately or together. Thus, three types of biclusters with coherent values are distinguished (Fig. 3.4):

- *Shifting:*
$$b_{gs} = C_0 r_g + C_1 c_s \tag{3.3}$$
- *Scaling:*
$$b_{gs} = C_0 r_g C_1 c_s \tag{3.4}$$
- *Shifting and scaling:*
$$b_{gs} = C_0 r_g C_1 c_s + C_2 c_s, \tag{3.5}$$
where r_g and c_s are the effects of row and column corresponding gene g and sample s, and C_0, C_1, C_2 are constants.

The method by Cheng and Church [47] aimed at finding biclusters with shifting pattern utilizes Mean Squared Residue (MSR):

$$MSR(B(G',S')) = \frac{1}{|S'||G'|}\sum_{g\in G'}\sum_{g\in S'}(b_{gs} - b_{gS'} - b_{G's} + b_{G'S'})^2, \tag{3.6}$$

where $b_{G's}$ and $b_{gS'}$ are row (gene) and column (sample) means of $B(G',S')$, respectively. For biclusters with a perfect shifting pattern, MSR equals 0. However, in independent benchmarks carried out by Eren et al. and Padilha et al. the method demonstrated the higher performance on constant biclusters (which is a special case of shift pattern), rather than on biclusters with shift pattern.

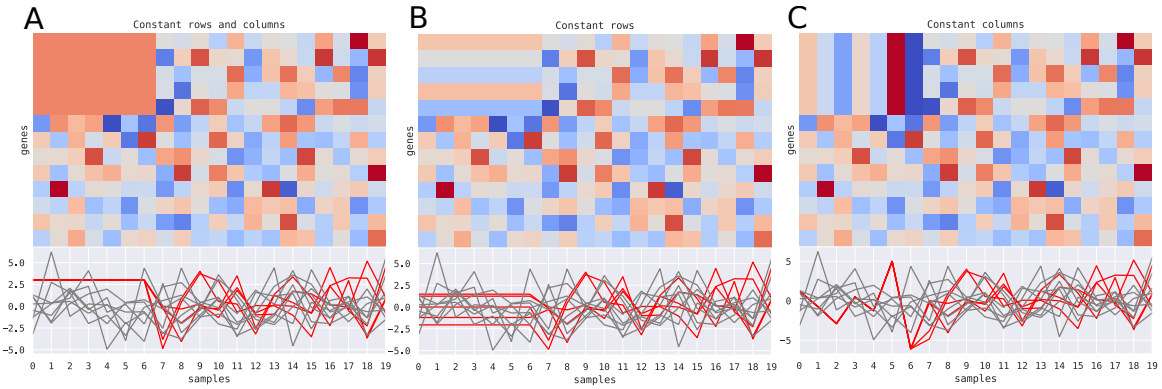

Fig. 3.3 The example of biclusters with constant values on columns and rows (A), only on rows (B) and only on columns (C). The bottom panel of each plot shows genes in parallel coordinates. Red color highlights genes that belong to the bicluster.

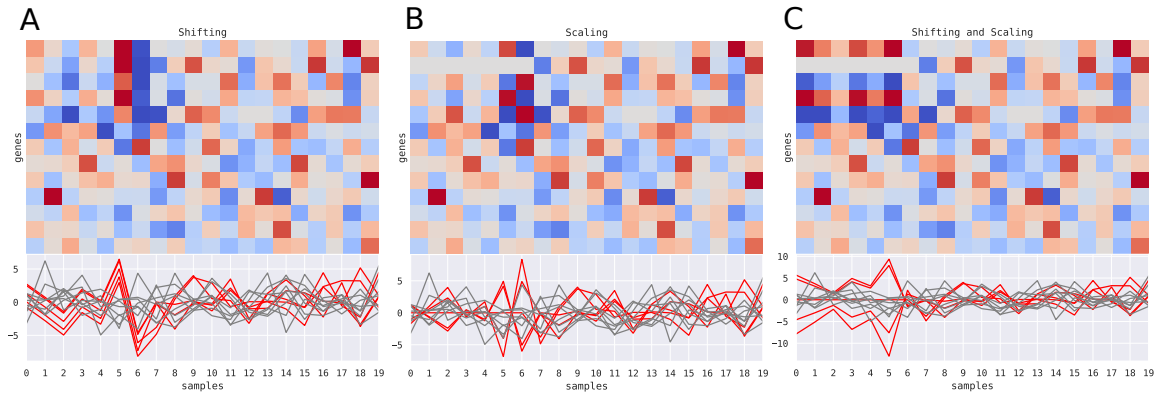

Fig. 3.4 The example of biclusters with coherent values. A. Shifting pattern. B. Scaling pattern. C. Shifting and scaling pattern.

Mukhopadhyay et al. [195] proposed a measure similar to MSR for the detection of shifting patterns, called Scaling MSR (SMSR):

$$SMSR(B(G',S')) = \frac{1}{|S'||G'|} \sum_{g \in G'} \sum_{s \in S'} \frac{(b_{gS'} b_{G's} - b_{gs} b_{G'S'})^2}{b_{gS'}^2 b_{G's}^2}, \tag{3.7}$$

MRS and SMRS measures are not capable of identifying patterns combining shifting and scaling. This problem was later solved by Pontes et al., 2013 [222] and Ahmed et al., 2014 [2]. The first proposed Evolutionary biclustering method Evo-Bexpa and a developed new evaluation measure called Virtual Error [68] (VE):

$$VE = \frac{1}{|S'||G'|} \sum_{g \in G'} \sum_{s \in S'} |\hat{b}_{gs} - \hat{b}_{gS'}|, \tag{3.8}$$

where $\hat{b}_{gs}$ and $\hat{b}_{gS'}$ are standardized values of b_{gs} and $b_{gS'}$. Owing to standardization, more similar shapes of g patterns would give lower VE.

A year later Ahmed et al. [2] published the Intensive Correlation Search (ICS) algorithm which uses Shifting and Scaling Similarity (SSSim) measure. SSSim measure defines the pairwise similarity between patterns of two genes g_1 and g_2 in $k = |S'|$ samples.

$$SSSim(g_1, g_2) = 1 - \frac{1}{k-2} \sum_{i=2}^{k-1} \frac{b_{g_1 s_{i+1}} - b_{g_1 s_i}}{b_{g_1 2} - b_{g_1 1}} - \frac{b_{g_2 s_{i+1}} - b_{g_2 s_i}}{b_{g_2 2} - b_{g_2 1}} \times \frac{1}{2max(lmean_i - \frac{b_{g_1 s_{i+1}} - b_{g_1 s_i}}{b_{g_1 s_2} - b_{g_1 s_2}}, lmean_i - \frac{b_{g_2 s_{i+1}} - b_{g_2 s_i}}{b_{g_2 s_2} - b_{g_2 s_2}})}, \tag{3.9}$$

where

$$lmean_i = \begin{cases} mean(\frac{b_{g_1 s_{i+1}} - b_{g_1 s_i}}{b_{g_1 2} - b_{g_1 1}}, \frac{b_{g_2 s_{i+1}} - b_{g_2 s_i}}{b_{g_2 2} - b_{g_2 1}}, \frac{b_{g_1 s_{i+2}} - b_{g_1 s_{i+1}}}{b_{g_1 2} - b_{g_1 1}}, \frac{b_{g_2 s_{i+2}} - b_{g_2 s_{i+1}}}{b_{g_2 2} - b_{g_2 1}}), \text{ if } i = 2, \\ mean(\frac{b_{g_1 s_i} - b_{g_1 s_{i-1}}}{b_{g_1 2} - b_{g_1 1}}, \frac{b_{g_2 s_i} - b_{g_2 s_{i-1}}}{b_{g_2 2} - b_{g_2 1}}, \frac{b_{g_1 s_{i+1}} - b_{g_1 s_i}}{b_{g_1 2} - b_{g_1 1}}, \frac{b_{g_2 s_{i+1}} - b_{g_2 s_i}}{b_{g_2 2} - b_{g_2 1}}), \text{ if } i = k-1, \\ mean(\frac{b_{g_1 s_i} - b_{g_1 s_{i-1}}}{b_{g_1 s_2} - b_{g_1 s_2}}, \frac{b_{g_2 s_i} - b_{g_2 s_{i-1}}}{b_{g_2 2} - b_{g_2 1}}, \frac{b_{g_1 s_{i+1}} - b_{g_1 s_i}}{b_{g_1 2} - b_{g_1 1}}, \frac{b_{g_2 s_{i+1}} - b_{g_2 s_i}}{b_{g_2 2} - b_{g_2 1}}, \frac{b_{g_1 s_{i+2}} - b_{g_1 s_{i+1}}}{b_{g_1 2} - b_{g_1 1}}, \frac{b_{g_2 s_{i+2}} - b_{g_2 s_{i+1}}}{b_{g_2 2} - b_{g_2 1}}), \\ \text{otherwise} \end{cases} \tag{3.10}$$

$SSSim(g_1, g_2) = 1$ when g_1 and g_2 demonstrate a perfect scaling and shifting pattern. ICS searches for maximal subspaces for which SSSim exceeds the user defined threshold.

Coherent evolutions. This group includes biclusters, whose rows and columns demonstrate a similar tendency not described by any of the above models. For example, biclusters composed of genes accordingly up- and/or down-regulated in a subset of samples, i.e. differentially expressed biclusters (Fig. 3.5). Biclusters, whose rows or columns demonstrate a higher pairwise correlation compared to the background also fall into this group. The difference of this group from biclusters with coherent values is that a perfect pattern of such bicluster is not formalized. It is important to note that biclusters with coherent evolutions may be a good approximation for biclusters with shifting, scaling and shifting-and-scaling patterns. For example, in synthetic benchmark [206] CPB [26] and OPSM [18] aimed at biclusters with coherent evolutions showed a good performance on shifting and scaling patterns.

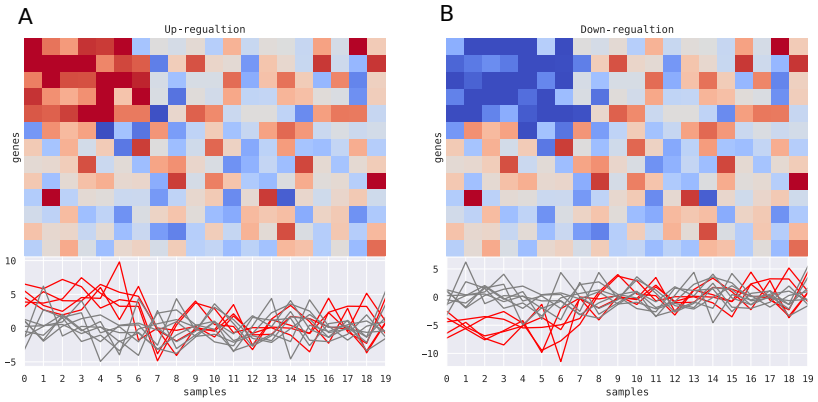

Fig. 3.5 The example of biclusters with coherent evolutions. A. Up-regulation. B. Down-regulation.

3.2.3 The usage of additional data sources

Although most of the biclustering methods require input only expression (or multi-omics [137]) data matrix, some of them can additionally incorporate orthogonal biological data to improve biclustering results. For example, COALESCE can optionally accept sequences of gene regulatory regions and perform *de novo* motif search jointly with biclustering [114]. It searches for biclusters composed of genes whose regulatory regions are enriched by the same motifs. Another biclustering method, cMonkey2, in addition to motif enrichment data, considers functional associations between genes in its scoring function [228]. However, as well as many other biclustering methods, cMonkey2 is more suitable for searching for differentially co-expressed biclusters, rather than differentially expressed. A novel version of QUBIC can utilize the data on known gene relationships when it ranks gene pairs before constructing biclusters [157].

The usage of orthogonal data about gene relationships drives the solution towards more biologically reliable biclusters. However, in isolation from the concrete task, it's hard to say how useful such biological constraints are. In any case, each method takes into account only a fixed data set, which may be far from complete and correct. For instance, the network may miss existing connections between poorly annotated genes and contain false connections between well-studied ones. Moreover, some connections may be true, but irrelevant to the studied tissue or cell type. Finally, biclusters observed into expression data may be explained by the third reason not related to the chosen data. Summarizing the above, using additional biological constraints may bias the results and lead to a systematic loss of some correct solutions.

3.2.4 Other constraints

In order to find a solution in a shorter time, some methods apply other constraints on biclusters or on the whole biclustering result. It is necessary to consider them when comparing biclustering methods because the usage of these restrictions also simplifies the problem definition and biases the results towards certain solutions.

- Number of biclusters
- Size of biclusters
- Row and column exhaustiveness
- Whether the method assumes any global data structure, e.g. checkerboard

3.3 Related works

As it was already mentioned above, biclustering methods differ first of all in the desired pattern they are searching for. Not all biclustering methods are suitable for the detection of differentially expressed biclusters. Since the differential expression is in the focus of this thesis, eleven state-of-the-art biclustering methods [19, 47, 108, 114, 151, 157, 196, 234, 242] were chosen based on their good performance on synthetic datasets with differentially expressed biclusters in a recent benchmark by [206]. However, SAMBA [261] failed to be installed and BiMAX [227] did not return any result on the real-world dataset after a week of running. Of the nine remaining methods, only QUBIC was able to take into account network data. Therefore, in addition to these methods, yet another tool able to perform network-constrained, called cMonkey2 [228] method was initially added to the baselines. However, the experiments on real data have shown that cMonkey2 was not successful in the detection of bicluster with pronounced differential expression and therefore it was not included in the benchmark. Table 3.1 compares nine selected computational tools for biclustering. The descriptions of the approaches used by the selected methods are also provided below.

3.3.1 Cheng and Church

In 2000 Cheng and Church became the first who applied biclustering for the analysis of gene expression data [47]. The proposed algorithm was aimed at maximal biclusters with Mean Squared Residue (see equation 3.6) below a user-defined threshold. It was shown to be

method	**output**	**data transformation**	**deterministic**	**prior knowledge**	**requires to specify the number of biclusters**	**theoretical runtime complexities**
Cheng & Church	biclusters with constant values	none	yes	not used	yes	$O(\|G\|\|S\|)$
xMOTIFs	coherent and constant biclusters	discretization	no	not used	no	$O(\|G\|n_s n_d)$
QUBIC	nonzero constant columns biclusters	discretization	yes	network (optional)	yes	$O(\|G\|^3\|S\|)$
Plaid	additive biclusters with coherent values	none	no	not used	yes	$O(\|G\|\|S\|Kn_{iter})$
FABIA	multiplicative biclusters with coherent values	none	no	not used	yes	$O(\|G\|\|S\|p^2 n_{iter})$
ISA2	biclusters with row and columns averages higher than T_c and T_r	none	no	not used	no	$O(n_{seeds}\|G\|\|S\|n_{iter})$
DeBi	differentially expressed biclusters	binarization	yes	not used	no	
COALESCE	up- and down-regulated biclusters	none	no	motifs (optional)	no	$O(\|G\|^2(\|G\|+\|S\|)n_{bics})$
BiBit	binary biclusters	binarization	no	not used	max. number (optional)	$O(\|G\|^2\|S\|)$

Table 3.1 Comparison of nine tested biclustering methods.

efficient only for biclusters with constant value [76, 206]. To date, the method received more than 2500 citations and remains a popular baseline for novel biclustering methods.

In addition to the expression matrix, the algorithm requires three input parameters: upper MSR threshold δ, scaling factor α and the expected number of biclusters n. The algorithm performs n iterations, each including three steps:

1. **Multiple node deletion.** At the initialization, the algorithm assigns the whole matrix to $B(G',S')$ and computes its MSR. If $MSR(B(G',S')) \leq \delta$, $B(G',S)$ it is returned; otherwise the algorithm iterates through all genes G' and samples S' and removes those whose MSR exceeds $\alpha \times MSR(B(G',S'))$, where $\alpha > 1$.

2. **Single node deletion.** Either row or columns with the largest MSR are deleted, until $MSR(B(G',S')) \leq \delta$.

3. **Node addition.** All non-bicluster rows and columns with MSR not exceeding δ are joined to B one by one.

At every iteration, the algorithm detects one bicluster and overwrites the bicluster with random values at the end of the iteration, thus removing it from the matrix.

3.3.2 Plaid

Plaid assumes that the observed expression matrix results from the sum of biclusters and background effects [151]. The value of each element e_{gs} from the expression matrix E is assumed to be composed of background expression and the effects of K biclusters:

$$\hat{e}_{gs} = \theta_{gs0} + \sum_{k=1}^{K} \theta_{gsk}\rho_{gk}\kappa_{sk}, \tag{3.11}$$

where θ_{gs0} is the background expression, ρ_{gk} and κ_{sk} are binary indicators of column and row membership, and θ_{gsk} is the impact of bicluster, defined as $\theta_{gsk} = a_{gk} + b_{sk} + m_k$, where a_{gk}, b_{sk}, and m_k denote row, column and background impacts of the biclsuter k.

Plaid is aimed to minimize the sum of squared errors between the observed and modeled expressions, assuming K biclusters in the data:

$$MSE = \sum_{g}^{G}\sum_{s}^{S}(e_{gs} - \theta_{gs0} - \sum_{k=1}^{K} \theta_{gsk}\rho_{gk}\kappa_{sk},)^2. \tag{3.12}$$

To solve this problem, plaid employs an iterative procedure, where it identifies one layer at time and removes it from E at the end of iteration. For each layer k, plaid fixes ρ_{gk}, and κ_{sk} and finds θ_{gsk} corresponding minimum squared error. Next, it similarly identifies ρ_{gk}, fixing κ_{sk} and θ_{gsk} and κ_{sk} fixing ρ_{gk} and θ_{gsk}. For each layer, plaid performs n steps of optimization of ρ, κ and θ. Initial values for ρ_{ik}, and κ_{jk} were set to 0.5 plus a small random number (distribution was not specified). Optimizations are performed until either

- k reaches K_{max}, or
- the *importance* of k-th layer defined as $\sigma_k^2 = \sum_n \sum_p \theta_{gsk}^2 \rho_{gk}\kappa_{sk}$ does not exceed $\tilde{\sigma}_k^2$ – the maximal importance of the layer obtained in r shuffled versions of E.

3.3.3 xMOTIFs

The method by Murali et al. [196] is based on the assumption that observed gene expression values may be interpreted as a small number of gene states. To reduce search space, the authors suggest considering only those intervals, which contain significantly more samples than expected by chance, assuming a uniform distribution of samples over all intervals. They determine such intervals using a one-sided hypergeometric test.

In their paper, Murali et al. introduced a term *conserved gene expression* to designate the cases when expression values of a gene are bound to a certain narrow interval in a subgroup

of samples. They represented expression profiles of samples as points in |G|-dimensional space and suggested to find *xMOTIFs* – maximal hyperrectangles in this multidimensional space. These rectangles are bound in the dimensions of conserved genes and open in all others. In other words, the definition xMOTIF may be also understood as an approximation of a bicluster $B(G', S')$ with coherent evolutions.

The searched xMOTIF must cover not less than α-fraction of all samples and to be maximal in terms of genes. Maximality in genes means that none of $G \setminus G'$ genes of the xMOTIF is not conserved in more than β samples from S'. Besides α and β, the user must specify the number $n \geq 2$ of equal-sized bins used to discretize all gene expressions, and three algorithm parameters n_s, n_d and s_d explained below.

When expressions are discretized, the algorithm randomly chooses n_s seed samples and performs n_d attempts to grow a xMOTIF from each s_{seed}:

- Randomly selects S' samples from S, $|S'| = s_d$;
- Finds all G' genes, which expressions fall into the same bin as s_{seed} in all S' samples;
- Extends xMOTIF by all samples from $S \setminus S'$ with the same pattern;
- If the resulting xMOTIF includes less than $\alpha|S|$ samples, it is discarded; otherwise, it is extended in genes and returned.

3.3.4 ISA

The first version of the Iterative Signature Algorithm (ISA) was published by Bergmann et al. [19] in 2003. ISA searches for Transcriptional Modules (TMs) – subsets of genes and samples, such that column and row averages of corresponding submatrices in normalized expression matrices would exceed user-defined thresholds T_G and T_S respectively. Normalization is performed independently for rows and columns of E. ISA centers and rescales to unit length gene and sample vectors of expression matrix E resulting in two matrices E^G and E^S. As the reader can see from the definition, TM corresponds to an up-regulated bicluster. Similarly, ISA can also find down-regulated biclusters requiring row and column averages to be below given thresholds.

ISA grows TMs from n_{seeds} randomly chosen sets of genes. Each TM may be described as a pair of gene and sample sets, e.g. G'^0 and S'^0 at initialization. Equivalently, it can be represented as a pair of binary vectors of gene and sample memberships $\overline{\mathbf{g^0}}$ and $\overline{\mathbf{s^0}}$:

$$g_i^0 = \begin{cases} 1, \text{ if } g_i \in G'^0, \\ 0, \text{ otherwise} \end{cases} \quad s_i^0 = \begin{cases} 1, \text{ if } s_j \in S'^0, \\ 0, \text{ otherwise} \end{cases}$$

ISA randomly chooses G'^0 from G and computes $\overline{\mathbf{g^1}}$ and $\overline{\mathbf{s^1}}$ for given $\overline{\mathbf{g^0}}$:

$$\overline{\mathbf{s^1}} = f(E_G^T * \overline{\mathbf{g^0}}, T_G), \overline{\mathbf{g^1}} = f(E_S * \overline{\mathbf{s^1}}, T_S), \text{ where} f(x,T) = w(x)\Theta(\bar{x} - T) \tag{3.13}$$

Here $\bar{x}$ is centered and scaled x, Θ is a function, which sets 0 to all non-positive components of a vector, and $w(x)$ is a weight function. By default, weights of all genes (samples) are simply set to 1, although they can be changed if the user wants to incorporate prior knowledge, e.g. add weight to some important genes. Iterations performed until the convergence when the $\overline{\mathbf{g}}$ change becomes smaller than tolerance ε during n last steps. The vector $\overline{\mathbf{g}^*}$ to which the system is converged and corresponding vector $\overline{\mathbf{s}^*}$ are termed in the paper "fixed point" and defines a TM candidate.

Finally, when many candidate TMs generated from multiple runs of the method, similar TMs are united. Two TMs are considered similar if correlations between their row and column vectors exceed a threshold (the authors suggested the threshold of 0.8 [120]). ISA restarts from the average of all the same TMs in order to merge them into a single one.

Later in 2010, Csardi et al. [63] published ISA2, a reimplementation of ISA in R, which is used in this thesis. In contrast with the first version, it applies z-score for normalization of expression matrix and uses slightly different thresholds $T_g\sigma_g$ and $T_s\sigma_s$, where σ denotes standard deviation for gene g and sample s.

3.3.5 COALESCE

Huttenhower et al. 2009 [114], proposed a new method for the detection of regulatory modules, given expression and sequence data, called COALESCE (Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction). In contrast with many other biclustering methods, COALESCE is able to perform biclustering jointly with motif search and enrichment analysis. It searches for biclusters formed of co-expressed genes and can optionally consider regulatory sequences of these genes. If sequences are provided, it favors biclusters composed of genes whose regulatory sequences are enriched by regulatory motifs.

COALESCE requires a matrix E of expression z-scores and can optionally accept a list of regulatory sequences associated with genes from E. If sequences are provided, COALESCE builds a matrix M of frequencies for the detected motifs before biclustering. COALESCE

extracts one bicluster at a time and therefore requires the user to specify the expected number of biclusters. Each round starts from a pair of genes with the highest correlation across all the samples. First, COALESCE determines samples whose expressions of both genes significantly different from the background. For that COALESCE applies Z-test and includes samples whose p-values do not exceed the threshold p_e. This results in a seed bicluster $B(G',S')$, thus far composed of two genes G' and S' samples. If sequence data is provided, COALESCE identifies sequence motifs and tests the hypothesis of their enrichment in sequences regulating genes from the bicluster.

COALESCE applies Bayesian integration to define the probability of each of the remaining genes to be a part of this bicluster, conditioned on the observed expression and motif enrichment data:

$$P(g \in G'|E,M) \propto P(E|g \in G')P(M|g \in G')P(g \in G'), \tag{3.14}$$

where M and E are expression and motif enrichment data, and priors are set proportional to $|G'|$. Genes with $P(g \in G'|E,M)$ exceeding probability threshold p_g join the bicluster. Thus S' and G' are iteratively updated until the convergence. Finally, the profile of the resulting bicluster (average of columns) is subtracted from E before a new round of search.

3.3.6 QUBIC

Li et al. [157] published a graph-based method for QUalitative BIClustering (QUBIC), performing in two phases. First, QUBIC discretizes input expression data and represents it as an edge-weighted network of genes. In this network, weights of edges correspond to the size of a sample subset in which expression profiles are similar. Second, it searches for heavy connected subgraphs in the network, one at time.

Discretization. For each gene, QUBIC converts its expression levels to $2r+1$ integers: $-r,...,0,...,r$. First, it determines samples in which expressions of genes are not altered (i.e. belong to background) and sets their expressions to 0. For that, QUBIC arranges all expression values in ascending order and calculates the size $2d$ of the background group, which is controlled by a user-defined parameter $0 < q < 0.5$:

$$d = min(E_{gs_c} - E_{gs_s}, E_{gs_{|S|-s+1}} - E_{gs_c}), \tag{3.15}$$

where c is the index of the median sample, and $s = |S|q+1$. A gene g is considered to be not altered in a sample s if its expression value E_{gs} belongs to the interval $E_{gs_c} - d, E_{gs_c} + d)$.

Gene expression values exceeding $E_{gs_c} + d$ are considered to be up-regulated. QUBIC splits into r equal-sized bins and encodes them as $r, ..., 1$. Gene expressions below $E_{gs_c} - d$ are discretized similarly and encoded as $-1, ..., -r$.

QUBIC considers gene expressions similar if they fall into the same bin. The similarity of two gene profiles simply is the number of samples with similar expressions, except those set to 0. At the end of this phase, expression data is represented as a network of genes with edges weighted according to the number of samples with the same expressions.

Identifying biclusters. Heavy subnetworks in the resulting network may define promising bicluster candidates. However, as the authors noted, such subgraphs do not necessarily define good biclusters. For example, two pairs of genes may share patterns manifesting in two different groups of samples. To avoid finding subnetworks composed of heavy, but inconsistent edges, the authors introduced a consistency threshold $0 < c \leq 1$. The consistency of a bicluster $B(G', S')$ equals a minimum fraction of matching expressions among all columns (samples).

QUBIC performs multiple rounds of search and tries to identify one heavy subnetwork at time. Since the method is aimed at maximal biclusters, at each round, QUBIC starts from the edge with the highest weight and grows the subnetwork in four steps, until joining new genes is possible without violation of the consistency condition.

1. QUBIC selects a pair of genes g_1, g_2 with the highest weight. At the first round of search, when no biclusters are detected, all edges are considered as seeds. Later, when some biclusters are detected, seeds are additionally checked for the following conditions

   - at least one of its genes g_1 or g_2 is not in any bicluster, or
   - g_1 and g_2 are in different biclusters $B(G'_1, S'_1)$ and $B(G'_2, S'_2)$ not overlapping in genes and the weight of this pair exceeds $\max(|G'_1|, |G'_2|)$.

   For a chosen pair of seed genes $G' = g_1, g_2$, a set of samples S' with matching non-zero expressions is determined.

2. Iteratively adds to G' all genes, such that the consistency of a bicluster does not decrease. When g is added to G' resulting in a new gene set $G'' = G' \cup \{g\}$, a new set of samples S'' preserving perfect consistency of $B(G'', S'')$ is identified. The joining of a new gene g is only allowed if the minimal dimension of a bicluster does not decrease: $min(|G''|, |S''|) \geq min(|G'|, |S'|)$.

3. Adds new samples to S' starting from those, whose joining results in maximal consistency of B. Stops when joining new samples is no longer possible without dropping the consistency below the threshold of c.

4. If the user is interested in finding biclusters with genes demonstrating inverse patterns, step (3) is repeated considering the same expressions with opposite signs.

In the paper by Li et al. published in 2009 provides the implementation of QUBIC in C. In 2017 Zhang et al. [294] released a new version of QUBIC implemented in R. Besides functions for data discretization, heatmap visualization and building of co-expression network, the new version of QUBIC allows the user to incorporate prior knowledge in the form of the weighted gene network. This query-based version of QUBIC sums edge weights of the input network with the network obtained at the first step, thus increasing the weights of gene pairs considered to be more relevant.

3.3.7 FABIA

Hochreiter et al., 2010 [108], proposed the biclustering method called FABIA (Factor Analysis for Bicluster Acquisition). Given the normalized expression matrix, FABIA searches for biclusters with a scaling pattern, which is modeled as a product of two sparse vectors λ and z (Fig. 3.6).

Different from other methods, FABIA assumes that all background genes not participating in any bicluster are deleted from the input data, which may be hard to achieve in practice. Same as plaid, FABIA tries to decompose the input expression matrix into a sum of $p+1$ layers, where p layers correspond to biclusters and one is the noise ε:

$$E = \sum_{i=1}^{p} \lambda_i z_i^T + \varepsilon. \tag{3.16}$$

Thus, the biclustering problem is reduced to the task of learning all z_i and λ_i resulting in maximum *a posteriori*. FABIA sets Laplace distributions as priors for z and λ and applies variational EM to find the maximum of the posterior. The estimated z_i and λ_i defines the fuzzy gene and sample membership in the i-th bicluster. If necessary, FABIA applies a threshold to identify crisp memberships of genes and samples in a bicluster.

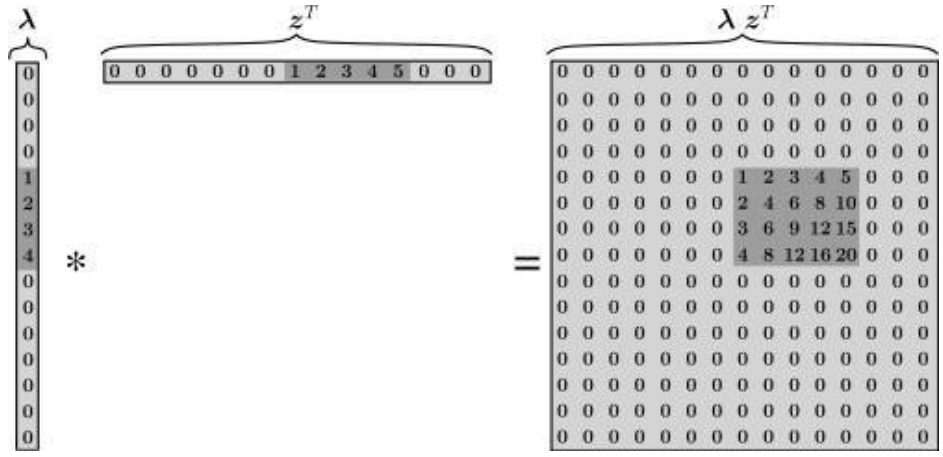

Fig. 3.6 The product of gene and sample effect vectors λ and z resulting in a bicluster with a scaling pattern. From Hochreiter et al., 2010 [108].

3.3.8 BiBit

Rodriguez-Baena et al. [234] proposed a binary biclustering method called BiBit (Fig. 3.7). BiBit accepts a binary matrix, and two parameters min_r and min_c determining minimal numbers of rows and columns in a bicluster. For binarization, the authors use a two-step approach. First, all gene expressions were standardized and those which laid outside the interval [-3;3] were set to -3 or 3. Then, expressions were split into 12 equal-sized bins and those which fall into the first six bins considered to be down-regulated and the rest – up-regulated.

In order to compress the data, Rodriguez-Baena et al. suggest grouping columns by n_{bits} and encode every row as an integer and perform a further search on the encoded matrix.

After the encoding phase, the method iterates over all pairs of rows and forms a seed pattern applying logical “AND” operation on their integer representations. Further, this seed pattern is compared with all remaining rows in the same way. If the result of “row AND pattern” again matches that pattern, the row is included in the bicluster. The authors also provided an extended version of BiBit, able to tolerate a specified proportion of zeros in the resulting bicluster.

3.3.9 DeBi

In 2011 Serin and Vingron [242] proposed a method for the discovery of differentially expressed biclusters called DeBi. The method binarizes the expression matrix and utilizes the Frequent Itemset Approach (MAFIA) [32] algorithm to detect maximal binary biclusters.

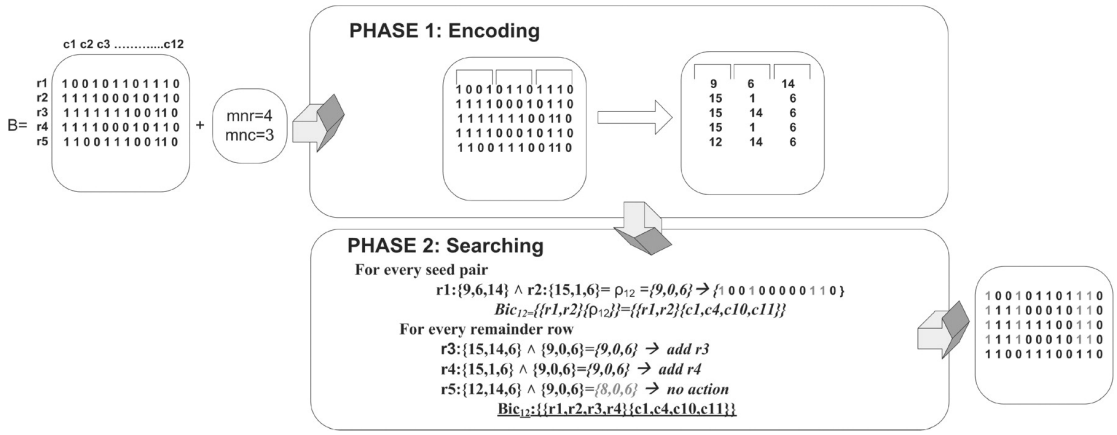

Fig. 3.7 A scheme of BiBit workflow. From [234].

According to the choice of the user, DeBi can search for up-regulated, down-regulated biclusters, or biclusters combining both patterns. For binarization of expression data, DeBi applies a fold-change threshold on expression values. Given a binary matrix, where 1 reflects expressions matching the desired pattern, DeBi enumerates all maximal binary biclusters in three steps:

1. **Finding seed biclusters.** At the first step, DeBi is aimed at finding maximal perfect binary biclusters. This task is equivalent to finding maximal c-frequent gene sets. Gene set G' is c-frequent if its support is larger. The support of a gene set is just the proportion of samples, in which expressions of all genes are ones. If no superset of G' is c-frequent, G' is called maximal. Given that each subset of a c-frequent gene set is also at least $c-$frequent, instead of enumeration of all possible gene sets, DeBi iteratively applies MAFIA algorithm with varying support thresholds arranged in descending order. At each iteration, MAFIA searches for *maximally frequent gene sets* with support exceeding a given threshold. Initially, the support threshold c is set equal to the maximal support among all individual genes. c is further decreased by $\frac{1}{|S|}$ per iteration until it reaches minimal support specified by the user.

2. **Expanding seed biclusters.** The first step of DeBi results in a set of perfect binary biclusters. Given that it uses an arbitrary threshold for binarization, some genes might be binarized incorrectly, and therefore missing in biclusters. To recover such genes, DeBi tests all genes for association with each of seed biclusters. All genes for which the significance of the overlap of non-zero samples and S' exceeds a p-value threshold of α, are joined to the bicluster. To reduce computations, DeBi precomputes the sizes of overlaps that yield a p-value higher than α.

3. **Removal of overlapping biclusters.** Finally, starting from the largest bicluster, DeBi removes all biclusters which overlap in more than $0 < o < 1$-fraction of its area with a bigger one.

3.4 Validation approaches

Validation approaches can be classified as supervised when ground truth is available and unsupervised when it is not. In the first case, method performance may be calculated directly, comparing a set of found biclusters with a set of known biclusters. For the evaluation of biclustering results, many similarity measures suitable for comparison of two sets of biclusters have been developed [110]. The choice of an optimal metric depends on the task, such as the tolerance to first and second type errors, redundancy of the results, etc. In this thesis, Relevance and Recovery scores proposed by Prelic et al. [227] and used in previous benchmarks [76, 258] were chosen (see subsection 4.5.1).

Direct performance evaluation is complicated by the fact that ground truth data may be unavailable or not reliable. Breast cancer chosen for this thesis is known to have several well-characterized subtypes, distinguishable at the level of gene expression. However, even if some molecular subtypes of breast cancer are specified based on gene expressions, we cannot refer to them as absolute ground truth, because:

- unknown disease subtypes determined by expressions of different genes may exist along with known;
- these known subtypes may be defined imprecisely. Indeed, some recent works suggested the extensions [224, 226] of PAM50 molecular classifications [209, 216].

Similar considerations may concern almost every biological dataset. This means that the evaluation of biclustering on real data may result in a biased performance estimate.

An alternative way of supervised evaluation in the absence of ground truth is a benchmark on synthetic data. This approach also allows the direct computation of performances and therefore is widely used by the community for the evaluation of biclustering methods. In this setting, the experimenter has full control of the data. This allows investigating the dependence of various data properties such as the number of biclusters, overlap, level of noise, etc. on the method performance. The results of benchmarks performed by Bozdag et al. [26], Eren et al. [76], and Padilha et al. [206] have demonstrated that method performances may vary widely depending on these characteristics of the data.

The main disadvantage of this approach is that simulated data may not reflect the complexity of real-world data or miss some of its important aspects. This may lead to the over- or underestimation of method performances (see the discussion in chapter 5 for the details).

In the absence of reliable ground truth, the indirect validation of the results obtained on real data is still possible. Genes falling into the same bicluster are demonstrating a similar pattern of expression are expected to be functionally related. Therefore to obtain indirect evidence of method performance, the resulting biclusters are tested for biological significance. Almost all of the above methods discussed in this chapter test gene sets for overlap with Gene Ontology (GO) categories. GO is a controlled vocabulary of gene attributes, providing annotations of genes with molecular functions they perform, biological processed they participate and cellular components in which they work. Overrepresentation of genes labeled with the same GO term in a bicluster compared to background genes points to their functional coherence and supports the reliability of this bicluster.

A similar idea can be applied for the evaluation of patients groupings obtained in the result of biclustering. They can be tested for associations with various biological variables like known disease subtypes or survival. Of course, the absence of association of bicluster with any functional group or clinical variable result does not necessarily mean that it is defined incorrectly.

Chapter 4

Methods

The lack of biclustering methods specifically aimed at the detection of differentially expressed biclusters motivated the development of a novel biclustering method called [298]. To reduce search space and obtain more robust biclusters, we suggested adding gene network to the problem definition and searching for network-constrained differentially expressed biclusters. This chapter starts from the formal problem definition (section 4.1, published in [298]), represents the first version of DESMOND (section 4.2, published in [298]), and introduces the second version of the method (section 4.3). Theoretical analyzes of runtime complexity for both versions of DESMOND are provided in section 4.4. Sections 4.5, and 4.6 (also adapted from [298]) explain data preprocessing and validation approaches respectively. The details on the implementation of the methods are provided in section 4.7.

4.1 Problem definition

The problem addressed in this thesis is the discovery of connected groups of genes differentially expressed in an unknown subgroup of samples, given a network of gene interactions and a matrix of gene expression profiles (Fig. 4.1). This problem can be classified as network-constrained biclustering, or, alternatively, as unsupervised active subnetwork detection, when the desired sample subgroups are unknown.

Formally speaking, given expressions of genes in G measured in the samples of set S, and an undirected and unweighted graph $N = (G, I)$, representing I interactions between the G genes, the aim is to find subsets of $G' \subset G$ genes and $S' \subset S$ samples, such that genes G' are differentially expressed in a subset of samples S' compared to the background samples $\overline{S'} = S \setminus S'$; and G' forms a connected component in the network N. Such pairs (G', S') are called *modules* which is a synonym of bicluster in the context of this thesis. A gene g is

differentially expressed in a set of samples $S' \subset S$ compared to $\overline{S'} = S \setminus S'$, if $\mu_{g,S'}$, its median expression in S', is different from the median expression $\mu_{g,\overline{S'}}$ in $\overline{S'}$. Since the aim of this thesis is the discovery of gene subsets that differentiating putative disease subtypes, it is important to find biomarkers which expressions in S' would be well-separated from the background. To control how well the expression of the gene g distinguishes the group of samples S' from the background, one can employ the signal-to-noise ratio (SNR) [100, 187]. The SNR for expression of gene g in S' samples is defined as

$$SNR(g,S') = \frac{\mu_{g,S'} - \mu_{g,\overline{S'}}}{\sigma_{g,S'} + \sigma_{g,\overline{S'}}}, \tag{4.1}$$

where μ and σ denote mean and standard deviation of gene expression in a subgroup of samples.

Similarly, a set of genes G' is also called differentially expressed in the samples of set S' if $\forall$ gene $g \in G'$, g differentially expressed in S'. The average of absolute SNR over all genes G' is used as a measure of differential expression of a bicluster $B(G',S')$:

$$avg.|SNR(B(G',S'))| = \frac{1}{|G'|} \sum_{g \in G'} |SNR(g,S')| \tag{4.2}$$

A higher average absolute SNR value indicates that a subset of samples S' is well-separated from the background in a subspace of G'. Such gene sets are promising biomarker candidates for distinguishing unknown but biologically relevant subtypes of samples.

In the standard setting of differential expression analysis, all genes are tested in two given groups, e.g. disease vs control. In contrast, in the biclustering problem, the groups of samples are undefined and are to be discovered. If genes are up-regulated in more than half of all samples, the remaining samples also form a down-regulated module and *vice versa*. Therefore, it makes sense to search for groups of samples of size not bigger than $|S|/2$. Furthermore, the desired module should not be too small in terms of samples, because a smaller module has a higher probability to appear just by chance. To avoid finding too small modules, the user can select an appropriate s_{min} value based on the size of the dataset and intended downstream analysis.

4.2 DESMOND Algorithm

To solve the problem formulated above, a new method for identification of **D**ifferentially **E**xpre**S**sed gene **MO**dules i**N** **D**iseases (DESMOND) has been developed. The first version of

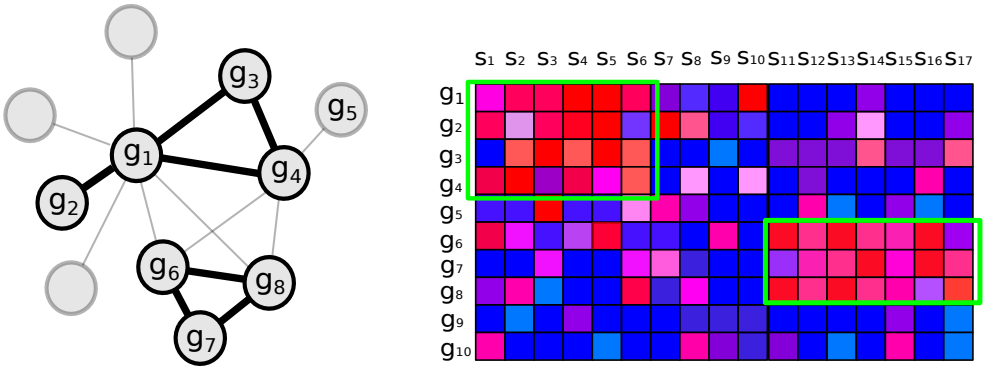

Fig. 4.1 The scheme illustrating the searched network-constrained biclusters on the example of a toy gene network (on the left, nodes represent genes, edges connect functionally related gene) and expression matrix (on the right, rows, and columns correspond genes and samples respectively). Genes connected in the network and differentially expressed (up-regulated) in subgroups of samples are shown bold. Biclusters including the up-regulated and connected genes and samples, in which these genes are overexpressed, are highlighted by green frames in the expression matrix.

DESMOND is published in [298] and described in deeper detail below. To identify network-constrained biclusters, potentially representing disease modules, DESMOND performs three phases (Fig. 4.2):

1. **Identifying samples, in which genes demonstrate an altered level of expression compared to the background.** A similar problem is faced by other biclustering methods, such as BiMAX [227], BiBit [234], or DeBi [242]. As discussed in subsection 3.2.2, the disadvantage of the binarization approaches utilized by these methods is that they apply the same cutoff on all genes, regardless of the distribution of expression. This approach may not work well, for example, when genes dysregulated in groups of samples of different sizes. To avoid this obstacle, DESMOND searches for a group of samples, in which a pair of genes are concordantly (both up- or down-regulated). In contrast with other methods, for each pair of genes, DESMOND identifies an individualized pair of binarization thresholds, such that

    - the overlap between samples demonstrating concordantly altered expressions is significantly larger than random;
    - differential expression is pronounced, i.e. average *SNR* computed for the bicluster and the background is high.

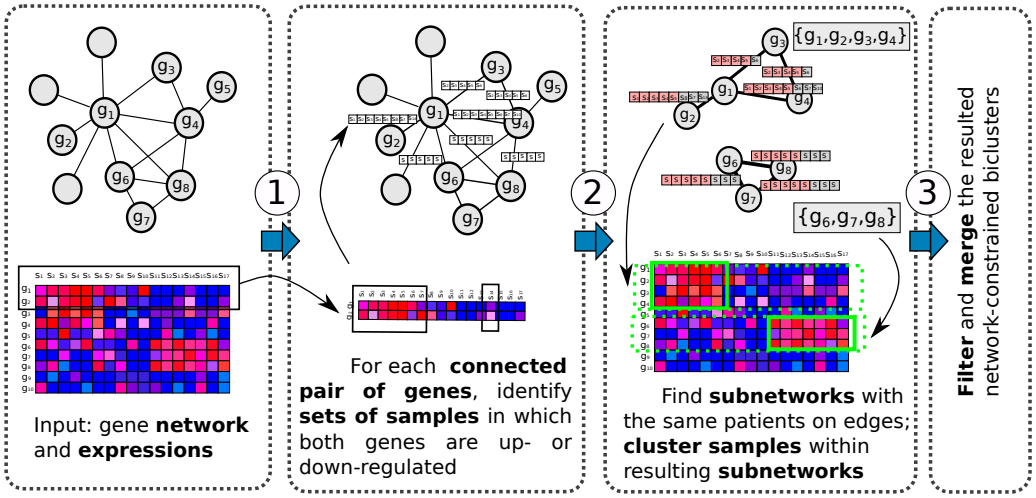

Fig. 4.2 Three phases of the DESMOND algorithm. 1. For each connected gene pair, identifying sample groups, in which genes demonstrate concordantly altered expressions. 2. Grouping of gene pairs (edges) which are dysregulated in similar sets of samples into subnetworks and identifying biclusters in the subspaces of these subnetworks. 3. Post-processing – merging biclusters overlapping in samples and removing biclusters with too few genes or too low *SNR*.

To reduce computations, DESMOND considers only connected pairs of genes at this phase.

2. **Grouping pairs of genes that are dysregulated in similar sets of samples.** On the second step, DESMOND performs probabilistic clustering of edges, associated with a non-empty set of samples. Edges are assembled into subnetworks, composed of edges associated with similar sets of samples. Each subnetwork gives rise to a bicluster, obtained when samples are split into two groups in a subspace of subnetworks.

3. **Post-processing.** Merging biclusters overlapping in samples and discarding biclusters with less than three genes or weakly differentially expressed.

4.2.1 Step 1. Assigning sample sets to edges

In the first step, for each interaction edge i connecting genes u and v, DESMOND identifies a maximal set of samples $S_i^{shared} = \{s_1 \ldots s_n\}$ in which both u and v are differentially expressed compared to $S \setminus S_i^{shared}$. For that, it employs a modification of the Rank-Rank Hypergeometric Overlap (RRHO) method [221] (Figure 4.3), originally developed for comparison of differential expression profiles obtained in two experiments. It searches for a group of

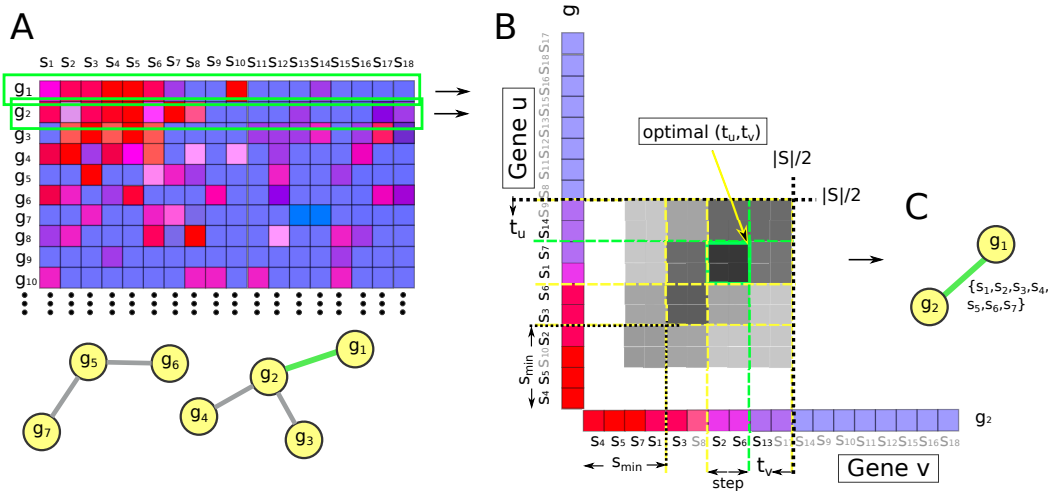

Fig. 4.3 Modified RRHO method used to find the maximal set of samples, in which two interacting genes g_1 and g_2 are up-regulated. A. Input network and expression matrix, red and blue respectively indicate higher and lower expressions. B. Two lists of samples arranged in decreasing order of the expression values of g_1 and g_2. Two thresholds t_1 and t_2 move from $\frac{|S|}{2}$ to s_{min} with step size 2. The intensity of the cell color shows overlap significance for corresponding thresholds. For the case of down-regulation, the same procedure applies, but gene profiles are sorted in ascending order. C. A set of samples S^{shared} assigned to the edge connecting g_1 and g_2. From [298].

genes significantly enriched in the tops or bottoms of two ranked lists. Basically, this method finds an optimal pair of thresholds, for which the enrichment in tops (bottoms) of the ranked list is the most significant and returns a set of genes with expressions above both thresholds. For two ranked lists of genes, the method creates a 2D-heatmap of the one-sided Fisher's exact test p-values showing the significance of every pair of threshold values t_u, t_v, picking a combination corresponding to the most significant overlap.

DESMOND uses a modification of the RRHO method to find for a given connected pair of genes u and v a group of samples of size between s_{min} and $\frac{|S|}{2}$, such that both genes are concordantly dysregulated in that sample group. Different from the original RRHO method, DESMOND moves the thresholds from the middle of the lists to the top and stop when achieving the first significant overlap and averaged $|SNR|$ value above SNR_{min}. This SNR_{min} threshold value could be explicitly defined by the user or estimated based on the data. A maximal set of samples S_i^{shared} in whose expressions of u and v are both above the thresholds and whose $avg.|SNR| > SNR_{min}$ is assigned to the edge i. If no significant overlap bigger than s_{min} found, the edge is excluded from further consideration.

4.2.2 Step 2. Probabilistic edge clustering

In the result of the first step, every edge is assigned a set of samples in which the pair of genes connected by this edge is up-regulated (or down-regulated). In step two, the algorithm groups edges into connected components, such that each component contains edges with similar sets of samples.

The output of the first step may be also represented as a binary matrix $X = [x_{ji}]_{n\times m}$ for n edges and m samples, such that $x_{ji} = 1$ if sample i is assigned to edge j and $x_{ji} = 0$ otherwise. For clustering the rows of this matrix (i.e. edges) into expression modules, DESMOND models constrained Bayesian mixture of Bernoulli distributions. The underlying distributions of the mixture model are as follows:

$$\begin{aligned} x_{ji}|\theta_{ic}, s_j &\sim Bernoulli(x_{ji}|\theta_{is_j}), \\ \theta_{ic}|\alpha &\sim Beta(\theta_{ic}|\alpha/2, \alpha/2), \\ s_j|\pi &\sim Categorical(s_j|\pi), \\ \pi|\beta &\sim Dirichlet(\pi|\underbrace{\beta/K, ..., \beta/K}_{K \text{ of them}}) \end{aligned} \tag{4.3}$$

In the above model, the assignments of samples to the edges are modeled as a Bernoulli distribution with parameter θ_{ic} and a Beta prior, for each sample $1 \leq i \leq m$ and module $1 \leq c \leq K$. The number of modules is set to K, equal the number of non-empty edges of the network resulting from step 1. s_j, $1 \leq s_j \leq K$, indicates the module to which edge j belongs and follows a categorical distribution with parameter π and a Dirichlet prior. The model initializes with each edge assigned to a separate module.

Further, DESMOND performs collapsed Gibbs sampling for parameter learning. Each iteration of Gibbs sampling goes over all edges and samples the edge indexes $s_j (1 \leq j \leq n)$. The Gibbs sampling includes two phases: (1) burn-in, consisting of several consecutive iterations for initialization of s_j, and (2) sampling, which consists of several iterations throughout which the values of s_j are recorded for further analysis for identification of modules.

At each Gibbs sampling iteration (either in the burn-in or in the sampling phase), to sample the value of s_j, DESMOND first computes the marginal conditional probability of each s_j belonging to a module k as follows:

$$P(s_j = k|X, s_{-j}, \alpha, \beta) \propto P(X, s_j = k, s_{-j}, \alpha, \beta) = \int_\pi \int_\theta [P(X|\theta, s_j = k, s_{-j}) P(\theta|\alpha) d\theta] P(s_j = k, s_{-j}|\pi) P(\pi|\beta) d\pi \tag{4.4}$$

where s_{-j} indicates the current assignment of all edges except edge j to the modules. Because the method uses conjugate priors (Beta and Dirichlet) the products are in closed form and integrations over π and θ are straightforward. Keeping the terms that vary with k, conditional probability is expressed as follows:

$$\begin{aligned} P(s_j = k|X, s_{-j}, \alpha, \beta) \propto & \\ & \prod_{i:x_{ji}=1} \left[\frac{\alpha/2 + \sum_{l:s_l=k, l\neq j} x_{li}}{\alpha + |\{l : s_l = k, l \neq j\}|} \right] \\ & \times \prod_{i:x_{ji}=0} \left[\frac{\alpha/2 + \sum_{l:s_l=k, l\neq j} (1 - x_{li})}{\alpha + |\{l : s_l = k, l \neq j\}|} \right] \\ & \times \frac{|\{l : s_l = k, l \neq j\}| + \beta/K}{n - 1 + \beta} \end{aligned} \tag{4.5}$$

No information is stored about s_j during the burn-in phase. During the sampling phase, which consists of the last 20 iterations before the convergence, the values of s_j are recorded. We assume convergence when edge transition probabilities stabilize. Specifically, edge transition probability matrices P_i from the previous 20 model states are computed starting from $i+1$-th iteration. Sampling stops when $RMS(Pi, P_{i+1})$ reaches a plateau. Achieving the plateau is detected based on the slope of a line fitting the curve. When the slope remains between $-t$ and t during the last r iterations. By default, t and r are set to 0.1 and 5 respectively, although the user has an opportunity to change these parameters. The example on Fig. 4.4 shows the dynamics of the number of oscilating edges and $RMS(Pi, P_{i+1})$ during DESMOND run. The final modules are computed as the most frequent value of s_j for each j in the last 20 iterations.

Candidate modules obtained in the result of probabilistic edge clustering contain from zero to many edges and can overlap in genes and samples. Each non-empty module represents a subnetwork, defining a subspace of genes in which samples could be split into two groups differentially expressing these genes. To split all samples into the aforementioned two groups, DESMOND performs 2-means clustering of samples in a subspace of genes representing each module.

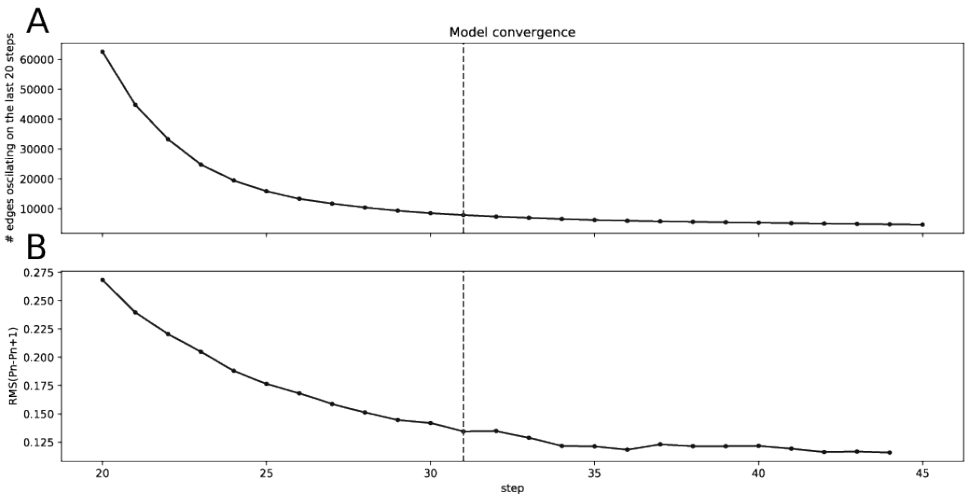

Fig. 4.4 The convergence of the model build for TCGA-micro dataset (see subsection 4.4.2) with $\alpha = 0.5$, $\frac{\beta}{K} = 1.0$ and $p = 0.005$ on step 31. The dynamics of the total number of edges changing their module membership during the last 20 steps (A), and $RMS(Pi, P_{i+1})$ (B). Dashed lines show the border between the burn-in and sampling phases.

4.2.3 Step 3. Post-processing

Since DESMOND aims to discover subnetworks of differentially expressed genes distinguishing unknown disease subtypes, all the modules with less than two edges and too low $avg.|SNR|$ must be removed. The user can either explicitly define the SNR_{min} threshold or draw a certain quantile q from the distribution of $avg.|SNR|$ values computed for 1000 "minimal" biclusters – randomly chosen network edges.

Finally, to find more complete gene modules, DESMOND merges interconnected modules, dysregulated in the same samples. This is necessary because

- reference biological networks are incomplete [168],
- local structure of the network, e.g. changes of network connectivity, may force the method to detect parts of a large bicluster as separate smaller biclusters.

Therefore, DESMOND recursively merges modules, starting from the pair with the most significant overlap in samples (Bonferroni-adjusted p-value < 0.05). The merge is only allowed if $avg.|SNR|$ of the resulting bicluster exceeds SNR_{min}. The procedure is repeated until no merge is possible.

4.3 DESMOND2

After the implementation of the first version of DESMOND, another idea of how to binarize gene expressions came up. This problem may be solved by fitting the observed distributions of gene expressions with a mixture of two distributions, e.g. two Gaussians. Based on this idea, the second version of DESMOND was developed. In contrast with the first version, which must be run independently for the detection of up- and down-regulated biclusters, DESMOND2 can also detect biclusters mixing up- and down-regulated genes.

Same as the first version, DESMOND2 performs three steps, of which only first is changed. In the first step, DESMOND2 models the distribution of expression of each individual gene as a mixture of two Gaussians. For that, it uses *GaussianMixture* function from the python library *scikit-learn* v 0.19.1, which implements expectation-maximization (EM) algorithms to learn model parameters from the data. The initial distribution of sample memberships was obtained from the results of 2-means clustering of expressions. The maximal number of iterations was restricted to 300 and the other parameters were set to default values.

After fitting the model, each sample was assigned to one of the two components. Depending on the expression pattern, chosen by the user, DESMOND2 determines background and bicluster samples:

- **Mixed pattern.** Samples from the component including less than $\frac{|S|}{2}$ samples are assigned to the bicluster and marked as 1, and the rest are assigned to the background.
- **Up-regulation.** Samples from the component with a higher median are assigned to the bicluster and marked as 1, and the rest are assigned to the background.
- **Down-regulation.** Samples from the component with a lower median are assigned to the bicluster and marked as 1, and the rest are assigned to the background.

Same as in the first version, genes with an average *SNR* between bicluster and background groups or less than s_{min} samples were excluded.

On the second step, DESMOND2 clusters connected genes instead of gene pairs, using the same approach as the first version. This is advantageous at least for computational complexity because the number of edges is 1-2 orders of magnitude greater than the number of nodes. The third step of the algorithm remained unchanged compared to the first version of DESMOND.

dataset (genes x samples interactions)	parameters	**Simulated data (2000x200, 14863)**	**TCGA-micro (11959x529, 179514)**	**TCGA-RNAseq (11959x1081, 179514)**	**METABRIC (11959x1904, 179514)**
DeBi	D	49.4	26928	87575	NA
DeBi	O	57.8	12953	7228	16714
COALESCE	D	21.2	38819	85498	44309
COALESCE	O	15.8	5260	32552	2430
FABIA	D	47.6	180	391	610
FABIA	O	39.6	2971	6062	9860
ISA	D	58.1	380	357	734
ISA	O	79.8	267	500	659
QUBIC	D	36.1	7299	9332	7523
QUBIC	O	50.8	8890	7488	7713
DESMOND	O	35.5	10229	74841	49465
DESMOND2	O	76.9	1449	2553	2964

Table 4.1 Algorithm runtimes in seconds.

4.4 Analysis of the runtime complexity

Table 4.1 reports runtimes demonstrated by both versions of DESMOND and their competitors measured on a synthetic dataset of 200 samples, 2000 genes connected by 14863 edges, with 10 implanted 100x100 biclusters, and three real datasets of almost 12 thousand genes connected by 179514 interactions and 529 – 1904 samples (see the details in subsections 4.5.1 and 4.5.2 respectively). All runtimes were estimated on Dell Latitude E5470 laptop with Intel core i5 vPro, and 16GB RAM for methods run with default and optimized parameters (see subsection 4.6.1).

The current implementation of DESMOND demonstrates one of the longest runtimes among the compared methods. Its second version was much faster and showed the best runtime among network-based methods on large real-world datasets.

Theoretical analysis of DESMOND and DESMOND2 runtime complexities, assuming input with G genes, S samples, and I interactions, are provided below.

4.4.1 DESMOND

Step 1. The complexity of the first step depends on the number of edges and the number of samples. The direct approach would require $(\frac{|S|}{2})^2$ computations of average SNR and exact Fisher's test p-values for each of |I| edges. Instead, the current implementation of DESMOND uses several approximations: DESMOND does not check every possible pair of

thresholds but creates a "grid" of thresholds at the distance $max(1, 0.01 \times |S|)$. This makes the number of Fisher's exact tests independent from $|S|$. Fisher's exact test is not computed for every pair of thresholds. Instead, in the beginning DESMOND precomputes a table of critical overlap sizes for each pair of threshold positions, given a p-value cutoff. It uses a precomputed table for every edge and thus avoids running multiple exact Fisher's tests for each edge. Computing average SNR linearly depends on the number of samples. With the above approximation, total complexity of the first step is $O(|I||S|)$.

Step 2. Input of the second step is a binary matrix of size $|I| \times |S|$ reflecting whether each sample is dysregulated for each edge. Every of Gibbs Sampler rounds includes iteration over all the edges, and sampling of a new module for an edge. The latter requires computation of joining probabilities for all neighbouring edges, if their modules were changed. Computing the probability of merging an edge with a module requires an iteration over samples and costs $O(|S|)$. It does not depend on module size because the number of ones for each patient in a module are stored in a separate matrix and updated when the module is changed.

The number of neighbours of each edge may vary widely depending on the network topology. A fully connected undirected network of $|G|$ genes without duplicated edges has $\frac{|G|(|G|-1)}{2}$ edges, although constraining on a fully connected network would make no sense. Therefore the complexity of one round of sampling is upper bounded by $O(|S||I||G|)$. In real biological networks, e.g. in PPI networks, the actual number of neighbours is much smaller than the total number of nodes. Moreover, the majority of nodes are located on network periphery and have a small degree.

Step 3. In the third step, DESMOND tests overlap all of K modules with at least 2 edges, resulting in a previous step. In the worst case, if all edges grouped in modules by two $K = |I|/2$, although in reality K is smaller. Module merging is performed iteratively, starting from the pair with the most significant overlap in samples, until no pair can be merged without dropping $avg.|SNR|$ below SNR threshold. For that, K^2 comparisons are needed for each iteration and there may be $K-1$ merges in the worst case. Every attempt of merging require running 2-means of samples, which complexity is $O(2n_{iter}|S|)$. The total complexity of Step 3 therefore is limited by $O(|S||K|^3)$ which equals $O(|S||I|^3)$ in the worst case. Considering that $|I|$ is not less than $|G|-1$, and in most real biological networks $|I| > |G|$, the overall worst-case complexity of the algorithm is defined by the complexity of the third step and is cubic of $|I|$.

4.4.2 DESMOND2

Step 1. On this step, a mixture of two Gaussians is fitted for the distribution of expressions of each individual gene. It is done by EM algorithm, which runtime linearly depends on the number of data points $|S|$, therefore the whole step take $O(|S||G|)$.

Steps 2 is nearly unchanged compared to the first version. The main difference is that DESMOND2 clusters genes instead of edges, and $|G|$ is much less than $|I|$. Therefore the complexity of step 2 is upper bounded by $O(|S||G|^2)$, by analogy with step 2 in the first version. It results in a smaller K, than the second step of DESMOND. The third step remains unchanged. This step is the most expensive For DESMOND2 and determines overall runtime complexity of the method.

As one can see from theoretical analysis, DESMOND2 has lower complexity than DESMOND and must be faster. This agrees with running times observed on three real datasets, but not on the synthetic dataset Table 4.1. On the synthetic dataset, running time of DESMOND2 is approximately twice longer, than running time of DESMOND. This is explained by the fact that, besides $|G|$,$|S|$,$|I|$, the actual runtime is strongly influenced by α. High values of alpha facilitate faster convergence and favour a quick formation of large and unspecific modules. DESMOND achieves the convergence faster than DESMOND2 on this dataset.

4.5 Datasets

4.5.1 Generation of synthetic datasets

A strategy similar to the described in literature [76, 206] was chosen for synthetic expression data generation. For every gene, its expression value was sampled from normal distribution $\mathcal{N}(2,1)$ if the gene and sample belonged to a bicluster, or from $\mathcal{N}(0,1)$ otherwise. Since no assumption of the prevailing bicluster sizes in real data is made, 20 expression matrices with implanted biclusters of varying shapes were generated. Each matrix subjected to the insertion of 10 biclusters with the size of 5, 10, 20, 50 or 100 genes and 10, 20, 50 or 100 samples in every matrix. For each implanted bicluster, gene and sample sets were chosen randomly from all genes and samples, i.e. overlaps in genes and samples were allowed.

For each synthetic expression dataset, a scale-free network of 2000 nodes was created using *scale_free_graph* function from Networkx 1.10 python package, implementing the procedure proposed by [23]. Setting the parameters $\beta = 0.9$, and $\alpha = \gamma = 0.05$ resulted in a

scale-free networks with the exponent equal 2.44. Gene labels were assigned to the network nodes in a way so that genes from the same bicluster would be connected in two steps:

1. Initially, genes belonging to exclusively to each bicluster were assigned to the network, adopting the approach proposed by Ghiassian et al., 2015 [89]. They have shown that disease-associated genes form compact but not densely connected components on PPI and developed DIAMOnD, a disease module detection method based on this idea. A subnetwork corresponding to an exclusive part of each bicluster was initialized from a random unlabelled node. Further, on every step, a node adjacent to the subnetwork with the highest *connectivity* was added to the growing subnetwork. Connectivity p-value of a neighbor node g was computed applying hypergeometric test on the observed numbers of nodes (i) connected with g and already included in the subnetwork, (ii) connected with g but not included in the subnetwork, (iii) not connected with g and included in the subnetwork, and (iv) neither connected with g nor included in the subnetwork. Fig. 4.5 illustrates three consequent iteration of the DIAMOnD algorithm.

2. Next, genes shared by multiple biclusters were assigned to the network, randomly choosing unlabelled nodes from the set of all unlabelled nodes connecting the desired biclusters. Finally, background genes were assigned to unlabelled nodes.

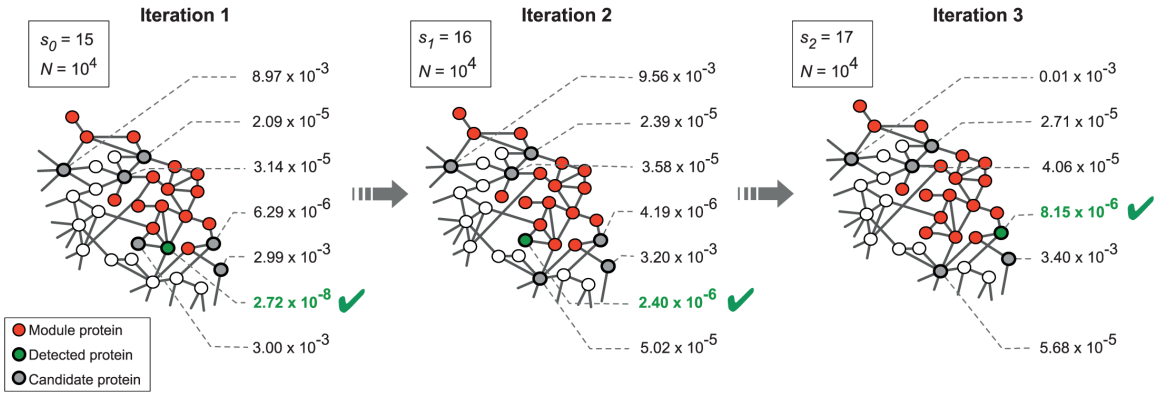

Fig. 4.5 Three iterations of the DIAMOnD algorithm. Red highlights the nodes already included in the subnetwork (e.g. disease module), grey shows candidate nodes. Green highlights the best candidate with the lowest connectivity p-value. N is the total number of nodes in the network, s is the number of genes included in the network at the corresponding iteration. From [89] with changes.

4.5.2 Obtaining and preprocessing of real data

Normalized gene expression profiles from TCGA and METABRIC cohorts were downloaded from cBioPortal [36] (http://www.cbioportal.org/). Only genes expressed in more than 5% of samples in the cohort were kept. The expressions of the remaining genes were log2-transformed and standardized. Samples from both cohorts were annotated with patient age at diagnosis, stage of the tumor, and molecular subtype. All clinical information was downloaded from cBioPortal and converted into the same format.

A human gene network derived from the BioGRID [256] published by Huand et al. [112] was used in this thesis. This network consisted of 258,257 interactions between 16,702 genes. BioGRID was chosen for this thesis because it is one of the most comprehensive and frequently updated gene interaction networks for *Homo sapiens*. It comprises curated genetic and protein interactions which are more reliable than composite networks, containing computationally predicted interactions. While it provides good coverage of human genes, BioGRID is not too dense and resembles the scale-free property, which characterizes many natural networks [12]. Although most of the edges in this network represent protein interactions, BioGRID still suits our problem because genes with interacting protein products are functionally related.

4.6 Experiments

4.6.1 Evaluation with synthetic data and the choice of parameters

Previous studies have shown the importance of appropriate parameter setting for method performance [76, 258]. Therefore each method was applied on all 20 simulated datasets multiple times with different combinations of parameter values to find an optimal, i.e. resulting in maximal performance on average. Table 4.2 reports parameter values tested for each method, which were chosen based on recommendations of method developers and the results of [76] and [258]. All possible combinations of these parameters listed in the table were tested and all other parameters were set to default. The expected number of biclusters was set to 10 when possible.

To compare the set of biclusters B_{pred} obtained in the result of each run with the ground truth set B_{true}, Relevance and Recovery scores proposed by Prelic et al. [227] were calculated:

$$Relevance(B_{pred}, B_{true}) = \frac{1}{|B_{pred}|} \sum_{B_p \in B_{pred}} \max_{B_t \in B_{true}} J(B_p, B_t) \tag{4.6}$$

method	**parameters tested**	**optimal parameters**
DeBi	b =[0.5, 0.75, **1.0**, 1.25, 1.5, 2.0, 2.5] o = [**0.5**, 0.75, 1.0]	b=1.5; o=0.5
ISA2	no.seeds = [2,5,10, 20, 30, ..., **100**, 125, 150, 200] thr.row* = **[0.5, 1.0, 1.5, 2.0]** thr.column* = **[0.5, 1.0, 1.5, 2.0]**	no.seeds=20
xMOTIFs	discr_levels = [2, 3, 5,10,15, 20, 30, 40, 50] α = [0.001, 0.01, **0.05**, 0.10, 0.15] ns = [5, **10**, 25, 50, 100] nd = [**10**,100,1000] sd = [**5**,10,20,100]	discr_levels=5; α=0.1; ns=100
Cheng & Church	α =[1.0, 1.1, 1.2, 1.3, 1.4, **1.5**] δ = [0.1,0.2, ... , **1.0**]	α=1.4; δ=1.0

Plaid	row.release = [0.5, 0.55, 0.60, 0.65, **0.7**]
	col.release = [0.5, 0.55, 0.60, 0.65, **0.7**]
	back_fit_values = [**0**,10,100]
	iter_startup_values = [**5**, 10,100,1000]
	iter_layer_values = [**10**,100,200,500,1000]
		row_release=0.5;
		col_release=0.55;
		back_fit=10;
		iter_startup=100;
		iter_layer=100

FABIA	α = [0.005, **0.01**, 0.05, 0.1, 0.5] spl = [**0**, 0.5, 0.75, 1.0, 1.5, 2.0] spz = [0, **0.5**, 0.75, 1.0, 1.5, 2.0]	α=0.05; spl=0.5; spz=0.75

COALESCE	prob_gene = [0.99, **0.95**, 0.9]
	pvalue_cond = [0.01, **0.05**, 0.1]
	pvalue_correl = [0.01, **0.05**, 0.1, 0.2]
	zscore_cond = [0.005, 0.01, **0.05**, 0.1]
		prob_gene=0.95;
		pvalue_cond=0.1;
		pvalue_correl=0.5;
		zscore_cond=0.05

QUBIC	r = [**1**, 2, 3, 4, 5, 7,10]
	q = [0.05,**0.06**, 0.1, 0.25, 0.5]
	c = [0.5, 0.65, 0.70, 0.75, 0.8, 0.85, 0.9, **0.95**]
	P = [TRUE, **FALSE**]
	with or **without** network
		r=1;
		q=0.25;
		c=0.65;
		P=FALSE

BiBit	max_n_bics=[**0**,10,15,100] pattern_bitsize=[8,**16**,32] max_discr_value=[0.5,**1.0**,1.5,2.0]	max_n_bics=100 pattern_bitsize=16 max_discr_value=1.0
DESMOND	$\alpha = [5.0, 1.0, 0.5, 0.1, 0.05]$ $\frac{\beta}{K} = [10000, 1.0, 0.0001]$ $p = [0.001, 0.005, 0.01, 0.05]$ $q = [0.25, 0.5, 0.75]$	$\alpha = 0.5$; $p = 0.01$; $q = 0.5$
DESMOND2	$\alpha = [10.0, 5.0, 2.5, 1.0, 0.5]$ $\frac{\beta}{K} = [10^8, 10^6, 10^5, 10^4, 10^3, 10^2, 10, 10^-4]$ $SNR_{min} = [0.5, 0.75, 1.0]$	$\alpha = 5.0$; $\frac{\beta}{K} = 10^4$; $SNR_{min} = 0.75$

Table 4.2 The results of hyperparameter tuning on ten synthetic datasets. In case of ties (e.g. for DeBi and BiBit), parameter combination closer to the default is reported. Default parameter values are highlighted by bold text font. ISA2 accepts several row and columns thresholds column thresholds and automatically determines the most appropriate combination of parameters marked by (*).

$$Recovery(B_{pred}, B_{true}) = \frac{1}{|B_{true}|} \sum_{B_t \in B_{true}} \max_{B_p \in B_{pred}} J(B_p, B_t) \tag{4.7}$$

where $J(B_p, B_t)$ denotes Jaccard similarity of two biclusters $B_p(G'_p, S'_p)$ and $B_t(G'_t, S'_t)$:

$$J(B_p, B_t) = \frac{|B_p \cap B_t|}{|B_p \cup B_t|} = \frac{|G'_p \cap G'_t| \times |S'_p \cap S'_t|}{|G'_p \cup G'_t| \times |S'_p \cup S'_t|} \tag{4.8}$$

As seen from the above equations, relevance and recovery scores are very similar and aimed at the quantification of type I and type II errors. Relevance score reflects how well the predicted biclusters match with the biclusters from the ground truth set. It is high if, for each of the predicted biclusters, there is the best match among the true biclusters with a strong overlap. However, relevance score will be still high, in the case when all predicted biclusters match some but not all of the true biclusters. Recovery score shows to what extent true biclusters are recovered by the set of predicted biclusters and becomes high when each of true biclusters strongly with at least one of the predicted. Since none of the error types takes precedence over the other, the overall performance score combines Relevance and Recovery taking their geometric mean:

$$Performance(B_{pred}, B_{true}) = \sqrt{Relevance(B_{pred}, B_{true}) \times Recovery(B_{pred}, B_{true})} \tag{4.9}$$

A combination of parameters was considered optimal if it resulted in the highest Performance score, averaged over all 20 synthetic datasets. For non-deterministic methods, average Performance scores in 10 runs were compared, since their results vary from run to run.

4.6.2 Evaluation with breast cancer data

DESMOND and baseline methods were evaluated on the data collected in two large breast cancer studies, TCGA-BRCA [164] and METABRIC [213]. In METABRIC cohort, all 1904 gene expression profiles were measured by microarray technology. TCGA-BRCA data comprised of two datasets: 1081 expression profiles were measured by RNA-Seq (TCGA-RNAseq) and 529 by microarrays (TCGA-micro). TCGA-micro and TCGA-RNAseq cohorts were not independent: 517 expression profiles were obtained from the same samples. Since microarray and RNA-seq platforms employ different technologies to estimate gene expression levels, their measurements in the same samples and genes may differ [232]. This may affect

the results of biclustering, therefore biclustering was performed independently on TCGA-RNAseq and TCGA-micro datasets.

For evaluation purposes, only 11959 genes presented in all three expression datasets and in the BioGRID network were kept. Also, the nodes corresponding to genes absent in expression profiles and their adjacent edges were removed from the network before using it (179514 edges remained).

The discovered biclusters were tested for associations with Gene Ontology [57] (GO) terms and known cancer subtypes using the one-sided exact Fisher's test, to evaluate their biological significance. All gene sets used in this thesis were downloaded from the EnrichR [143] website (http://amp.pharm.mssm.edu/Enrichr/). Overall survival (OS) analysis was performed using the Cox proportional hazards model implemented in Lifelines v0.23.0 [67] with age at diagnosis and stage as covariates. All other statistical tests were performed in python using Scipy 1.1.0. Benjamini and Hochberg's procedure implemented in the gseapy 0.9.9 [41] python library was applied for multiple testing correction.

4.7 Code Availability

The latest version of DESMOND is implemented in python 3.8 and available at https://github.com/ozolotareva/DESMOND/. This repository also provides the code for the generation of artificial expression and network data and for the preprocessing of real data. Both versions of DESMOND used in this thesis were implemented in python 2.7.15 and remain available at https://github.com/ozolotareva/DESMOND/DESMOND_py2.

R package *biclust* (https://cran.r-project.org/web/packages/biclust/) version 2.0.1 was used to run the method by Cheng & Church, xMOTIFs, QUBIC, Plaid, and FABIA. ISA2 was available as a separate R package of the same name (https://cran.r-project.org/web/packages/isa2/, version 0.3.5). DeBi, COALESCE, and BiBit were run via *JbiclustGE* wrapper (https://jbiclustge.github.io/).

Chapter 5

Results and Discussion

This chapter presents and discusses the results of experiments introduced in section 4.6. Parts of the results concerning the first version of DESMOND and five baseline methods are originally published in *Bioinformatics (Oxford University Press)* [298].

5.1 Evaluation on synthetic data

Method performances varied widely among tools and different bicluster shapes (Fig. 5.1). Almost all methods benefited from parameter optimization. DeBi, FABIA, COALESCE, and QUBIC greatly improved their average performances (Fig. 5.2, Table 5.1).

Interestingly, classic and query-based versions of QUBIC demonstrated similar performances, despite the fact that the later took into account network information. Although no method outperformed others in all cases, COALESCE had the best overall performance in this benchmark (on average, 0.63 (the third) with default and 0.72 (the first) with tuned parameters). DESMOND and DESMOND2 were the second and the fourth top-performing methods with an average performance of 0.64 and 0.58.

DESMOND outperformed all other methods for biclusters of sizes 100x100, 50x100 and with $\alpha = 0.5$, RRHO p-value threshold $p = 0.01$ and $q = 0.5$. The first version of DESMOND was not sensitive to changes of β/K (Wilcoxon signed-rank test p-values 0.11 and 0.67 for comparison of performances obtained with β/K set to 1.0 versus 10^4 and 10^{-4} and other parameters fixed) and therefore β/K was set to 1.

DESMOND2 demonstrated the best performance with $\alpha = 5.0$, $\beta/K = 10^4$ and $SNR_{min} = 0.75$ and beaten all the other methods on biclusters of shapes 5x50 and 5x100.

Both versions of DESMOND did not perform well on biclusters with a small number of samples. When considering only datasets with biclusters of 20 or more samples, DESMOND

on average outperforms all methods including COALESCE. Therefore, given that both versions of DESMOND could not accurately detect the biclusters small in terms of samples, we set s_{min} to 10% of the whole cohort size in all subsequent experiments.

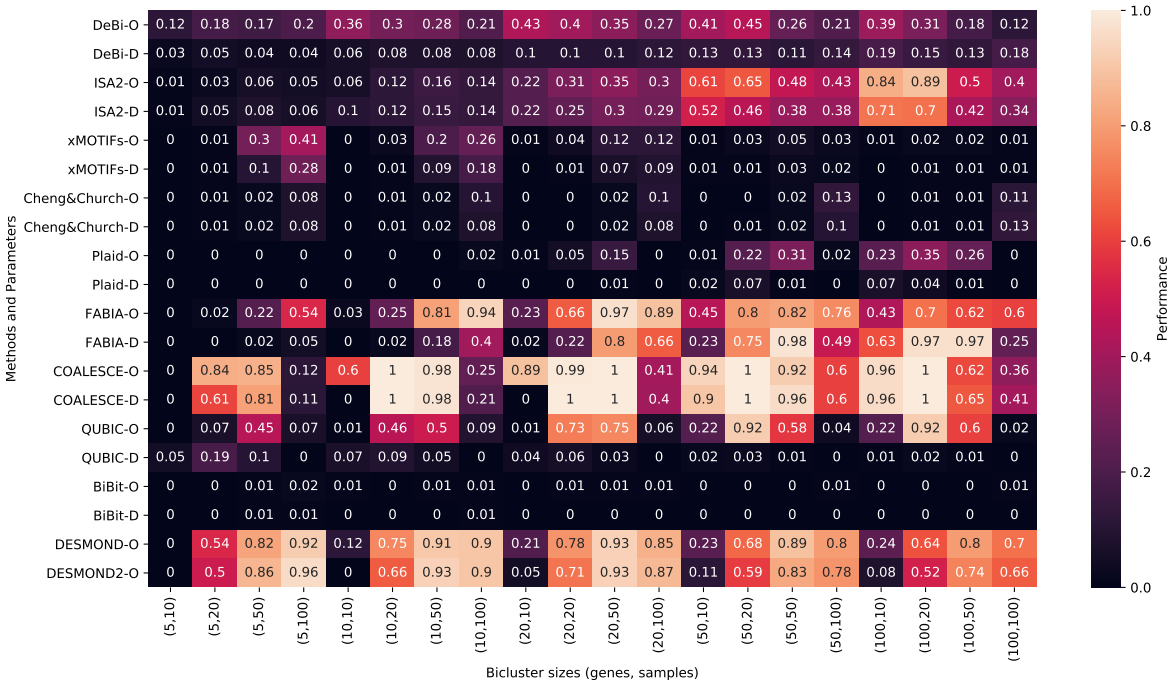

Fig. 5.1 Performance scores demonstrated by DESMOND and baseline methods on 20 synthetic datasets containing biclusters of different shapes. For non-deterministic methods, average performance in 10 runs is reported. For each of baselines, performance scores for default (D) and optimized (O) parameters are reported.

	with optimized parameters				with default parameters			
	Relevance	**Recovery**	**Performance**	**Biclusters on average**	**Relevance**	**Recovery**	**Performance**	**Biclusters on average**
DeBi	0.214	0.374	0.28	51.5	0.026	***0.426***	0.102	614.0
ISA2	0.222	0.522	0.33	58.9	0.147	***0.586***	0.284	203.7
xMOTIFs	0.088	0.08	0.084	9.1	0.051	0.046	0.084	10.0
Cheng & Church	0.032	0.034	0.033	4.9	0.028	0.032	0.03	4.8
Plaid	0.148	0.052	0.082	1.5	0.028	0.005	0.011	0.4
FABIA	0.536	0.539	0.537	10.2	0.38	0.383	0.382	10.0
COALESCE	**0.722**	**0.720**	**0.716**	11.0	0.633	0.63	0.629	10.5
QUBIC	0.351	0.322	0.336	10.0	0.05	0.031	0.039	10
BiBit	0.004	0.011	0.006	100	NA	NA	NA	NA*
DESMOND	0.616	0.669	0.635	8.2	NA	NA	NA	NA
DESMOND2	0.628	0.557	0.584	7.1	NA	NA	NA	NA

Table 5.1 Average Relevance, Recovery, Performance, and the number of reported biclusters computed for the results obtained by each method on synthetic data with the default and optimized parameters. COALESCE with optimized parameters demonstrated the highest Relevance, Recovery, and Performance. For ISA and DeBi, parameter combination resulting in the highest performance had decreased Recovery, compared to default parameters.

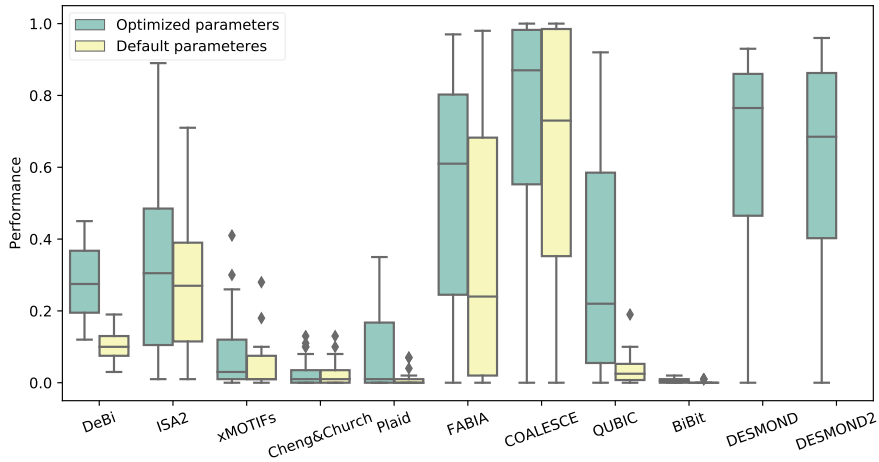

Fig. 5.2 Average performance scores demonstrated by DESMOND, DESMOND2 and nine baseline methods on 20 synthetic datasets with the default and optimal parameters.

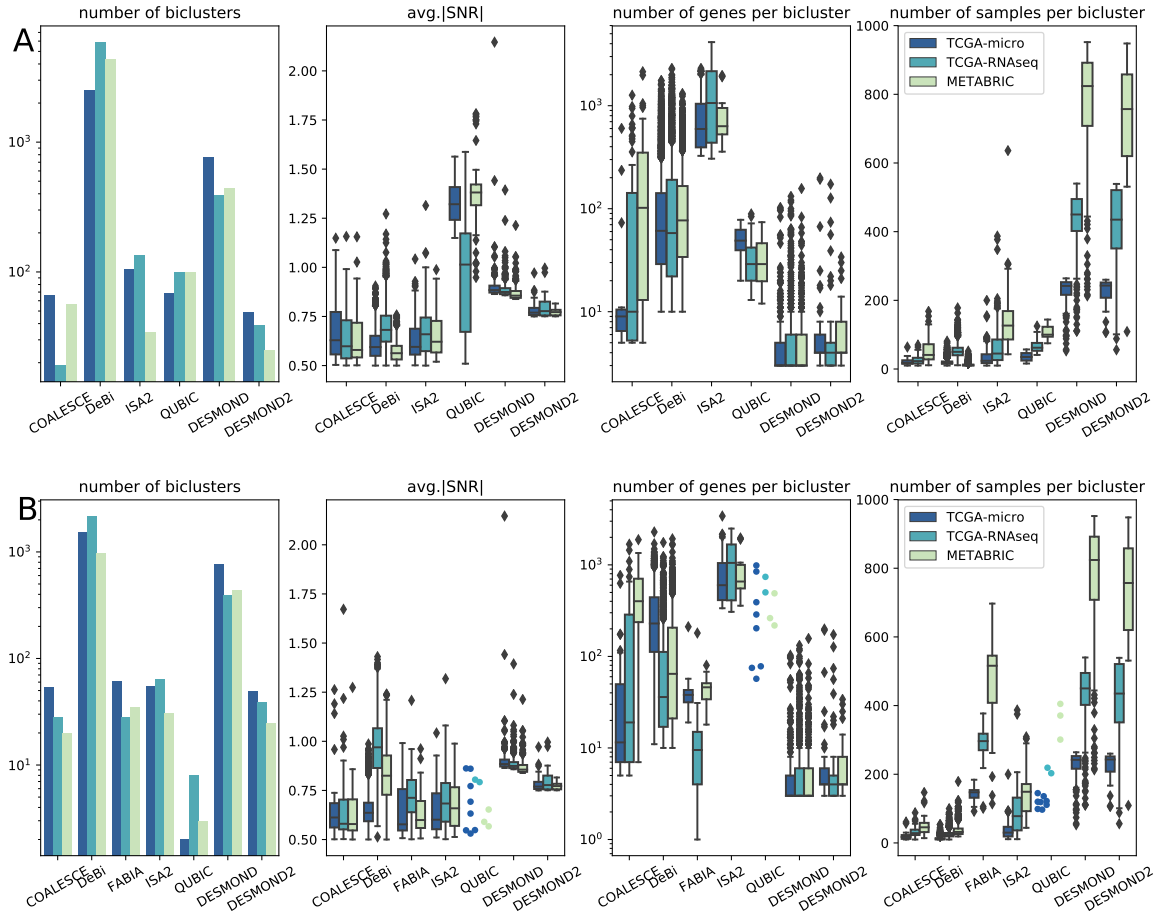

Fig. 5.3 Characteristics of differentially expressed biclusters produced by DESMOND, DESMOND2 and baseline methods on TCGA and METABRIC data with default (A) and optimized (B) parameters. Since QUBIC produced less than 10 biclusters on all real datasets with optimized parameters, its results are represented by dots instead of boxplots.

5.2 Evaluation on real breast cancer data

Five baselines (COALESCE, DeBi, ISA, FABIA, and QUBIC) demonstrated their ability to detect differentially expressed biclusters in synthetic data were chosen and applied on three real-world datasets: TCGA-micro, TCGA-RNAseq, and METABRIC. Each method was run twice: with default parameters and with parameters optimized on synthetic data. Since we were interested in differentially expressed biclusters, we excluded from further analyses all biclusters with $avg.|SNR| < 0.5$ (this corresponds to SNR between $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$) and less than 2 genes or 10 samples.

All methods produced different numbers of biclusters, demonstrating diverse distributions of bicluster shapes and $avg.|SNR|$ values (Fig. 5.3). FABIA run with default parameters

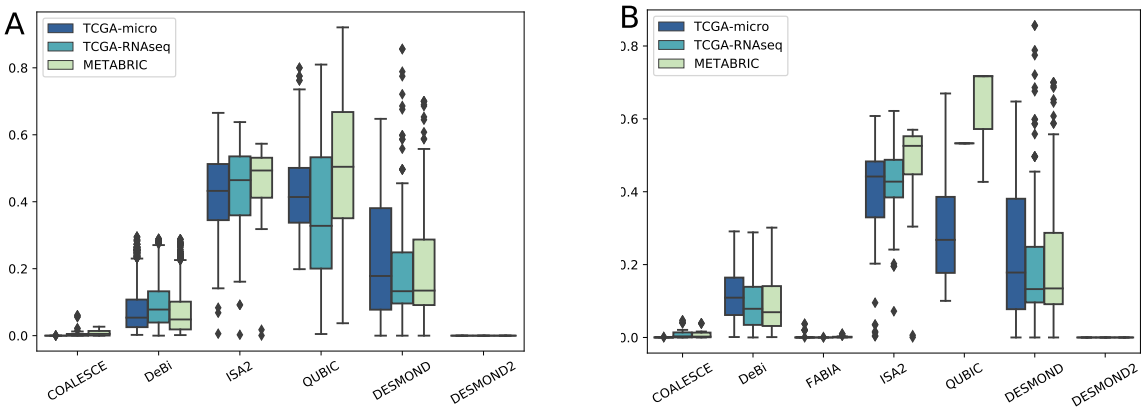

Fig. 5.4 Distributions of bicluster redundancies computed for the output of each method.

identified no biclusters with average $|SNR|$ above 0.5. DeBi did not finish after a week of running with default parameters on the METABRIC dataset and therefore was run on the subset of 500 randomly chosen samples. In contrast, QUBIC identified biclusters with weaker differential expressions, when running with optimized parameters than with defaults. Only 8, 2 and 3 biclusters found with optimized parameters in TCGA-micro, TCGA-RNAseq, and METABRIC respectively passed SNR threshold of 0.5. Given that the effect of parameter tuning was controversial, the results obtained with default and optimized parameters are reported here and below.

DESMOND identified 390, 763, and 442 biclusters in TCGA-micro, TCGA-RNAseq, and METABRIC respectively. DESMOND2 found much fewer biclusters than the first version: 39, 49, and 25 in TCGA-micro, TCGA-RNAseq, and METABRIC datasets. This difference is explained by the higher redundancy of the biclusters found by DESMOND Fig. 5.4. Redundancy of each individual bicluster B_p from a set of biclusters B_{pred} was calculated as Jaccard similarity of B and its best match from B_{pred}. Since DESMOND clusters edges, it returns many biclusters overlapping in genes and samples. The second version of DESMOND clusters genes, and therefore outputs a small number of non-redundant biclusters.

Biclusters produced by both versions of DESMOND tended to be smaller in terms of genes and bigger in terms of samples than biclusters found by other methods. DESMOND, DESMOND2, QUBIC with default and DeBi with optimized parameters identified biclusters with more pronounced differential expression, compared to the other methods.

In contrast with the synthetic data benchmark, no ground truth was available for real-world breast cancer datasets. Therefore, to evaluate the results of all methods, gene and sample sets defined by the produced biclusters were further tested for biological significance.

5.2.1 Associations with GO terms

To demonstrate the identified biclusters are composed of functionally coherent genes, the obtained gene sets were tested for overlap with known functionally related gene sets from Gene Ontology (GO) and pathways from KEGG. Most of the biclusters identified by DESMOND and DESMOND2 were significantly enriched with at least one GO term. Owing to network constraints, the proportion of DESMOND biclusters significantly overlapping with KEGG pathways was higher than for all other methods. The proportion of GO-enriched DESMOND biclusters was also high, although it was slightly lower than the proportion of significant biclusters found by ISA2 on TCGA-RNAseq and by QUBIC on METABRIC (Fig. 5.5).

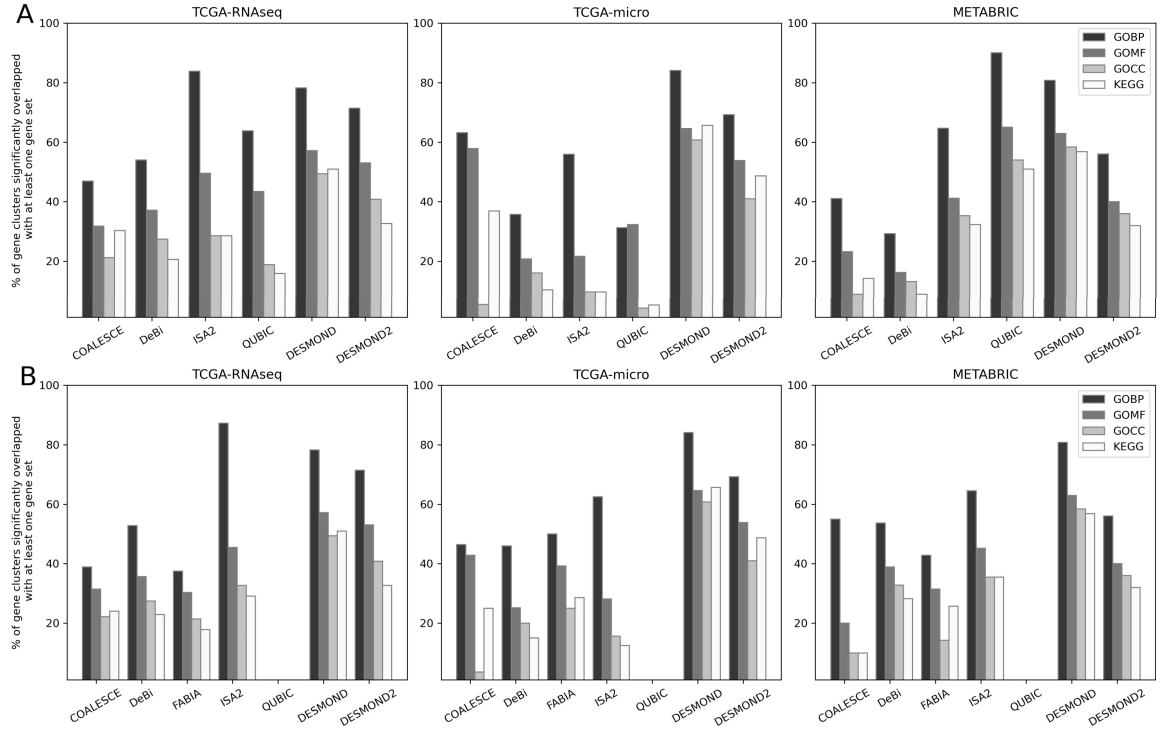

Fig. 5.5 Percent of gene clusters significantly (BH-adjusted p-value<0.05) overlapping with at least one functionally related gene set from GO Biological Process (GOBP), GO Molecular Function (GOMF), GO Cellular Component (GOCC) and KEGG pathways. Results obtained with default (A) and tuned (B) parameters. Only overlaps including more than one gene were taking into account.

	GOBP		GOCC		GOMF		KEGG	
	DESMOND	***DESMOND2***	***DESMOND***	***DESMOND2***	***DESMOND***	***DESMOND2***	***DESMOND***	***DESMOND2***
TCGA-RNAseq	0.78 (0.56±0.02)	0.73 (0.62±0.06)	0.49 (0.27±0.01)	0.39 (0.3±0.07)	0.55 (0.34±0.02)	0.53 (0.39±0.07)	0.5 (0.24±0.02)	0.33 (0.28±0.06)
TCGA-micro	0.84 (0.55±0.03)	0.69 (0.65±0.07)	0.61 (0.26±0.02)	0.41 (0.33±0.07)	0.63 (0.33±0.03)	0.54 (0.42±0.07)	0.65 (0.23±0.02)	0.49 (0.3±0.07)
METABRIC	0.8 (0.57±0.02)	0.56 (0.63±0.08)	0.58 (0.27±0.02)	0.36 (0.3±0.08)	0.62 (0.35±0.02)	0.4 (0.39±0.09)	0.56 (0.24±0.02)	0.32 (0.28±0.08)

Table 5.2 The proportion biclusters found by DESMOND and DESMOND 2 significantly overlapping with at least one gene set from GO. Values in brackets represent the mean and standard deviation for the proportion of randomly chosen subnetworks, significantly overlapping with any GO gene set.

To prove that DESMOND performance in this test was superior not only due to network constraints, we generated 100 sets of random subnetworks of the same sizes as DESMOND biclusters. Percent of enriched gene sets was always much higher for DESMOND biclusters than for any random set of subnetworks (empirical p-value<0.01) (Table 5.2).

For DESMOND2, the percentage of enriched biclusters was lower than for biclusters produced by the first version. It was almost always the second or the third. However, in half of all tests, the percentage of enriched DESMOND2 biclsuters was not significantly higher than the null model.

Surprisingly, the percentage of GO-enriched biclusters in QUBIC with default parameters was significantly higher from random subnetworks only on the METABRIC dataset (all p-values are < 0.01). In contrast, its results on TCGA-RNAseq and TCGA-micro contained even less GO-enriched biclusters, than expected by chance.

It is important to add, that gene set libraries of EnrichR [143] used in this thesis include only GO terms of level four or higher. Less specific GO terms were removed from these databases. Using the whole GO database would give much higher percentage of GO term-associated biclusters in all cases.

5.2.2 Reproducibility of found biclusters

Yet another way to prove that the methods identify similar biclusters would be demonstrating that their findings reproduce on independent datasets with the same biology. To check, whether the methods identify the same biclusters on different datasets, two questions were formulated:

- Do the methods identify biclusters composed of the same genes in different datasets?
- How similar are biclusters found in TCGA-RNAseq and TCGA-micro datasets, which share 517 patients?

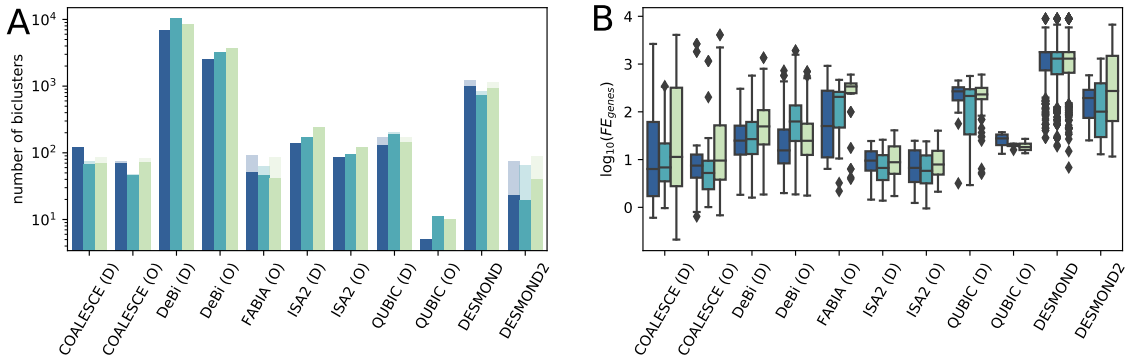

Fig. 5.6 Gene similarities of biclusters found in different breast cancer datasets. (A) The total number of matched pairs of biclusters. The transparent part of the bar shows biclusters for which no best match was found. (B) Distributions of log-transformed fold-enrichments of Jaccard similarity, computed for best matches.

To answer the first question, pairwise comparisons of biclusterings obtained on TCGA-micro, TCGA-RNAseq, and METABRIC have been done. For every pair of biclusterings, non-reciprocal best matches in genes were identified. A bicluster was marked as "unmatched", if no bicluster from the target set shared any gene with it. In the result of matching of two biclusterings B_1 and B_2 maximal of $|B_1| + |B_2|$ pairs of biclusters were established. Since the size of biclusters in genes varied greatly, instead of Jaccard similarities of best matches, the ratios of observed to expected overlaps were compared.

Figure 5.6 shows the number of unmatched biclusters in every comparison (A), and the distributions of log-fold enrichment computed for observed Jaccard similarities compared to the expected given bicluster sizes (B). DESMOND2, FABIA, and DESMOND produce a certain proportion of unmatched biclusters, but the matched biclusters tend to overlap stronger than matched biclusters found by other methods. Other methods produce no or nearly no unmatched biclusters, but fold enrichments for the Jaccard similarities were lower.

On one hand, the presence of unmatched biclusters may point to the high rate of false findings and therefore a poor agreement between the results. DESMOND2 and (at lesser extent) FABIA produce a much higher fraction of unmatched biclusters than the other methods and what raises concerns about the reliability of their results. On the other hand, the unmatched biclusters may reflect the existing variation between datasets.

In the above experiment, biclusters were compared in genes, but not in samples. Given that TCGA-RNAseq and TCGA-micro datasets share 517 patients, the biclusters may be also compared considering both genes and samples. In this experiment, the same approach

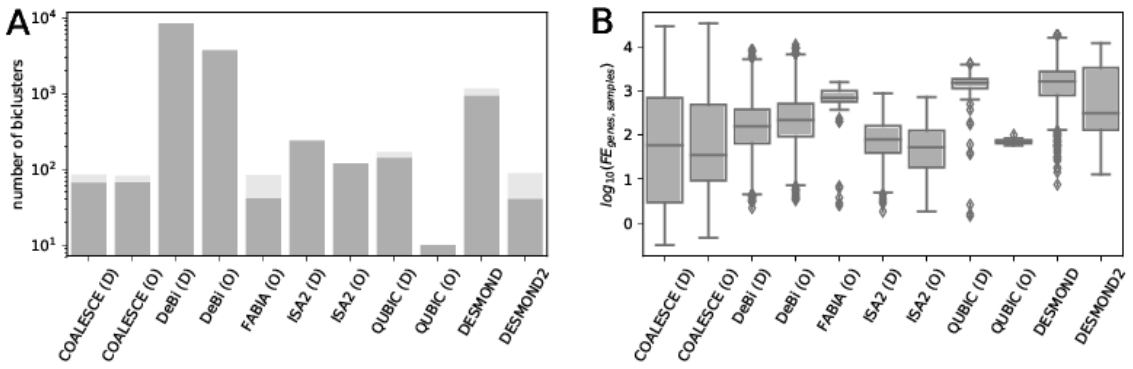

Fig. 5.7 Similarities of biclusters found in TCGA-BRCA datasets profiled by RNA-seq and microarrays, computed considering genes and samples. (A) Total number of biclusters tested. Transparent part of the bar represents biclusters without any best match. (B) Distributions of log-transformed fold-enrichments computed for best matches.

is used, but best matches were identified based on maximal Jaccard similarity in genes and samples.

DESMOND, FABIA with optimized parameters, and QUBIC with default, demonstrated the highest gain of the observed overlap size compared to random overlap (Figure 5.7). DESMOND2 on average, found less similar biclusters on TCGA-micro and TCGA-RNAseq datasets. Again, this may be explained by either lower robustness of DESMOND2 predictions, or its ability to detect the true platform-specific variation.

5.2.3 Associations with clinical variables

Breast cancer subtypes

All methods were able to identify many biclusters, significantly (BH-adjusted hypergeometric p-value<0.05) enriched by samples annotated with known breast cancer subtypes (Fig. 5.8). However, although many biclusters were significantly associated with one or several subtypes, only a few of them demonstrated a strong overlap with the associated subtype in terms of the Jaccard similarity. All the methods except QUBIC with optimized parameters found biclusters overlapping Luminal A (LumA) subtype in TCGA (Jaccard similarities about 0.5-0.9). ISA2 and both versions of DESMOND found biclusters strongly overlapping with Basal subtype in TCGA datasets (Jaccard similarity above 0.8). ISA2 applied with default but not with optimized parameters identified biclusters strongly (Jaccard similarity about 0.5) overlapping with Her-2 subtype in TCGA. For all other subtypes in TCGA and all subtypes

in METABRIC, overlaps with the most significantly enriched biclusters were even weaker. DESMOND managed to find biclusters with stronger overlaps with LumA, LumB, and Basal subtypes in TCGA than its competitors. Although DESMOND2 found much fewer biclusters than DESMOND, its most strongly overlapping biclusters had the same or just a bit smaller overlap with LumA, LumB, and Basal subtypes.

Almost all biclusters found by DESMOND were associated with at least one molecular subtype of breast cancer. Only 0.5-3.3% of DESMOND biclusters showed no significant over- or under-representation of any molecular subtype. In contrast, COALESCE and DeBi produced larger fractions of biclusters not associated with any subtype, up to 68% and 91% of all reported biclusters.

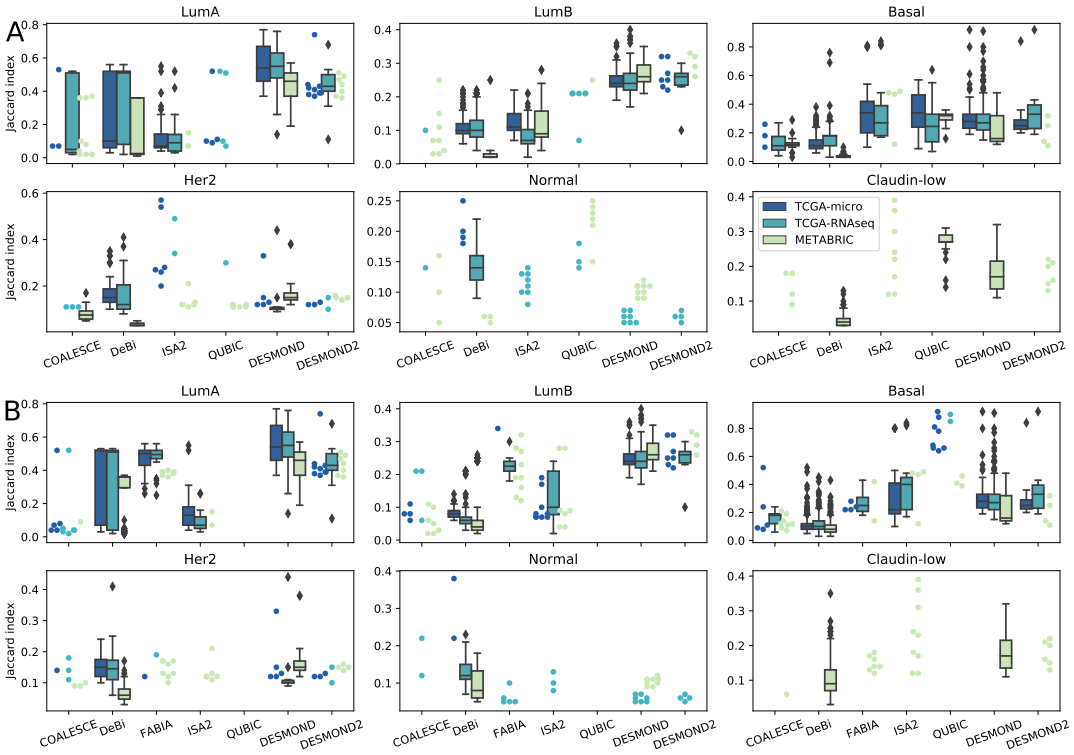

Fig. 5.8 Distributions of Jaccard similarities of known breast cancer subtypes and sample sets defined by biclusters produced by each method. For each bicluster, over- and under-representation of each subtype was evaluated using the hypergeometric test. Each bicluster was annotated with the subtype based on a minimal adjusted p-value passing threshold of 0.05. The results obtained with default parameters and with parameters optimized on synthetic data are shown in figures A and B respectively. When the group contains less than 10 biclusters, the results are shown as dots instead of a boxplot. Claudin-low subtype was annotated only in METABRIC dataset and therefore biclusters found in TCGA data sets were not tested for overlap with this subtype.

Overall survival

All identified biclusters were further tested for association with overall survival (OS) using Cox proportional hazards model. DESMOND detected 47 and 96 OS-associated biclusters in TCGA-RNAseq and METABRIC. It produced more biclusters significantly associated with overall survival on TCGA-RNAse and METABRIC datasets compared to the other methods (Figures 5.9). DeBi was the only method managed to identify any OS-associated biclusters in TCGA-micro data. It also found 37 and 11 biclusters in TCGA-RNAseq and METABRIC with default parameters and 24 and 5 with optimized. However, the similarity between OS-associated biclusters found by DeBi on TCGA-micro and TCGA-RNAseq was not high: pairs of biclusters with the strongest overlap in genes never shared more than two samples.

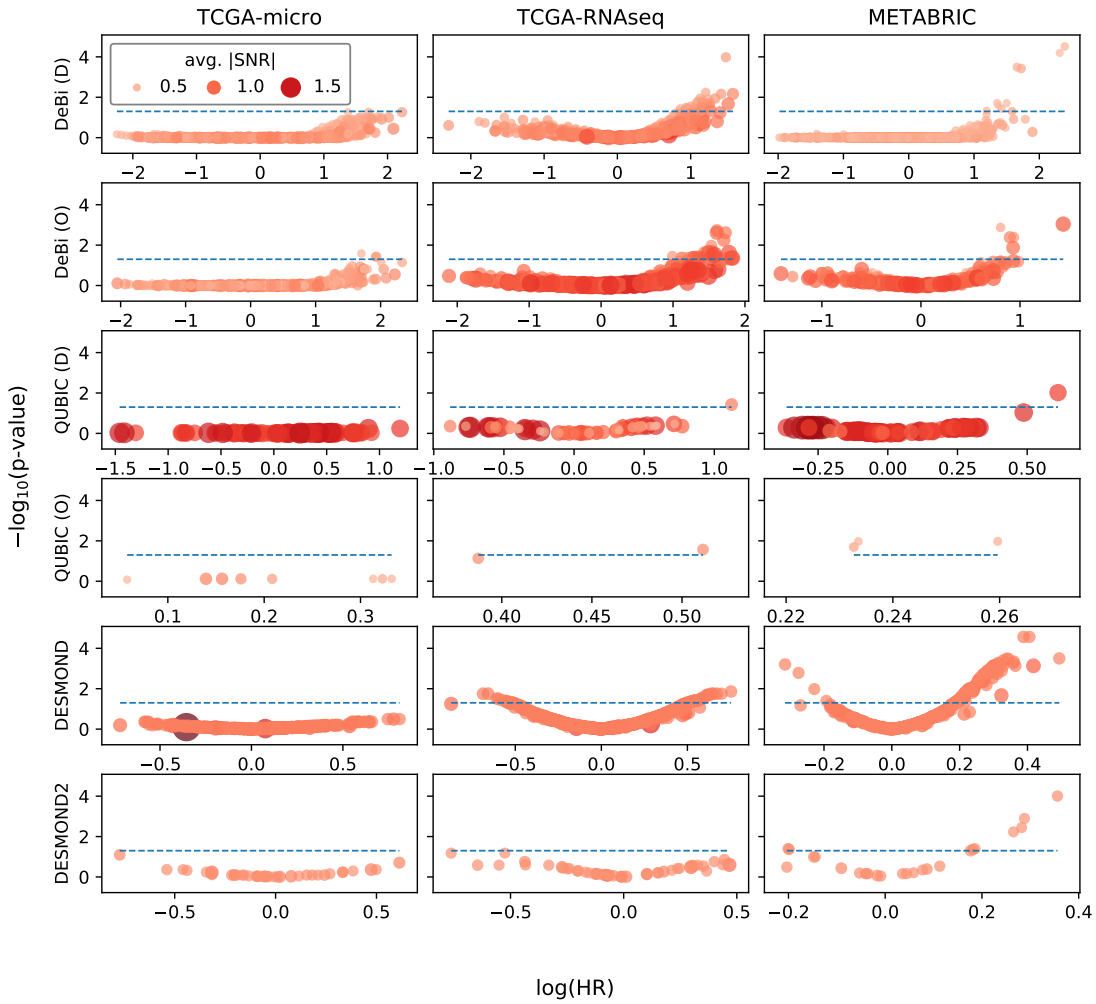

Fig. 5.9 Association of biclusters found by DeBi, QUBIC, DESMOND, and DESMOND2 with overall survival. Every circle represents a bicluster, with size and color intensity proportional to $avg.|SNR|$. The X and Y axes show a negative logarithm of adjusted p-values and coefficients (logarithm of Hazard Ratio) of Cox regression models fitted for patient sets defined by biclusters. The best biomarkers have higher $avg.|SNR|$ and larger positive or negative regression coefficients.

Of all methods, only DESMOND, DeBi and QUBIC identified OS-associated biclusters in both TCGA-RNAseq and METABRIC. OS-associated biclusters found by each method in these two cohorts were tested for similarity in genes.

DeBi and DESMOND identified multiple OS-associated biclusters in TCGA-RNAseq and METABRIC. Although DeBi identified biclusters with higher HR than DESMOND, the latter produced more similar OS-associated biclusters in TCGA and METABRIC. To

demonstrate this, for each bicluster found in one dataset, its best match in another was identified based on the maximum Jaccard similarity of their gene sets. Distributions of Jaccard similarities for all pairs of best matches are shown in Fig. 5.10.

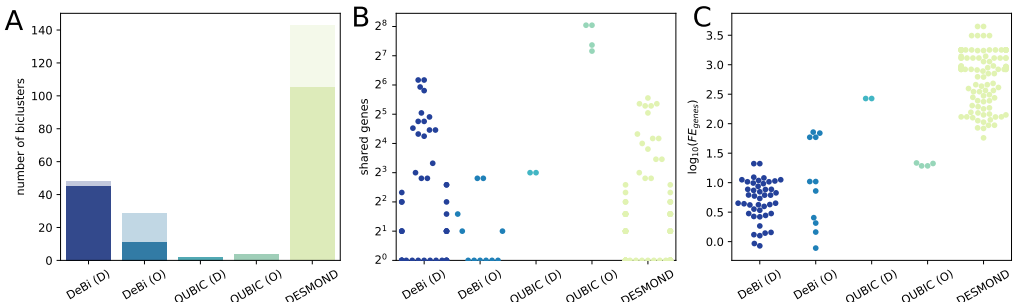

Fig. 5.10 A. The number of OS-associated biclusters tested. Transparent part of each bar corresponds to unmatched biclusters. B. The number of genes shared between the best matches in genes between OS-associated biclusters found in TCGA-RNAseq and METABRIC. C. Logarithms of observed Jaccard similarities divided by expected Jaccard similarities.

QUBIC applied with default parameters found one at each bicluster in TCGA-RNAseq and METABRIC. QUBIC found several isolated OS-associated biclusters in TCGA-RNAseq and METABRIC (one at each with default settings and one and three with optimized). All of the OS-associated biclusters found by QUBIC strongly overlapped each other in genes and samples and were associated with the Basal subtype. QUBIC and DESMOND, thus identify different but reproducible biclusters between TCGA and METABRIC. Such biclusters might be promising biomarker candidates and potentially define uncharacterized subgroups within known subtypes of breast cancer.

Nine of 25 biclusters found by DESMOND2 in METABRIC were significantly OS-associated. Unfortunately, in two other datasets, none of the biclusters found were significantly associated with OS. COALESCE and FABIA found only isolated biclusters either in TCGA-RNAseq or METABRIC.

The better reproducibility of OS-associated biclusters found by DESMOND and QUBIC may be explained by the network constraints applied to the modules. Higher stability is desirable for the discovery of gene signatures reproducible in independent studies, regardless of the expression profiling method used.

Chapter 6

Conclusions

This thesis is devoted to the development of a novel method for network-constrained biclustering of gene expressions. The new method called DESMOND is aimed at the detection of the differentially expressed gene modules – connected groups of genes up- or down-regulated in unknown subgroups of samples. The formulated problem has a great significance for biomedical research, in particular, for the identification of previously unknown disease subtypes and subtype-specific biomarkers.

This thesis presents two versions of DESMOND, which differ in a way they determine and represent differentially expressed genes. Two factors distinguish DESMOND from most biclustering methods: (i) it searches for differentially expressed biclusters, rather than biclusters with co-expression, (ii) it performs a network-constrained search when the majority of biclustering methods are unconstrained. Both versions of DESMOND were applied to synthetic and real-world datasets. Their performances were compared with state-of-the-art biclustering methods.

Another contribution of this thesis besides the development of new methods is the creation of a synthetic dataset with differentially expressed network-constrained biclusters. In contrast to the previous benchmarks, in this thesis biclusters were modeled with less prominent, but more realistic differential expression and had more diverse shapes.

The experiment results demonstrated the capability of all evaluated methods to identify biclusters representing biologically meaningful subsets of genes and samples. Interestingly, all methods produced very diverse biclusters. None of the tested methods outperformed all others in all experiments. DESMOND was on average the second of the best performing on synthetic datasets and was inferior only to COALESCE. However, the advantage of COALESCE over the other methods was not confirmed in experiments on real data. DESMOND,

in turn, tended to produce more GO-enriched gene clusters on the breast cancer datasets than the competitors, owing to its ability to consider gene interactions.

Yet another important outcome of this thesis is the identification of several OS-associated biclusters in TCGA and METABRIC, which were similar in genes. This may point to the presence of new molecular subtypes, characterized by differential expression of these genes. Replication of such expression patterns in independent cohorts confirms that they are less likely to be false findings. Although all such biclusters demonstrated the same OS with known molecular subtypes and significantly overlapped with them, they did not match well. Instead, replicated biclusters represented distinguishable subgroups within known subtypes, suggesting the presence of molecular heterogeneity within known PAM50 subtypes. These promising biomarker candidates are subject to further investigation, validation, and evaluation of clinical significance.

The main disadvantage of DESMOND is its running time. Motivated by the necessity to simplify the method and reduce runtime, the second version called DESMOND2 has been developed. DESMOND2 was much faster than the first version, but demonstrated inferior performance compared to DESMOND and therefore needs to be improved.

It is important to note that this thesis has several limitations, discussed in this chapter below. These limitations can be split into two groups: limitations of the methods and limitations of the experimental design. Addressing these limitations highlights the direction of future research.

6.1 Limitations

6.1.1 Limitations of the methods

The main weakness of the DESMOND algorithm is its high computational complexity, which results in a long runtime when input is large. It took more than a day for DESMOND to process the largest dataset in this study, comprising almost 2,000 samples and more than 13,000 genes. QUBIC, which also considers gene interactions, processes the same data in hours. Although runtime demonstrated by DEMOND is comparable to some other methods, like DeBi (unconstrained) or cMonkey2 (network-constrained), its reduction remains one of the main priorities for future development.

Similar to some other biclustering methods, DESMOND produced many modules overlapping in their gene sets. This happens because DESMOND clusters pairs of interacting genes, and tends to produce strongly overlapping but different gene clusters from densely

connected regions of the network. To partially address this issue, DESMOND merges strongly overlapping modules in the post-processing step. However, further reduction of the redundancy between modules in the first steps of DESMOND remains a direction for future development.

DESMOND2 clusters genes instead of gene pairs, and therefore the resulting biclusters never overlap in genes. On the one hand, obtaining a lesser amount of non-redundant biclusters is advantageous, because it simplifies their downstream analysis and interpretation. On the other hand, biclusters overlapping in genes may be not rare in real-world data. On the contrary, some genes participate in multiple biological processes and may be dysregulated under various conditions. Therefore, changing the algorithm in a way that each gene may be assigned to multiple biclusters on the second phase may be advantageous. For example, instead of assigning a gene to the most probable module, the gene can be assigned to any module it visited in the sampling phase. Since not many genes oscillate after the burn-in, the redundancy of the resulting biclusters will not be high.

Unfortunately, compared to the first version, DESMOND2 tends to demonstrate lower performance on real and synthetic data. There are at least three possible reasons for this performance decline:

- As was already mentioned above, real biclusters indeed overlap in genes, and biclusters non-overlapping at all might be not realistic.
- The network may become a too strict constraint when single genes are clustered instead of gene pairs
- The current version of DESMOND2 models gene expression as a mixture of Gaussians, when expression distributions may be better described by a mixture of heavy-tailed distributions.

6.1.2 Limitations of the experimental design

Three important limitations of the proposed experimental design should be noted:

- First, synthetic expression and network data created in this thesis still do not reflect all the aspects of real-world data. For example, we did not model correlations between background genes, although in reality multiple nested co-expression modules present in the data [148]. This simplified the task for methods using gene correlations for the initialization of biclusters, e.g COALESCE. Besides that, the effect of noise and variation in the level of differential expression was not investigated.

- Second, the effect of the gene network on the results of biclustering was not investigated. In this thesis, both versions of the method were tested only with synthetic networks and the BioGRID network. The method, however, may not perform well on a regulatory network, in which co-regulated genes are not connected directly. Also, the network should not be too dense, e.g. like composite functional networks. If dysregulated genes already form a connected component, adding more edges to this component would only increase runtime. On the other hand, the network should not be too sparse, otherwise many biclusters may be lost due to the network constraint.

- Third, the methods were tested only on the expression profiles of breast tumors. Although no reason to think that the methods will not perform well on the data from a different biological context, this must be checked experimentally. In the future, I am going to apply biclustering on data from the other cancer type (e.g. prostate) and for the search of drug response biomarkers in cell line expression profiles. Yet another intriguing experiment in the context of cancer would be testing biclusters discovered in expression data for association with genomic alterations (e.g. SNA, CNA) and drug response. Besides cancer, DESMOND is suitable for any other heterogeneous disease or phenotype. For example, as a follow-up of the project on the comorbidity of asthma and hypertension, DESMOND can be applied on expression data relevant for these diseases and disease-specific networks described in section 2.8. However, for this experiment appropriate expression datasets are necessary.

6.2 Future Work

Besides the improvements of DESMOND method proposed above, the conclusion of this thesis also highlights some thoughts about possible directions of development of the biclustering field. At the end, this section summarizes the ideas applicable to a broader list of computational approaches for investigation of complex and heterogeneous disease on the example of gene prioritization methods.

6.2.1 Development of biclustering

Since 1972 and to date, biclustering remains a developing field and one or several new methods get published every year. However, biclustering seems to be less popular than conventional clustering among the researchers whose major task is the application of existing tools rather than the development of novel methods. This may be partly explained by the

fact that the majority of works presenting biclustering methods do not discuss the connection between the patterns they aimed (e.g. shift-scale or constant) and the patterns searched by biologists (e.g. differential expression and differential co-expression). Another problem confusing and repelling potential users may be the large number of non-obvious parameters, which alterations affect the result. Finally, high computational cost of biclustering must be justified on multiple examples demonstrating a clear advantage of biclustering over conventional or two-way clustering.

Considering the above, an important direction of the field development may be the popularization of biclustering methods among the target audience. The rise of confidence and interest in biclustering methods may be achieved by

- demonstration that biclustering methods are capable to achieve the comparable performance in the detection of differential expression and differential co-expression patterns as conventional methods;
- the development of tools for consensus biclustering;
- evaluation of biclustering as a dimensionality reduction technique, alternative to clustering, gene set enrichment-based approaches, and autoencoders.

6.2.2 Investigation of complex diseases

By definition, complex diseases develop in the results of interactions between multiple molecular genetic and non-genetic factors. The effects of these factors may be weak and non-additive, which complicates the detection of individual factors and analysis of their impacts. Despite this, some works still utilize the same approaches for the search and analysis of associations in Mendelian and complex disorders. For instance, this concerns some gene prioritization methods discussed in section 2.2. Such methods usually lose the competition with more advanced ones, which can model known or unknown interactions and borrow the evidence of associations between interacting partners.

Heterogeneity is another intrinsic feature of non-Mendelian phenotypes and in particular of complex diseases and is taken into account less often. This thesis and many other works demonstrate high heterogeneity of breast tumors beyond known molecular subtypes. Nevertheless, to date, many fundamental questions on disease heterogeneity remain unanswered. How frequently complex diseases are composed of distinct subtypes? How many mechanistically distinct subtypes are hidden in the guise of every single diagnosis? How large are the differences between them?

The creation of large well-annotated datasets of patient-specific omics profiles and the development of computational tools for the investigation of disease heterogeneity will shed light on these questions. In the future, knowledge about disease subtypes and subtype-specific biomarkers may improve diagnostics, guide the choice of the most appropriate treatments, and therefore maintain health and even save lives for many patients. Besides that, investigation of individual disease subtypes with minimal molecular heterogeneity may give a more clear picture of disease mechanism than joint analysis of all disease cases taken together.

References

[1] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., and et al. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544.

[2] Ahmed, H. A., Mahanta, P., Bhattacharyya, D. K., and Kalita, J. K. (2014). Shifting-and-scaling correlation based biclustering algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(6):1239–1252.

[3] Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764.

[4] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

[5] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

[6] Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's online mendelian inheritance in man (OMIM(r)). *Nucleic Acids Research*, 37(Database):D793–D796.

[7] Amberger, J. S., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2018). OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47(D1):D1038–D1043.

[8] Antonarakis, S. E., Youssoufian, H., and Kazazian, H. H. (1987). Molecular genetics of hemophilia A in man (factor VIII deficiency). *Mol. Biol. Med.*, 4(2):81–94.

[9] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

[10] Bader, G. D. (2006). Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(90001):D504–D506.

[11] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18.

[12] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[13] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2010). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

[14] Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112.

[15] Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics*, 26(22):2924–2926.

[16] Beisser, D., Klau, G. W., Dandekar, T., Muller, T., and Dittrich, M. T. (2010). BioNet: an r-package for the functional analysis of biological networks. *Bioinformatics*, 26(8):1129–1130.

[17] Bellot, P., de los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics*, 210(3):809–819.

[18] Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2003). Discovering local structure in gene expression data: The order-preserving submatrix problem. *Journal of Computational Biology*, 10(3-4):373–384.

[19] Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67(3).

[20] Bertram, L. and Tanzi, R. E. (2008). Thirty years of alzheimer's disease genetics: the implications of systematic meta-analyses. *Nature Reviews Neuroscience*, 9(10):768–778.

[21] Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., Jensen, L. J., Nicolae, D., Shah, N. H., Grossman, R. L., Cox, N. J., White, K. P., and Rzhetsky, A. (2013). A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell*, 155(1):70–80.

[22] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[23] Bollobás, B., Borgs, C., Chayes, J., and Riordan, O. (2003). Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pages 132–139, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[24] Bomprezzi, R. (2003). New approaches to investigating heterogeneity in complex traits. *Journal of Medical Genetics*, 40(8):553–559.

[25] Borish, L. and Culp, J. A. (2008). Asthma: a syndrome composed of heterogeneous diseases. *Annals of Allergy, Asthma & Immunology*, 101(1):1–9.

[26] Bozdağ, D., Parvin, J. D., and Catalyurek, U. V. (2009). A biclustering method to discover co-regulated genes using diverse gene expression datasets. In *Bioinformatics and Computational Biology*, pages 151–163. Springer Berlin Heidelberg.

[27] Bragina, E. Y., Tiys, E. S., Rudko, A. A., Ivanisenko, V. A., and Freidin, M. B. (2016). Novel tuberculosis susceptibility candidate genes revealed by the reconstruction and analysis of associative networks. *Infection, Genetics and Evolution*, 46:118–123.

[28] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

[29] Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E. W., Brinkman, F. S. L., and Lynn, D. J. (2012). InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Research*, 41(D1):D1228–D1233.

[30] Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1).

[31] Browne, F., Wang, H., and Zheng, H. (2015). A computational framework for the prioritization of disease-gene candidates. *BMC Genomics*, 16(Suppl 9):S2.

[32] Burdick, D., Calimlim, M., and Gehrke, J. (2001). MAFIA: a maximal frequent itemset algorithm for transactional databases. In *Proceedings 17th International Conference on Data Engineering*. IEEE Comput. Soc.

[33] Calderone, A., Castagnoli, L., and Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. *Nature Methods*, 10(8):690–691.

[34] Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2006). Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, 7(1):78.

[35] Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., and Karp, P. D. (2015). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480.

[36] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., and Schultz, N. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data: Figure 1. *Cancer Discovery*, 2(5):401–404.

[37] Chan, C., Law, B., So, W., Chow, K., and Waye, M. (2017). Novel strategies on personalized medicine for breast cancer treatment: An update. *International Journal of Molecular Sciences*, 18(11):2423.

[38] Chapman, S. J. and Hill, A. V. S. (2012). Human genetic susceptibility to infectious disease. *Nature Reviews Genetics*, 13(3):175–188.

[39] Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.-J., Dolinski, K., and Tyers, M. (2016). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379.

[40] Chatterjee, S. and Davies, M. J. (2018). Accurate diagnosis of diabetes mellitus and new paradigms of classification. *Nature Reviews Endocrinology*, 14(7):386–387.

[41] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N. R., and Ma'ayan, A. (2013a). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128.

[42] Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(Web Server):W305–W311.

[43] Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G. (2007). Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 8(1):392.

[44] Chen, Y., Wu, X., and Jiang, R. (2013b). Integrating human omics data to prioritize candidate genes. *BMC Medical Genomics*, 6(1).

[45] Chen, Y.-A., Tripathi, L. P., and Mizuguchi, K. (2011). TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS ONE*, 6(3):e17844.

[46] Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., and Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, 36(Web Server):W399–W405.

[47] Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press.

[48] Choobdar, S., , Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., Natoli, T., Narayan, R., Subramanian, A., Zhang, J. D., Stolovitzky, G., Kutalik, Z., Lage, K., Slonim, D. K., Saez-Rodriguez, J., Cowen, L. J., Bergmann, S., and Marbach, D. (2019). Assessment of network module identification across complex diseases. *Nature Methods*, 16(9):843–852.

[49] Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., Chiew, M.-Y., Tai, C.-S., Wei, T.-Y., Tsai, T.-R., et al. (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302.

[50] Chowdhury, S. A. and Koyutürk, M. (2009). Identification of coordinately dysregulated subnetworks in complex phenotypes. In *Biocomputing 2010*, pages 133–144. WORLD SCIENTIFIC.

[51] Christiansen, S. C., Schatz, M., Yang, S.-J., Ngor, E., Chen, W., and Zuraw, B. L. (2016). Hypertension and asthma: A comorbid relationship. *The Journal of Allergy and Clinical Immunology: In Practice*, 4(1):76–81.

[52] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3.

[53] Chumbalkar, V. C., Subhashini, C., Dhople, V. M., Sundaram, C. S., Jagannadham, M. V., Kumar, K. N., Srinivas, P. N. B. S., Mythili, R., Rao, M. K., Kulkarni, M. J., Hegde, S., Hegde, A. S., Samual, C., Santosh, V., Singh, L., and Sirdeshmukh, R. (2005). Differential protein expression in human gliomas and molecular insights. *PROTEOMICS*, 5(4):1167–1177.

[54] Cogill, S. and Wang, L. (2016). Support vector machine model of developmental brain gene expression data for prioritization of autism risk gene candidates. *Bioinformatics*, page btw498.

[55] Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.

[56] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1).

[57] Consortium, G. O. (2016). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338.

[58] Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194.

[59] Cooper, D. (1998). The human gene mutation database. *Nucleic Acids Research*, 26(1):285–287.

[60] Cornish, A. J., David, A., and Sternberg, M. J. E. (2018). PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics*, 34(12):2087–2095.

[61] Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammaduddin, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Kaski, S., Gray, J. W., and Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212.

[62] Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551–562.

[63] Csardi, G., Kutalik, Z., and Bergmann, S. (2010). Modular analysis of gene expression data with r. *Bioinformatics*, 26(10):1376–1377.

[64] Daemen, A., Griffith, O. L., Heiser, L. M., Wang, N. J., Enache, O. M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J. E., Griffith, M., Hur, J. S., Huh, N., Chung, J., Cope, L., Fackler, M., Umbricht, C., Sukumar, S., Seth, P., Sukhatme, V. P., Jakkula, L. R., et al. (2013). Modeling precision treatment of breast cancer. *Genome Biology*, 14(10):R110.

[65] Dao, P., Colak, R., Salari, R., Moser, F., Davicioni, E., Schönhuth, A., and Ester, M. (2010). Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics*, 26(18):i625–i631.

[66] Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., and Sahinalp, S. C. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13):i205–i213.

[67] Davidson-Pilon, C., Kalderstam, J., Zivich, P., Kuhn, B., Fiore-Gartland, A., AbdealiJK, Moneda, L., Gabriel, WIlson, D., Parij, A., Stark, K., Anton, S., Besson, L., Jona, Gadgil, H., Golland, D., Hussey, S., Kumar, R., Noorbakhsh, J., et al. (2019). Lifelines: v0.23.0.

[68] Divina, F., Pontes, B., Giráldez, R., and Aguilar-Ruiz, J. S. (2012). An effective measure for assessing the quality of biclusters. *Computers in Biology and Medicine*, 42(2):245–256.

[69] Dogra, S., Ardern, C. I., and Baker, J. (2007). The relationship between age of asthma onset and cardiovascular disease in canadians. *Journal of Asthma*, 44(10):849–854.

[70] Doncheva, N. T., Kacprowski, T., and Albrecht, M. (2012). Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(5):429–442.

[71] Doncheva, N. T., Morris, J. H., Gorodkin, J., and Jensen, L. J. (2018). Cytoscape StringApp: Network analysis and visualization of proteomics data. *Journal of Proteome Research*, 18(2):623–632.

[72] Dongen, S. V. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141.

[73] Edwards, S. L., Beesley, J., French, J. D., and Dunning, A. M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *The American Journal of Human Genetics*, 93(5):779–797.

[74] Edwards, S. L., Brough, R., Lord, C. J., Natrajan, R., Vatcheva, R., Levine, D. A., Boyd, J., Reis-Filho, J. S., and Ashworth, A. (2008). Resistance to therapy caused by intragenic deletion in BRCA2. *Nature*, 451(7182):1111–1115.

[75] Emad, A., Cairns, J., Kalari, K. R., Wang, L., and Sinha, S. (2017). Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. *Genome Biology*, 18(1).

[76] Eren, K., Deveci, M., Kucuktunc, O., and Catalyurek, U. V. (2012). A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3):279–292.

[77] Erlich, Y., Edvardson, S., Hodges, E., Zenvirt, S., Thekkat, P., Shaag, A., Dor, T., Hannon, G. J., and Elpeleg, O. (2011). Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1a in hereditary spastic paraparesis. *Genome Research*, 21(5):658–664.

[78] Erten, S., Bebek, G., and Koyutürk, M. (2011). Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *Journal of Computational Biology*, 18(11):1561–1574.

[79] Evans, C., Hardin, J., and Stoebel, D. M. (2017). Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792.

[80] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2017). The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655.

[81] Fang, M., Hu, X., He, T., Wang, Y., Zhao, J., Shen, X., and Yuan, J. (2014). Prioritizing disease-causing genes based on network diffusion and rank concordance. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

[82] Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I. J., Vellai, T., Csermely, P., and Korcsmáros, T. (2013). SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology*, 7(1):7.

[83] Ferguson, S., Teodorescu, M. C., Gangnon, R. E., Peterson, A. G., Consens, F. B., Chervin, R. D., and Teodorescu, M. (2014). Factors associated with systemic hypertension in asthma. *Lung*, 192(5):675–683.

[84] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

[85] Gan, M., Li, W., Zeng, W., Wang, X., and Jiang, R. (2017). Mimvec: a deep learning approach for analyzing the human phenome. *BMC Systems Biology*, 11(S4).

[86] Gebregiworgis, T. and Powers, R. (2012). Application of NMR metabolomics to search for human disease biomarkers. *Combinatorial Chemistry & High Throughput Screening*, 15(8):595–610.

[87] Geeleher, P., Cox, N. J., and Huang, R. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, 15(3):R47.

[88] George, R. A., Liu, J. Y., Feng, L. L., Bryson-Richardson, R. J., Fatkin, D., and Wouters, M. A. (2006). Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Research*, 34(19):e130–e130.

[89] Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLOS Computational Biology*, 11(4):e1004120.

[90] Gill, N., Singh, S., and Aseri, T. C. (2014). Computational disease gene prioritization: An appraisal. *Journal of Computational Biology*, 21(6):456–465.

[91] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[92] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

[93] Gradishar, W. J., Anderson, B. O., Balassanian, R., Blair, S. L., Burstein, H. J., Cyr, A., Elias, A. D., Farrar, W. B., Forero, A., Giordano, S. H., Goetz, M. P., Goldstein, L. J., Isakoff, S. J., Lyons, J., Marcom, P. K., Mayer, I. A., McCormick, B., Moran, M. S., O'Regan, R. M., et al. (2017). NCCN guidelines insights: Breast cancer, version 1.2017. *Journal of the National Comprehensive Cancer Network*, 15(4):433–451.

[94] Guala, D., Sjölund, E., and Sonnhammer, E. L. L. (2014). MaxLink: network-based prioritization of genes tightly linked to a disease seed set. *Bioinformatics*, 30(18):2689–2690.

[95] Guala, D. and Sonnhammer, E. L. L. (2017). A large-scale benchmark of gene prioritization methods. *Scientific Reports*, 7(1).

[96] Guney, E. and Oliva, B. (2012). Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS ONE*, 7(9):e43557.

[97] Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129.

[98] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(S6761):C47–C52.

[99] He, X., Folkman, L., and Borgwardt, K. (2018). Kernelized rank learning for personalized drug recommendation. *Bioinformatics*, 34(16):2808–2816.

[100] He, Z. and Zhou, J. (2008). Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. *Applied and environmental microbiology*, 74 10:2957–66.

[101] Heck, S., Al-Shobash, S., Rapp, D., Le, D. D., Omlor, A., Bekhit, A., Flaig, M., Al-Kadah, B., Herian, W., Bals, R., Wagenpfeil, S., and Dinh, Q. T. (2017). High probability of comorbidities in bronchial asthma in germany. *npj Primary Care Respiratory Medicine*, 27(1).

[102] Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K. E., Jones, L. P., Assefnia, S., Chandrasekharan, S., Backlund, M. G., Yin, Y., Khramtsov, A. I., Bastein, R., Quackenbush, J., Glazer, R. I., Brown, P. H., Green, J. E., Kopelovich, L., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology*, 8(5):R76.

[103] Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353.

[104] Himmelstein, D. S. and Baranzini, S. E. (2015). Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. *PLOS Computational Biology*, 11(7):e1004259.

[105] Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6.

[106] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

[107] Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., van't Veer, L. J., Lopez-Bigas, N., Laird, P. W., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944.

[108] Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Sanden, S. V., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527.

[109] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2014). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(D1):D512–D520.

[110] Horta, D. and Campello, R. J. (2014). Similarity measures for comparing biclusterings. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(5):942–954.

[111] Hu, J. X., Thomas, C. E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629.

[112] Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., and Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell Systems*, 6(4):484–495.e5.

[113] Hung, M.-C. and Link, W. (2011). Protein localization in disease and therapy. *Journal of Cell Science*, 124(20):3381–3392.

[114] Huttenhower, C., Mutungu, K. T., Indik, N., Yang, W., Schroeder, M., Forman, J. J., Troyanskaya, O. G., and Coller, H. A. (2009). Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274.

[115] Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., Obar, R. A., Guruharsha, K. G., Li, K., Artavanis-Tsakonas, S., Gygi, S. P., and Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509.

[116] Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., Kolippakkam, D., Mintseris, J., Obar, R. A., Harris, T., Artavanis-Tsakonas, S., Sowa, M. E., Camilli, P. D., Paulo, J. A., Harper, J. W., and Gygi, S. P. (2015). The BioPlex network: A systematic exploration of the human interactome. *Cell*, 162(2):425–440.

[117] Ideker, T. and Nussinov, R. (2017). Network approaches and applications in biology. *PLOS Computational Biology*, 13(10):e1005771.

[118] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl 1):S233–S240.

[119] Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Research*, 18(4):644–652.

[120] Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003.

[121] Isakov, O., Dotan, I., and Ben-Shachar, S. (2017). Machine learning–based gene prioritization identifies novel candidate risk genes for inflammatory bowel disease. *Inflammatory Bowel Diseases*, 23(9):1516–1523.

[122] Ivanisenko, V. A., Saik, O. V., Ivanisenko, N. V., Tiys, E. S., Ivanisenko, T. V., Demenkov, P. S., and Kolchanov, N. A. (2015). ANDSystem: an associative network discovery system for automated literature mining in the field of biology. *BMC Systems Biology*, 9(Suppl 2):S2.

[123] Janssens, A. C. J. and van Duijn, C. M. (2008). Genome-based prediction of common diseases: advances and prospects. *Human Molecular Genetics*, 17(R2):R166–R173.

[124] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database):D412–D416.

[125] Jiang, R. (2015). Walking on multiple disease-gene networks to prioritize candidate genes. *Journal of Molecular Cell Biology*, 7(3):214–230.

[126] Joehanes, R., Zhang, X., Huan, T., Yao, C., xia Ying, S., Nguyen, Q. T., Demirkale, C. Y., Feolo, M. L., Sharopova, N. R., Sturcke, A., Schäffer, A. A., Heard-Costa, N., Chen, H., ching Liu, P., Wang, R., Woodhouse, K. A., Tanriverdi, K., Freedman, J. E., Raghavachari, N., Dupuis, J., Johnson, A. D., O'Donnell, C. J., Levy, D., and Munson, P. J. (2017). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology*, 18(1).

[127] Johnson, M., Nriagu, J., Hammad, A., Savoie, K., and Jamil, H. (2010). Asthma, environmental risk factors, and hypertension among arab americans in metro detroit. *Journal of Immigrant and Minority Health*, 12(5):640–651.

[128] Jostins, L. and Barrett, J. C. (2011). Genetic risk prediction in complex disease. *Human Molecular Genetics*, 20(R2):R182–R188.

[129] Jourquin, J., Duncan, D., Shi, Z., and Zhang, B. (2012). GLAD4u: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics*, 13(Suppl 8):S20.

[130] Kacprowski, T., Doncheva, N. T., and Albrecht, M. (2013). NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471–1473.

[131] Kalathur, R. K. R., Pinto, J. P., Sahoo, B., Chaurasia, G., and Futschik, M. E. (2017). HDNetDB: A molecular interaction database for network-oriented investigations into huntington's disease. *Scientific Reports*, 7(1).

[132] Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2012). The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D793–D800.

[133] Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2008). Consensus-PathDB—a database for integrating human functional interaction networks. *Nucleic Acids Research*, 37(suppl_1):D623–D628.

[134] Kanehisa, M. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

[135] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462.

[136] Keith, B. P., Robertson, D. L., and Hentges, K. E. (2014). Locus heterogeneity disease genes encode proteins with high interconnectivity in the human protein interaction network. *Frontiers in Genetics*, 5.

[137] Khakabimamaghani, S. and Ester, M. (2015). Bayesian biclustering for patient stratification. In *Biocomputing 2016*. World Scientific.

[138] Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van̂der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266.

[139] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

[140] Kobayashi, M., Yokoyama, K., Shimizu, E., Yusa, N., Ito, M., Yamaguchi, R., Imoto, S., Miyano, S., and Tojo, A. (2017). Phenotype-based gene analysis allowed successful diagnosis of x-linked neutropenia associated with a novel WASp mutation. *Annals of Hematology*, 97(2):367–369.

[141] Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958.

[142] Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., Pierleoni, A., Pignatelli, M., Platt, T., Rowland, F., Wankar, P., Bento, A. P., Burdett, T., Fabregat, A., et al. (2016). Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45(D1):D985–D994.

[143] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97.

[144] Kumar, A. A., Laer, L. V., Alaerts, M., Ardeshirdavani, A., Moreau, Y., Laukens, K., Loeys, B., and Vandeweyer, G. (2018). pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. *Bioinformatics*, 34(13):2254–2262.

[145] Kuruvilla, M. E., Lee, F. E.-H., and Lee, G. B. (2018). Understanding asthma phenotypes, endotypes, and mechanisms of disease. *Clinical Reviews in Allergy & Immunology*, 56(2):219–233.

[146] Lacouture, M. and Sibaud, V. (2018). Toxic side effects of targeted therapies and immunotherapies affecting the skin, oral mucosa, hair, and nails. *American Journal of Clinical Dermatology*, 19(S1):31–39.

[147] Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868.

[148] Langfelder, P. and Horvath, S. (2008). WGCNA: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1).

[149] Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T. A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., and Gray, J. W. (2010). Exon-level microarray analyses identify alternative splicing programs in breast cancer. *Molecular Cancer Research*, 8(7):961–974.

[150] Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2014). MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Research*, 43(D1):D321–D327.

[151] Lazzeroni, L. and Owen, A. (2000). Plaid models for gene expression data. *Stat Sin.*, 12:61–86.

[152] Le, D.-H. and Kwon, Y.-K. (2012). GPEC: A cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection. *Computational Biology and Chemistry*, 37:17–23.

[153] Le, D.-H. and Kwon, Y.-K. (2013). Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. *Computational Biology and Chemistry*, 44:1–8.

[154] Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121.

[155] Lee, Y., Li, H., Li, J., Rebman, E., Achour, I., Regan, K. E., Gamazon, E. R., Chen, J. L., Yang, X. H., Cox, N. J., and Lussier, Y. A. (2013). Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *Journal of the American Medical Informatics Association*, 20(4):619–629.

[156] Letovsky, S. and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(Suppl 1):i197–i204.

[157] Li, G., Ma, Q., Tang, H., Paterson, A. H., and Xu, Y. (2009). QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research*, 37(15):e101–e101.

[158] Li, H., Fan, J., Vitali, F., Berghout, J., Aberasturi, D., Li, J., Wilson, L., Chiu, W., Pumarejo, M., Han, J., Kenost, C., Koripella, P. C., Pouladi, N., Billheimer, D., Bedrick, E. J., and Lussier, Y. A. (2018). Novel disease syndromes unveiled by integrative multiscale network analysis of diseases sharing molecular effectors and comorbidities. *BMC Medical Genomics*, 11(S6).

[159] Li, J., Lin, X., Teng, Y., Qi, S., Xiao, D., Zhang, J., and Kang, Y. (2016). A comprehensive evaluation of disease phenotype networks for gene prioritization. *PLOS ONE*, 11(7):e0159457.

[160] Li, L. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.

[161] Li, Y. and Patra, J. C. (2010). Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224.

[162] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425.

[163] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. (2011). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861.

[164] Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Hu, H., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11.

[165] Liu, Y., Liang, Y., and Wishart, D. (2015). PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Research*, 43(W1):W535–W542.

[166] López, Y., Nakai, K., and Patil, A. (2015). HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species. *Database*, 2015:bav117.

[167] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12).

[168] Luck, K., Sheynkman, G. M., Zhang, I., and Vidal, M. (2017). Proteome-scale human interactomics. *Trends in Biochemical Sciences*, 42(5):342–354.

[169] Luo, J. and Liang, S. (2015). Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. *Journal of Biomedical Informatics*, 53:229–236.

[170] Lysenko, A., Boroevich, K. A., and Tsunoda, T. (2017). Arete – candidate gene prioritization using biological network topology with additional evidence types. *BioData Mining*, 10(1).

[171] Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics*, 27(9):1290–1298.

[172] Maass, P. G., Aydin, A., Luft, F. C., Schächterle, C., Weise, A., Stricker, S., Lindschau, C., Vaegler, M., Qadri, F., Toka, H. R., Schulz, H., Krawitz, P. M., Parkhomchuk, D., Hecht, J., Hollfinger, I., Wefeld-Neuenfeld, Y., Bartels-Klein, E., Mühl, A., Kann, M., Schuster, H., et al. (2015). PDE3a mutations cause autosomal dominant hypertension with brachydactyly. *Nature Genetics*, 47(6):647–653.

[173] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (2016). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, 45(D1):D896–D901.

[174] Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

[175] Mahurkar, S., Moldovan, M., Suppiah, V., and O'Doherty, C. (2013). Identification of shared genes and pathways: A comparative study of multiple sclerosis susceptibility, severity and response to interferon beta treatment. *PLoS ONE*, 8(2):e57655.

[176] Marabti, E. E. and Younis, I. (2018). The cancer spliceome: Reprograming of alternative splicing in cancer. *Frontiers in Molecular Biosciences*, 5.

[177] Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, 6(1):287–303.

[178] Martínez, V., Cano, C., and Blanco, A. (2014). ProphNet: A generic prioritization method through propagation of information. *BMC Bioinformatics*, 15(Suppl 1):S5.

[179] Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334.

[180] McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

[181] McClellan, J. and King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, 141(2):210–217.

[182] Melamed, R. D., Emmett, K. J., Madubata, C., Rzhetsky, A., and Rabadan, R. (2015). Genetic similarity between cancers and comorbid mendelian diseases identifies candidate driver genes. *Nature Communications*, 6(1).

[183] Meldal, B. H., Forner-Martinez, O., Costanzo, M. C., Dana, J., Demeter, J., Dumousseau, M., Dwight, S. S., Gaulton, A., Licata, L., Melidoni, A. N., Ricard-Blum, S., Roechert, B., Skyzypek, M. S., Tiwari, M., Velankar, S., Wong, E. D., Hermjakob, H., and Orchard, S. (2014). The complex portal - an encyclopaedia of macromolecular complexes. *Nucleic Acids Research*, 43(D1):D479–D484.

[184] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabasi, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601–1257601.

[185] Meshkin, A., Shakery, A., and Masoudi-Nejad, A. (2018). GPS: Identification of disease genes by rank aggregation of multi-genomic scoring schemes. *Genomics*.

[186] Mirkin, B. G. (1996). *Mathematical classification and clustering*. Kluwer Academic Publishers, Dordrecht Boston.

[187] Mishra, D. and Sahu, B. (2011). *A signal-to-noise classification model for identification of differentially expressed genes from gene expression data*. IEEE.

[188] Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

[189] Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., Miyamoto, T., Miyashita, A., Kuwano, R., and Tanaka, H. (2012). AlzPathway: a comprehensive map of signaling pathways of alzheimer's disease. *BMC Systems Biology*, 6(1):52.

[190] Morales, D. R., Lipworth, B. J., Donnan, P. T., Jackson, C., and Guthrie, B. (2017). Respiratory effect of beta-blockers in people with asthma and cardiovascular disease: population-based nested case control study. *BMC Medicine*, 15(1).

[191] Mordelet, F. and Vert, J.-P. (2011). ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12(1):389.

[192] Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536.

[193] Mosca, E., Bersanelli, M., Gnocchi, M., Moscatelli, M., Castellani, G., Milanesi, L., and Mezzelani, A. (2017). Network diffusion-based prioritization of autism risk genes identifies significantly connected gene modules. *Frontiers in Genetics*, 8.

[194] Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(Suppl 1):S4.

[195] Mukhopadhyay, A., Maulik, U., and Bandyopadhyay, S. (2009). A novel coherence measure for discovering scaling biclusters from gene expression data. *Journal of Bioinformatics and Computational Biology*, 07(05):853–868.

[196] Murali, T. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium of Biocomputing*, pages 77–88.

[197] Murdoch, J. D. and State, M. W. (2013). Recent developments in the genetics of autism spectrum disorders. *Current Opinion in Genetics & Development*, 23(3):310–315.

[198] Narrandes, S. and Xu, W. (2018). Gene expression detection assay for cancer clinical use. *Journal of Cancer*, 9(13):2249–2265.

[199] Navlakha, S. and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063.

[200] Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B., and Moreau, Y. (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460.

[201] Nurgali, K., Jagoe, R. T., and Abalo, R. (2018). Editorial: Adverse effects of cancer chemotherapy: Anything new to improve tolerance and reduce sequelae? *Frontiers in Pharmacology*, 9.

[202] Ohn, J. H. (2017). The landscape of genetic susceptibility correlations among diseases and traits. *J Am Med Inform Assoc*, 24(5):921–926.

[203] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., et al. (2013). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363.

[204] Östlund, G., Lindskog, M., and Sonnhammer, E. L. L. (2009). Network-based identification of novel cancer genes. *Molecular & Cellular Proteomics*, 9(4):648–655.

[205] Ozgur, A., Vu, T., Erkan, G., and Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285.

[206] Padilha, V. A. and Campello, R. J. G. B. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18(1).

[207] Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., Mardis, E., Kupfer, D., Wilson, R., Kris, M., and Varmus, H. (2004). EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences*, 101(36):13306–13311.

[208] Parente, L. and Solito, E. (2004). Annexin 1: more than an anti-phospholipase protein. *Inflammation Research*, 53(4):125–132.

[209] Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

[210] Peeters, R. (2003). The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics*, 131(3):651–654.

[211] Peltonen, L., Perola, M., Naukkarinen, J., and Palotie, A. (2006). Lessons from studying monogenic disease for common disease. *Human Molecular Genetics*, 15(suppl_1):R67–R74.

[212] Peng, X., Chen, Z., Farshidfar, F., Xu, X., Lorenzi, P. L., Wang, Y., Cheng, F., Tan, L., Mojumdar, K., Du, D., Ge, Z., Li, J., Thomas, G. V., Birsoy, K., Liu, L., Zhang, H., Zhao, Z., Marchand, C., Weinstein, J. N., Bathe, O. F., et al. (2018). Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell Reports*, 23(1):255–269.e4.

[213] Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D. W. Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., McKinney, S., Oloumi, A., Shah, S., Rosenfeld, N., Murphy, L., Bentley, D. R., Ellis, I. O., Purushotham, A., Pinder, S. E., Børresen-Dale, A.-L., Earl, H. M., Pharoah, P. D., Ross, M. T., Aparicio, S., and Caldas, C. (2016). The somatic mutation profiles of 2, 433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7(1).

[214] Perfetto, L., Briganti, L., Calderone, A., Perpetuini, A. C., Iannuccelli, M., Langone, F., Licata, L., Marinkovic, M., Mattioni, A., Pavlidou, T., Peluso, D., Petrilli, L. L., Pirrò, S., Posca, D., Santonico, E., Silvestri, A., Spada, F., Castagnoli, L., and Cesareni, G. (2015). SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Research*, 44(D1):D548–D554.

[215] Perou, C. M. (2011). Molecular stratification of triple-negative breast cancers. *The Oncologist*, 16(S1):61–70.

[216] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

[217] Pers, T. H., Timshel, P., Ripke, S., Sullivan, P. F., O'Donovan, M. C., Franke, L., and Hirschhorn, J. N. (2016). Comprehensive analysis of schizophrenia-associated loci highlights ion channel pathways and biologically plausible candidate causal genes. *Human Molecular Genetics*, 25(6):1247–1254.

[218] Pinero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839.

[219] Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015(0):bav028–bav028.

[220] Piro, R. M. and Cunto, F. D. (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS Journal*, 279(5):678–696.

[221] Plaisier, S. B., Taschereau, R., Wong, J. A., and Graeber, T. G. (2010). Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Research*, 38(17):e169–e169.

[222] Pontes, B., Giráldez, R., and Aguilar-Ruiz, J. S. (2013). Configurable pattern-based evolutionary biclustering of gene expression data. *Algorithms for Molecular Biology*, 8(1).

[223] Pontes, B., Giráldez, R., and Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180.

[224] Prado-Vázquez, G., Gámez-Pozo, A., Trilla-Fuertes, L., Arevalillo, J. M., Zapater-Moros, A., Ferrer-Gómez, M., Díaz-Almirón, M., López-Vacas, R., Navarro, H., Maín, P., Feliú, J., Zamora, P., Espinosa, E., and Vara, J. Á. F. (2019). A novel approach to triple-negative breast cancer molecular classification reveals a luminal immune-positive subgroup with good prognoses. *Scientific Reports*, 9(1).

[225] Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772.

[226] Prat, A. and Perou, C. M. (2010). Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, 5(1):5–23.

[227] Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

[228] Reiss, D. J., Plaisier, C. L., Wu, W.-J., and Baliga, N. S. (2015). cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism. *Nucleic Acids Research*, 43(13):e87–e87.

[229] Relogio, A. (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research*, 30(11):51e–51.

[230] Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinformatics*, 12(1):480.

[231] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

[232] Robinson, D. G., Wang, J. Y., and Storey, J. D. (2015). A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. *Nucleic Acids Research*, page gkv636.

[233] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

[234] Rodriguez-Baena, D. S., Perez-Pulido, A. J., and Aguilar-Ruiz, J. S. (2011). A biclustering algorithm for extracting bit-patterns from binary datasets. *Bioinformatics*, 27(19):2738–2745.

[235] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2009). CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research*, 38(suppl_1):D497–D501.

[236] Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications*, 9(1).

[237] Saik, O. V., Demenkov, P. S., Ivanisenko, T. V., Bragina, E. Y., Freidin, M. B., Dosenko, V. E., Zolotareva, O. I., Choynzonov, E. L., Hofestaedt, R., and Ivanisenko, V. A. (2018a). Search for new candidate genes involved in the comorbidity of asthma and hypertension based on automatic analysis of scientific literature. *Journal of Integrative Bioinformatics*, 15(4).

[238] Saik, O. V., Demenkov, P. S., Ivanisenko, T. V., Bragina, E. Y., Freidin, M. B., Goncharova, I. A., Dosenko, V. E., Zolotareva, O. I., Hofestaedt, R., Lavrik, I. N., Rogaev, E. I., and Ivanisenko, V. A. (2018b). Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Medical Genomics*, 11(S1).

[239] Salwinski, L. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(90001):449D–451.

[240] Schmitt, T., Ogris, C., and Sonnhammer, E. L. L. (2013). FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Research*, 42(D1):D380–D388.

[241] Ségalat, L. (2007). Loss-of-function genetic diseases and the concept of pharmaceutical targets. *Orphanet Journal of Rare Diseases*, 2(1):30.

[242] Serin, A. and Vingron, M. (2011). DeBi: Discovering differentially expressed biclusters using a frequent itemset approach. *Algorithms for Molecular Biology*, 6(1).

[243] Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6(7):639–645.

[244] Shannon, P. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

[245] Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C., and Ester, M. (2020). AITL: Adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *BioRXiv*.

[246] Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509.

[247] Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., Sahni, N., Thibault, D., Voung, L., Guo, F., Ghiassian, S. D., Gulbahce, N., Baribaud, F., Tocker, J., Dobrin, R., Barnathan, E., Liu, H., Panettieri, R. A., Tantisira, K. G., Qiu, W., Raby, B. A., Silverman, E. K., Vidal, M., Weiss, S. T., and Barabasi, A. L. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.*, 24(11):3005–3020.

[248] Shen, H., Cheng, X., Cai, K., and Hu, M.-B. (2009). Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706–1712.

[249] Shim, J. E., Hwang, S., and Lee, I. (2015). Pathway-dependent effectiveness of network algorithms for gene prioritization. *PLOS ONE*, 10(6):e0130589.

[250] Shoshi, A., Hofestädt, R., Zolotareva, O., Friedrichs, M., Maier, A., Ivanisenko, V. A., Dosenko, V. E., and Bragina, E. Y. (2018). GenCoNet – a graph database for the analysis of comorbidities by gene networks. *Journal of Integrative Bioinformatics*, 15(4).

[251] Shrestha, R., Hodzic, E., Sauerwald, T., Dao, P., Wang, K., Yeung, J., Anderson, S., Vandin, F., Haffari, G., Collins, C. C., and Sahinalp, S. C. (2017). HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. *Genome Research*, 27(9):1573–1588.

[252] Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., Strawbridge, R. J., Pers, T. H., Fischer, K., Justice, A. E., Workalemahu, T., Wu, J. M. W., Buchkovich, M. L., Heard-Costa, N. L., Roman, T. S., Drong, A. W., Song, C., Gustafsson, S., Day, F. R., Esko, T., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196.

[253] Snow, O., Noghabi, H. S., Lu, J., Zolotareva, O., Lee, M., and Ester, M. (2019). BD-KANN – biological domain knowledge-based artificial neural network for drug response prediction. *BioRXiv*.

[254] Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E., and Borresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

[255] Srihari, S. and Leong, H. W. (2013). A survey of computational methods for protein complex prediction from protein interaction networks. *Journal of Bioinformatics and Computational Biology*, 11(02):1230002.

[256] Stark, C. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(90001):D535–D539.

[257] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[258] Sun, P., Speicher, N. K., Röttger, R., Guo, J., and Baumbach, J. (2014). Bi-force: large-scale bicluster editing and its application to gene expression data biclustering. *Nucleic Acids Research*, 42(9):e78–e78.

[259] Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., and Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular Systems Biology*, 4.

[260] Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P., and Khatri, P. (2016). Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Research*, 45(1):e1–e1.

[261] Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(Suppl 1):S136–S144.

[262] Tatusov, R. L. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637.

[263] Thienpont, B., Zhang, L., Postma, A. V., Breckpot, J., Tranchevent, L.-C., Loo, P. V., Møllgård, K., Tommerup, N., Bache, I., Tümer, Z., van Engelen, K., Menten, B., Mortier, G., Waggoner, D., Gewillig, M., Moreau, Y., Devriendt, K., and Larsen, L. A. (2010). Haploinsufficiency of TAB2 causes congenital heart defects in humans. *The American Journal of Human Genetics*, 86(6):839–849.

[264] Tranchevent, L.-C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., and Moreau, Y. (2016). Candidate gene prioritization with endeavour. *Nucleic Acids Research*, 44(W1):W117–W121.

[265] Tranchevent, L.-C., Barriot, R., Yu, S., Vooren, S. V., Loo, P. V., Coessens, B., Moor, B. D., Aerts, S., and Moreau, Y. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, 36(Web Server):W377–W384.

[266] Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967.

[267] Turner, B. M., Gimenez-Sanders, M. A., Soukiazian, A., Breaux, A. C., Skinner, K., Shayne, M., Soukiazian, N., Ling, M., and Hicks, D. G. (2019). Risk stratification of ER-positive breast cancer patients: A multi-institutional validation and outcome study of the rochester modified magee algorithm (RoMMa) and prediction of an oncotype DX recurrence score < 26. *Cancer Medicine*.

[268] Turner, H., Bailey, T., and Krzanowski, W. (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis*, 48(2):235–254.

[269] Ulitsky, I., Krishnamurthy, A., Karp, R. M., and Shamir, R. (2010). DEGAS: De novo discovery of dysregulated pathways in human diseases. *PLoS ONE*, 5(10):e13367.

[270] van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, page bbw139.

[271] van der Valk, R. J., Kreiner-Møller, E., Kooijman, M. N., Guxens, M., Stergiakouli, E., Sääf, A., Bradfield, J. P., Geller, F., Hayes, M. G., Cousminer, D. L., Körner, A., Thiering, E., Curtin, J. A., Myhre, R., Huikari, V., Joro, R., Kerkhof, M., Warrington, N. M., Pitkänen, N., Ntalla, I., et al. (2014). A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human Molecular Genetics*, 24(4):1155–1168.

[272] van Uitert, M., Meuleman, W., and Wessels, L. (2008). Biclustering sparse binary genomic data. *Journal of Computational Biology*, 15(10):1329–1345.

[273] Vandin, F., Clay, P., Upfal, E., and Raphael, B. J. (2011). Discovery of mutated subnetworks associated with clinical data in cancer. In *Biocomputing 2012*. WORLD SCIENTIFIC.

[274] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1):e1000641.

[275] Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700.

[276] Wagner, A. H., Taylor, K. R., DeLuca, A. P., Casavant, T. L., Mullins, R. F., Stone, E. M., Scheetz, T. E., and Braun, T. A. (2013). Prioritization of retinal disease genes: An integrative approach. *Human Mutation*, 34(6):853–859.

[277] Walhout, A. J. (2000). Protein Interaction Mapping in C.elegans Using Proteins Involved in Vulval Development. *Science*, 287(5450):116–122.

[278] Wallstrom, G., Anderson, K. S., and LaBaer, J. (2013). Biomarker discovery for heterogeneous diseases. *Cancer Epidemiology Biomarkers & Prevention*, 22(5):747–755.

[279] Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854.

[280] Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl_2):W214–W220.

[281] White, S. and Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. ACM Press.

[282] Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, 9(4):326–332.

[283] Woess, W. (1994). Random walks on infinite graphs and groups - a survey on selected topics. *Bulletin of the London Mathematical Society*, 26(1):1–60.

[284] Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Molecular Systems Biology*, 4.

[285] Xie, J., Ma, A., Fennell, A., Ma, Q., and Zhao, J. (2018). It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Briefings in Bioinformatics*.

[286] Xu, Q., Li, K., Sun, Q., Ding, D., Zhao, Y., Yang, N., Luo, Y., Liu, Z., Zhang, Y., Wang, C., Xia, K., Yan, X., Jiang, H., Shen, L., Tang, B., and Guo, J. (2017). Rare GCH1 heterozygous variants contributing to parkinson's disease. *Brain*, 140(7):e41–e41.

[287] Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, 12(9):841–843.

[288] Yoo, M., Shin, J., Kim, J., Ryall, K. A., Lee, K., Lee, S., Jeon, M., Kang, J., and Tan, A. C. (2015). DSigDB: drug signatures database for gene set analysis: Fig. 1. *Bioinformatics*, 31(18):3069–3071.

[289] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):e59.

[290] Yu, S., Falck, T., Daemen, A., Tranchevent, L.-C., Suykens, J. A., Moor, B. D., and Moreau, Y. (2010). L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, 11(1):309.

[291] Zampieri, G., Tran, D. V., Donini, M., Navarin, N., Aiolli, F., Sperduti, A., and Valle, G. (2018). Scuba: scalable kernel-based gene prioritization. *BMC Bioinformatics*, 19(1).

[292] Zhang, M., Yao, C., Guo, Z., Zou, J., Zhang, L., Xiao, H., Wang, D., Yang, D., Gong, X., Zhu, J., Li, Y., and Li, X. (2008). Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*, 24(18):2057–2063.

[293] Zhang, Y., Liu, J., Liu, X., Fan, X., Hong, Y., Wang, Y., Huang, Y., and Xie, M. (2018). Prioritizing disease genes with an improved dual label propagation framework. *BMC Bioinformatics*, 19(1).

[294] Zhang, Y., Xie, J., Yang, J., Fennell, A., Zhang, C., and Ma, Q. (2016). QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics*, page btw635.

[295] Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE*, 9(1):e78644.

[296] Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024.

[297] Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *IN ICML*, pages 912–919.

[298] Zolotareva, O., Khakabimamaghani, S., Isaeva, O. I., Chervontseva, Z., Savchik, A., and Ester, M. (2020). Identification of differentially expressed gene modules in heterogeneous diseases. *Bioinformatics*.

[299] Zolotareva, O. and Kleine, M. (2019). A survey of gene prioritization tools for mendelian and complex human diseases. *Journal of Integrative Bioinformatics*, 16(4).

[300] Zolotareva, O., Saik, O. V., Königs, C., Bragina, E. Y., Goncharova, I. A., Freidin, M. B., Dosenko, V. E., Ivanisenko, V. A., and Hofestädt, R. (2019). Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. *Scientific Reports*, 9(1).