Università degli Studi di Padova

Head Office: Università degli Studi di Padova
Department of Biology

Ph.D. Course in Biosciences
Curriculum: Genetics, Genomics and Bioinformatics
Series XXX

Computational methods for the discovery of molecular signatures from Omics Data

Coordinator: Prof.ssa Ildikò Szabò

Supervisor: Prof.ssa Chiara Romualdi

Co-Supervisor: Prof.ssa Monica Chiogna

Ph.D. student: Elisa Salviato

Abstract

I biomarcatori molecolari, ottenuti attraverso l'utilizzo di piattaforme high-throughput sequencing, costituiscono le basi della medicina personalizzata di nuova generazione. Nonostante un decennio di sforzi e di investimenti, il numero di biomarcatori validi a livello clinico rimane modesto. La natura di "big-data" dei dati omici infatti ha introdotto nuove sfide che richiedono un miglioramento sia degli strumenti di analisi che di quelli di esplorazione dei risultati. In questa tesi vengono proposti due temi centrali, entrambi volti al miglioramento delle metodologie statistiche e computazionali nell'ambito dell'individuazione di firme molecolari.

Il primo lavoro si sviluppa attorno all'identificazione di miRNA su siero in pazienti affetti da carcinoma ovarico impiegabili a livello diagnostico. In particolare si propongono delle linee guida per il processo di analisi e una normalizzazione *ad-hoc* per dati di microarray da utilizzarsi nel contesto di molecole circolanti.

Nel secondo lavoro si presenta un nuovo approccio basato sui modelli grafici Gaussiani per l'identificazione di firme molecolari funzionali. Il metodo proposto è in grado di esplorare le informazioni contenute nei pathway biologici e di evidenziare la potenziale origine del comportamento differenziale tra due condizioni sperimentali.

Abstract

Molecular biomarkers, derived from high-throughput technologies, are the foundations of the "next-generation" precision medicine. Despite a decade of intense efforts and investments, the number of clinically valid biomarkers is modest. Indeed, the "big-data" nature of omics data provides new challenges that require an improvement in the strategies of data analysis and interpretation.

In this thesis, two themes are proposed, both aimed at improving the statistical and computational methodology in the field of signatures discovery. The first work aim at identifying serum miRNAs to be used as diagnostic biomarkers associated with ovarian cancer. In particular, a guideline and an *ad-hoc* microarray normalization strategy for the analysis of circulating miRNAs is proposed.

In the second work, a new approach for the identification of functional molecular signatures based on Gaussian graphical models is presented. The model can explore the topological information contained in the biological pathways and highlight the potential sources of differential behaviors in two experimental conditions.

Table of contents

List of figures

List of tables

Chapter 1

OVERVIEW

In the last two decades, high-throughput technologies have allowed an unprecedented vision of molecular biological systems. The research fields that are involved in obtaining and understanding these measures - such as genomics, transcriptomics, and proteomics - are commonly aggregated in the so-called "omics" sciences.

Given a molecular measurement, a natural objective is the development of statistical models that use these measures to predict the clinical outcome of interest, such as disease state, survival time, response to therapy. Often, patients with the same symptoms and signs of cancer have different results, or patients subject to an identical medical treatment have different reactions to the toxicity of a drug. The improvement of prognosis and diagnosis regardless of the approach and the data chosen is the goal of "precision medicine".

Precision medicine, in its more recent sense, is based on finding a set of signature molecules. A molecular signature is thus defined as a set of biomolecular features (e.g., DNA variations, DNA copy number, mRNAs, proteins and metabolite abundances) together with a predefined computational procedure that is able to predict a phenotype of clinical interest.

Biomarkers are the foundations of precision medicine, and they can be grouped on the basis of a variety of characteristics. For example, according to the purpose, they can be correlational (i.e., only associated with the disease) or functional (i.e., they have an identified mechanism of action related to disease). In general, they can be measured and used individually or in groups to infer the risk, diagnosis, prognosis, or therapeutic response. Classification may also depend on the type of molecule used.

DNA, RNA, protein metabolites can all function as biomarkers.

The statistical model used for biomarkers identification and phenotype prediction, independent of the univariate or multivariate setting, can be divided into two major categories of approaches: i) evidence-based approaches (EBAs) and ii) knowledge-based approaches (KBAs). EBAs are commonly associated with hypothesis-generating paradigms and allow interrogating essentially all the collected data by providing the full spectrum of possible associations between genotype and phenotype. In contrast, KBAs try to infuse expert knowledge into omics data analysis through hypothesis-testing paradigms. The first category is frequently used in the screening phase, where the goal is the diagnosis and/or the prognosis, without any particular interest in functionality. Whereas, the second class is more targeted to providing mechanistic evidence, rather than the prediction of a phenotype. Although this different application, the EBAs, and KBAs are not mutually exclusive, but they must be thought of as complementary tools to facilitate and complement various goals.

Precision medicine in the last decade has accumulated a long list of terms with similar meanings, including personalized medicine, P4 medicine, genomic medicine, predictive medicine, and individual medicine. Although its evolution, the overall approach to the development of signatures has not changed considerably. Indeed, despite a decade of intense efforts and a substantial investment in labor and funds, the number of clinically valid biomarkers approved by FDA (Food and Drug Administration) are modest, creating in the scientific community a certain mistrust for the "biomarker" term.

The success of molecular signature discovery has mainly been hampered by several factors that characterize all "big data" disciplines. The first difficulty can be attributed to the low number of samples collected in omics experiments, compared to the number of measured molecules. In fact, most biomarkers are identified at the tissue-level and therefore require invasive procedures such as biopsy, slowing down the possibility of expanding existing cohorts. Although tissue-level studies are often crucial in recognizing a mechanistic link between biomarkers and carcinogenesis, they are not practical for assessing cancer risk or monitor response to treatment in the clinic of large populations over time.

Also, the heterogeneous nature of omics data presents a new challenge that requires a deep understanding of the underlying biological mechanisms and the algorithms to perform data integration and interpretation. Thus, it becomes crucial to understand

the advantages and disadvantages of different platforms to carry out explorations and to draw inference models that can decipher and translate the molecular mechanisms expressed by molecular profiles, in an automatic and functional way.

In this thesis, two central themes are proposed, both intended to contribute to the improvement of statistical and computational methodology in the field of signature discovering.

In the first part (Part I), I present a study aimed at identifying circulating miRNAs to be used as diagnostic biomarkers associated with ovarian cancer, based on an evidence-based approach. In particular, a refinement of the normalization strategy, specifically designed to adapt to the properties of circulating molecules, is proposed. Particular attention has also been paid to the experimental design to ensure i) consistent homogeneity of the samples using clinical information and ii) reproducibility of the obtained results, using three profiling technologies and two independent cohorts.

In the second part (Part II), a new knowledge-based approach to identify functional molecular signatures is proposed . In particular, an innovative perspective is used to exploit the topological information contained in biological pathways through Gaussian graphical models, moving from marginal-based to conditional-based approach. Through this model, it is possible to zoom on the potential sources of the different behaviors of two experimental conditions, identify its causes and translate the results into biological hypotheses. The natural application of this model is aimed at clarifying carcinogenesis processes and developing therapeutic targets.

Part I

Circulating miRNA landscape identifies *miR-1246* as promising biomarker in high-grade serous ovarian carcinoma

Chapter 2

INTRODUCTION

2.1 High-Grade Serous Ovarian Carcinoma

Ovarian Cancer (OC) is the seventh most diagnosed tumor among women and the most lethal gynecology malignancy. Annually, 239.000 new cases and 152.000 deaths are estimated worldwide. The highest rate is recorded in East and Central Europe [30].

The majority of ovarian cancer - benign and malignant forms - are of epithelial origin (90%), whereas fewer develop from other cell types, such as sex-cord stromal (5-6%), germ cell (2-3%) or mixed cell-type tumors [60]. Within the most common type of ovarian cancer - the epithelial ovarian cancer (EOC) - five main subtypes are identified: high-grade serous (70%), endometrioid (10%), clear cell (10%), mucinous (3%) and low-grade serous (<5%) [64]. These histotypes reflect the strong heterogeneity of the EOC regarding cellular origin, pathogenesis, molecular alterations, gene expressions, and prognosis, as to be considered essentially distinct disease.

The most common histological EOC, high-grade ovarian cancer (HGSOC) is characterized by a high aggressiveness and rapid development, with a five-year survival rate after diagnosis of less than 30%. In fact, in 90% of cases, despite an aggressive surgery and early chemotherapy, patients develop resistance to platinum agents and die from an incurable disease. Indeed, because of its asymptomatic nature, 70% of HGSOC are diagnosed in the advanced stage when the neoplasm is diffused into the peritoneum (stage III) and to distant organs organ (stage IV).

The lack of therapeutic measures that can efficiently improve survival also depends on the fact that the precursor lesion and the pathogenesis of the disease are still being discussed within the scientific community. Although, in the last decade, a growing body of evidence suggests that the majority of EOC could develop from other gynecological

tissues [76] - involving only overseas ovary - morphological and genetic studies do not have a clear pattern of progression. In particular, HGSOCs could originate from the epithelium of the fibrial end of the fallopian tube. At present, the most effective strategy to reduce mortality in HGSOC patients is the complete surgical debulking of ovaries and fallopian tubes in women carrying germline BRCA1/BRCA2 mutations or with a strong family history of breast and/or ovarian cancer.

From this overview, it emerges the importance of optimizing clinical outcomes through early diagnosis detection in order to limit invasive surgeries and to achieve survival improvement.

2.2 Diagnosis and screening

The often asymptomatic nature of primary lesions, along with the lack of adequate screening and diagnosis technique, make the discovery of HGSOC in early stages of the disease, an infrequent event.

At present, the serum Carbohydrate Antigen 125 (CA-125), in combination with transvaginal ultrasound, is the most common biomarker used for the diagnosis of the HGSOC. However, CA-125 cannot be adequately characterized as a screening test because of the high incidence of false positives [100] among benign gynecological conditions in premenopausal women (such as endometriosis).

The ineffectiveness of CA-125 screening properties has been illustrated in an extensive multicenter prospective study, called PLCO (Prostate, Lung, Colorectal and Ovarian Cancer Screening) [112], where the predictive value of CA-125 resulted in less than 4%. In general, at the diagnostic level, the sensitivity of CA-125 in the distinction between benign and malignant mass varies between 61% and 90%, while the specificity ranges between 35% and 91%. The wide variation is due to different inclusion criteria for premenopausal women in diverse studies [74]. To endorse these findings, a more recent UK Cancer Cancer trial (UKCTOCS) [43] analyzed more than 200.000 post-menopausal women to assess the predictive capacity of CA-125 in combination with ultrasound scanning. The results showed a modest reduction in mortality, for an estimated 20% after 14 years, where 50% of cases were detected by multimodal screening or ultrasound alone.

A mention should be made for another diagnostic serum marker that has emerged in recent years, that is the Human Epididymis protein 4 (HE4) [26]. Based on encouraging

results, Moore et al. [69] developed the Risk of Ovarian Malignancy Algorithm (ROMA) based on a scoring system that combines levels of HE4 and CA-125. This algorithm has proved useful in detecting the EOC and is becoming increasingly widespread in clinical practice for the discrimination of benign masses.

However, in the last two decades, none of these predictors impacted - in a consistent way - mortality rate reduction of this disease, because of their limited sensitivity and specificity. In fact, based on current results, widespread screening is not yet justified using the diagnostic methods presented.

In this scenario, technological advances in high-throughput technologies could improve the discovery of new screening and diagnosis biomarkers, aimed at defining the individual risk of contracting a specific disease. One of the most promising research fields relies on microRNA expression profiles (miRNA) in liquid biopsies, such as serum and plasma.

The choice of circulating molecules is motivated mainly by the two central limits that distinguish the tissue samples. First, tissue samples are difficult to acquire during patient follow-up and require invasive procedures, so the evaluation of their role in longitudinal analysis is limited. Secondly, since HGSOC is a systemic disease, the molecular portraits obtained in the ovary do not necessarily reflect those obtained from synchronous lesions in other anatomic sites [77].

Liquid biopsies are becoming a new resource for developing new biomarkers for diagnostic purposes

2.3 Circulating miRNAs in cancer

MicroRNAs are highly conserved single-stranded small RNA molecules that play a key role in post-transcriptional regulation of genes [2]. These small RNA molecules bind to the 3 'UTR region of their target mRNAs, inducing post-transcriptional regulation of the gene by inhibiting its translation or degradation [4].

Recent studies have shown that miRNAs can move (encapsulated in exons and/or bound to lipoproteins) from tumor tissue to circulation, following apoptotic and necrotic cell death or as active release [101]. As a result, these molecules can be found in a variety of body fluids (such as blood, serum, plasma, urine, saliva, seminal fluid and

pleural effusion) [18] in a stable form - protected by endogenous RNAs - thus making miRNA levels circulating well suited for minimally invasive patient analysis [66].

Although the expression levels of circulating miRNAs reflect the cumulative effects of several underlying pathways, not yet fully clarified [17], independent studies have shown, in the context of malignancy, a potential as molecular instruments for diagnosis, prognosis, and choice of treatment for various cancers.

One of the first studies that demonstrated their involvement in cancer patients was by Lawrie et al. [55]. The authors pointed out that the levels of three microRNAs (*miR-155*, *miR210*, and *miR-21*) were significantly higher in the serum of patients with large B-cell lymphoma, and among them, *miR-21* was associated with relapsed free survival of patients. Subsequently, serum miRNAs were tested as biomarkers for disease monitoring in prostate cancer, and for early diagnosis of both lungs and colorectal carcinomas. Besides, similar studies have been performed in many types of cancer including gastric, breast and ovarian cancer [111, 89, 72]. Currently, the miR-test, based on a 13-miRNA serum signature, is one of the most promising instruments for screening lung cancer in high-risk individuals [68].

2.4 Motivation problem

In the last decade, even in ovarian cancer, several circulating miRNAs have been identified as biomarkers with early diagnosis implications in association with clinical pathology and prognosis (Table 2.1). However, the assessment of the specificity and reproducibility of the individual studies reported suggests that the realization of this promise for EOC remains work in progress. Although it is impossible to directly compare the results of all studies - given the distinct questions and the small miRNAs panel they have focused on - there is a general inconsistency due to the large variety of methodological parameters that compromise evaluation.

Several are the pre-analytic and analytic aspects that may interfere with the accurate quantification of the proposed circulating miRNAs biomarkers, such as individual factors, detection platforms, independent validation and data analysis [98]. The lack of overlap between studies suggests, therefore, the urgency of further investigations on well-characterized patients with ovarian cancer and a larger panel of miRNAs, that could consistently address all technical challenges affecting the analysis of miRNA circulating.

Table 2.1 Studies on circulating miRNAs as potential biomarkers of ovarian cancer

Authors	Years	Type of tissue	Histology	Control	Up-regulated	Down-regulated	Discovery platform
Taylor et al. [95]	2008	Serum exosome	Serous	Benign ovarian tumor	*miR-21, miR-141, miR-200a,b,c, miR-203, miR-205, miR-214*	-	Custom microRNA arrays
Resnick et al. [81]	2009	Serum	EOC	Healthy control	*miR-21, miR-29a, miR-92, miR-93, miR-126*	*miR-127, miR-155 miR-99b*	TaqMan array RT-qPCR
Hausler et al. [37]	2010	Whole blood	EOC	Healthy control	*miR-30c1*	*miR-342-3p, miR-181a, miR-450-p*	Geniom Biochip
Kan et al. [46]	2012	Serum	HGSC	HOSE/healthy control	*miR-200a,b,c*	-	TaqMan assays
Chung et al. [15]	2013	Serum	Serous	Healthy control	-	*miR-132, miR-26a, let-7b, miR-145*	Microarray RT-qPCR
Zheng et al. [110]	2013	Plasma	EOC	Healthy control	*miR-205*	*let-7f*	TaqMan low-density array RT-qPCR
Xu et al. [108]	2013	Serum	EOC	Healthy control	*miR-21*	-	RT-qPCR
Hong et al. [38]	2013	Serum	EOC	Healthy control	*miR-221*	-	RT-qPCR
Ji et al. [45]	2014	Serum	EOC	Healthy control / benign ovarian tumors	*miR-22, miR-93*	-	Solexa sequencing
Shapira et al. [88]	2014	Plasma	Serous	Healthy control / benign ovarian tumors	-	*miR-106a, miR-126, miR-146-a, miR-150, miR-16, miR-17, miR-19b, miR-20a, miR-223, miR-24, miR-92a*	TaqMan Open Array MicroRNA
Langhe et al. [53]	2015	Serum	Serous	Benign ovarian tumor	-	*let-7i-5p,miR-122, miR-152-5p, miR-25-3p*	Exiqon panel RT-qPCR
Gao et al. [32]	2015	Serum	EOC	Healthy control	*miR-200c, miR-141*	-	RT-qPCR
Zuberi et al. [114]	2015	Serum	EOC	Healthy control	*miR-200a,b,c*	-	RT-qPCR
Zuberi et al. [113]	2016	Serum	EOC	Healthy control	-	*miR-199a*	RT-qPCR
Meng et al. [65]	2016	Serum exosomes	EOC	Healthy control	*miR-373, miR-200a,b,c*	-	TaqMan assay

For this reason, we propose to measure miRNA expression levels in HGSOC patients on microarray technology, to select the most promising microRNAs among the largest set of possible candidates sourced from miRNome. We collected serum of 168 HGSOC patients and 65 healthy controls, by two independent collections. We divided them into a training set, to identify candidate biomarkers, and validation set, to confirm their reproducibility. During the second phase, we used RT-PCR and ddPCR, in order to improve the accuracy and reduce the bias due to the technology used. Finally, we proposed an innovative statistical approach to normalize microarray data, designed to best fit the characteristics of the circulating molecules in the high-throughput context.

In the following chapters, all the pre-analytical and analytical aspects of the selection and analysis process are described in detail and the comments on the obtained results - with particular emphasis on statistical aspects - are presented.

Chapter 3

MATERIALS and METHODS

3.1 Serum samples collection

The following sections describe the optimized protocols used, including collection, handling, storage, miRNAs extraction and hemolysis monitoring of serum samples.

In particular, the cohort of serum samples consists of a total of 233 patients, collected by two independent Italian collections, between 2003 and 2013. The first, collected at the Department of Obstetrics and Gynecology of ASST Spedali Civili (University of Brescia, Brescia - Italy), consists of 110 patients with stage III-IV HGSOC and 52 healthy individuals of comparable age. The second, collected at the Oncological Division of Oncology of the A.Gemelli Clinical Hospital (Catholic University, Rome - Italy), is made up of 58 specimens of patients with HGSOC and 13 samples of comparable healthy individuals.

3.1.1 Sample

7.5 ml of blood were collected in S-Monovette with clot activator (Sarstedt AG & Co., Numbrecht, Germany) and centrifuged after half an hour at 3000 rpm for 10 minutes at room temperature. The serum was then aliquoted and stored at $-80°C$ within an hour. Free hemoglobin concentration was analyzed using *miR-451a* and *miR-23a-3p* expression ratio to exclude hemolyzed samples from downstream analyses [53]. All samples showed a very homogeneous level of expression ratio, markedly below level 5, demonstrating the absence of hemolysis in the collected samples.

Species	RNA Oligo	Sequence
Kaposi sarcoma associated herpesvirus	kshv-miR-K12-2	5'-rGrUrC rCrGrGrGrUrCrGrArU rCrUrG-3'
Human Citomegalovirus	hcmv-miR-UL112	5'-rCrGrG rUrGrArGrArUrCrCrArGrGrC rU-3'
Epstein-Barr virus	ebv-miR-BART8	5'-rCrGrG rUrUrUrCrCrUrArGrArUrUrGrUrArC rArG-3'
C.elegans	cel-miR-39-5p	5'-AGCUGAUUUCGUCUUGGUAAUA-3'
C.elegans	cel-miR-54-5p	5'-AGGAUAUGAGACGACGAGAACA-3'
C.elegans	cel-miR-238-3p	5'-UUUGUACUCCGAUGCCAUUCAGA-3'
Arabidopsis thaliana	ath-miR-160a	5'-UGCCUGGCUCCCUGUAUGCCA-3'
Arabidopsis thaliana	ath-miR-171b	5'-UUGAGCCGUGCCAAUAUCACG-3'
Arabidopsis thaliana	ath-miR-416	5'-GGUUCGUACGUACACUGUUCA-3'
Arabidopsis thaliana	ath-miR771	5'-UGAGCCUCUGUGGUAGCCCUCA-3'

Table 3.1 Ten synthetic spike-in RNA oligo sequences

3.1.2 Total RNA extraction

Total RNA was extracted from 200μl of serum using miRNeasy Mini kit (Qiagen, Milan Italy). In particular, serum samples were thawed in ice, then 1 ml of QIAzol Lysis Reagent (Qiagen) was added to the samples, and they were kept at room temperature for 5 minutes. Ten synthetic spike-in RNA oligos, without sequence homology to known human miRNAs, were added to samples to control for variations during the preparation of total RNA and subsequent steps (Table 3.1). The spike-in RNA oligos were introduced in serum samples as a mixture of 12.5 fmol in a total volume of 2.5 μl. All the last steps of purification were performed following the manufacturer's instructions. RNA was eluted from spin columns in 35 μl of nuclease-free water, and 15 μl were used for miRNA expression profiling.

3.1.3 miRNA profiling by microarray

We used the commercially available G4872A-046064 human miRNA Microarray (Agilent Technologies), customized with probes for the detection of specific RNA spike-in oligos. For circulating miRNA profile, we hybridized fixed volume of eluted total RNA, derived from fixed serum volumes, for all samples tested. The arrays were washed and scanned with a laser confocal scanner (G2565BA, Agilent Technologies) according to the manufacturer's instructions. miRNA microarrays underwent standard post-hybridization processing, and the intensities of fluorescence were calculated by Feature Extraction software version 11 (Agilent Technologies).

3.1.4 cDNA synthesis and Real-Time PCR

Real-Time-reverse transcription PCR (RT- qPCR) was performed starting from 5 ml of total RNA, purified and reverse transcribed into cDNA, following manufacturer's instructions (miScript Reverse Transcription Kit, Qiagen). We used a fixed volume of eluted RNA sample as input for RT-qPCR, rather than using a fixed quantity of input RNA [50]. Two microliters of cDNA were used for RT-qPCR experiments in triplicate using Rotor-Gene Thermal Cycler (Qiagen). Operations were run in triplicate and plates were prepared by automatic liquid handling station in a final volume of 10 ml (QiaAgility).

3.1.5 Digital droplet PCR

Each EvaGreen amplification mixture (20 ml) was loaded into a disposable droplet generator cartridge (Bio-Rad) and mixed with 70 ml of droplet generator oil into the QX200 droplet generator (Bio-Rad), thus portioning each sample into 20,000 nL-sized droplets. Emulsified samples were then transferred into a 96-well PCR plate to perform PCR, using a conventional thermal cycle.

The cycling steps were set as follow: 95°C for 5min,(95°C for 30 min, 58°C for 1min) 40 cycles, 4°C for 5 min, 90° C for 5 min and infinite 4°C holding. The PCR plate was then loaded into the QX200 droplet reader (Bio-Rad) for sample automated analysis. A no template control (no cDNA in PCR) and a negative control for each reverse transcription reaction (RT-neg) were included in every assay run.

3.2 Tissue samples collection

The following sections describe the optimized protocols used, including collection, handling, storage and miRNAs extraction of tissue samples.

In particular, the cohort of tissue samples consists of 76 out of 110 HGSOC patients enrolled at the Department of Obstetrics and Gynecology of ASST Spedali Civili (University of Brescia, Brescia - Italy) (see Section 3.1). As controls, 28 samples of normal ovarian and fallopian tubes are used from patients undergoing a hysterectomy and bilateral salpingo-oophorectomy for benign pathologies, in the same institution. Control samples were collected between 2011 and 2013.

3.2.1 Sample

Neoplastic tissue specimens were sharp-dissected and frozen in liquid nitrogen within 30 minutes of debulking surgery. For each sample, a specular hematoxylin-eosin section was reviewed by a staff pathologist to check for epithelial purity and only samples containing at least 70% tumor epithelial cells were used for the following RNA extraction.

Normal luminal fallopian tube epithelial cells and normal ovarian surface epithelial (HOSE) cells were collected by scraping in 1ml of physiological saline solution immediately after surgery and centrifuged at 1000 rpm for 10 minutes. The cell pellet was then resuspended in 200 μl of TRIzol Reagent (Life Technologies, Carlsbad, CA, USA) and stored at -80°C. All the normal samples were verified to be free of any neoplastic pathology before using for total RNA extraction.

3.2.2 Total RNA extraction

Total RNA from tissue samples was isolated using TRIZOL reagent (Life Technologies) and further purified using RNeasy MiniElute Cleanup kit (Qiagen), with a modified protocol for co-purification of small RNAs according to the manufacturer's instructions. RNA concentration and 260/280 absorbance ratio (A260/280) were measured with Infinite M200 spectrophotometer (TECAN). RNA integrity was assessed with RNA 6000 Nano LabChip kit using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). RNA integrity number (RIN), generated with Agilent 2100 Expert software, was superior to 8 for all RNA samples.

Only samples with good RNA yield and no RNA degradation were retained for further experiments. RNA samples were diluted at 75 ng/mul and 50 ng/μl for gene expression and miRNA expression profiling, respectively.

3.2.3 miRNA profiling by microarray

For tissue miRNA profiling, 100 ng of RNA, enriched in miRNA fraction, were Cyanine 3-pCp labeled and hybridized to the commercially available G4871A human miRNA Microarray, using a miRNA labeling and hybridization kit according to the manufacturer's instructions (Agilent Technologies).

3.3 Statistical analysis

All the analyses were made using statistical software R (version 3.3.1) [96] and several libraries, freely available at CRAN and Bioconductor repositories. In this section, we provide some essential notions and relevant references about the methods used to perform the analysis.

3.3.1 Cyclic lowess

Cyclic locally weighted smoothed spline (cyclic lowess) method was used to normalize microarray data, both for serum and tissue sample.

Lowess [16] is a normalization method for microarray data - initially conceived for dual channel data - that combines multiple regression models into a k-nearest-neighbor-based meta-model. It is based on the "M versus A" (MA) plot, where M represents the difference between the log expression values of one observation in two experiments, and A is the average of the log expression of these values. The MA plot for normalized data would show a points cloud scattered around the $M = 0$ axis.

The cyclic lowess normalizes two samples at a time by applying a correction factor obtained by estimating locally weighted polynomial regression, based on the points trend of the MA plot. A loop - or iteration - of the algorithm consists of performing this pairwise normalization on all distinct pairs of samples. For a complete outline of the algorithm, refer to Cleveland et al. [16].

A drawback of cyclic lowess is the amount of time required to normalize a set of data, in fact, the computational time grows exponentially with the number of samples. For this reason, Bolstad et al. [8] propose a fast version - that grows linearly - where each sample is normalized with respect to a reference, which is constructed as the average of all samples. This version of the algorithm is implemented in the `normalizeCyclicLoess` function of the R `limma` package [82].

In general, the lowess algorithm requires five parameters: i) the polynomial order; ii) the number of algorithmic iterations; iii) the weight function; iv) the span of the loess smoothing window, and iv) the weight of each probe. While for the first three parameters there are standard choices suggested by the literature, the last two are often arbitrarily chosen. In fact, the way they are defined is highly subject to interpretation depending on the actual data. In the next chapter (Section 4.3.2) we will describe the way we set these parameters to deal with circulating miRNAs.

3.3.2 eBayes test

Empirical Bayes test (eBayes) [91] has been used in testing for differential expression microarray data, for both the serum and tissue sample.

Although t-test is a popular choice for detecting differential expression genes (DEGs) in microarray experiments, is documented to not work well with low expression levels and outliers. For these reasons, a class of moderate t-test statistics able to address the problem of variance instability has been proposed. In particular, eBayes shrinks the probe-wise sample variance towards a common value: the obtained variance for each gene is a compromise between the gene-wise estimator - derived from the data for that gene alone - and the global variation across all genes - estimated by pooling the ensemble of all the genes [91].

The `limma` package [82] includes a robust estimation strategy for the shrinkage parameters of the model, and it is implemented in the `eBayes` function. This function has been used to rank genes in order of evidence for differential expression.

3.3.3 ROC curve

The Receiver Operating Characteristic (ROC) curves have been used to illustrate the diagnostic capacity of microRNA profiles as binaries classifier.

A ROC curve is created by representing on the y-axis the true positive rate (*sensitivity*) and on the x-axis the false positive rate (*1-specificity*), across a series of thresholds. The random choice is represented by the bisection between the first and third quadrants.

The total Area Under the Curve (AUC) is a single value used to summarize the overall performance of the test. In our case, the larger the AUC, the better is the ability of one microRNA to correctly distinguish affected and unaffected subjects. It should be stressed that two identical AUC, may have very different ROC curve. An asymptotically exact method to evaluate the uncertainty of an AUC has been proposed by Delong et al. [20].

The optimal cut-off is obtained by first calculating the distance of each observed cut-off point from the point (0.1) (i.e., the perfect classifier), and finally choosing the threshold that minimizes this distance. This criterion gives equal weight to sensitivity

and specificity and imposes no ethical, cost, and no prevalence constraints.

The package `pRoc` [83] contains a collection of tools for calculating, visualizing and comparing receiver operating characteristic curves. In particular, we used `roc` and `auc` functions.

3.3.4 Hierarchical clustering

To evaluate the discriminatory capacity of differential expressed microRNAs - between control and case - we used an unsupervised learning method. Precisely, we used a distance-based hierarchical clustering approach.

The first step in any method of cluster analysis consists of defining a rule of similarity between two objects, in this case, the samples of our experiment. The most common choice is to measure the pairwise Euclidean distance, among all possible couples of samples, and record these measurements in a distance matrix. The hierarchical clustering algorithm begins by assigning each sample to a separate cluster and then iteratively proceeds to merge the two most similar clusters at each step, as long as a single cluster is obtained. At each stage distances between clusters are recomputed by the Lance-Williams dissimilarity update formula.

In particular, the `dist` function was used to calculate the distance matrix, while for the clustering analysis the `hclust` function was used, with a complete method. Both are implemented in the `stats` R package [96].

Chapter 4

ANALYSIS and RESULTS

4.1 Workflow

The focus of this work is the detection of the levels of circulating miRNAs in sera from patients with HGSOC as a first step in the evaluation process of their role as diagnostic biomarkers.

The workflow of the circulating miRNAs analysis process is divided essentially into three phases (Figure 4.1):

(i) the cohort selection phase (both for training and validation set) (Section 4.2);

(ii) the explorative discovery phase, using microarray data (Section 4.3);

(iii) the candidates selection and validation phase, using RT-qPCR (Section 4.4).

In particular, the principal source of analytical variation in the procedure is derived from the normalization strategies related to the different used technologies, corrected adopting *ad hoc* procedures, as shown in Section 4.3.2 (for microarray data) and Section 4.4.2 (for RT-PCR data).

In the following, we present the obtained results. Moreover, we discuss in detail the pre-analytical and analytical aspects of the study which, in our opinion, represents an improvement compared to previous studies concerning this disease, and can be used as a guideline for future studies aimed at the same purpose.

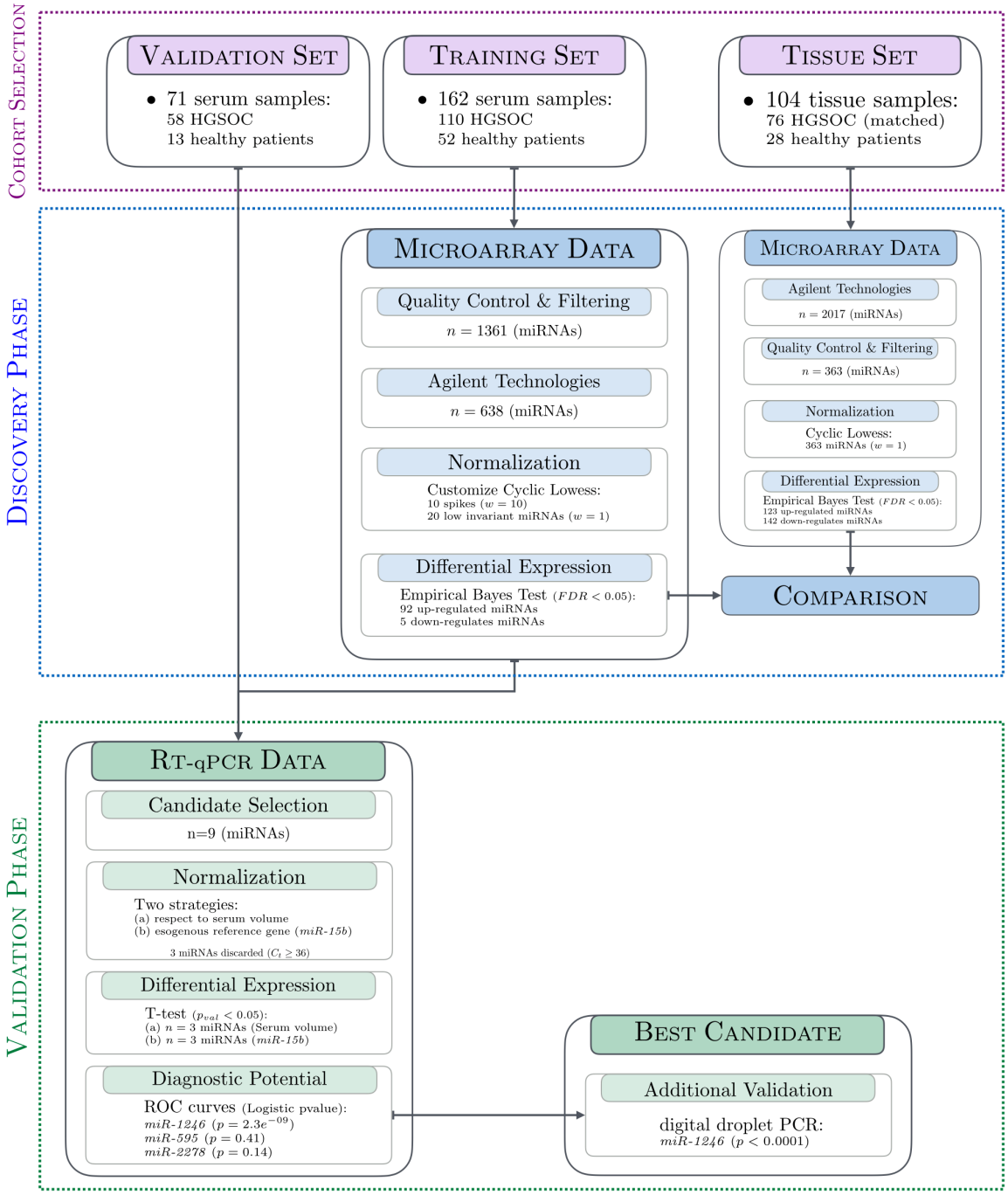


Fig. 4.1 Flowchart summarizing the different steps and the main results of the circulating miRNAs analysis for the HGSOC datasets.

4.2 Cohort selection

The main variables that may have profound implications for the accurate detection of biomarkers are those related to intrinsic inter-individual variability [98].

Facing circulating miRNAs as ovarian cancer biomarker molecules, the first source of variability to be considered is related to the tumor itself. In fact, the literature suggests that each EOC histotypes should be viewed as a distinct disease, as indicated by their differences in epidemiological and genetic risk factors, precursor lesions, patterns of spread, molecular events during oncogenesis, response to chemotherapy and prognosis [78]. So, we focused exclusively on HGSOC in patients, which is the most frequent and aggressive subtype.

Moreover, since many studies show how individual variability can contribute to affecting miRNA levels, we have created a cohort as homogeneous as possible, considering various clinical-pathological features such as age, FIGO stage, ascites, and metastasis. Refer to Supplementary Materials by Todeschini et al. [99] for a complete view of the considered variables.

In particular, serum samples were collected at the time of diagnosis before each treatment, and come from two independent Italian tumor serum collections (training and validation set), as described in Section 3.1. In contrast, tissue samples were collected at the first surgery (Section 3.2). The training set consists of 162 samples (110 HGSOC, 68% - 52 controls, 32%), while the validation set consists of 71 samples (58 cases, 82% - 13 control, 18%).

All patients were staged according to FIGO (Federation International of Gynecology and Obstetrics) guidelines as stage III-IV stage [79], with high-grade serous histological type. The median age at diagnosis was 61 and 58 years for the training and the validation set, respectively. The vast majority of women were in postmenopausal status (77% and 71% for the training and the validation set, respectively). Some patients showed the presence of ascites (82% training set and 54% validation set, respectively) and lymph node metastasis (39% for the training set and 36% for the validation set).

4.3 Discovery phase

To identify the entire repertoire of known miRNAs - sourced from miRBase version 19 - expressed in patients with stage III-IV HGSOC, a microarray profiling was performed in the training set. We chose the Agilent system as it emerged as the one of those obtaining the highest performances among hybridization-based methods [11], and it is probably the most commonly used. Accurate measurement of circulating miRNAs, through high-throughput technologies, poses different challenges due to both their short lengths and the low abundance of these molecules in body fluids.

Here, we focus on three aspects: i) filtering (Section 4.3.1), ii) normalization (Section 4.3.2) and iii) differential expression analysis (Section 4.3.3).

4.3.1 Quality control and filtering

The first step in analyzing microarray data is to filter microRNAs that are not expressed. The Agilent system contains a flag for each spot that states whether the molecule is expressed or indeterminate (*gisPosandSig* flag) and can be used to select only miRNAs with stable expression values within an array.

Serum raw microarray data comprised an initial number of 1361 miRNAs (included controls and spikes, see Section 3.1.2) repeated 40 times. The expression of these 40 replicates per miRNA has been used as technical replicates and quality control. Specifically, within each experiment, we selected only miRNAs with at least 20% of good quality measures among the 40 replicates. Otherwise, they were considered NA (not available). Then, the second filter was applied to the arrays. Specifically, we selected only miRNAs with at least 75% of good quality measures (not NA) across samples. After these filtering steps, we remain with 638 miRNAs, including the ten spikes. In this passages, we paid attention to not exclude miRNAs present (or absent) only in one of the two groups of patients considered (i.e., case and control).

Finally, the miRNA replicates within samples were summarized using the median. A few missing values still present after the filtering step was imputed with k-nearest neighborhood method. The distributions of the raw expression values are reported in the first panel of Figure 4.2.

4.3.2 Normalization

Global normalization methods developed for gene expression analysis are routinely applied to circulating miRNA expression profiling. However, the inappropriate assumptions and the reduced number of circulating miRNA profiles raises some doubts about the appropriateness of such methods.

In our study, in order to control the variability caused by experimental artifacts, we added ten exogenous synthetic microRNAs (also called spikes) in a constant amount across samples. These microRNAs are the best representation of variability due to issues such as variation in starting materials, RNA extraction, or reaction efficiency. In principle, after the normalization it is expected i) that spikes distributions were stable at high expression levels across samples and with low variability and ii) that the sample distributions be centered on the same mean.

To identify the best normalization technique, we performed a comparative evaluation using different normalization approaches: quantile [8], variance stabilization normalization (*vsn*)[33], classic cyclic loess [91] and custom cyclic loess (*CLWs*) in which spike controls are used as weights (see Section 3.3.1). Figure 4.2 shows the normalized distributions. As you can see, none of the classic standardization models (vsn, quantile or cyclic loess) is able to correct the technical variability of the experiment, represented by the colored spike lines across samples. Conversely, the use of only spikes as standardization factors - characterized by large values - leads to a bias of data transformation around high expression levels. However, in this context, an appropriate set of weights should contain both highs and low invariant expressed features.

For this reason, we included among the weights - in addition to the ten spikes - 20 invariant low expressed miRNAs, selected among those with expression values less than 3 (in log scales) and with the smallest difference in mean between cases and controls. In general, the number of normalization factors (n) will depend on the total number of profiles remained after the filtering step, and the weight values (w) rely on the noise contained in the raw data. Different combinations of parameters (n and w) have been tried but with small differences in performance (data not shown).

Summarizing, we assigned:

- high weights ($w = 10$) to ten synthetic, non-human, spike-in miRNAs to ensure the correction of technical variability introduced in the pre-analytic steps;

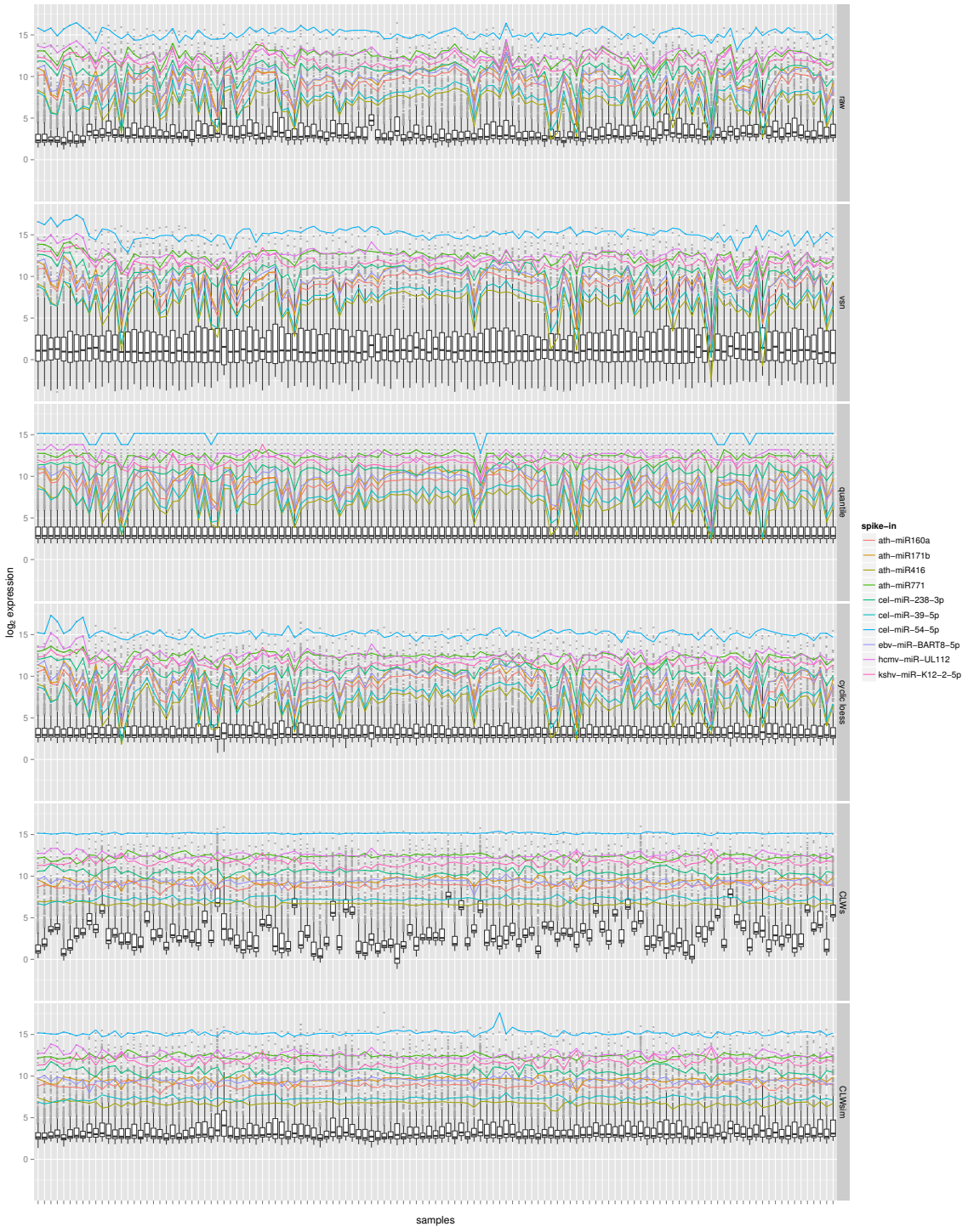


Fig. 4.2 Distribution of raw and normalized expression values with highlighted spike-in data (colored lines). The type of normalization is reported on the right gray bar of each panel.

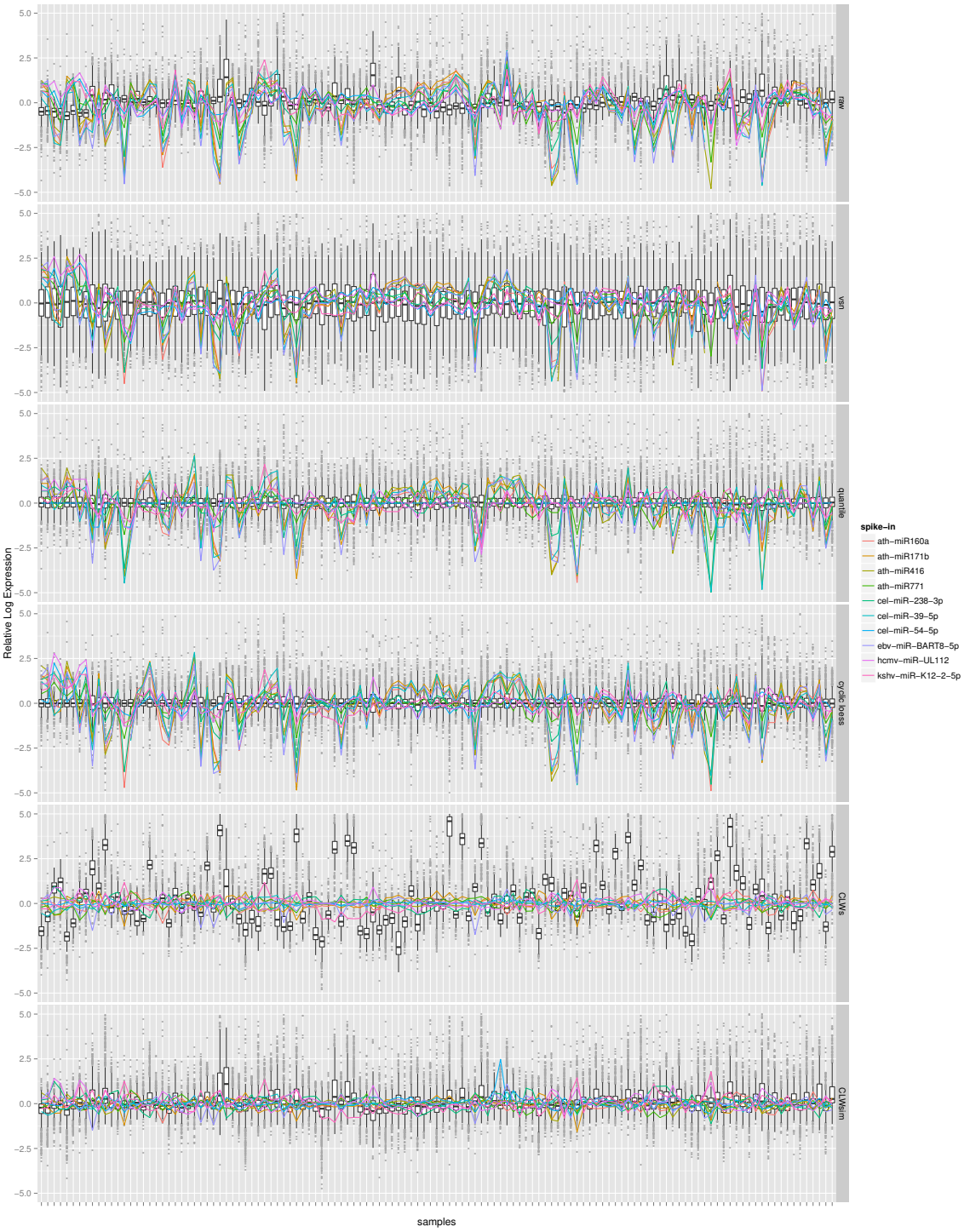


Fig. 4.3 RLE plots of raw and normalized expression values with highlighted spike-in data (colored lines). The type of normalization is reported on the right gray bar of each panel.

- low weights ($w = 1$) to 20 context-specific endogenous miRNAs to remove differences due to sampling selection and experimental quality;
- no weights for the remaining miRNAs, indeed the reduced number of features measured tend to introduce high redundancy with a corresponding removal of the interesting trends in the normalized data.

The application of this last strategy, hereafter called *CLWsim*, leads to the best compromise between a stable distribution of spikes expression values across samples and centered array boxplots, as shown in the last panel of Figure 4.2.

Another basic plot to verify the goodness of normalization is the Relative Log Expression (RLE) plot. The RLE plot shows the distribution of the ratio between the expression of a miRNA and the median expression of this miRNA across all arrays of the experiment. It is assumed that most miRNAs are not changed across the arrays, so it is expected that these ratios are around 0 on a log scale. The boxplots presenting the distributions should then be centered near 0 and have a similar spread. Another behavior would be a sign of low quality.

The RLE plots obtained after each normalization is reported in Figure 4.3. As expected classic loess, quantile, and vsn have more homogeneous RLEs, but a misguided distribution of spikes. However, CLWsim normalization has the best compromise between a stable distribution of spike expression and uniform distribution of RLE plots. Then we decide to use CLWsim approach, as the best in our case.

4.3.3 Differential expressed miRNAs

The 638 microRNAs obtained after filtering and normalization steps were used for differential expression analysis, using eBayes method (Section 3.3.2).

A total of 97 microRNAs (15%) were identified as differential expression (DEM) testing HGSOC serum versus healthy controls, 92 (95%) were up-regulated, while five (5%) down-regulated. It should be noted that, unlike gene expression analysis where an equal number of up-regulated and down-regulated genes is expected, the unbalanced finding is in line with scientific evidence on circulating miRNAs. In fact, in studies aimed at detecting cancer-related circulating biomarkers, primarily a growth in circulating miRNAs is expected, with respect to unaffected individuals [73].

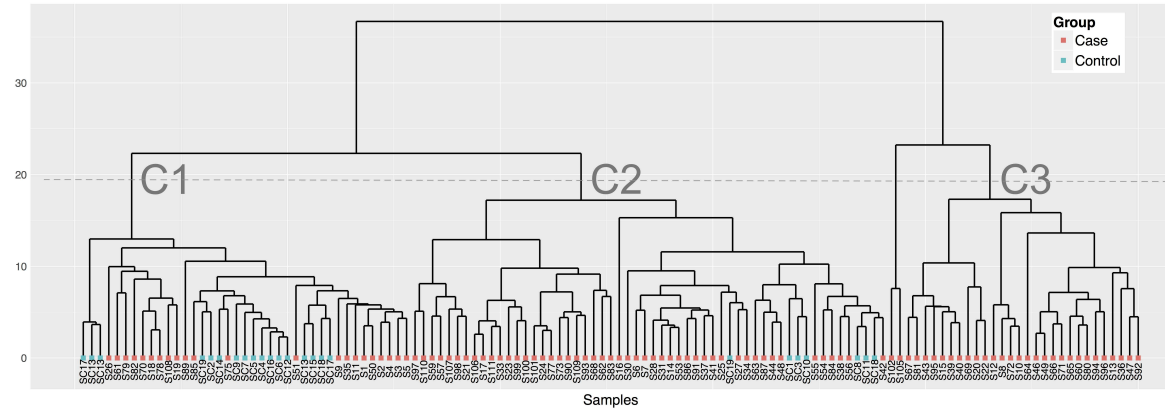


Fig. 4.4 Cluster analysis using all differentially expressed miRNAs.

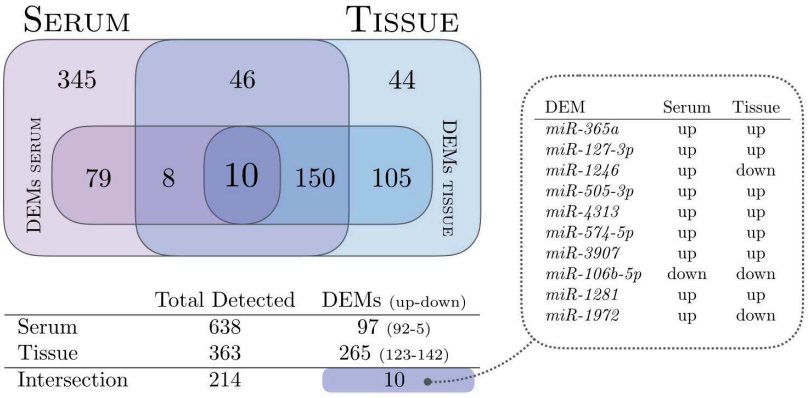
Similarity across samples was further investigated by unsupervised cluster analysis (see Section 3.3.4) using DEM expression levels. The dendrogram depicted in Figure 4.4 shows a clear separation of three groups of patients, called C_1, C_2, and C_3.

Except for seven healthy patients, cluster C_2 and C_3 are mainly composed of HGSOC patients, while cluster C_1 is largely composed of healthy controls. No significant differences have been observed regarding clinical characteristics of the C_2 and C_3 groups.

4.3.4 Tissue comparison

To investigate the origin of tissue of the selected DEMs, we examined miRNA expression profiles in those patients for whom matched serum and tissues were available (Section 3.2). To obtain consistent results, we employed the same models for both normalization and differential expression analysis in tissue data. In this case, however, given the rich concentration of expression levels and the lack of measured endogenous spikes, in the cyclic loess, the same weight was given to all microRNAs.

In particular, tissue raw microarray data comprised 2017 microRNAs. A filter has been applied to select those miRNAs with reliable expression values across arrays. Specifically, we selected only miRNAs with at least 75% of good quality measures (as *gIsPosAndSignificant* Agilent flag) across samples. After these filtering steps, we remained with 363 miRNAs. Few missing values still present after the filtering step were imputed with k-nearest neighborhood method. Classic cyclic loess and empirical Bayes test have been applied to normalize raw profiles and to identify differentially expressed miRNAs, respectively. Analysis revealed 265 DEMs (71%) - 123 up-regulated



	Total Detected	DEMs (up-down)
Serum	638	97 (92-5)
Tissue	363	265 (123-142)
Intersection	214	10

DEM	Serum	Tissue
miR-365a	up	up
miR-127-3p	up	up
miR-1246	up	down
miR-505-3p	up	up
miR-4313	up	up
miR-574-5p	up	up
miR-3907	up	up
miR-106b-5p	down	down
miR-1281	up	up
miR-1972	up	down

Fig. 4.5 The Venn diagram shows the partition of miRNAs found in serum, tissue, and how many of these resulted differentially expressed in HGSOC samples compared to healthy control (internal squared panel, DEMs). The bottom-left table shows a summary of these partitions, while the right table shows name and trend (*up* and *down* adjusted) of the ten DEMs common to both serum and tissue.

(46%) and 142 down-regulated (54%) - in HGSOC biopsies compared to 28 normal tissues.

Only ten miRNAs resulted differentially expressed in both matched tissue and serum samples (Figure 4.5). Of these, eight miRNAs shared the same trend of regulation in both tumors and sera, while two miRNAs (*miR-1246* and *miR-1972*) displayed an opposite pattern.

We further explored *miR-1246* expression trend in HGSOC versus separately ovarian and fallopian tube epithelia. Using our microarray data, we observed a significant down-regulation of *miR-1246* compared to ovarian epithelial ($p < 0.0001$), while the comparison between *miR-1246* levels in HGSOC tissues and normal fallopian tube epithelia was not significant ($p = 0.42$). These results were confirmed by RT-qPCR, using *SNORD48* as reference for proper data normalization [6].

These discordances between tissue and serum miRNAs, both regarding the lack of overlap between DEMs and the different profile trends, are in line with the results shown in Jarry et al. [44]. In fact, detected alterations in circulating miRNAs reflect the systemic response to the presence of cancer and therefore can result not only from the primer tumor but also from other types of cells, including immune cells [70].

Rank	miRNAs	Case	Control	Adjust p-value	*log*-FC
1	*miR-483-3p*	4.45	3.42	0.00001	1.03
5	*miR-4290*	4.06	3.10	0.00001	0.96
12	*miR-595*	5.03	3.54	0.0004	1.49
15	*miR-2278*	4.41	3.19	0.0005	1.22
18	*miR-32-3p*	5.19	3.37	0.0008	1.46
24	*miR-3148*	4.57	3.23	0.0008	1.34
50	*miR-1246*	7.18	6.24	0.0070	0.94
64	*miR-574-5p*	7.27	5.83	0.0154	1.44
73	*miR-4281*	11.37	10.50	0.0218	0.87

Table 4.1 Results of the microarray data analysis of the nine microRNAs emerged as best candidates among DEMs for the validation phase.

4.4 Validation phase

miRNA analysis strongly depends on both the selected cohort and the technology used in the discovery phase. To exclude the possibility that the performance of the proposed biomarkers may be related to an unknown "study bias" and the intrinsic weaknesses of the profiling method used, we decided to employ pre-operative sera from an independent cohort, and a second technology to measure miRNAs concentration.

In the following, we present the choices we made in the selection of candidates to be validated, starting from the set of DEM found in the discovery phase, and the results obtained by RT-qPCR on this collection (Section 4.4.1). Also, a study on the diagnostic potential of the best candidates is presented in Section 4.4.3.

4.4.1 Candidate selection and validation

Although the high-throughput nature of microarrays allows to analyze up to thousands of miRNAs in one assay, typically has a low dynamic range and specificity. Conversely, the quantitative reverse transcription PCR (RT-qPCR) is a well-established method, considered as the "gold standard" for miRNA profiling with high specificity, dynamic range, and sensitivity. For this reasons, it has been used during the validation phase.

However, due to the limits of PCR-based approaches, and the low abundance of miRNA species in the sera of HGSOC patients, validation experiments were performed only on a selection of DEM according to the following criteria: i) we selected only miRNAs with at least 75% of good quality measures (not NA) across samples or within patients or healthy controls, ii) highest log fold change, measured in patients compared

to healthy controls, iii) lower adjusted p-value.

Following these criteria, we identified a panel of nine microRNAs (Table 4.1: *miR-1246*, *miR-574-5p*, *miR-483-3p*, *miR-4290*, *miR-595*, *miR-2278*, *miR-32-3p*, *miR- 4281* and *miR-3148*), two of which are DEM in matched tissue samples. Each of these has been quantified by qRT-PCR, and the differential expression has been calculated on the 2^{-C_t} value, by t-test. Since there are no established endogenous miRNAs acting as a reference for serum miRNAs, the raw C_t value has been normalized to the volume of biological material. Finally, only miRNA expressions with $C_t < 36$ were considered reliable. Table 4.2 summarizes the results obtained for both training and validation set.

In the training set, *miR-1246*, *miR-4290*, *miR-595*, and *miR-2278* ($FC = 7.78$, $FC = 1.97$, $FC = 2.08$, $FC = 7.01$, respectively) are the most significantly up-regulated miRNAs in the serum of patients compared to healthy controls. *miR-574-5p* and *miR-483-3p* were not confirmed. RT-qPCR C_t values for *miR-32-3p*, *miR-4281* and *miR- 3148* resulted above the selected cut-off and therefore were discarded from downstream validation.

In the validation set, accordant to previous results, the expression levels of *miR-1246*, *miR-595* and *miR-2278* displayed a significant over-expression ($FC = 3.11$, $FC = 2.96$, $FC = 1.1$, respectively) in the serum of HGSOC patients compared to healthy controls. Conversely, *miR-4290* showed an opposite trend. Collectively, these results suggest that circulating *miR-1246*, *miR-595* and *miR-2278* in serum may serve as candidate biomarkers for diagnosis of HGSOC (Figure 4.6).

4.4.2 A note on C_t normalization

To further support the robustness of our results, we have decided to standardize raw C_t values through housekeeping.

However, housekeeping used for tissue miRNA analysis, such as RNU6, RNU48 or SNORD48 cannot be detected in circulation for their extensive RNAse-mediated degradation [106]. The most popular alternative in the literature is undoubtedly the *miR-16*, but it has recently recognized to be one of the most affected by hemolysis and therefore can not be considered a reliable reference [80]. So, due to the lack of a global consensus, independent studies have proposed other miRNAs as candidate

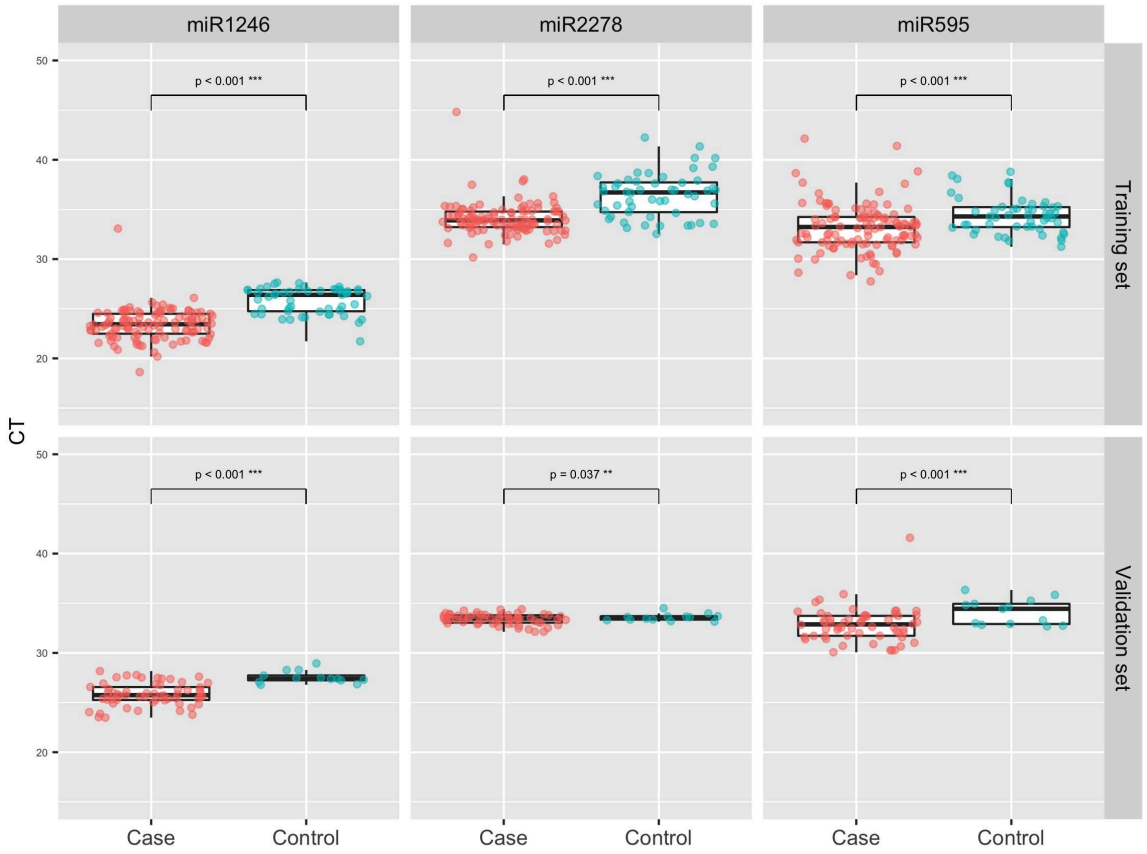


Fig. 4.6 Boxplot diagrams showing the raw C_t values of *miR-1246*, *miR-595*, *miR-2278* measured by RT-qPCR in sera of the training set (upper panel) and validation set (lower panel). *Control* group refers to healthy controls and *Case* group to HGSOC patients.

Fig. 4.7 Boxplot diagrams of the RT-qPCR C_t values in sera of training (upper panel) and validation (lower panel) sets patients, for *miR-15b* housekeeping. *Control* group refers to healthy controls and *Case* group to HGSOC patients.

	Training Set			Validation Set		
	Case (C_t)	Control (C_t)	p-value (2^{-C_t})	Case (C_t)	Control (C_t)	p-value (2^{-C_t})
miR-1246	25.83	23.41	< 0.00001	27.56	25.82	< 0.00001
miR-574-5p	28.16	28.32	0.6637	30.67	29.60	< 0.00001
miR-483-3p	32.44	32.72	0.9799	31.84	32.28	0.4145
miR-4290	33.40	32.62	0.0002	30.21	31.08	0.0344
miR-595	34.37	33.17	0.0002	34.13	32.85	< 0.0001
miR-2278	36.44	34.06	< 0.00001	33.58	33.39	0.0373
miR-32-3p	36.16	37.16	-	38.34	37.29	-
miR-4281	38.75	36.81	-	37.80	37.61	-
miR-3148	39.83	37.83	-	37.55	38.35	-

Table 4.2 Raw RT-qPCR data analysis results for each of the nine selected miRNAs. The miRNAs that showed a significant over-expression in the serum of HGSOC patients compared to healthy controls, both in training and validation sets, are reported highlighted.

	Training Set			Validation Set		
	Case (ΔC_t)	Control (ΔC_t)	p-value ($2^{-\Delta C_t}$)	Case (ΔC_t)	Control (ΔC_t)	p-value ($2^{-\Delta C_t}$)
miR-1246	0.957	0.868	< 0.00001	0.953	0.910	0.0041
miR-574-5p	1.022	1.038	0.1768	1.081	1.039	0.0068
miR-483-3p	1.200	1.214	0.3256	1.124	1.133	0.5754
miR-4290	1.221	1.191	0.0270	1.055	1.093	0.0132
miR-595	1.265	1.227	0.0098	1.197	1.155	0.0268
miR-2278	1.354	1.260	< 0.00001	1.167	1.176	0.4605
miR-32-3p	1.336	1.370	-	1.360	1.307	-
miR-4281	1.433	1.368	-	1.336	1.319	-
miR-3148	1.460	1.382	-	1.318	1.348	-

Table 4.3 Normalized RT-qPCR data analysis analysis results for each of the nine selected miRNAs. *miR-15b* has been used has reference for the normalization. The miRNAs that showed a significant over-expression in the serum of HGSOC patients compared to healthy controls, both in training and validation sets, are reported highlighted.

housekeeping. Among these, the work of Bianchi et al. [5] suggests the *miR-15b*: it resulted as the most invariant in our cohort of samples as well (Figure 4.7).

Using ΔC_t method, and *miR-15b* as a reference, we normalized the expression levels of candidate miRNAs across all HGSOC samples and healthy controls. As shown in Table 4.3, *miR-1246* and *miR-595* were also successfully validated by this further normalization approach, either in training and validation cohort of samples. *miR-4294* maintained its opposite trend. On the contrary, *miR-2278* maintained its differential expression between HGSOC and controls in the training set ($p < 0.0001$), while not in the validation set ($p = 0.461$).

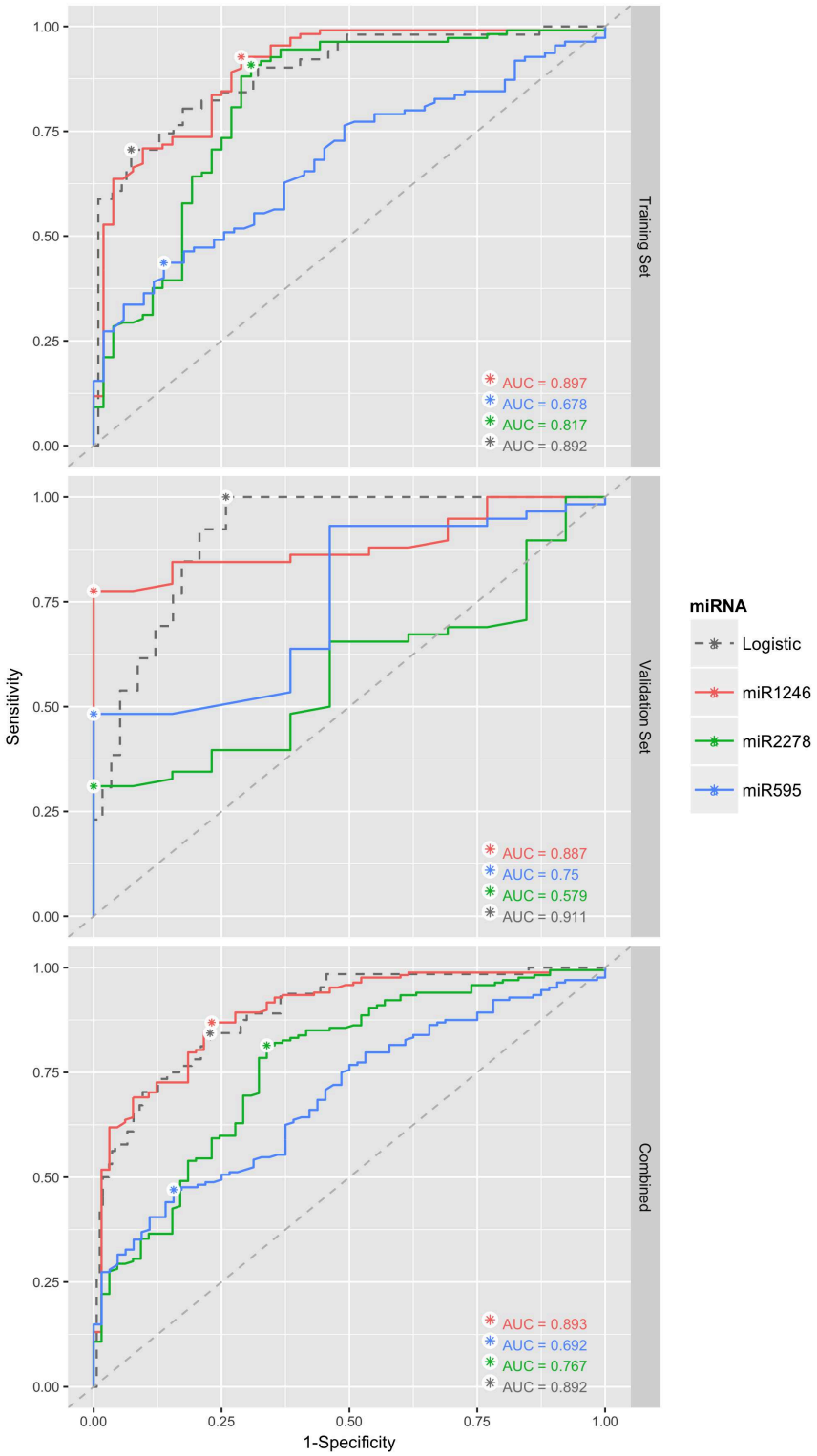


Fig. 4.8 ROC curves showing the diagnostic performance of each single miRNA markers in training and validation set, and the combination of the two of them. The curves were estimated using raw RT-qPCR C_t values. Stars indicate the combination of sensitivity and specificity with the highest AUC. In dashed dark gray, the model integrating the three miRNAs. In dashed light gray the random classification.

(a) *miR-1246* ROC performance

	Threshold	Specificity	Sensitivity	Accuracy
Training Set	24.95	0.71	0.93	0.86
Validation Set	26.77	1	0.77	0.82
Combined	0.41	0.77	0.87	0.84

(b) *miR-595* ROC performance

	Threshold	Specificity	Sensitivity	Accuracy
Training Set	32.47	0.86	0.44	0.57
Validation Set	32.65	1	0.48	0.58
Combined	-0.29	0.84	0.47	0.57

(c) *miR-2278* ROC performance

	Threshold	Specificity	Sensitivity	Accuracy
Training Set	35.38	0.69	0.91	0.84
Validation Set	33.13	1	0.31	0.44
Combined	0.26	0.66	0.81	0.77

Table 4.4 Diagnostic performance of selected miRNA biomarkers in the training set, in the validation set and in the combination of both sets. Specificity, sensitivity and accuracy have been calculated at the selected threshold as described in Section 3.3.3.

4.4.3 Diagnostic potential

As we can observe (Figure 4.8), all three miRNAs perform best than the random classification of patients (gray dotted line). In particular, for *miR-1246* (Table 4.4a) the sensitivity is 87%, the specificity is 77%, and the accuracy is 84%, with an AUC (Area Under the Curve) of 0.89. For *miR-595* (Table 4.4b), the sensitivity is 47%, the specificity is 84%, and the accuracy is 57%, with an AUC of 0.69. For *miR-2278* (Table 4.4c), the sensitivity was 81%, the specificity is 66%, and the accuracy is 77%, with an AUC of 0.76.

Then, we tested the diagnostic value of the integration of the three biomarkers, using multivariate logistic regression. We found that *miR-1246* remains the strongest biomarker ($p = 2.3e^{-09}$), while *miR-595* ($p = 0.41$) and *miR-2278* ($p = 0.14$) resulted to be not significant. Moreover, the combination of the three biomarkers resulted in a moderate increase of AUC only in the validation set.

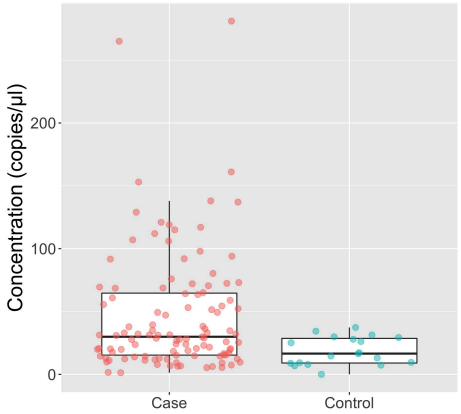


Fig. 4.9 Boxplot of the absolute quantification of *miR-1246* by ddPCR in HGSOC patients compared to controls, for both training and validation sets. Results are presented as copies per microliter of the amplification reaction mixture.

These results indicated that *miR-1246*, which showed the greatest ability in differentiating HGSOC patients from controls, could act as a suitable biomarker for detecting HGSOC patients.

4.4.4 Additional validation

As miR-1246 appeared the most promising diagnostic serum biomarker, we decided to validate its expression levels with an additional and more sensitive technique, EvaGreen-based ddPCR technology. The quantification by ddPCR expressed as copies/μl (Figure 4.9), confirmed the diagnostic potential of miR-1246 ($p < 0.0001$) in discriminating HGSOC patients and healthy controls.

Chapter 5

CONCLUSIONS

Considering the lack of ovarian cancer screening tests that can significantly reduce patient mortality, developing new strategies for early diagnoses, such as identifying new biomarkers, is one of the possible strategies to be pursued. Serum tumor biomarkers are currently considered one of the best tools to improve early diagnosis, help to predict prognosis and possibly therapeutic response. They are particularly relevant for a neoplastic disease such as ovarian cancer, which is often asymptomatic at its start and tissue samples are not always accessible during clinical follow-up.

In recent years, circulating miRNAs have been discovered and found highly stable in a variety of body fluids that can be minimally invasive. Although the expression levels of circulating miRNAs reflect the cumulative effects of different negative pathways, which have not yet been fully elucidated, the levels and composition of miRNAs in blood, serum or plasma have been found to reflect the presence of several malignant diseases. Many technical challenges in the analysis of circulating miRNAs complicated the comparison of independent data sets and delayed their entry into clinical environments.

In the presented study, published in *Cancer Letters* journal [99], we used microarray technology to achieve effective selection of the most promising miRNAs among the thousands of possible candidates coming from miRNome (miRBAse version 19). In addition, we have developed a new bioinformatic approach to identify specific circulating miRNAs that characterize HGSOC patients.

The miRNA profiles of the training set initially allowed us to identify 97 miRNAs with different levels of expression (DEM) between HGSOC patients and healthy controls. In line with previous studies, we found a modest overlap between miRNA expression

pattern in serum and tissue [44, 111], suggesting that circulating miRNAs could derive from a contribute of inflammation-related and tumor-specific miRNAs, selectively and actively secreted through microvesicles and exosomes as a novel mechanism of genetic exchange between cells.

Among DEMs, nine of these were further validated in a completely independent data set, of which *miR-595* and *miR-1246* were confirmed in both. The ROC curve confirmed *miR-1246* as the most promising diagnostic biomarker, as it was able to accurately classify cancer patients concerning healthy controls, both in training and validation cohorts.

Currently, serum CA-125 is the most frequently used biomarker for EOC detection, showing the best performance in advanced-stage HGSOC, while exhibiting both a low specificity and sensitivity to detect early-stage disease. Consequently, prospective studies on a larger cohort of serum samples are needed either to test *miR-1246* potential clinical utility in late-stage HGSOC or to assess its value in early-stage diagnosis. Although a detailed biological analysis of *miR-1246* is far from the scope of this study, there are some data previously reported in the literature that is warranted to be discussed in detail. Its expression has been largely reported as upregulated in various cancer tissues [36, 49], and as circulating marker, it has been proposed for the detection of several carcinomas [75, 89]. Moreover, *miR-1246* has been associated with stemness in non-small cell lung cancer [49] and pancreatic carcinoma [36]. According to our knowledge, *miR-1246* has not previously been associated with ovarian cancer, either at tissue level or serum level. Indeed, despite the abundance of published articles on miRNA circulating in the diagnosis of ovarian cancer, there is a high level of inconsistency in the studies. Pre-analytical and analytical challenges in circulating miRNA experiments, data analysis and normalization, statistical power, and validation of results, are the principal causes of this poor overlap in the outcomes.

Within this complex scenario, we believe that our study displays several improved features compared to previous studies, including:

(i) the focus on HGSOC, the most frequent and aggressive ovarian carcinoma subtype;

(ii) optimized protocols including collection, handling, hemolysis monitoring, storage and miRNAs extraction of serum samples;

(iii) the inclusion of two cohorts of HGSOC patients and controls, gathered from independent serum collections;

(iv) the use of an innovative and effective statistical approach of microarray data normalization, combining synthetic spike-in RNA oligos and the most invariant endogenous miRNAs;

(v) the use of two RT-qPCR techniques for miRNA validation and, in particular, of Exiqon primer sets with LNA technology, which maximizes sensitivity and specificity in detecting miRNA amplicons.

This rigorous approach makes us confident in our results, reporting *miR-1246* as a novel diagnostic biomarker in HGSOC. Moreover, we believe that the presented experimental design can be used as a guideline for other studies concerning circulating miRNAs.

In fact, we are currently adopting the same strategy in a second study of saliva samples from a cohort of Head and Neck Squamous Cell Carcinoma (HNSCC) patients. In particular, also in saliva samples, the proposed customized normalization approach showed the ability to correct confounder variables. Further evaluations on the list of differentially expressed miRNAs, should be made to confirm the validity of the results and to identify possible biomarkers for HNSCC.

Part II

Primary genes detection in perturbed biological pathways

Chapter 6

INTRODUCTION

6.1 Omics data and system biology

In the last two decades, biology has become a "big-data" science. With the first bacterial genome sequenced and the consequent optimism generated [31] by the Human Genome Project [103, 52], we have seen the begginning of an era characterized by the extensive collection of biological information and the consequent birth of the so-called "omics" sciences.

Modern high-throughput technologies (HTs) - such as genomics, proteomics, and transcriptomics - are providing tons of data that grow in a multidimensional way over time and lead to more and more detailed information. In fact, HTs allow researchers to perform complete measures of the molecular status of biological samples and furnish lots of information on gene association with particular phenotypes.

Regardless of the technology used, high-throughput data analysis typically yields long lists of genes or proteins whose expression change in different experimental conditions (DEGs) [59]. There are several univariate statistical methods to establish the differential expression of these molecules, such as t-test, non-parametric tests, and Bayesian models [28].

These lists are handy for identifying genes that may have a role in a given phenotype. However, for many investigators, they don't often provide mechanistic insights into the biological condition contributing to a limited understanding of complex diseases.

Indeed, the adaptable nature of living systems constitutes a significant challenge to derive accurate and predictive models from genomic data. Because cellular processes are governed by networks of molecular interactions, critical alterations to these systems

may arise at different points, yet result in similar phenotypes. Typical gene-level analyses of HT data, such as tests of differential expression, are unable to capture these effects [10].

There is growing consensus that the genetic risk for a complex disease is mainly due to multiple genes with small but coordinated interactions that act in a modular fashion, rather than by the differential expression of a single gene [97]. As a result, the advent of HTs has increased the interest in the systems-level analysis of genomic data.

6.2 Gene Set Enrichment Analysis

One of the most promising and widely used computational approaches analyzing data coming from HTs is the Gene Set Enrichment Analysis (GSEA). To reduce the complexity of the analysis, the idea is to move from the gene-to-gene view, towards the analysis of a group of functionally related genes. Researchers have developed a large number of knowledge bases to facilitate this task. These knowledge bases describe biological processes, components, or structures in which individual genes are involved, and how and where gene products interact with each other. An example of this idea is to identify groups of genes that work in the same pathway.

The pathway-based analysis (PA) - also known as functional enrichment analysis - is desirable for two reasons. First, grouping thousands of genes, proteins and/or other biological molecules reduces complexity to a hundred pathways per experiment. Secondly, identifying pathways that differ between two conditions may increase the explanatory power of the simple list of genes or proteins. This procedure has shown a better understanding of the molecular mechanisms that cause complex diseases [106].

The development of PA techniques has been made possible by the growth of databases that describe the functional networks of interactions. Among these are KEGG [47], Reactome [29], BioCarta [71], NCI Pathway Interaction Database (NCI-PID) [86] and WikiPathways [51]. Also, to address the challenge of querying these databases using a common framework, Markup languages such as KGML and BioPAX [21] have been developed to describe pathways using a consistent format. Finally, several tools have been proposed for the conversion of a biological pathway into suitable graphical and mathematical structure [84, 104, 109].

Over the past decade, a considerable number of methods have been developed to integrate this information, both in the univariate and multivariate case. According to the statistical test adopted and the hypotheses tested, PA methods can be divided into two broad categories. The first group comprises models designed to identify pathways in which significant genes are overrepresented. Some examples are the GSEA [93] and the Ingenuity Pathway Analysis. The second category of methods is based on global and multivariate approaches that summarize the variation of the genes across the pathway and test for pathway level differences, without relying on single gene association statistics. Goeman and Buhlmann [34] term these approaches competitive and self-contained, respectively.

Both approaches have their limitations. On the one hand, competitive methods assume the independence among genes and use a stringent cut-off for the selection of DEGs, leading to a reduction in statistical power. On the other hand, global and multivariate approaches relax the assumption of independence of genes from the same pathway - identifying moderate but coordinated differences - but tend to have high power, leading too many significant tests. For the latter, the number of replicas in experiments is often too low for multivariate models.

6.3 Topological Pathway Analysis

Despite these advances, most of PA methods manage pathways as a simple list of genes, ignoring the fact that they are structured in a network with explicit interactions. Moreover, the contribution of the network topology to biological functions has long been appreciated and proven [92]. So, even if our understanding of biological functions is continually improving - and pathways are regularly updated by adding, removing or re-mapping links in the diagrams - as long as they involve the same set of genes, they will produce identical results. In recent years, efforts have been made to consider topological information within self-contained methods, thus seeing the emergence of the third generation of models, called Topological Pathway Analysis (TPA).

The seminal paper of Draghici et al. [25] proposes one of the first approaches (Impact Analysis, SPIA [94]). They attempt to capture and combine two different aspects of data: i) the overrepresentation of DEGs in a given pathway through fold-change and ii) the abnormal perturbation of the pathway, measured by propagating

expression changes across its topology. Since then, this approach has become very popular, resulting in the publication of several pathway topology based algorithms [48].

Other attempts to incorporate topological information into pathway analysis make use of graph theory methods. Isci et al. [40] propose a Bayesian pathway analysis that models each biological pathway as a Bayesian network; Vaske et al. [102] use a probabilistic graphical model framework for learning the underlying causal networks compliant; Jacob et al. [41] develop a graph-structured two-sample test for means. Finally, Massa et al. [62] introduce a Gaussian graphical model approach that examines both differences in means and in covariance matrices between two experimental conditions.

Although it's hard grouping these methods into macro categories because of the heterogeneity of the proposals, they can be categorized by different criteria, including the type of input required, the mathematical model used, the chosen implementation, and the provided output [67]. From the output, the result is typically a list of ordered pathways.

In most of the methods described above, individual values that represent gene expression are combined - following the pathway-defined internal structure - into a single global score that results in activation/deactivation. Then, the pathways are used as whole functional units in the interpretation of phenotype association experiments. However, stating that a pathway is activated (or deactivated) is not very informative by itself. Indeed, partial activation (or deactivation) within the same pathway can have very different and sometimes opposite biological implications [87].

For this reason, some methods propose a subsequent refinement of the analysis by identifying sub-networks consistent with the condition under study [42, 90, 63]. This ability is essential when the pathway contains hundreds of genes. In fact, the significance of a large pathway could be misleading, hiding significant parts that are mostly involved in the biological process under exam.

Some of the most recently proposed approaches seek to model significant sub-pathways as signal paths or perturbation chains that can guide systematic differences. They can result in a more detailed and realistic description of functional consequences of gene up and down regulations within the context of each pathway.

Li et al. [57], attempt to identify biologically significant pathways by using a minimum spanning tree and an extension of the SPIA mentioned above. In this case, the signal is represented by differentially expressed and related genes through

linear structures. Instead, Vrahatis et al. [105] use the union of pathways to extract differentially expressed gene modules characterized by five specific topological structures (components, streams, cascades, and neighborhoods).

From an entirely different angle, Sebastian-Leon et al. [87] adopt a probabilistic model in which gene expression measurements are used to calculate the probability of activation of stimulus-response circuits within pathways. Finally, Martini et al. [61], use chains of modules (cliques), obtained through a structure - called junction tree - to identify the signal path most associated with the phenotype.

6.4 Motivation Problem

Although authors tend to interpret genes that are part of perturbed sub-graphs as disease-causing biomarkers and likely targets for therapy, this may not be true. Indeed, all above mentioned models, using marginal approaches, are unable to distinguish between the genes that are the real sources of perturbation (for example due to mutations, number variations, epigenetic changes, etc.) and those that merely respond to signal dysregulation, due to the so-called network propagation.

Let us take the example of intervention studies, such as knock-out and knock-in. Here, the expression of a single gene is altered. Marginal approaches will identify all pathways or sub-pathways that include the gene - as a consequence of the intervention - without giving any information about the direct target of the intervention. For example, although Ansari and colleagues [3] propose a formulation that explicitly distinguishes between "primary disregulation" and "secondary disregulation", are incapable of identifying the knock-down gene. They merely use this information to improve pathway perturbation scores. Also, although in the intervention studies the gene that generates the phenotype is known, it may be not annotated in any pathway, or its function may be undiscovered. In this case, the identification of primary regulators can help the experimenter to identify the direct targets of the perturbed gene. Another compelling application - which goes in the direction of precision medicine - is *case-control* studies: the identified set of genes will be responsible for the differences that cause the observed perturbation, and then may be the upstream regulators.

To fill this gap, we propose to pursue the identification of the set of variables driving the difference in different experimental condition (i.e., the *primary genes*) within a graphical model context [23]. We present a new way to address this question, which

uses the idea of simultaneously looking at the differences between two multivariate normal distributions (i.e., two phenotypes) in all marginal and conditional distributions implicated by Markov properties, associated with a non-directed and decomposable graph representing the pathway under exam. This model can be thought as the natural extension of the method proposed by Massa et al. [62] for identifying primary genes. We had also implemented the algorithm and made it available as an R package with additional graphical devices, to guide the user in the exploration of research findings.

In the first chapter (Chapter 7), we present all the fundamental notions regarding the proposed method. A deepening discussion about the used strategies to overcome the limitations inherited by high-dimensionally and the multiple tests issue can be found in Section 7.3. The aspects related to the implementation of the algorithm are listed in Chapter 8. Finally, the evaluation of the model performances on both *in-silico* and real biological data are presented in Chapter 9.

Chapter 7

METHODS AND RATIONALE

7.1 Theoretical Background

When a set of genes functionally related to gene set categories (i.e., pathways) are flagged to be associated with a certain phenotype, the fundamental biological question is to identify - among the genes involved - those that are potentially responsible for differential behavior to determine possible biomarkers or therapeutic targets.

An approach based on graphical models, allowing to search for both differential expression and co-expression behavior, was proposed by Massa et al. [62]. After converting the pathway to an appropriate graph where the nodes and edges represent genes and biochemical interactions respectively, the authors assume to model the data through two Gaussian graphical models with the same undirected graph.

Within the context of graphic models, data is considered as coming from Gaussian multivariate distributions with a structured concentration matrix (inverse of the covariance matrix), which reflects dependencies between the variables derived from the pathway topology conversion. Formally:

$$(X_1^{(1)}, ..., X_p^{(1)}) = \{Y \sim \mathcal{N}_p(\mu^{(1)}, \Sigma^{(1)}), (\Sigma^{(1)})^{-1} \in \mathcal{S}^+(G)\}$$
$$(X_1^{(2)}, ..., X_p^{(2)}) = \{Y \sim \mathcal{N}_p(\mu^{(2)}, \Sigma^{(2)}), (\Sigma^{(2)})^{-1} \in \mathcal{S}^+(G)\}$$

where p is the number of variables (number of genes) and $\mathcal{S}^+(G)$ is the set of symmetric positive definite matrices with null elements corresponding to the missing edges of G.

The authors can formulate a topological approach to gene set analysis that transforms the encoded dependency structure contained in pathways into undirected graphs

and models them as Gaussian graphical models. In this context, the assessment of whether the pathway expression changes in different experimental conditions naturally fits within the hypothesis tests on the mean and variance/covariance parameters for which the inference procedures are well characterized [54]. Which is:

$$H_0 : \Sigma^{(1)} = \Sigma^{(2)} \text{ vs } H_1 : \Sigma^{(1)} \neq \Sigma^{(2)}$$
$$H_0 : \mu^{(1)} = \mu^{(2)} \text{ vs } H_1 : \mu^{(1)} \neq \mu^{(2)}$$

The first one tests if the strength of the connections among genes is altered in different experimental conditions; the second test is more traditionally designed for testing differential expression, nevertheless in a multivariate setting.

Once the hypothesis of equal distribution is rejected, the pathway is defined as perturbed (activated or deactivated). The modular nature of Gaussian graphical models allows identifying the portion of the graph that is mostly associated with the phenotype under study. So, authors partition the graph into smaller units - the so-called maximal cliques - and tests are performed for each of them. In fact, when the underlying graph is decomposable, cliques induce saturated models that can be tested separately and compared marginally.

The main criticism of the method proposed by Massa et al.[62], concerns the dependence of the tests associated with the cliques being these non-disjoint and overlapping sets of variables. It is common for biological networks that some nodes have a degree of connectivity higher than others (i.e., hub genes): when the pathway is translated into a graphical structure, these nodes tend to be in almost all cliques, dramatically amplifying the dependence of the tests.

This issue can be addressed by decomposing the global test in moving from marginal-based to conditional-based approach, to obtain independent hypothesis. Although the conceptual framework needed to achieve this formulation has already been introduced in the seminal work of Dawid and Lauritzen [19] - under the name of Hyper Markov Laws - the implementation of this theory has never been applied in the biological context. The ultimate goal is to provide the researchers with a tool that allows to zoom on the potential sources of differential behaviors, identify the causes, and translate results into biological hypotheses that can be experimentally validated.

Determining the set of variables that are the actual source of the differences between two phenotypes - what we will call from now on *source set* - is the purpose of the proposed method.

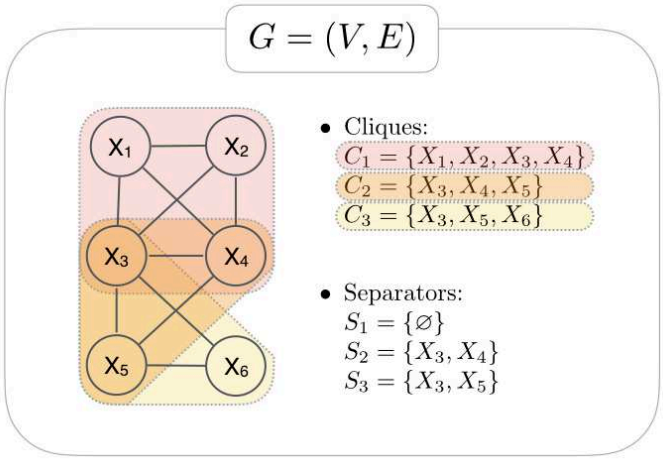


Fig. 7.1 Decomposable graph G consisting of six nodes ($|V|$) and three cliques (k).

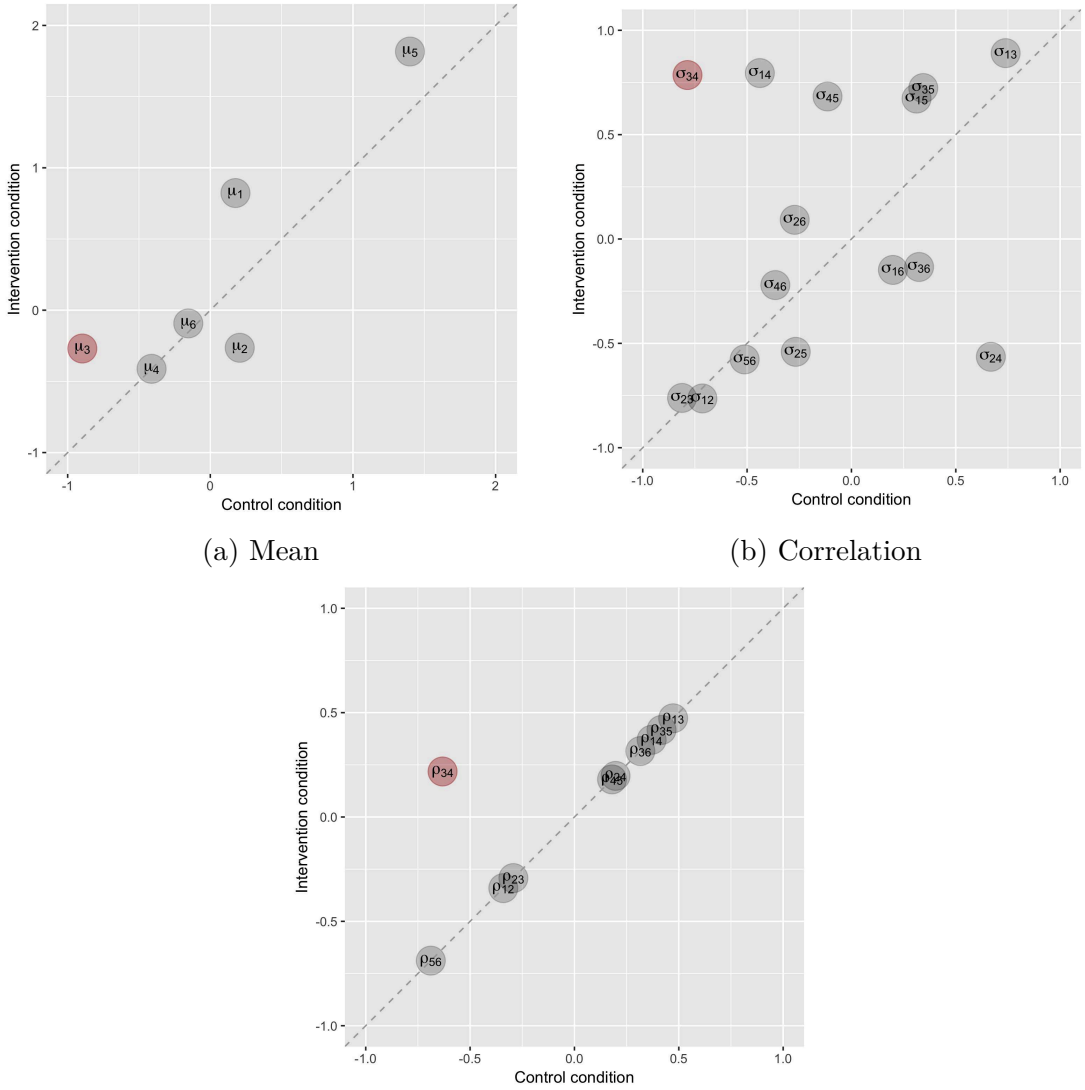
7.2 The source set

In this chapter, we will introduce the idea of *source set* - the set of variables that are the true source of the differences observed between two conditions - through a toy example (Section 7.2.1). This example will serve to fully understand the difference between using a marginal approach rather than conditional, and it will be utilized throughout the chapter to introduce different concepts. We will formalize this notion in (Section 7.2.2). Finally, we will show how the modularity of the graphical models can be used to decompose the global hypothesis of equality of two distributions (Section 7.2.3) into a set of local independent hypothesis that can be exploited to estimate the source set (Section 7.2.4). Some typical issues of inference on high-throughput data, such as small sample size and multiple testing, are addressed in (Section 7.3).

All the fundamental notions regarding the graphical models, relevant to the understanding of the following paragraphs, can be found in Appendix A. For a detailed view, see Lauritzen [54]. All the theorems, the demonstrations and further details regarding the source set are contained in the work of Djordjilović et al. [23]. Instead, for information on pathway topology conversion into an undirected decomposable graph, refer to Massa et al. [62]. For a more detailed view of Sales et al. [84].

7.2.1 Marginal and conditional distribution

Let us consider the graph $G = (V, E)$, represented in (Figure 7.1), and a vector of causal variables $(X_1, ..., X_6)$ with the associated normal multivariate distribution $\mathcal{N}_6(\mu, \Sigma)$,



(a) Mean

(b) Correlation

(c) Partial correlation

Fig. 7.2 Mean (a), correlation (b) and partial correlation (c) parameters in control and intervention conditions. Parameters directly affected by the intervention are highlighted (red). Line $y = x$ added for reference (gray dashed line).

where $\mu \in \mathbb{R}^6$ is arbitrary, and $\Sigma^{-1} \in \mathcal{S}^+(G)$ obeys to the conditional independence encoded in G. Two variables are said to be conditionally independent when their partial correlation - that is, the corresponding element of Σ^{-1} - is zero, and G is an undirected decomposable graph. We can think of $X_1, \ldots, X_6$ as the expression levels of six genes, and G as the dependency structure among these in a given pathway.

We assume that variables X_3 and X_4 are the true sources of perturbation (i.e., due to mutations, epigenetic changes, etc.), while the remaining variables - X_1, X_2, X_5 X_6 - merely respond to the perturbation of the signal. More specifically, compared to the control condition, the mean of X_3 decreases by 70%, and as a result, the partial correlation between X_3 and X_4 reduces in turn. In other words, the intervention influences the mechanism underlying the joint distribution by acting on the two variables, but it leaves unaltered the conditional distribution of the remaining variables (Figure 7.2c). It affects all the marginal distributions considered, both for the mean and for the variance matrix, as illustrated in (Figure 7.2b) and (Figure 7.2a).

We generate a random sample of size 100 for each condition (i.e., before and after intervention on variable X_3 and X_4). Following the strategy presented in Section 7.1, we decompose the graph into its maximal cliques and perform the marginal test of equality of distributions on each of these. Note that in this case we are considering the null hypothesis $H_0 : \Sigma^{(1)} = \Sigma^{(2)}$ and $\mu^{(1)} = \mu^{(2)}$. As expected, all cliques are highly significant (Table 7.1), in fact, they all contain variables on which we intervened. Since all marginal distributions between the two conditions are different, we would conclude that the condition under study has an impact on all variables. Although correct, this view fails in identifying the special role of the variables X_3 and X_4, which are the primary genes of the perturbation. If we adopt the terminology proposed by Ansari et al. [3], we are not able to distinguish between the "primary dysregulation" of a given set of genes itself and the effect of signal propagation, that is the "secondary dysregulation". Moreover, the authors assume that it is the leading cause of the large number of false positive which the TPA methods proposed so far presently facing.

7.2.2 Definition

Resuming the idea shown in the previous paragraph, we can give a more formal definition of the set of our primary genes. Suppose we have p variables, and two normal random vectors $X^{(1)}$ and $X^{(2)}$, which represent the distributions of the variables in the two experimental conditions.

Element	A	$\lambda(A)$	Gdl	p-value
C_1	$\{X_1, X_2, X_3, X_4\}$	231.95	14	1.5×10^{-41}
C_2	$\{X_3, X_4, X_5\}$	225.58	9	1.4×10^{-43}
C_3	$\{X_3, X_5, X_6\}$	48.93	9	1.7×10^{-7}
S_2	$\{X_3, X_4\}$	224.71	5	1.4×10^{-46}
S_3	$\{X_3, X_5\}$	43.74	5	2.6×10^{-8}

Table 7.1 Test of equality distribution of cliques and separators induced by G. Significant p-values are highlighted.

Definition 1. *We call the set D the minimal* source set*, if:*

1. *the distribution of $X_D^{(1)}$ differs from that of $X_D^{(2)}$;*
2. *the conditional distributions $X_{\bar{D}}^{(1)}|X_D^{(1)}$ and $X_{\bar{D}}^{(2)}|X_D^{(2)}$ coincide*

where D is a subset of the p variables, and $\bar{D}$ its complementary.

In our toy example, the source set D is equal to the set (X_3, X_4). In fact, the distribution of the set of variables $\bar{D} = \{X_1, X_2, X_5, X_6\}$ conditioned to the source set will be the same in the two conditions, for construction.

A naive strategy to identify the set D from data would require testing all possible subsets of V, but the number of potential source sets grows with the power of p, making the search space too large for many practical applications. This definition is general and does not refer to the graph structure.

However, when comparing two normal distributions, we can take advantage of the decomposable graphs and the modularity implicated by hyper Markov properties to obtain a linear solution.

7.2.3 Decomposition of the global hypothesis

Let $G = (V, E)$ be a decomposable undirected graph on p vertices. Let $C_1, ..., C_k$ be a sequence of its cliques satisfying the *running intersection property*, and let $S_2, ..., S_k$ be an associated sequence of *separators* (for definition, see Appendix 7.4). Let's remember that a clique is a complete maximal subgraph, that is not a subset of any other complete subgraph, and a subgraph is said to be complete if all its vertices are connected to G.

Let $X_1^{(1)}, ..., X_{n_1}^{(1)}$ and $X_1^{(2)}, ..., X_{n_2}^{(2)}$ be two random samples from two multivariate normal distributions $X^{(1)} \sim N_p(\mu^{(1)}, \Sigma^{(1)})$ and $X^{(2)} \sim N_p(\mu^{(2)}, \Sigma^{(2)})$ and consider the

hypothesis of equality distributions

$$H_0 : \Sigma^{(1)} = \Sigma^{(2)} \text{ and } \mu^{(1)} = \mu^{(2)} \tag{7.1}$$

and the hypothesis of equality distributions for the satured marginal model induced by the subset of variable $A \subseteq V$

$$H_A : \Sigma_A^{(1)} = \Sigma_A^{(2)} \text{ and } \mu_A^{(1)} = \mu_A^{(2)} \tag{7.2}$$

Theorem 1. *Let $\lambda(V)$ denote the log likelihood criterion for testing (7.1) and $\lambda(A)$ the log likelihood criterion for testing (7.2). Thanks to the Hyper markov laws associated to the graph G, the global hypothesis of equality of $X^{(1)}$ and $X^{(2)}$ decomposes into a set of independent local tests, as follow:*

$$\lambda(V) = \lambda(C_1) + \sum_{i=2}^{k} [\lambda(C_i) - \lambda(S_i)] \tag{7.3}$$

Moreover, under the null hypothesis, the k terms on the right hand side are asymptotically independent chi-squared distributions.

At this point, it is necessary to explain the interpretation of the k components of the decomposition (7.3) to clarify how this formulation plays a key role in estimating the source set. The first term $\lambda(C_1)$ corresponds to the log likelihood ratio (LLR) criterion for testing the equality of the marginal distributions of the C_1 clique, i.e. X_{C_1}; while the $(k-1)$ terms on the right, correspond to the LLR for testing the equality of the conditional distribution of the variable belonging to C_i, given the variable belonging to the associated separators S_i, i.e. $X_{C_i \setminus S_i} | X_{S_i}$. Note that, thanks to this decomposition, we do not explicitly estimate any conditioned dependence, but we only need to perform marginal tests in small models induced by cliques and separators, associated to the graph G.

Looking at the toy example, graph G consists of three cliques and two separators. A proper sequence is (C_1, C_2, C_3) and the associated sequence of separators is (S_2, S_3). Following (7.3), the global statistics can be decomposed as:

$$\lambda^{C_1}(V) = \lambda(C_1) + [\lambda(C_2) - \lambda(S_2)] + [\lambda(C_3) - \lambda(S_3)]$$

where the first component tests the equality of the marginal distribution of X_1, X_2, X_3, X_4, the second tests the equality of the conditional distribution of X_5 given (X_3, X_4), and

the third of X_6 given $(X3, X5)$. We are dealing with different partitions of the set of vertices and each variable enters a single time in the hypothesis system. It should be noticed that such an ordering naturally leads to avoid the dependence among tests. However, despite the fact that graph G determines the set of cliques and separators, the orderings that satisfy the *running intersection property* are not unique.

Given the uniqueness of the separators, it is easily demonstrable that there is precisely one decomposition for each choice of the root clique of the sequence (i.e., the clique that corresponds to the marginal test of the decomposition) leading to a total of k decompositions, where k is the number of cliques. Other permissible, and equally valid, decompositions, are:

$$\lambda^{C_2}(V) = \lambda(C_2) + [\lambda(C_1) - \lambda(S_2)] + [\lambda(C_3) - \lambda(S_3)]$$
$$\lambda^{C_3}(V) = \lambda(C_3) + [\lambda(C_2) - \lambda(S_3)] + [\lambda(C_1) - \lambda(S_2)]$$

Since each of these orderings corresponds to a different factorization of the same distribution, this multiplicity can be exploited to obtain a set of alternative views on the joint distribution under study and to estimate the source set.

7.2.4 Estimate

We can estimate the source set by inspecting the decompositions obtained by exploiting the structure of the graph and the modularity implicated by Hyper markov laws as follows.

To identify the i-th decomposition, obtained when C_i is set as the root clique, we let $C_{i,1}, ..., C_{i,k}$ denote a sequence of cliques satisfying the *running intersection property*. Let $S_{i,2}, ..., S_{i,k}$ be an associated sequence of separators, and set $S_{i,1} = \varnothing$, $i = 1, ..., k$. The (7.3) can be reformulated as:

$$\lambda^{C_i}(V) = \sum_{j=1}^{k} [\lambda(C_{i,j}) - \lambda(S_{i,j})] \tag{7.4}$$

Let assume that we collect n_1 and n_2 observations from $X^{(1)}$ and $X^{(2)}$, respectively. For each i-th ordering and for each components $\lambda(C_{i,j}) - \lambda(S_{i,j})$ defined in (7.4), we test the $H_{i,j}$ hypothesis of the equality of the two distributions. Finally, we save the result of the $(k \times k)$ tests in a vector $\phi_i = (\phi_{i,1}, ..., \phi_{i,k}) \in \{0, 1\}^k$, where $\phi_{i,j} = 1$ if the null hypothesis is rejected, and $\phi_{i,j} = 0$ otherwise. Note that, since $S_{i,1}$ is always

Decomposition	Component	$\lambda(C_i) - \lambda(S_i)$	Gdl	p-value	$\hat{D}_{G,i}$
$\lambda^{C_1}(V)$	C_1	231.94	14	1.5×10^{-41}	$\{ C_1 \}$
	$C_2 \mid S_2$	0.87	4	0.93	
	$C_3 \mid S_3$	5.19	4	0.27	
$\lambda^{C_2}(V)$	C_2	225.58	9	1.4×10^{-43}	$\{ C_2 \}$
	$C_1 \mid S_2$	7.23	9	0.61	
	$C_3 \mid S_3$	5.19	4	0.27	
$\lambda^{C_3}(V)$	C_3	48.93	9	1.7×10^{-7}	$\{ C_3 \cup C_2 \}$
	$C_2 \mid S_3$	181.84	4	3.0×10^{-38}	
	$C_1 \mid S_2$	7.23	9	0.61	

Table 7.2 Marginal and conditional tests for the k decompositions of G. Significant p-values are highlighted.

the empty set, the first component is the LLR test on the marginal distribution of the root-clique of the i-th sequence.

Definition 2. *The random set $\hat{D}_G$ defined as:*

$$\hat{D}_G = \bigcap_{i=1}^{k} \bigcup_{j:\phi_{i,j}=1} C_{i,j} \tag{7.5}$$

is an estimator of the minimal source set D.

In other words, the estimated source set consists of the set of shared variables, associated with marginal and/or conditional significant tests for all the orderings induced by the decomposition allowed by the graphical structure G.

It should be stressed that D_G - the graphical source set - could be not equal to the minimum group of variables that explain the differences between the two experimental conditions. The level of detail depends on the size of the cliques and the separators of the graph G. The graphical source set D_G and the minimum source set D will be equal in certain situations, such as every time the set of the primary genes is a separator within the graph. It can be thought of as the price to pay for not considering all possible subsets of $\{X_1, .., X_p\}$. For more details, see Djordjilović et al. [23].

7.2.5 A guided illustration

To summarize the proposed procedure, we consider the toy example of (Section 7.2.1). We generated a random sample of 100 for each condition.

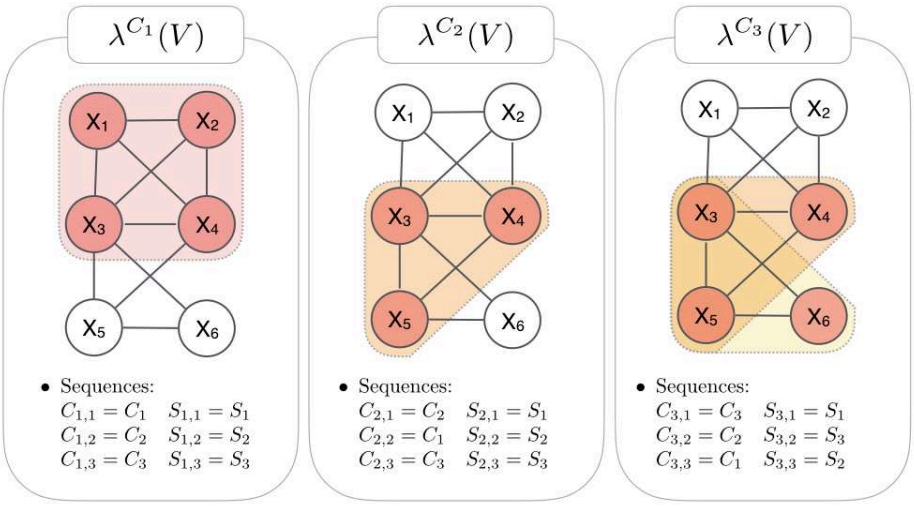


Fig. 7.3 The k possible decompositions of G for different choices of the root clique. Subsets corresponding to $\hat{D}_{G,i}$, $i = 1, 2, 3$ are highlihted. The sequence of cliques $C_{i,1}, ..., C_{i,k}$ and the associated sequence of separators $S_{i,1}, ..., S_{i,k}$ for each decomposition are reported on the bottom.

All possible decompositions for the graph G are three (Table 7.2), i.e., the number of cliques. Each of these decompositions consists of a marginal test (i.e., the root-clique) and two conditional tests. Since the conditional test statistic is just the difference between marginal tests, we can limit ourselves to performing tests on cliques and separators. Each marginal test statistic has a chi-square distribution with $p \times (p+3)/2$ degree of freedom (gdl), where p is the number of variables of the considered set (Table 7.1), i.e., the cardinality.

For each decomposition, the three obtained test statistics are asymptotically independent, and we can calculate the asymptotic p-value from a chi-square distribution. The degrees of freedom for the conditioned tests will be equal to the difference between the gdl of the test on the clique and the separator. The results are shown in (Table 7.2). For example, if we look at the decomposition $\lambda^{C_1}(V)$, the first test is significant while the other two do not, although C_2 and C_3 are marginally significant (Table 7.1). It means that the difference in marginal distributions is fully explained by the changes in the marginal distribution of the first clique. These conclusions can be extended to the remaining two decompositions.

We combine the results of the three decompositions according to (7.5) and obtain $\phi_1 = \{1, 0, 0\}$, $\phi_2 = \{1, 0, 0\}$ and $\phi_3 = \{1, 1, 0\}$, which leads to $\hat{D}_G = \{X_3, X_4\}$ (Figure

7.3). Thus we have identified the true source set D, which, being a separator, coincides with the minimum source set.

7.3 Practical Issue

When applied to data coming from HT technologies, our method suffers from all limitations inherited by high-dimensionality - the small number of samples compared with the huge number of genes - and multiple tests issue - the system-level analysis - making the proposed method applicable to a limited number of real applications.

In the next sections, we discuss these two issues and describe the methods proposed to solve them properly.

7.3.1 Small sample size

The basic block of our method is the likelihood ratio statistic. A necessary condition for the existence of the maximum-likelihood estimate is that the number of samples for the smallest group $n = min(n_1, n_2)$ is greater than the cardinality of the largest clique $p = max(|C_1|, ..., |C_p|)$. In fact, when we estimate the sample covariance matrices Σ, $\Sigma^{(1)}$ and $\Sigma^{(2)}$ these must be positive definite.

Moreover, even when the number is sufficient and the estimate of maximum likelihood exists, given the asymptotic nature of the log likelihood test the distribution for finite sample may be far from its asymptotic distribution. In fact, even for moderate number (i.e., $n \approx p$), the sample covariance matrix can no longer be considered a good approximation of the true covariance matrix.

In these cases, we suggest using a ridge estimator (i.e., the addition of a penalty to the log-likelihood) by summing a small quantity to the diagonals of these matrices. Then, we calculate the observed variance distributions in the two conditions and the pooled sample, compute the fifth percentile of each of these distributions, and use the minimum as the regularized quantity [39]. This approach allows us both to stabilize the estimates of the covariance matrices and to make comparable the LLR criterion among distributions.

However, if a regularized estimator for the covariance matrices is employed, permutation methods have to be applied to derive an approximation of the null distribution

of the test statistics. It must be emphasized that the permutational approach can be used because of the exchangeability property of the observations under the equality of the two distributions. This is not the case when means or variance equality hypotheses are considered.

7.3.2 Multiple testing correction

The estimated graphical source set $\hat{D}_G$ is a random variable, and its sampling properties hinge on the employed likelihood ratio test. In particular, there are k asymptotically independent tests within each decomposition of the global hypothesis. If we control the family-wise error rate (FWER) at a level α, when there are no differences between the two conditions (i.e., under the null hypothesis) $\hat{D}_G$ will be an empty set with probability converging to $(1-\alpha)$. We are thus protected against the inclusion of false positives.

The described procedure performs a total collection of m tests, of which k are marginal and $\sum_{i=1}^{k} v(C_i)$ are conditional tests, where $\sum_{i=1}^{k} v(C_i)$ denote the number of unique separators contained within the k cliques. While the p-values corresponding to hypothesis within each ordering are independent, this is not true for p-value across decompositions, and this definitely calls for multiple testing error correction.

The naive approach to control the FWER would be to apply the Bonferroni correction but in general intricate relationships among subgroups of hypothesis lead to high dependence on the associated p-values. Under this circumstance, Bonferroni correction is known to be conservative, and this means that the true FWER can be significantly lower than the chosen nominal level α.

To address this problem we use a method proposed by Westfall and Young [107], which uses permutations to obtain the joint distribution of the p-values, and by considering their dependency system and thus attenuating the conservativeness of Bonferroni.

The procedure proposed by these authors, also called *maxT*, starts with the creation of T permuted datasets and calculates the m test statistics for each of these. The results can be arranged in a $(T+1) \times m$ matrix P, where the first row is filled with the statistics calculated on the original data, while the remaining T store the test statistics of the permuted datasets.

Fixed a level alpha, we proceed as follows:

step 1) for each column of P, we calculate the asymptotic p-values for the hypotheses;

step 2) for each row of P, we calculate the minimum for each $(T+1)$ dataset;

step 3) the corrected threshold θ_i is the α-quantile of the permutational distribution of the p-values obtained in the previous step.

To get a threshold as close as possible to α, we decided to use a *step-down* version of the algorithm. Then, at the end of the three steps, we remove from the matrix P all the columns associated with rejected tests using the corrected threshold θ_i. Steps (2) and (3) are repeated on the resized P matrix until no hypothesis is rejected, considering the θ_i threshold for each i-th iteration.

As described in the previous paragraph (Section 7.3.1), if a regularized estimate of the covariance matrix is used, asymptotic distribution is no longer valid, and the *minP* version must be used. In this case, we compute permutational p-values per-hypothesis to obtain the $\widetilde{P}$ matrix, where each $\widetilde{p}_{i,j}$ elements is defined as $\#\{l : p_{l,j} \leq p_{ij}\}$. The matrix $\widetilde{P}$ replaces the P matrix in the $maxT$ algorithm.

The number T of permutations is always an issue with permutation-based multiple testing. It depends on the method, the alpha level chosen, and the number of hypotheses m. Although it would be best always to use the collection of all possible permutations, this is computationally not feasible even for a moderate dataset. For this reason, a collection of randomly generated permutations is often used.

The $minP$ method usually requires more permutations than $maxT$, due to the discrete nature of the permutation p-values. In fact, the minimum observed p-value will be equal to the minimum possible p-value for most of the permuted datasets - unless the number of permutations is very large - resulting in zero power for the method. For this reason, Goeamann et al. [35] recommend using m/α permutations as an absolute minimum. For $maxT$ the authors ensure appreciable performance with only $1/\alpha$ permutations when these can be enumerated. While with random permutations a higher number is suggested: 1.000 permutations are sufficient when the threshold is set at 0.05.

7.4 Appendix: lexicon and notation

Here, we briefly review key notions regarding Gaussian graphical models, relevant for our work.

Consider an undirected graph $G = (V, E)$ where V is a set of nodes and E is a set of edges. A subset of vertices A defines an induced subgraph $G_A = (A, E \cap A \times \mathrm{A})$. A clique is a maximal complete subgraph, that is, it is not a subset of any other complete subgraph. Two disjoint subsets $A, B \in V$ are said to be *separated* by a subset S (disjoint from A and B) if all paths from A to B contain vertices from S.

A graph G is decomposable if and only if the set of cliques of G can be ordered so as to satisfy the *running intersection property*, that is, for every $i = 2, ..., k$:

$$\text{if } S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \text{ then } S_i \in C_l \text{ for some } l < i - 1$$

although this ordering is generally not unique, the structure of the graph G uniquely determines the set of cliques $\{C_1, ..., C_k\}$ and the set of separators $\{S_2, ..., S_k\}$. For ease of notation, it is often set $S_1 = \{\varnothing\}$, so that the set of separators become $\{S_1, ..., S_k\}$.

For simplicity, we consider only graph consisting of a simple connected component, although most of the presented notions remain valid for more general graphs. We also restrict our attention to decomposable graphs, and this assumption is central to our approach. We assume throughout that cliques have been ordered in an order satisfying the running intersection property. Since, we deal with different partitions of the set of vertices, we note that such an ordering naturally leads to several *partitions* of V. Recall that (A, S, B) is said to be a partition of V if A, S and B are disjoint and $V = A \cup S \cup B$. Partitions of V that correspond to *decompositions* of the graph G are of particular interest. For a graph $G = (V, E)$, a partition (A, S, B) of V is a decomposition of G if A and B are separated by S in G, and S is complete.

Denote $p = |V|$ and let $X \sim \mathcal{N}(\mu, \Sigma)$ be a p-variate norm vector indexed by vertices of G. If Σ is invertible and such that its inverse, $K = \Sigma^{-1}$, has zeroes corresponding to missing edges of G we say that X is a Gaussian graphical model. Let $\mathcal{S}^+$ denote the set of all symmetric $p \times p$ positive definite matrices with zeros corresponding to the missing edges of G. Moreover, for $A \subset V$, let Σ_A denote the corresponding block submatrix of Σ. In gaussian graphical models, decompositions of the graph G correspond to special properties of the induced statistical models and associated inference procedures (see, Djordjilovic et al. [23]).

Chapter 8

IMPLEMENTATION

The programming language and style used for implementation play an important role in the diffusion of a method. Many of the developed GSEA approaches are implemented in `R` programming language [96] and are available as software packages either from Bioconductor and `CRAN` repositories, or the author's website. Their popularity depends not only on their usefulness but also on their availability as `R` package and their maintenance [67].

For this reason, the presented method (Section 7.2) has been implemented in an `R` package, called `SourceSet` (soon available on one of the above-reported repositories). In particular, the model has been extended to fit into the more traditional PA framework, where the interest is in considering more than one pathway at a time.

Thus, the functions contained in the package:

(i) use a list of pathways and a matrix of values that represents the gene expressions, to identify - for each graph - a set of variables consisting of potential sources of differential behavior between two experimental conditions;

(ii) perform a global meta-analysis on the entire set of input pathways, to provide replicable summaries of research findings through additional visualization tools and statistics.

Although our focus is on gene set analysis, it should be stressed that the developed methodology and its `R` packages are readily applicable in a wide range of other contexts. More precisely, it suits situations where the graphical structure is given a priori and remains constant across to two experimental conditions - allowing only the strength of relations between variables to change - and whenever the data can be assumed to be Gaussian. In the biological context, this includes, among others, log-transformed

microarray gene expression data, squared root reads counts or RPKM/FPKM in next-generation sequencing experiments and protein abundance.

In the following (Section 8.1), we present the algorithm in detail, with some notes on the extension to the case of N pathways and the issues discussed in (Section 7.3). A basic introduction to the usage of the package and its features, with particular emphasis on meta-analysis and personalization of graphical display of results, is given in (Section 8.2).

8.1 Algorithm

Motivated by the differential analysis of gene expression data, we proposed a method for identifying the set of genes responsible for the difference between two multivariate normal distributions Markov with respect to the same pathway.

Given a graph G - representing the dependency structure encoded in a pathway - and a matrix of values - that contains the expression levels of the measured genes in the collected samples for two experimental conditions - a general scheme of the procedure can be outlined as follows (Figure 8.1):

(*step* 1) decompose graph G in the set of the maximal cliques and the set of separators. G must be decomposable, so it may be necessary to moralize and/or triangulate the starting graph.

(*step* 2) identify the cliques orderings, and the associated separators, that satisfy the running intersection property, using each clique as root.

(*step* 3)
- (a) calculate marginal test statistics for the cliques and the separators, for both the original and the permutated datasets;
- (b) compute the conditional test statistics for the unique components, calculated as the difference between clique and separator marginal test statistics;
- (c) correct the α level to control the FWER, using the test statistics matrix of the previous point.

(*step* 4) make the union of the sets of variables belonging to cliques that are associated with a significant test, for each ordering defined in (*step* 2).

(*step* 5) derive the source set, defined as the intersection of the set of variables obtained in (*step* 4).

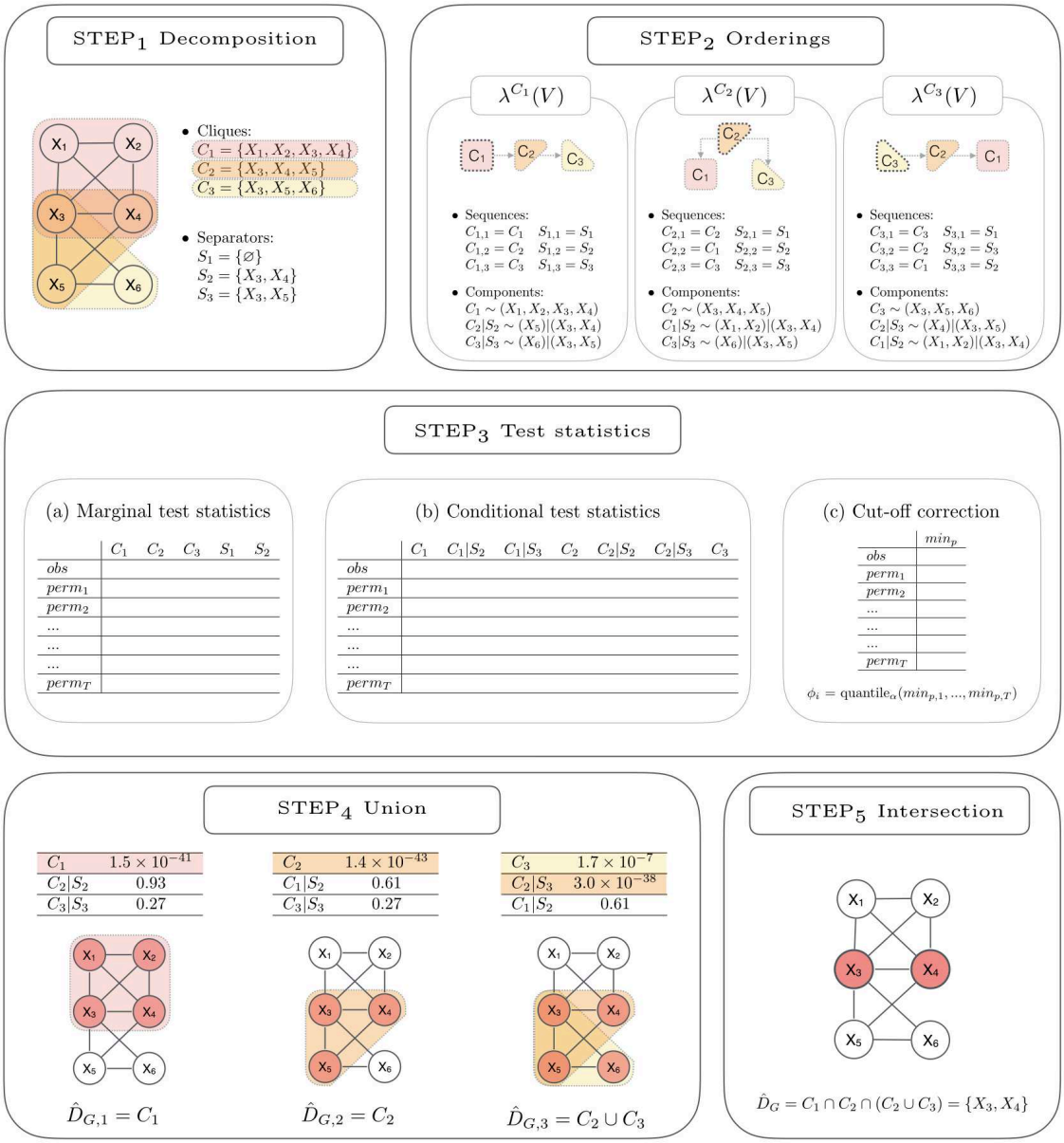


Fig. 8.1 Basic outline of `sourceSet` algorithm for the analysis of a single graph.

Looking at the broad picture of the N pathways case, to make the procedure consistent some major clarifications have to be made. In the following, we point out some computational remarks about the basic outline.

(*Note 1*) List of pathways

The primary interest of our work is not in the detection of the structure of a pathway because we consider it as fixed *a priori*. Various research groups have tried different strategies to address this challenge that has led to the development of many knowledge databases. To incorporate pathways into graphical models, the diagram needs to be translated into a mathematical graph, either directed or undirected. Due to the descriptive nature of pathways and their inherent complexity, there is no simple recipe for conversion that can be applied in every situation. For this reason, close collaboration with biologists is preferred at this step [24].

In general, we give full freedom to the user in providing the underlying graph, constraining only to accept a specific input data format (i.e., a graphNEL object). So, the user can provide a list of manually curated pathways, or use developed software to translate the bases of knowledge. To date, the most complete software available for this task is `graphite` R package [84]. `graphite` provides easy access to six different databases for a total of 14 different species. The resulting networks represent a uniform resource for the pathway analysis.

Regardless of the type of graph, obtained at the end of the translation (i.e., undirected or directed), our method works only with decomposable structure. However, it should be stressed that starting from a valid graph we can always obtain a decomposable one in a few steps (i.e., moralization and triangulation).

(*Note 2*) Setting the parameters

The input pathways normally have heterogeneous size and degrees of connectivity. To make the results obtained from each graph comparable, and to conduct a meta-analysis, particular attention is needed for the choice of the parameters. In particular, two parameters have to be set:

- the estimation method for the covariance matrices, i.e., sample or regularized;
- the number of permutations for the multiple testing threshold correction.

The estimation method must be the same for all pathways. If the user wants to use the sample covariance matrix, all cliques - in all pathways - must satisfy the

$n > p_i$ condition, where n is the number of samples for the smallest class and p_i the cardinality of the largest clique in the i-th pathway. If even one clique does not satisfy this requirement, the regularized estimate must be used. To obtain a reliable power, for the sample covariance estimate, it is recommended to use as criterion $n \gg p_i$ (see Section 9.1.3). Indeed, the distribution used to calculate the p-values of the performed tests is only asymptotically valid.

The number of permutations T_i, whether it's the $maxT$ or $minP$ correction, is naturally suggested from the α threshold and the number m_i of unique tests in the i-th pathway (see Section 7.3.2). Using different thresholds allows us to simultaneously control the "local" FWER and achieve comparable power among graphs. A "global" FWER control is possible by using a single $(T_M + 1) \times M$ matrix P, where M is the number of unique tests performed in each pathway, that is $M = \sum_{i=1}^{N} m_i$. The main problem is that the number of T_M permutations is generally very large, making the algorithm computationally onerous. Besides, the results may be lost reproducibility as the threshold - and the power - depends on the number and the degree of connectivity of the input graphs. For these reasons, the "global" option is not considered in the implemented algorithm.

(*Step 1-2*) Decomposition and orderings

For each path, the first step requires identifying the maximal cliques and all possible decompositions of the global distribution induced by the decomposable graph G. Generally speaking, the *clique problem* is NP-complete, indeed it is fixed-parameter intractable and hard to approximate. Listing all the maximal cliques can take an exponential time. Therefore, much of the theory about the clique problem is devoted to identifying appropriate types of graph that admit more efficient algorithms. In our model, a consistent computational relief is possible because of decomposable graphs - also called chordal graphs - fall into this last category. Also, the detection of permissible decompositions is closely related to the identification of perfect orderings, and such as problem may be solved in polynomial time when the input is chordal.

Specifically, we decided to use the `rip` function implemented in the `gRbase` package [22]. It identifies a sequence of the set of cliques that satisfies the running intersection property by first ordering variables by the maximum cardinality search algorithm. The root argument is used to check which clique will be the first to enter in the rip ordering.

Function	Description
`sourceSet`	Main function
`infoSource`	Resume informations about graphs and variables
`easyLookSource`	Summarise the results through a ggplot
`sourceSankeyDiagram`	Create a D3 Javascript Sankey diagram

Table 8.1 `SourceSet` main functions

In the `ripAllRootsClique` function (implemented in the `SourceSet` package) we extended the search space to get all possible orderings, that is, using as root all maximal cliques induced by the graph G. Given a graph, the function will provide:

- a list of elements: consisting of k maximal cliques and the associated k separators, and the m unique components;
- a list of k orderings: each of them will contain a proper subset k of the m unique components.

8.2 `SourceSet` R package

The `SourceSet` package consists principally of four functions (Table 8.1): the first one, which is the main function, implements the algorithm seen in the previous section, while, the other three functions guide the user in interpreting the obtained results through a meta-analysis, providing additional statistics and graphical device.

The following sections describe the arguments required for the use of each function, and the outputs provided.

8.2.1 Main function

Let's start exploring the package through the `sourceSet` main function. The function necessarily requires the following arguments:

`graphs`	a list of graphNEL objects, that represent the pathways to be analyzed.
`data`	a matrix of expression levels with column names for genes and row names for samples.

`classes`	a vector of 1 and 2 indicating the classes of samples. This vector must be matched with the rows of the data matrix, and can not contain more than two classes.
`alpha`	the p-value threshold
`shrink`	if set to `TRUE`, the algorithm will use the regularized estimate of the covariance matrices; otherwise, it will use the sample covariance matrices.
`permute`	if set to `TRUE`, permutational p-values will be computed for the significance of the tests; otherwise, the asymptotic distribution will be used.

If the last two parameters violate the assumptions required for the existence of the source set estimate (Section 7.3.1), the algorithm reserves the possibility to change the user settings through internal controls. A progress bar will show, for each pathway, the permutations status and the elapsed time.

The output of the main function is an object of the `sourceSetList` class. It contains as many lists as the input graphs, and provides the following variables:

`sourceSet`	a vector that contains the name variables belonging to the estimated source set $\hat{D}_G$
`marginalSet`	a list of vectors, one for each ordering, that contain the name variables belonging to the estimated source set $\hat{D}_{G,i}$
`Components`	a data frame that contains all the information about the m unique conditioned tests of the form $C_i \setminus S_i \vert S_i$, including the associated p-values
`Decompositions`	a list of data frame, one for each identified ordering. Each data frame is a subset of size k, of the `Components` elements
`Elements`	all the sets of cliques and separators induced by the used decomposable graph (see `Graph`)

`Threshold`	a list that contains all information regarding the threshold correction for multiple testings. It includes: • `alpha`: the input threshold; • `value`: the corrected threshold; • `type`: the used procedure (*minP* or *maxT*); • `iterations`: the number of iterations for the *step-down* procedure; • `nperms`: the number of T permutations.
`Graph`	the used decomposable graph. It should be pointed out that it may not be the same as the input graph. In fact, if it is not decomposable, the function will internally provide to moralize and triangulate it.

8.2.2 Meta-analysis and visualization

Although the interpretation of the source set for a single graph is intuitive, the analysis of the whole collections of results obtained from the N pathways might be complex. For this reason, we propose a guideline for the meta-analysis providing descriptive statistics and predefined plots. The key input argument of the meta-analysis functions is an object of the sourceSetList class, that is the output of the `sourceSet` function. Additional parameters may be needed to customize the display.

`infoSource`

The `infoSource` provides a summary of the results by focusing on either nodes or pathways, in fact, it supplies two different lists that are composed as follows:

	`$graph`
`n.source`	number of genes belonging to the source set, that is $\lvert\hat{D}_G\rvert$
`n.marginal`	number of genes belonging to at least one of the ordering source set, that is $\lvert\bigcup_{i=1}^{k}\hat{D}_{G_i}\rvert$
`n.graph`	number of genes in the graph, that is $\lvert V\rvert$
`n.cluster`	number of disconnected graph in G

source.impact	percentage of genes in the source set compared to the total genes in the graph. This index quantifies the proportion of the graph impacted by the primary dysregulation.
marginal.impact	percentage of genes in at least one of the ordering source set compared to the total genes in the graph. This index quantifies the proportion of the graph impacted by the secondary disregulation
p.value	mean of the k p-values obtained for each ordering by the Fisher's method combination of the k independent p-values.

$variable

n.graph	number of pathways in which the gene is contained.
specificity	percentage of pathways in which the gene appears, compared to the total number of analyzed pathways.
source.impact	percentage of time in which the gene is in the source sets, compared to the total number of pathways in which it appears.
marginal.impact	percentage of time in which the gene is in at least one ordering source sets, compared to the total number of pathways in which it appears.
relevance	percentage of times in which the gene is in the source sets, compared to the total number of analyzed pathways. From a general measure of the importance of the node based on the chosen pathways.
score	logarithm of the mean of the scores for the gene in all analyzed pathways changed in sign. The lower bound is zero (i.e., no significance), while the upper bound is $+\infty$ (i.e., maximum significance). The score depends on the p-values of the $H_{i,j}$ tests calculated in all orderings of the input pathways. Formally, it is defined as $-\log(\sum_{p=1}^{P} score_p^x/P)$, for $p = 1, .., P$, where P is the number of pathways and x is the considered gene. Instead, $score_p^x = \max(p_{i,j}^x)$ for $i = 1, ..k_p$, where p_i^x is the p-value for the j-th component in the i-th ordering, such that $C_{i,j}/S_{i,j}$ contains x.

`easyLookSource`

The function `easyLookSource` allows to summarize the results of the analysis through an heatmap (Figure 8.2b). The plot is composed of a matrix in which, on the rows, are represented the pathways and, on the columns, the genes.

Each $\text{cell}_{i,j}$ can take one of the following configurations:

(2) *blue* color, if the i-th gene is in the source set of the j-th pathway;

(1) *light blue* color, if the i-th gene is in at least one of the ordering source set of the j-th pathway;

(0) *gray*, if the i-th gene belong to the j-th pathway;

(NA) *white*, if the i-th gene does not belong to the j-th pathway.

In the plot, the pathways are vertically ordered - top to bottom - according to the numbers of nodes in the source set. Instead, genes are horizontally ordered (from left to right) based on the number of times they appear in a source set.

`sourceSankeyDiagram`

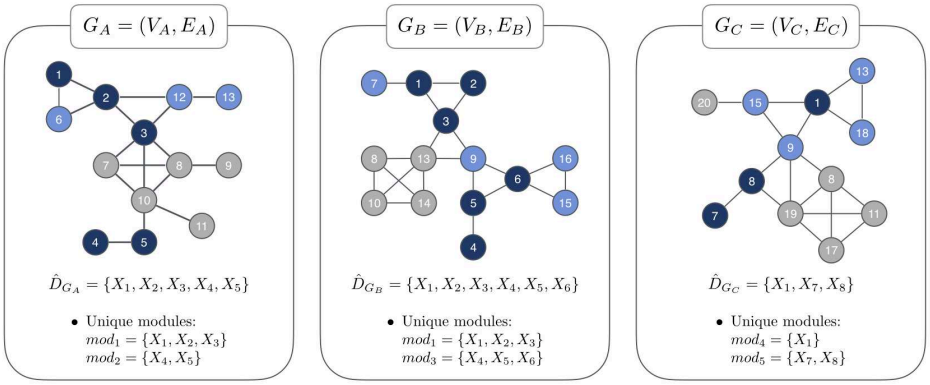
Another way to summarize the results in a visual manner is through a Sankey diagram (Figure 8.2c). It allows to highlight the relationships among nodes, graphs, and source sets.

The layout is organized on three levels:

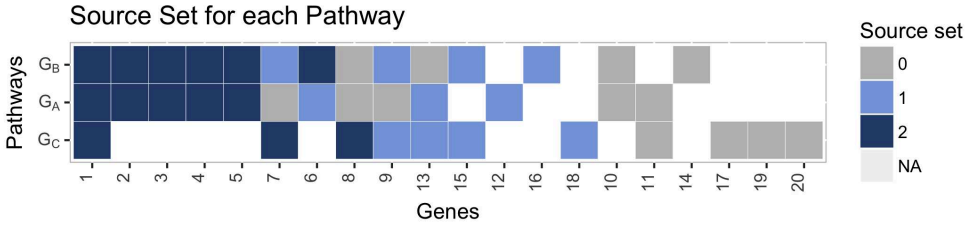
- the first level (on the left) consists of nodes that appear in at least one of the N source set.
- the second level (central) is made up of modules (Figure 8.2a). A module is defined as a set of nodes belonging to a connected subgraph of one pathway, that is also contained in associated source set. A pathway can have multiple modules, and at the same time, one module can be contained in multiple pathways.
- the third level (on the right) consists of pathways.

A link between two elements a and b must be interpreted - from left to right - as "*element a is contained in element b*".

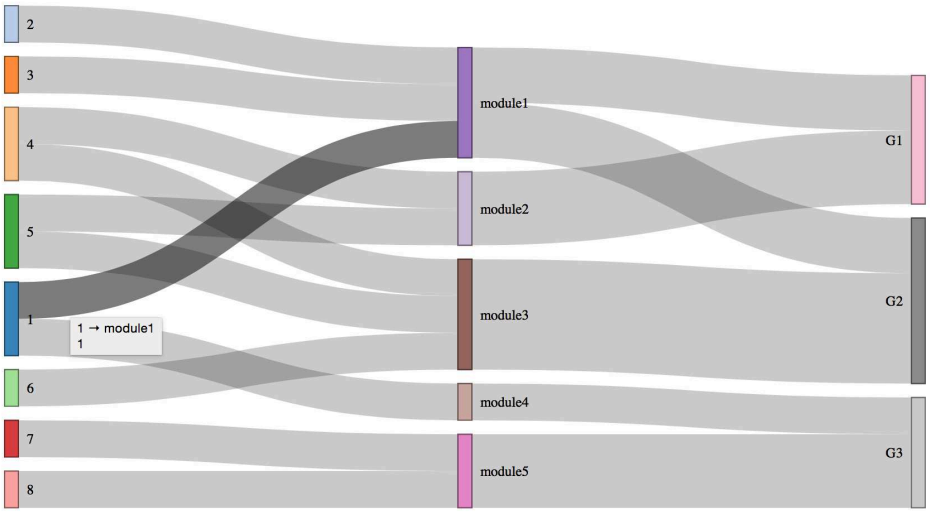
The implementation of the `sourceSankeyDiagram` function takes advantage of the D3 library [9, 1] (JavaScript), making the plot interactive. In fact, it is possible to vertically shift the displayed elements, and to view some usefull information positioning the cursor over items and links .



(a) Unique modules



(b) `easyLookSource` function



(c) `sourceSankeyDiagram` function

Fig. 8.2 Visualization of the results obtained from the `sourceSet` analysis, on an example database consisting of three pathways and twenty genes. For each graph, the modules are identified on the basis of the estimated source set (Figure 8.2a). The modules that are unique sets are used in the representation of the interactive Sankey graph (Figure 8.2c). The `easylookSource` graph is depicted in Figure 8.2b. The heat map highlights which sets the nodes belong to, in a specific pathway. For example, the gene *7* is in the source set of the G_C graph (blue rectangle) and in the marginal set of the G_A graph (light blue rectangle). The gene *10* is contained in both the G_A and G_B graphs, but never appears in either the source and the marginal sets (gray rectangles); while it does not appear in the G_C graph (white rectangle).

Chapter 9

VALIDATION

9.1 Simulated data

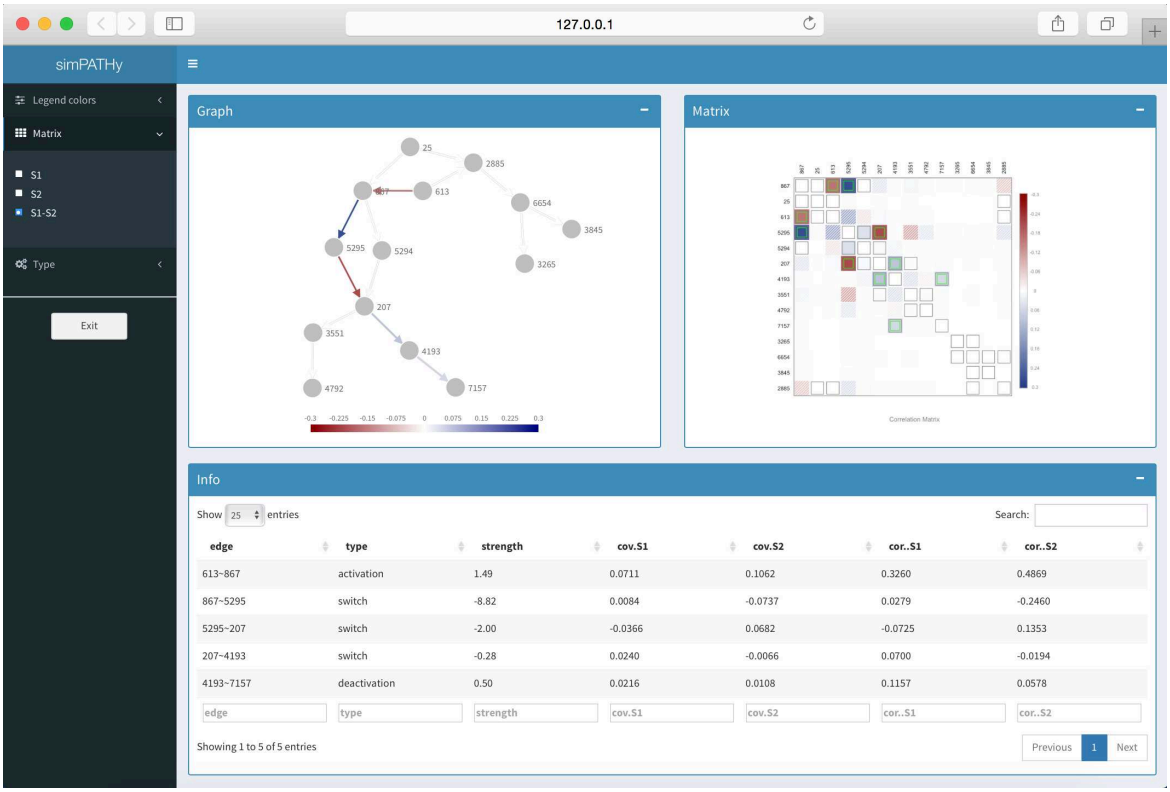
In the context of exploratory analysis, you may be concerned about the so-called screening property of the procedure, that is the guarantee of covering the source set with a high probability. For our procedure, the asymptotic guarantee is given by the consistency of the log-likelihood criterion test [23]. In the finite case, this property is closely related to the power of the underlying test and the magnitude of the differences between the two conditions.

We then studied the properties of the proposed method with a simulation study, both under the null hypothesis and under the alternative hypothesis. In the first part (Section 9.1.1), we illustrate some critical points about *in silico* data generation from biological networks, and briefly outline the proposed strategy. In the last sections, we describe the simulation settings (Section 9.1.2) and the results obtained (Section 9.1.3).

9.1.1 `simPATHy` package

Generating synthetic data that mimics the real biological dysregulation and that can be used as a benchmark dataset should be the first step of a proper validation strategy for any GSA tools. Nevertheless, almost all the methods proposed so far limited their attention to verifying performance under the null hypothesis of equality between the two populations, where the simulation strategy is trivial (i.e., the random sample of group labels).

Fig. 9.1 Shiny app allowing for an interacting exploration of the *simPATHy* output.



Creating synthetic data from two experimental conditions that differ only for a subset of genes that represent primary genes and emulate signal propagation following the topology of a pathway is not a simple task. In fact, to the best of our knowledge, there were no tools that simulate data according to this design.

For this reason, during my Ph.D. I have worked on the development of a new method for simulating data from perturbed biological pathways based on probabilistic graphical models. *simPATHy* is implemented in an *R* package and freely available on CRAN. In this framework, we assume to model the data of the same pathway in different experimental conditions through two undirected graphical models that share the same structure G.

The model assumes that the disregulation mechanism is the effect of a set of genes, and as a consequence of this perturbation it propagates on the remaining variables through the connections described in the structure of a pathway. Intuitively, the idea is to emulate a chain of reactions.

Given a set of primary genes d and a subset of edges e - contained in the induced subgraph of the genes in d - the model defines a perturbation in three different ways:

- a change in the force of the connections in e (i.e., the pairwise covariances);
- a change in the variability of the genes in d (i.e., the variances);
- a change in the expression levels of the genes in d (i.e., the means).

Containing the covariance matrix all the information on the pathway topology, this is the key element on which `simPATHy` works.

The algorithm is able to:

(i) obtain an estimate of a covariance matrix compatible with a given graph starting from a sample covariance matrix;

(ii) modify the strength of selected variances/covariances elements that emulate the signal perturbation;

(iii) provide an appropriate repair mechanism based on the spectral decomposition of the correlation matrix for indefinite matrices.

In fact, even small changes to the covariance matrix elements result in an indefinite matrix and a lost of the structure of G in the corresponding concentration matrix. Adjusting the mean, however, consists in just decreasing/increasing the original values.

The covariance matrix and the vector of the starting mean (i) represent the parameters of the reference condition; while the modified covariance matrix (iii) and the vector of the - eventually - changed means represent the parameters of the perturbed condition. These parameters are used to generate random samples from normal multivariate distributions. Formally, starting from the parameters related to the control group, the procedure act on means, variances, and covariances so that the conditional distribution of the variables on which it doesn't intervene remains unchanged under the two conditions. However, this action affects the entire global joint distribution, thus creating the propagation effect.

For more technical details, please refer to the original work [85], while referring to the `R` package vignettes for instructions on the software.

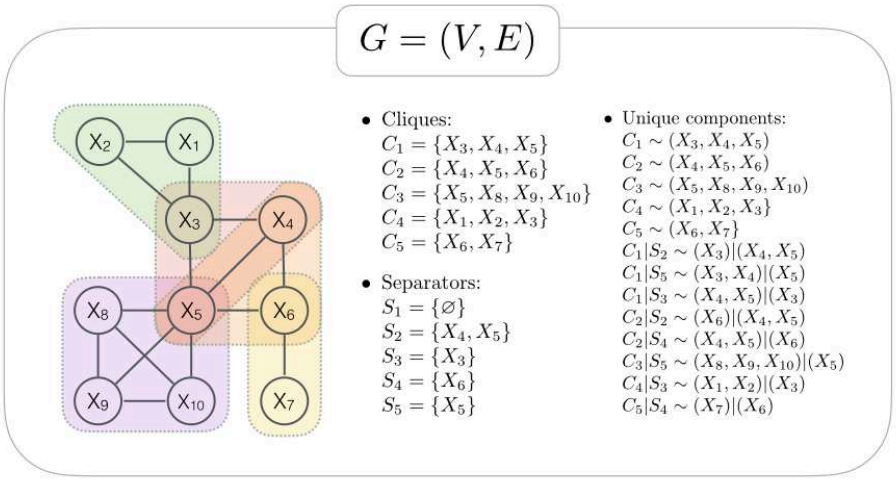


Fig. 9.2 Decomposable graph G consisting of 10 nodes, 5 cliques and 13 unique components.

9.1.2 Settings

We studied the finite case behavior of our algorithm through a simulation study under different scenarios, following the algorithm described in the previous paragraphs. To better understand the results of the simulation, we assumed that the dysregulation mechanism directly affects only a single gene, so we considered the perturbation defined as under (ii) and (iii), i.e., as an increase of the mean and the variance parameters of the chosen gene.

We used an a priori fixed graph G (Figure 9.2) and we considered three different scenarios:

(*scenario* 1) *there are no differences between the two conditions*: the source set is the empty set and so we are under the null hypothesis, $D = \{\varnothing\}$;

(*scenario* 2) *the differences between the two conditions are driven by a node that is a separator within the graph* G: the real source set is the variable X_5, $D = \{5\}$;

(*scenario* 3) *the differences between the two conditions are driven by a node that is contained in only a clique of the graph* G: the real source set is variable X_{10}, $D = \{10\}$.

For the last two scenarios, which are under the alternative hypothesis, the perturbation may be due to *mild*, *moderate* or *strong* intervention. It implies an increase in

the mean and the variance parameters of 20%, 60%, and 100%, respectively.

To verify the power of our procedure when we move away from the asymptotic distribution, we considered several numbers of samples for each class ($n = n_1 = n_2 \in \{25, 10, 5\}$). Since the cardinality of the largest clique is 4 (case $n > p$), the sample covariance matrix is well defined so we can use both the standard and the regularized approach. To observe the power of the two estimator methods, we rely on the permutational p-values corrected with the *step-down minP* algorithm with α equal to 0.05 and a number of permutation T set to 1.000 ($T \gg m/\alpha = 260$).

For each combination of source set and disregulation intensity, we obtained the mean and the variance parameters for the two conditions as described in Section 9.1.1. The parameters are used to generate 500 datasets derived from multivariate normal multivariate, for each number of samples. All the parameters used in this simulation can be found in the `SourceSet` package, through the `data(simulation)` command.

To evaluate the performance of our procedure, we considered the power of the test, defined as the number of times that the estimated source set $\hat{D}_G$ is contained in the true minimal source set D.

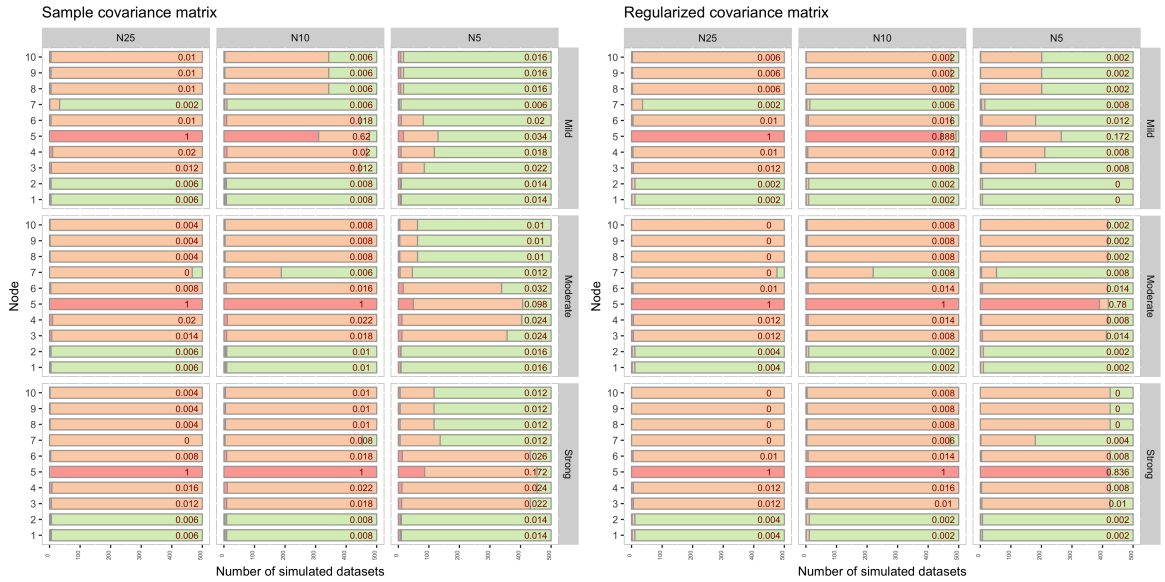
9.1.3 Results

H_0: Null hypothesis

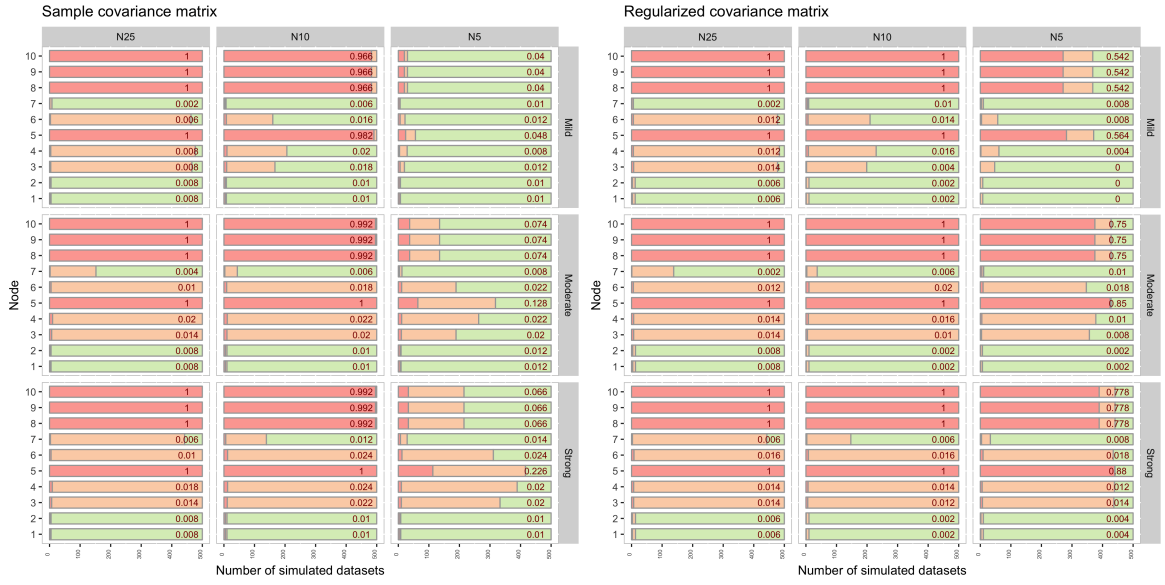
Under the null hypothesis, the algorithm demonstrates an excellent control of Type I error (Table 9.1) regardless of sample size and the choice of covariance matrix estimation method. Note that the procedure seems very conservative under the global null hypothesis. This behavior can be explained by the fact that when there are no differences between the two conditions, at least two false rejections are needed in order to obtain a non-empty source set estimate. This is, however, characteristic only for the global null hypothesis: when among considered hypotheses some are false, the control of the FWER is more accurate.

H_1: Alternative hypotesis

The results under the alternative hypothesis, for the sample covariance estimated (left panel) and the regularized one (right panel), are shown in (Figure 9.3). Each small



(a) $D = \{X_5\}$



(b) $D = \{X_{10}\}$

Fig. 9.3 Simulation study results under the alternative hypothesis, when the true source of the perturbation is the separator node *5* (Figure 9.3a) and the non-separator node *10* (Figure 9.3b), for the graph shown in Figure 9.2. In the left panels, the results for the sample covariance matrix estimate, while in the right panels, the results for the regularized estimate. Each small panel represents a combination of the parameters (number of samples and intensity of disregulation). Each bar is proportional to the number of times that the node appear in the source set (red bar), marginal set (orange bar), otherwise (green bar).

$D = \{\varnothing\}$	sample covariance matrix	regularized covariance matrix
$n = 25$	0.004	0.008
$n = 10$	0.008	0.022
$n = 5$	0.022	0.012

Table 9.1 Type I error for simulated data for different sample sizes ($n = n_1 = n_2$) and different estimators for the covariance matrix.

panel represents a configuration of the parameters (intensity and number of samples). Within them, there are three colored bars for each node:

- the red bar is proportional to the number of times the node d appears in the estimated source set, $d \in \hat{D}_G$;
- the orange bar is proportional to the number of times the node d is in the estimated source set of at least one ordering but not in all, $\exists i$ such that $d \in \hat{D}_{G,i}$ but $d \notin \hat{D}_G$;
- the green bar in the remaining case.

The length of each bar is 500, that is the number of simulated datasets.

Not surprisingly, the results depend on the setting used in the scenario. If we consider the scenario 2 where $D = \{5\}$, when the intervention is *moderate* or *strong*, and the sampling number is $n \geq 10$, the power is equal to 1 in all the considered cases. On the other hand, when the number of samples is close to the theoretical limit for the existence of the LLR criterion (i.e., 5), the regularized estimator outperform and correctly identifies the source set about 80% of the times compared to the 13% reached by the sample covariance matrix estimator. When the intervention is *mild*, the performances are lower for both the approaches, although the improvement that is obtained through the stabilization operation remains.

All the above considerations can be extended to the scenario 3, where the true source set is a non-separator node. In this case, however, a remark is mandatory. While for the second scenario the $\hat{D}_G$ corresponds exactly to D, in the third $\hat{D}_G$ only contains D. In fact, as anticipated (Section 7.2.4), our estimator tests the subsets of hypotheses that are induced by the structure of G, and hence it is not necessarily minimal. However, this behavior should not be considered as a false positive in our simulation. Moreover, if we consider the biological context, this behaviour should not be a cause for concern, in fact, i) pathways are represented by undirected decomposable

graphs that are - often - very connected, and therefore most nodes are separators; (ii) highly connected genes (hub genes) are more likely to be the cause of the disease as it has been shown to be the most lethal.

Finally, although the used graphical model is simple and the perturbation restricted to one node at a time, we are confident that the conclusions can be extended to more complex models and scenarios, including the $n \ll p$ case.

9.2 Biological validation

The goal of classical TPA methods is to identify the most perturbed pathway in a given experimental condition. As, to date, there is no universally accepted technique for the validation of the results, it is common practice to select one (or more) dataset such that there is a specific pathway that model the investigated condition. For example, *breast cancer pathway* in KEGG will be the target pathway in a breast cancer dataset. Analyzing and ranking all (or a subset) pathways, according to the score provided by the method, it is expected that the target will be as high as possible.

Intuitively, a pathway should be more significantly impacted if it hosts several genes that are the real source of perturbation [3]. But, as we pointed out in Section 7.2.1, most of the currently available methods cannot distinguish between primary and secondary disregulation. As a result of the complexity of biological phenomena, a large number of pathways are virtually implicated in all conditions leading to a bias in the results.

Although the objective of the proposed method is different (i.e., identifying genes responsible for the perturbation, and not the most perturbed pathways), we can use a similar validation technique, with a focus on genes. Moreover, we can illustrate how a marginal approach can hide the role of primary genes. For this reason, the assessment will focus on the ability of the source set algorithm to identify genes that are involved - or for which there are documented evidence - in the origin of phenotypes under study. In particular, two validation approaches are used.

The first dataset (Section 9.2.1) refers to the knock-down of STAT3 gene in patients affected by High-Grade Glioma. In this experiment, the exact source of perturbation is known, that is, the specific gene that has been knock-down. For this reason, all pathways that include this gene will be selected for validation, and STAT3 gene will

be expected to be included in the source set of each graph.

In the second approach (Section 9.2.2), we consider a well-known benchmark dataset on the ABL/BCR Acute Leukemia Lymphocytic (ALL) chimera, already analyzed by many other authors [3, 13, 27, 61, 62]. Although - unlike knock-down experiments - there is no true source of disregulation, some genotype abnormalities are known to be responsible for different transformation mechanisms of ALL and, as a consequence, of different response to treatment. Comparing patients with and without the B-cell receptor (ABL/BCR) gene rearrangement and analyzing all available pathways, we expect that the genes of the chimera will be present in the source set of the pathways that contained them. Moreover, we foresee that the chimera genes result among the most relevant genes in the meta-analysis. In this disease, the target pathway is *Chronic Myeloid Leukemia*.

To derive the list of underlying pathways needed as input for the `sourceSet` function, we used the `graphite` R package [84], which transforms the pathways contained in KEGG database [47] into graph objects. Each of these objects has been moralized and triangularized to obtain decomposable graphs (`gRbase` package [22]). The number associated with each node is a unique gene identifier from the Entrez Gene database at National Center for Biotechnology Information [58].

In the next sections, we present the results of these two case studies.

9.2.1 Silencing of STAT3 in brain tumors

The High-Grade Glioma (HGG) is the most common brain tumor in humans. Despite multimodal treatment with surgery, radiotherapy, and chemotherapy, these patients cannot be cured. The median survival of patients with glioblastoma (GBM) - the most frequent and malignant HGG - is only 15 months.

It is known that the over-expression of a mesenchymal gene expression signatures (MGSE) is associated with a poor prognosis in glioma patients. Carro et al. [12] identified 6 TFs that controlled the expression of $> 74\%$ of the MGSE genes. Two of them (STAT3 and C/EBPβ) emerge as synergistic initiators and master regulators (MRs) of this specific cancer signatures. To further investigate the role of both the TFs, the authors silence STAT3 and CREB both independently, and in combination.

Here we report the analysis of only STAT3 silencing dataset. The dataset includes 22 samples (11 knockdown and 11 control patients) and 19.292 measured gene expres-

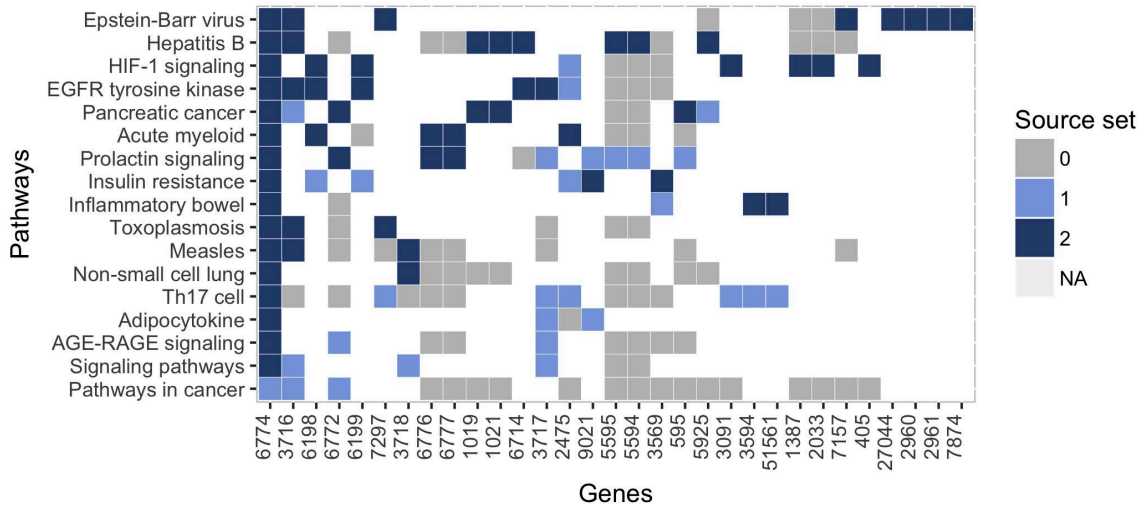


Fig. 9.4 `easyLookSource` visualization of the source set analysis results for STAT3 dataset. All the KEGG pathways that contained the STAT3 genes ($n_G = 17$) and the genes that appeared in at least one source set, are represented. The number associated with each node is a unique gene identifier from the Entrez Gene database. In particular, the STAT3 gene has the Entrez ID *6774*. A gene (column) can be in the source set (blue rectangle), in the marginal set (light blue rectangle) of a pathway (row), or otherwise (gray rectangle). If the gene is not present in a pathway, the rectangle is white.

sion levels. The original dataset is available in the GEO database, with the accession number GSE19114.

A total of 17 biological pathways from KEGG contain STAT3 gene (Table 9.2a). The analysis has been performed only on these pathways to validate the performance of our method to identify the primary dysregulation. The number of DEG (using Empirical Bayes test as implemented in `limma` package with $FDR < 0.05$) is 1.029, and as expected, STAT3 is the most differentially expressed. Forty out of 1.029 DEGs are mapped to the 17 biological pathways considered.

Using our method we identify all pathways with STAT3 as significant. In particular, looking at Figure 9.4, it is worth to note that, apart from *Pathway in Cancer*, STAT3 is always included in the source set. Indeed, it has been found on 16 over 17 pathway (source.impact=0.941). We highlight that in four out of 16 pathways (Table 9.2b), STAT3 is the only element of the source set, although marginal regulation involves many more genes. *Th17 cell differentation* pathway gives the most obvious example: the source set algorithm identifies STAT3 as the primary source of disregulation while classifying the remaining 39 genes as perturbed by the effect of signal propagation (i.e.,

(a) General pathways information

Name	$\|V\|$	$\|E\|$	n.cluster	k	$max(\|C_k\|)$	$\|DEG\|$
Acute myeloid leukemia	55	193	1	30	11	4
Adipocytokine signaling pathway	62	286	1	39	11	4
AGE-RAGE signaling pathway in diabetic complications	87	582	2	43	17	8
EGFR tyrosine kinase inhibitor resistance	79	468	1	33	19	6
Epstein-Barr virus infection	81	268	7	38	11	7
Hepatitis B	129	597	4	69	17	11
HIF-1 signaling pathway	97	459	1	61	13	8
Inflammatory bowel disease (IBD)	47	131	1	30	7	2
Insulin resistance	91	675	1	32	17	10
Measles	99	375	9	41	15	4
Non-small cell lung cancer	56	228	2	23	10	4
Pancreatic cancer	62	233	4	31	13	6
Pathways in cancer	304	2878	3	131	28	22
Prolactin signaling pathway	70	404	1	31	13	3
Signaling pathways regulating pluripotency of stem cells	107	751	1	37	28	6
Th17 cell differentiation	90	650	2	36	20	5
Toxoplasmosis	87	259	6	41	13	6

(b) Results of source set analysis

Name	n.graph	n.source	n.marginal	source impact	marginal impact	p-value
Acute myeloid leukemia	55	5	13	0.091	0.236	≈ 0
Adipocytokine signaling pathway	62	1	29	0.016	0.468	≈ 0
AGE-RAGE signaling pathway...	87	1	15	0.011	0.172	≈ 0
EGFR tyrosine kinase inhibitor...	79	6	19	0.076	0.241	≈ 0
Epstein-Barr virus infection	81	8	22	0.099	0.272	≈ 0
Hepatitis B	129	8	12	0.062	0.093	≈ 0
HIF-1 signaling pathway	97	7	19	0.072	0.196	≈ 0
Inflammatory bowel disease	47	3	12	0.064	0.255	$8.30e^{-06}$
Insulin resistance	91	3	21	0.033	0.231	≈ 0
Measles	99	3	3	0.030	0.030	≈ 0
Non-small cell lung cancer	56	2	2	0.036	0.036	≈ 0
Pancreatic cancer	62	5	18	0.081	0.290	≈ 0
Pathways in cancer	304	0	9	0.000	0.030	≈ 0
Prolactin signaling pathway	70	4	39	0.057	0.557	$7.94e^{-05}$
Signaling pathways regulating...	107	1	7	0.009	0.065	≈ 0
Th17 cell differentiation	90	1	40	0.011	0.444	$1.00e^{-07}$
Toxoplasmosis	87	3	7	0.034	0.080	≈ 0

Table 9.2 Pathways meta-analysis results for STAT3 dataset, provided by `infoSource` function. Pathway where STAT3 gene is the only element of the source set are highlighted. For more details about the interpretation of each index, see table `$graph` in Section 8.2.2 .

Entrez	Symbol	n.graph	specificity	source impact	marginal impact	score	relevance
6774	STAT3	17	1.00	0.94	1.00	7.07	0.94
3716	JAK1	9	0.53	0.56	0.90	0.51	0.30
6772	STAT1	9	0.53	0.22	0.44	0.85	0.12
6776	STAT5A	8	0.47	0.25	0.25	2.11	0.12
6777	STAT5B	8	0.47	0.25	0.25	1.72	0.12
6198	RPS6KB1	4	0.24	0.75	1.00	1.82	0.18
1019	CDK4	4	0.24	0.50	0.50	2.39	0.12
1021	CDK6	4	0.24	0.50	0.50	2.39	0.12
6199	RPS6KB2	4	0.24	0.50	0.75	1.71	0.12

Table 9.3 Genes meta-analysis results for STAT3 dataset, provided by `infoSource` function. Knock-down gene is highlighted. For more details about the interpretation of each index, see table `$variable` in Section 8.2.2.

(a) General pathways information

Name	$\|V\|$	$\|E\|$	n.cluster	k	$max(\|C_k\|)$	$\|DEG\|$
Axon guidance	126	793	3	66	18	2
Cell cycle	111	1024	1	35	21	3
Chronic myeloid leukemia	67	256	3	27	14	2
ErbB signaling pathway	78	294	1	43	10	2
Neurotrophin signaling pathway	98	509	2	52	13	3
Pathways in cancer	263	1960	4	127	24	5
Ras signaling pathway	170	1924	2	70	50	4

(b) Results of the source set analysis

Name	n.graph	n.source	n.marginal	source impact	marginal impact	p-value
Axon guidance	126	0	49	0	0.39	0
Cell cycle	111	16	30	0.14	0.27	0
Chronic myeloid leukemia	67	2	18	0.03	0.27	0
ErbB signaling pathway	78	3	8	0.04	0.10	0
Neurotrophin signaling pathway	98	4	15	0.04	0.15	0
Pathways in cancer	263	3	61	0.01	0.23	0
Ras signaling pathway	170	2	4	0.01	0.02	0

Table 9.4 Pathways meta-analysis results for ALL dataset, provided by `infoSource` function. Only pathways that contain chimera genes are shown. Pathways that contain both chimera genes in the estimated source set are highlighted. For more details about the interpretation of each index, see table `$graph` in Section 8.2.2.

secondary disregulation).

In addition to STAT3, meta-analysis tools provide a panel of four other genes (Table 9.3) with attractive characteristics (high/moderate relevance and score indices): one member of the Janus kinase family (JAK1), other members of the STAT family (STAT1, STAT5A, STAT5B), some ribosomal proteins (RPS6KB1 and RPS6KB2) and some cyclin-dependent kinases (CDK6, CDK4). All of these genes have direct protein-protein interactions with the knock-down gene. As a consequence of this proximity, they could capture the effect of primary disregulation in all those pathways where the STAT3 gene is not annotated.

9.2.2 ABL/BCR chimera in acute leukemia

Several distinct genetic mechanisms lead to ALL malignant transformations deriving from different lymphoid precursor cells that have been committed to either T-lineage or B-lineage differentiation. In particular, chromosome translocations and molecular rearrangements are frequent events in B-lineage ALL and reflect distinct mechanisms

Entrez	Symbol	n.graph	specificity	source impact	marginal impact	score	relevance
5330	PLCB2	37	0.25	0.24	0.62	1.90	0.06
5331	PLCB3	37	0.25	0.24	0.62	1.87	0.06
5332	PLCB4	37	0.25	0.24	0.62	1.85	0.06
23236	PLCB1	37	0.25	0.23	0.62	1.84	0.06
25	ABL1	7	0.05	0.86	1.00	7.24	0.04
217	ALDH2	6	0.04	0.83	0.83	5.81	0.03
3678	ITGA5	5	0.03	1.00	1.00	7.01	0.03
8900	CCNA1	5	0.03	0.80	1.00	6.81	0.02
857	CAV1	3	0.02	1.00	1.00	7.60	0.02
613	BCR	2	0.01	1.00	1.00	7.60	0.01

Table 9.5 Genes meta-analysis results for ALL dataset, provided by `infoSource` function. Chimera genes are highlighted. For more details about the interpretation of each index, see table `$variable` in Section 8.2.2.

of transformation. The relative frequencies of specific molecular rearrangements differ in children and adults with B-lineage ALL. The BCR breakpoint cluster region and the c-abl oncogene 1 (BCR/ABL) gene rearrangement occurs in about 25% of cases in adult ALL, and much less frequently in pediatric ALL.

The dataset we used here was published by Chiaretti et al. [14] and characterizes gene expression signatures in acute lymphocytic leukemia cells associated with known genotypic abnormalities in adult patients. Expression values (as available in the `ALL` BioC package [56]), appropriately normalized according to robust multiarray analysis (rma) and quantile normalization consist of $n_1 = 37$ observations from one experimental condition (BCR/ABL, presence of gene rearrangement) and $n_2 = 42$ observations from control condition (NEG, absence of rearrangement) and 8.595 genes. Using classical inferential analysis, we found 159 DEGs (only ABL1 is present).

In this case study, we decided to perform the analysis on the whole set of KEGG pathway ($n_G = 148$). Given the presence of the BCR/ABL chimera we expected that i) all pathways including BCR and/or ABL1 genes will be found as significant, and ii) the chimera genes will be included in the source set. Specifically, we require that the source set of *Chronic myeloid leukemia* (i.e., the pathway that describes the impact of the fusion genes in the cell) would be composed only by the chimera.

On the whole set of pathways, 56 (36%) have a non-empty source set, with a median size of four genes. The total number of genes in the source sets is 218 (Figure 9.5).

Fig. 9.5 `easyLookSource` visualization of the source set analysis results for ALL dataset. All the pathways contained in KEGG database ($n_G = 148$), and the first 50 genes, sorted in descending order with respect to the number of times they appear in source sets, are represented. The number associated with each node is a unique gene identifier from the Entrez Gene database. In particular, the chimera genes ABL1 and BCR have the Entrez ID *25* and *613*, respectively. A gene (column) can be in the source set (blue rectangle), in the marginal set (light blue rectangle) of a pathway (row), or otherwise (gray rectangle). If the gene is not present in a pathway, the rectangle is white.

ABL1 is annotated in seven pathways (Table 9.4a), and apart from one (*Axon guidance*), it is always identified in the source set (Table 9.4b). While BCR is annotated in two pathways and for both, it is detected in the source set. It is worth noting that, as required, ABL and BCR are the only genes in the source set of the target pathway. In this case, the source set algorithm highlights the fundamental role of the chimera, which with the marginal methods described in Section 7.1, would be hidden because of the propagation of the perturbation, involving 18 out of 63 genes.

Moreover, some members of the protein family of phosphocloresters PLCB (PLCB1, PLCB2, PLCB3, PLCB4) emerge as genes involved in several pathways of the considered collection. However, they obtain a moderate score ($score \in [1.84, 1.90]$). Other interesting genes are ITGA5, ALDH2, CCNA1 and CAV1 ($score \in [5.81, 7.60]$). In particular, although CAV1 is not noted in any pathway where the ABL1 chimera gene is present, there is evidence to support their protein-protein interaction [7]. Since CAV1 has been found to be significant in all pathways in which it is annotated (*source.impact=1*), this gene could play a key role in capturing the effect of primary disregulation due to ABL1 in those pathways.

Chapter 10

CONCLUSIONS

The high-throughput "omics" technologies are providing tons of data which are growing in size over time. The statistical analysis and the interpretation of such a complex and dynamic biologic systems have become a major challenge nowadays.

The goal of topological pathway analysis (TPA) is to identify the most perturbed pathways in a given condition. For this purpose, TPA methods estimate a score for a whole pathway that represents the activation/inactivation of the corresponding biological function. Pathways associated with significant scores are used as whole functional units in the interpretation of phenotype association experiments.

In the perspective of identifying genes that are responsible for the differences in a phenotype, TPA methods fail. Indeed, they are unable to distinguish between the real source of perturbation and the genes that merely respond to the perturbing signal.

Motivated by the analysis of differential expression genes, we proposed to model two experimental conditions with multivariate normal distributions with the same graph, which represents the information encoded in a pathway. Our approach exploits the idea of simultaneous looking at the differences in all marginal and conditional distributions implied by the Markov properties and uses the resulting evidence to infer the source set, that is a set of primary genes consisting of the potential source of differential behavior.

Moreover, the global hypothesis reformulation allows us to solve two of the main problems that characterize omics data. Thanks to the property of interchangeability of the observations under the null hypothesis, we improved the power of the test by a permutational approach. Also, we made the algorithm applicable even in cases where the number of observations is far below from the number of variables, by adopting a

ridge strategy for estimating the covariance matrix.

Finally, we extended the algorithm to fit into the more traditional PA framework, (where the interest is in considering more than one pathway at a time) and we implemented it in the `SourceSet` R package. Inside, we also provided additional statistics and graphical device to guide the user in interpreting the obtained results through a meta-analysis.

The results for both the simulated data (obtained through `simpathy`, a simulation method that we published in *Bioinformatics* journal [85]) and real dataset show that, when a disregulation exist, the `SourceSet` algorithm leads to an excellent sensitivity and specificity in all the possible scenarios considered, even with a low number of samples.

Although our focus is topologic pathway analysis, we feel that much of the methodology developed here is readily suitable for a wide range of other contexts. Specifically, this method could be applied to the metabolome data where patients are divided into groups according, for instance, to the presence of specific polymorphisms.

Currently, this is a working progress project born during my visiting period at Stanford University, in the Sabatti group. The study is focused on a subset of selected SNPs that characterize diabetes, in two cohorts coming from population studies for which are collected serum NMR-based characterization of lipoprotein and lipid measures along with other metabolic variables.

Taking as a starting point that the metabolic network can be represented as a graph, where metabolites and their interactions are nodes and edges, the SourceSet model could provide the portions that are effectively associated with a group of subjects with a particular mutation. In this way, it is possible to test groups of metabolites within a multivariate framework and highlight only those are deregulated by a mutation.

Moreover, the `SourceSet` approach assumes that the graph is known a priori, but we can relax this assumption generalizing our approach to learn the dependence structure with an estimation procedure that uses a portion of randomly sampled data. We are currently testing this generalization using the metabolome data.

References

[1] Allaire, J., Gandrud, C., Russell, K., and Yetman, C. (2017). *networkD3: D3 JavaScript Network Graphs from R*, r package version 0.4 edition.

[2] Ambros, V. (2001). microrna: Tiny regulators with great potential. *Cell*, 107(7):823 – 826.

[3] Ansari, S., Voichita, C., Donato, M., Tagett, R., and Draghici, S. (2017). A novel pathway analysis approach based on the unexplained disregulation of genes. *Proceedings of the IEEE*, 105(3):482–495.

[4] Bartel, D. P. (2017). Microrna: Target recognition and regulatory functions. *Cell*, 136(2):215–233.

[5] Bianchi, F., Nicassio, F., Marzi, M., Belloni, E., Dall'Olio, V., Bernard, L., Pelosi, G., Maisonneuve, P., Veronesi, G., and Di Fiore, P. P. (2011). A serum circulating mirna diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. *EMBO Molecular Medicine*, 3(8):495–503.

[6] Bignotti, E., Calza, S., Tassi, R. A., Zanotti, L., Bandiera, E., Sartori, E., Odicino, F. E., Ravaggi, A., Todeschini, P., and Romani, C. (2016). Identification of stably expressed reference small non-coding rnas for microrna quantification in high-grade serous ovarian carcinoma tissues. *Journal of Cellular and Molecular Medicine*, 20(12):2341–2348.

[7] Boettcher, J. P., Kirchner, M., Churin, Y., Kaushansky, A., Pompaiah, M., Thorn, H., Brinkmann, V., MacBeath, G., and Meyer, T. F. (2010). Tyrosine-phosphorylated caveolin-1 blocks bacterial uptake by inducing vav2-rhoa-mediated cytoskeletal rearrangements. *PLOS Biology*, 8(8):1–12.

[8] Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

[9] Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

[10] Braun, R. and Shah, S. (2014). Network Methods for Pathway Analysis of Genomic Data. *ArXiv e-prints*.

[11] Callari, M., Tiberio, P., Cecco, L. D., Cavadini, E., Dugo, M., Ghimenti, C., Daidone, M. G., Canevari, S., and Appierto, V. (2013). Feasibility of circulating mirna microarray analysis from archival plasma samples. *Analytical Biochemistry*, 437(2):123 – 125.

[12] Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., Anne, S. L., Doetsch, F., Colman, H., Lasorella, A., Aldape, K., Califano, A., and Iavarone, A. (2010). The transcriptional network for mesenchymal transformation of brain tumors. *Nature*, 463(7279):318–325.

[13] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, pages 808–835.

[14] Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Wang, K. S., Mandelli, F., Foà, R., and Ritz, J. (2005). Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clinical Cancer Research*, 11(20):7209–7219.

[15] Chung, Y.-W., Bae, H.-S., Song, J.-Y., Lee, J. K., Lee, N. W., Kim, T., and Lee, K.-w. (2013). Detection of microrna as novel biomarkers of epithelial ovarian cancer from the serum of ovarian cancer patient. *International Journal of Gynecological Cancer*, 23(4):673–679.

[16] Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression - An Approach to Regression-Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610.

[17] Cortez, M. A., Bueso-Ramos, C., Ferdin, J., Lopez-Berestein, G., Sood, A. K., and Calin, G. A. (2011). Micrornas in body fluids-the mix of hormones and biomarkers. *Nature*, 8:467 EP –.

[18] Cortez, M. A. and Calin, G. A. (2009). Microrna identification in plasma and serum: a new tool to diagnose and monitor diseases. *Expert Opinion on Biological Therapy*, 9(6):703–711. PMID: 19426115.

[19] Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317.

[20] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*.

[21] Demir, E., Cary, M. P., and Paley, e. a. (2010). The biopax community standard for pathway data sharing. *Nat Biotech*, 28(9):935–942.

[22] Dethlefsen, C. and Hojsgaard, S. (2005). A common platform for graphical models in R: The gRbase package. *Journal of Statistical Software*, 14(17):1–12.

[23] Djordjilović, V. and Chiogna, M. (2017). Searching for a source of difference: a graphical model approach. Submitted to the Annals of Applied Statistics.

[24] Djordjilović, V., Chiogna, M., and Vomlel, J. (2017). An empirical comparison of popular structure learning algorithms with a view to gene network inference. *International Journal of Approximate Reasoning*, 88(Supplement C):602 – 613.

[25] Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545.

[26] Drapkin, R., von Horsten, H. H., Lin, Y., Mok, S. C., Crum, C. P., Welch, W. R., and Hecht, J. L. (2005). Human epididymis protein 4 (he4) is a secreted glycoprotein that is overexpressed by serous and endometrioid ovarian carcinomas. *Cancer Research*, 65(6):2162–2169.

[27] Dudoit, S. and van der Laan, M. J. (2008). Multiple tests of association with biological annotation metadata. *Multiple Testing Procedures with Applications to Genomics*, pages 413–476.

[28] Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86.

[29] Fabregat, A., Sidiropoulos, K., and Garapati, e. a. (2016). The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487.

[30] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5):E359–E386.

[31] Fleischmann, R. D., Adams, M. D., White, O., and et al, R. A. C. (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512.

[32] Gao, Y.-c. and Wu, J. (2015). Microrna-200c and microrna-141 as potential diagnostic and prognostic biomarkers for ovarian cancer. *Tumor Biology*, 36(6):4843–4850.

[33] Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, 19(14):1817–1823.

[34] Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

[35] Goeman, J. J. and Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978.

[36] Hasegawa, S., Eguchi, H., Nagano, H., Konno, M., Tomimaru, Y., Wada, H., Hama, N., Kawamoto, K., Kobayashi, S., Nishida, N., et al. (2014). Microrna-1246 expression associated with ccng2-mediated chemoresistance and stemness in pancreatic cancer. *British journal of cancer*, 111(8):1572–1580.

[37] Hausler, S. F. M., Keller, A., Chandran, P. A., Ziegler, K., Zipp, K., Heuer, S., Krockenberger, M., Engel, J. B., Honig, A., Scheffler, M., Dietl, J., and Wischhusen, J. (2010). Whole blood-derived mirna profiles as potential new tools for ovarian cancer screening. *Br J Cancer*, 103(5):693–700.

[38] Hong, F., Li, Y., Xu, Y., and Zhu, L. (2013). Prognostic significance of serum microrna-221 expression in human epithelial ovarian cancer. *Journal of International Medical Research*, 41(1):64–71. PMID: 23569131.

[39] Huang, Y.-T. and Lin, X. (2013). Gene set analysis using variance component tests. *BMC Bioinformatics*, 14(1):210.

[40] Isci, S., Ozturk, C., Jones, J., and Otu, H. H. (2011). Pathway analysis of high-throughput biological data within a bayesian network framework. *Bioinformatics*, 27(12):1667–1674.

[41] Jacob, L., Neuvial, P., and Dudoit, S. (2010). Gains in Power from Structured Two-Sample Tests of Means on Graphs. *ArXiv e-prints*.

[42] Jacob, L., Neuvial, P., and Dudoit, S. (2012). More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.*, 6(2):561–600.

[43] Jacobs, I. J., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M., and Kalsi, J. K. (2016). Ovarian cancer screening and mortality in the uk collaborative trial of ovarian cancer screening (ukctocs): a randomised controlled trial. *The Lancet*, 387(10022):945–956.

[44] Jarry, J., Schadendorf, D., Greenwood, C., Spatz, A., and van Kempen, L. (2014). The validity of circulating micrornas in oncology: Five years of challenges and contradictions. *Molecular Oncology*, 8(4):819–829.

[45] Ji, T., Zheng, Z.-G., Wang, F.-M., Xu, L.-J., Li, L.-F., Cheng, Q.-H., Guo, J.-F., and Ding, X.-F. (2014). Differential microrna expression by solexa sequencing in the sera of ovarian cancer patients. *Asian Pacific journal of cancer prevention: APJCP*, 15(4):1739–1743.

[46] Kan, C. W., Hahn, M. A., Gard, G. B., Maidens, J., Huh, J. Y., Marsh, D. J., and Howell, V. M. (2012). Elevated levels of circulating microrna-200 family members correlate with serous epithelial ovarian cancer. *BMC Cancer*, 12(1):627.

[47] Kanehisa, M. and Furumichi, e. a. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361.

[48] Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):1–10.

[49] Kim, G., An, H.-J., Lee, M.-J., Song, J.-Y., Jeong, J.-Y., Lee, J.-H., and Jeong, H.-C. (2016). Hsa-mir-1246 and hsa-mir-1290 are associated with stemness and invasiveness of non-small cell lung cancer. *Lung Cancer*, 91:15–22.

[50] Kroh, E. M., Parkin, R. K., Mitchell, P. S., and Tewari, M. (2010). Analysis of circulating microrna biomarkers in plasma and serum using quantitative reverse transcription-pcr (qrt-pcr). *Methods (San Diego, Calif.)*, 50(4):298–301.

[51] Kutmon, M., Riutta, A., and Nunes, e. a. (2016). Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494.

[52] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[53] Langhe, R., Norris, L., Saadeh, F. A., Blackshields, G., Varley, R., Harrison, A., Gleeson, N., Spillane, C., Martin, C., O'Donnell, D. M., D'Arcy, T., O'Leary, J., and O'Toole, S. (2015). A novel serum microrna panel to discriminate benign from malignant ovarian disease. *Cancer Letters*, 356(2, Part B):628 – 636.

[54] Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

[55] Lawrie, C. H., Gal, S., Dunlop, H. M., Pushkaran, B., Liggins, A. P., Pulford, K., Banham, A. H., Pezzella, F., Boultwood, J., Wainscoat, J. S., Hatton, C. S. R., and Harris, A. L. (2008). Detection of elevated levels of tumour-associated microrna in serum of patients with diffuse large b-cell lymphoma. *British Journal of Haematology*, 141(5):672–675.

[56] Li, X. (2009). *ALL: A data package*, r package version 1.18.0 edition.

[57] Li, X., Shen, L., Shang, X., and Liu, W. (2015). Subpathway analysis based on signaling-pathway impact analysis of signaling pathway. *PLOS ONE*, 10(7):1–19.

[58] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 33:D54–D58.

[59] Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1):34.

[60] Marquez, R. T., Baggerly, K. A., Patterson, A. P., Liu, J., and Broaddus, R. (2005). Patterns of gene expression in different histotypes of epithelial ovarian cancer correlate with those in normal fallopian tube, endometrium, and colon. *Clinical Cancer Research*, 11(17):6116–6126.

[61] Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. (2013). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research*, 41(1):e19.

[62] Massa, M. S., Chiogna, M., and Romualdi, C. (2010). Gene set analysis exploiting the topology of a pathway. *BMC Systems Biology*, 4:121–121.

[63] Massa, S. and Sales, G. (2016). *topologyGSA: Gene Set Analysis Exploiting Pathway Topology*, r package version 1.4.6 edition.

[64] McCluggage, W. G. (2011). Morphological subtypes of ovarian carcinoma: a review with emphasis on new developments and pathogenesis. *Pathology*, 43(5):420–432.

[65] Meng, X., Müller, V., Milde-Langosch, K., Trillsch, F., Pantel, K., and Schwarzenbach, H. (2016). Diagnostic and prognostic relevance of circulating exosomal mir-373, mir-200a, mir-200b and mir-200c in patients with epithelial ovarian cancer. *Oncotarget*, 7(13):16923–16935.

[66] Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., and Tewari, M. (2008). Circulating micrornas as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, 105(30):10513–10518.

[67] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichiţa, C., and Drăghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4:278.

[68] Montani, F., Marzi, M. J., Dezi, F., Dama, E., Carletti, R. M., Bonizzi, G., Bertolotti, R., Bellomi, M., Rampinelli, C., Maisonneuve, P., Spaggiari, L., Veronesi, G., Nicassio, F., Di Fiore, P. P., and Bianchi, F. (2015). mir-test: A blood test for lung cancer early detection. *JNCI: Journal of the National Cancer Institute*, 107(6):djv063.

[69] Moore, R. G., McMeekin, D. S., Brown, A. K., DiSilvestro, P., Miller, M. C., Allard, W. J., Gajewski, W., Kurman, R., Bast, Robert C., J., and Skates, S. J. (2009). A novel multiple marker bioassay utilizing he4 and ca125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecologic Oncology*, 112(1):40–46.

[70] Nagy, Z. B., Barták, B. K., Kalmár, A., Galamb, O., Wichmann, B., Dank, M., Igaz, P., Tulassay, Z., and Molnár, B. (2017). Comparison of circulating mirnas expression alterations in matched tissue and plasma samples during colorectal cancer progression. *Pathology & Oncology Research*.

[71] Nishimura, D. (2001). Biocarta. *Biotech Software & Internet Report*, 2(3):117–120.

[72] Nowak, M., Janas, Ł., Stachowiak, G., Stetkiewicz, T., and Wilczyński, J. R. (2015). Current clinical application of serum biomarkers to detect ovarian cancer. *Przeglad Menopauzalny = Menopause Review*, 14(4):254–259.

[73] nucleic acids, E., their potential as diagnostic, p., and predictive biomarkers (2007). O'driscoll, lorraine. *Anticancer research*, 27(3A):1257–1265.

[74] of Obstetricians, A. C. and Gynecologists (2007). Acog practice bulletin no. 83: Management of adnexal masses.

[75] Ogata-Kawata, H., Izumiya, M., Kurioka, D., Honma, Y., Yamada, Y., Furuta, K., Gunji, T., Ohta, H., Okamoto, H., Sonoda, H., et al. (2014). Circulating exosomal micrornas as biomarkers of colon cancer. *PloS one*, 9(4):e92921.

[76] Ohman, A., Hasan, N., and Dinulescu, D. (2014). Advances in tumor screening, imaging, and avatar technologies for high-grade serous ovarian cancer. *Frontiers in oncology*, 4:322.

[77] Paracchini, L., Mannarino, L., Craparotta, I., Romualdi, C., Fruscio, R., Grassi, T., Fotia, V., Caratti, G., Perego, P., Calura, E., et al. (2016). Regional and temporal heterogeneity of epithelial ovarian cancer tumor biopsies: implications for therapeutic strategies. *Oncotarget*, 5.

[78] Prat, J. (2012). Ovarian carcinomas: five distinct diseases with different origins, genetic alterations, and clinicopathological features. *Virchows Archiv*, 460(3):237–249.

[79] Prat, J. and on Gynecologic Oncology, F. C. (2015). Figo's staging classification for cancer of the ovary, fallopian tube, and peritoneum: abridged republication. *Journal of Gynecologic Oncology*, 26(2):87–89.

[80] Pritchard, C. C., Kroh, E., Wood, B., Arroyo, J. D., Dougherty, K. J., Miyaji, M. M., Tait, J. F., and Tewari, M. (2012). Blood cell origin of circulating micrornas: A cautionary note for cancer biomarker studies. *Cancer Prevention Research*, 5(3):492–497.

[81] Resnick, K. E., Alder, H., Hagan, J. P., Richardson, D. L., Croce, C. M., and Cohn, D. E. (2009). The detection of differentially expressed micrornas from the serum of ovarian cancer patients using a novel real-time pcr platform. *Gynecologic Oncology*, 112(1):55 – 59.

[82] 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.

[83] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1):77.

[84] Sales, G., Calura, E., and Romualdi, C. (2017). *graphite: GRAPH Interaction from pathway Topological Environment*, r package version 1.22.0 edition.

[85] Salviato, E., Djordjilović, V., Chiogna, M., and Romualdi, C. (2016). simpathy: a new method for simulating data from perturbed biological pathways. *Bioinformatics*, 33(3):456–457.

[86] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Pid: the pathway interaction database. *Nucleic Acids Research*, 37(suppl1):D674–D679.

[87] Sebastian-Leon, P., Vidal, E., Minguez, P., Conesa, A., Tarazona, S., Amadoz, A., Armero, C., Salavert, F., Vidal-Puig, A., Montaner, D., and Dopazo, J. (2014). Understanding disease mechanisms with models of signaling pathway activities. *BMC Systems Biology*, 8(1):121.

[88] Shapira, I., Oswald, M., Lovecchio, J., Khalili, H., Menzin, A., Whyte, J., Dos Santos, L., Liang, S., Bhuiya, T., Keogh, M., Mason, C., Sultan, K., Budman, D., Gregersen, P. K., and Lee, A. T. (2014). Circulating biomarkers for detection of ovarian cancer and predicting cancer outcomes. *Br J Cancer*, 110(4):976–983.

[89] Shimomura, A., Shiino, S., Kawauchi, J., Takizawa, S., Sakamoto, H., Matsuzaki, J., Ono, M., Takeshita, F., Niida, S., Shimizu, C., Fujiwara, Y., Kinoshita, T., Tamura, K., and Ochiya, T. (2016). Novel combination of serum microrna for detecting breast cancer in the early stage. *Cancer Science*, 107(3):326–334.

[90] Shojaie, A. and Ma, J. (2016). *netgsa: Network-Based Gene Set Analysis*, r package version 3.0 edition.

[91] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:1–25.

[92] Stelling, J. (2004). Mathematical models in microbial systems biology. *Current Opinion in Microbiology*, 7(5):513 – 518.

[93] Subramanian, A. and Tamayo, e. a. (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.

[94] Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82.

[95] Taylor, D. D. and Gercel-Taylor, C. (2008). Microrna signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. *Gynecologic Oncology*, 110(1):13 – 21.

[96] Team, R. C. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[97] Thomas, D. (2010). Gene environment-wide association studies: emerging approaches. *Nat Rev Genet*, 11(4):259–272.

[98] Tiberio, P., Callari, M., Angeloni, V., Daidone, M. G., and Appierto, V. (2015). Challenges in using circulating mirnas as cancer biomarkers. *BioMed Research International*, 2015:10.

[99] Todeschini, P., Salviato, E., Paracchini, L., Ferracin, M., Petrillo, M., Zanotti, L., Tognon, G., Gambino, A., Calura, E., Caratti, G., Martini, P., Beltrame, L., Maragoni, L., Gallo, D., Odicino, F. E., Sartori, E., Scambia, G., Negrini, M., Ravaggi, A., D'Incalci, M., Marchini, S., Bignotti, E., and Romualdi, C. (2016). Circulating mirna landscape identifies mir-1246 as promising diagnostic biomarker in high-grade serous ovarian carcinoma: A validation across two independent cohorts. *Cancer Letters*, 388:320–327.

[100] VA, M. and on behalf of the U.S. Preventive Services Task Force* (2012). Screening for ovarian cancer: U.s. preventive services task force reaffirmation recommendation statement. *Annals of Internal Medicine*, 157(12):900–904.

[101] Valadi, H., Ekstrom, K., Bossios, A., Sjostrand, M., Lee, J. J., and Lotvall, J. O. (2007). Exosome-mediated transfer of mrnas and micrornas is a novel mechanism of genetic exchange between cells. *Nat Cell Biol*, 9(6):654–659.

[102] Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245.

[103] Venter, J. C., Adams, M. D., and Myers, e. a. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[104] Voichita, C., Ansari, S., and Draghici, S. (2017). *ROntoTools: R Onto-Tools suite*, r package version 2.4.0 edition.

[105] Vrahatis, A. G., Balomenos, P., Tsakalidis, A. K., and Bezerianos, A. (2016). Desubs: an r package for flexible identification of differentially expressed subpathways using rna-seq experiments. *Bioinformatics*, 32(24):3844–3846.

[106] Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, 81(6):1278–1283.

[107] Westfall, P. and Young, S. (2017). Resampling-based multiple testing : examples and methods for p-value adjustment / peter h. westfall, s. stanley young. *Wiley*.

[108] Xu, Y.-Z., Xi, Q.-H., Ge, W.-L., and Zhang, X.-Q. (2013). Identification of serum microrna-21 as a biomarker for early detection and prognosis in human epithelial ovarian cancer. *Asian Pacific Journal of Cancer Prevention*, 14(2):1057–1060.

[109] Zhang, J. D. (2017). *KEGGgraph: Application Examples*, r package version 1.38.1 edition.

[110] Zheng, H., Zhang, L., Zhao, Y., Yang, D., Song, F., Wen, Y., Hao, Q., Hu, Z., Zhang, W., and Chen, K. (2013). Plasma mirnas as diagnostic and prognostic biomarkers for ovarian cancer. *PLOS ONE*, 8(11):1–9.

[111] Zhu, C., Ren, C., Han, J., Ding, Y., Du, J., Dai, N., Dai, J., Ma, H., Hu, Z., Shen, H., Xu, Y., and Jin, G. (2014). A five-microrna panel in plasma was identified as potential biomarker for early detection of gastric cancer. *British Journal of Cancer*, 110(9):2291–2299.

[112] Zhu, C. S., Huang, W.-Y., and Pinsky, P. F. (2016). The prostate, lung, colorectal and ovarian cancer (plco) screening trial pathology tissue resource. *Cancer Epidemiol Biomarkers Prev; Cancer Epidemiology and Prevention Biomarkers*, 25(12):1635–1642.

[113] Zuberi, M., Khan, I., Gandhi, G., Ray, P. C., and Saxena, A. (2016). The conglomeration of diagnostic, prognostic and therapeutic potential of serum mir-199a and its association with clinicopathological features in epithelial ovarian cancer. *Tumor Biology*, 37(8):11259–11266.

[114] Zuberi, M., Mir, R., Das, J., Ahmad, I., Javid, J., Yadav, P., Masroor, M., Ahmad, S., Ray, P. C., and Saxena, A. (2015). Expression of serum mir-200a, mir-200b, and mir-200c as candidate biomarkers in epithelial ovarian cancer and their association with clinicopathological features. *Clinical and Translational Oncology*, 17(10):779–787.