Evry-Val-d'Essonne University
Statistique et Génome laboratory
Pharnext

PhD Thesis

presented by

Marine Jeanmougin

to get the diploma of Doctor in Applied Mathematics

Statistical methods for robust analysis of transcriptome data by integration of biological prior knowledge

Defended on November 16, 2012

Jury:

Anne-Laure Boulesteix (Prof.)	Munich University	(Reviewer)
David Causeur (Prof.)	Agrocampus Ouest	(Reviewer)
Laurence Tiret (Dr. Inserm)	Pierre & Marie Curie University	(Examiner)
Jean-Philippe Jais (MD, MdC)	CHU Necker - Paris-Descartes Univ.	(Examiner)
Christophe Ambroise (Prof.)	Evry-Val-d'Essonne University	(Supervisor)
Mickaël Guedj (PhD)	Pharnext	(Co-supervisor)

A ma famille…

Méthodes statistiques pour une analyse robuste du transcriptome par l'intégration d'*a priori* biologique

Résumé Au cours de la dernière décennie, les progrès en Biologie Moléculaire ont favorisé l'essor de techniques d'investigation à haut-débit. En particulier, l'étude du transcriptome à travers les puces à ADN ou les nouvelles technologies de séquençage, a permis des avancées majeures dans les sciences du vivant et la recherche médicale. Dans cette thèse, nous nous intéressons au développement de méthodes statistiques robustes, dédiées au traitement et à l'analyse de données transcriptomiques à grande échelle.

Nous abordons dans ces travaux le problème de sélection de signatures de gènes à partir de méthodes d'analyse de l'expression différentielle. Dans un premier temps, nous proposons une étude de comparaison de différentes approches issues de la littérature, basée sur plusieurs stratégies de simulations ainsi que sur des données réelles. Afin de pallier les limites des méthodes d'analyse différentielle qui s'avèrent peu reproductibles en pratique, nous présentons, dans un second temps, un nouvel outil appelé `DiAMS` (*DIsease Associated Modules Selection*). `DiAMS` est dédié à la sélection de modules fonctionnels de gènes par l'intégration de données d'expression et de données d'interactions protéiques et repose sur une extension du score-local.

Par la suite, nous nous intéressons au problème d'inférence de réseaux de régulation de gènes. Nous proposons une méthode de reconstruction à partir de modèles graphiques Gaussiens, basée sur l'introduction d'*a priori* biologiques sur la structure des réseaux. Cette approche nous permet d'étudier les interactions entre gènes et d'identifier d'éventuelles altérations des mécanismes de régulation, qui peuvent conduire à l'apparition ou à la progression d'une maladie.

Enfin, l'ensemble des développements méthodologiques décrits précédemment sont intégrés dans un *pipeline* d'analyse que nous appliquons à l'étude de la rechute métastatique dans le cancer du sein.

Mots-clefs Transcriptome; Puces à ADN; RNA-seq; Analyse différentielle; Information *a priori*; Intégration de données hétérogènes; Modèles graphiques Gaussiens; Cancer du sein.

Statistical methods for robust analysis of transcriptome data by integration of biological prior knowledge

Abstract Recent advances in Molecular Biology have led biologists toward high-throughput genomic studies. In particular, the investigation of the human transcriptome offers unprecedented opportunities for understanding cellular and disease mechanisms. In this PhD, we put our focus on providing robust statistical methods dedicated to the treatment and the analysis of microarray and RNA-seq data.

We discuss various strategies for differential analysis of gene expression levels and propose a comparison study. We provide practical recommendations on the appropriate method to be used based on various simulation models and real datasets. With the eventual goal of overcoming the inherent instability of existing tools for differential analysis, we present an innovative approach called `DiAMS`, for DIsease Associated Modules Selection. This method was applied to select functional modules of genes rather than individual genes and involves the integration of both transcriptome and Protein-Protein Interactions data in a local-score strategy.

We then focus on the development of a framework to infer gene regulatory networks by integration of a biological informative prior over network structures, using Gaussian graphical models. This approach offers the possibility of exploring the molecular relationships between genes, leading to the identification of altered regulations potentially involved in disease processes.

Finally, we apply our statistical developments to study the metastatic relapse of breast cancer.

Keywords Transcriptome; Microarrays; RNA-seq; Differential analysis; Prior knowledge; Integration of heterogeneous data; Gaussian graphical models; Breast cancer.

Remerciements

Durant ces années de thèse, j'ai eu la chance d'être entourée et soutenue par un certain nombre de personnes auxquelles je souhaiterais adresser mes plus sincères remerciements. En tout premier lieu, j'aimerais remercier mes deux directeurs : Christophe Ambroise, pour son encadrement scientifique de qualité qui m'a permis de m'épanouir dans mes travaux de recherche mais aussi pour sa grande gentillesse et le soutien qu'il m'a apporté pendant ces trois années, ainsi que Mickaël Guedj, qui est à l'origine de ce projet de thèse, pour ses conseils avisés à la fois en Statistique et en Génomique, sa disponibilité et l'engouement qu'il a manifesté tout au long de ma thèse. Mes remerciements vont aussi à Bernard Prum pour m'avoir chaleureusement accueillie au laboratoire Statistique et Génome et à Daniel Cohen qui m'a donné l'opportunité de travailler au sein de la société Pharnext.

J'exprime également mes remerciements à Anne-Laure Boulesteix et David Causeur pour l'intérêt qu'ils ont porté à mes recherches en acceptant de rapporter cette thèse et pour le temps qu'ils y ont consacré. Merci également à Laurence Tiret et Jean-Philippe Jais qui m'ont fait l'honneur d'examiner mes travaux.

Je voudrais remercier très chaleureusement tous les membres passés et présents du laboratoire Statistique et Génome : Carène pour la bonne humeur, l'enthousiasme (et les maisons en pain d'épices) qu'elle apporte au labo, Julien pour ses encouragements et pour l'appui scientifique qu'il m'a apporté pendant ces trois années en répondant toujours avec patience à mes questions, Michèle pour sa bienveillance et parce que nos conversations me manqueront, notre irremplaçable Claudine, Etienne, Pierre, Cyril, Maurice (pour m'avoir approvisionnée en baume du tigre les lendemains d'entraînements difficiles), Nicolas, Yolande, Mikaël, Bernard (en souvenir des enseignements en MB21), Edouardo, Ivan, Alex, Edith, Cécile, Guillem, Catherine (notre *big sister* à tous), Marie-Luce, Anne-Sophie et Gilles. Je voudrais souhaiter bon courage et bonne continuation aux thésards du laboratoire : Justin, Jonathan, Van Hanh, Marius, Morgane, Alia, Sarah que je remercie de s'être régulièrement enquise de mon état psychologique, Camille, pour nos petits bavardages et nos grandes discussions qui ont rendu cette thèse plus agréable et Matthieu, sans qui ni monsieur Lego ni le chaînon manquant n'auraient probablement jamais vu le jour.

Mes remerciements s'adressent également à tous mes collègues de Pharnext, en particulier le service Biostat': Fabrice (grand maître incontesté des cannelés bordelais), Jonas (et ses fameux voilà voilà !), Sory (notre source intarissable de conversations et de débats), Jan (et ses blagues... douteuses), Claudia qui nous a rejoint plus récemment ainsi que Caro en souvenir de nos séances d'escalade mais aussi Aude, Yannick, Nicolas, Viviane, Esther, Aurélie pour leurs encouragements, surtout pendant ces derniers mois.

Je remercie aussi les personnes avec qui j'ai eu l'occasion de collaborer, à savoir le consortium StatOmique, notamment Julie Aubert, Christelle Hennequet-Antier, Marie-Agnès Dillies ainsi que les membres du groupe SSB.

Je conclurai par une note plus personnelle pour mes amis et ma famille. Un grand merci bien évidemment à Clem, qui vadrouille à mes côtés depuis de nombreuses années (et qui n'a pas peur de faire 800 km pour venir m'entendre parler de choses incompréhensibles), à Iliane pour sa précieuse amitié, Antoine (qui doit être en train de terminer la rédaction de sa thèse à l'heure où j'écris ces lignes...courage!), Alex, Pierre, Antoine, Coco et Nico avec qui j'ai partagé de très bons moments : les soirées insaliennes, les séjours au ski, les week-ends à Flacy, les barbec' à la Feyssine. Muito obrigado à Audrey et Verão mes acolytes de capoeira. Un tendre merci à mes deux petits frères, Nico et Mathieu à qui je dois mes plus beaux fous-rires, à mes grands-parents, à ma grand-mère pour qui j'ai beaucoup d'admiration, ainsi qu'à Josette pour avoir toujours été une oreille attentive. Je remercie affectueusement mes formidables parents pour leur amour et leur soutien inconditionnel. Enfin, merci Nico, pour nos sorties d'escalade et nos entrainements nocturnes de course à pied quand il fallait décompresser des journées de rédaction, merci d'avoir retrouvé le bon sentier toutes les fois où tu nous as perdu en rando, merci d'être là chaque jour et de rendre la vie un peu plus douce.

Contents

Contents xiii

1 Preface 1

2 Introduction 5
2.1 Biological background 7
2.1.1 Central Dogma of Molecular Biology 7
2.1.2 Regulation of gene expression 8
2.1.3 Transcriptome analysis 11
2.2 Gene expression microarrays 15
2.2.1 Hybridization 15
2.2.2 Microarray protocol 16
2.2.3 From microarray to gene expression data 16
2.2.4 Data preprocessing 17
2.3 RNA deep sequencing (RNA-seq) 20
2.3.1 Data production workflow 20
2.3.2 Image processing 22
2.3.3 Mapping and summarization of reads 22
2.3.4 Normalization 25
2.4 Statistical background 27
2.4.1 Concepts of hypothesis testing 27
2.4.2 Differential analysis for transcriptome data 31
2.4.3 Multiple-testing 33

3 Statistical modeling for the identification of molecular signatures 39
3.1 Comparison of hypothesis testing strategies for microarray data 41
3.1.1 Statistical tests for microarray data 41
3.1.2 Comparison process 46
3.1.3 Results 49
3.1.4 Conclusion on hypothesis testing strategies 54
3.2 A local-score approach for the identification of differentially expressed modules in molecular networks 58
3.2.1 Local-score 59
3.2.2 Disease associated modules selection 62
3.2.3 Evaluation strategy 67
3.2.4 Discussion 69
Conclusion 71

4 Interpretation of gene signatures through regulatory networks inference 73
4.1 Gaussian Graphical Models (GGMs) 75
4.1.1 Preliminaries relating to GGMs 75
4.1.2 Conditional independence and partial correlation 75
4.1.3 Covariance selection 78
4.1.4 Likelihood inference of partial correlations 79
4.2 Application of GGMs to transcriptome data 79
4.2.1 Background of high-dimensional inference 79
4.2.2 Inference of sparse regulatory networks 80
4.2.3 Multiple inference under various experimental conditions 83
4.3 Inferring high-dimensional regulatory networks from biological prior knowledge 85
4.3.1 Motivations 85
4.3.2 Network Inference from a biological prior 87
4.3.3 Structure prior definition 88
Conclusion 93

5 Application to breast cancer 95
5.1 Breast carcinomas data 97
5.1.1 Breast cancer: a brief overview 97
5.1.2 Transcriptome data 97
5.1.3 Protein-Protein Interaction data (PPI) 98
5.2 Metastatic relapse study 98
5.2.1 Signature selection 98
5.2.2 Pathway analysis 101
5.2.3 Comparison of regulatory networks 102
Conclusion 105

6 Collaborative projects 107
6.1 Review and comparison of normalization approaches for RNA-seq data 109
6.1.1 Normalization methods 109
6.1.2 Comparison on real datasets 110
6.1.3 Evaluation on simulated data 114
6.1.4 Discussion on normalization methods 117
6.2 Introduction to differential analysis for RNA-seq data 117
6.2.1 Poisson distribution-based strategies 117
6.2.2 NB distribution-based strategies 118
6.2.3 Evaluation of differential analysis methods: preliminary results 118
6.3 Tree-Lasso: sparse estimation for tree structures . . . 119
6.3.1 Tree-Lasso model 120
6.3.2 Validation and application 120
Conclusion 123

7 Conclusion 125
7.1 General conclusion 125
7.2 Perspectives 127

Communications 131

1 Preface

Context

This thesis is based on the research which I carried out part-time in the *Statistique et Génome* (SG) laboratory and in Pharnext, a French biopharmaceutical company. For the three year duration of my research project, I was supervised by Professor Christophe Ambroise, director of the SG laboratory as well as Mickaël Guedj, head of the Department of Bioinformatics and Biostatistics of Pharnext. This PhD was half funded by a BDI (*Bourse Docteur Ingenieur*) grant issued by the CNRS (*Centre national de la recherche scientifique*) and half funded by Pharnext.

The SG laboratory, founded by Bernard Prum, develops statistical tools for processing genomic and genetic data, principally those used for analyzing biological sequences, high-throughput profiling of RNA/DNA, Single-Nucleotide Polymorphism (SNP) genotyping or Comparative Genomic Hybridization (CGH) data. In addition, the main research areas of the laboratory include the inference and study of biological network. During my thesis, I collaborated with Julien Chiquet and Camille Charbonnier on the latter thematic.

Pharnext, founded in 2007 by Professor Daniel Cohen and his main collaborators, aims to identify Pleotherapy-based drug candidates by combining several mini-doses of drugs which have already been approved for treatment of other diseases. This innovative approach allows targeting several molecular "nodes" in disease-perturbed pathways and thus helps to increase the treatment efficacy and safety. By the end of 2011, Pharnext has received clinical trial authorization for a phase II study with the first drug, PXT3003, owing to its Pleotherapy technology for the treatment Charcot-Marie-Tooth disease.

Motivations

The way each cell deploys its genome affects its function and the rate at which a transcript is synthesized. Thus, quantifying the amount of transcripts in a given cell under specific conditions enables us to determine which genes are expressed, providing clues to their possible role in the cell. The study of transcriptome, *i.e* the collection of transcripts, has become a major tool for biomedical research thanks to the development of high-throughput technologies, which provide a comprehensive picture of the set of transcripts in cells. The apparition of such technologies has revealed a striking need for statistics. Indeed, a typical high-throughput experiment measures the expression of several thousands of genes on a relatively small number of samples, which requires rigorous procedures for extracting relevant information. During the last decade, an incredible number of statistical tools have emerged for studying the transcriptome. A key motivating factor is the selection of genes, often referred to as the "molecular signature", whose combination is characteristic of a biological condition. Signatures give rise to new clinical opportunities, for understanding disease predispositions, improving diagnostics or prognostics, and providing new therapeutic targets as well as individualized treatment regimens. Their identification has become a topic of much interest in the medical research area, with several applications emerging, particularly in the field of Oncology. However, it turns out that the signatures resulting from classical tools proposed in the literature suffer from a lack of reproducibility and are not statistically generalizable to new cases. They appear therefore not to be reliable enough for translation to clinical practice. In this context, we put our focus on providing methods dedicated to the identification of molecular signatures with increased stability and easier interpretability, from high-throughput transcriptome data.

Thesis outlines and contributions

The first chapter consists of an introduction to transcriptome data and statistical tools to which we will refer throughout the thesis. After a summary of the basic concepts of Molecular Biology, we detail the process of transcriptome data production by focusing on microarray and RNA-seq platforms. Then, we explore the notions associated with hypothesis tests and detail the procedure of differential analysis for transcriptome data. Finally, we address the issue of multiple-testing and provide a review of the main approaches in this field.

The second chapter progresses to the issue of selection of robust molecular signatures. We first discuss various methods for differential analysis of gene expression levels. We then present the results of a comprehensive study to assess the performance of these tools based on statistical and practical criteria. It yields to the emergence of a test, namely `limma` from Smyth (2004), that exhibits the overall best results in terms of power, false-positive rate and ease of use. Despite its good performance, we highlight that `limma`, as with all other tests included in the study, suffers from low stability. In relation to this major drawback, we introduce a promising approach under the name of `DiAMS`, for Disease Associated Modules Selection. It involves a local-score strategy to detect modules of functionally related genes and allows the integration of transcriptome and Protein-Protein Interaction data. We demonstrate that `DiAMS`, suitable for both microarray or RNA-seq data, yields to power and reproducibility improvements in comparison to classical approaches. In the case of robust analysis, these results are very encouraging.

In the third chapter we address the challenging issue of signature interpretation in biologically meaningful terms and propose a framework based on regulatory network inference. The major advantage of using networks is that it enables the study of biological processes on a systems level, in order to understand how cellular components work together to produce system behaviors rather than focusing on individual components. Our approach builds on previous works from Ambroise et al. (2009) and Chiquet et al. (2011) and allows the statistical inference of regulatory networks from prior knowledge in a Graphical Gaussian Models framework. It offers the possibility of identifying altered regulations potentially involved in disease mechanisms.

In the fourth chapter, we apply the statistical developments we introduced throughout this thesis to study the metastatic relapse of breast cancer. We detail an analysis pipeline that includes the `DiAMS` approach and the network inference strategy described in chapters 3 and 4 respectively. We illustrate which kind of insights can be reasonably expected from these methods and demonstrate that the provided results can be a starting point for applications at clinical levels.

Finally, in the last chapter we present the collaborative works which were conducted in relation with our PhD research project. We first introduce two studies which was undertaken with the French StatOmique Consortium and which tackle the problem of RNA-seq data normalization and differential analysis. Our main contribution in this field is to provide an evaluation and a comparison of such approaches. In the third section of this chapter we mention a project carried out by Gen Yang, a trainee which I co-supervised with Christophe Ambroise and Julien Chiquet. During its internship in the *Statistique et Génome* laboratory, we were interested in a regularization technique, the tree-Lasso, for the purpose of variable selection.

We summarize the thesis outline as well as our main contributions in Figure 1.1.

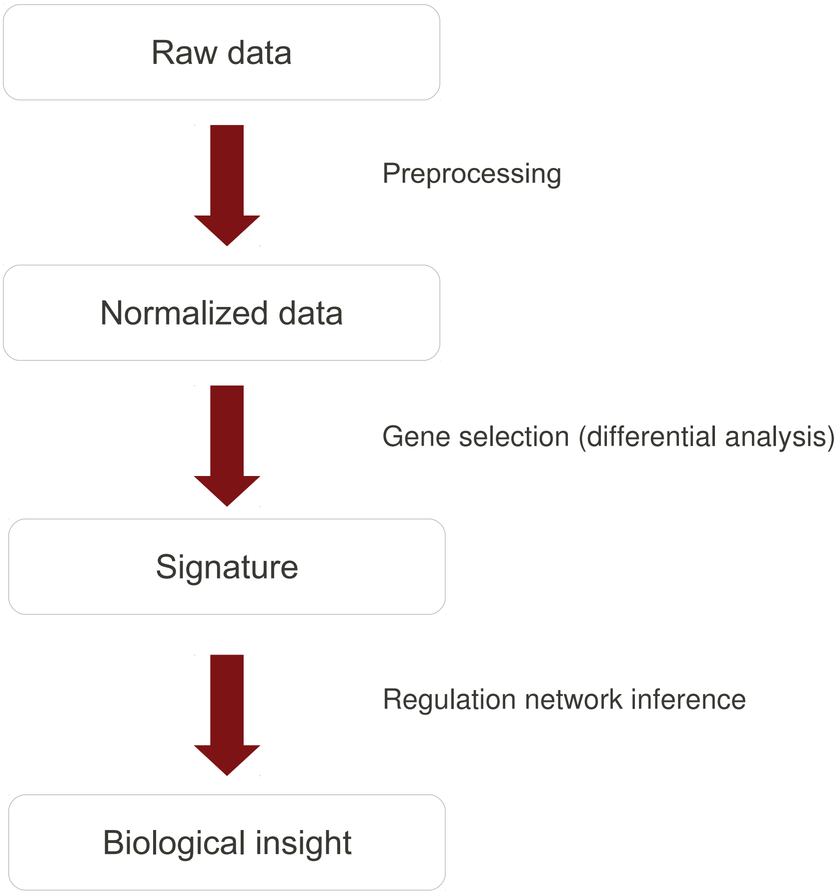

Raw data
Preprocessing
Normalized data
Gene selection (differential analysis)
Signature
Regulation network inference
Biological insight

Figure 1.1 – ***Thesis outline and contributions.***
We discuss in this thesis various themes of transcriptome data analysis. In particular, we conduct evaluation studies of statistical methods for normalization and differential analysis. In addition, we provide new statistical developments dedicated to gene selection and inference of regulatory networks.

Introduction 2

The human body contains about 10^{14} cells. Each of these cells has the same constitutional DeoxyriboNucleic Acid (DNA) sequence which is further organized intro short segments, called genes, that carry the genetic information. The gene expression refers to the process by which a protein is synthesized from the DNA molecule through the intermediate of the RiboNucleic Acid (RNA). The regulation of gene expression, *i.e.* the way in which different genes are turned on or off in specific cells, allows cells to achieve a wide range of functions and to generate the variety of phenotype we observe in Humans.
The study of the transcriptome, the entire repertoire of RNA molecules, represents an essential step towards a better understanding of the link between the genetic information encoded in DNA and the phenotype. In particular, the study of gene expression profiling has many applications in the area of biomedical research for understanding the mechanisms involved in the genesis of diseases, for diagnosis, prognosis or even to predict the response to treatment.

Although, the tools for transcriptome profiling have been available for years, the rapid quantification of transcripts in high-throughput became a possibility with the development of microarrays a decade ago. Expression profiling by microarray has provided a way to simultaneously investigate the abundance of thousands of genes in a single experiment. An equally revolutionary technology has recently emerged, known as RNA deep sequencing or RNA-seq. The first studies using RNA-seq to obtain transcriptome data were published in 2008. Since then, it has been successfully applied to a wide range of organisms and became a serious alternative to microarray for profiling the transcriptome.
The advent of new technologies that provide a vast amount of data gives rise to new statistical challenges. The most eminent arises from high-dimensionality, when the number of variables greatly exceeds sample size. Such a setting amplifies the need for effective statistical methods dedicated to inference, estimation, prediction or classification. Particular emphasis will be placed in this thesis on the problems of model or gene selection.

After introducing basic biomolecular and cell biological concepts that are used throughout the thesis, we review the regulation mechanisms that occur in cells and show how they modulate gene expression. We then briefly present the various existing tools for quantifying mRNA abundance and provide an overview of major biomedical applications. To conclude this first section, we discuss the statistical challenges associated with transcriptome data analysis in a high-throughput context. In the second and third sections, we introduce in more detail microarrays and RNA-seq as well as their respective data production workflows. We also address the issue of data normalization. In the final section, we recall basic statistical tools for transcriptome data analysis. We first present the hypothesis testing framework and illustrate the differential analysis procedure for transcriptome data. We then tackle the problem of multiple-testing by discussing the Family-Wise Error Rate (FWER) and the False-Discovery Rate (FDR), two error control criteria that have received much attention in the literature.

This chapter is associated with the following publication:

1. Bouaziz, Jeanmougin, et Guedj (2012). **Multiple-testing in large-scale genetic studies**, In "Data Production and Analysis in Population Genomics" (Bonin A. and Pompanon F. eds), *Methods in Molecular Biology Series, Humana Press*.

2.1 Biological background

2.1.1 Central Dogma of Molecular Biology

Basic concepts

The basics of Molecular Biology have been encapsulated in a concept called the Central Dogma that states that the DeoxyriboNucleic Acid (DNA) molecules carry the genetic information of cell coded in an alphabet of four letters: A, T, C and G. Each letter refers to a small molecule within a group known as nucleotides (usually referred to as bases), namely Adenine, Thymine, Cytosine and Guanine. A DNA molecule has two strands wrapped around one another in a helix (see Figure 2.1-A). The two strands are held together by the hydrogen bonding between their bases. As illustrated in Figure 2.1-B, Adenine forms two hydrogen bonds with Thymine and Cytosine can form three hydrogen bonds with Guanine.

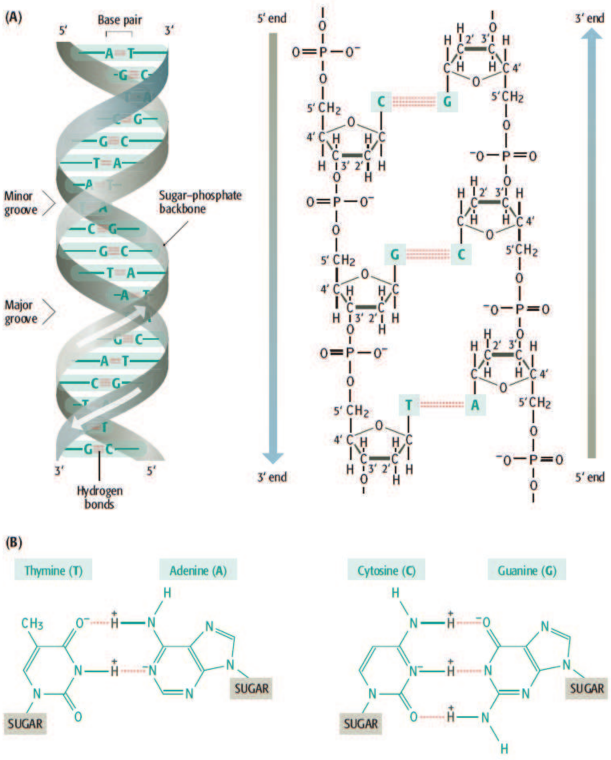


Figure 2.1 – ***DNA double helix.***
This figure is reproduced from Brown (2006). (A) DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone. (B) Rules of base pairing: Adenine forms two hydrogen bonds with Thymine, shown as dotted lines, and Cytosine can form three hydrogen bonds with Guanine.

In the nucleus, DNA is packed into a chromatin fiber whose primary structural unit is the nucleosome, a complex in which DNA firmly binds to a histone octamer. This compaction of DNA molecules is necessary to fit the large genomes inside cell nuclei.

The DNA is further organized into segments, called genes that code and control the synthesis of proteins. The proteins can be viewed as the major active tools of cells as they catalyze biochemical reactions and are involved in many mechanisms such as cell signaling or signal transduction. The process of information transmission from DNA to proteins is called *gene expression* and can be roughly summarized into two main parts: the transcription and the translation.

Transcription

The synthesis of proteins is mediated by RiboNucleic Acid (RNA) molecules. RNA, like DNA, is defined by a sequence of four nucleotides, but Thymine (T) is replaced by a similar molecule called Uracil (U). There are various types of RNA molecules in cells but here we are particularly interested in messengerRNA, or mRNA, that carries protein-building information.
The transcription is the process of RNA synthesis that is initiated at a certain signal sequence, called the promoter. It is mediated by a protein complex containing the RNA polymerase enzyme that copies the nucleotide sequence of the DNA into a complementary molecule: the precursor mRNA or pre-mRNA. After going through various post-transcriptional modifications, the pre-mRNA is spliced in the nucleus to remove non coding sequences called "introns" and transported to the cytoplasm through the nuclear pores. The result of splicing is a mature mRNA molecule. The corresponding mRNA to a certain gene is called "transcript". The transcriptome is thus defined as the collection of RNA molecules present in the cell.

Translation

Translation, the second major step in gene expression, leads to the synthesis of an amino acid sequence, called the protein. During this process, the mRNA is read in triplet (3 adjacent nucleotides) according to the genetic code, which relates the RNA sequence to the amino acid sequence in proteins. Proteins can then undergo post-translational modifications, such as cleavages or addition of functional groups, that will affect their function. The cell repertoire of proteins is called the proteome.

2.1.2 Regulation of gene expression

There are many different types of cells that serve a wide range of functions. If all of the cells within the human organism contain the same DNA information then what makes a brain cell so different from a breast tissue cell ? The answer lies in the regulation of gene expression, *i.e.* the way in which different genes are turned on and off, that occurs at different levels through various mechanisms. The Figure 2.2 reprinted from the book of Brown (2006) provides a good overview of the complexity of regulation processes. In the following sections we provide a non-exhaustive overview of regulation mechanisms that occur in the cells.

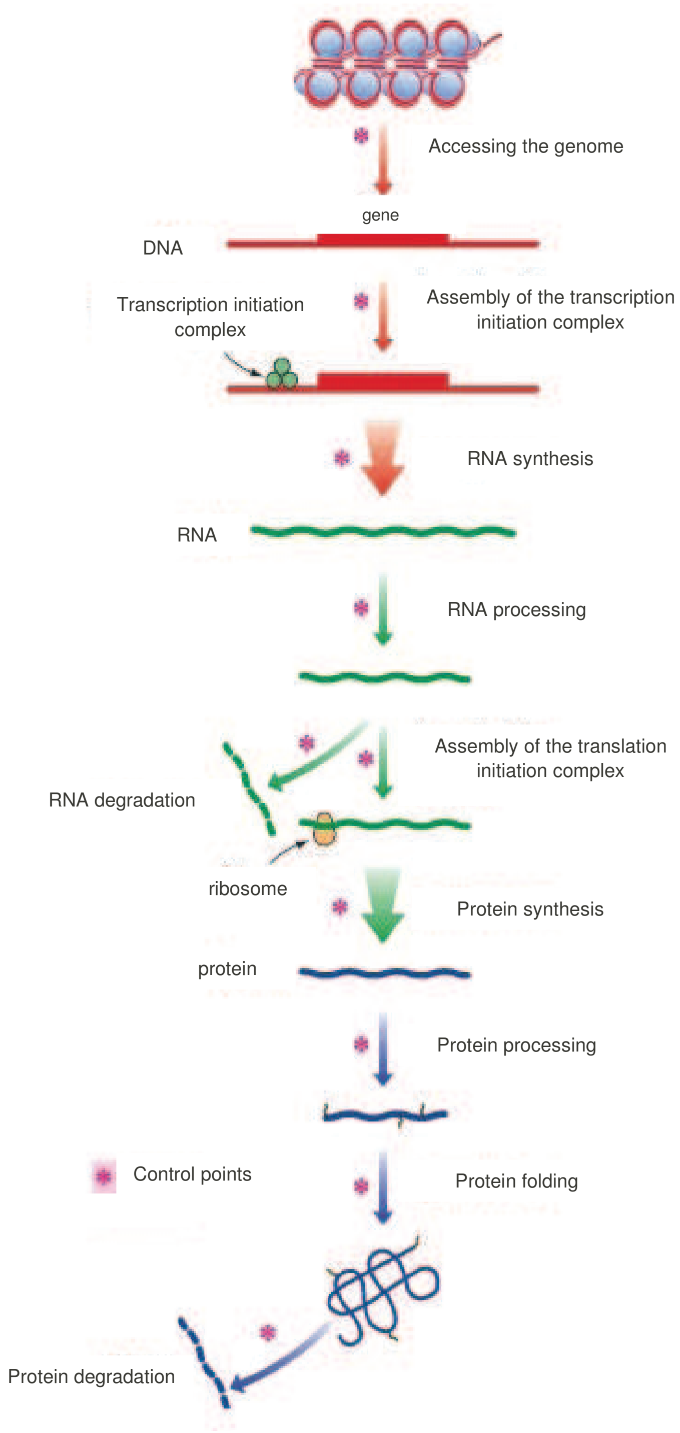


Figure 2.2 – ***Gene expression: regulation mechanisms.***
Overview of regulation mechanisms involved in the gene expression process from chromatin remodeling to post-translational modifications.

Transcriptional regulation

Firstly, transcription is controlled by limiting the amount of mRNA that is produced through various processes. The main mechanism of regulation at this level involves regulatory proteins. These proteins, called Transcription Factors (TFs), activate or inhibit the transcription by binding to specific promoter sequences.

However even when TFs are present in order to induce RNA synthesis in a cell, transcription does not always occur due to TFs being unable to reach their target sequences. Indeed, when histones and DNA are tightly bound in the chromatin fiber, they limit the access of TFs and RNA polymerase to promoters. This phenomenon of dynamic modification of chromatin architecture is called *chromatin remodeling*.

DNA methylation, which refers to the addition of a methyl group to the cytosine bases of DNA, also plays critical roles in gene silencing through chromatin remodeling. It occurs at CpG (Cytosine-phosphate-Guanine) sites, which are regions of the DNA sequence where a cytosine is directly followed by a guanine. Effects of DNA methylation are mediated through proteins known as Methyl-CpG-Binding Domain (MBDs) proteins. Such proteins are able to recruit chromatin remodeling proteins and may induce a transcriptional repression mechanism. The DNA methylation status is thus closely associated with chromatin structure.

Post-transcriptional regulation

Secondly, gene expression is subject to regulation at the post-transcriptional level. Alternative splicing of pre-mRNA is a fundamental mechanism of post-transcriptional regulation. It enables the production of different alternative splice variants, in other words the same pre-mRNA induces the synthesis of different mRNAs. The ability to generate a variety of mRNA molecules allows a single gene to encode various protein isoforms. Thus, by regulating which splice patterns occur in a given tissue or under a specific condition, a single gene can achieve different functions in the cell. Alternative splicing can also repress expression of a gene by introducing a premature stop codon that triggers mRNA decay. As with splicing, other modifications, namely the capping or polyadenylation (addition of a Poly(A)-tail), play a role in the post-transcriptional regulation of gene expression.

Finally, more recent research highlights the role of small non-coding RNAs (ncRNAs) molecules, known as microRNAs or miRNAs, in modulating gene activity. It has been shown that miRNAs participate in the regulation of a wide range of cellular processes. By base pairing to mRNAs, microRNAs mediate translational repression or the degradation of mRNAs. A variety of other non-coding RNAs, including long intergenic ncRNAs (lincRNAs) or small interfering RNAs (siRNAs), has become a subject of intense research since evidence of their implication in regulation mechanisms was highlighted.

Post-translational regulations

Post-translational modifications play a role in the regulation of gene activity and can occur shortly after translation or later in the life cycle of the protein. They enable to modulate the gene expression by altering the protein function or localization as well as their interactions with other proteins. For instance, they may induce chemical modifications to proteins, for instance by involving covalent addition of one or more groups. The most common types of covalent additions to proteins are phosphorylation, acylation, alkylation, glycosylation, and oxidation.

In addition to genetic and epigenetic mechanisms mentioned in the previous sections, increasing evidence suggests that the interactions between genes and environment might play a critical role in the regulation of gene expression. Gene-environment interactions are extremely complex and are thought to be mediated by epigenetic modifications of the genome. For instance, the degree of DNA methylation may be perturbed by environmental factors such as nutrients.

2.1.3 Transcriptome analysis

Measuring gene expression

The tools for quantifying and analyzing gene expression have been available for years. Until recently it was only possible to analyze a single gene at a time. The standard methods in Molecular Biology include Northern blots or Reverse Transcription-Polymerase Chain Reaction (RT-PCR) as well as sequence-based approaches such as Serial Analysis of Gene Expression (SAGE) or comparative Expressed Sequence Tag (EST). Microarray first made the analysis of the whole transcriptome in high-throughput possible. The principle and technology of microarrays is further detailed in section 2.2. Briefly, it consists of a small solid support onto which the sequences of tens of thousands or millions of different genes are immobilized. Transcripts, called targets, are extracted from samples to be investigated and labeled. They are then deposited onto the surface of the support using a robotic spotting device. The general idea is that targets will be hybridized to their complementary sequence on the microarray. Scanning the array results in expression levels of thousands of genes. Over the past 5 years, the emergence of Next-Generation Sequencing (NGS) methods has provided a new approach for quantifying the full transcriptome, termed RNA-seq. An RNA-seq experiment first involves isolating, fragmenting at random positions, and copying into complementary DNA (cDNA) the transcript content from cells. After an amplification step, the cDNA is then sequenced and the resulting reads are aligned to a reference genome. Finally, the number of reads mapped to each gene provide a measure of gene expression level. The experimental workflow for data production and the preprocessing steps are discussed in depth in section 2.3.

Applications

High-throughput expression profiling can be used to compare the mRNA abundance in samples under various conditions. This approach, known under the name of *differential analysis* or *class comparison* in the literature, aims to identify genes which behave differently, *i.e.* genes that are not regulated in the same way, across conditions. Such set of genes are often referred to as the "molecular signature". For instance, let us focus on the bone metastatic relapse status in a dataset provided by Bos et al. (2009), which is referred to as the *Bos dataset* in the following sections. These data[1] consist of gene expression values from 204 Affymetrix Human Genome U133 Plus 2.0 Arrays, each with gene expression values for over $54,000$ probes. We compare gene expression levels for patients who experience a bone relapse (BR) *versus* patients who do not (notBR) in order to identify the genes that may contribute to bone relapse. The differential analysis constitutes the core tool for such study as discussed in Chapter 3. Here we apply a Welch t-test to highlight genes whose expression levels significantly differ between BR and notBR patients. Approximately 150 genes were selected as differentially expressed. We visualized the expression profiles of these genes with a heat map displayed in Figure 2.3. Each colored cell in the map represents the expression intensity of a gene in a given sample: green indicates low expression, while red indicates high expression or up-regulation. Such representation allows us to observe genes that are differentially regulated between conditions. This signature of genes may be further investigated using pathway analysis (see Section 4.3.3) or network inference (see Chapter 4) in order to identify the mechanisms underlying the bone relapse.

In addition, gene expression profiling may help to define subtypes of samples within a population. These types of approach are referred to as *class discovery*. Cluster methods, and in particular hierarchical clustering, are widely used to detect groups of samples with similar expression profiles. For instance, Guedj et al. (2011) determined a classification of breast cancer tumors using unsupervised classification methods. Based on transcriptome data, they identify six homogeneous molecular subtypes displayed in Figure 2.4-A, associated with significant differences in clinical outcome, site of metastatic relapse or genomic anomalies.

Finally, measuring gene activity is of great interest to develop statistical models dedicated to *class prediction*, also called *discrimination* or *supervised learning* in the literature. The most common approaches for class prediction include classification methods such as discriminant analysis, k-nearest neighbor and random forests as well as support vector machines (SVM).
In their study, Guedj et al. (2011) provided a 256 genes signature able to predict the breast tumor subtype of a given sample. A classical distance-to-centroid method, implemented in the `citbcmst` R package [2] allows a given sample to be assigned to one of the subgroups. Using their ap-

[1] Series GSE12276 at the NCBI Genebank GEO

[2] `http://cran.r-project.org/web/packages/citbcmst/index.html`

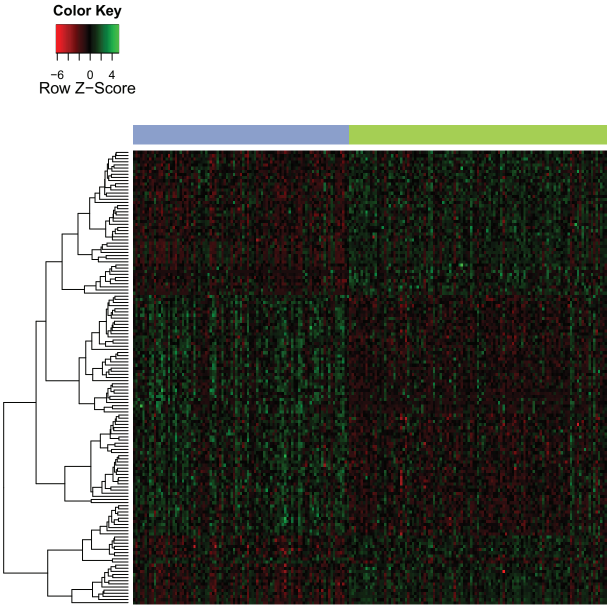


Figure 2.3 – ***Br vs. notBR patients Heat Map.***
The map represents expression levels for the genes (in rows). The columns contain samples from two types of patients, BR (in green) and notBR (in blue). Red represents high expression, while green represents low expression.

proach on the Bos dataset, we obtained the classification represented in Figure 2.4-B.

Throughout this thesis we focus on the issue of class comparison. We highlight in the first paragraph that molecular signatures are useful tools for identifying genes of interest, which exhibit differential expression patterns. The selection of relevant and robust signatures is a critical step towards a better understanding of disease genesis and progression. It requires reliable statistical methods at each step of the data analysis process.

Statistical challenges

The first statistical challenge when analyzing transcriptome lies in the preprocessing of raw data. Indeed, the various technologies induce systematic biases due to experimental variations such as intensity effect in the scanning process as well as non-specific hybridization for microarray or sequencing depth for RNA-seq. In the literature, this area of research is collectively referred to as normalization. This is a crucial step for comparing various samples; we address this point in sections 2.2.4 and 6.1 for microarray and RNA-seq data. The issue of experimental design is not discussed in this manuscript, although in relation to the normalization problem, it should be noted that the fundamental design aspects of data collection and analysis have to be treated with caution. Indeed, biological and technical variations should not be confounded to be able to partition the two sources of variations. As an illustration, let us return to the exam-

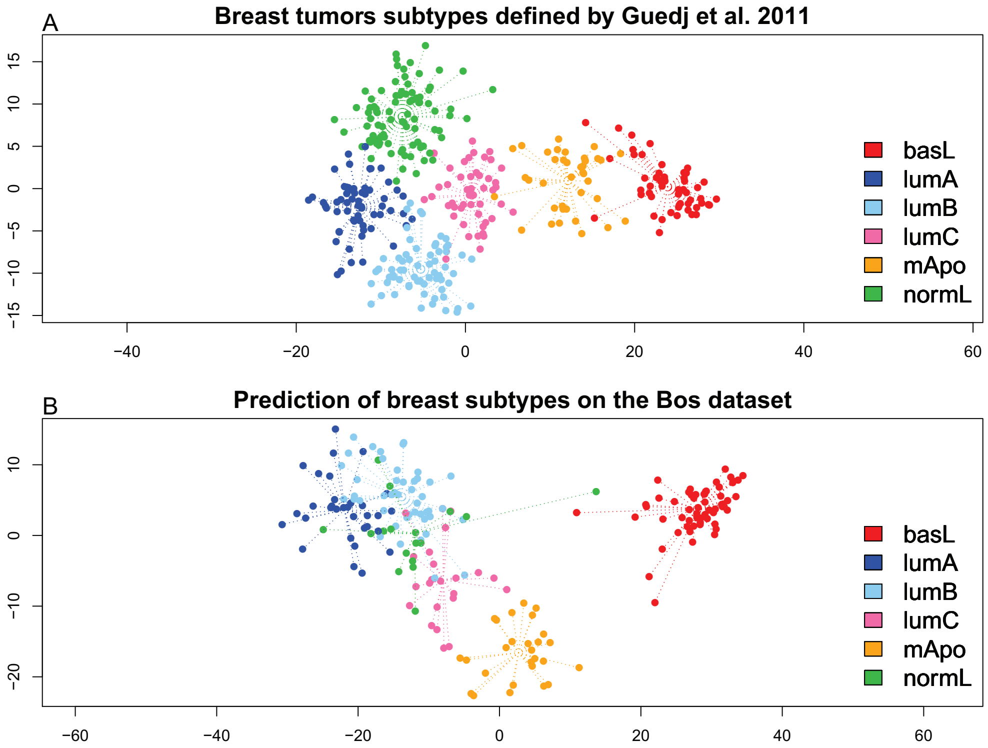


Figure 2.4 – ***Breast tumors classification.***
(A) Principal component analysis (PCA) of the six breast tumors subtypes defined by Guedj et al. (2011) (B) Prediction of tumor subtypes of the Bos dataset samples.

ple of the bone relapse study in the Bos dataset. If all samples of patients who experience a bone marrow relapse are hybridized in one day, and all samples of notBR patients are done in another day, the batch effects are mixed with real biological variations that we are interested in. Thus, we will not be able to make a conclusion about a biological effect. In this case, it would have been more reasonable to consider a design where BR and notBR samples are randomly distributed into different hybridization days. The principles of good design have been formalized by Fisher (1935) into three points, namely randomization, replication, and blocking. The interested reader may refer to the articles by Kerr et al. (2000), Lee et al. (2000) or Churchill (2002) for further detail on experimental designs in microarray or to Auer et Doerge (2010) and Fang et Cui (2011) for a review of concepts and designs in RNA-seq experiments.

Once systematic biases are removed, in relation to the issue of class comparison, the statistical challenge becomes selecting statistically significant genes. In a high-throughput context, the data consists of tens of thousands of variables with generally tens or at best hundreds of samples. So, in this framework, the dimensionality of the problem, denoted p, exceeds the sample size, n. In this "large p, small n" world, the traditional statistical theory is no more valid.
High dimensional data have commonly emerged in a wide variety of areas, from Biomedical Imaging and the large numbers of magnetic resonance images, to Astronomy and the vast amounts of data generated at a

high rate by telescopes all around the world. In this context, the ability to reduce dimensionality and to extract relevant information is fundamental. In the literature of microarray and RNA-seq data analysis, various testing procedures have been proposed as discussed in section 3.1.1. In carrying out thousands of tests simultaneously, a non-negligible proportion will be spuriously declared as significant. Classical methods for controlling the probability of false-positive findings are no longer relevant. Statistical approaches reviewed in section 2.4.3, generally referred to as multiple-testing procedures, have been developed to deal with multiple-testing and the inherent problem of false-positives. They consist in reassessing probabilities obtained from statistical tests by considering more interpretable and suited statistical confidence measures.

Regularization approaches, also known as penalized least squares or penalized maximum likelihood, are other types of method widely used to cope with variable selection in high-dimension. A current strategy of such approaches is to exploit the *sparsity*. The main assumption underlying sparse modeling is that only few genes carry relevant information over thousands of genes. By focussing on relevant variables, sparse approaches allow the complexity of the model to be reduced, consequently enhancing the interpretability of the resulting models. In Chapter 4 of this thesis, we focus on penalized likelihood approaches in the context of network inference. In particular, we discuss the Lasso (Least Absolute Shrinkage and Selection Operator) regularization technique introduced by Tibshirani (1996) and various extensions.

2.2 Gene expression microarrays

Microarray is a general term that refers to a set of various chip-based technologies used to measure biological components in high-throughput. Most common applications of microarray technology include gene expression profiling but also single-nucleotide polymorphism (SNP) detection, CGH microarrays to identify genomic copy number variations as well as ChIP-chip, dedicated to the investigation of protein binding sites or more recently DNA methylation arrays to assay methylation patterns. In the following section, we focus on gene expression microarrays that enable the simultaneous measurement of mRNA abundance of tens of thousands of genes.

2.2.1 Hybridization

The core principle behind the microarray technology is based on a physico-chemical property of DNA and RNA molecules. As we have previously discussed in section 2.1.1, the DNA exhibits a double helix structure, whose strands form hydrogen bonds between their complementary bases, Adenine/Thymine or Cytosine/Guanine. Like in the DNA molecules, base-pairing occurs in RNA between Adenine and Uracil or Cytosine and Guanine. This property of nucleotides is known under the term of *hybridization*.

2.2.2 Microarray protocol

A microarray consists of a solid support, typically a glass chip or a nylon membrane, and a set of oligonucleotide strands that are immobilized (attached to the support or directly synthesized) on its surface in known positions. These short sequences are called *probes* and each gene is normally represented by a set of probes or a *probeset*, that map to different gene regions. A variety of expression microarrays are used in the literature but basically, they are all made using the same three-steps protocol:

1. The mRNA samples are first extracted from the cells of interest.
2. The oligonucleotides in the samples, called *targets*, are then labeled either radioactively or fluorescently.
3. Finally, they are hybridized to the array.

In the context of comparing gene expression levels between two samples, two strategies can be distinguished: (i) hybridize a mixture of the two mRNA preparations simultaneously to a common array by labeling the samples either with a green or a red fluorescent dye (this involves a competitive hybridization of targets) or (ii) hybridize each sample separately on two different arrays with a single label. We will particularly focus on the latter strategy, whose one of the main representative are Affymetrix GeneChip arrays [3], for which the probes are synthesized in situ on the support using a technique called photolithography. This process allows a stepwise synthesis $25-30$ base long probes. A mRNA molecule is typically represented on Affymetrix arrays by a probeset containing from 11 to 20 unique probes. Millions of copies of a single oligonucleotide are synthesized in probe cells called *spots*. For instance, the GeneChip Human Genome U133 Plus 2.0 Array used for profiling the Bos dataset is composed of about $54,000$ probesets composed of 11 different probes that are randomly distributed across the array and comprised more than $1,300,000$ distinct oligonucleotide features.

2.2.3 From microarray to gene expression data

Image analysis

Once the hybridization step is performed, the array is scanned at a high-spatial resolution. The resulting image exhibits a set of spots of fluorescence such that each spot is represented by many pixels. A step of *segmentation* allows to identify and characterize pixels as signals, inside a given spot or as background, outside the spot. Finally the *quantification* enables to obtain a single overall intensity value, that reflects the abundance of mRNA, simultaneously for each spot. The intensity values are then saved in a so called CEL-file.

Gene expression data

A typical microarray experiment generates an $n \times p$ matrix of expression levels, denoted $X = (X_{ig})$, where each column g corresponds to a variable,

[3] www.affymetrix.com

Probe set ID/samples	sample 1	sample 2	sample 3	sample 4
229819_at	5.1	5.3	4.5	3.8
204639_at	5.9	5.9	6.6	5.7
203440_at	5.3	4.5	4.8	4.3
212607_at	8.7	5.8	8.6	5.0
235473_at	4.6	5.9	4.6	5.7

Table 2.1 – ***Microarray data.***
Expression levels associated to each probe (in row) for samples (in column).

roughly a gene, and each row i corresponds to an array, also referred to as *sample* in this manuscript. In statistical terms, we have n observations, each being a realization of a p-dimensional random variable, see Table 2.1.

The expression levels generally exhibit asymmetric long-tailed distributions. Thus, rather than working with raw values, we usually consider log-transformed expression levels, using a logarithm to base 2 in order to make the distribution more symmetrical. Traditionally, the log-transformed expression measurements are seen to follow rough Gaussian distributions.

2.2.4 Data preprocessing

Preprocessing constitutes the initial stage in the analysis of microarray data to ensure the reproducibility and reliability of the downstream analysis. The ultimate goal of preprocessing is to control the effects of systematic error while retaining full biological variation. Indeed, non-biological factors can contribute to the variability of data. Regarding the case of microarray, we observe many sources of variation that are not solely due to the biological effect of interest. The first source of error in microarray experiments is due to non-specific hybridization, *i.e.* hybridization of other sequences than the intended target of a given probe. This phenomenon is problematic as it adds a background intensity, which is not related to the gene expression level. Systematic technical biases such as variation in preparation and hybridization of samples (temperature fluctuation, dye incorporation...) or measurement errors in scanning contribute to variability in microarray data. Thus, to ensure highly reproducible analyses and accurate estimates of signal intensities, these technical biases should be corrected through a normalization step. Even if the preprocessing issues are comparable for other types of microarrays, the procedures are often platform specific. Therefore, we dedicate this section to the particular case of Affymetrix gene expression data as we mainly worked on this technology. In the sequel, the term microarray will thus refer to Affymetrix GeneChips arrays .

There is a considerable amount of literature regarding microarray preprocessing techniques among which Microarray Suite 5 approach (MAS5.0) from Affymetrix (2002) and the Robust Multi-array Analysis (RMA) procedure from Irizarry et al. (2003) are the most popular for Affymetrix data.

Comparison studies highlight that RMA exhibits better results than MAS5.0, see Barash et al. (2004) or Harr et Schlatterer (2006). Thus, we detail here the RMA method and its extension GC-RMA, that we use in the remainder of the thesis, and particularly in Chapter 5, to preprocess microarray data.

RMA background correction

The idea behind the RMA background correction is to define the background signal as a combination of non-specific hybridizations and optical noise. They assume a common mean background level on each array, meaning that there is no spatial effect. The authors state that the measured intensity can be modeled as a sum of a normally distributed background and a signal of interest. Let us I_{ijg}, B_{ijg} and S_{ijg} respectively denote the intensity measure, the background component and the signal in sample i for the jth probe of the gene g. The exponential-normal convolution model from Irizarry et al. (2003) can be expressed as follows:

$$I_{ijg} = B_{ijg} + S_{ijg},$$

where $\mathbb{E}(B_{ijg}) = \beta_i$. In addition, S_{ijg} is assumed to be exponential, and hence positive, with rate parameter α.

Given this model, the authors are able to derive an expression of the expected true signal, $\mathbb{E}(S_{ijg}|I_{ijg})$, which is used as the background corrected intensity for the jth probe of the gene g in each array.

GC-RMA, an extension of RMA developed by Wu et al. (2004), uses a more sophisticated background correction. It relies on the observation that probes often display different affinities to the target sequences. In particular, GC content, *i.e.* the proportion of G and C bases of probe sequences, can significantly affect the intensity level. Thus, GC-rich probes exhibit a higher non-specific hybridization. Consequently, GC-RMA introduces a background correction that incorporates a probe affinity term that is dependent on base composition and the position of each base along the probe.

Quantile normalization

The RMA procedure includes a quantile normalization described by Bolstad et al. (2003). This step of normalization is crucial when comparing intensity levels between various arrays. Indeed, it allows intensities to be made comparable across arrays by removing potential unwanted effects. The assumption behind the quantile normalization is that the intensity levels of each array originate from the same distribution because the RNA populations hybridized to the arrays should be identical. Indeed, in practice we expect only a few genes to be differentially expressed. The idea is then to give each array the same distribution by forcing the values of quantiles to be equal.

A Quantile-Quantile plot, or QQ-plot, can be used as a tool to determine if two samples come from the same distribution. If they are from the same distribution then the quantiles line up on the diagonal. The method is driven by the concept that a QQ-plot shows that the distribution of two data vectors is the same if the plot is a straight diagonal line and different if it is other than a diagonal line. When working with a large number of samples, the boxplot is more appropriate to visualize the effect of normalization as illustrated in Figure 2.5.

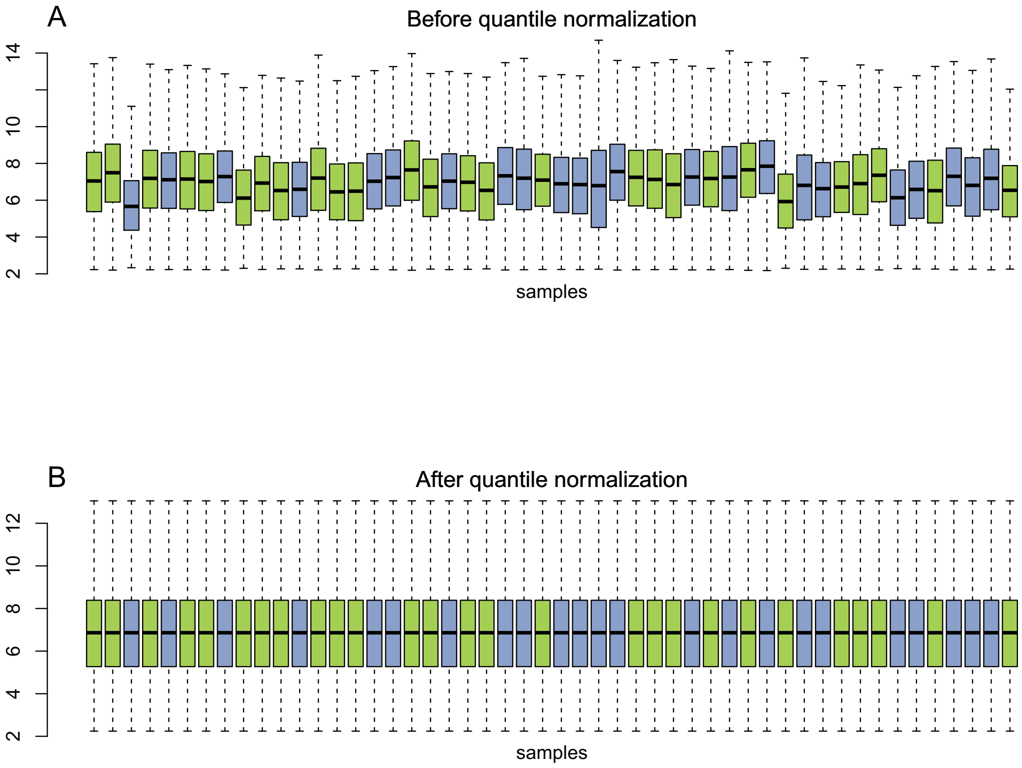


Figure 2.5 – ***Effects of Quantile normalization.***
Boxplots of gene expression distribution for a subset of BR (in green) and notBR (in blue) samples of the Bos dataset, before (A) and after (B) normalization by the quantile approach.

A possible issue with quantile normalization is the strong assumption that the distributions of probe intensities are identical. However, in specific cases, for instance when comparing different tissues, the quantities of RNA transcripts vary a lot. A quantile normalization might be likely to drastically weaken the biological effects between samples. In this case, other normalization methods should be indicated.

2.3 RNA deep sequencing (RNA-seq)

Next-generation sequencing (NGS) technologies have recently emerged as a revolutionary tool in Molecular Biology and have completely transformed the way whole transcriptome analyses can be done. Deep-sequencing of RNA, or RNA-seq, allows the assessment and quantification of mRNA by generating the sequence of transcripts in high-throughput. This technique has therefore, over the past 5 years, become an attractive alternative to microarrays for cell transcriptome investigation. RNA-seq offers the advantage to query all transcripts on a genome-wide scale, allowing the identification of previously unknown exons. This is not the case in microarray experiments for which a transcript can be detected only when there is a corresponding target on the array. RNA-seq gives rise to a wide range of novel applications, including detection of alternative splicing isoforms, transcript fusion detection as well as strand-specific expression analysis. The concepts and principle of RNA-seq will be further discussed in the following sections. As done for microarrays, we will detail the process to generate expression data from deep sequencing technologies.

2.3.1 Data production workflow

Roche, Illumina (initially Solexa) and Life Technologies, among others, have developed well-established platforms for deep sequencing. As most of the published work in RNA-seq studies has taken place primarily for the Illumina platform we focus on this technology in the following section. However despite their technological differences, the three platforms rely on similar workflows for the production of data and the process detailed in the next paragraphs shares common steps with Roche and Life Technologies processes. Where possible, we illustrate the workflow described in the next few paragraphs using a dataset described in Strub et al. (2011) and called the *Strub dataset*. In this study, the authors investigate the effect of Micropthalmia Transcription Factor (MITF) on a human melanoma cell line (501Mel), where gene expression in this cell line was observed following small interfering RNA-mediated MITF knockdown (siMITF) as compared to control siluciferase (siLUC) cells. A goal of this survey is to identify of MITF-regulated genes. Data were sequenced on the Illumina GAII platform. The authors provide gene expression profiles for about 28,000 genes under two conditions labeled "MITF" and "siLUC", each with two replicates.

Library preparation

In a similar process to microarray, the RNA content is first extracted from a tissue of interest or a culture, as is the case with the Strub dataset. The RNA population is then converted into cDNA, representative of the RNA molecules, for stabilization issues. The next step consists of randomly fragmenting cDNA sequences and ligating *adapters* that typically have a known sequence. This collection of cDNA is termed the *library*. Complementary sequences to the adapters are attached to a glass surface, called a *flow cell*, and enable to hybridize the library within the flow cell. In

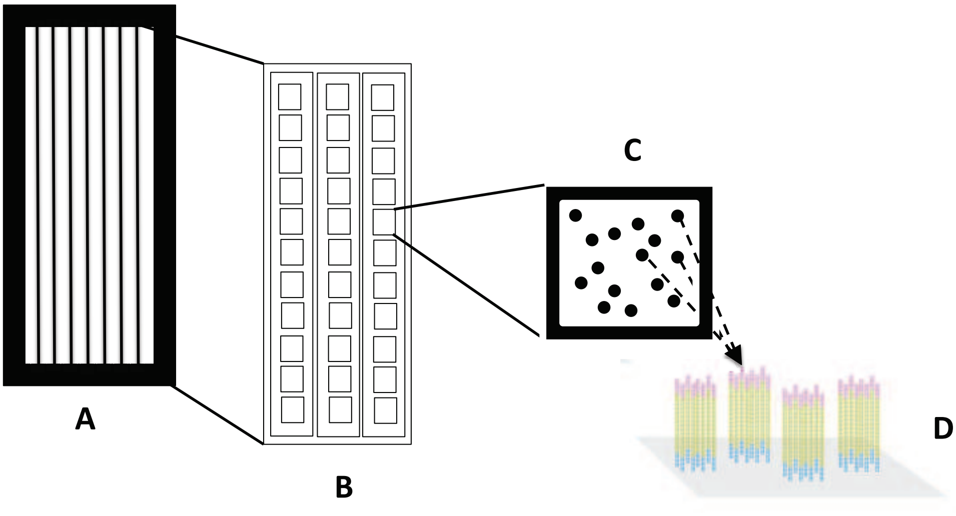


Figure 2.6 – ***Flow cell.***
(A) A flow cell, a small glass slide, consists of eight lanes physically separated from each other. (B) There are three columns of tiles in each lane and (C) each column contains one thousand tiles. (D) A tile holds up to millions of DNA clusters, which consists of identical copies of a template molecule.

addition, to initiate the sequencing, the DNA polymerase enzyme needs a *primer* to incorporate the first nucleotide. These primers are short chemically synthesized oligonucleotides that are added to cDNA sequences.

The Illumina flow cell contains eight independent sequencing areas called *lanes* that are physically separated from each other, allowing eight separate samples to be processed at the same time. Lanes are further broken down into *tiles* in which individual hybridized library fragments are amplified, generating up to 1,000 identical copies, called a *cluster*. In the Strub dataset, a PCR amplication was performed using the following protocol: (i) the activation of polymerases and initial denaturation is accomplished in 30 sec at 98°C, (ii) then 13 cycles of 10 sec at 98°C, 30 sec at 65°C and 30 sec at 72°C, (iii) followed by 72°C for 1 min. After PCR amplification, PCR products were purified using AMPure beads (Agencourt Biosciences Corporation). This amplification step aims to increase the signal intensity that will be measured. Each tile contains hundreds of thousands of clusters spatially distributed as illustrated on Figure 2.6.

In some studies, an alternative process is used to generate short fragments from both ends of each cDNA sequences, resulting in a "paired-end" library. The library preparation is identical to what we described in the previous paragraphs but with a modified template. Indeed, two unique priming sites are introduced into the cDNA fragments during preparation to allow the hybridization of two sequencing primers. Thus, sequencing can occur from both ends.

Parallel sequencing

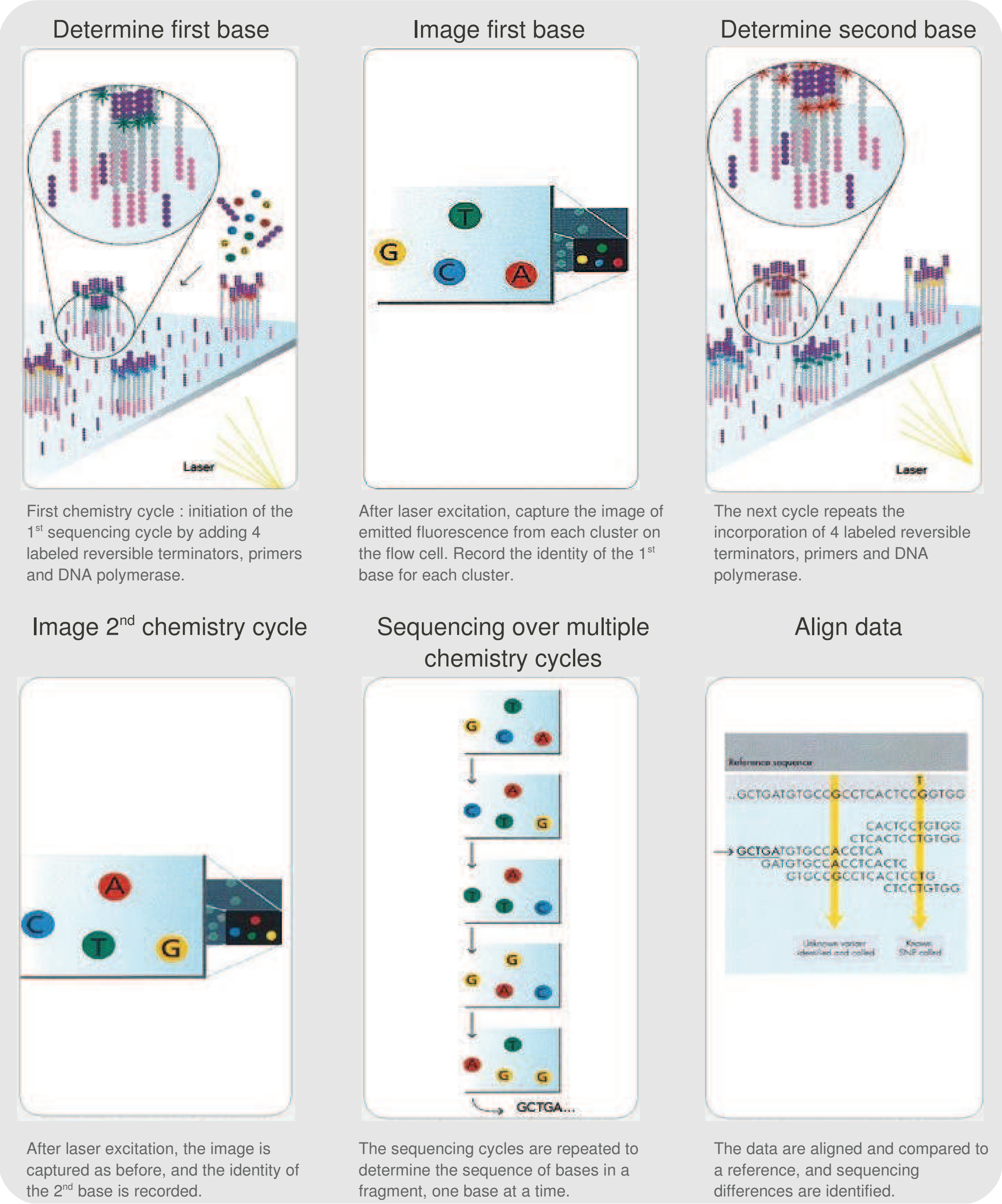
Illumina sequencing is based upon *sequencing-by-synthesis* technology from Solexa to monitor the extension of millions of DNA strands in parallel. This technology, illustrated on Figure 2.7, uses modified nucleotides that are designed with fluorescently-labeled reversible terminators. In the first step, all four modified bases (A, T, G, C) are added in the reaction mix. Polymerase enzymes are able to select the single correct base to incorporate for all clusters simultaneously and the unincorporated nucleotides are removed. Then, after a laser excitation, the image of emitted fluorescence is captured from each cluster on the flow cell, to identify which of the four bases was added at that position. Labeled terminators are finally cleaved to allow incorporation of the next base. This cycle is repeated, one base at a time, generating a series of images, each representing a single base extension at a specific cluster. As each sequencing cycle provides one base of sequence, the length of the synthesized sequence is determined by the number of sequencing cycles.
Fragments of the Strub dataset were sequenced on the Illumina Genome Analyzer II platform as single-end 54 bases reads.

2.3.2 Image processing

The sequencing run generates data in the form of a series of images, which are analyzed by means of Illumina's Pipeline software package. First, the positions of clusters are identified for each raw image file. Then, both intensity signal and noise levels are calculated. Thereafter, images are filtered, and, in this way, clusters are sharpened, background noise is mitigated, and the scale is adjusted. Raw intensities are then converted into discretized sequences or *reads*, a process known as base-calling. In the Strub dataset, the base-calling was performed using the Illumina pipeline v1.8.0.

2.3.3 Mapping and summarization of reads

In order to quantify the abundance of transcript, the location from which each read originated has to be identified. Two strategies are used, either by mapping the reads onto a genome of reference when an annotated genome is available (for instance, the Human genome) or by assembling RNA-seq reads in the absence of a reference genome or annotations. Alignment and assembly of reads are classic problems in bioinformatics, which gives rise to a large amount of literature. For instance, solutions like BLAST, initially introduced by Altschul et al. (1990) are widely used for long reads such as those generated by conventional sequencing. But, due to the short size of reads produce by NGS technologies, such algorithms do not perform well, which presents new challenges to the bioinformatic community. In particular, it induces only limited overlaps between sequences, making the assembly very difficult and leading to large error rates. Furthermore, short reads are likely to map equally well to several


Determine first base

First chemistry cycle : initiation of the 1st sequencing cycle by adding 4 labeled reversible terminators, primers and DNA polymerase.

Image first base

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the 1st base for each cluster.

Determine second base

The next cycle repeats the incorporation of 4 labeled reversible terminators, primers and DNA polymerase.

Image 2nd chemistry cycle

After laser excitation, the image is captured as before, and the identity of the 2nd base is recorded.

Sequencing over multiple chemistry cycles

The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

Align data

The data are aligned and compared to a reference, and sequencing differences are identified.

Figure 2.7 – ***Sequencing-by-synthesis.***
Figure reprinted from the Illumina website.

Gene annotation/samples	siMITF1	siMITF2	siLUC1	siLUC2
$ENSG00000117748$	373	390	3649	3595
$ENSG00000172399$	1461	1861	193	103
$ENSG00000125148$	292	130	829	2443
$ENSG00000108691$	1205	635	13	12
$ENSG00000113369$	600	652	45	14

Table 2.2 – ***RNA-seq data.***
Expression levels associated with genes, in rows, for each sample of the Strub dataset. Each value corresponds to the number of reads mapped to the corresponding gene.

regions of the genome. Various methods have been developed to deal with new alignments challenges. A comparison of these techniques is beyond the scope of this thesis, but can be found in Grant et al. (2011). Regarding their recommendations, reads of the Strub dataset were mapped onto the hg19 assembly of the human genome using GSNAP (Genomic Short-read Nucleotide Alignment Program) from Wu et Nacu (2010). The number of mapped short reads is referred to as the *library size* or *sequencing depth*.

Once the genomic locations of reads are identified, the next step is to summarize reads mapped into a gene or transcript level count. The basic idea involves counting the number of reads overlapping the exons in a gene by ignoring reads that map to genomic regions outside annotated exons. More sophisticated approaches enable the inclusion of junction reads and unannotated transcripts, see for instance Trapnell et al. (2010). Quantification of gene expression in the Strub dataset was carried out on uniquely aligned reads using HTseq-count [4] with gene annotations from Ensembl release 64.

The summarization step provides a single value of expression for each gene. Let us consider n experiments such as each experiment measures expression levels of p genes. The resulting data can be written as an $n \times p$ matrix called X, for which X_{ig} denotes the expression level for gene g in sample i, see Table 2.2.

[4] `http://www-huber.embl.de/users/anders/HTSeq/doc/count.html`

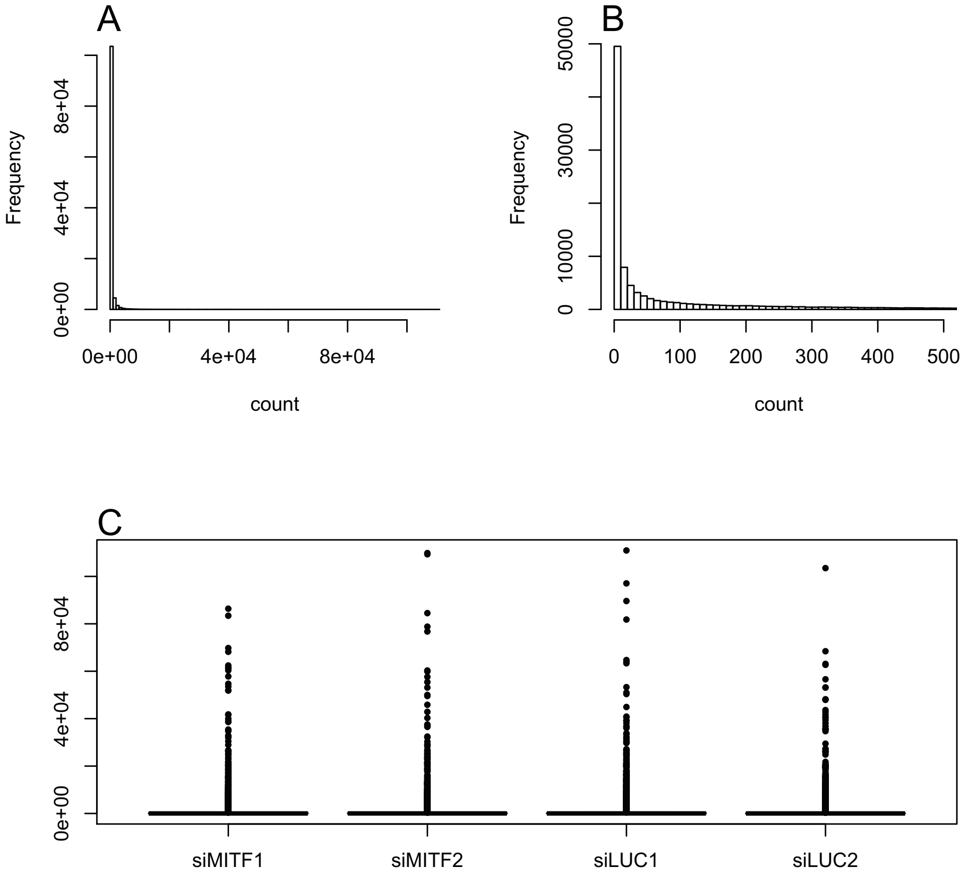


Figure 2.8 – ***RNA-seq data distribution for the Strub dataset.***
(A) The count distribution exhibits a large amount of low counts and only few high count genes. (B) A detailed view of read count distribution between 0 *and* 500*. (C) Boxplots highlight the presence of genes with high read counts for each of the two replicates under the siMITF and siLUC conditions.*

For RNA-seq data, X_{ig} is a nonnegative integer, that is, the number of reads mapped to this gene, resulting in a discrete measurement for gene expression that is usually regarded to follow a Poisson or a negative binomial distribution as illustrated in Figure 2.8.

2.3.4 Normalization

Experience with microarray data has repeatedly shown that normalization is an essential step in the analysis of gene expression. An important advantage of RNA-seq is their ability to allow direct access to sequences of mRNA, avoiding biases due to hybridization and labeling. However, other sources of systematic variation have been reported: (i) between-sample differences such as library size: larger library sizes result in higher counts for the entire sample (ii) within-sample gene-specific effects related to gene length or GC-content. Thus, the basic problem is still the same: how to remove unwanted variations such that any differences in expres-

sion between samples are due solely to biological effects.
Mortazavi et al. (2008) defined the widely used "reads per kilobase per million" or (RPKM), which is a normalized measure of read counts. It involves both a normalization for RNA length and for the total read number in the measurement:

$$RPKM_{ig} = \frac{X_{ig}}{S_i \times L_{ig}},$$

where, X_{ig} is the number reads that have been mapped to a region in which an exon is annotated for the gene g, S_i is the library size for sample i (in millions) and L_{ig} denotes the sum of the lengths of all exons annotated for the gene g, measured in kilobases. Various authors showed that sequencing depth is not a stable scaling factor and a number of more robust alternatives were suggested. We discuss some of the inter-normalization approaches proposed in the literature in section 6.1 and provide a comprehensive comparison of these methods.

2.4 Statistical background

2.4.1 Concepts of hypothesis testing

Null and alternative hypotheses

Basically, statistical hypothesis testing is a process whereby one tests a claim made about one or more population parameters, which is formulated from a (biological) question of interest. Let us take an example, where we are interested in determining if the prevalence of a disease, the diabetes for instance, is the same in two populations. From a statistical point of view, the biological question can be translated into two hypotheses, called H_0 and H_1, concerning the proportion of people with disease:

$$\begin{cases} H_0: & \text{"the proportions of cases are equal in both populations"} \\ H_1: & \text{"the proportions of cases differ between both populations"} \end{cases}$$

The hypothesis denoted H_0 is called the *null hypothesis* and merely claims that "nothing is going on". It will generally involve an equality of population parameters. Let us return to the example of diabetes and denote μ_1 and μ_2 the numbers of cases estimated from the sample of two populations denoted 1 and 2. The null hypothesis is defined as follows:

$$H_0 : \mu_1 = \mu_2.$$

H_1 refers to the *alternative hypothesis* and is expressed as an inequality or an inequation and has one of theses three forms:

$$\begin{array}{lll} H_1: & \mu_1 \neq \mu_2 & (1) \\ H_1: & \mu_1 < \mu_2 & (2) \\ H_1: & \mu_1 > \mu_2 & (3) \end{array}$$

The first form (1) refers to a two-sided test. It is performed if there is no prior information regarding the direction of the alternative, while a one-sided test, such as expressed in (2) or (3), specifies in its alternative hypothesis that the parameter is either greater than or less than the value specified in the null hypothesis.

Test statistic

The approach of hypothesis testing assumes that H_0 is true, and then looks for evidence that it is not true, by examining the likelihood of observing such data under the null hypothesis. The decision whether to accept or reject the null, is based on a function of the observed data $T(X)$, called the test statistic, which measures the distance between H_0 and the data.
In case of a right-tailed test for instance, the chosen test statistic should be greater under H_1 than under H_0.

Null distribution

The null hypothesis is used to derive the *null distribution*, *i.e.* the probability distribution of the test statistic $T(X)$ under H_0. In standard cases, the null distribution of the statistic is known or may be approximated. In other cases, when the null distribution is unknown, it has to be estimated from the data by resampling procedures or permutation tests that make use of rearrangements of sample labels. The value of the test statistic is then calculated for all possible rearranged samples (or for a large random sample thereof), providing an empirical null distribution.

Rejection region

The hypothesis testing procedure compares the observed statistic, computed from the data, to the null distribution and enables to specify a rejection region W, *i.e.* a set of values of the test statistic for which the null hypothesis is rejected. For instance, in a two-sided test, rejection occurs for both large and small values of $T(X)$ and the rejection rule can be therefore defined as:

$$W_1 = \{X : T(X) > c_1 \text{ or } T(X) < c_2\},$$

where c_1 and c_2 are critical values for rejecting the null hypothesis. In a one-sided test rejection occurs either for large or small values of the test statistic (but not both) as dictated by the alternative hypothesis. The corresponding rejection region are of the form: $W_2 = \{X : T(X) > c_1\}$ for a right-tailed test or $W_3 = \{X : T(X) < c_2\}$ for a left-tailed test.

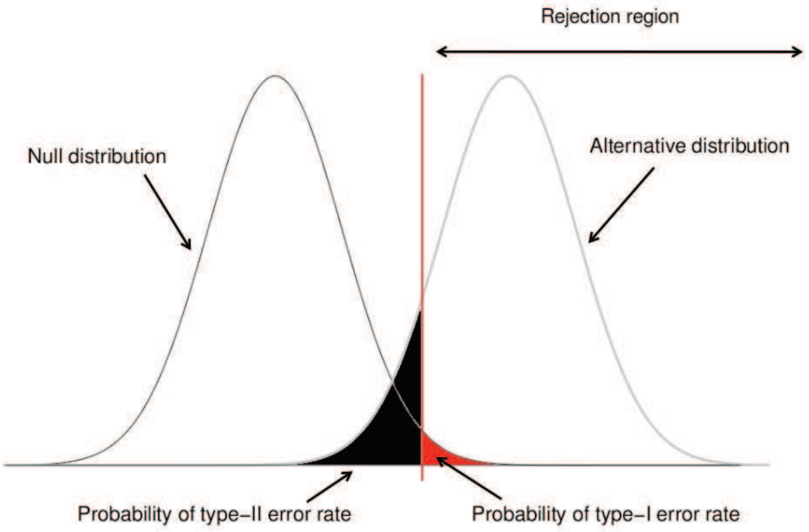

Rejection region
Null distribution
Alternative distribution
Probability of type–II error rate
Probability of type–I error rate

Figure 2.9 – ***Probability of errors and rejection region.***
Illustration of the notions of type-I error rate (red area), type-II error rate (black area) and rejection region in a right-tailed test. The red line represents the critical value for rejecting the null hypothesis.

Errors

There are two possible outcomes to a test: H_0 is rejected because the statistic is in the rejection region or H_0 is not rejected because there is not enough evidence to reject it in favor of H_1. Naturally, if the test either fails to reject a true null hypothesis or rejects a false null hypothesis, it acts correctly. But, when performing a hypothesis test, two types of errors can arise: (i) a type-I error (or a false positive), if the test rejects a true null hypothesis (ii) a type-II error (or a false negative), if the test does not reject a false null hypothesis. The true state and the decision to accept or reject a null hypothesis are summarized in Table 2.3. The probability of making a type-I error, α, is the probability that the data is in the rejection region conditional upon assuming the null hypothesis, while β is the probability of making a type-II error, *i.e.* the probability of being out of the rejection region given that the null hypothesis is false.

Ideally, one would minimize both α and β to eliminate the possibility of false-positive and false-negative results. However the type-I and II errors are inversely related: the smaller the risk of one, the higher the risk of the other, making it impossible to set both parameters at zero. Thus, there is a necessary trade-off between type-I and II errors to accept a reasonable risk of committing either type of error. In most cases, the research questions make it particularly important to avoid a type-I error. That is why it is common to first control type-I error probability at a specified level α such as the maximal probability of making a type-I error is less than or equal to α. This level is referred to as the *significance level*.

Once the significance level has been fixed, it is used to identify both the critical value and the critical region of the test statistic. For instance, in a two-tailed test, c_1 and c_2 are chosen so that:

$$\mathbb{P}_{H_0}(T(X) > c_1) + \mathbb{P}_{H_0}(T(X) < c_2) \leqslant \alpha.$$

There may be many ways in which the sum of these two terms can satisfy the inequality and one has to make a decision regarding how to divide the probability α between the two terms. Usually, when one has no prior on the direction of the alternative hypothesis, it seems appropriate to divide this total probability symmetrically between the two tails. This is, the condition $\mathbb{P}_{H_0}(T(X) > c_1) = \mathbb{P}_{H_0}(T(X) < c_2)$ is imposed and therefore:

$$\begin{aligned} \mathbb{P}_{H_0}(T(X) > c_1) &= \mathbb{P}_{H_0}(T(X) < c_2) \\ &\leqslant \tfrac{\alpha}{2}. \end{aligned}$$

The null hypothesis is rejected in favor of the alternative hypothesis if the observed statistic is greater than c_1 or less than c_2. In case of a symmetrical distribution of the statistic under the null hypothesis, the critical values are equal in absolute value.

When the null hypothesis is tested using an α-level one-sided test to the left, the critical value c_2 is defined so that $\mathbb{P}_{H_0}(T(X) < c_2) \leqslant \alpha$. In a right-tailed test the critical value c_2 is simply determined such as $\mathbb{P}_{H_0}(T(X) > c_1) \leqslant \alpha$.

Reality \ Decision	H_0 not rejected	H_0 rejected
H_0 true	true-negative $(1-\alpha)$	false-positive / type-I error (α)
H_0 false	false-negative / type-II error (β)	true-positive $(1-\beta)$

Table 2.3 – ***Outcomes of a statistical test performed at the level α.***

Power

The quantity $1-\beta$, called the power, denotes the ability to reject H_0 when it is actually false:

$$\begin{aligned} \text{Power}(\alpha) &= \mathbb{P}_{H_1}(H_0 \text{ rejected at the } \alpha \text{ level}) \\ &= 1-\mathbb{P}_{H_1}(H_0 \text{ not rejected at the } \alpha \text{ level}) \\ &= 1-\beta \end{aligned}$$

The power depends directly on three factors:

1. Firstly, it is influenced by the level of significance, α. Indeed, a larger α results in a smaller probability of committing a type II error which thus increases the power.

2. Statistical power is also a function of the sample size. Larger sample sizes enable the standard error to be decreased, resulting in an increased power. However, the practical realities of conducting research, particularly with regards to financial costs, restrict the size of samples for most researchers. Power analysis can be used at the research design stage to calculate the minimum sample size required to ensure that the test will have a sufficient power to detect biologically important effects.

3. Finally, power is affected by the "size effect", *i.e.* the magnitude of the effect under the alternative hypothesis. The greater the difference between the null and alternative hypotheses distributions, the greater the power of the test. That is to say, large effect sizes will increase the power of the test. In practice, the goal of hypothesis testing is to maximize power, thus an appropriate balance among these factors must be found.

p-value

The hypothesis testing procedure enables us to decide whether or not to reject H_0 by calculating a test statistic and a rejection region. But this decision is not so informative and there is a need to quantify how much evidence there is against the null hypothesis for a given test statistic. It is usually expressed in terms of a probability called the p-value. The concept of p-value is defined as the probability under H_0 of observing a more "extreme" value of the test statistic. The sense of the term "extreme" varies depending on the type of test which is performed. For instance, for a right-tailed test the p-values is given by the formula:

$$\text{pv} = \mathbb{P}_{H_0}(T \geqslant t),$$

such as t is the observed statistic value. In this case, "extreme" means "larger" but in a left-tailed test it would have meant "smaller". A p-value smaller than α is said to be *significant* and implies the rejection of the null hypothesis at the α level of significance.

2.4.2 Differential analysis for transcriptome data

The purpose of differential analysis is the identification of differentially expressed genes, *i.e.* genes whose expression levels differ from one condition to another. To know which genes are differentially expressed between conditions is of crucial importance for any biological interpretation. The core tool for differential analysis consists of statistical tests. Based on the various concepts previously detailed, we will describe in the following section, the procedure of differential analysis.

The first step in a hypothesis testing procedure is to specify the model as well as the null and alternative hypothesis. Differential analysis involves testing the null hypothesis (H_0) that the expected values of expression for a given gene are equal between the two (or more) conditions to compare, against the alternative hypothesis (H_1) that they differ. The strategy of testing is the same for microarray and RNA-seq data. However as we need to define a statistical model for the data, we illustrate the testing procedure using the Bos microarray dataset. We are thus interested in identifying genes that are implicated in tumors relapsing to bone in breast cancer.

Let $X_{ig}^{(c)}$ be the expression level of the ith sample for gene g under condition c, with $c \in \{1,2\}$. We arbitrarily denote by 1, the patients who experienced a bone relapse and by 2, the patients who do not. Under the assumption of heteroscedasticity (*i.e.* inhomogeneity of variance) between conditions, the general model is given by:

$$\mathbb{E}(X_{ig}^{(c)}) = \mu_g^{(c)} \quad \text{and} \quad \mathbb{V}(X_{ig}^{(c)}) = (\sigma_g^{(c)})^2, \tag{2.1}$$

where $\mu_g^{(c)}$ and $(\sigma_g^{(c)})^2$ are respectively the expected level of expression and the variance of gene g under condition c. So defined, the null hypothesis to test for two conditions comes down to:

$$\begin{cases} H_0 : & \mu_g^{(1)} = \mu_g^{(2)}, \\ H_1 : & \mu_g^{(1)} \neq \mu_g^{(2)}. \end{cases}$$

In order words, under the null hypothesis the expected values of expression for a given gene are equal for the two conditions, whereas they differ under the alternative hypothesis.

The second step in the procedure consists in selecting a significance level α. In practice, levels of 0.05 or 0.01 are the accepted standards.

The third step is to define the test statistic. The t-statistic is certainly the most popular statistic and merely consists in a normalized difference of means. There are two versions of the t-test that have been widely used in the literature of microarray data analysis. The first one, the Student t-statistic, assumes that the two populations being compared have the same variance: $(\sigma_g^{(1)})^2 = (\sigma_g^{(2)})^2$ and uses a pooled variance estimator, which consists in a weighted average of the two sample variances. The second version, called the t-statistic of Welch and denoted t_g^{welch}, is given by the formula:

$$t_g^{\text{welch}} = \frac{\bar{x}_{.g}^{(1)} - \bar{x}_{.g}^{(2)}}{\sqrt{\frac{(S_g^{(1)})^2}{n_1} + \frac{(S_g^{(2)})^2}{n_2}}},$$

where n_1 and n_2 are the number of samples relative to conditions and $\bar{x}_{.g}^{(1)}$ is the natural estimator of $\mu_g^{(1)}$, *i.e.* the average expression level for gene g under condition 1. $(S_g^{(1)})^2$ and $(S_g^{(2)})^2$ are the usual unbiased estimators of the variance $(\sigma_g^{(1)})^2$ and $(\sigma_g^{(2)})^2$, respectively, heterogeneous between conditions. The Welch t-statistic is calculated and displayed in Table 2.4, for the first genes of the Bos dataset.

Gene	Statistic	p-value
$A1BG$	0.48	0.63
ADA	-3.7	0.00031$*$
$CDH2$	-0.89	0.37
$AKT3$	0.12	0.9
$MED6$	2.5	0.013$*$

Table 2.4 – ***Differential analysis results at a* 5% *level***
Results of a differential analysis conducted on the first genes of the Bos dataset at a 5% *level. The Welch t-statistic and the p-value are provided for each gene. Significant p-values are flagged with a "*".*

In the fourth step, two strategies are possible: (i) identify the rejection region and compare the test statistic to the critical values or (ii) define the rejection rule with reference to the calculation of the p-value associated to the test statistic. We usually go through the second procedure for two main reasons:

1. while the rejection region provides a dichotomous outcome (rejection or not), the p-value enables to reflect the strength of results,

2. the use of p-value is more convenient as we do not need to redefine the rejection rule if the significance level α is changed.

Moreover, the p-value allows the comparison across different statistics.

Once the p-value computed, the last step of the procedure is to reach a conclusion about the rejection of H_0 or not. Significant results, *i.e.* those with p-value less than α, will lead to the rejection of the null hypothesis. Non-significant results do not allow the acceptance of the null hypothesis because they do not necessarily imply that there is no difference of means between the two populations. Indeed, as Greenwald (1975) pointed out, there are "*many ways (including incompetence of the researcher), other than the null hypothesis being true, for obtaining a null result*". This includes insufficient sample sizes and small effect sizes as discussed in the previous subsection.

2.4.3 Multiple-testing

The procedure described in the previous section shows how a difference of expression can be scored, and how the decision to declare a gene differentially expressed can be taken for one gene, controlling the probability of having a false positive at the α-level of significance. Nevertheless, the reality of microarray data is much more complicated, since thousands of hypothesis tests are simultaneously evaluated making error rates substantially harder to control. Indeed, as the number of hypotheses being tested increases, so does the overall chance of making an error and the simple use of a significance test without adjustment for multiple comparisons could lead to a large chance of false-positive findings. Let us illustrate this point with an example where m tests are conducted with the α-level. Depending on whether each hypothesis tested is true or false and whether the statistical tests reject or does not reject the null hypotheses, each of the m results will fall in one of four possible outcomes: (i) failing to reject when H_0 is true, (ii) failing to reject when H_0 is false, (iii) rejecting when H_0 is true, or (iv) rejecting when H_0 is false. We define the frequency at which each occurs to be: TN, FN, FP and TP respectively, as outlined in Table 2.5.

Reality \ Decision	H_0 not rejected	H_0 rejected	total
H_0 true	TN	FP	m_0
H_0 false	FN	TP	m_1
total	m_U = TN + FN	m_R = FP + TP	m

Table 2.5 – ***Outcome when testing*** m ***hypotheses.***
TP: true-positives; TN: true-negatives; FP: false-positives; FN: false-negatives.

Assuming that m_0 of the null hypotheses are true such as $m = m_0$, the number of false positives FP is a random variable, with Binomial probability distribution such as:

$$FP \sim \mathcal{B}(m, \alpha).$$

The expected number of false-positives is $\mathbb{E}(FP|H_0) = m\alpha$ and the probability of having at least a false-positive, assuming that all tests are independent, is given by:

$$\begin{aligned}\mathbb{P}_{H_0}(FP \geqslant 1) &= 1 - \mathbb{P}_{H_0}(FP = 0) \\ &= 1 - \tbinom{m}{0}\alpha^0(1-\alpha)^m \\ &= 1 - (1-\alpha)^m.\end{aligned}$$

For $\alpha = 0.05$, the function $(1-\alpha)^m$ decreases monotonically and tends quickly to zero as m grows larger. In consequence, the probability of having at least one false-positive grows rapidly as m increases. For instance, for $m = 10$, $\mathbb{P}(FP \geqslant 1) = 40\%$. Thus, when more than $10,000$ tests are performed, as is the case with the Bos dataset, the probability of having at least a false-positive is equal to one. The goal of multiple-testing procedures is to provide more appropriate measures of error in order to control the global risk rather than the individual risk associated with each test. We detail both of these in the following subsections, namely the Family-Wise Error Rate (FWER) and the False-Discovery Rate (FDR), which are widely used by the community. Before that we focus on the analysis of

p-value distribution which provides an intuitive approach to qualitatively assess the evidence of true-positives.

Analyzing the distribution of p-values

Investigating the distribution of p-value is a necessary first step when dealing with multiple tests as it provides a qualitative way for assessing the proportion of tests declared under the null and alternatives hypotheses.

A fundamental property of a statistical hypothesis test is that p-values follow under the null hypothesis the standard Uniform distribution. This property allows for precise, unbiased evaluation of error rates and statistical evidence in favor of the alternative. On the other hand, the alternative distribution of p-values corresponds to a distribution that tends to accumulate toward 0. In practice, the density function of p-values, denoted $f(p)$, is then expressed as a two-component mixture of null and alternative densities, respectively named f_0 and f_1:

$$f(p) = \pi_0 f_0(p) + \pi_1 f_1(p),$$

where π_0 and π_1 are the proportions of p-values generated under H_0 and H_1 such as $\pi_0 + \pi_1 = 1$. Note that, $\forall p, f_0(p) = 1$. We provide on Figure 2.10 a graphical representation of a mixture distribution of p-values along with the potential outcomes of the corresponding statistical tests. In comparison, we display on Figure 2.11 the distribution of p-values resulting from a Welch t-test on the Bos dataset. It exhibits a significant accumulation of p-values close to zero, suggesting a large number of differentially expressed genes.

Controlling the Family-Wise Error Rate

The first alternative confidence measure involves intuitively controlling the probability of falsely rejecting at least one null hypothesis over the collection (or "family") of hypotheses, denoted $\{H_0^g\}_{1 \leqslant g \leqslant m}$, that is being considered for joint testing at the level α. This definition referred to as the Family-Wise Error Rate (FWER) criterion:

$$\text{FWER}(\alpha) = \mathbb{P}_{H_0}(FP \geqslant 1).$$

Thus, the FWER is equal to the type-I error rate when testing only a single hypothesis. In practice, as the number of tests increases, the type-I error rate remains fixed at the level α whereas the FWER tends toward 1. Recall that the the FWER can be computed directly for m independent tests such as:

$$\text{FWER}(\alpha) = 1 - (1 - \alpha)^m.$$

By solving this equation for α, Šidák concluded that performing each test at the level $1 - (1 - \alpha)^{\frac{1}{m}}$ ensures the global control of the FWER at level α. This procedure is called the Šidák correction. The main criticism of this procedure is that it is based on the assumption that the tests are independent, whereas it is obviously not the case in transcriptomic studies

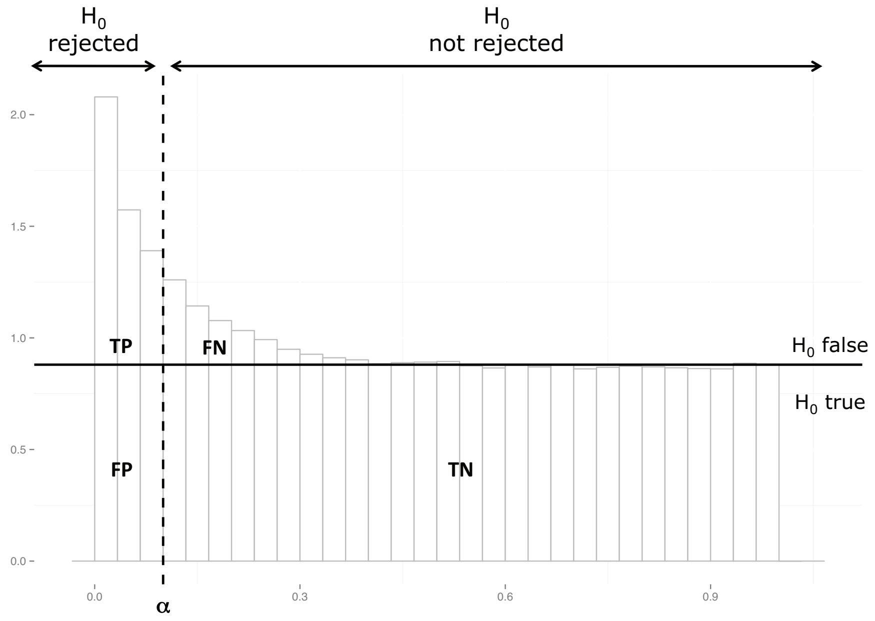


Figure 2.10 – ***Mixture distribution of p-values***
Simulated distribution of p-values as a mixture of a $\mathcal{U}(0,1)$ distribution, corresponding to true nulls and a $\mathcal{B}(0.8,7)$, corresponding to false nulls, with corresponding proportions of true-positives (TP), true-negatives (TN), false-positives (FP), and false-negatives (FN) at the level α.

as many variables are related.
For this reason, Bonferroni (1935) developed another procedure based on the following inequality: $\mathbb{P}\left\{\bigcup_j \left(E_j\right)\right\} \leqslant \sum_j \mathbb{P}(E_j)$ such as E_j is an event in a given probability space. In the context of multiple hypothesis testing of transcriptomic data, we denote FP_j the event: "the jth test is a false-positive at the α level". Thus, for a family of m comparisons, the Bonferroni inequality enables the upper bound of the FWER to be defined:

$$\begin{aligned} FWER(\alpha) &= \mathbb{P}\left\{\bigcup_{j=1}^{m} \mathrm{FP}_j\right\} \\ &\leqslant \sum_{j=1}^{m} \mathbb{P}(\mathrm{FP}_j) \\ &\leqslant m\alpha. \end{aligned}$$

Performing each test at the individual level α/m guarantees the probability of rejecting at least one true hypothesis to be less than or equal to α without assuming that the m tests are independent. The major advantage of this procedure is that it is simple and straightforward to apply and can easily be used in any multiple-testing application. However, the Bonferroni correction suffers from low power, as it leads to very conservative decisions, mainly when the number of hypotheses is very large. Improvements in power may be achieved by considering step-wise procedures such as the step-down version of Bonferroni and Šidák proposed by Holm (1979). In step-wise procedures, p-values are ordered and rejection of tests depends on the outcome of the previous tests of other hypotheses. In particular, in step-down procedures hypotheses that correspond to the

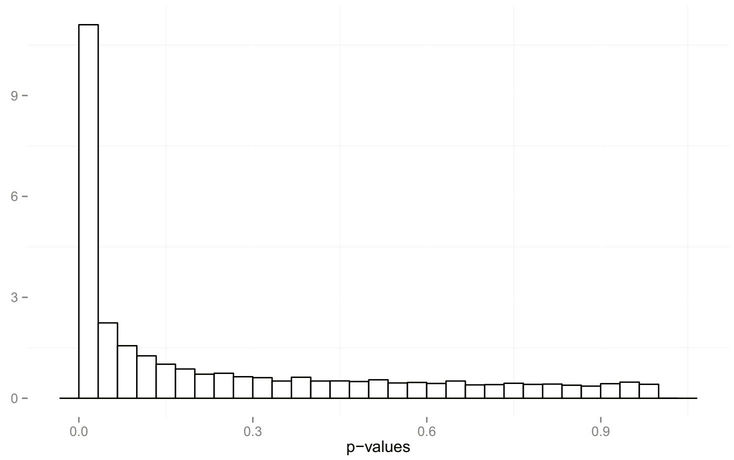


Figure 2.11 – ***Distribution of p-values resulting from a Welch t-test on the Bos dataset.***

most significant p-values are first considered and as soon as one fails to reject a null hypothesis, no further hypotheses are rejected. Another attempt to gain more power is due to Westfall et Young (1993) who designed a step-down resampling algorithm. However, the resultant gain from both improvements is quite small and less stringent criteria was designed to find a proper balance between false-positive and false-negatives.

Controlling the False-Discovery Rate

To overcome the limitations of the FWER, many new concepts of error rate have been developed in the literature of multiple-testing. Benjamini et Hochberg (1995) introduced a more liberal approach: the False-Discovery Rate (FDR), based on the principle that most researchers would tolerate some false-positives in return for a greater statistical power, provided their number is small in relation to the number of rejected hypotheses. Thus, the FDR is defined as the expected proportion of truly null hypotheses that are falsely rejected at the α level:

$$\text{FDR}(\alpha) \quad = \quad \mathbb{E}\left(\frac{FP}{m_R}\mathbb{I}_{\{m_R>0\}}\right),$$

where $\mathbb{I}_{\{m_R>0\}}$ is the indicator function, which equals 1 when $m_R = FP + TP$, the number of rejected hypothesis at the level α, is non-null and 0 otherwise.

Benjamini and Hochberg also proposed a step-up procedure to control the FDR for independent tests. Let $p_1 \leqslant ... \leqslant p_m$ be the ordered p-values of m independent tests. The aim is to identify the maximal threshold i^* such as the null hypothesis is rejected, by considering the ordered p-values successively in a step-up procedure (from the least significant p-value to the most). If the procedure stops at the threshold p_i, then $m_R = rg(p_i) = i$, such as $rg(p_i)$ is the rank of the ith p-value. Moreover, the expected number of false-positives at the level p_i is:

$$\begin{aligned} \mathbb{E}\left(FP(p_i)\right) &= m_0 * p_i \\ &= m\pi_0 * p_i. \end{aligned}$$

Thus, if the procedure threshold is p_i, the FDR is defined as:

$$FDR(p_i) = \frac{m\pi_0 * p_i}{i}.$$

The strategy proposed by Benjamini and Hochberg is to use the fact that $\pi_0 \leqslant 1$ to provide a upper bound of the FDR:

$$FDR(p_i) \leqslant \frac{m * p_i}{i}.$$

To control the FDR at the level α, the hypotheses H_0^g should be rejected if $g \leqslant i^*$, where i^* is defined as:

$$i^* = \max_{1 \leqslant i \leqslant m} \left\{ i : p_i \leqslant \frac{i}{m}\alpha \right\}.$$

As soon as the null hypothesis i^* is rejected, all further hypotheses are also rejected.

One can argue that an upper bound of 1 for π_0 leads to a loss of precision in the estimation of the FDR. Such estimations are actually probably conservative with respect to the proportion of test statistics drawn under H_0 and H_1; that is, if the classical method estimates that the FDR associated with a collection of p-values is 5%, then on average the true FDR is lower than 5%. Consequently, a variety of more sophisticated methods introducing the estimation of π_0 have been developed for achieving more accurate FDR estimations. For instance, Storey (2001) introduced a procedure based on the work of Schweder et Spjotvoll (1982), which estimated the proportion π_0 by the density of p-values exceeding a tuning parameter λ. The minimum bias of $\hat{\pi}_0$, the π_0 estimation, is obtained for $\lambda = 1$, while the variance of $\hat{\pi}_0$ increases as λ tends to 1. In consequence, a compromise has to be made between variability and bias to assess the λ value.

Finally, most FDR-based corrections strongly depend on the assumption that the p-values are uniformly distributed under the null hypothesis, which may not always be the case in practice, for instance when variables are strongly correlated. The impact of dependence between variables on p-values distribution and on the estimation of the proportion of true null is extensively investigated in Friguet et Causeur (2011) by using factor analysis models. The authors show that high levels of dependence lead to instability in multiple testing procedures and to biased estimations of π_0.

To tackle this issue, various methods relying on more advanced statistical and algorithmic notions exist. For instance, Efron et Tibshirani (2002) developed a local version of the FDR, called *local-FDR*, which quantify the probability for a given null hypothesis to be true according to the specific p-value of each gene tested. The semi-parametric approach developed by Robin et al. (2007) and implemented in the R package `kerfdr` from Guedj et al. (2009), uses the null distribution to provide a flexible kernel-based estimation of the alternative distribution.
Causeur et al. (2011) introduce a framework for high-dimensional multiple testing procedures which involves capturing the components of dependence into a low dimensional set of latent variables that are integrated in the calculation of new test statistics. The eventual goal of this approach is to restore independence among tests in order to apply multiple testing

procedures initially derived for independent variables. The method is implemented in the R package `FAMT`.

The statistical concepts described in this section, will be employed throughout the following chapters. In particular, the notions of power and false-positive rate will be widely used as evaluation criteria of the various approaches discussed in this manuscript.

3 Statistical modeling for the identification of molecular signatures

As discussed in the previous chapter, the identification of molecular signatures is of great interest for diagnosis, treatment recommendations or early disease detection. A major statistical issue in high-throughput transcriptome experiments is how to select relevant and robust signatures given the large number of genes under study. In this chapter, we focus on robust gene selection through differential analysis approaches.
In the first section, we introduce a comprehensive comparison of eight hypothesis testing approaches. In particular, we focus on evaluating the relevance of variance modeling strategies. Based on this comparison study, we propose practical recommendations on the appropriate test to be used for the analysis of gene expression data. This work was initiated during my internship at the *Ligue Nationale contre le Cancer* and at the University of *Paris Descartes*, under the supervision of Mickaël Guedj and Gregory Nuel. In the second section we present a new approach, `DiAMS` (Disease Associated Modules Selection), that aims at improving the robustness of signatures across studies. The proposed methodology integrates both Protein-Protein Interactions (PPI) and gene expression data in a local-score approach.

This chapter is associated with the following two publications:

1. Jeanmougin et al. (2010). **Should we abandon the *t*-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies.** *PLoS ONE, 5(9),*

2. Jeanmougin et al. (2012a). **Improving gene signatures by the identification of differentially expressed modules in molecular networks : a local-score approach.** *Proceedings, JOBIM.*

3.1 Comparison of hypothesis testing strategies for microarray data

Once raw data have been preprocessed, there are many different treatments that can be performed on them, but most fall under some category of differential expression analysis. In this section we focus on the case of microarray technology to review and evaluate statistical methods for detecting genes significantly differentially expressed.

3.1.1 Statistical tests for microarray data

The most intuitive heuristic used to identify differentially expressed genes is known as the Fold-Change estimation (FC). It consists in evaluating the average log-ratio of two expression levels under two conditions and considers as differentially expressed all the genes that differ by more than an arbitrary cut-off. Classically, a change of at least two-fold (up or down) was considered meaningful. Using a log2 transformation, a two-fold up- or down-regulation in gene expression is equivalent to log-ratios of $+1$ or -1 respectively (see Figure 3.1). The simplicity of the method established its popularity. The general belief is that greater the FC, the higher the significance of the genes. So defined, FC lacks of a solid statistical footing: it does not take the variance of the samples into account. This point is especially problematic since variability in gene expression measurements is partially gene-specific, even after the variance has been stabilized by data transformation, as carried out in Zhou et Rocke (2005) and Simon et al. (2003). In this context, one should prefer hypothesis testing which provides a formal and efficient framework for identification of differentially expressed genes that enables to overcome FC limitations.

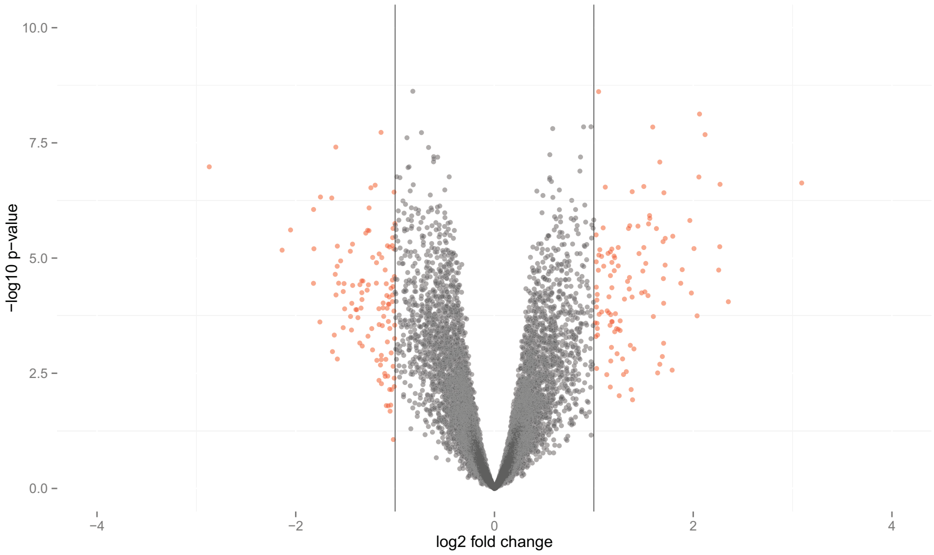


Figure 3.1 – ***Volcano plot.***
The figure highlights in red the genes selected as differentially expressed by the FC approach with cut-off values of -1 and 1, i.e. a two-fold up- or down-regulation in gene expression.

In the literature of microarray, the Welch t-test is a long-time standard for differential analysis. Let $X_{ig}^{(c)}$ be the expression level of the ith sample for gene g under condition c and $(\sigma_g^c)^2$ be the variance of gene g expression levels. Given the general model 2.1 defined in section 2.4.2, the estimator of the expected level of expression, called $\hat{\mu}_g^{(c)}$ is given by:

$$\hat{\mu}_g^{(c)} = \bar{x}_{.g}^{(c)} = \frac{\sum_{i=1}^{n_c} x_{ig}^{(c)}}{n_c},$$

where n_c is the sample size of condition c. For any given gene, the Welch statistic, denoted t_g^{Welch}, corresponds to a normalized difference between the means of expression levels in both conditions:

$$t_g^{\text{Welch}} = \frac{\bar{x}_{.g}^{(1)} - \bar{x}_{.g}^{(2)}}{\sqrt{\frac{(S_g^{(1)})^2}{n_1} + \frac{(S_g^{(2)})^2}{n_2}}},$$

where n_1 and n_2 are number of replicates of conditions 1 and 2 and $\bar{x}_{.g}^{(1)}$ is the average expression level for gene g under condition 1 (across all possible replicates). The Welch t-statistic does not assume equal variances across conditions (heteroscedastic hypothesis). The variances $(S_g^{(1)})^2$ and $(S_g^{(2)})^2$ are then estimated independently in both conditions for each gene such as:

$$(\hat{\sigma}_g^{(c)})^2 = (S_g^{(c)})^2 = \frac{\sum_{i=1}^{n_c} (x_{ig}^{(c)} - (\bar{x}_{.g}^{(c)})^2}{(n_c - 1)}. \tag{3.1}$$

In contrast homoscedastic tests make the assumption of equality of variances across conditions. The estimated variance $\hat{\sigma}_g^2$, is thus given by:

$$\hat{\sigma}_g^2 = S_g^2 = \frac{(n_1 - 1)\left(S_g^{(1)}\right)^2 + (n_2 - 1)\left(S_g^{(2)}\right)^2}{n_1 + n_2 - 2}, \tag{3.2}$$

where S_g^2 denotes the unbiased pooled estimate of the variance.

A major drawback of the Welch t-test is that the variance estimates can be skewed by genes having (artifactually) a very low variance. These genes are associated to a large t-statistic and falsely selected as differentially expressed, as shown in Tusher et al. (2001). As highlighted by Murie et al. (2009), another drawback comes from its application on small sample sizes which implies low statistical power. Consequently, the efficacy of the t-test have been seriously called into question. It has led to the development of many innovative alternatives dedicated to the analysis of microarray data, with hope of improved variance estimation accuracy and power.

In a parametric context, these alternatives fall into a few nested categories of variance modeling. The first strategy makes the assumption that homogeneity of the variability falls across all genes. Thus, a pooled estimator across genes, defined as the mean of gene variances, is attributed to all the genes. This approach enables the computation of a robust estimator of the variance over a large set of data. However, the biological assumption underlying this statistical model is not realistic, leading to a loss of power as the number of false-negative results increases. The second solution is to define a gene-specific variance such as proposed in the standard Welch *t*-test and in the `ANOVA`, which has been first applied to microarray data by Kerr et al. (2000). Among both of these approaches, the `ANOVA` assumes that the variance of the error term is constant across condition (homoscedasticity), whereas the Welch statistic enables us to test for equality of means under heteroscedasticity. However, the number of replicates has a considerable influence on the estimation of the variance and a generally limited number of replications does not allow an accurate estimation of gene-specific variances leading to spurious small values of the variance due to errors of estimation. Consequently, intermediate modelings, most of which consists of extended *t*-test approaches, have been proposed in the microarray literature to tackle variance estimation issues. We detail in the following paragraphs various shrinkage strategies as well as two alternatives methods, namely `VarMixt` and the test of `Wilcoxon`, based respectively on a mixture model on the variances and a non-parametric ranking approach. We compare the performance of the Welch *t*-test and all the methods detailed in the following, in the section 3.1.

a - Shrinkage approaches

A key strategy to improve the variance estimation is to borrow information across genes. This is the idea underlying shrinkage approaches, which consist in estimating each individual gene value by taking into account information from all genes of the experiment. In general, the shrinkage estimator is written as a function of a gene-by-gene estimator and a common estimator of the whole population. Shrinkage methods aim to prevent the statistic from becoming too large when the variance is close to zero (which can lead to false positives).

RVM Wright et Simon (2003) introduce a Random Variance Model (RVM) and assume that the gene-specific variances are random variables with an inverse Gamma distribution. The statistic is described as follows:

$$t_g^{\text{RVM}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{S_g^{\text{RVM}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

where the variance is given by:

$$(S_g^{\text{RVM}})^2 = \frac{(n_1+n_2-2)S_g^2 + 2a(ab)^{-1}}{(n_1+n_2-2)+2a}.$$

It consists in a weighted average of (i) the pooled variance S_g^2 (with (n_1+n_2-2) degrees of freedom) and (ii) the mean of the fitted inverse gamma distribution $(ab)^{-1}$ (with $2a$ degrees of freedom).

The probability of getting the observed t_g^{RVM} value under the null hypothesis is calculated using the Student distribution.

`Limma` Like the `RVM` model, the `limma` statistic from Smyth (2004) assumes a random variance model where the variances are drawn from an inverse chi-square distribution. A Bayesian estimator of the variance $(S_g^{\text{limma}})^2$ has been substituted for the usual empirical variance S_g^2, into the classical t-statistic:

$$t_g^{\text{limma}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{S_g^{\text{limma}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

The posterior variance is a combination of an estimate obtained from a prior scale-inverse-chi-square distribution and the estimator S_g^2:

$$(S_g^{\text{limma}})^2 = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g},$$

where d_0 and d_g respectively are the residual degrees of freedom for the prior estimator S_0^2 and for the linear model for gene g.
An empirical Bayes approach is adopted in `limma`, estimating the hyperparameter S_0^2 from the data. We refer to Smyth (2004) for details of this estimation. It appears that the estimated value of S_0^2 is usually a little less than the mean of the S_g^2's. Both the `RVM` and `limma` statistics shrink the observed variances towards a prior variance estimate.

`SMVar` To provide a shrunk estimate of the variance, Jaffrezic et al. (2007) employe a structural mixed models defined as follows:

$$\ln\left((S_g^{*(c)})^2\right) = m_c + \delta_g^{(c)} + \epsilon_g^{(c)},$$

where m_c is a condition effect (assumed fixed) and $\delta_g^{(c)}$ is the gene effect under condition c, assumed independent and normally distributed. $\epsilon_g^{(c)}$ is a sampling error due to the estimation of the true variances by the empirical variances. `SMVar` is an heteroscedastic test, whose test statistic is given by:

$$t_g^{\text{SMVar}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{\sqrt{\frac{(S_g^{*(1)})^2}{n_1} + \frac{(S_g^{*(2)})^2}{n_2}}},$$

where $(S_g^{*(1)})^2$ and $(S_g^{*(2)})^2$ are estimations of the variance under the structural model. Such a model usually requires the use of stochastic estimation procedures based on MCMC methods such as Gibbs sampling that are quite time-consuming. The authors therefore propose an approximate method to obtain estimates of the parameters, based on the empirical variances.
The probability of obtaining the observed t_g^{SMVar} value under the null hypothesis is calculated using the Student distribution, which leads to a shrinkage of gene variance towards the condition effect.

SAM Tusher et al. (2001) propose SAM, a non-parametric solution which consists in adding a stabilizing constant, called the *fudge* factor, denoted s_0, to the pooled estimate of the standard deviation:

$$t_g^{\text{SAM}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{s_0 + S_g\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where S_g is the pooled estimate of the standard deviation, as defined previously. Thus, SAM shrinks the gene-specific standard deviation toward s_0. The value of s_0 is some constant performed only once for the entire data. The authors have initially proposed to determine this factor by a complex procedure based on minimizing the coefficient of variation of t_g^{SAM}. Thereafter, simplified versions have been proposed. For instance, s_0 has been computed as the 90^{th} percentile of the standard errors of all genes.
The probability of obtaining the observed t_g^{SAM} value under the null hypothesis is calculated using a permutation procedure.

b - Other approaches

VarMixt An other strategy introduced by Delmar et al. (2005) relies on a mixture model on the variances. They make the assumption that the set of all genes can be divided into classes based on similar responses to the various sources of variability, with all the genes in a particular class having equal variance. The idea is thus to estimate a variance for each gene class from a large number of observations in place of the individual gene variance. Given that each gene is partially assigned to variance classes, the denominator of the statistic is a weighted sum of the variance of all the classes it belongs to:

$$(S_g^{\text{VM}})^2 = \sum_{k=1}^{K} \pi_{g,k} S_{G_k}^2,$$

where $\{S_{G_1}^2, S_{G_2}^2, ..., S_{G_K}^2\}$ denote the K variance classes of the model. The weight $\pi_{g,k}$ is the posterior probability that the true variance of gene g is $S_{G_k}^2$.
For a given gene, using the group variance in place of individual gene variance, the statistic is given by the following expression:

$$t_g^{\text{VM}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{S_g^{\text{VM}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Delmar et al. (2005) use an EM approach to determine the number of groups and their associated variances $S_{G_k}^2$.
The probability of getting the observed t_g^{VM} value under the null hypothesis is calculated using the standard Gaussian distribution.

Wilcoxon This test involves the calculation of a non-parametric statistic, W_g, based on rank summation. The gene expression values are first ranked

Test	P / NP	Homosc.		Variance estimation
		Yes	No	
Welch	P		X	gene-specific
ANOVA	P	X		gene-specific
SAM	NP			shrinkage (fudge factor)
RVM	P	X		shrinkage (prior distribution)
limma	P	X		shrinkage (prior distribution)
SMVar	P		X	shrinkage (mixed model)
VarMixt	P	X		mixture model
Wilcoxon	NP			rank sum

Table 3.1 – ***Variance modeling strategies of tests for differential analysis*** - *This table summarize the eight tests of the comparison study as well as their variance modeling strategies: Parametric (P) vs. Non-Parametric (NP), Homoscedastic vs. Heteroscedastic and finally we detail the approach used for variance estimation.*

for each gene across all conditions (an average rank is assigned for ties). The rank sums are then used to calculate the W_g statistic:

$$W_g = R_g^{(1)} - \frac{(n_1(n_1+1))}{2},$$

where n_1 is the sample size for condition 1 and $R_g^{(1)}$ is the rank sums in the same sample.
The idea is that ranks should be randomly arranged between the two conditions if observations are drawn from the same underlying distribution. W_g is then compared to a table of all possible distributions of ranks to calculate the p-value.

3.1.2 Comparison process

In this section we focus on the choice of the statistic used to score the difference of expression. As detailed in the third step of the procedure of hypothesis testing, the most widely used statistic is certainly the Welch t-statistic. However, in microarray studies where there are only small sample sizes, it can lead to poor results and inaccurate statistical tests of differential expression. This has led to the development of a wide range of alternative approaches. However, a critical issue is that selecting a different test usually leads to a different gene list. In this context, identifying the most efficient approach in practice remains crucial. In an attempt to address this, we conducted a comparison study of the eight tests described in the previous section. The comparison process relies on four steps: gene list analysis, simulations, spike-in data and re-sampling. Our aim is to benefit from the specificity of each evaluation strategy, to make our results comparable to previous studies and to ease the formulation of general, robust and reproducible conclusions. At each step of the process, tests are compared in terms of statistical power assessed at the same false-positive rate. Control of the false-positive rate to the desired value is checked for each test which is, to our opinion, too rarely considered in the literature. Eventually, in addition to an efficacy comparison, we found it relevant to confront each test in terms of practical consideration such as execution time and ease of use.

Gene list analysis

An intuitive first step to compare the tests is to investigate the consistency between gene lists resulting from the application of each test on real data. Here we apply this approach to five publicly available datasets (Table 3.2) to assess the overlap between gene lists and to identify similar behaviors among the variance modeling strategies.

In addition to the eight tests, we define a "control" test that draws from a Uniform distribution between 0 and 1 for each gene a p-value. We then applied the tests to the five data-sets to identify genes differentially expressed by setting a p-value threshold of 0.05.
Gene list similarities between tests are assessed and displayed using both a Hierarchical Clustering (HC) and a Principal Component Analysis (PCA). The HC was built using a binary metric and the Ward's aggregation algorithm, both available in the R package `stats`. We performed the PCA from the R function `dudi.pca` of the package `ade4` developed by Chessel et al. (2004).

Simulation study

The purpose of this study is to estimate power and false-positive rate on a large range of simulated datasets, in order to compare the tests under simple and sometimes extreme situations. We define a reference model (denoted M_1), frequently adopted in the literature and that matches the assumptions of the t-test. Under M_1, gene expressions for the conditions 1 and 2 are drawn from Gaussian distributions of same variance ($\sigma^2 = 1$):

$$\begin{cases} X_{ig}^{(1)} & \sim \quad \mathcal{N}(\mu_g^{(1)}, \sigma^2), \\ X_{ig}^{(2)} & \sim \quad \mathcal{N}(\mu_g^{(2)}, \sigma^2). \end{cases}$$

Under H_0: $\mu_g^{(1)} = \mu_g^{(2)}$, while under H_1: $\mu_g^{(2)} = \mu_g^{(1)} + \delta$, with $\delta = 0.5$.

We then applied three extensions of M_1 (denoted M_2, M_3 and M_4) designed to be less to the t-test advantage. M_2 is quite similar but expression levels are now drawn from a Uniform distribution of same parameters. M_3 applies a mixture model on variances and corresponds to the `VarMixt` hypothesis; genes are then divided into three classes of variance. Under M_4, 10% of the genes are simulated with small variances ($\sigma^2 = 0.05$) since they can lead to an increase of false-positive rate when the t-test is applied.

For each model we simulated $10,000$ independent genes under H_0 to assess the false-positive rate attached to each test , and $10,000$ under H_1 to compute their respective power. False-positive rate and power are both assessed at a p-value threshold of 0.05. Sample size ranges from 5 to 100 samples per condition.

Spike-in dataset

The Human Genome U133 dataset is used to test and validate microarray analysis methods (`http://www.affymetrix.com`). The dataset con-

sists of 14 hybridizations of 42 spiked transcripts in a complex human background at concentrations ranging from 0.125 pM to 512 pM. Each condition includes three replicates. We perform the 13 pairwise comparisons for which spike-in genes have a true fold-change of two.
The whole dataset contains 22, 300 genes. The 42 spike-in genes are designed to be differentially expressed (under H_1) and used for power estimation. To be able to calculate the false-positive rate, the 22, 258 remaining genes are forced to be under H_0 by permutation of the condition labels. False-positive rate and power are both assessed at a p-value threshold of 0.05.

Re-sampling approach

This approach is inspired from Wright et Simon (2003). Its main idea is to assess the ability of a test to select genes determined as differentially expressed from the full dataset, in small subsamples ($n = 5$ and $n = 10$). The strategy can be summarized in four steps:
Step 1: from the 500 samples dataset described in Guedj et al. (2011), we define as differentially expressed the genes whose p-value $\leqslant 10^{-4}$, with the Welch t-test. This set of genes is considered in Step 3 as the "truth" to estimate power.
Step 2: n samples are drawn from each condition and the eight tests are performed on this subset of the initial data. We apply the Benjamini and Hochberg correction at a 0.1 FDR level.
Step 3: from Step 2 we estimate power as the proportion of genes defined as differentially expressed at Step 1 and detected at Step 2.
Step 4: Steps 2 and 3 are iterated 1, 000 times. Finally power is averaged over the 1, 000 iterations.

Data-set	**Conditions**	**Sample size**	**Publication**
Lymphoid tumors	Disease staging	37	Lamant et al. (2007)
Liver tumors	TP53 mutation	65	Boyault et al. (2007)
Head and neck tumors	Gender	81	Rickman et al. (2008)
Leukemia	Gender	104	Soulier et al. (2005)
Breast tumors	ESR1 expression	500	Guedj et al. (2011)

Table 3.2 – ***Datasets used for the gene list analysis.***
The five datasets come from the Cartes d'Identité des Tumeurs (CIT, `http://cit.ligue-cancer.net`) program and are publicly available. All the microarrays are Affymetrix U133A microarrays with 22, 283 genes.

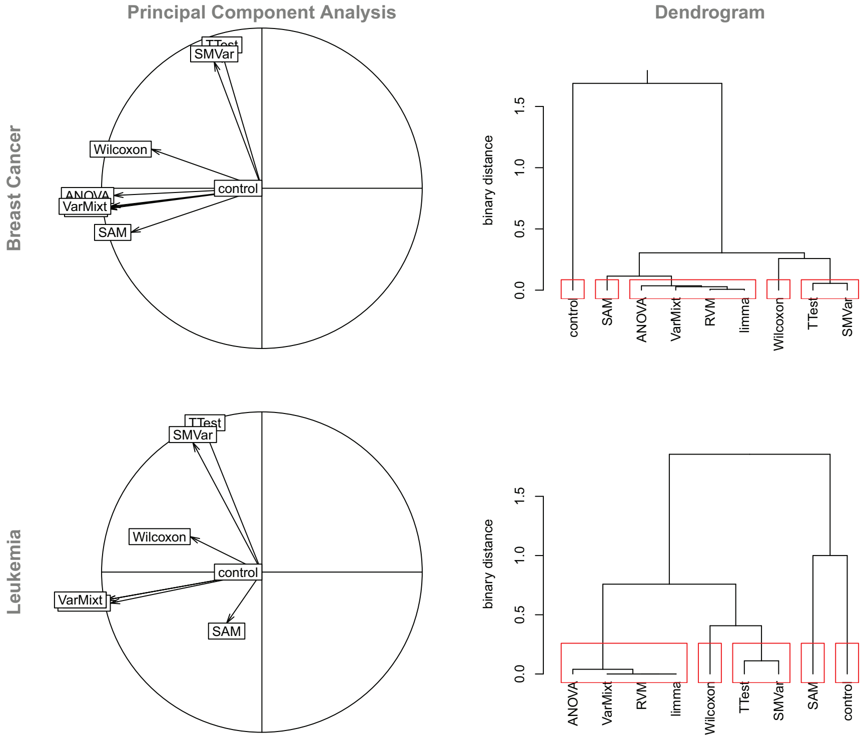


Figure 3.2 – ***Gene list analysis.***
PCAs and dendrograms are generated based on the gene lists resulting from the application of the eight tests of interest and the control-test. Here we show results for the breast cancer dataset of Guedj et al. (2011) and the leukemia dataset of Soulier et al. (2005). Both outline five clusters of tests.

3.1.3 Results

Gene list analysis

The figure 3.2 represents PCAs and dendrograms resulting from gene list analysis. The cumulative inertia explained by the two first axes of PCA is approximately 80%. Both representations underline the same tendencies.

As expected, gene lists resulting from the control-test are clearly independent from the others, since it selects genes (differentially expressed or not) uniformly. The eight tests show various behaviors. Six tests clusterize in two distinct groups: {Welch *t*-test; `SMVar` } and {`VarMixt`; `limma`; `RVM` ; `ANOVA`}. The proportion of common genes selected by two tests of the same cluster is approximately 90%. On the other hand, `Wilcoxon` and `SAM` do not clearly fall in one of the two main groups: `Wilcoxon` tends to consistently lie between them, whereas `SAM` does not exhibit a reproducible behavior.

To summarize, homoscedastic (`VarMixt`, `limma`, `RVM` and `ANOVA`) and heteroscedastic (Welch *t*-test and `SMVar`) variance modeling strategies are well discriminated by an analysis of similarities between gene lists. It outlines the interesting property that similar modeling strategies in theory imply similar results in practice.

Simulation study

First, we evaluate power according to sample size under the simulation model M_1 (Figure 3.3). On Figure 3.3-A, we notice little difference between the tests (less than 0.08), particularly for large samples as expected. `Wilcoxon` is not as good as the other tests in most cases. `SAM` and `ANOVA` show equivalent performance to the *t*-test. `VarMixt`, `RVM` and `limma` tend to provide an increase in power, and `SMVar` slightly outperforms all the tests (Figures 3.3-A and 3.3-B).

As we know, these preliminary results are valid only if all the tests meet the theoretical 5% false-positive rate when applying a *p*-value threshold of 0.05. Table 3.3 gives the observed false-positive rate for each test under small and large sample sizes and sheds light on the fact that some tests clearly deviate from the 5% level and return biased *p*-values. Observed deviations are more accentuated for small sample sizes compared to large ones. `SMVar` and `RVM` inflate the expected number of false-positives whereas `Wilcoxon` and the Welch *t*-test tend to be conservative; `ANOVA`, `SAM`, `limma` and `VarMixt` show no deviation.

Regarding these observations, the tests which are inefficient in controlling the false-positive rate at the expected 5% level have to be adjusted by a time consuming Monte-Carlo procedure. Figures 3.3-C and 3.3-D present power results as adjusted and hence valid false-positive rates. Differences are clearly reduced compared to Figures 3.3-A and 3.3-B which confirms that part of the difference in power observed is due to an actual difference in false-positive rate, particularly concerning `SMVar` . After adjustment `VarMixt`, `RVM` and `limma` tend to be the best tests although they provide an insignificant gain compared to the *t*-test; `Wilcoxon` remains the less powerful. `ANOVA` has performance comparable to the Welch *t*-test which is interesting: under the same variance between the two conditions, tests that make the corresponding homoscedastic assumption (`ANOVA`) do not show improved power compared to those which are heteroscedastic (Welch *t*-test).

Surprisingly, model M_2 leads to the same conclusions (data not shown). Here, expression values follow a Uniform distribution instead of a Gaussian one, which does not suit the hypothesis of parametric approaches, which assume that the expression levels follow a standard normal distribution under H_0. Compared to model M_1, we were expecting to observe a notable increase in power for `Wilcoxon` , which was not observed. This result confirms that *t*-test and assimilated approaches are quite robust to the Gaussian assumption. Indeed the Central Limit Theorem implies that even if expression values are not Gaussian, the *t*-statistic resulting from the comparison of two conditions is likely to be. It should be noted that the structural model of `SMVar` is not able to provide results for the uniform model.

Finally models M_3 and M_4 also lead to the same conclusion, with an overall loss of power (data not shown).

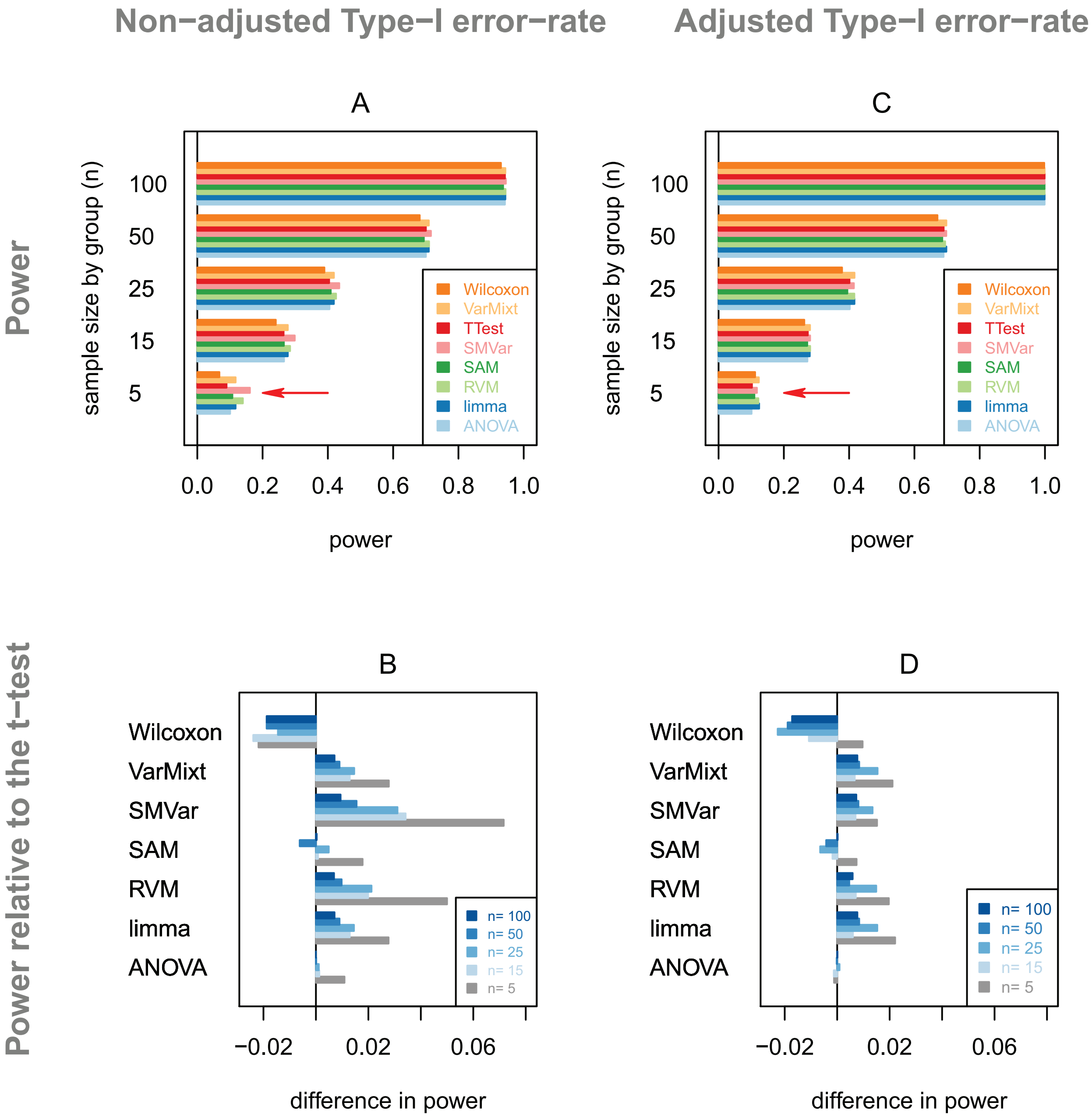


Figure 3.3 – ***Power study from simulations (Gaussian model, M1).***
Power values are calculated at the 5% *level and displayed according to the sample size. Figures A and C represent power values. Red arrows highlight the effect of false-positive rate adjustment on power values. Figures B and D represent power values relative to t-test. Figures A and B concern power values calculated at the actual false-positive rate. Figures C and D concern power values calculated at the adjusted false-positive rate.*

	M1		M2		M3		M4	
sample size	**n = 5**	**n = 100**	**n = 5**	**n = 100**	**n = 5**	**n = 100**	**n = 5**	**n = 100**
t-test▼	3.8 – 4.6	4.5 – 5.4	4.0 – 4.8	4.6 – 5.5	3.8 – 4.6	4.7 – 5.6	3.9 – 4.7	4.4 – 5.3
ANOVA	4.5 – 5.2	4.5 – 5.4	4.7 – 5.6	4.6 – 5.5	4.5 – 5.4	4.7 – 5.6	4.5 – 5.3	4.4 – 5.3
Wilcoxon▼	2.8 – 3.5	4.6 – 5.5	2.6 – 3.3	4.5 – 5.4	2.8 – 3.5	4.7 – 5.6	2.7 – 3.4	4.5 – 5.4
SAM	4.6 – 5.5	4.5 – 5.3	4.2 – 5.1	4.5 – 5.4	4.7 – 5.6	4.7 – 5.6	4.3 – 5.2	4.4 – 5.3
RVM▲	5.7 – 6.7	4.5 – 5.4	5.6 – 6.5	4.5 – 5.4	5.4 – 6.3	4.7 – 5.6	5.3 – 6.2	4.7 – 5.5
limma	4.6 – 5.5	4.6 – 5.5	4.2 – 5.1	4.5 – 5.4	4.7 – 5.6	4.7 – 5.6	4.4 – 5.3	4.3 – 5.1
SMVar▲	7.0 – 8.1	4.7 – 5.6	–	–	5.9 – 6.8	4.8 – 5.7	4.6 – 5.5	4.5 – 5.3
VarMixt	4.7 – 5.5	4.6 – 5.5	4.3 – 5.2	4.6 – 5.5	4.8 – 5.6	4.6 – 5.5	4.5 – 5.4	4.5 – 5.3

Table 3.3 – ***False-positive rate study from simulations.***
For small and large samples, this table presents the 95% *confidence-interval of false-positive rate obtained by applying a threshold of* 0.05 *to the p-values. Up triangles ▲ (resp. down triangles ▼) indicate an increase (resp. a decrease) of the false-positive rate compared to the expected level of* 5%. *Two triangles denote a deviation in both small and large sample sizes*

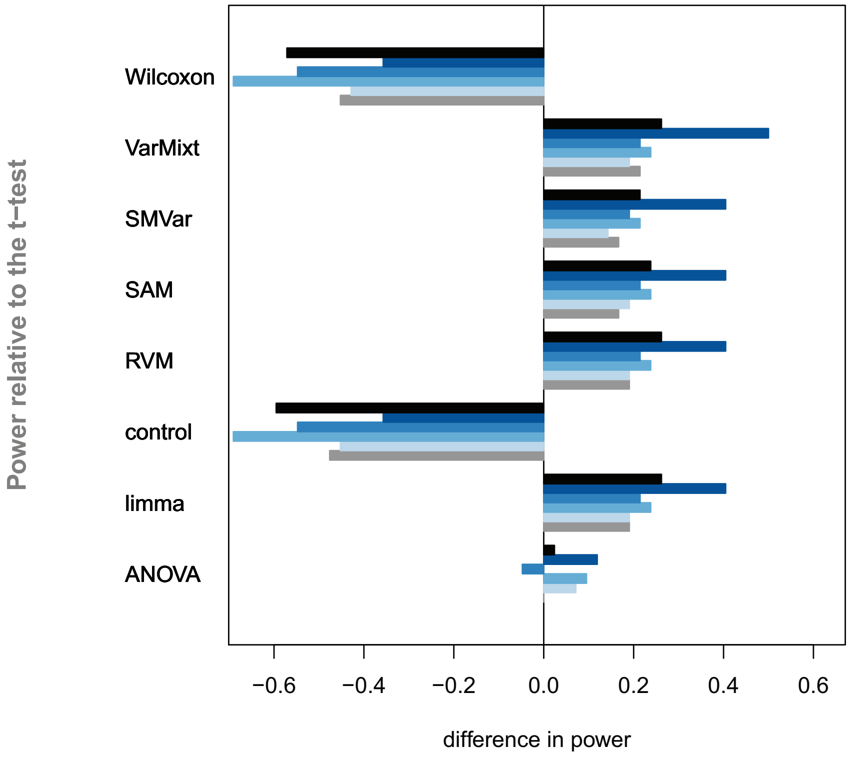


Figure 3.4 – ***Spike-in dataset.***
Power values are calculated at the 5% *level and displayed according to* 6 *of the* 13 *pairwise comparisons.*

Spike-in dataset

Spike-in data confirm observations made on the simulations. `SMVar` and `RVM` inflate the expected number of false-positives whereas `Wilcoxon` and the *t*-test tend to be conservative. Power values adjusted to a valid false-positive rate present more significant differences than in simulations (Figure 3.4): with an average decrease of almost 0.6, `Wilcoxon` is the least powerful and similar to the control test; `ANOVA` shows equivalent performance than the Welch *t*-test; `VarMixt`, `RVM` , `SMVar` and `limma` provide a significant increase in power with an average gain of 0.25. With performance comparable to the best tests, `SAM` has a different behavior than in simulations.

Re-sampling approach

This approach corroborates tendencies obtained with simulations and spike-in data (Figure 3.5): `limma`, `VarMixt` and `RVM` perform much better than other tests in identifying differentially expressed genes, while `SMVar` is somewhat less efficient than the three top-tests. `ANOVA` and the *t*-test still show equivalent performance, although `ANOVA` presents here a

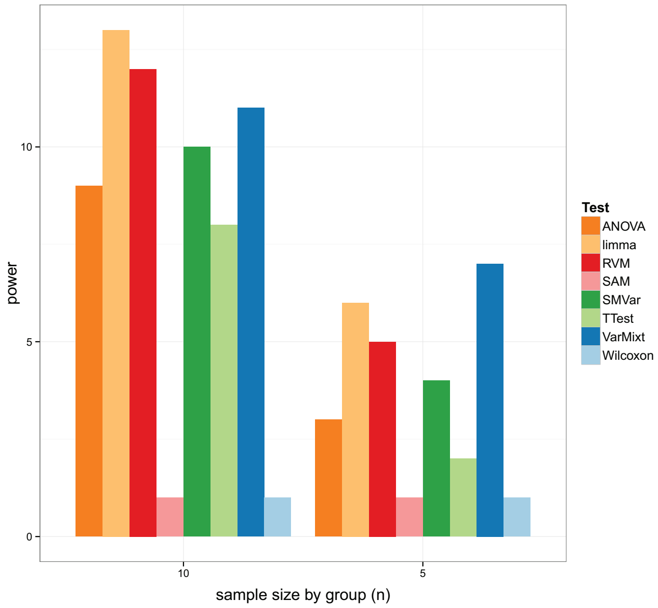


Figure 3.5 – ***Re-sampling approach.***
Power values are calculated at a 0.1 *FDR level and displayed according to the sample size.*

slight but significant improvement.

Wilcoxon and SAM were never able to detect genes determined as differentially expressed. Indeed the calibration performed can not reach *p*-value lower than 10^{-3} for small sample sizes. After the Benjamini-Hochberg correction at a 0.1 FDR level (corresponding here to a 10^{-6} *p*-value threshold), they do not detect any gene as differentially expressed.

Practical comparison

Concerning time of execution and ease of use, the Welch *t*-test and ANOVA are the most efficient as they rely on standard statistical considerations and have benefited from improved implementations. On real high-throughput data, both take few seconds to treat tens of thousands of genes. In terms of execution time, limma appears as efficient as the *t*-test and ANOVA, which is a noteworthy point. SMVar , RVM and SAM for a longer, yet still reasonable time (up to 8 minutes in our case). VarMixt turns out to be the slowest approach (up to 80 minutes) as it relies on a time consuming EM algorithm.

3.1.4 Conclusion on hypothesis testing strategies

Given the current tendency to apply the *t*-test to gene expression data and the wealth of available alternatives, finding the most appropriate approach to handle differential analysis is critical.

In order to provide some solutions to this problem, we developed a comparison process of eight tests for differential expression. It is based on gene list analysis, simulations, spike-in data and re-sampling, with the intention of benefitting from the specificity and advantages of each strategy.

Gene list analysis does not properly compare test performance and hence lead to limited conclusions. However it is an appropriate preliminary approach that focuses on similarities between test results. An analysis of the consistency between gene lists outlines general tendencies that can help in interpreting differential analysis results. In our case, we observed comparable results between tests based on similar variance modeling strategies.

The three other approaches (simulations, spike-in data and re-sampling) propose a direct comparison of power values. Simulations represent a convenient statistical framework as genes under H_0 and H_1 are known in advance. Additionally, different hypotheses on data structure can be specified under different simulation models. Here, the three further models (M_2, M_3 and M_4) actually lead to the same conclusions as the reference Gaussian one (M_1). If simulations do not allow to observe significant differences in power between the tests, they still reveal reproducible tendencies. In addition, simulations turn out to be the gold standard against which to check for possible deviations from the expected false-positive rate. However, it is unclear whether simulated datasets can sufficiently and realistically reflect the noise inherent in real microarray data as highlighted by Wu (2005).

More empirical alternatives include the use of spike-in data and re-sampling. Spike-in genes can represent gene expression better than simulations. In our case it confirms conclusions from simulations with more significant differences in power. Regarding the Affymetrix dataset we used, a criticism of this approach could be that the small number of actual spike-in genes does not allow a very accurate power estimation. Moreover, variation across technical replicates is likely to be lower than that typically observed across true biological replicates, and many biological effects of interest may be smaller than two-fold.

In this context, a re-sampling approach takes advantage of the complexity found in real data. Differentially expressed genes are not known but determined from a large dataset (500 samples in our case); power is then evaluated on a subset of the data. Results are comparable to those obtained with simulations and spike-in data. However this approach can be considered as limited in that it assumes that gene lists generated on the full dataset are correct; besides it is fastidious to implement and extremely time consuming.

By applying four distinct comparison strategies with specific advantages and drawbacks: **(i)** we ensure to offset the limitations of each strategy and **(ii)** we provide robust conclusions on the performances of each test.

We applied this comparison process to eight tests representative of different variance modeling strategies. Results are summarized in Table 3.4. A first important result concerns the control of the false-positive rate, which is often disregarded in the literature. Under H_0, distribution of p-values is supposed to be uniform and the false-positive rate resulting from a p-value threshold of 0.05 should be controlled at 5%. Deviation from this major assumption may indicate biased p-values. In both simulations and spike-in data, some tests deviate from the expected false-positive rate, which partly explains some differences in power (namely `SMVar` , `RVM` and `Wilcoxon`). For the purpose of our study, we performed a Monte-Carlo based adjustment of the false-positive rate to formulate comparable conclusions across all the tests. However in practice this adjustment remains fastidious to implement. In consequence, we strongly advocate to avoid using these tests until a proper corrected version is made available.

Overall, `Wilcoxon` and `SAM` show weak performance. One of our simulation models (M_2) clearly outlines the robustness of parametric tests to the Gaussian assumption. Concerning `SAM`, it not possible to formulate clear conclusions from our results, and indeed they serve to highlight existing doubts about its efficacy, as shown Zhang (2007).

Compared to the t-test, `limma` and `VarMixt` consistently show real improvement, particularly on small sample sizes. `Limma` has often been discussed in the biostatistical field and its positive performance has been reported in Kooperberg et al. (2005), Jeffery et al. (2006) and Murie et al. (2009). Surprisingly `VarMixt` does not appear as weak as similar methods evaluated by Kooperberg et al. (2005). Presumably it benefits from a more realistic mixture model on variances which are less likely to generate false-positives.

If `limma` and `VarMixt` are equivalent in relation to both power and false-positive rate, then in practice `limma` presents several further advantages in terms of execution time. In addition, `limma` can be generalized to more than two conditions which makes it relevant to many broader situations.

In conclusion, we have developed a comprehensive process to compare statistical tests dedicated to differential analysis. This approach can be used as the basis on which to evaluate performance of methods developed in the near future. Furthermore, in response to our question "Should we abandon the t-test", `limma` provides a substantial improvement compared to the t-test, particularly for small sample sizes. However the t-test remains easy to apply through a wide-range of tools for genomic analysis whereas `limma` can initially appear more difficult to implement.

	False-positive rate		**Power**		**In practice**	
	Small samples	Large samples	Small samples	Large samples	Ease of use	Execution time
***t*-test**	+	+ + +	+	+ + +	+ + +	+ + +
ANOVA	+ + +	+ + +	+	+ + +	+ + +	+ + +
Wilcoxon	+	+	+	++	+ + +	++
SAM	+ + +	+ + +	+	++	++	++
RVM	+	++	+ + +	+ + +	++	+
limma	+ + +	+ + +	+ + +	+ + +	++	+ + +
VarMixt	+ + +	+ + +	+ + +	+ + +	+	+
SMVar	+	+	++	+ + +	++	+ + +

Table 3.4 – ***Summary table.***
This table summarizes the results of our study in terms of false-positive rate, power and practical criteria. The number of "+" indicates the performance, from poor (+), to excellent (+++).

3.2 A local-score approach for the identification of differentially expressed modules in molecular networks

Classical tools for differential expression analysis of microarray experiments suffer from a lack of reproducibility across studies. In practice, signatures selected in comparable experiments share only a few genes in common as shown by Ein-Dor et al. (2005) or Solé et al. (2009), which is a major drawback in making gene signatures a standard tool for clinical studies. A possible explanation is the use of different tissue acquisition methods, non-standardized platforms or the variation in patient sampling. Besides these sources of variability which are identified and can be controlled, the instability of molecular signature is also due to (i) the high-dimension context and (iii) the complexity of gene expression regulation mechanisms. In the high-dimensional setting, the expression levels of the genes are often collected on relatively few samples which makes the use of classical tools inappropriate for genes selection. In addition, gene expression is the result of the coordinated interactions of multiple proteins and is influenced by internal factors such as hormones or metabolisms as well as environmental factors including lifestyle or nutrition, which are not clearly identified and renders the problem of gene selection even harder. For the purpose of addressing this, there has been growing interest in developing approaches that try to improve signature reproducibility by integrating additional information on the relationships between genes, over the past few years. Several attempts in this direction tried to integrate knowledge on Protein-Protein Interactions (PPIs) as well as pathways or functional annotations with the hope of making signatures more stable and more interpretable. One of the first approaches was described in Guo et al. (2005), which mapped genes to GO functional categories. Instead of considering individual gene expression levels, they compute a summary measure for each GO category significantly enriched with differentially expressed genes. Thus, they provide a way of reducing the dimensionality of the data by treating the gene expressions within a functional category for further analysis. Using this method as a starting point, various improvements have been proposed for dealing with both pathways or functional annotations and gene expression data, see for instance Yu et al. (2007) or Kammers et al. (2011). Another kind of method, based on the identification of differentially expressed subnetworks, was first introduced by Chuang et al. (2007). In the latter paper, the authors employed a sliding window model. This type of model uses the mutual information to measure the association between a subnetwork expression profile and a given phenotype, selecting significantly differentially expressed subnetworks by comparing their discriminative potentials to those of random networks. In a machine learning framework, kernel-based techniques or regularization methods have been applied. For instance, the strategy developed by Haury et al. (2010) is based upon an extension of the Lasso regression, namely the graph Lasso. Instead of using a classical ℓ_1-norm penalty, they proposed a new penalty to incorporate the gene network information, leading to the selection of genes that are often connected to each other in PPI networks or pathways.

In this section, we introduce a new approach in the line with the work of Chuang et al. (2007). Our method is motivated by the observation that genes causing the same phenotype are likely to interact together. We therefore explore an approach for identifying modules, *i.e.* genes that are functionally related, rather than individual genes, by integrating topological features to classical transcriptome analysis. One of the strongest manifestations of functional relations between genes is Protein-Protein Interactions (PPIs). Thus, the general idea of our approach is to map gene expression levels onto the PPI network in order to detect modules of connected genes, or subnetworks, associated with the disease or phenotype of interest. This approach, named `DiAMS` for Disease Associated Modules Selection, involves a local-score strategy and provides a set of candidate modules with a measure of statistical significance.

The section is outlined as follows. First, we introduce the local-score statistic and we propose an extension dedicated to module selection in biological networks. We then detail the module scoring strategy used to compute the local-score. In a second subsection, we present the global approach of `DiAMS`, for the selection of modules significantly enriched in disease associated genes. We specify the input parameters of the method, the algorithm for module ranking and how to assess the significance of modules by Monte-Carlo simulations. Finally, we evaluate the performance of `DiAMS` in terms of power, false-positive rate and reproducibility.

3.2.1 Local-score

Definition

The local-score statistic is a matter of interest in biological sequence analysis. It found many applications in pattern identification to locate transmembrane or hydrophobic segments, DNA-binding domains as well as regions of concentrated charges. The literature on the subject of local-score includes, but is not limited to: Karlin et al. (1991), Brendel et al. (1992) or Karlin et Brendel (1992). More recently, Guedj et al. (2006) proposed to extend the local-score approach to Genetic Epidemiology to capture dependences between markers in association studies.
To define the local-score statistic, let us consider an example where $\mathbb{A} = (A_i)_{1 \leqslant i \leqslant n}$ is a peptide sequence and Θ the alphabet corresponding to this sequence. Let us suppose we are interested in the detection of hydrophobic regions. We denote $f : \Theta \to \mathbb{Z}$, the scoring function that assigns positive or negative scores to amino acids according to the polarity of their side-chains, and $X = (X_i)_{1 \leqslant i \leqslant n}$ the corresponding score sequence such as $X_i = f(A_i)$. The local-score, L, can be expressed as follows:

$$\mathrm{L} = \max_{1 \leqslant i \leqslant j \leqslant n} \sum_{k=i}^{j} X_k.$$

The local-score of the sequence is thus defined as the value of the subsequence with the maximal sum of scores. In the example, the segment that realized the local-score is the most hydrophobic region of the sequence. It is referred to as the "maximal scoring subsequence" or the "local highest scoring segment" or even "locally optimal subsequence" in the literature.

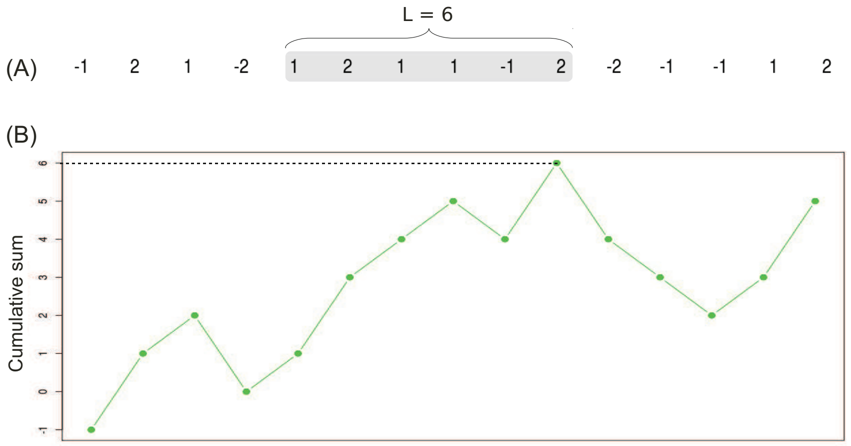


Figure 3.6 – ***Local-score illustration.***
(A) A score sequence: we highlight in grey the locally optimal subsequence and indicate its corresponding local-score, L = 6. (B) Cumulative sum of sequence scores.

Extended version for module discovery

We propose to extend the local-score to the discovery of high-scoring modules of genes in a PPI network. Let us consider here that we have enumerated all the possible modules of the network in a list called $\mathcal{M}$. Obviously, it is not possible in large-scale networks and we dedicate the section 3.2.2 to the development of an alternative approach. In analogy to the peptide score X_i, we denote W_g, the score of a given gene g. The local-score is thus defined as the value of the highest scoring module (*i.e.* the module whose sum of gene score is maximal):

$$\mathrm{L} = \max_{M \in \mathcal{M}} \left(\sum_{g \in M} W_g \right).$$

Note that a module is maximal in the sense that it can not be extended or shortened without reducing the local-score statistic.

This definition of the local score restricts our search to the highest scoring module. However, the next highest scoring modules may be potentially interesting for the study. We therefore rank all modules of the initial network, such that the kth best module is defined as the module with the kth best local-score denoted L_k such as $\mathrm{L}_1 > ... > \mathrm{L}_m$, and identify significant ones. Such an approach will probably yield to the identification of overlapping modules. For instance, the second best module will likely include or be contained in the first highest scoring module. To avoid such situations that provide limited information, we look at disjoint modules.

Thus, once the best module has been identified, each gene included in it is thus removed from the remaining modules.

Module scoring

The local-score statistic relies on gene scores, denoted W_g, that reflects the association of a given gene to the phenotype of interest. We define the scoring function as follows: $W_g = Z_g - \delta$, such as Z_g is the individual score of each gene g and δ a constant specified in the following paragraph. In this work, we derive the individual score Z_g of gene g from its p-value, denoted p_g, resulting from a statistical test: `limma` for microarray data, `TSPM` or `SAMseq` for RNA-seq data as they exhibit the most promising results in our preliminary analysis mentioned in section 6.2.
Given that a high score Z_g should denote a high chance of association with phenotypes of interest, the p-values need a transformation such as $Z_g = -log_{10}(p_g)$, to be used as an individual score for each gene.
A constraint of the strategy is to have expected negative individual scores, *i.e.* $\mathbb{E}\left(W_g\right) \leq 0$, otherwise the module with the highest score would easily span the entire network. Consequently, a constant δ must be subtracted to obtain corrected scores, see Figure 3.7. Genes with a score higher than δ will improve the cumulative score of a given module whereas genes with a score below the threshold will penalize it. We set the value of δ equal to the significance level $\alpha = 0.05$.

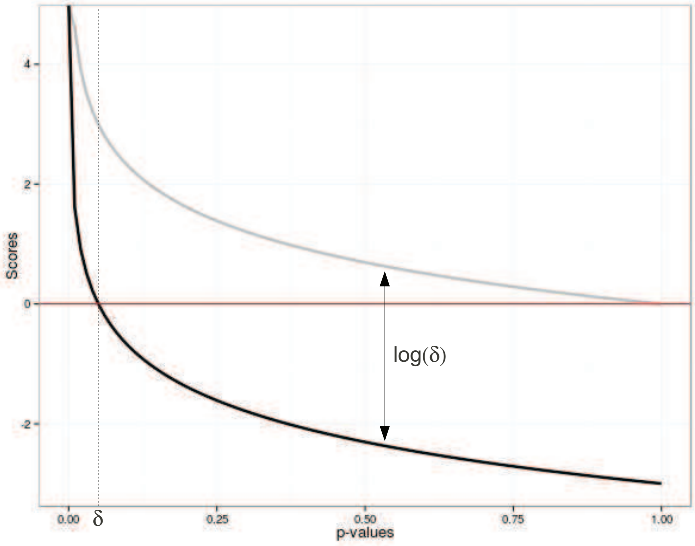


Figure 3.7 – ***Distribution of gene scores in function of p-values.***
In light gray we illustrate the distribution of gene scores Z_g. We observe that they all have positive values. In this case, the highest scoring module would be the entire network. To avoid such situations, a constant δ is subtracted from each gene score. The curve in black displays the distribution of $Z_g - \delta$ values.

3.2.2 Disease associated modules selection

In the present subsection we detail the global strategy, summarized in Figure 3.10, to search for functional modules presenting unexpected accumulations of genes associated to a phenotype of interest in a PPI network.

Input parameters

The first input parameter that is passed to `DiAMS` is a PPI network. The main issue when working with biological networks lies in the impossibility of exploring the huge space of possible gene subnetworks. In Chuang et al. (2007), the authors strategy was to define an initial "seed", *i.e.* starting points for candidate subnetworks and to look at the effect of the addition of a gene in a module within a specified distance $d = 2$ from the seed. In practice, it leads to the identification of modules made up of only a few genes. Moreover, it is required to define a limited set of starting points for the algorithm. Here, we propose an alternative strategy which allows the entire network to be screened without constraints on module sizes by converting the network into a tree structure using a clustering algorithm. This is driven by the observation that biological graphs are globally sparse but locally dense, *i.e.* there exist groups of vertices, called communities, highly connected within them but with few links to other vertices. Therefore, by applying a strategy of clustering which enables to obtain a hierarchical community structure we are able to capture much information about the network topology. The main advantage is that the hierarchical structure renders it relatively easy to go through it instead of exploring all possible subnetworks. Thus the preliminary step of our approach is to convert the network structure into a relevant tree structure. For this purpose, we use the approach of Pons et Latapy (2004), named `walktrap`. The authors employed a random walk strategy through the network for detecting dense modules, introducing a similarity measure based on short walks, which is used to define a distance matrix between any two genes (or nodes) of the network. According to Ward's criterion, they are able to infer a tree structure. A module is no longer defined as a subnetwork but as a subtree of the hierarchical structure (see Figure 3.8). In analogy with the definition of the local-score for network, we define it for a hierarchical community structure, denoted $\mathcal{H}$, as follows:

$$L = \max_{H \subseteq \mathcal{H}} \left(\sum_{g \in H} W_g \right),$$

such as H is included in $\mathcal{H}$ if H is a subtree of $\mathcal{H}$, *i.e.* H can be obtained from $\mathcal{H}$ by deleting nodes in $\mathcal{H}$.

The second parameter that has to be passed to the method is a vector of scores Z_g, that quantifies for each gene its association to the disease. In this study the scoring function is related to the differential expression of the gene such as significant genes, *i.e.* those that are significantly differentially expressed, have a higher score than non-significant genes. However, other kinds of scoring approaches may also be suitable as well

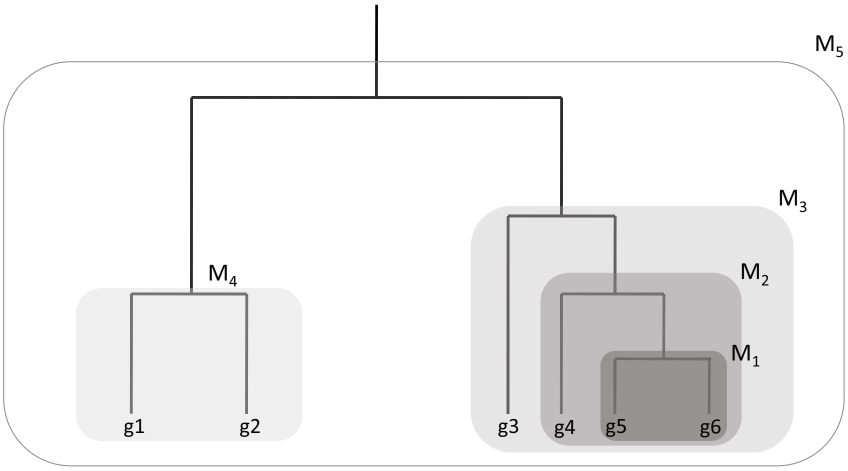


Figure 3.8 – ***Module description.***
A module is defined as a subtree of the hierarchical structure. Leaves, i.e. genes, are also considered as modules. Thus, in this figure we count eleven modules: six modules of size one and five modules of size greater than one. For instance, the module M_3 is composed of four genes. Its score is the sum of each individual gene score, W_{g3}, W_{g4}, W_{g5} and W_{g6}.

as high scores, denoting a strong association to the disease.

Module ranking through a local-score strategy

Once both the tree structure and the score vector have been defined, we search for accumulation of high-scoring genes in the tree. The strategy for the selection of significant modules can be described in the following three-step algorithm:

1. ***Initialization*** - The first step consists of enumerating modules of the tree in a list and assigning them a score, which is defined as the sum of individual scores, denoted W_g, of all the genes that constitute it, see 3.8.

2. ***Module ranking*** - The second step, detailed on Figure 3.9, involves an iterative local-score algorithm: (i) the highest-scoring module is identified (ii) then, it is removed from the list of modules. Steps (i) and (ii) are then repeatedly applied until all disjoint modules have been enumerated. Thus, we obtain a ranked list of m modules and their respective local-scores $L_1, ..., L_m$ such as $L_1 > ... > L_m$ with the ith best module being disjoint from the preceding $i\text{th} - 1$ best modules.

3. ***Module significance assessment*** - Given $L_1, ..., L_m$, the last step proposes a way to select a set of modules significantly enriched in disease associated genes. The global significance of each module is

assessed via Monte-Carlo simulations, discussed in the next paragraph. Through this permutation procedure we obtain a p-value for each module and are able to make a conclusion about its significance of at a given level.

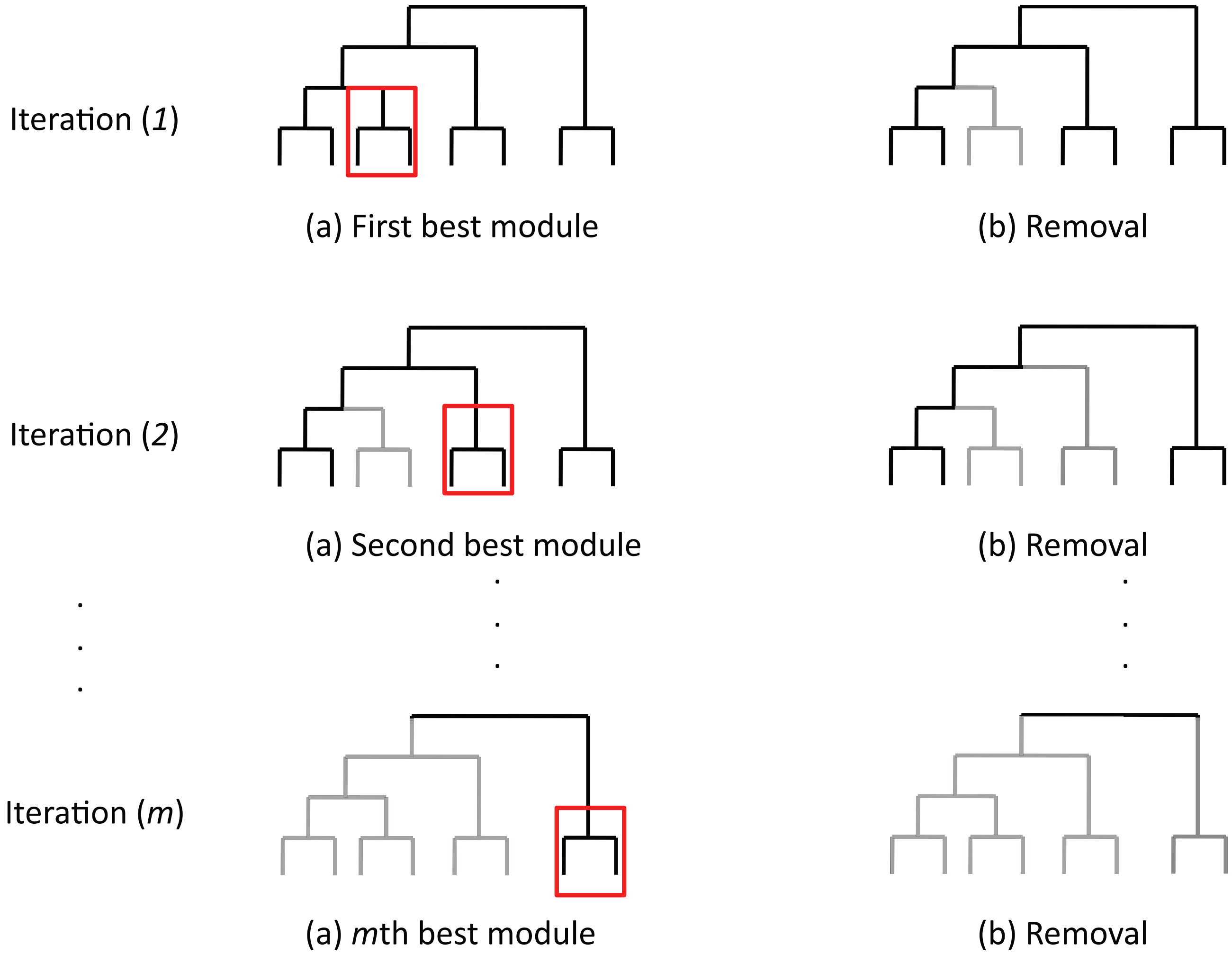


Figure 3.9 – ***Iterative module ranking.***
Once the highest-scoring module identified, each gene included in it has to be removed from the remaining modules. The process is repeated iteratively m times, until all disjoint modules have been enumerated.

Monte-Carlo approach

To evaluate the significance of modules we derive their distributions under the null hypothesis of no accumulation of high-scoring genes, using Monte-Carlo permutations.
For i from 1 to B, we iterate the following process:

1. Permute the sample labels for all the genes of the initial expression matrix, denoted $X_{(0)}$.
2. Calculate the p-values for each of the permuted dataset, $X_{(i)}$, and update the score $W_g^{(i)}$ for each gene.
3. Compute and save the module scores, $L_m^{(i)}$.

Finally, the p-value of the whole procedure is given by:

$$p_m = \frac{\text{card}\{i, L_m^{(i)} \geqslant L_m^{(\text{obs})}\}}{B},$$

such as $L_m^{(\text{obs})}$ is the observed score of the mth best module. Note that for each module of size one, the p-value is set equal to the `limma` p-value.

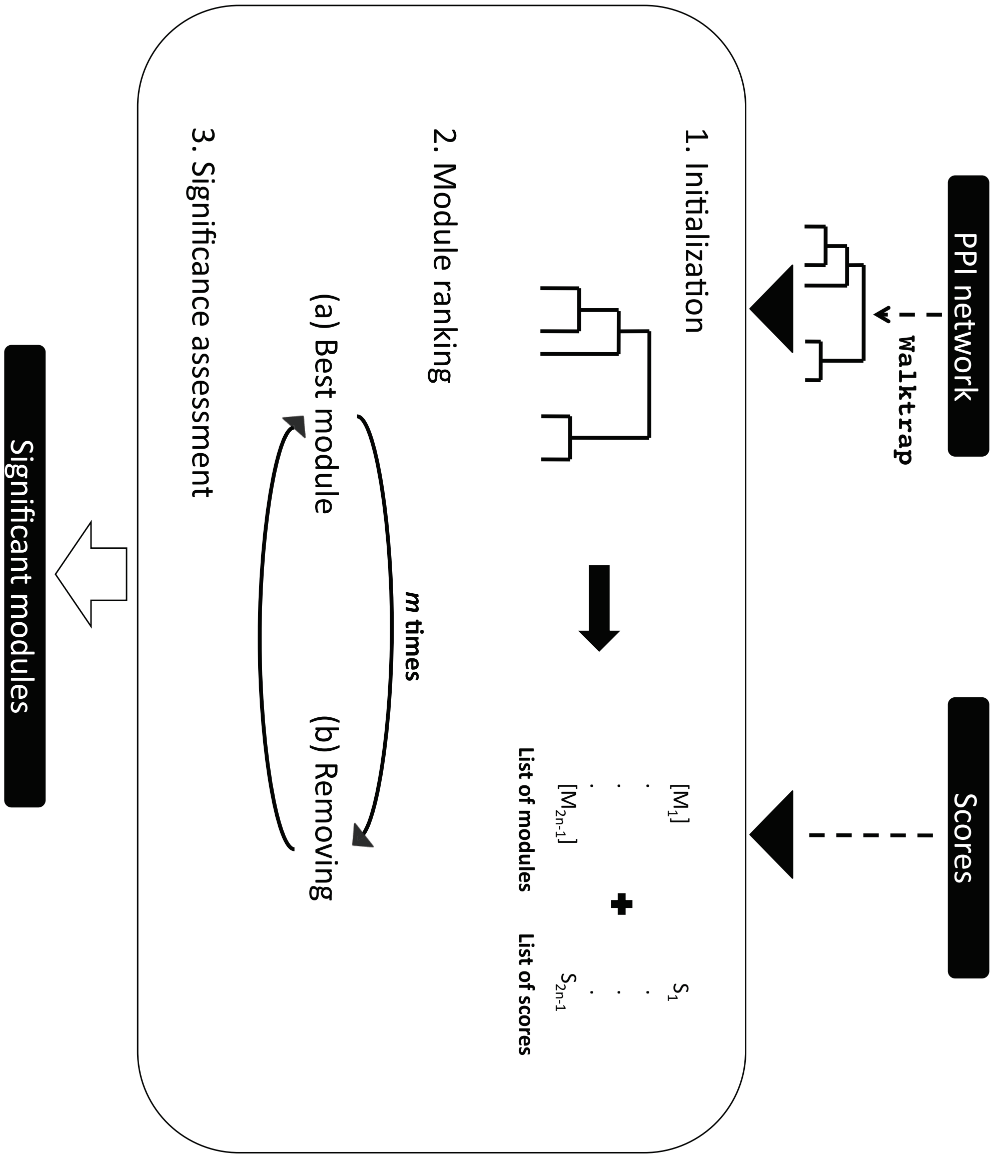


Figure 3.10 – ***DiAMS*** ***global approach.***

A PPI network and a vector of gene scores are passed as input parameters to DiAMS*. The PPI network is then converted into a tree structure using the* walktrap *approach. The three-step algorithm described in 3.2.2 is then performed: (i) initialization (ii) module ranking and (iii) module significance assessment.*

3.2.3 Evaluation strategy

In this section we detail the strategy adopted to evaluate DiAMS. Each evaluation criterion, namely the power, the type-I error rate and the reproducibility, are compared with our modular strategy and its individual scoring counterpart, limma. Here, we perform the simulations under a Gaussian model, *i.e.* for data produced by a microarray experiment.

Power Study Recent results from Gandhi et al. (2006), Lage et al. (2007) or Oti et Brunner (2007), which have motivated the development of DiAMS, suggest that genes involved in the molecular mechanisms of genetic diseases interact together in functional modules. Therefore, to evaluate our approach, we designed a simulation study under this hypothesis of a modular activity of genes. Firstly, it involves randomly sampling significant modules in the tree structure. Secondly, according to the model M_1 described in Section 3.1.2, we simulate a gene expression matrix. The genes belonging to non-significant modules are simulated under the null hypothesis of equality across the mean expression levels for both conditions: $\mu_g^{(1)} = \mu_g^{(2)}$, while genes of significant modules are simulated under H_1, such as $\mu_g^{(2)} = \mu_g^{(1)} + \Delta$, with Δ in $\{0.5, 0.75, 1, 1.25, 1.5, 2, 3\}$. The p-values obtained from Monte-Carlo permutations are then adjusted using the Benjamini-Hochberg procedure to control the FDR criterion at a level of 5%.

The Figure 3.11 illustrates the results of the power analysis for both DiAMS and limma. As expected, the curve describing the statistical power converges to 1 with increasing values of Δ. For $\Delta = 0.5$, it appears that the power is very similar for both approaches , although DiAMS is slightly more powerful. For all values of Δ in $\{0.75, 1, 1.25, 1.5, 2\}$, we observe large differences in power between the two approaches with DiAMS outperforming limma.

We also consider a scenario where genes are simulated independently under H_1, *i.e.* without assuming a modular activity. The power values obtained are identical for both methods, due to the fact that the p-values of individual genes are exactly the same as those resulting from limma. At worst, if the hypothesis of a functional relationship between disease genes is wrong, the power results are equivalent to limma.

False-Positive Rate Using the same simulation strategy as described in the previous subsection, we assess the false-positive rate. A statistical test conducted at a significance level of 0.05 should control the false-positive rate at 5%. Thus, by simulating an entire dataset under H_0, *i.e.* $\forall g : \Delta = 0$, we evaluate the proportion of genes spuriously selected as significant. Both the limma and DiAMS false-positive rates are estimated for various sample sizes ranging from 5 to 50 samples per condition.

Figure 3.12 shows the estimated false-positive rates for both selection methods and various sample sizes. It appears that the rates are similar for

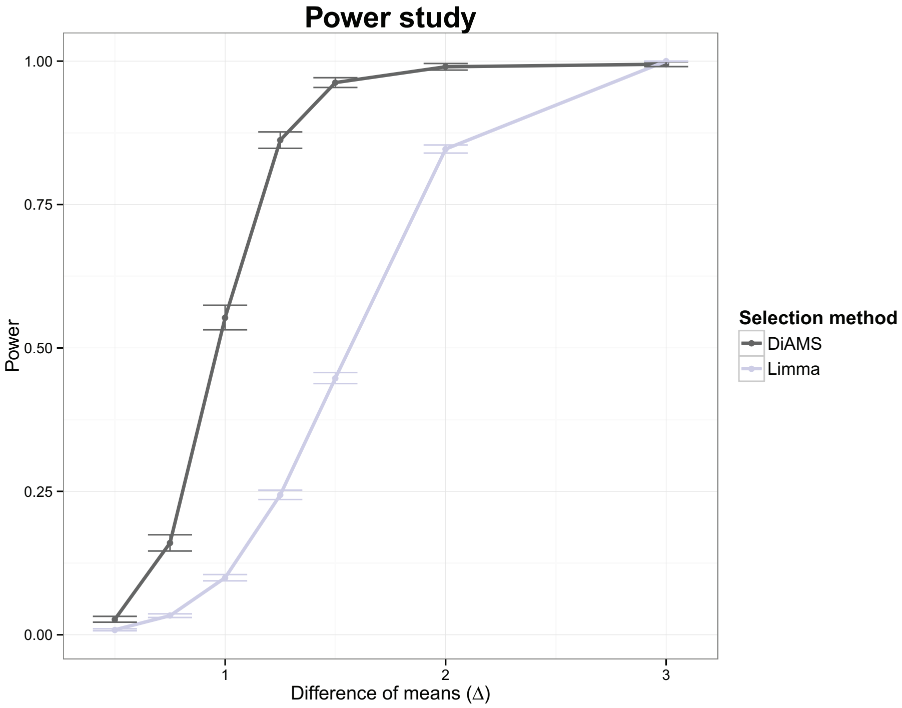


Figure 3.11 – ***Power study.***
The mean of power values over the 1,000 *simulations and its* 95% *confidence interval are calculated at a* 0.05 *FDR level for the* `DiAMS` *method (in dark gray) and the* `limma` *statistic (in light gray) and displayed according to* Δ*, the difference of mean expression levels under* H_0 *and* H_1.

both approaches and they lie within the 95% confidence interval. For each sample size, `limma` and `DiAMS` meet the theoretical false-positive rate.

Reproducibility study Next, we examined the agreement between signatures using a subsampling procedure. As described in the power study, we simulated modules under H_1 as well as the corresponding expression matrix and compute a signature of reference. Then, we randomly subsampled the replicates of the initial matrix with replacement and estimate the signature again. The reproducibility is calculated as the overlap between the reference signature and the signature of subsampled expression matrices. This procedure is performed for various subsample sizes from an initial dataset containing 50 samples for two conditions.

The reproducibility results are averaged over 10,000 simulations and displayed in Figure 3.13. For the larger sample size, the initial matrix has been re-sampled with replacement. Even if the sample size is the same, meaning that the noise added to the initial dataset is relatively low, the percentage of reproducibility for `limma` is only 90% while `DiAMS` almost reaches 100%. All the results displayed in Figure 3.13 show that `limma` is very sensitive to the noise in data while `DiAMS` results appear to be more consistent. This is especially true for small sample sizes, for which the reproducibility of the signature is about 95% with the `DiAMS` approach while the percentage is almost null (0.3%) with the `limma` selection method. The

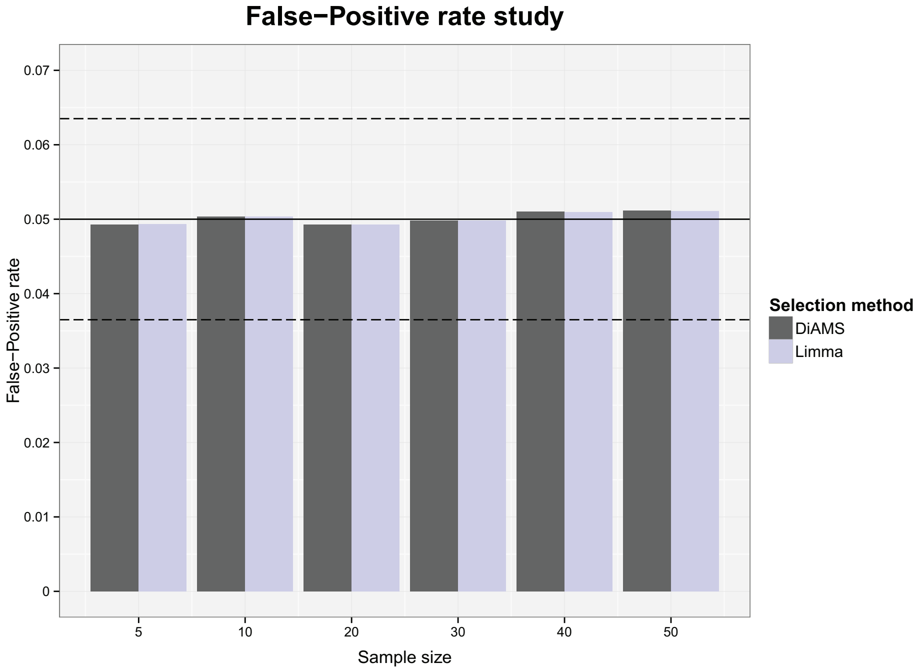


Figure 3.12 – ***False-positive rate study.***
The estimated false-positive rate over the 1,000 *simulations are displayed for the* `DiAMS` *method (in dark gray) and the* `limma` *statistic (in light gray) for various sample sizes. The plain black line represents the* 5% *level at which the tests were conducted. The dashed black lines are the* 95% *confidence intervals for this level.*

gap remains very large for the other sample sizes and `DiAMS` clearly provides significantly better results than `limma` in terms of reproducibility.

3.2.4 Discussion

We developed a network-based approach named `DiAMS` for the selection of gene signatures. We demonstrated through simulations that, under the assumption of a modular activity of genes, `DiAMS` is more efficient in terms of power and reproducibility than the moderated *t*-statistic strategy used in `limma`. The application on breast cancer data in Chapter 5 is a good illustration of the potential of our method for highlighting relevant biological phenomena and shows promising results. In particular, such an approach facilitates the ease of the interpretation of the resulting signature by providing information on molecular mechanisms through the extraction of PPI subnetworks.

However the quality and the coverage of the PPI data is one of the main limitations of this approach. In 2008, estimates of the proportion of known Protein-Protein Interactions suggest that in human, only approximately 10% of interactions have been identified. Moreover, the PPI data are biased towards some particular biological interests. Indeed, some proteins are studied more extensively than others. For instance, the human epidermal growth factor receptor (HER) family of receptor tyrosine kinase is the subject of intense research in breast cancer, like are the tumor suppressor genes BRCA1 and BRCA2. It results in a bias

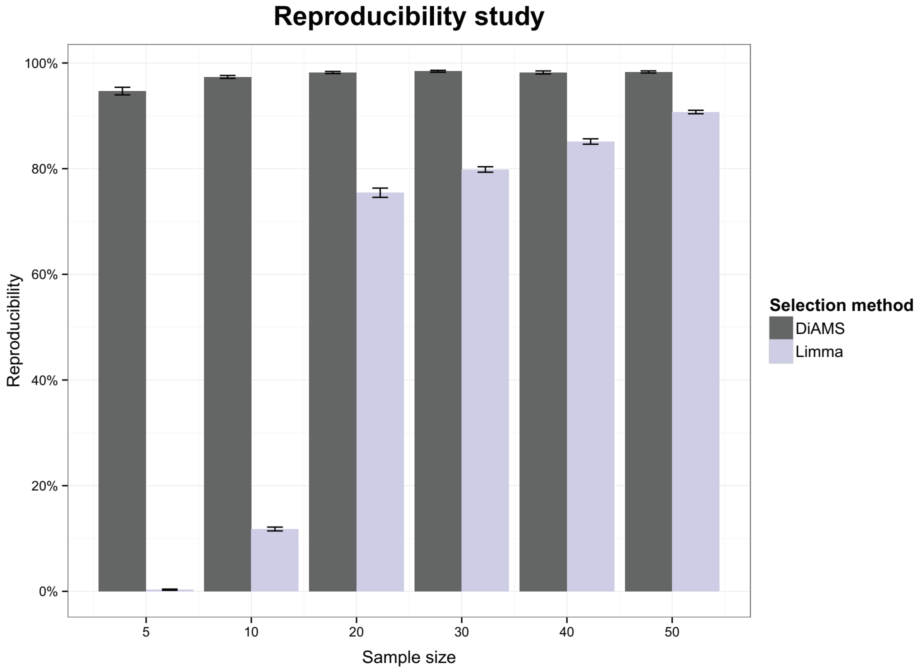


Figure 3.13 – ***Reproducibility study.***
This barplot displays the results of the reproducibility analysis for which we compute the mean of the overlap between a signature of reference and signatures of subsampled expression matrices over 10,000 *simulations. We represent the* 95% *confidence interval for each sample size.*

towards our approach by selecting the most documented proteins. Gene selection approaches based on PPI networks are hence highly dependent on the quality and the amount of available information.

`DiAMS` has the advantage of being easily adjusted to suit various types of data. Indeed, the vector of scores passed to the method could be extracted from genetic association tests, for instance. Such applications to genetic data have been conducted in Pharnext to identify modules of genes associated with Alzheimer's disease. Moreover the input network could be reconstructed from heterogeneous data such as gene regulation information or genetic interactions.

Chapter Conclusion

Molecular signatures yield valuable insight into cell biology as well as relevant information about the mechanisms of human diseases. The identification of sets of genes gives rise to a wide range of statistical developments, particularly in the areas of hypothesis testing. Given the large number of methods available in the literature, the first step towards robust analysis and relevant biological results is to evaluate and compare their performances. Among all of the tests included in the study, we found that `limma` is the best alternative to the classical *t*-test for differential expression analysis of microarray data. In addition to providing recommendations to the scientific community about which methods to employ, we define a standardized comparisons process, relying on both real and simulated data, that can be used as a reference for the integration and evaluation of novel approaches. A comparable study is being carried out for tests dedicated to RNA-seq experiments in Chapter 6.

Regardless of their power, in practice, existing methods fail to replicate gene signatures in independent studies, casting doubts on their reliability and limiting their translation to clinical applications. Indeed, we demonstrate that even `limma` yields to highly unstable signature. We therefore developed `DiAMS`, a method that integrates gene expression and Protein-Protein Interaction data. `DiAMS` holds great promise for discovering reliable molecular signatures by drastically improving their reproducibility. In addition, `DiAMS` provides substantial gains in power while preserving false-positive rate control. The statistical framework developed herein is flexible and potentially applicable to a wide range of data.

4 Interpretation of gene signatures through regulatory networks inference

Once a molecular signature is identified, the challenge lies in its biological interpretation. Indeed, understanding the underlying regulation mechanisms that lead to alterations of gene expression cannot be achieved with a simple examination of the individual genes in the signature. Further biological insights can be gained by investigating the interactions between genes or proteins at the cellular level rather than considering them one at a time. This approach, which entails looking at a biological problem from a larger perspective, refers to the concept of Systems Biology. Due to the development of high-throughput technologies, such systems-wide approaches are now possible. In this field, the study of regulation patterns between genes through the inference of regulatory networks has received much attention. Gaussian Graphical Models (GGMs) provide a well-researched framework to describe conditional dependencies between RNA measurements through the concept of partial correlation. An eventual goal of such networks is to highlight potential functional interactions between genes.

In this chapter we focus on regulatory network inference as a systems approach to interpret molecular signatures. After a preliminary subsection on GGMs, we recall the notions of conditional independence and partial correlation which are central concepts of Graphical Models. We then introduce covariance selection problem that is further formulated as a maximum likelihood estimation problem. In a second section, we underline some extensions proposed in the literature which are dedicated to the inference of gene regulatory networks in a high-dimensional setting under various experimental conditions. Finally, in a third part we present a new statistical development to introduce biological prior knowledge in order to drive the inference of gene regulatory networks.

Two publications are associated with this chapter:

1. Jeanmougin et al. (2011). **Defining a robust biological prior from Pathway Analysis to drive Network Inference.** *J-SFdS, 152(2)*.
2. Jeanmougin et al. (2012b). **Network inference in breast cancer with Gaussian graphical models and extensions** *"Probabilistic Graphical Models for Genetics"*.

4.1 Gaussian Graphical Models (GGMs)

4.1.1 Preliminaries relating to GGMs

Graphical Models are a very useful framework to investigate and represent conditional independence structures among a collection of random variables. This consists of the combination of a graph (or a network) and a probability distribution of the random variables. They have been a matter of interest with the advent of high-throughput experiments due to their ability to capture the complexity of biological systems and provide an intuitive representation of underlying probabilistic models. In particular, graphical models are widely employed in the biomedical research area to describe and to identify interactions between genes and gene products, with the eventual aim to better understand the disease mechanisms mediated by changes in the network structure. In the context of regulatory networks modeling, a graph is a structure consisting of a set of vertices $V = \{1, ..., p\}$, also called nodes, that represents genes and a set of edges E that models interactions between genes. Let us represent the expression levels of the p genes by a Gaussian random vector $X = (X_1, \ldots, X_p)^\intercal \in \mathbb{R}^p$ which follows a multivariate Gaussian distribution: $X \sim \mathcal{N}_p(0, \Sigma)$, with unknown covariance matrix Σ. No loss of generality is involved when centering X, thus we assume that X is drawn from a normal distribution with mean zero in order to simplify the notations. The model detailed above, which assumed a Normal distribution of random variables, is known under the name of Gaussian Graphical Model (GGM). A GGM is represented by an undirected graph, denoted $G = (V, E)$ in the following sections. "Undirected" means that $g \sim h \in E$, *i.e.* an edge between the gth and hth vertices, is equivalent with $h \sim g \in E$.

Network inference consists in estimating the structure of a graph, this involves to determine the set of edges of G defined as follows:

$$E := \{g, h \in V | g \neq h, g \sim h\}.$$

In GGMs, partial correlations are used as a measure of independence of any two genes. As a consequence, two vertices are connected by an edge when the corresponding variables are dependent given the other variables of the graph. Thus the absence of an edge connecting two vertices indicates conditional independence of the two corresponding variables given the other variables. An extensive description of the GGMs theory can be found in Whittaker (1990), Lauritzen (1996) and Edwards (2000).

4.1.2 Conditional independence and partial correlation

In order to elucidate functional interactions from expression data, a popular and simple strategy is to infer *relevance networks* that allow the dependency structure of the data to be visualized. They are based on correlation as a measure of dependency between gene expression levels and enable to find genes which are somehow similar across various experimental conditions. The most widely used measure is the Pearson product moment

correlation that describes the linear relationship between variables. Although the advantages of the relevance network are its straight-forward approach and low computational cost, this approach is only of limited use for understanding gene interaction. Indeed, if correlation highlights co-expressed genes, it does not provide any indication of how the chain of information passes from gene to gene. For instance, the observation of a strong dependency between two variables may be due to the action of another variable. In the context of regulatory network inference, this means that an edge may be identified between two genes that are regulated by the same third variable (see Figure 4.1). Thus, the notion of dependency is too limited to highlight regulation events between genes.
To cope with this problem, we need to consider conditional rather than marginal dependencies between genes. The conditional independence relationships can be inferred from partial correlations in the particular case where random variables follow a multivariate normal distribution. The partial correlation coefficient between the variables X_g and X_h, conditional on all other variables indexed by $V\backslash\{g,h\}$, can be formulated as a normalized expression of the conditional covariance:

$$\rho_{X_g,X_h|X_{V\backslash\{g,h\}}} = \frac{\mathrm{cov}(X_g, X_h|X_{V\backslash\{g,h\}})}{\sqrt{\mathrm{var}(X_g|X_{V\backslash\{g,h\}})\mathrm{var}(X_h|X_{V\backslash\{g,h\}})}}, \tag{4.1}$$

with,

$$\begin{aligned}\mathrm{cov}(X_g, X_h|X_{V\backslash\{g,h\}}) = \mathbb{E}(X_g, X_h|X_{V\backslash\{g,h\}}) \\ - \mathbb{E}(X_g|X_{V\backslash\{g,h\}})\mathbb{E}(X_h|X_{V\backslash\{g,h\}}).\end{aligned}$$

Note that, in the following, we will write $\rho_{g,h|V\backslash\{g,h\}}$ instead of $\rho_{X_g,X_h|X_{V\backslash\{g,h\}}}$ to simplify the notations.

A particularly convenient property of partial correlation is that it allows a distinction to be made between the correlation of two genes due to direct causal relationships, and the correlation that originates *via* intermediate genes. In Figure 4.1, we illustrate the notion of partial correlation. We consider three random variables, X_1, X_2 and X_3 and display their respective variances. This kind of representation enables us to examine the relationship between X_1 and X_3 after removing the effect of X_2. Thus, calculating $\rho_{1,3|V\backslash\{1,3\}}$ boils down to quantifying the shared variance between X_1 and X_3, denoted c on Figure 4.1. It can be formulated in terms of linear regression. When regressing the two random variables X_1 and X_3 on the remaining variable, the partial correlation coefficient between X_1 and X_3 is given by the Pearson correlation of the residuals from both regressions. Intuitively speaking, we remove the linear effects of X_2 on X_1 and X_3 and compare the remaining signals to determine if the variables are still correlated.

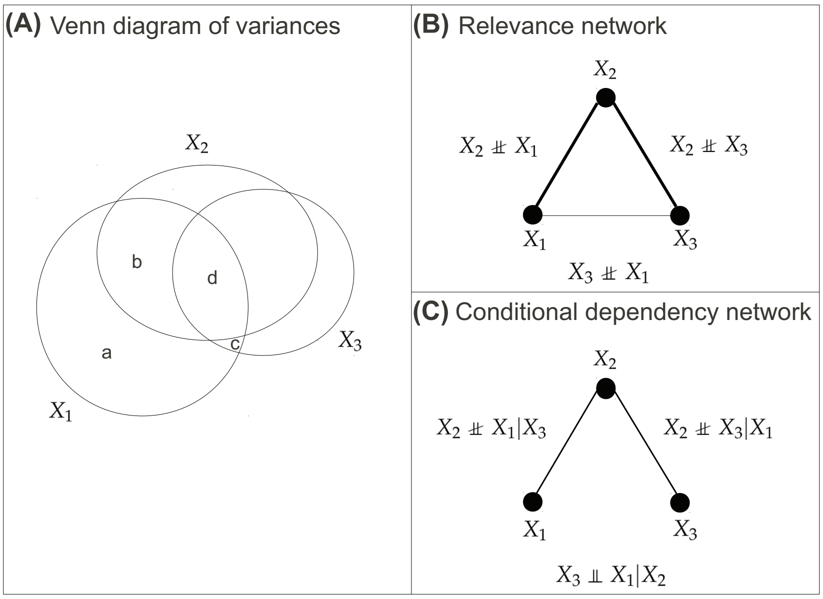
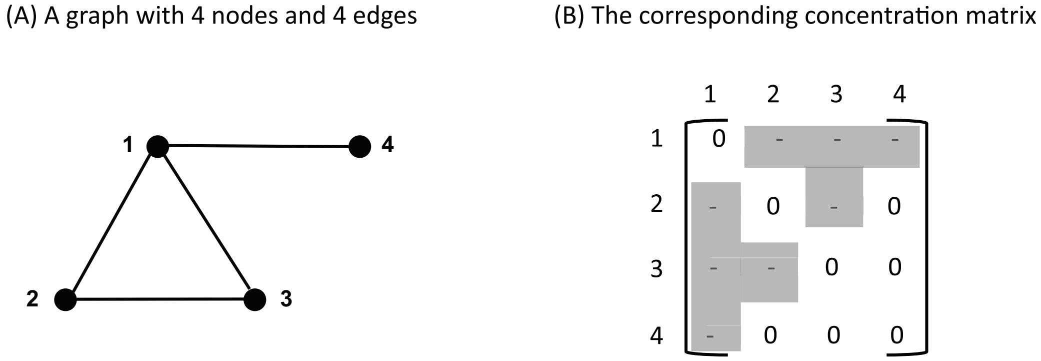

(A) Venn diagram of variances
X_2
b
d
a
c
X_3
X_1
(B) Relevance network
X_2
$X_2 \not\perp\!\!\!\perp X_1$
$X_2 \not\perp\!\!\!\perp X_3$
X_1
X_3
$X_3 \not\perp\!\!\!\perp X_1$
(C) Conditional dependency network
X_2
$X_2 \not\perp\!\!\!\perp X_1|X_3$
$X_2 \not\perp\!\!\!\perp X_3|X_1$
X_1
X_3
$X_3 \perp\!\!\!\perp X_1|X_2$

Figure 4.1 – ***Concepts of correlation and partial correlation.***
(A) Venn diagram of variances: the variance of each variable X_i is represented by a unit circle. The total variability of X_1 is given by $a + b + c + d$ and $b + d + c$ is the variability of X_1 explained by X_2 and X_3. We show that the covariance between X_1 and X_3, denoted by $d + c$, is mainly due to the covariance between X_2 and X_3. Using a simple correlation measure, it yields to the inference of an edge between X_1 and X_3 in the relevance network shown in (B). The partial correlation enables us to measure what the correlation between X_1 and X_3 would be if they were not each correlated with the variable X_2. This area, denoted c, is almost null. Consequently, no edge between X_1 and X_3 is inferred in the conditional dependency network displayed in (C).


(A) A graph with 4 nodes and 4 edges
1
4
2
3
(B) The corresponding concentration matrix

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & - & - & - \\ 2 & - & 0 & - & 0 \\ 3 & - & - & 0 & 0 \\ 4 & - & 0 & 0 & 0 \end{array}$$

Figure 4.2 – ***Covariance selection***
(A) In a GGM, a random variable X_i is associated with each vertex $i \in V$. (B) Zero pattern of the concentration matrix associated with the graph in (A). The set of edges corresponds to the nonzero elements of matrix, in grey.

4.1.3 Covariance selection

Certainly, in a GGM, an edge will be drawn between any two genes in the graph, if they have a nonzero partial correlation coefficient, meaning that their expressions are dependent conditional on all other gene expression levels in the dataset. In other words, the absence of an edge between the gth and hth vertices represents an independence between the variables X_g and X_h, being fixed all other variables:

$$\begin{aligned} \forall(g,h) \in V, X_g \perp\!\!\!\perp X_h | X_{V\setminus\{g,h\}} \quad &\Leftrightarrow \quad \rho_{g,h|V\setminus\{g,h\}} = 0 \\ &\Leftrightarrow \quad g \nsim h. \end{aligned}$$

A result originally emphasized in Dempster (1972) claims that partial correlations are proportional to the corresponding off-diagonal elements of the inverse covariance matrix, called the *concentration matrix* and denoted by $\Theta = (\theta_{gh})_{g,h \in V}$ such as $\Theta = \Sigma^{-1}$. The partial correlation between X_g and X_h given $X_{V\setminus\{g,h\}}$ can be written as follows:

$$\rho_{g,h|V\setminus\{g,h\}} = -\frac{\theta_{gh}}{\sqrt{\theta_{gg}\theta_{hh}}}.$$

We note that the partial correlation is zero, if and only if, the corresponding element of the concentration matrix is zero. Therefore, recovering nonzero entries of Θ is equivalent to inferring the conditional independence graph and after a simple rescaling, Θ can be interpreted as the adjacency matrix of the graph. Consequently, the problem of Graphical Model selection is equivalent to recovering the off-diagonal zero-pattern of the inverse covariance matrix, as illustrated on Figure 4.2, and is often referred as the *covariance selection* problem in the literature. This term makes clear that reconstructing the GGM is a variable selection problem.

4.1.4 Likelihood inference of partial correlations

Let us consider n observations that are mutually independent and distributed according to a multivariate Gaussian distribution $\mathcal{N}_p(0, \Sigma)$. Likelihood inference about the concentration matrix Θ is based on the multivariate normal log-likelihood function:

$$\mathcal{L}(\Theta; S) = \frac{n}{2} \log\det(\Theta) - \frac{n}{2}\mathrm{Trace}(S\Theta) - \frac{np}{2}\log(2\pi),$$

with $S = n^{-1}X^{\intercal}X$ be the empirical variance-covariance matrix. $\det(\cdot)$ and $\mathrm{Trace}(\cdot)$ are the determinant and the trace of a matrix, respectively.
The coefficients of Θ are estimated by recovering the elements which maximize the log-likelihood such as $\widehat{\Theta}^{\mathrm{MLE}} = \arg\max_{\Theta}\{\mathcal{L}(\Theta; S)\}$. Omitting irrelevant factors and constants, the Maximum Likelihood Estimator (MLE) of Θ is defined by:

$$\widehat{\Theta}^{\mathrm{MLE}} = \arg\max_{\Theta} \quad \log\det(\Theta) - \mathrm{Trace}(S\Theta). \tag{4.2}$$

The condition $n \geqslant p$ implies the existence of the global maximum of $\mathcal{L}(\Theta; S)$. Thus, when n is larger than p, the maximization problem 4.2 admits a unique solution S^{-1}.

4.2 Application of GGMs to transcriptome data

4.2.1 Background of high-dimensional inference

There are two major limitations with the MLE regarding the objective of graph reconstruction by recovering the set of nonzero entries of the estimate of Θ from transcriptome data. First, it provides an estimate of the saturated graph: all genes are connected to each other. However, a growing body of biological evidence suggests that among all $p(p-1)/2$ possible interactions between genes, only a few actually take place. This property, called *sparsity*, holds in a wide variety of biological applications and is usually well justified. Indeed, it has been demonstrated in the literature that most molecular networks are not fully connected, see for instance Gardner et al. (2003). In other words they contain many genes with few interactions and a few genes with many interactions. For example, a transcription factor controls only a few genes under specific conditions. In this context, an effective variable selection procedure is needed in order to determine which are the estimated partial correlations that represent actual linear dependencies.

The second major problem with the maximum likelihood approach resides in the data *scarcity*, because n must be larger than p to be able to even define the MLE of Θ. Indeed, in order for an inverse matrix to exist, it must have the property of being full rank. Dykstra (1970) establishes that the covariance matrix has full rank, with probability 1, if and only if $n > p$. This is never the case in high-throughput transcriptome studies because microarray or RNA-seq experiments typically measure the expression level of a huge number of genes across a small number of samples. Thus classical GGM theory is not valid in a small sample setting and its application to transcriptome data is quite challenging.

To overcome the high-dimensional issue, various strategies have been proposed in the literature. The most intuitive approach consists in reducing the number of genes under study in order to satisfy $n < p$ and thus avoiding the dimensionality issue. However, the restriction to a limited number of genes risks that the estimated network topology is seriously distorted because important genes may have been excluded from the analysis. More sophisticated methods include the use of regularized estimates for the covariance matrix and its inverse. In the following subsection we review some state-of-the-art regularization techniques and are particularly interested in lasso-type regularizers for undirected GGMs that ensure both the sparsity of the solution and the existence of the inverse of the covariance matrix.

4.2.2 Inference of sparse regulatory networks

Introduction to norm regularizers

The inversion of the covariance matrix Σ is an ill-posed estimation problem for $n < p$. Regularization approaches turn an ill-posed problem into well-posed one, by ensuring the existence, uniqueness, and stability of its solution. The idea behind regularized estimators is to impose a prior on the model parameters through a penalty term. Here, we are interested in regularizers that impose a norm constraint on the coefficients of the concentration matrix using ℓ_q penalties. In the following, the ℓ_q-norm of the matrix Θ, called $||\Theta||_{\ell_q}$, will refer to the entry-wise norm defined as:

$$||\Theta||_{\ell_q} = \left(\sum_{i=1}^{n} \sum_{j=1}^{p} |\theta_{ij}|^q \right)^{\frac{1}{q}}.$$

We denote respectively by $\|\Theta\|_{\ell_0}$, $\|\Theta\|_{\ell_1}$ and $\|\Theta\|_{\ell_2}$ its ℓ_0, ℓ_1 and ℓ_2-norms, defined as follows:

$$\|\Theta\|_{\ell_0} = \sum_{i=1}^{n} \sum_{j=1}^{p} \mathbb{I}_{\{\theta_{ij} \neq 0\}}, \qquad \|\Theta\|_{\ell_1} = \sum_{i=1}^{n} \sum_{j=1}^{p} |\theta_{ij}|, \qquad \|\Theta\|_{\ell_2} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} \theta_{ij}^2}.$$

For $q = 2$, the norm of the matrix Θ is also known as the Frobenius norm.

Regularization techniques maximize the problem 4.2 subject to a bound on the norm of coefficients of the concentration matrix. This can be written by making an explicit formulation of the size constraint on the coefficients:

$$\begin{aligned} &\arg\max_{\Theta} \log\det(\Theta) - \text{Trace}(S\Theta), \\ &\qquad\qquad \text{subject to} : ||\Theta||_{\ell_q} \leqslant c, \end{aligned}$$

with $c > 0$ or by including the penalty term in the log-likelihood function:

$$\arg\max_{\Theta} \quad \log\det(\Theta) - \text{Trace}(S\Theta) - \lambda||\Theta||_{\ell_q}, \tag{4.3}$$

where $\lambda > 0$. These two formulations are equivalent but for the sake of clarity, in the sequel, we will introduce the various models in terms of penalized maximum likelihood as written in equation 4.3.

The parameter λ is a positive penalty parameter that governs the complexity of the selected model. A large value of λ tends to indicate a simple model, whereas a small value of λ indicates a complex model. In particular, when $\lambda = 0$, all variables are selected and the model is even unidentifiable when $p > n$. The interesting cases lie between these two choices. The tuning of λ is typically done through cross-validation or information criteria (BIC, AIC).

The maximization problem 4.3, involving a penalty term on the ℓ_q norm of the parameters with $q \geqslant 0$, is known as the Bridge regularization and is discussed in Hastie et al. (2001). In the following paragraphs we introduce various regularization techniques, namely subset selection ($q = 0$), ridge regularization ($q = 2$) and finally the lasso ($q = 1$), in the context of variable selection.

Regularization with ℓ_0 penalties

Subset selection using ℓ_0 regularizers is a standard statistical method which computes the following estimator:

$$\arg\max_{\Theta} \quad \log\det(\Theta) - \mathrm{Trace}(S\Theta) - \lambda||\Theta||_{\ell_0}. \tag{4.4}$$

The ℓ_0-norm controls the number of nonzero coefficients in Θ by penalizing the dimensionality of the model. Given $||\Theta||_{\ell_0} = k$, the solution to 4.4 is the subset with the largest maximum likelihood among all subsets of size k.

ℓ_0-regularization leads to interpretable models by producing sparse solutions. However, solving 4.4 would require the exploration of all possible subgraphs which is computationally too expensive for applications to network inference from high-throughput transcriptome data. To tackle this problem other penalty functions should be used when working with high-dimensional models.

Regularization with Ridge penalties

The Ridge regularization or Tikhonov regularization, first introduced by Hoerl et Kennard (1970), adds an ℓ_2-regularization term in the optimization problem that encourages the sum of the absolute values of the parameters to be small:

$$\arg\max_{\Theta} \quad \log\det(\Theta) - \mathrm{Trace}(S\Theta) - \lambda||\Theta||_{\ell_2}. \tag{4.5}$$

The Ridge regularization yields a well-conditioned solution to the inverse problem. Thus, the inverse matrix will always exist as long as λ is strictly positive. In addition, it has been demonstrated that the use of ℓ_2 penalties yields to a stable estimate of the concentration matrix. However, due to the nature of the ℓ_2-constraint displayed on Figure 4.3-B, the Ridge

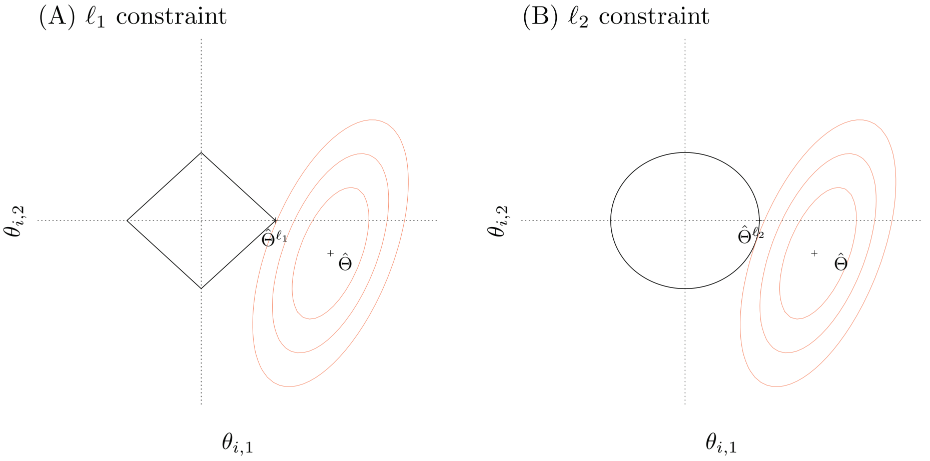


Figure 4.3 – ***Unit ball for (A) ℓ_1 and (B) ℓ_2 constraints***.
$\hat{\Theta}$ denotes the MLE. The orange curve represent the loss function, defined as the negative log-likelihood, that we seek to minimize. (A) The zone bounded by straight black lines shows the ℓ_1 constraint area. The $\hat{\Theta}^{\ell_1}$ minimizing the penalized function generally lies on one of the singularities of the ball as illustrated here. Thus, the constrained solution typically has zero coordinates. (B) The spherical shape of the ℓ_2 constraint area does not favor sparse solutions. Indeed the loss function does not hit the constraint area on one of the axes.

regularization produces a very dense graph. Indeed, even if it shrinks the coefficients towards zero, they will never become exactly zero. So, when the number of predictors is large, Ridge regression does not provide any interpretable model and is inappropriate for network inference in a sparse context.

Regularization with Lasso penalties

To combine both, the stability of the ridge regularization and the interpretability of the ℓ_0 regularization, Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator, known as the *Lasso* or *basic pursuit* in the signal processing literature, see Chen et al. (2001). The Lasso estimator proposed in Banerjee et al. (2008) directly consider the following penalized log-likelihood problem:

$$\arg\max_{\Theta} \quad \log\det(\Theta) - \mathrm{Trace}(S\Theta) - \lambda||\Theta||_{\ell_1}. \tag{4.6}$$

The Lasso has the desirable property of encouraging many parameters to be exactly zero due to the nature of the ℓ_1-constraint, see Figure 4.3. In other words, it encourages conditional independence among variables. Larger values of λ lead to more entries of Θ being estimated as zero. Thus, in comparison to subset selection, the Lasso performs a kind of continuous subset selection.

Equation 4.6 can be efficiently solved using the Graphical Lasso, based on an algorithm initially introduced in Banerjee et al. (2008) and revisited in Friedman et al. (2008), which maximizes the penalized log-likelihood function. It relies on fast coordinate descent algorithms to solve the lasso problem. Meinshausen et al. (2006) take a more naive approach to this problem; they estimate a sparse Graphical Model by linearly regressing X_g with an ℓ_1 penalty on the rest of the nodes, $X_{V\backslash\{g\}}$. Thus, it consists in solving p independent ℓ_1-penalized regression problems and successively estimating each gene neighborhood. The component θ_{gh} of the concentration matrix is estimated by the nonzero elements of $\hat{\beta}_g$ solving problem (4.7).

$$\hat{\beta}_g = \arg\min_{\beta \in \mathbb{R}^{p-1}} \frac{1}{n} \left\| \mathbf{X}_g - \beta \mathbf{X}_{V\backslash\{g\}} \right\|_{\ell_2}^2 + \lambda \left\| \beta \right\|_{\ell_1}. \tag{4.7}$$

This method is known by the name of *neighborhood selection*. The main drawback of such a procedure is that a symmetrization step is required to obtain the final network. It might, for instance, be the case that the estimated coefficient of the regression coefficient of variable g on h is zero, whereas the the estimated coefficient of variable h on g is nonzero. However, this procedure has been reported to be very accurate in terms of edge detection.

4.2.3 Multiple inference under various experimental conditions

Context

The focus so far in the previous sections has been on estimating a single Gaussian Graphical Model. However, in transcriptome experiments, we usually have to deal with data generated under various conditions. From a statistical point of view, we consider n observations collected under C different conditions. A common practice in GGM-based inference method consists in merging the C different experimental conditions. This strategy has the advantage of enlarging the number of observations available for inferring regulations. However, GGMs assume that the observed data form an independent and identically distributed sample which is obviously wrong when data are collected in different conditions. Thus, such a strategy is likely to have detrimental effects on the estimation process. In addition, we may be interested in comparing the regulation patterns under the different conditions and merging the data leads to the inference of a unique network which is hardly interpretable in practice. Another strategy is to merely ignore the relationship between data and infer a network separately in each condition. Thus, each sample is assumed to be drawn independently from a Gaussian distribution such as: $X^{(c)} \sim \mathcal{N}_p(0, \Sigma^{(c)})$. By following the approach described in subsection 4.2.2, the objective function exhibits the same form as the equation 4.6:

$$\arg\max_{\{\Theta^{(c)}\}_{c=1}^{C}} \sum_{c=1}^{C} \left(\mathcal{L}(\Theta^{(c)}; S^{(c)}) - \lambda ||\Theta^{(c)}||_{\ell_1} \right),$$

where $\mathcal{L}(\Theta^{(c)}; S^{(c)})$ denotes the Gaussian log-likelihood function in condition c. We expect that the regulation patterns under various conditions are not exactly the same but in practice, this approach leads to c graphs which exhibit dramatically different structures. This is partly due to the noise inherent in microarray data and to the generally small amount of available data. However, from a biological point of view, subpopulations are assumed to share a large common core of edges and only differ by a small subset of edges. Thus, when recovering the structure for one graph under a given condition, we would like to use evidence from other conditions as supporting information. This becomes particularly important in settings with limited amount of data, such as in transcriptome studies. In this context, jointly estimating the models allows for a more efficient use of data which is available for multiple related conditions. This can be achieved by either modifying the log-likelihood function or the penalizer as proposed in Chiquet et al. (2011). In the following paragraphs we detail the use of structured penalizers and in particular mixed norms for the inference of multiple GGMs.

Mixed norms

In the context of mutliple GGM inference, the use of mixed norms aims to encourage similar sparsity patterns across conditions. In other words, mixed norms favor graphs with common regulations, *i.e.* common edges, by grouping each partial correlation coefficient across conditions instead of performing independent estimations.
We call mixed norm of Θ the $\ell_{(q,r)}$-norm defined as:

$$||\Theta^{(c)}||_{\ell_{(q,r)}} = \left(\sum_{g \neq h} \left(\sum_{c=1}^{C} |\theta_{gh}^{(c)}|^r \right)^{\frac{q}{r}} \right)^{\frac{1}{q}}.$$

It can be seen as a two-stage penalization. First, groups are penalized by a ℓ_q-norm , then for variables within each group, an ℓ_r-norm is applied. The coupling is strongly dependent of the choice of r and q. Hence, one can favor sparsity across the groups (with q close to 1) or inside each group (with r close to 1). This approach, which consists in coupling the estimation the estimation of $\Theta^{(1)}...\Theta^{(C)}$ across various conditions (or tasks), is termed of *multi-task* learning in machine learning literature.

Group-Lasso penalty

One example of mixed norms is the Group-Lasso introduced by Yuan et Lin (2006), which uses a $\ell_{(1,2)}$ penalization. The MLE of the Group-Lasso, denoted $\hat{\Theta}^{\text{GL}}$, is given by:

$$\hat{\Theta}^{\text{GL}} = \underset{\{\Theta^{(c)}\}_{c=1}^{C}}{\arg\max} \sum_{c=1}^{C} \mathcal{L}(\Theta^{(c)}; S^{(c)}) - \lambda \sum_{g \neq h} \left(\sum_{c=1}^{C} (\theta_{gh}^{(c)})^2 \right)^{\frac{1}{2}}. \tag{4.8}$$

We simplify the previous equation as following:

$$\hat{\Theta}^{\text{GL}} = \underset{\{\Theta^{(c)}\}_{c=1}^{C}}{\arg\max} \sum_{c=1}^{C} \mathcal{L}(\Theta^{(c)}; S^{(c)}) - \lambda \sum_{g \neq h} ||\theta_{gh}^{[1:C]}||_{\ell_2}, \tag{4.9}$$

where $\theta_{gh}^{[1:C]} = \left(\theta_{gh}^{(1)}, .., \theta_{gh}^{(C)}\right)^T \in \mathbb{R}^C$ is the vector of the θ_{gh}'s across tasks.

The Group-Lasso norm introduces a sparse selection of groups through the ℓ_1 penalization and preserves all group members due to ℓ_2-norm properties. Thus, variables enter or leave the support group-wise. If no threshold is applied to the concentration matrix the learning problem 4.9 will lead to a common structure of graphs across the different conditions. Regarding the comparison of regulation pattern between conditions, this formalization is not really satisfactory.

Cooperative-Lasso penalty

Chiquet et al. (2011) developed an alternative strategy, implemented in the R package `SIMoNe` and known under the name of Cooperative-Lasso or Coop-Lasso, which is built on the Group-Lasso penalty described in the previous subsection. The motivation behind the Coop-Lasso is to preserve the type of regulation, namely activation or repression, by encouraging solutions with similar sign patterns across conditions. Thus, the Coop-Lasso disconnects the selection of up and down regulations by applying a Group-Lasso constraint separately on positive and negative coefficients of the concentration matrix:

$$\hat{\Theta}^{\text{Coop}} = \arg\max_{\{\Theta^{(c)}\}_{c=1}^{C}} \sum_{c=1}^{C} \mathcal{L}(\Theta^{(c)}; S^{(c)}) - \lambda \sum_{g \neq h} \left(||(\theta_{gh}^{[1:C]})_+||_{\ell_2} + ||(\theta_{gh}^{[1:C]})_-||_{\ell_2} \right),$$

where $(\theta_{gh})_+ = \max(0, \theta_{gh})$ and $(\theta_{gh})_- = \max(0, -\theta_{gh})$. In this way, the Coop-Lasso allows various regulation patterns across conditions. For instance, the resulting graphs may activate an up-regulation under a given condition while this regulation disappears under the other ones, as illustrated on Figure 4.4-B.

4.3 Inferring high-dimensional regulatory networks from biological prior knowledge

4.3.1 Motivations

Gaussian Graphical Models are promising probabilistic tools for regulatory network inference that have been widely used in the literature over the past decade, see for instance Wille et al. (2004), Scutari et Strimmer (2011) or Kramer et al. (2009). However, recovering network structures based on high-throughput transcriptome data remains a major challenge in Systems Biology for several reasons. Firstly, even if the use of regularization approaches enables us to restrict the estimation problem by enforcing parsimony in the model, the space of possible networks is often too large compared to the limited size of available data. Secondly, the noise inherent in the gene expression measurements leads to graphs with poor robustness. Finally, the difficulty of inferring relevant regulatory networks lies in the complexity of regulation processes that involved a large number of actors and interactions but also a myriad of mechanisms that occur at various levels. Indeed, the gene activity is influenced by transcription

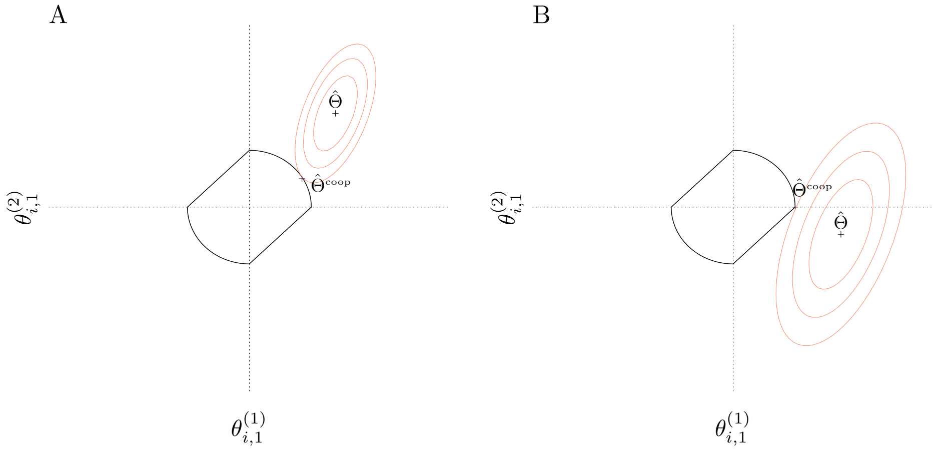


Figure 4.4 – ***Unit ball for the Coop-Lasso****.*

Illustrations of the Coop-norm with a group of coefficients $\mathcal{G} = \{\theta_{i,1}^{(c)}\}$ *under two conditions* $c \in \{1,2\}$ *. (A) The Coop-Lasso encourages solutions where signs match within a group, such as all coefficients are either non-negative or non-positive. (B) The Coop-norm allows to activate an up-regulation under a given condition,* $\theta_{i,1}^{(2)} > 0$*, while this regulation disappears under the other one,* $\theta_{i,1}^{(1)} = 0$.

factors but also by the degradation of proteins and transcripts as well as the post-translational modification of proteins as mentioned in Chapter 2. Thus, using only transcriptome data is not sufficient and enables to understand only a limited part of regulation mechanisms.

Therefore the quality of network reconstruction could be significantly improved by the integration of heterogeneous data. It may help to limit the set of candidate networks and to infer more robust structures. Additionally, it provides a better understanding of the complex regulation behavior of cells. Some research has already been done to associate biological informative prior to drive network inference in a Bayesian network framework Mukherjee et Speed (2008) or by the use of dynamic Bayesian networks Bernard et Hartemink (2005) from time series data. In the field of Machine Learning other types of approach have been proposed in Yamanishi et al. (2004) and Vert et Yamanishi (2005) based on kernel metric learning. GGMs provide a convenient probabilistic framework to integrate biological knowledge as a prior information. In this section we propose an extension of the Coop-Lasso regularization devoted to the incorporation of a prior on the graph structure in order to drive the network inference. Roughly, this prior is defined from biological pathways and is based on the assumption that genes involved in similar cellular mechanisms are likely to be connected in the network. The following subsections are dedicated to the description of the model and to the construction of the structure prior.

4.3.2 Network Inference from a biological prior

The idea underpinning the integration of a biological prior is to bias the estimation of the network structure towards a given topology. Thus, we wish to estimate the correct graph in a more robust way. In the following model 4.10, we incorporate the prior information into the maximization problem by adding an additional constraint whose entries depend on the prior structure. Let us assume that the graph we wish to infer is endowed with a structure Z which clusters the genes into a set $\mathcal{Q} = \{1, \ldots, Q\}$ of given overlapping clusters. For any gene g, the indicator variable Z_{gq} is equal to 1 if $g \in q$ and 0 otherwise, hence describing to which cluster the gene g belongs. The structure of the graph is thus described by the matrix $\mathbf{Z} = (Z_{gq})_{g \in V, q \in \mathcal{Q}}$. The group structure over the Q gene clusters combined to the multi-task inference strategy lead to estimating the C concentration matrices which are the solutions of the following penalized log-likelihood maximization problem:

$$\underset{\{\Theta^{(c)}\}_{c=1}^{C}}{\arg\max} \sum_{c=1}^{C} \mathcal{L}(\Theta^{(c)}; S^{(c)}) - \lambda \sum_{\substack{g,h \in V \\ g \neq h}} \rho_{\mathbf{Z}_g \mathbf{Z}_h} \left(||(\theta_{gh}^{[1:C]})_+||_{\ell_2} + ||(\theta_{gh}^{[1:C]})_-||_{\ell_2} \right), \tag{4.10}$$

where the coefficients of the penalty are defined as:

$$\rho_{\mathbf{Z}_g \mathbf{Z}_h} = \begin{cases} \sum_{q,\ell \in \mathcal{Q}} Z_{gq} Z_{h\ell} \frac{1}{\lambda_{\text{in}}}, & \text{if } g \neq h, \text{ and } q = \ell, \\ \\ \sum_{q,\ell \in \mathcal{Q}} Z_{gq} Z_{h\ell} \frac{1}{\lambda_{\text{out}}}, & \text{if } g \neq h, \text{ and } q \neq \ell, \\ \\ 1, & \text{otherwise.} \end{cases} \tag{4.11}$$

The second part of the criterion is a penalty, which considers two types of edges: edges between two genes belonging to the same cluster are penalized with a coefficient $1/\lambda_{\text{in}}$ and edges between two genes which are never present together in a cluster are penalized with a coefficient $1/\lambda_{\text{out}}$. The intuition behind this model is that the presence of an edge between two genes of the network will be promoted or penalized depending on whether the genes belong to the same cluster or not. The basic form of the penalty has been proposed by Ambroise et al. (2009), which drive the penalty matrices from a topological prior inferred from the data rather than integrating a prior information from biological knowledge. They use the Stochastic Bloc Model (SBM) framework, which provides mixture models for random graphs, to estimate the prior. The idea is similar to what we propose: the elements of the concentration matrix are penalized according to the unobserved clusters to which the nodes belong.

SBM structures are integrated within the network inference strategy: (i) firstly, an initial graph is inferred via a usual ℓ_1 regularized GGM (ii) secondly, the latent structure is estimated via an SBM algorithm (iii) finally, the penalty matrix is derived from SBM parameters and the corresponding network is inferred. Inference of such models has been implemented in the R package `mixer` and included in `SIMoNe`. Ambroise et al. (2009)

have shown via simulations that the knowledge of an existing group structure was indeed improving the estimation of the nonzeros entries of the concentration matrix. Details about a large panel of methods to infer SBM can be found for instance in Daudin et al. (2008) or Latouche et al. (2011).

4.3.3 Structure prior definition

Various sources of biological information are available in the literature. It includes different types of 'omics' data such as genomic or proteomic data as well as other kind of technologies. For instance, ChIP-on-chip experiments allows to identify interactions between transcription factors and genes and enables to derive potential gene regulatory effects. As proposed in the previous chapter the use of Protein-Protein Interactions may also be of great interest for supporting the reconstruction of gene networks. Here we investigate the integration of biological pathways to define clusters of genes as detailed in section 4.3.2.

Pathways

Biological pathways are defined as sets of genes, or more precisely gene products, that interact in order to achieve a specific cellular function.Three types of pathways are distinguished in the literature: metabolic pathways, signaling pathways and transcription and protein synthesis pathways. Metabolic pathways characterize biochemical reactions that achieve basic cellular functions such as energy metabolism or fatty acid synthesis. Signaling pathways are responsible for transmitting information within and between cells and for coordinating metabolic processes as well as molecular activities. Often, signaling pathways result from the interaction of various components of different pathways. Finally transcription and protein synthesis pathways involve the mechanisms of protein synthesis from DNA.

Over-representation tests

Testing for over-representation of pathways, as mentioned in Draghici et al. (2003), is becoming especially popular in the field of gene expression data analysis. Starting with a set of differentially expressed genes, an over-representation test aims to identify the pathways that are over-represented in the set of genes of the signature, as shown in Manoli et al. (2006). An eventual goal of this approach is to highlight pathways that are targeted by the molecular signature. For instance, in a study where patients with disease are compared to healthy controls, it allows pathways to be found that are likely to be involved in the disease mechanisms.

Given the p genes measured on a microarray, the signature is defined as a subset of s genes, and a given pathway is defined as another subset of length t of these p genes. Let us assume that we observe y of these t genes that are differentially expressed. The probability of having y genes of a given pathway in the list of differentially expressed genes is modeled by the hypergeometric distribution $Y \sim H(s, p, t)$ such as:

$$\mathbb{P}(Y = y) = \frac{\binom{s}{y}\binom{p-s}{t-y}}{\binom{p}{t}}.$$

Under the null hypothesis of no over-representation, the probability of observing at least y genes of a pathway of size t in the signature can be calculated by

$$\begin{aligned} \mathbb{P}(Y \geq y) &= 1 - \mathbb{P}(Y \leq y) \\ &= 1 - \sum_{i=0}^{y} \frac{\binom{s}{i}\binom{p-s}{t-i}}{\binom{p}{t}}. \end{aligned}$$

The probability $\mathbb{P}(Y \geq y)$ corresponds to the p-value of a one-sided test. A pathway is said to be significant if the null hypothesis of no over-representation is rejected.

Core pathways definition

The difficulty in over-representation analysis lies in the interpretation of the list of significant pathways. Indeed, there is not yet a standardized definition for pathways and they do not clearly represent distinct entities. Thus, two pathways can involve common genes and hence share common biological information. Therefore, we propose to summarize the list of pathways found significant because of the same genes into a reduced set of "core pathways", each core pathway being constituted of a set of pathways. In practice we apply a hierarchical clustering algorithm on a binary matrix, denoted by $M = (m_{u,v})_{1 \leqslant u \leqslant y,\ 1 \leqslant v \leqslant k}$, where y is the length of the signature and k the number of significant pathways, such that:

$$m_{u,v} = \begin{cases} 1 & \text{if the gene } u \text{ belongs to the pathway } v, \\ 0 & \text{otherwise.} \end{cases}$$

Dissimilarity between pathways, which accounts for pairwise differences between two given pathways (denoted v_1 and v_2 in the following), is assessed by using a binary metric, also known as the Jaccard distance:

$$J_\delta = 1 - \frac{\sum_{u=1}^{y} \mathbb{I}_{\{m_{u,v_1}=1,\ m_{u,v_2}=1\}}}{y - \sum_{u=1}^{y} \mathbb{I}_{\{m_{u,v_1}=0,\ m_{u,v_2}=0\}}}.$$

This metric measures the percentage of nonzero elements of two binary vectors that differ. In our case, it corresponds to the percentage of genes that belong exclusively to either one of the two pathways of interest. Finally, from this dissimilarity matrix we perform a Hierarchical Agglomerative Clustering (HCA) using Ward's criterion. The HCA allows us define clusters of pathways, or core pathways, that are used to construct the penalty matrix $\rho_{\mathbf{Z}_g \mathbf{Z}_h}$ as suggested in the model 4.10. A major advantage to the use of core pathways is that our analysis is not database dependent as it is not limited to the strict definition of pathways found in the databases.

Robustness study

In this section, we study the effects of small sample size and noise in the expression data on the robustness of network estimation. In particular, we compare how both parameters impact the inference with and without priors. Although the present study is still in progress, we set out the general idea and the preliminary results here.

Simulations
The overall simulation process relies on the Hess expression dataset described in the next paragraph. It consists in five steps:

1. Determine a molecular signature from the original expression dataset. In the sequel, the expression dataset associated to the genes of the signature will be called the *signature dataset*.

2. Conduct an over-representation analysis on the resulting signature and define the core pathways that will be used as a prior knowledge to drive the network inference.

3. Infer a network, denoted $G_0^{(c)}$, from the expression levels of the signature dataset and under each condition c. The inference is conducted according to the model proposed in 4.3.2 and implemented in the R package `SIMoNe`, using the core-pathways previously defined. The C networks inferred will be referred to as the *reference networks* in the following.

4. In the signature dataset, sample or subsample with replacement the observations for each condition. Note that the proportion of observations is kept the same across conditions.

5. From the (sub)sampled dataset, proceed as described in step 3 to infer networks, denoted $G_i^{(c)}$, under each condition. Then, assess the percentage overlap between the adjacency matrices of $G_i^{(c)}$ and $G_0^{(c)}$, *i.e.* the number of common edges.

The steps 4 and 5 are run $1,000$ times and give rise to a set of graphs $G_1^{(c)}, ..., G_{1000}^{(c)}$. The sampling procedure was done for various sample sizes. The practical issues of over-representation analysis and definition of the core pathways are only briefly mentioned here but we discuss it in deeper details in Chapter 5.

In order to compare the results of robustness, the same process is applied to infer networks without prior knowledge and with a topological prior inferred from the data, by removing the second step and conducting the network inference without introducing the core-pathway information. In the latter case, the estimation of the topological prior is based on the Stochastic Bloc Model (SBM) framework mentioned in 4.3.2.

Data

The dataset we use to evaluate the robustness of network estimation, called the *Hess dataset*, is described in Jeanmougin et al. (2011) and was initially published by Hess et al. (2006), who study the response to chemotherapy in breast cancer patients. They look at the pathologic complete response, defined as the absence of disease in both the breast and lymph nodes, as an early surrogate marker of treatment efficacy. The data included 29 breast cancer patients for which the pathologic complete response status is known: 15 of them achieved a pathologic complete response and are denoted pCR while the remaining 14 patients, called notpCR patients, did not achieve a pathologic complete response.

The differential analysis of the gene expression profiles between the pCR and notpCR samples was performed using `limma` and yields about 100 genes with statistically significant differences at a 10^{-3} level. The over-representation test was conducted using the `KEGG` (Kyoto Encyclopedia of Genes and Genomes) database from Kanehisa et al. (2006). It led to the identification of 22 significant pathways at a 5% level, summarized in 6 core pathways.

Preliminary results and perspectives

The overlap with the reference networks is calculated for each sampled dataset and displayed in Figure 4.5 for the networks inferred (i) without biological prior information ("Noprior") (ii) with the information of core pathways ("BioPrior") and (iii) with the topological prior ("TopoPrior"). We found that introducing biological priors offers substantial gains in term of robustness. Indeed, the results highlight an overall higher overlap with the reference network: in the case where a biological prior was used, the estimation exhibits a higher degree of reproducibility than the network estimation done without informative prior. In addition, we note that the introduction of a topological prior inferred from the data yields to similar results than those we obtain without any prior information. This suggests that the addition of *a priori* knowledge is not sufficient to ensure the robustness of the network. This prior has to be relevant otherwise it has no effect or, worse still, adverse consequences on the graph estimation.

However, in the best case the mean overlap value only reaches an unsatisfactory 24%. This can be explained by the difficulty of the estimation problem due to the small sample available in this dataset. We are currently replicating these simulations using larger sample sizes. In addition, we are investigating the sensitivity to prior strength, tuned by the parameter λ_{in}.

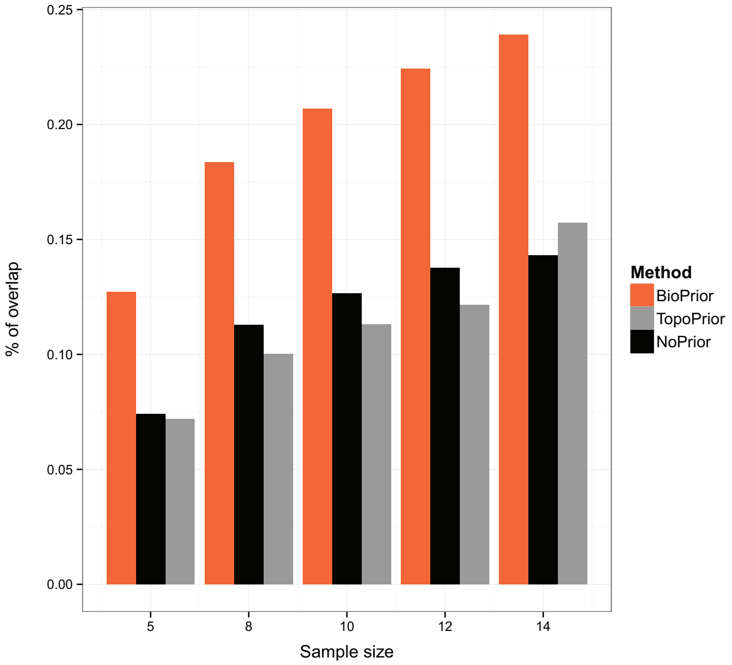


Figure 4.5 – ***Robustness of network inferences***.

We assess the overlap between the reference networks and the networks inferred from sampled expression matrices. Here we show the median overlap value over $1,000$ *simulations for various sample sizes. Both the estimation with prior knowledge.*

Chapter Conclusion

The focus of this chapter was the inference of regulatory networks, as tools for interpreting molecular signatures. Network inference is a very challenging issue in Systems Biology for which GGMs constitute a promising tool, under active development. From the initial works of Ambroise et al. (2009) and Chiquet et al. (2011), we implemented a global framework to infer robust networks on the basis of a biological informative prior over network structures. It has the advantage to reduce the space of possible network structures to investigate, and aims to propose a more interpretable network as illustrated in the next chapter. In addition, we found that introducing prior knowledge to drive the inference provides gains in term of robustness of the network estimation. This method is implemented in the R package SIMoNe.

There are various interesting questions and possible extensions to this work. First of all, there is an essential need for high-dimensional testing frameworks in order to derive confidence intervals on estimated networks and statistically validate the differences observed between inferred networks. From a methodological point of view, the crucial issue of how to tune the overall amount of penalty is still a matter of discussion for which cross-validation is a popular solution. However, its own construction makes it more suited to prediction problems than selection problems. Finally, even though we have provided a way to improve the robustness of network inference, the learning process remains quite unstable due to the high-dimensional setting and the high level of noise. Thus, there is a need for further development, such as bootstrap-based studies, for instance, to address this issue.

Application to breast cancer 5

To illustrate the statistical developments that we have introduced in this thesis, we applied the flowchart displayed in Figure 5.1 to study the metastatic relapse of Estrogen Receptor negative breast cancers. A metastatic relapse of breast cancer occurs when the patient experiences a recurrence of cancer in other parts of the body. It is a common occurrence after surgical tumor removal, most frequently resulting from the spread and the outgrowth of minimal residues of the primary tumor, through the lymphatic or blood system in bones, liver, lungs or even brain. Another possibility for explaining the recurrence of distant disease in patients who were free of overt metastases after initial treatments, is the existence of occult micrometastases, present at the time of diagnosis and surgery. Although trends indicate that survival is improving in patients with metastatic breast cancer, their prognosis remains generally poor. The possibility of prevention and early detection of the occurrence of metastatic relapse could therefore lead to a higher chance of survival and better quality of life. In this chapter, we aim to highlight the mechanisms underlying metastatic relapse by identifying a relevant molecular signature and searching for potential altered gene regulations. First of all, we provide a brief overview of breast cancer disease. In addition, we detail the transcriptome and Protein-Protein Interaction data used to conduct this survey as well as the preprocessing steps. In the second section, we apply `DiAMS` to select a signature of metastatic relapse and we evaluate the resulting set of genes for interpretability and relevance. We then perform a pathway analysis, including an over-representation test and the identification of core pathways, in order to highlight the major mechanisms underlying the relapse of breast tumors. Finally, we compare patterns of gene regulation between patients who relapsed (MR patients) and those who have not experienced a relapse of cancer (notMR patients).

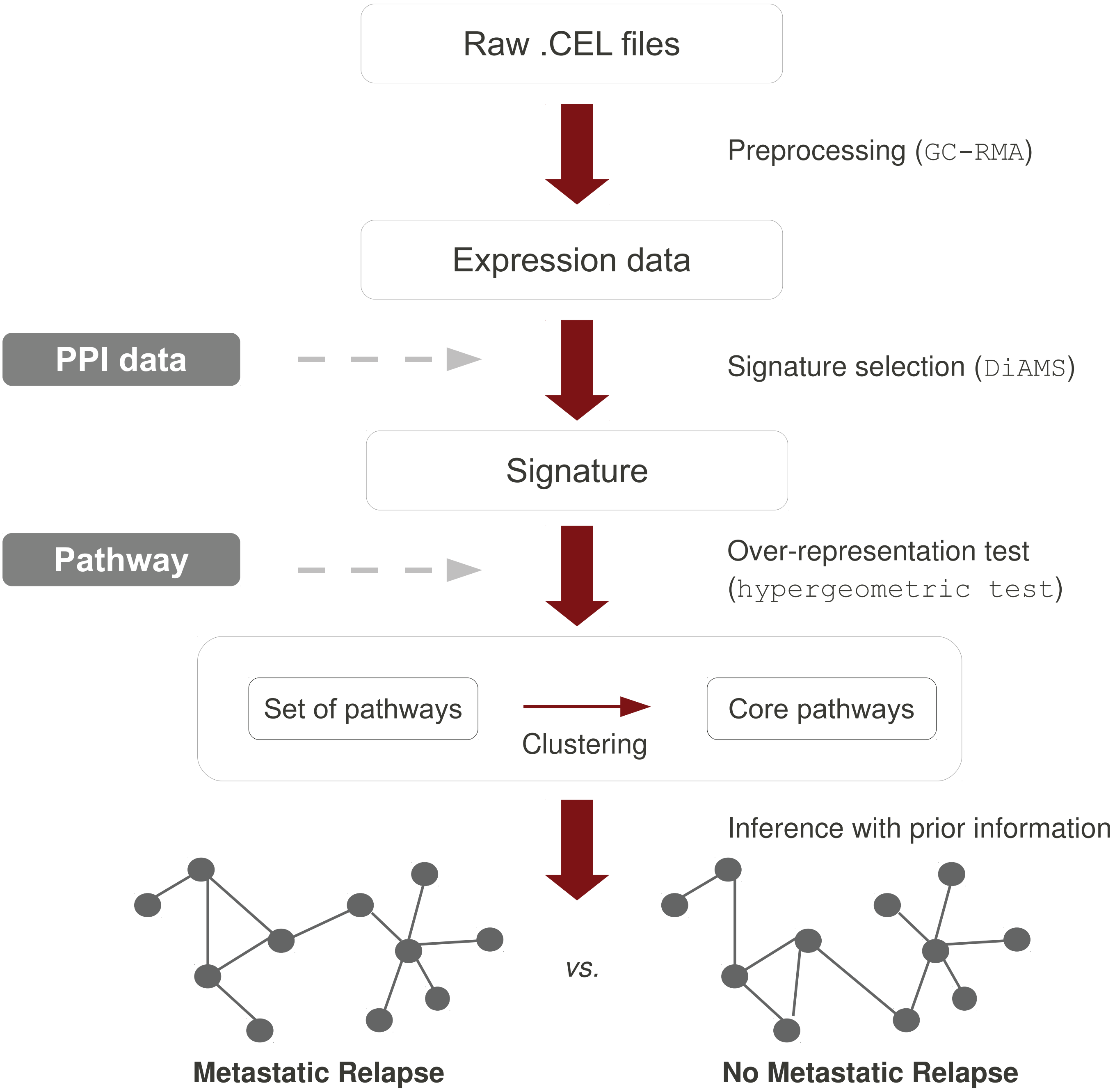


Figure 5.1 – *Analysis pipeline.*

5.1 Breast carcinomas data

5.1.1 Breast cancer: a brief overview

Cancer is a disease caused by the progressive and uncontrolled growth of the progeny of a single transformed cell. Breast cancer is a particular type of cancer originating from breast tissue. The great majority of breast cancers are *carcinomas*, *i.e.* they arise from the epithelial cells of the breast. Breast cancer is the most common cancer in women worldwide. In 2011, an estimated $53,000$ new cases of breast cancer were expected to be diagnosed in French women, along with $11,500$ deaths. However, breast cancer death rates have been going down over the past decade in France. This is definitely due to the growing understanding of the disease. Numerous risk factors have been identified as associated with the development of breast cancer, including genetic, environmental, hormonal, and nutritional influences as well as lifestyle choices. For instance, the literature shows that women drinking excess alcohol, obesity or giving birth to a first child at a late age have a higher risk of breast cancer. In addition to the high number of factors involved in carcinogenesis, breast tumors are also highly heterogeneous with many different clinical and pathological characteristics. This comprises of various subtypes defined by their amplification status of the epidermal growth factor receptor-2 gene (ERBB2) and the presence of hormone receptors. In this study we are particularly interested in Estrogen Receptor (ER) status which is an essential parameter of the pathological analysis of breast cancer. Estrogen is a female sex hormone that may stimulate the growth of cancer by triggering particular proteins (receptors) in the damaged cells. If breast cancer cells have Estrogen Receptors, the cancer is said to be ER positive (ER+) whereas if it has not, the tumor is said to be ER$^-$. These subtypes differ markedly in prognosis and in the repertoire of therapeutic targets they express. In this study we focus on ER$^-$ breast tumors that exhibit relatively homogeneous clinical and pathologic features compared to ER$^+$ tumors. In addition, ER$^-$ patients have a worse prognosis than ER$^+$ individuals. In particular, many more ER$^-$ cases will have relapsed early and develop metastatic recurrence. Thus, there is an urgent need to identify novel targets for the treatment of metastatic relapse in ER$^-$ breast cancer.

5.1.2 Transcriptome data

Microarray expression levels

Expression data were collected in the frame of the *Cartes d'Identité des Tumeurs* (CIT) program from the *Ligue Nationale contre le Cancer* and published by Guedj et al. (2011). All tumors were analyzed for expression profiling on Affymetrix U133 plus 2.0 chips and scanned with a Affymetrix GeneChip Scanner 3000. We downloaded the CEL files from the ArrayExpress website[1], from the following accession number: E-MTAB-365.
Gene expression levels are available for $54,675$ probesets in a set of 537 breast carcinomas. Among them, we are interested in the 91 ER$^-$ tumors. An annotation file is also available, which provides various features on

[1] http://www.ebi.ac.uk/arrayexpress/

samples, such as their labels, the biomaterial collection method, their grade, etc. We focus on the variable which denotes the metastatic relapse after 5 years. A total of 82 ER^- tumors are annotated for this variable.

Preprocessing

The raw data are normalized with GC-RMA, described in section 2.2.4, using the function justGCRMA from the R package gcrma. This provides gene expression measured in log base 2 scale. A filtering procedure is then performed to remove genes that exhibit little variation across samples and are not of interest. Finally a second filter is carried out in order to restrict our analysis to genes present in the Protein-Protein Interaction dataset describes in the next paragraph, which yields to an expression matrix containing $19,798$ genes and 82 tumors.

5.1.3 Protein-Protein Interaction data (PPI)

Due to the important efforts made to integrate and unify protein interaction information in public repositories, a number of PPI databases are currently available to the scientific community. About forty human PPI databases are listed in Pathguide (http://www.pathguide.org). They differ in terms of coverage, annotations, format and in the source of information they use. For instance, some of them collect only experimentally proven PPIs, while others integrate computational inference information. A comparison was conducted in Pharnext, as part of an internship project, in order to evaluate the various databases, based on coverage and topological criteria. In addition, data representation and updates were also considered. From this study it appears that combining the HPRD (Human Protein Reference Database) and a filtered version String (Search Tool for the Retrieval of Interacting Genes/Proteins) based on a 0.95 threshold applied to its confidence score, which is a good compromise.
Thus, we reconstruct from both databases a human PPI network. After removing PPIs which were duplicated and those containing proteins which are not mapped into a gene symbol, we select the largest connected component of the graph in order that all genes in the network are reachable by all other genes. We finally obtained a network comprising approximately $60,000$ interactions and more than $10,000$ proteins.

5.2 Metastatic relapse study

5.2.1 Signature selection

Description and properties

Once the data has been preprocessed, the challenge lies in identifying a relevant signature associated to the metastatic relapse. From both the expression and PPI data previously described, we aimed to identify modules of connected genes using DiAMS. The association of a gene to the metastatic relapse was scored using the limma p-value. The PPI network was converted into a tree structure using the walktrap algorithm available in the R package igraph. $20,439$ initial modules ($10,220$ modules of size 1, *i.e.* individual genes, and 10, 219 modules described by the tree structure) were tested for association to the metastatic relapse. Finally 19 modules were selected as significant at a 5% FDR level.

It is first interesting to note that the DiAMS signature includes genes that would not be selected individually by a classical approach, as their *p*-values are not significant at a 5% FDR level. As an example, the first module, detailed in Table 5.1, contained individually non-significant genes such as PSMB8 or TAP1. Including genes that play important roles in biological mechanisms, without showing a large differential expression, may be of great interest to define a more interpretable signature. In particular, genes that correspond to the disease phenotype, also called *driver* genes, are not necessarily highly differentially expressed but lead to a cascade of dysregulations of other genes. Missing such genes can seriously compromise the interpretation of the signature.

Second, we observe that DiAMS selects modules of varying sizes. In this application, sizes fall in the range of 2 to 17. In comparison to the approach of Chuang et al. (2007), DiAMS generates relatively large modules, that are less likely to be spurious.

Finally, a major advantage of our approach is that it directly provides information on molecular mechanisms through the extraction of PPI subnetworks. Compared to classical approaches which generate a list of individual genes, our approach facilitates the ease of interpretation of the resulting signature.

Gene symbol	EntrezGene ID	*p*-value
β2M	β2 microglobulin	0.00367
FCER1G	Fc fragment of IgE, high affinity I, receptor for γ polypeptide	0.0335
HLA-A	major histocompatibility complex, class I, A	0.0162
HLA-B	major histocompatibility complex, class I, B	0.0201
HLA-C	major histocompatibility complex, class I, C	0.0141
HLA-E	major histocompatibility complex, class I, E	0.0514
HLA-F	major histocompatibility complex, class I, F	0.0121
HLA-G	major histocompatibility complex, class I, G	0.0140
KIR2DL3	KIR, two domains, long cytoplasmic tail, 3	0.0798
KIR2DL2	KIR, two domains, long cytoplasmic tail, 2	0.0802
KIR2DL4	KIR, two domains, long cytoplasmic tail, 4	0.0231
KIR3DL1	KIR, three domains, long cytoplasmic tail, 1	0.00203
LILRB1	leukocyte IR, subfamily B, member 1	0.0292
LILRB2	leukocyte IR, subfamily B, member 2	0.0102
PSMB8	proteasome subunit, β type, 8	0.0743
TAP1	transporter 1, ATP-binding cassette, sub-family B	0.0873
TAP2	transporter 2, ATP-binding cassette, sub-family B	0.0368

Table 5.1 – ***Genes included in the most significant module identified by DiAMS.*** *Abbreviation - Ig: immunoglobulin; KC: killer cell; KIR: KC Ig-like receptor; IR: Ig-like receptor;*

Biological significance

We focus on the most significant module, described in Table 5.1 to illustrate the biological significance of the results obtained. It consists of 17 genes, whose 6 genes are related to the Major Histocompatibility Complex (MHC), also termed Human Leukocyte Antigen (HLA) in Humans. They code for cell surface proteins that are involved in the immune system. There are two major families of genes in the HLA complex, here the module highlights class-I HLA molecules that are expressed on the surface of all cell types. Another gene, called the $\beta 2$ microglobulin or $\beta 2$m is part of the HLA complex as illustrated on the Figure 5.2.

The HLA complex plays a pivotal role in immune responses against tumor cells. Indeed, recognition of breast cancer cells by lymphocytes requires presentation by HLA class-I molecules of peptides derived from tumor associated antigens, via the TAP genes (transporters associated with antigen processing) which are responsible for delivering these peptides to class-I molecules in the endoplasmic reticulum. HLA class-I genes are recognized by two distinct lymphocytes: natural killer (NK)

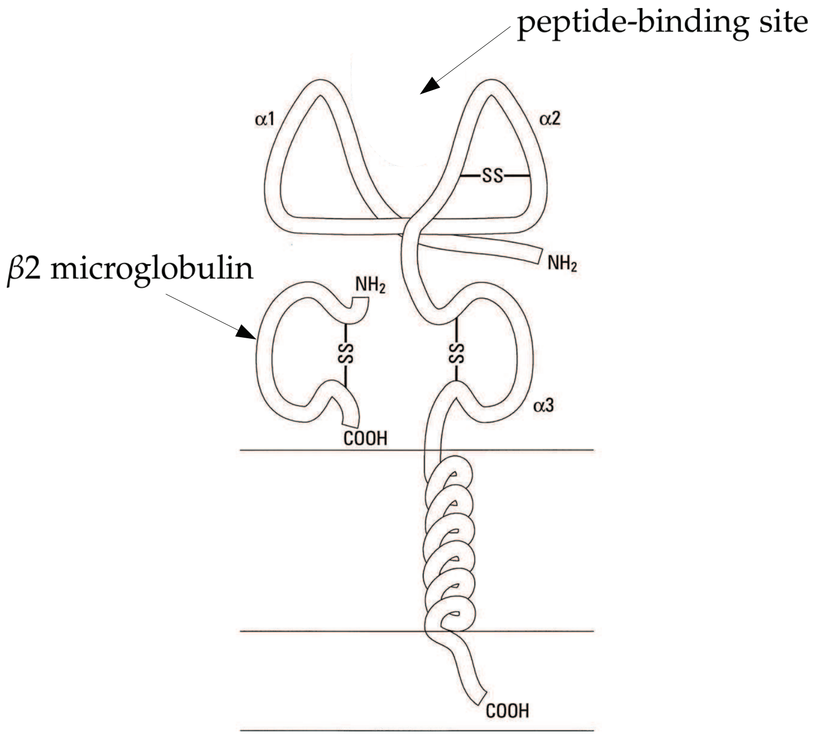


Figure 5.2 – ***HLA structure.***
This figure is inspired from Hoek et al. (1997), it displays the structure of the HLA class I complex that consists of two polypeptide chains, α and β. The α-chain is encoded in the HLA genes; the β-chain is termed β2m and is encoded on a separate chromosome. The two chains are linked non-covalently via interaction of β2m and the α3 domain. The molecular confirmation of the chains forms a groove in which the antigenic peptide is presented.

and cytotoxic T cells (CTL). NK cells are included as part of the innate immune system and provide a first line of defense by lysing tumor cells. They express killer cell immunoglobulin-like receptors (KIR), a family of HLA class I receptors, which regulate their killing functions. The other type of lymphocytes, the CTLs, are effectors of the adaptive immune responses. Once tumors' antigens are recognized, the CTLs are able to eliminate tumor cells by inducing a programmed cell death. Both the innate and adaptive immune responses are the two complementary arms of cell-mediated cytotoxicity that govern response to infection.

All the immune mechanisms previously described take place in a cell that is able to induce a immune response to cancer. However, in our data, all the genes belonging to the first module are down-regulated in MR patients. The down-regulation of HLA class-I expression has been extensively reported in studies of primary breast carcinomas, see for instance Cabrera et al. (1996) and suggests a role in preventing the mobilization of an adequate immune response. The involvement of the immune system in the host response to tumors has been a topic of intense research for more than a century. Today, it has been demonstrated that if the cellular or humoral effectors of the immune system do not recognize tumor antigens, the cancer may develop due to tumor cells escaping the host immune surveillance, or worse still, the tumor growth can actually be stimulated by inappropriate immune responses. Thus, the host immune system plays an essential role in cancer development and it is not surprising that is also involved in the metastatic relapse.

5.2.2 Pathway analysis

In order to provide a functional interpretation of the molecular signature, we conduct an over-representation analysis. This approach first requires a pre-defined set of pathways to analyze. In this study, the test was done using KEGG, the Kyoto Encyclopedia of Genes and Genomes developed by Kanehisa et al. (2006), and the BioCarta[2] database. The corresponding sets of pathways were downloaded on the Broad Insitute website[3], which gathers a collection of annotated gene sets in the Molecular Signatures Database (MSigDB). The over-representation test yields to the selection of thirty-one pathways significantly over-represented at a 5% level in the signature and summarized in three core pathways displayed in Figure 5.3. Most of the pathways identified are strongly associated to immune response and highlight some of mechanisms previously discussed: the NK cells pathway, the pathway of antigen presentation or even the CTL pathway.

We relate the first core pathway to *cell-to-cell signaling and interaction*. It includes the chemokine signaling pathway that mediate a wide range of trafficking function essential for the immune system, or mechanisms such as host recognition of pathogen-associated molecular patterns. The second core pathway is associated to *cell death and survival*. For instance,

[2] http://www.biocarta.com

[3] http://www.broadinstitute.org

it involves the pathway of the CTLs, key factors of the immune response that mediate the destruction of target cells by various mechanisms. Another major pathway is the one related to Interleukin 12 or IL-12, known to modulate the cytotoxic activity of NK cells and CTLs. Finally, the last core pathway is more difficult to characterize as it involves more heterogeneous pathways. We identify two main categories of pathways: (i) those associated with *cell signaling* such as the Cdc42 pathway that control diverse cellular functions and (ii) those related to the *regulation of the inflammatory response*. The association between inflammation, a physiologic process in response to tissue damage, and cancer was demonstrated by clinical studies. It is one of the first responses of the immune system to cancer and involves the recruitment of immunocompetent cells to inflammatory sites. In particular, the cooperative dialogue, or "crosstalk", of Dendritic cells (DCs) and NK cells play a critical role in early defenses against cancer and influences both innate and adaptive immune responses.

5.2.3 Comparison of regulatory networks

Once core pathways have been defined, they are used as a biological prior for driving the network inference of notMR and MR patients using the R package `SIMoNe`. The resulting networks are summarized in one single network, whose one subpart containing 51 nodes is displayed on Figure 5.4. 122 edges are inferred in both notMR and MR networks, while 6 regulation events take place only in the MR patients.

The gene RYR3 which encodes ryanodine receptor, is involved in two differential regulations between notMR and MR networks. In the MR network, an edge is inferred between the mitogen-activated protein kinase kinase kinase-13 (MAP3K13) and RYR3. MAP3K13 plays a role in the mechanisms of regulation of the Jun kinase, which enhances breast tumor cell growth, invasion and metastasis. In addition, in a recent study of Stephens et al. (2012), driver mutations were identified in MAP3K13. Most were protein truncating or non-synonymous mutations which can alter the signaling pathway that activate Jun kinases. Thus, a dysregulation of MAP3K13 may have dramatic consequences on tumor relapse and metastasis development. The second differential regulation of RYR3 involves the MMP16 genes which codes for the matrix metalloproteinase-16 enzyme. The family of matrix metalloproteinases (MMPs) is over-expressed in various human malignancies and Stetler-Stevenson et al. (1993) highlight their contribution to tumor invasion and metastasis.
The ADP-ribosyltransferase-1 (ART1) is also a key gene of this study. In MR patients, it interacts with the HAMP gene, which encodes the hepcidin protein, a central regulator of iron homeostasis recently identified by Pinnix et al. (2010) as a marker of metastasis-free survival of women after definitive primary treatment of their breast cancer. The MR patients also exhibit a differential regulation between ART1 and EID2, a gene which inhibits the differentiation of EP300. EP300 regulates transcription via chromatin remodeling and is important in the processes of cell proliferation and differentiation. It has also been identified as a co-activator of HIF1A (hypoxia-inducible factor 1 alpha), and thus plays a role in the

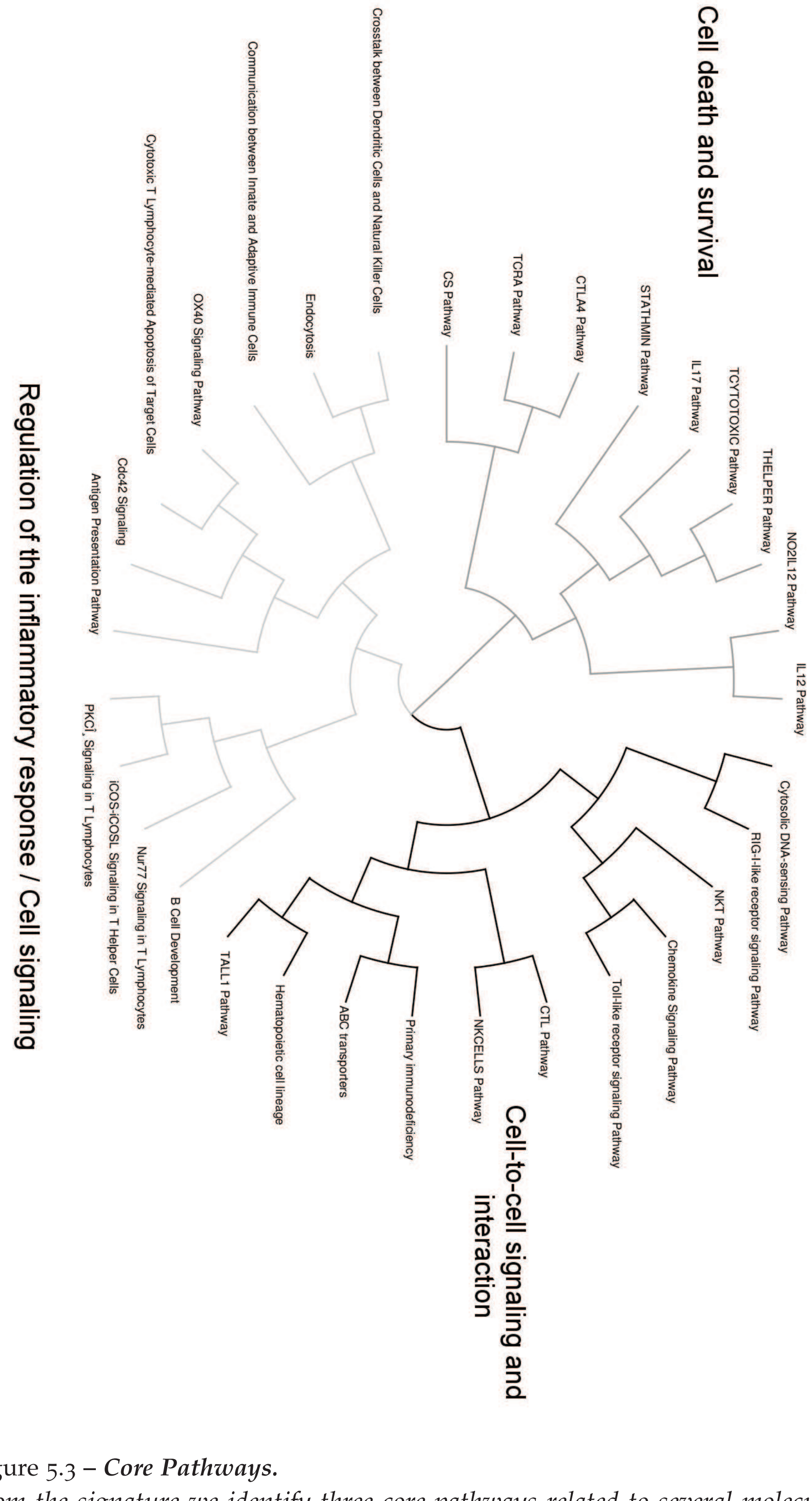

Cell death and survival
Crosstalk between Dendritic Cells and Natural Killer Cells
Communication between Innate and Adaptive Immune Cells
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells
OX40 Signaling Pathway
Endocytosis
CS Pathway
TCRA Pathway
CTLA4 Pathway
STATHMIN Pathway
IL17 Pathway
TCYTOTOXIC Pathway
THELPER Pathway
NO2IL12 Pathway
IL12 Pathway
Cytosolic DNA-sensing Pathway
RIG-I-like receptor signaling Pathway
NKT Pathway
Chemokine Signaling Pathway
Toll-like receptor signaling Pathway
CTL Pathway
NKCELLS Pathway
Primary immunodeficiency
ABC transporters
Hematopoietic cell lineage
TALL1 Pathway
B Cell Development
Nur77 Signaling in T Lymphocytes
iCOS-iCOSL Signaling in T Helper Cells
PKCî, Signaling in T Lymphocytes
Antigen Presentation Pathway
Cdc42 Signaling
Regulation of the inflammatory response / Cell signaling
Cell-to-cell signaling and interaction

Figure 5.3 – ***Core Pathways.***

From the signature we identify three core pathways related to several molecular and cellular functions. The core pathway associated with cell death and survival is displayed in sky gray. The one related to cell-to-cell signaling and interaction is colored in black. Finally, the last core pathway (in light gray) targets functions of regulation of the inflammatory response and cell signaling.

promotion of tumor growth through the stimulation of hypoxia-induced genes such as the vascular endothelial growth factor. Other dysregulations affect the killer cell immunoglobulin-like receptors, KIR2DL3 and chemokine motif ligand-4 (CCL4) as well as the chemokine receptor type-7 (CXCR7) and the retinoic acid-inducible gene-1 receptor encoded by the gene DDX58.

The investigation of the regulatory networks obtained under notMR and MR conditions aims to elucidate the mechanisms of metastatic relapse at a systems level. It enables structural changes to be highlighted in the MR network. Indeed, 6 novel regulations that are biologically reasonable were predicted in the latter network. To assess the relevance of our findings, it would be interested to experimentally validate the genes involved in such altered regulations through RT-PCR (Reverse transcription polymerase chain reaction).

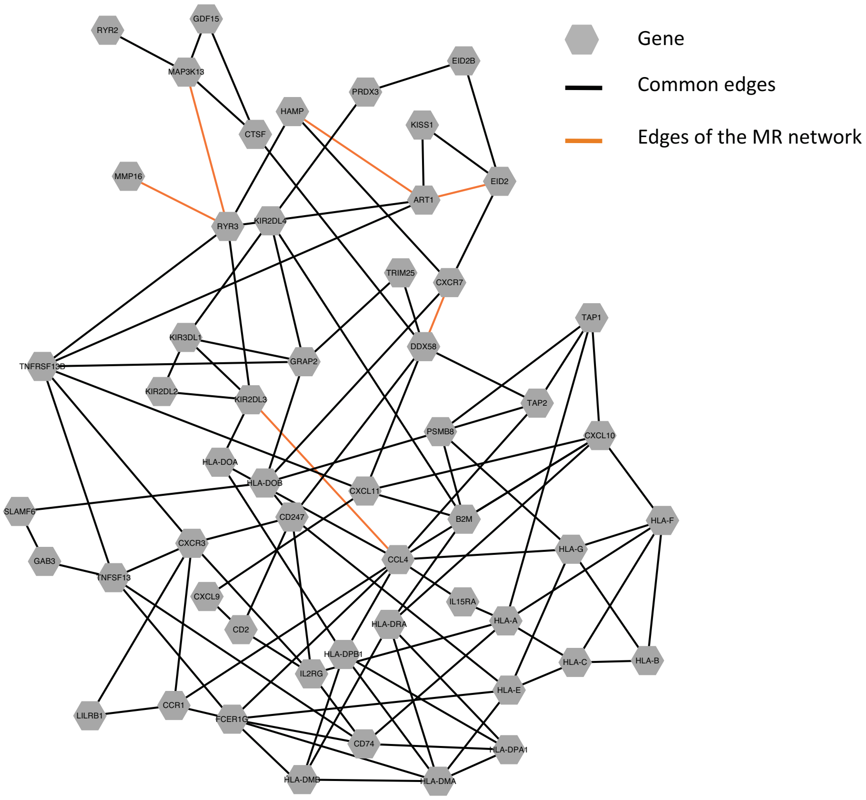


Figure 5.4 – ***Network associated with the metastatic relapse signature.***
The figure summarizes a subpart of notMR and MR networks. The black edges are common to both networks. In orange are the edges inferred only under the MR condition.

Chapter Conclusion

This study of metastatic relapse in ER$^-$ breast tumors represents a concrete application of the developments proposed in this thesis and gathered in an integrated pipeline. We demonstrated the relevance of signatures obtained by `DiAMS` and illustrate how regulatory network inference may help to highlight disrupted regulations under certain conditions.

In terms of biological interpretation, the signature we obtained is consistent with previous findings in the literature and involved immune response processes that are known to play a pivotal role in cancer development. In addition, we spotted genes involved in altered regulations in the network associated with the metastatic relapse status. This provides good insights in disease mechanisms that should be further investigated.

6 Collaborative projects

In this chapter we introduce collaborative works, in relation with our PhD research project.

In the first section, we present a study which was undertaken within the french StatOmique Consortium, in which we address the data preprocessing issues for RNA-seq technologies. In particular, we evaluate the performance of normalization approaches proposed in the literature by investigating the impact and the characteristics of each method on real datasets. In addition, a simulation study allows a further evaluation of the approaches. Such comparison studies are particularly important for RNA-seq technology which is a relatively new methodology, still under active development. Thus, it will take some time to define appropriate standard tools for analyzing resulting expression data.
A similar study is currently conducted on differential analysis methods dedicated to RNA-seq data. The various approaches found in the literature fall into two categories: (i) Poisson distribution-based and (ii) negative binomial distribution based approaches. We mention, in the second section, some preliminary results on the performances of each method, based on simulated data.
In the third section, we introduce a research project conducted as part of the Master's thesis of Gen Yang, which I co-supervised with Christophe Ambroise and Julien Chiquet. Gen Yang implements an extension of the Lasso, called the tree-lasso, which is a tree-structured sparse regularization approach dedicated to variable selection. During his internship in the *Statistique et Génome* laboratory, I worked with him on the validation and the application of the method to transcriptome data.

This chapter is associated with the following publication as a co-first author:

1. Dillies, Rau, Aubert, Hennequet-Antier, Jeanmougin, Servant, Keime, Marot, Castel, Estelle, Guernec, Jagla, Jouneau, Laloe, Le Gall, Schaeffer, Le Crom, Guedj, et Jaffrezic (2012)*. **A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.** *Briefings in Bioinformatics*. *This work has been done within the StatOmique Consortium.

6.1 Review and comparison of normalization approaches for RNA-seq data

In the present section, we focus on the issue of preprocessing for RNA-seq data. In particular, we provide a comprehensive comparison of normalization methods. This section is largely inspired by the paper from Dillies et al. (2012).

6.1.1 Normalization methods

Experience with microarray data has repeatedly shown that normalization is an essential step in the analysis of gene expression. An important advantage of RNA-seq is their ability to allow direct access to sequences of mRNA, avoiding biases due to hybridization and labeling. However, other sources of systematic variation have been reported: (i) between-sample differences such as library size: larger library sizes result in higher counts for the entire sample (ii) within-sample gene-specific effects related to gene length or GC-content. Thus, the basic problem is still the same: how to remove unwanted variations such that any differences in expression between samples are due solely to biological effects.

As this thesis focuses on the comparison of samples between various conditions, we only discuss inter-normalization approaches in the following section. When using RNA-seq for assessing differential expression, read counts need to be properly normalized between sample to extract meaningful expression estimates. Because the most obvious source of variation between lanes is the differences in library size, the simplest form of inter-sample normalization is achieved by scaling raw read counts in each lane by a single lane-specific factor reflecting its library size. We consider five different methods for calculating these scaling factors and two other normalization strategies, all described below.

a - Scaling normalization

***Total counts* `(TC)`**: Gene counts are divided by the total number of mapped reads associated with their lane and multiplied by the mean total counts across all the samples of the dataset.

***Upper Quartile* `(UQ)`**: Very similar in principle to `TC`, the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors, as done in Bullard et al. (2010).

***Median* `(Med)`**: Also similar to `TC`, the total counts are replaced by the median counts different from 0 in the computation of the normalization factors.

***Differential Expression analysis for SEQuence count data* `(DESeq)`**: A `DESeq` scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its geometric mean across all lanes. The underlying idea is that non differentially expressed genes should have similar read counts across samples, leading to a ratio of 1. Assuming most genes are not differentially expressed, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis. This factor is computed for each lane, and raw read counts are divided by the factor associated with their sequencing lane. See Anders et Huber (2010) for more details.

***Trimmed Mean of M-values* `(TMM)`**: This normalization method was introduced by Robinson et Oshlack (2010). It is also based on the hypothesis that most genes are not differentially expressed. The `TMM` factor is computed for each lane, with one lane being considered as a reference sample and the others as test samples. For each test sample, `TMM` is computed as the weighted mean of log ratios between this test and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios. According to the hypothesis of low differential expression, this `TMM` should be close to 1. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (and not the raw counts) in order to fulfill the hypothesis. The `calcNormFactors()` function in the `edgeR` Bioconductor package provides these scaling factors. To obtain normalized read counts, these normalization factors are re-scaled by the mean of the normalized library sizes. Normalized read counts are obtained by dividing raw read counts by these re-scaled normalization factors.

b - Other normalization strategies

***Quantile* `(Q)`**: In analogy with the quantile approach for normalizing microarray data, this method consists of matching distributions of gene counts across lanes.

***Reads Per Kilobase per Million mapped reads* `(RPKM)`**: This approach was initially introduced by Mortazavi et al. (2008) to facilitate comparisons between genes within a sample and combines between- and within-sample normalization, as it re-scales gene counts to correct for differences in both library sizes and gene length. However, it has been shown that attempting to correct for differences in gene length in a differential analysis actually has the effect of introducing a bias in the per-gene variances, in particular for lowly expressed genes. Despite these findings, the `RPKM` method continues to be a popular choice in many practical applications.

6.1.2 Comparison on real datasets

As illustrated in the previous paragraph, a number of normalization approaches to treat RNA-seq data have emerged in the literature differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data accumulate, there is still no clear indication of how the normalization method chosen impacts the downstream analysis and in particular the differential analysis. To this end, we propose a systematic

comparison of the seven normalization approaches described above. The raw unnormalized data, denoted by Raw Counts (RC), are also added to the comparison process.
The comparison first relies on real datasets sequenced using a Illumina sequencing machine, involving different species (*H. sapiens* from Strub et al. (2011), *A. fumigatus* (unpublished data), and *E. histolytica* (Weber et *al.*, submitted)), summarized in Table 6.1. Both the qualitative characteristics of normalized data and the impact of the normalization method on the results from a differential expression analysis.

Organism	Abbr.	♯ of genes	Rep. per cond.	Min LS	Max LS
H. sapiens	Hs	26437	3	2.0×10^7	2.8×10^7
A. fumigatus	Af	9248	2	8.6×10^6	2.9×10^7
E. histolytica	Eh	5277	3	2.1×10^7	3.3×10^7

Table 6.1 – ***Real datasets.***
Summary of datasets used for comparison of normalization methods, including the organism, the abbreviation used to design the dataset (Abbr.), number of genes, number of replicates per condition (Rep. per cond.), minimum and maximum library sizes (LS).

First, each normalization method is applied to the RNA-seq reference dataset. The Figure 6.1 displays the effect of normalization on read counts distribution for each sample of the Hs dataset. An effective normalization should result in a stabilization of read counts across samples. We note that most of the methods yield comparable results, except for RPKM and TC that do not improve over the raw counts in term of stabilization. Both methods scale the library sizes using the total number of reads. Thus, they both are very sensitive to high count genes.

In addition, the within-condition variability measure is assessed for all datasets, based on the coefficient of variation per gene. Here, we show the results for the reference dataset only but we observe similar patterns in other datasets. The boxplots in Figure 6.2 represents the distribution of this coefficient across samples for the two conditions of the Eh dataset. Little difference is observed among the normalization methods.

An investigation was then carried out on the average variation of a set of 30 housekeeping genes in the human data, assuming that these genes are similarly expressed across samples. The housekeeping genes were selected from a list previously described in Eisenberg et Levanon (2003) and presented the least variation across the 84 human cell types of the GeneAtlas data from Su et al. (2004) available on GEO [1] with the accession number GSE1133. Considering that these genes are assumed to have relatively constant expression, Figure 6.3 highlights that DESeq and TMM normalization methods lead to smallest coefficient of variation.

[1] GeneExpressionOmnibus:http://www.ncbi.nlm.nih.gov/geo

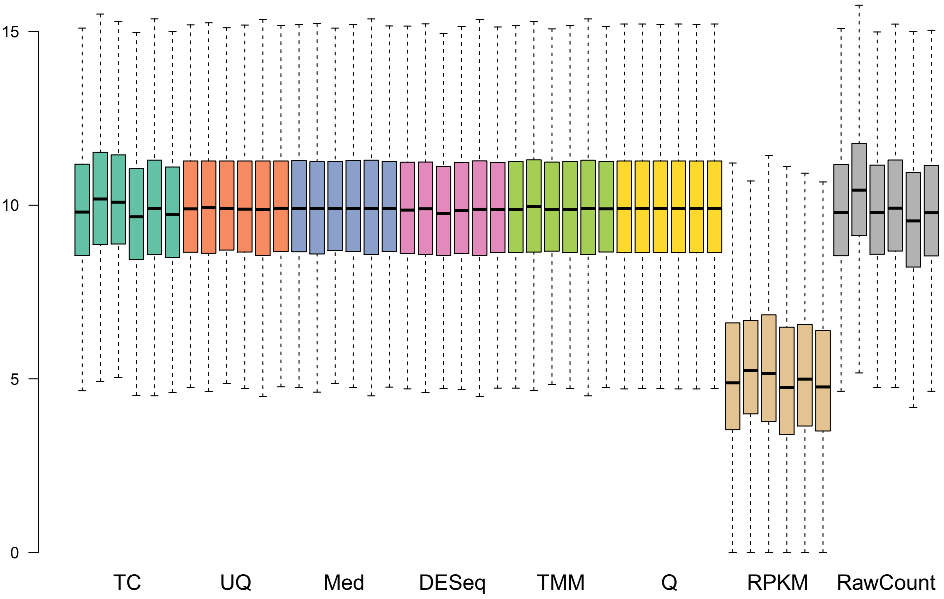


Figure 6.1 – ***Effects of normalization on E. histolytica data.***
Boxplots of log(counts + 1) for all samples in the E. histolytica data, by normalization method.

Finally, the seven normalization methods were compared based on results from a differential analysis performed with the Bioconductor package `DESeq` and the `TPSM` method. For each real dataset, we generated a dendrogram representing the similarity between the lists of differentially expressed genes obtained with each normalization method, based on the binary distance and the Ward linkage algorithm. The three dendrograms are subsequently merged into a consensus dendrogram, displayed on Figure 6.4. It results from the mean of the distance matrices obtained from each real dataset using `DESeq`. The advantage of such an approach is that it allows us to determine which methods perform similarly. The consensus dendrogram illustrates a trend, namely that in the results from a differential analysis, the `TC` normalization tends to group with `RPKM` and the unnormalized raw counts, while the remaining methods tend to group together. We note that the consensus dendrogram tree constructed using results from the `TSPM` (data not shown) is nearly identical to that constructed from the `DESeq` results, suggesting that the relationships identified among the normalization methods are not simply linked to the model used for the differential analysis.

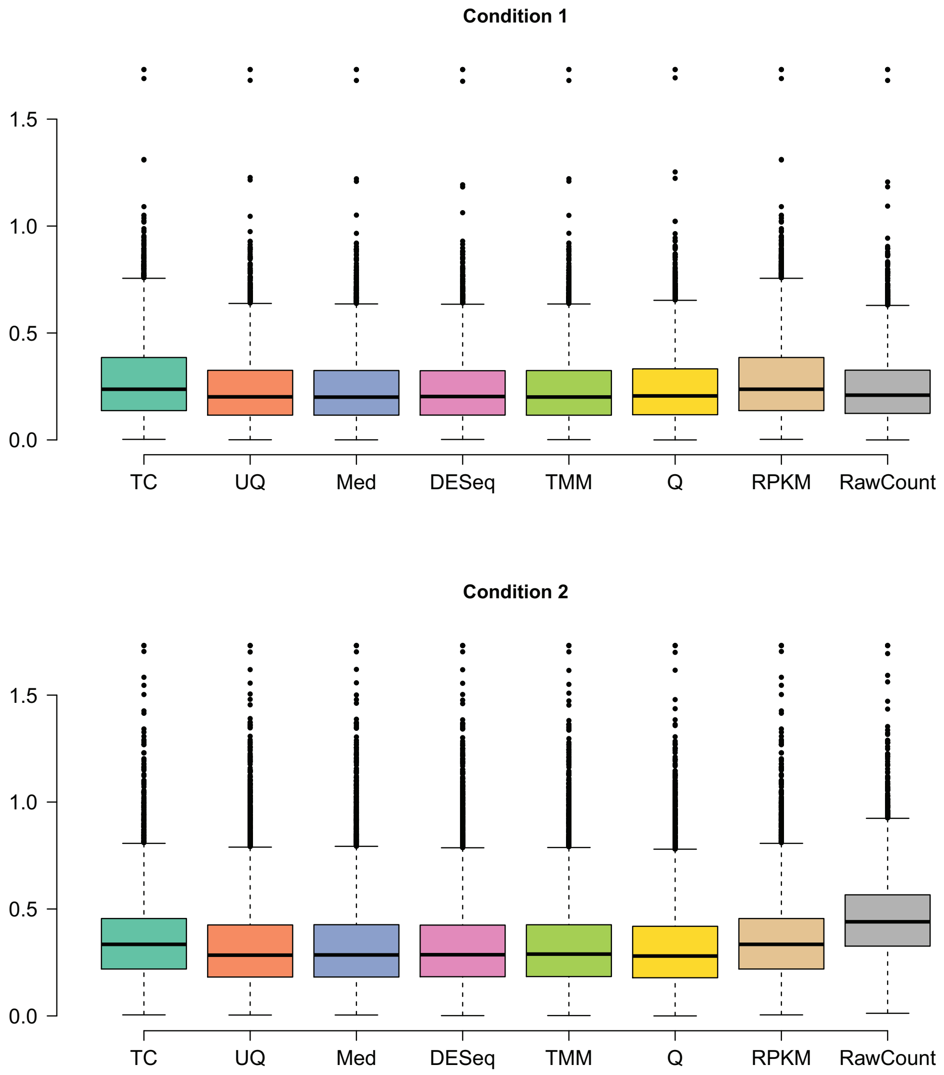


Figure 6.2 – ***Intra-group variance.***
Boxplots of intra-group variance for the 2 conditions of the Eh dataset.

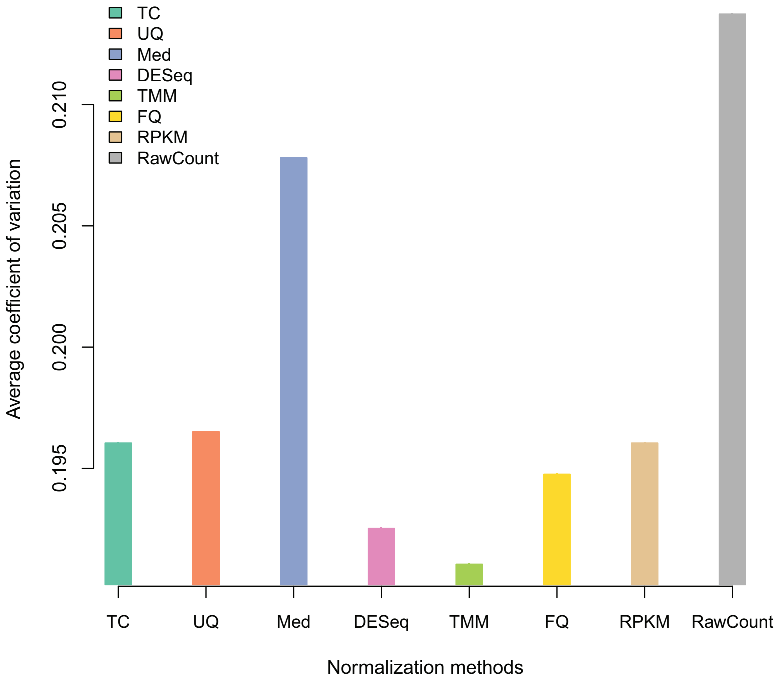


Figure 6.3 – ***Housekeeping genes studies.***
Average coefficient of variation of 30 known housekeeping genes in the Hs data.

6.1.3 Evaluation on simulated data

We used simulated data to assess the impact of normalization on differential analysis. The simulation model is similar to one we used in Jeanmougin et al. (2010) and adapted to counts. We assume $X_{ig}^{(c)}$, the expression level of gene g in sample i, follows a Poisson distribution of parameter $\lambda_g^{(c)}$ according to the condition c to which belongs sample i:

$$\left\{ \begin{array}{lcl} X_{ig}^{(1)} & \sim & \mathcal{P}(\lambda_g^{(1)}) \\ X_{ig}^{(2)} & \sim & \mathcal{P}(\lambda_g^{(2)}) \end{array} \right.$$

Under H_0: $\lambda_g^{(1)} = \lambda_g^{(2)}$ while under H_1: $\lambda_g^{(2)} = (1+\tau)\lambda_g^{(1)}$, with $\tau = 0.2$. Data are simulated with $p = 15,000$ and $n = 20$ (ten samples per condition) and the proportion of genes simulated on H_1 increasing from 0% to 30%. To assess the impact of non-equivalent library sizes and high count genes, various simulation models, described in Table 6.2, are considered. The parameter $\lambda_g^{(1)}$ is estimated from the datasets as the mean expression for each gene. For each simulated dataset, the false-positive rate and power are computed, based on the genes simulated under H_0 and H_1 respectively.

In situations where library sizes are simulated to be equivalent and no high count genes are present, all normalization methods considered perform nearly identically to the unnormalized raw counts in terms of

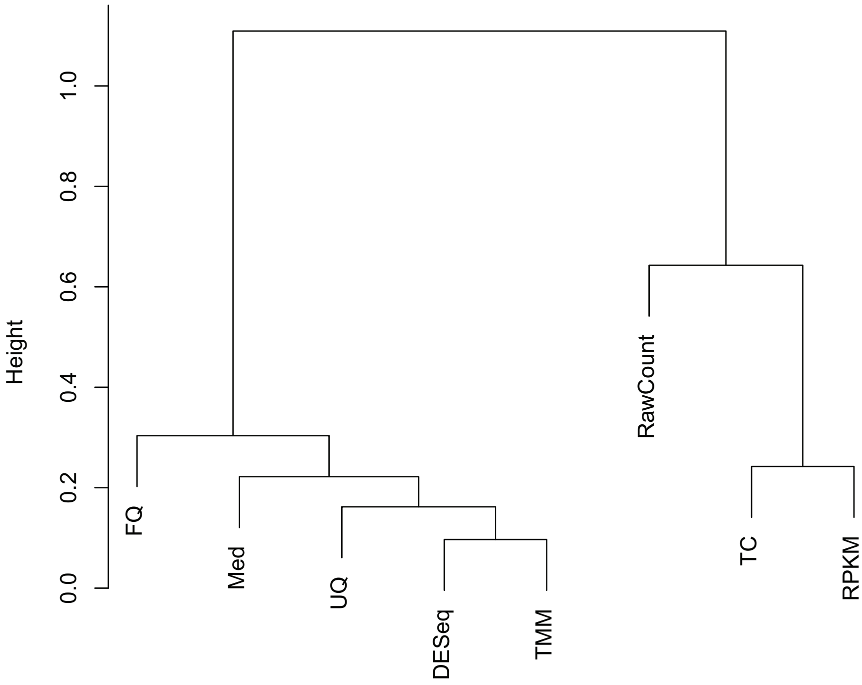


Figure 6.4 – ***Study of gene lists resulting from the differential analysis.***
Consensus dendrogram of differential analysis results, using the `DESeq` *Bioconductor package, for all normalization methods across the three datasets under consideration.*

	Library size		**High count genes**	
	Equiv.	Non equiv.	Yes	No
M1	X			X
M2		X		X
M3	X		X	

Table 6.2 – ***Simulation plan.***
Three simulation models are considered according to the equivalence or not of library sizes between samples and the presence or absence of high count genes in the dataset.

the false-positive rate and power. In situations where library sizes are different (see Figure 6.5-A), we note that the nominal false-positive rate is not maintained and the power is significantly decreased for the unnormalized data. All of the normalization methods are able to correct for these differences in library sizes, as they all control the false-positive rate and maintain a reasonable power. Figure 6.5-B presents results from the most discriminant simulation setting, where the library sizes are simulated to be equivalent for all samples with the presence of a few high count genes (model M3). This setting indicates that contrary to the situation with varying library sizes, the presence of high count genes does not impact the performance of raw counts; this seemingly contradictory result is due to the fact that the data are simulated under the model used for the differential analysis. However, the presence of these high count genes clearly results in an inflated false-positive rate for five of the normalization methods (`TC`, `UQ`, `Med`, `Q` and `RPKM`). Only `DESeq` and `TMM` are able to control the false-positive rate while also maintaining the power to detect differentially expressed genes.

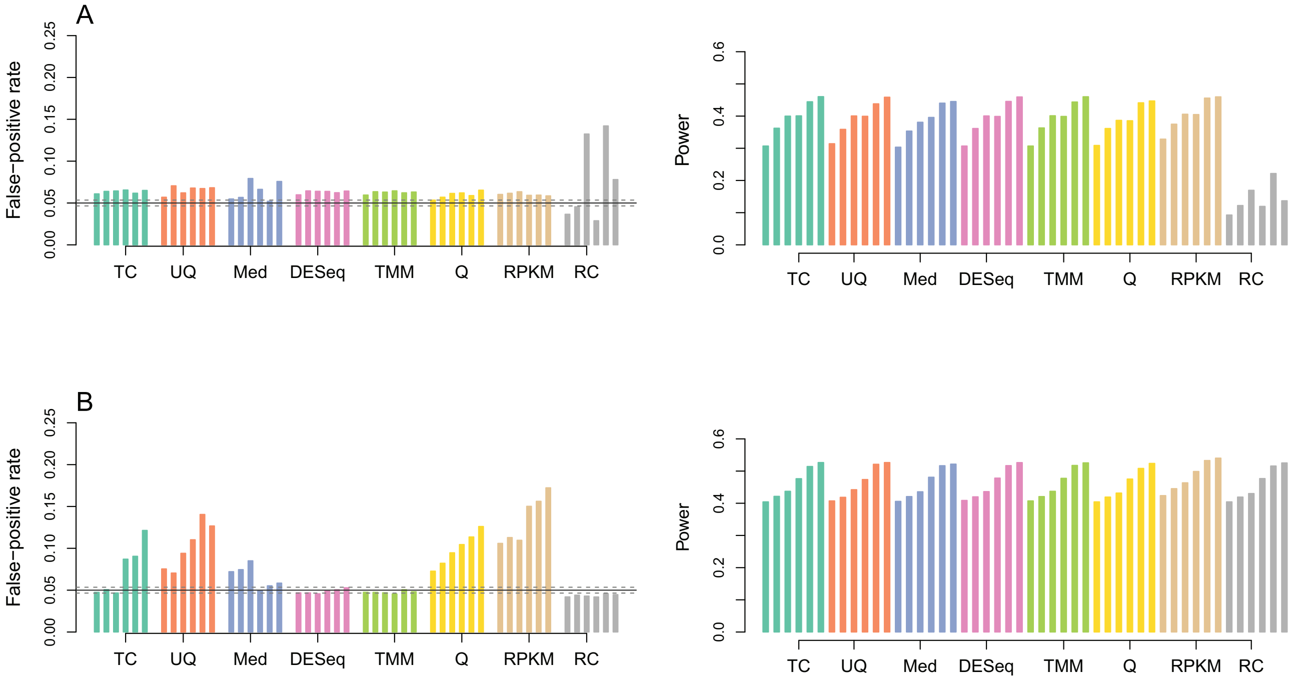


Figure 6.5 – ***False-positive rate and power for the simulation models (A) M2 and (B) M3.***

The false-positive rates and the power are averaged over 10 *independent datasets simulated with varying proportions of differentially expressed genes: (i) from* 0% *to* 30% *within each color grouping for false-positive values (ii) from* 5% *to* 30% *within each color grouping for power values.*

6.1.4 Discussion on normalization methods

Despite initial optimistic claims that RNA sequencing data do not require sophisticated normalization, in practice normalization remains an important issue since raw counts are often not directly comparable within and between samples. While this subject has received some attention in the literature, the increasing number of RNA-seq normalization methods makes it challenging for scientists to decide which method to use for their data analysis. Given the fact that the choice of normalization has a great influence on the subsequent statistical analyses, the quality and credibility of these methods need to be assessed fairly. To this end, our comparison study deals with seven representative normalization strategies compared on three real datasets involving different experimental designs, and on simulated datasets representing various scenarios. Based on our three real mRNA sequencing datasets, we confirm previous observations that `RPKM` and `TC`, both of which are still widely in use, are ineffective and should be definitively abandoned in the context of differential analysis. Similarly, the quantile normalization is based on the strong assumption that all samples must have identical read count distributions. As shown in our comparison, this may lead to increased within-condition variability and should be avoided. The other normalization methods (`UQ`, `Med`, `DESeq` and `TMM`) perform similarly on the varied datasets considered here, both in terms of the qualitative characteristics of the normalized data and the results of differential expression analyses. Simulations allow a further discrimination of the seven methods, in particular in the presence of high count genes, where it appears that only `DESeq` and `TMM` are able to maintain a reasonable false-positive rate without any loss of power. These two methods do not explicitly include an adjustment of count distributions across samples, allowing samples to exhibit differences in library composition. It is not surprising, then, that these two methods performed much better than the others for data with differences in library composition.

6.2 Introduction to differential analysis for RNA-seq data

Due to the short history of RNA-seq there is no clear ′gold standard′ for detecting differentially expressed genes. Several approaches have been published and it is expected that more will appear. Current statistical methods model the count data from RNA-seq experiments using Poisson or negative binomial (NB) distributions.

6.2.1 Poisson distribution-based strategies

Various strategies detect differentially expressed genes using a Poisson model. For instance, Wang et al. (2010) approach, called `DEGSeq`, relies on a Fisher's exact test or a likelihood ratio test to identify differentially expressed genes. In `GPseq`, Srivastava et Chen (2010) employ a generalized Poisson distribution to model the position-level read counts and assess the differential expression using a likelihood ratio test. However, one limitation with the Poisson distribution is that the variance of a read is often much greater than the mean value. Thus, RNA-seq data may exhibit more variability than what is consistent with Poisson distribution. In consequence, Poisson based methods lead to high false positive rates. This phenomenon is called over-dispersion. In this kind of situation, Poisson distribution with over-dispersed variances appear as an appropriate solution. Recently, Auer et Doerge (2011) introduce the two-stage Poisson model or `TSPM` to overcome classical Poisson based method limitations. The first stage of `TSPM` involves testing for over-dispersion for each gene. Two strategies are subsequently used according the results of the first step: (i) a quasi-Poisson likelihood approach is applied for genes displaying evidence of over-dispersion (ii) otherwise a likelihood ratio test is performed.

6.2.2 NB distribution-based strategies

Another way to deal with over-dispersed data is to assume a NB model since the NB distribution has a variance which is always larger than the mean. Hardcastle et Kelly (2010) propose in the `baySeq` package two distributions for the data: a Poisson or a NB. The NB model is recommended by the authors as it provides better fit for most RNA-seq data. An empirical Bayesian analysis is employed to identify differentially expressed genes by ranking them according to the estimates of the posterior probabilities. In the `DESeq` method, Anders et Huber (2010) use the same strategy as those proposed for microarray analysis, which entails borrowing information across genes in order to better estimate the dispersion parameter. A locally linear relationship between variance and the mean expression levels is assumed to estimate the over-dispersion parameter for genes with similar expression profiles using pooled data. The same strategy is used in the `edgeR` package from Robinson et Oshlack (2010), where the estimates are moderated towards a common dispersion. An empirical Bayes rule is employed to determine the moderation.

6.2.3 Evaluation of differential analysis methods: preliminary results

In analogy to the comparison we performed for statistical tests dedicated to the analysis of microarray data, we are currently conducting a study to evaluate the performances of statistical test methods designed for RNA-seq data with the StatOmique Consortium. In addition to the methods mentioned above, we include `limma`, the `Wilcoxon` statistic, `SAMseq` form Fahlgren et al. (2009) which is a modified version of `SAM` dedicated to sequencing data as well as `PoissonSeq` from Li et al. (2012) and noiseq developed by Tarazona et al. (2011). Preliminary results on simulated data are presented Figure 6.6. They suggest that `TSPM` and `SAMseq` provide better results in term of False-positive rate, power and area under the ROC curve (AUC).

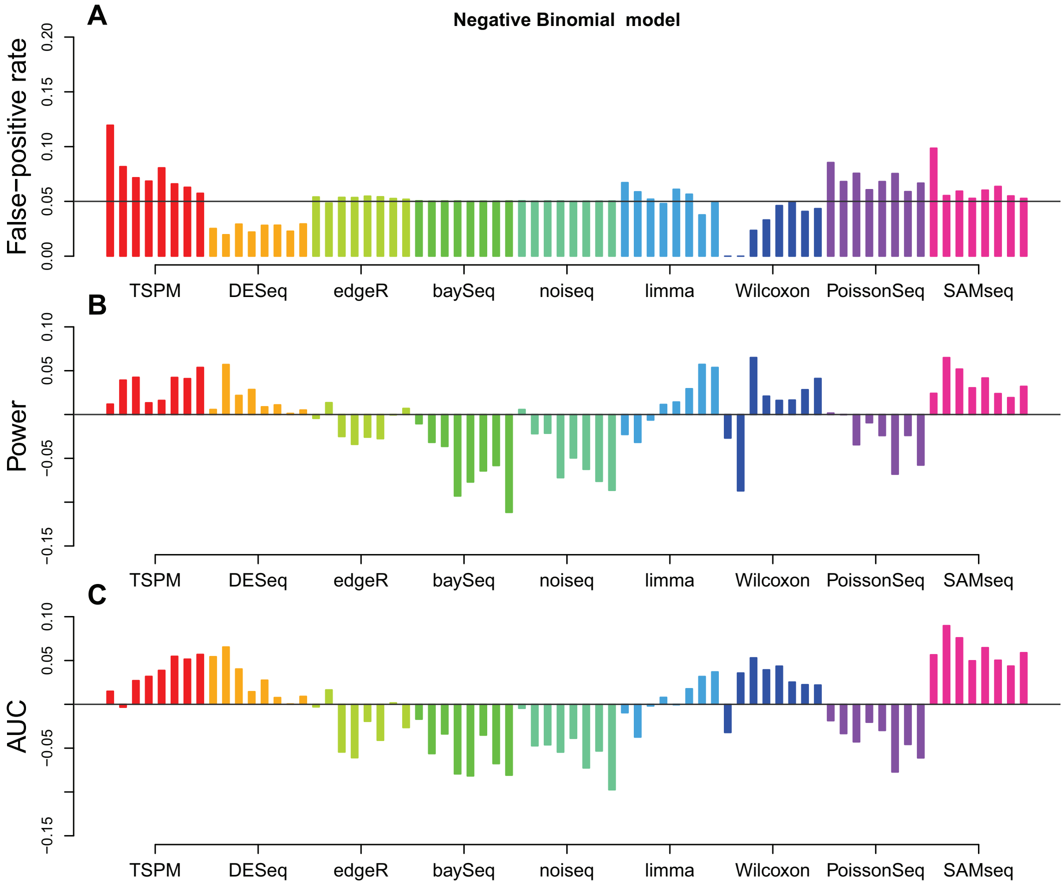


Figure 6.6 – ***Comparison of statistical tests for RNA-seq data.***
Count data were simulated under a negative binomial model for 2 *conditions. The false-positive rate and the power were evaluated at a* 5% *level for various sample sizes,* $n = \{2, 3, 4, 5, 6, 8, 10, 12\}$*. A ROC curve was then constructed by computing the sensitivity and specificity of increasing significance levels. The overall accuracy of tests is evaluated by calculating the area under the ROC curve (AUC). Both the power and the AUC are represented relative to the mean value over all tests.*

6.3 Tree-Lasso: sparse estimation for tree structures

In this section, we summarize a research project conducted during the three-month internship of Gen Yang, which I co-supervised with Christophe Ambroise and Julien Chiquet. In this project, we were interested in the problem of learning a sparse regression, where the structure in the variables can be represented as a tree. The regularization method which was studied is an extension of the usual ℓ_1 and the group-Lasso penalty, by allowing the groups of variables to overlap. The algorithms developed by Gen Yang were applied to transcriptome data, in order to select relevant molecular signatures.

6.3.1 Tree-Lasso model

Given a response vector $Y \in \mathbb{R}^n$ and a matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables, we consider the usual linear regression model:

$$Y = X\beta^* + \varepsilon.$$

We assume the data to be centered and consider the model without an intercept. ε is a zero-mean Gaussian error variable with variance σ^2 and $\beta^* = (\beta_1^*, ..., \beta_p^*)^T \in \mathbb{R}^p$ is the unknown vector of parameters we wish to estimate in the sparse case. For this purpose, let us define the following penalized estimator:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\{||Y - X\beta||^2 + \text{pen}(\beta)\},$$

where pen : $\mathbb{R}^p \to \mathbb{R}$ is a penalty function. For instance the Lasso takes pen(β) to be the ℓ_1 norm. This classical setting treats each variable independently from the others, while they may exhibit possible relationship between them. In this study, we are interested in the case where the variables can be organized in a tree structure, denoted $\mathcal{H}$, of depth d. Let $H_i = \{H_1^i, ..., H_{n_i}^i\}$ be the set of nodes corresponding to depth i. Figure 6.7 shows an example of tree structure. Each leaf is associated with a single variable, while internal nodes represent group of variables.

Given the tree structure, the penalty has the following form:

$$\text{pen}(\beta) = \sum_{i=0}^{d} \sum_{j=1}^{n_i} \lambda ||\beta_{H_j^i}||_{\ell_2}.$$

The regularization term is based on a group-Lasso penalty, where groups are defined with respect to the tree structure. In other words, this norm introduces a sparse selection of groups through the ℓ_1 penalization at each depth of the tree. An ℓ_2 penalty is then applied to all group members. The tree-Lasso can be seen as a special case of the overlapping group-Lasso, where groups are determined according to the tree structure. Gen implemented a proximal gradient method for solving the group-Lasso and the tree-Lasso. His work is based on the class of iterative shrinkage-thresholding algorithms (ISTA) initially proposed in De Mol et al. (2009).

6.3.2 Validation and application

With a focus on prediction assessment, we compare the Lasso and tree-Lasso in terms of error prediction. The simulations performed do not reveal significant differences between both approaches. But, as expected they better control the variance in the estimation than the Ordinary Least Squares. We assess the stability of the signature obtained with the tree-Lasso by evaluating its reproducibility. The process employed is similar to that described in section 3.2.3: (i) we identify a signature of reference in an expression dataset containing approximately 400 samples, (ii) the original expression matrix is then randomly subsampled and (iii) we estimate a

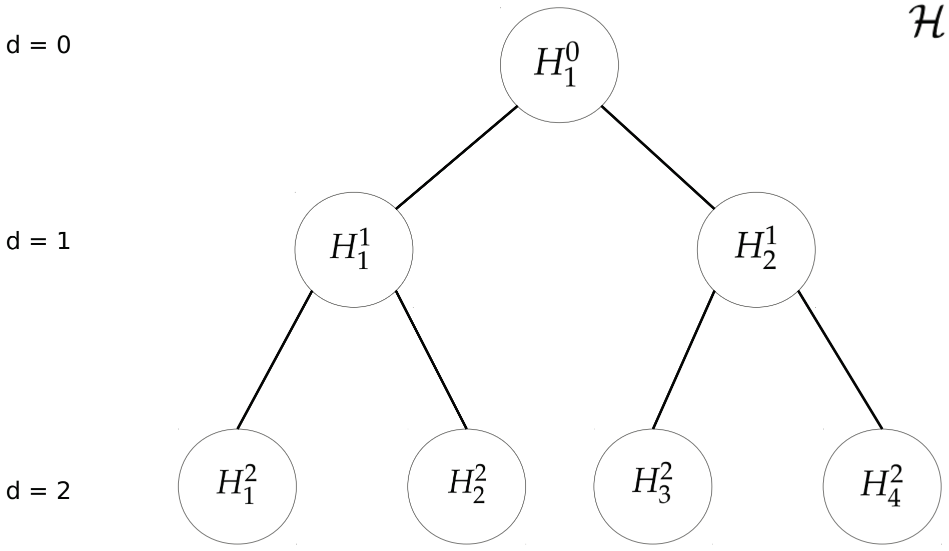


Figure 6.7 – ***Tree structured set of variables.***
Let $\mathcal{H}$ be a tree and $S = \{1, 2, 3, 4\}$ a set of variables. The leafs of the tree correspond to individual variables of S, and each internal node represents a cluster of variables. In particular, at the root, i.e. for depth $d = 0$, $H_1^0 = S$.

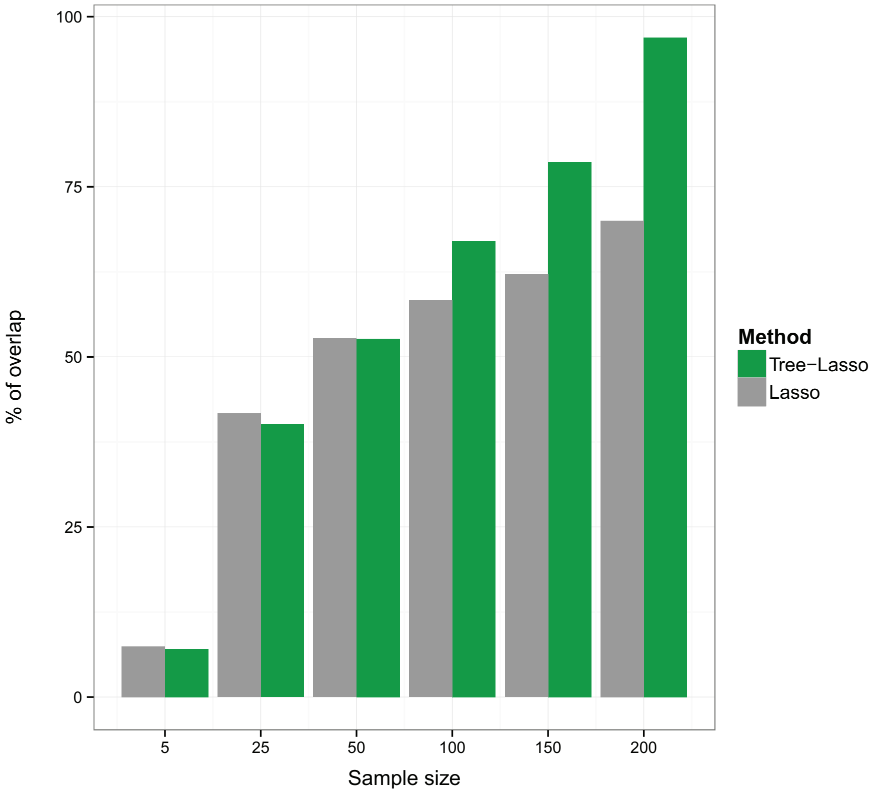


Figure 6.8 – ***Tree-lasso: reproducibility study.***
Results of the reproducibility analysis for the Lasso (in grey) and the tree-Lasso (in green). The mean of the overlap between a signature of reference and the signatures obtained from subsampled expression matrices was computed over $1,000$ *simulations.*

signature from the subsampled dataset. The reproducibility is calculated as the overlap between the reference signature and the signature of subsampled expression matrices. This procedure is performed for various subsample sizes and the results are displayed on Figure 6.8. We found that the tree-Lasso provides significantly better results than the classical Lasso for medium sample sizes. For very large samples, the stability results are comparable to those obtained with `DiAMS`. However, both the Lasso and the tree-Lasso exhibit a poor stability for small samples.

A study on real data was then conducted to assess the biological relevance of the signature obtained from the tree-Lasso approach. We aim to identify a gene expression signature of Estrogen Receptor status (ER^+ vs ER^-) in breast cancer, from the dataset provided in Guedj et al. (2011) and described in section 5.1.2. The tree structure was learned from the gene expression data by using the correlation between expression profiles as a similarity measure between genes. The `hclust` algorithm was then used to perform a hierarchical cluster analysis. 159 genes were selected by the tree-Lasso. The most significant genes of the signature include ESR1, a gene which encodes an Estrogen Receptor and IGF1R, which has been demonstrated to be a potent mitogens for ER^+ breast cancer cell lines.

KEGG pathways are tested for over-representation in the signature as described in 4.3.2. It yields to the identification of 29 significant pathways, which can be summarized in three mechanisms: (i) cellular communication (ii) cellular adhesion and (iii) signaling of ERBB proteins. The signature obtained from the classical Lasso approach does not over-represented pathways related to ERBB proteins signaling. However it is known that estrogen plays an important role in ERBB2-mediated signaling, which highlights the relevance of tree-Lasso results.

It should be noted that the tree structure of genes can either be estimated from the data using various clustering algorithms or be available as prior knowledge. For instance, it could be relevant to derive it from a PPI network as proposed in DiAMS. A recent study from Kim et Xing (2012) implements a tree-Lasso method with an application to expression quantitative trait locus mapping to identify SNPs with pleiotropic effects.

Chapter Conclusion

In this chapter, we discussed the collaborative works we conducted during our thesis. We detail the study of normalization approaches of RNA-seq data undertaken within the french StatOmique Consortium. A similar work on differential analysis methods dedicated to RNA-seq is currently being conducted. Thanks to our experience of comparison studies, we collaborate actively in these projects. We then introduced the work of Gen Yang, which we co-supervised during its three-month internship. Gen Yang implemented and assessed the performances of the tree-Lasso, for estimating tree structured sparsity. In addition to this projects, we collaborate with the GenHotel[2] laboratory to identify a molecular signature associated to the rheumatoid arthritis from blood RNA expression profiling. For this purpose various statistical analysis was performed, including normalization of technical biases, differential analysis or classification. This PhD also offered the opportunity to collaborate with researchers from other fields. In particular, we work with Carène Rizzon on the influence of the duplicated gene neighborhood on human gene expression.

Finally, throughout this PhD, we collaborate to Pharnext research projects. For instance, we applied DiAMS to select a signature associated with the Alzheimer's disease and include our developments in the analysis pipeline.

[2] http://www.genhotel.com/

Conclusion 7

7.1 General conclusion

The investigation of high-throughput genetic and genomic data offers unprecedented opportunities for understanding cellular and disease mechanisms. In particular, this thesis focused on the study of global changes in transcript abundance, which has turned out to be a highly promising tool for identifying genes associated with certain pathologies. Expression microarrays first made possible the analysis of the transcriptome on a genome-wide scale. During the last decade, microarrays have been the subject of intense research and they are nowadays considered as an advance technology. The process of data production is now well mastered and the biases inherent to microarray array technology are, for most of them, well identified. Despite the many advantages microarrays offer, the technology suffers from some limitations. For instance, it exhibits a poor accuracy for transcript in low abundance and it is not able to detect splice variants or previously unmapped genes. For this reason, today, researchers are turning to deep sequencing which uses direct sequence-based detection to quantify gene expression. In contrast with microarray technology, RNA-seq experiments are free of background hybridization and allow a more accurate expression level determination. This PhD project was conducted during a key period for the transcriptomic research field as the recent technological advances have led to an ongoing replacement of the classical microarrays with the RNA-seq technology.

Microarray and RNA-seq data differ from data associated with classical transcriptome profiling technologies in a critical way, which has challenged statisticians to develop new analytical methods. While conventional statistics typically deal with many observations made on few parameters, high-throughput gene expression data deal with relatively few observations made on many thousands of variables. A growing body of tools has emerged to help tackle these applications, but there is still little consensus about which to choose. In addition, some methods, which are considered as standards, suffer from important limitations that may compromise the relevance of the biological findings. In this thesis, we put our focus on providing robust statistical approaches dedicated to the analysis of data from high-throughput transcriptome experiments. For this purpose, our contributions can be summarized in three points: (i) identify the most relevant approaches from the literature (ii) propose novel developments that overcome the limitations of existing methods (iii) ensure the interpretability and the reliability of the resulting biological insights. Each of these points will be discussed in the following paragraphs.

Extensive research has shown that choice of methods employed to correct for bias in the transcriptome measurements or to select molecular signature can have a substantial impact on the subsequent analysis. Thus, the first step towards a robust investigation of transcriptome data involves choosing the most appropriate approach. In this context, the need for systematic evaluations of the existing methods is huge. Our contribution in this field was to conduct comparison studies but also to provide standardized processes for assessing the performance of novel statistical approaches proposed in the literature. We applied the proposed framework to compare differential analysis strategies as well as normalization methods. These studies enable us to provide useful guidelines to the community and to define tools that should be used as standards. In addition, such evaluation processes constitute the starting point for further developments as they allow not only to review what is proposed in the literature but also to clearly identify the weaknesses of existing methods. In this thesis, particular emphasis has been placed on improving stability of variable selection methods in the high-dimensional setting. This includes both selection of relevant set of genes and estimation of the edge set in graphical modeling.

In the third and fourth chapters, we detailed novel approaches that share a common rational: the integration of biological prior. Firstly, we introduced `DiAMS` (Disease Associated Modules Selection), a network-based approach dedicated to gene selection with the eventual goal of overcoming the inherent instability of differential analysis strategies. Motivated by the observation that genes causing the same phenotype are likely to interact together, we therefore explored an approach for identifying modules of functionally related genes, rather than individual genes. `DiAMS` involves an iterative algorithm based on an extended version of the local-score statistic to extract a molecular signature by integration of Protein-Protein Interaction and transcriptome data. We demonstrated through simulations that `DiAMS` not only outperforms standard differential analysis strategies in term of power but also produces significantly more reproducible signatures. The second statistical method proposed as part of this research project is described in Chapter 4 and involves the development of a framework to infer gene regulatory networks on the basis of a biological informative prior over network structures. This tool, included in the R package `SIMoNe` developed by Chiquet et al. (2009), offers the possibility of exploring the molecular relationships between genes, leading to the identification of altered regulations associated with a phenotype of interest. We found that introducing prior knowledge to drive the inference provides gains in terms of robustness of the network estimation.

The success of a novel approach depends not only on its performance on simulated datasets but also on its results on real data. It should ensure the interpretability and relevance of the biological findings and be implemented in an easy-to-use software. For this purpose, we apply the statistical developments that we have introduced in this thesis, to study the metastatic relapse of Estrogen Receptor negative breast cancers. We demonstrated the relevance of signatures obtained by DiAMS and illustrated how regulatory network inference may help to highlight disrupted regulations in patients who experienced a relapse of cancer. In addition, the tools we proposed are currently implemented in R packages that will be made available to the community. We also include our developments in the analysis pipeline of Pharnext.

7.2 Perspectives

Several statistical methods have been developed and evaluated in this thesis to ensure highly reproducible analyses by the integration of data and knowledge from disparate sources, such as Protein-Protein Interactions or pathways. Given the functional interdependencies between all of the molecular components in a cell, a disease is rarely a consequence of an abnormality in a single gene, but is governed by an intricate combination of transcription factors (TF), miRNAs, splicing factors and other complex processes occurring at transcriptomic, proteomic or metabolomic levels. Thus, data integration is a key part of conducting biological investigations with modern platform technologies. In particular, it appears to be crucial when inferring gene networks. Due to the large number of variables being investigated in a limited amount of samples, microarray datasets alone are indeed not enough to infer accurate gene regulatory networks. In addition, in a Gaussian Graphical Model framework edges are defined conditional on all other genes present in the dataset, the relevance of the inferred network greatly depends on the inclusion of all potential covariates in the analysis. On-going work gives promising first results in improving the robustness of the estimation and the interpretability of resulting networks but further work needs to be done to integrate heterogeneous data which could potentially help in the elucidation of the genomic system's regulatory mechanisms. For instance, the combination of interactions map between TF and their DNA binding locations with expression data or the integration of knockout and over-expression experiments could be relevant in elucidating regulation events.

Although the combination of data from different sources could help in the understanding and modeling of cellular mechanisms, this raises several issues. Firstly, the number of variables measured for a single gene may grow dramatically. This could be the case when integrating the data coming from various 'omics' technologies or from medical imaging techniques, such as magnetic resonance imaging, which produce very large amounts of data. Thus, this will require more sophisticated statistical developments for dimensionality reduction. In addition, as pointed out several times in this manuscript, a wide range of databases, which exhibit various format or annotations, are currently available for each domain. For instance, we mentioned the case of Protein-Protein Interaction databases which differ in terms of coverage, source of information or even in the type of interactions they include. This makes the integration of heterogeneous data very challenging for scientists in other domains. In this context, the work of comparison and standardization of databases is crucial.

In conclusion, understanding the cellular mechanisms at a system level remains challenging. For this purpose the integration of heterogeneous data is obviously necessary for advanced biomedical research and for providing better understanding of diseases, with the ultimate goal of improving health care. But, further statistical developments are needed to be able to fully exploit and analyze the volume and complexity of data available.

Communications

Publications

2012 [1] Dillies M-A*, Rau* A., Aubert* J., Hennequet-Antier* C., Jeanmougin* M., Servant* N., Keime* C., Marot G., Castel D., Jordi E., Guernec G., Jagla B., Jouneau L., Laloë D., Le Gall C., Schaëffer B., Le Crom* S., Guedj* M., Jaffrézic* F. A Comprehensive Evaluation of Normalization Methods for High-Throughput RNA Sequencing Data Analysis", *Briefings in Bioinformatics* , to be published. *These authors have contributed equally to this work.

2011 [2] Jeanmougin M. and Guedj, M. and Ambroise, C. Defining a robust biological prior from Pathway Analysis to drive Network Inference., *J-SFdS* , Vol 152(2).

2010 [3] Jeanmougin M. and de Reynies, A. and Marisa, L. and Paccard, C. and Nuel, G. and Guedj, M. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies , *PLoS ONE* , Vol 5 (9).

Book chapters

2012 [4] Jeanmougin M., Charbonnier C., Guedj M., Chiquet J. Network Inference in Breast Cancer with Gaussian Graphical Models and extensions, in *Probabilistic graphical models for genetics* , Oxford University Press, to be edited.

2011 [5] Bouaziz M., Jeanmougin M., Guedj M. Multiple testing in large-scale genetic studies, *In "Data Production and Analysis in Population Genomics" (Bonin A, Pompanon F eds)* , Methods in Molecular Biology Series, Humana Press.

Seminars and communications

Conference

2012 [1] Jeanmougin M., Ambroise C. and Guedj M. Improving gene signatures by the identification of differentially expressed modules in molecular networks: a local-score approach., *JOBIM* , Rennes.

2011 [2] Jeanmougin M., Ambroise C. and Guedj M. Network inference using pathway analysis results as a prior information: application to the study of treatment response in basal breast cancer , *SMPGD'11* , Paris.

2010 [3] Jeanmougin M., Ambroise C. and Guedj M. A comprehensive analysis combining network inference and pathway analysis for treatment response in basal breast cancer, *Cancer Bioinformatics Workshop* , Cambridge.

2009 [4] Jeanmougin M., Nuel G. and Guedj M. Should we abandon the t-test : a statistical comparison of 8 differential gene expression tests, *SMPGD'09* , Paris.

Seminar

2012 Guest to IRMA, Strasbourg

INRA "Inférence de Réseaux", Paris

"Génome tumoral" thematic school, Maffliers

2011 Guest to Ligue Contre le Cancer, Paris

"Génome tumoral" thematic school, Maffliers

2009 Statistics for Systems Biology, Paris

Posters

2011 Dillies M-A, and StatOmique consortium. RNA-seq Data Analysis: Lost in Normalization?, *JOBIM* , Paris.

2010 Jeanmougin M., Ambroise C. and Guedj M. A comprehensive analysis combining network inference and pathway analysis for treatment response in basal breast cancer, *Cancer Bioinformatics Workshop* , Cambridge.

Should We Abandon the *t*-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies

Marine Jeanmougin[1,2,3,4]*, Aurelien de Reynies[1], Laetitia Marisa[1], Caroline Paccard[2], Gregory Nuel[3], Mickael Guedj[1,2]

1 Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France, **2** Department of Biostatistics, Pharnext, Paris, France, **3** Department of Applied Mathematics (MAP5) UMR CNRS 8145, Paris Descartes University, Paris, France, **4** Statistics and Genome Laboratory UMR CNRS 8071, University of Evry, Evry, France

Abstract

High-throughput post-genomic studies are now routinely and promisingly investigated in biological and biomedical research. The main statistical approach to select genes differentially expressed between two groups is to apply a *t*-test, which is subject of criticism in the literature. Numerous alternatives have been developed based on different and innovative variance modeling strategies. However, a critical issue is that selecting a different test usually leads to a different gene list. In this context and given the current tendency to apply the *t*-test, identifying the most efficient approach in practice remains crucial. To provide elements to answer, we conduct a comparison of eight tests representative of variance modeling strategies in gene expression data: Welch's *t*-test, ANOVA [1], Wilcoxon's test, SAM [2], RVM [3], limma [4], VarMixt [5] and SMVar [6]. Our comparison process relies on four steps (gene list analysis, simulations, spike-in data and re-sampling) to formulate comprehensive and robust conclusions about test performance, in terms of statistical power, false-positive rate, execution time and ease of use. Our results raise concerns about the ability of some methods to control the expected number of false positives at a desirable level. Besides, two tests (limma and VarMixt) show significant improvement compared to the *t*-test, in particular to deal with small sample sizes. In addition limma presents several practical advantages, so we advocate its application to analyze gene expression data.

Citation: Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, et al. (2010) Should We Abandon the *t*-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. PLoS ONE 5(9): e12336. doi:10.1371/journal.pone.0012336

Editor: Kerby Shedden, University of Michigan, United States of America

Received April 8, 2010; **Accepted** June 24, 2010; **Published** September 3, 2010

Funding: This work was funded by the Ligue Nationale Contre le Cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: marine.jeanmougin@genopole.cnrs.fr

Introduction

During the last decade, advances in Molecular Biology and substantial improvements in microarray technology have led biologists toward high-throughput genomic studies. In particular, the simultaneous measurement of the expression levels of tens of thousands of genes has become a mainstay of biological and biomedical research.

The use of microarrays to discover genes differentially expressed between two or more groups (patients *versus* controls for instance) has found many applications. These include the identification of disease biomarkers that may be important in the diagnosis of the different types and subtypes of diseases, with several implications in terms of prognostic and therapy [7,8].

A first approach to identify differentially expressed genes is known as the Fold-Change estimation (FC). It evaluates the average log-ratio between two groups and considers as differentially expressed all genes that differ by more than an arbitrary cut-off. So defined, FC lacks of a solid statistical footing [9]: it does not take the variance of the samples into account. This point is especially problematic since variability in gene expression measurements is partially gene-specific, even after the variance has been stabilized by data transformation [10,11].

Rather than applying a FC cutoff, one should prefer statistical tests: they standardize differential expression by considering their variance [9,12]. Furthermore, corresponding effect sizes, confidence intervals and *p*-values are essential information for the control of false-positives [13] and meta-analysis [14].

The *t*-test is certainly the most popular test and has been matter of discussion. Computing a *t*-statistic can be problematic because the variance estimates can be skewed by genes having a very low variance. These genes are associated to a large *t*-statistic and falsely selected as differentially expressed [2]. Another drawback comes from its application on small sample sizes which implies low statistical power [12]. Consequently, the efficacy of a *t*-test along with the importance of variance modeling have been seriously called into question [15]. It has led to the development of many innovative alternatives, with hope of improved variance estimation accuracy and power.

These alternatives appear very diverse at a first sight, but fall into few nested categories relying on both statistical and biological hypotheses: parametric or non-parametric modeling, frequentist or Bayesian framework, homoscedastic hypothesis (same variance between groups of samples) and gene-by-gene variance estimation. Further propositions come from the field of machine-learning for instance [16], but lie beyond the scope of our study.

Journal de la Société Française de Statistique
Vol. 152 No. 2 (2011)

Defining a robust biological prior from Pathway Analysis to drive Network Inference.

Titre: Construction d'un a priori biologique robuste à partir de l'analyse de voies métaboliques pour l'inférence de réseaux.

Marine Jeanmougin [1,2] , Mickael Guedj [2] and Christophe Ambroise [1]

Abstract:
Inferring genetic networks from gene expression data is one of the most challenging work in the post-genomic era, partly due to the vast space of possible networks and the relatively small amount of data available. In this field, Gaussian Graphical Model (GGM) provides a convenient framework for the discovery of biological networks.
In this paper, we propose an original approach for inferring gene regulation networks using a robust biological prior on their structure in order to limit the set of candidate networks.

Pathways, that represent biological knowledge on the regulatory networks, will be used as an informative prior knowledge to drive Network Inference. This approach is based on the selection of a relevant set of genes, called the "molecular signature", associated with a condition of interest (for instance, the genes involved in disease development). In this context, differential expression analysis is a well established strategy. However outcome signatures are often not consistent and show little overlap between studies. Thus, we will dedicate the first part of our work to the improvement of the standard process of biomarker identification to guarantee the robustness and reproducibility of the molecular signature.

Our approach enables to compare the networks inferred between two conditions of interest (for instance case and control networks) and help along the biological interpretation of results. Thus it allows to identify differential regulations that occur in these conditions. We illustrate the proposed approach by applying our method to a study of breast cancer's response to treatment.

Résumé :
L'inférence de réseaux génétiques à partir de données issues de biopuces est un des défis majeurs de l'ère post-génomique, en partie à cause du grand nombre de réseaux possibles et de la quantité relativement faible de données disponibles. Dans ce contexte, la théorie des modèles graphiques gaussiens est un outil efficace pour la reconstruction de réseaux.
A travers ce travail nous proposons une approche d'inférence de réseaux de régulation à partir d'un *a priori* biologique robuste sur la structure des réseaux afin de limiter le nombre de candidats possibles.

Les voies métaboliques, qui rendent compte des connaissances biologiques des réseaux de régulation, nous permettent de définir cet *a priori*. Cette approche est basée sur la sélection d'un ensemble de gènes pertinents, appelé "signature moléculaire", potentiellement associé à un phénotype d'intérêt (par exemple les gènes impliqués dans le développement d'une pathologie). Dans ce contexte, l'analyse différentielle est la strategie prédominante. Néanmoins les signatures de gènes diffèrent d'une étude à l'autre et la robustesse de telles approches peut être remise en question. Ainsi, la première partie de notre travail consistera en l'amélioration de la stratégie d'identification des gènes les plus

[1] Statistics and Genome laboratory UMR CNRS 8071, University of Evry, Evry, France.
E-mail: marine.jeanmougin@genopole.cnrs.fr
[2] Department of Biostatistics, Pharnext, Paris, France.
E-mail: mickael.guedj@pharnext.com and E-mail: Christophe.Ambroise@genopole.cnrs.fr

Article Original

Improving gene signatures by the identification of differentially expressed modules in molecular networks : a local-score approach.

Marine JEANMOUGIN[1,2], Christophe AMBROISE[1], Matthieu BOUAZIZ[1,2] and Mickaël GUEDJ[2]

[1] STATISTICS AND GENOME, UMR8071 CNRS, 23 Boulevard de France, 91037 Évry, France
{marine.jeanmougin, christophe.ambroise}@genopole.cnrs.fr
[2] PHARNEXT, 11 rue des peupliers, 92130 Issy-les-moulineaux, France
{mguedj,mbouaziz}@pharnext.com

Abstract *Discovery of gene signatures for disease prognostic and diagnostic, has become a topic of much interest during the last decade. The identification of relevant and robust signatures is seen as a major step towards a better personalized medicine. Various methods have been proposed for that purpose. However, the detection of a set of genes from microarray data is a difficult statistical problem and signatures obtained from standard tools suffer from lack of reproducibility across studies. Therefore, it is difficult to achieve a clear biological interpretation from such signatures.*
We designed an approach for the selection of functional modules (i.e. subnetworks) of genes through the integration of interactome and transcriptome data. Using a strategy based upon the local-score, a statistic that found many applications in biological sequence analysis, we aim to identify subnetworks locally enriched in genes associated with phenotypes of interest.
We proved through simulations that the resulting modules are highly reproducible. In addition the method appears to be more powerful than classical strategies of gene selection. The potential of our method to highlight relevant biological phenomena is illustrated on breast cancer data to study the Estrogen Receptor (ER) status of tumors.

Keywords gene signatures; functional modules; protein-protein interactions network; differential analysis; breast cancer.

1 Introduction

The development of high throughput genomic and genetic technologies has provided insights into the biological mechanisms underlying diseases onsets and evolutions. Through the identification of biomarker genes (or so called signatures), the aim is to improve diagnosis, prognosis and clinical decisions about treatments of a given disease. One of the most widely used approach for the identification of such signatures consists in detecting differentially expressed genes whose expression levels change between two or more experimental conditions. Such strategies imply to compute an individual score for each gene under study. This score denotes the ability of a gene to discriminate between conditions of interest (patients *versus* controls for instance).
In practice, signatures selected in comparable studies share only few genes in common. Moreover it is generally difficult to achieve a clear biological interpretation of lists of differentially expressed genes as they focus on the level of genes instead of molecular functions or biological processes. Motivated by the observation that genes causing the same phenotype are likely to interact together, we explore an approach for identifying modules, *i.e.* genes that are functionally related, rather than individual genes. The idea is to combine topological features, extracted from Protein-Protein Interaction (PPI) networks for instance, and experimental data to detect modules of connected genes associated with disease or phenotypes of interest.
In recent years, there have been several attempts to integrate knowledge on PPIs, regulatory networks or canonical pathways into gene selection strategies. One of the first approaches was described in [1] and involves a sliding window model to identify differentially expressed subnetworks. It uses the mutual information statistic to measure the association between a subnetwork expression profile and a given phenotype and select significantly differentially expressed subnetworks by comparing their discriminative potentials to those of random networks. In [2], the authors introduced an approach to extract disease-specific gene networks from both DNA microarray measurements and an initial network constructed from protein-protein and genetic interactions as

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Abstract

During the last 3 years, a number of approaches for the normalization of RNA sequencing data have emerged in the literature, differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data continue to accumulate, there has been no clear consensus on the appropriate normalization method to be used or the impact of a chosen method on the downstream analysis. In this work, we focus on a comprehensive comparison of seven recently proposed normalization methods for the differential analysis of RNA-seq data, with an emphasis on the use of varied real and simulated datasets involving different species and experimental designs to represent data characteristics commonly observed in practice. Based on this comparison study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.

Keywords: *high-throughput sequencing; RNA-seq; normalization; differential analysis*

INTRODUCTION

During the last decade, advances in Molecular Biology and substantial improvements in microarray technology have enabled biologists to make use of high-throughput genomic studies. In particular, the simultaneous measurement of the expression levels of tens of thousands of genes has become a mainstay of biological and biomedical research. For example, microarrays have been used to discover genes differentially expressed between two or more groups of interest in a variety of applications. These include the identification of disease biomarkers that may be important in the diagnosis of the different types and subtypes of diseases, with several implications in terms of prognosis and therapy [1, 2].

In recent years, the continuing technical improvements and decreasing cost of next-generation sequencing technology have made RNA sequencing (RNA-seq) a popular choice for gene expression studies. Such sequence-based methods have revolutionized studies of the transcriptome by enabling a wide range of novel applications, including detection of alternative splicing isoforms [3, 4], genome-guided [5, 6] or *de novo* assembly of transcripts [7–9], transcript fusion detection [10] or strand-specific expression [11]. In addition, RNA-seq has become an attractive alternative to microarrays for the identification of differentially expressed genes between several conditions or tissues, as it allows for high coverage of the genome and enables detection of weakly expressed genes [12].

Corresponding author. Marie-Agnès Dillies, Institut Pasteur, PF2 Plate-forme Transcriptome et Epigénome, 28 rue du Dr Roux, Paris CEDEX 15, F-75724 France. Tel.: +33 (0) 145688651; Fax: +33 (0) 145688406; E-mail: marie-agnes.dillies@pasteur.fr

*These authors have contributed equally to this work.

The French StatOmique Consortium gathers more than 40 statisticians and bioinformaticians involved in high-throughput transcriptome data analysis in France. The objective of this group, created in 2008, is to share among researchers and practitioners knowledge of the statistical analysis of high-throughput data.

Chapter 13

Multiple Testing in Large-Scale Genetic Studies

Matthieu Bouaziz, Marine Jeanmougin, and Mickaël Guedj

Abstract

Recent advances in Molecular Biology and improvements in microarray and sequencing technologies have led biologists toward high-throughput genomic studies. These studies aim at finding associations between genetic markers and a phenotype and involve conducting many statistical tests on these markers. Such a wide investigation of the genome not only renders genomic studies quite attractive but also lead to a major shortcoming. That is, among the markers detected as associated with the phenotype, a nonnegligible proportion is not in reality (false-positives) and also true associations can be missed (false-negatives). A main cause of these spurious associations is due to the multiple-testing problem, inherent to conducting numerous statistical tests. Several approaches exist to work around this issue. These multiple-testing adjustments aim at defining new statistical confidence measures that are controlled to guarantee that the outcomes of the tests are pertinent.The most natural correction was introduced by Bonferroni and aims at controlling the family-wise error-rate (FWER) that is the probability of having at least one false-positive. Another approach is based on the false-discovery-rate (FDR) and considers the proportion of significant results that are expected to be false-positives. Finally, the local-FDR focuses on the actual probability for a marker of being associated or not with the phenotype. These strategies are widely used but one has to be careful about when and how to apply them. We propose in this chapter a discussion on the multiple-testing issue and on the main approaches to take it into account. We aim at providing a theoretical and intuitive definition of these concepts along with practical advises to guide researchers in choosing the more appropriate multiple-testing procedure corresponding to the purposes of their studies.

Key words: Multiple testing, Genetic, Association, Biostatistics, GWAS, Bonferroni, FWER, FDR

1. Introduction

During the last decade, advances in Molecular Biology and substantial improvements in microarray and sequencing technologies have led biologists toward high-throughput genomic studies. In particular, the simultaneous genotyping of hundreds of thousands of genetic markers such as single nucleotide polymorphisms (SNPs) on chips has become a mainstay of biological and

François Pompanon and Aurélie Bonin (eds.), *Data Production and Analysis in Population Genomics: Methods and Protocols*,
Methods in Molecular Biology, vol. 888, DOI 10.1007/978-1-61779-870-2_13, © Springer Science+Business Media New York 2012

Network inference in breast cancer with Gaussian graphical models and extensions

Marine Jeanmougin[1,2], Camille Charbonnier[1],
Mickaël Guedj[2], Julien Chiquet[1]

[1] *Laboratoire Statistique et Génome, UMR CNRS 8071, UCR INRA – 23 boulevard de France,*
91 037 Évry, France

[2] *Pharnext – 11 rue des peupliers, 92130, Issy-les-Moulineaux, France*

Bibliography

Affymetrix. Statistical algorithms description document. Rapport technique, 2002. (Cited page 17.)

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, et D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. ISSN 0022-2836. URL http://dx.doi.org/10.1006/jmbi.1990.9999. (Cited page 22.)

C. Ambroise, J. Chiquet, et C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009. (Cited pages 3, 87, and 93.)

S. Anders et W. Huber. Differential expression analysis for sequence count data. *Nature Precedings*, (713):R106, 2010. URL http://dx.doi.org/10.1038/npre.2010.4282.2. (Cited pages 110 and 118.)

P. L. Auer et R. W. Doerge. Statistical design and analysis of rna sequencing data. *Genetics*, 185(2):405–416, 2010. URL http://www.genetics.org/content/185/2/405.abstract. (Cited page 14.)

P. L. Auer et R. W. Doerge. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology*, 10(1): 26, 2011. (Cited page 118.)

o. Banerjee, L. El Ghaoui, et A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, March 2008. (Cited page 82.)

Y. Barash, E. Dehan, M. Krupsky, M. Franklin, W.and Geraci, N. Friedman, et N. Kaminski. Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. *Bioinformatics*, 20(6):839–846, Avril 2004. ISSN 1367-4803. URL http://dx.doi.org/10.1093/bioinformatics/btg487. (Cited page 18.)

Y. Benjamini et Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995. (Cited page 36.)

A. Bernard et A.J. Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput*, pages 459–470, 2005. (Cited page 86.)

B. M. Bolstad, R. A. Irizarry, M. Astrand, , et Speed T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003. (Cited page 18.)

C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. Dans *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, 1935. (Cited page 35.)

Paula D. Bos, Xiang H-F H. Zhang, Cristina Nadal, Weiping Shu, Roger R. Gomis, Don X. Nguyen, Andy J. Minn, Marc J. van de Vijver, William L. Gerald, John A. Foekens, et Joan Massagué. Genes that mediate breast cancer metastasis to the brain. *Nature*, 459(7249):1005–1009, Juin 2009. ISSN 1476-4687. URL `http://dx.doi.org/10.1038/nature08021`. (Cited page 12.)

M. Bouaziz, M. Jeanmougin, et M. Guedj. *Multiple testing in large-scale genetic studies. In 'Data Production and Analysis in Population Genomics*. Methods in Molecular Biology Series, Humana Press., 2012. (Cited page 6.)

S. Boyault, DS. Rickman, A. de Reynies, C. Balabaud, S. Rebouissou, E. Jeannot, A. Herault, J. Saric, J. Belghiti, D. Franco, P. Bioulac-Sage, P. Laurent-Puig, et J. Zucman-Rossi. Transcriptome classification of hcc is related to gene alterations and to new therapeutic targets. *Hepatology*, 45, 2007. URL `http://dx.doi.org/10.1002/hep.21467`. (Cited page 48.)

V. Brendel, P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, et S. Karlin. Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Sciences*, 89(6):2002–2006, 1992. URL `http://www.pnas.org/content/89/6/2002.abstract`. (Cited page 59.)

T.A. Brown. *Genomes 3*. Garland Science, 2006. (Cited pages 7 and 8.)

J. Bullard, E. Purdom, K. Hansen, et S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(1):94, 2010. ISSN 1471-2105. URL `http://www.biomedcentral.com/1471-2105/11/94`. (Cited page 109.)

T. Cabrera, M. A. Fernandez, A. Sierra, A. Garrido, A. Herruzo, A. Escobedo, A. Fabra, et F. Garrido. High frequency of altered hla class i phenotypes in invasive breast carcinomas. *Human Immunology*, 50(2):127 – 134, 1996. ISSN 0198-8859. URL `http://www.sciencedirect.com/science/article/pii/0198885996001450`. (Cited page 101.)

D. Causeur, C. Friguet, M. Houee-Bigot, et M.Kloareg. Factor analysis for multiple testing (FAMT): An R package for large-scale significance testing under dependence. *Journal of Statistical Software*, 40(14):1–19, 2011. URL `http://www.jstatsoft.org/v40/i14/`. (Cited page 37.)

S. S. Chen, D. L. Donoho, et M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, Janvier 2001. ISSN 0036-1445. URL `http://dx.doi.org/10.1137/S003614450037906X`. (Cited page 82.)

D. Chessel, AB. Dufour, et J. Thioulouse. The ade4 package - I : One-table methods. *R News*, 4:5–10, 2004. (Cited page 47.)

J. Chiquet, Y. Grandvalet, et C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011. (Cited pages 3, 84, 85, and 93.)

J. Chiquet, A. Smith, G. Grasseau, C. Matias, et C. Ambroise. Simone: Statistical inference for modular networks. *Bioinformatics*, 25(3):417–418, 2009. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/3/417. (Cited page 126.)

H-Y Y. Chuang, E. Lee, Y-T T. Liu, D. Lee, et T. Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1), Octobre 2007. ISSN 1744-4292. URL http://dx.doi.org/10.1038/msb4100180. (Cited pages 58, 59, 62, and 99.)

G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nat. Gen. Sup.*, 32:490-495, 2002. (Cited page 14.)

J-J. Daudin, F. Picard, et S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18:173–183, 2008. (Cited page 88.)

C. De Mol, E. De Vito, et L. Rosasco. Elastic-net regularization in learning theory. *J. Complex.*, 25(2):201–230, Avril 2009. ISSN 0885-064X. URL http://dx.doi.org/10.1016/j.jco.2009.01.002. (Cited page 120.)

P. Delmar, S. Robin, et J-J. Daudin. Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, 21(4):502–508, Feb 2005. URL http://dx.doi.org/10.1093/bioinformatics/bti023. (Cited page 45.)

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):pp. 157–175, 1972. ISSN 0006341X. URL http://www.jstor.org/stable/2528966. (Cited page 78.)

M-A.* Dillies, A.* Rau, J.* Aubert, C.* Hennequet-Antier, M.* Jeanmougin, N.* Servant, C.* Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S.* Le Crom, M.* Guedj, et F.* *These authors have contributed equally to this work. Jaffrezic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 2012. (Cited pages 107 and 109.)

S. Draghici, P. Khatri, R. Martins, G. Ostermeier, et S. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003. (Cited page 88.)

R. L. Dykstra. Establishing the positive definiteness of the sample covariance matrix. *Annals of Mathematical Statistics*, 41(6):2153–2154, 1970. (Cited page 79.)

David Edwards. *Introduction to Graphical Modelling*. Springer, Juin 2000. ISBN 0387950540. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387950540. (Cited page 75.)

B.y Efron et R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86, Jun 2002. URL http://dx.doi.org/10.1002/gepi.1124. (Cited page 37.)

L. Ein-Dor, I. Kela, G. Getz, D. Givol, et E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005. URL http://bioinformatics.oxfordjournals.org/content/21/2/171.abstract. (Cited page 58.)

E. Eisenberg et E. Y. Levanon. Human housekeeping genes are compact. *Trends Genet*, 19(7):362–365, Jul 2003. URL http://dx.doi.org/10.1016/S0168-9525(03)00140-9. (Cited page 111.)

N. Fahlgren, C. M. Sullivan, K. D. Kasschau, E. J. Chapman, J. S. Cumbie, T. A. Montgomery, S. D. Gilbert, M. Dasenko, T. W.H. Backman, S. A. Givan, et J. C. Carrington. Computational and analytical framework for small rna profiling by high-throughput sequencing. *RNA*, 15(5):992–1002, 2009. URL http://rnajournal.cshlp.org/content/15/5/992.abstract. (Cited page 118.)

Z. Fang et X. Cui. Design and validation issues in rna-seq experiments. *Briefings in Bioinformatics*, 12(3):280–287, 2011. URL http://bib.oxfordjournals.org/content/12/3/280.abstract. (Cited page 14.)

R.A. Fisher. *The Design of Experiments*. Olyver and Boyd Edinburgh, 1935. (Cited page 14.)

J. Friedman, T. Hastie, et R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, Juillet 2008. ISSN 1468-4357. URL http://dx.doi.org/10.1093/biostatistics/kxm045. (Cited page 83.)

C. Friguet et D. Causeur. Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Computational Statistics &amp; Data Analysis*, 55(9):2665 – 2676, 2011. ISSN 0167-9473. URL http://www.sciencedirect.com/science/article/pii/S0167947311001071. (Cited page 37.)

T. K B Gandhi, Jun Zhong, Suresh Mathivanan, L. Karthick, K. N. Chandrika, S. Sujatha Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, Goparani Mishra, Kannabiran Nandakumar, Beiyi Shen, Nandan Deshpande, Rashmi Nayak, Malabika Sarker, Jef D Boeke, Giovanni Parmigiani, Jrg Schultz, Joel S Bader, et Akhilesh Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293, Mar 2006. URL http://dx.doi.org/10.1038/ng1747. (Cited page 67.)

T. S. Gardner, D. di Bernardo, D. Lorenz, et Ja. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, 2003. URL http://www.sciencemag.org/content/301/5629/102.abstract. (Cited page 79.)

G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C.n J. Stoeckert, J. B. Hogenesch, et E. A. Pierce. Comparative analysis of rna-seq alignment algorithms and the rna-seq unified mapper (rum). *Bioinformatics*, 27(18):2518–2528, 2011. URL http://bioinformatics.oxfordjournals.org/content/27/18/2518.abstract. (Cited page 24.)

A.G. Greenwald. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82:1–20, 1975. (Cited page 32.)

M. Guedj, L. Marisa, A. de Reynies, B. Orsetti, R. Schiappa, F. Bibeau, G. MacGrogan, F. Lerebours, P. Finetti, M. Longy, P. Bertheau, F. Bertrand, F. Bonnet, A. L. Martin, J. P. Feugeas, I. Bieche, J. Lehmann-Che, R. Lidereau, D. Birnbaum, F. Bertucci, H. de The, et C. Theillet. A refined molecular taxonomy of breast cancer. *Oncogene*, 31(9):1196–1206, Mar 2011. URL http://dx.doi.org/10.1038/onc.2011.301. (Cited pages 12, 14, 48, 49, 97, and 122.)

M. Guedj, D. Robelin, M. Hoebeke, M. Lamarine, J. Wojcik, et G. Nuel. Detecting local high-scoring segments: a first-stage approach for genome-wide association studies. *Stat Appl Genet Mol Biol*, 5(1):Article22, 2006. (Cited page 59.)

M. Guedj, S. Robin, A. Celisse, et G. Nuel. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC bioinformatics*, 10:84+, Mars 2009. ISSN 1471-2105. URL http://dx.doi.org/10.1186/1471-2105-10-84. (Cited page 37.)

Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E.J. Topol, Q. Wang, et S. Rao. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6:58, 2005. URL http://dx.doi.org/10.1186/1471-2105-6-58. (Cited page 58.)

T. Hardcastle et K. Kelly. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, 2010. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/11/422. (Cited page 118.)

B. Harr et C. Schlatterer. Comparison of algorithms for the analysis of affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res*, 34(2), 2006. URL http://dx.doi.org/10.1093/nar/gnj010. (Cited page 18.)

T. Hastie, R. Tibshirani, et J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. (Cited page 81.)

A-C. Haury, L. Jacob, et J-P. Vert. Improving stability and interpretability of gene expression signatures. Rapport technique, 2010. (Cited page 58.)

K. R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, R. Rouzier, N. Sneige, J. S. Ross, T. Vidaurre, H. L. Gmez, G. N. Hortobagyi, et L. Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol*, 24(26):4236–4244, Sep 2006. URL `http://dx.doi.org/10.1200/JCO.2006.05.6861`. (Cited page 91.)

A. Hoek, J. Schoemaker, et H. A. Drexhage. Premature ovarian failure and ovarian autoimmunity. *Endocrine Reviews*, 18(1):107–134, 1997. URL `http://edrv.endojournals.org/content/18/1/107.abstract`. (Cited page 100.)

A. E. Hoerl et R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, Février 1970. ISSN 00401706. URL `http://dx.doi.org/10.2307/1267351`. (Cited page 81.)

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979. (Cited page 35.)

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, et T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4 (2):249–264, 2003. URL `http://biostatistics.oxfordjournals.org/content/4/2/249.abstract`. (Cited pages 17 and 18.)

F. Jaffrezic, G. Marot, S. Degrelle, I. Hue, et J-L. Foulley. A structural mixed model for variances in differential gene expression studies. *Genet Res*, 89(1):19–25, Feb 2007. URL `http://dx.doi.org/10.1017/S0016672307008646`. (Cited page 44.)

M. Jeanmougin, C. Ambroise, M. Bouzaziz, et M. Guedj. Improving gene signatures by the identification of differentially expressed modules in molecular networks : a local-score approach. Dans *JOBIM proceedings*, 2012a. (Cited page 39.)

M. Jeanmougin, C. Charbonnier, et J. Guedj, M. ANDChiquet. *Network inference in breast cancer with Gaussian graphical models and extensions*. Oxford University Press, 2012b. (Cited page 74.)

M. Jeanmougin, A. de Reynies, L. Marisa, C. Paccard, G. Nuel, et M. Guedj. Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE*, 5(9):e12336, 09 2010. (Cited pages 39 and 114.)

M. Jeanmougin, M. Guedj, et C. Ambroise. Defining a robust biological prior from pathway analysis to drive network inference. *J-SFdS*, 152(2): 97–110, 2011. (Cited pages 74 and 91.)

I. Jeffery, D. Higgins, et A. Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7(1):359, 2006. ISSN 1471-2105. URL `http://www.biomedcentral.com/1471-2105/7/359`. (Cited page 56.)

K. Kammers, M. Lang, JG. Hengstler, M. Schmidt, et J. Rahnenfuhrer. Survival models with preclustered gene groups as covariates. *BMC Bioinformatics*, 12:478, 2011. URL `http://dx.doi.org/10.1186/1471-2105-12-478`. (Cited page 58.)

M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, et M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–D357, Jan 2006. URL `http://dx.doi.org/10.1093/nar/gkj102`. (Cited pages 91 and 101.)

S. Karlin et V. Brendel. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257(5066):39–49, Juillet 1992. ISSN 0036-8075. URL `http://dx.doi.org/10.1126/science.1621093`. (Cited page 59.)

S. Karlin, P. Bucher, V. Brendel, et S. F. Altschul. Statistical methods and insights for protein and DNA sequences. *Annu Rev Biophys Biophys Chem*, 20:175–203, 1991. ISSN 0883-9182. URL `http://view.ncbi.nlm.nih.gov/pubmed/1867715`. (Cited page 59.)

MK. Kerr, M. Martin, et GA. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000. URL `http://www.liebertonline.com/doi/abs/10.1089/10665270050514954`. (Cited pages 14 and 43.)

S. Kim et E. P Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117, 2012. (Cited page 123.)

C. Kooperberg, A. Aragaki, AD. Strand, et JM. Olson. Significance testing for small microarray experiments. *Statistics in medicine*, 24:2281–2298, 2005. (Cited page 56.)

N. Kramer, J. Schafer, et A-L. Boulesteix. Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(1):384, 2009. ISSN 1471-2105. URL `http://www.biomedcentral.com/1471-2105/10/384`. (Cited page 85.)

K. Lage, E. O. Karlberg, Z. M. Størling, P. Í Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, et S. Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–316, Mar 2007. URL `http://dx.doi.org/10.1038/nbt1295`. (Cited page 67.)

L. Lamant, A. de Reynies, MM. Duplantier, DS. Rickman, F. Sabourdy, S. Giuriato, L. Brugieres, P. Gaulard, E. Espinos, et G. Delsol. Gene-expression profiling of systemic anaplastic large-cell lymphoma reveals differences based on ALK status and two distinct morphologic ALK+ subtypes. *Blood*, 109(5):2156–2164, 2007. URL http://bloodjournal.hematologylibrary.org/cgi/content/abstract/bloodjournal;109/5/2156. (Cited page 48.)

P. Latouche, E. Birmelé, , et C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12:93–115, 2011. (Cited page 88.)

S.L. Lauritzen. *Graphical models*. Clarendon Press, 1996. (Cited page 75.)

M-L. T. Lee, F. C. Kuo, G. A. Whitmore, et J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cdna hybridizations. *PNAS*, (97):9834–9839, 2000. (Cited page 14.)

J. Li, D. M. Witten, I. M. Johnstone, et R. Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538, 2012. URL http://biostatistics.oxfordjournals.org/content/13/3/523.abstract. (Cited page 118.)

T. Manoli, N. Gretz, H-J. Grone, M. Kenzelmann, R. Eils, et B. Brors. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22(20):2500–2506, 2006. URL http://bioinformatics.oxfordjournals.org/content/22/20/2500.abstract. (Cited page 88.)

N. Meinshausen, P. Buhlmann, et E. Zurich. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006. (Cited page 83.)

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, et B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7):621–628, Mai 2008. ISSN 1548-7091. URL http://dx.doi.org/10.1038/nmeth.1226. (Cited pages 26 and 110.)

S. Mukherjee et T. P. Speed. Network inference using informative priors. *PNAS*, 105(38):14313–14318, Sep 2008. URL http://dx.doi.org/10.1073/pnas.0802272105. (Cited page 86.)

C. Murie, O. Woody, A. Lee, et R. Nadon. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, 10(1):45, 2009. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/10/45. (Cited pages 42 and 56.)

M. Oti et H. G. Brunner. The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11, Jan 2007. URL http://dx.doi.org/10.1111/j.1399-0004.2006.00708.x. (Cited page 67.)

Z. K. Pinnix, L. D. Miller, W. Wang, R. D'Agostino, T. Kute, M. C. Willingham, H. Hatcher, L. Tesfay, G. Sui, X. Di, S. V. Torti, et F. M. Torti. Ferroportin and iron regulation in breast cancer progression and prognosis. *Sci Transl Med*, 2(43):43ra56, Aug 2010. URL http://dx.doi.org/10.1126/scisignal.3001127. (Cited page 102.)

P. Pons et M. Latapy. Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10:284–293, 2004. (Cited page 62.)

DS. Rickman, R. Millon, A. De Reynies, E. Thomas, C. Wasylyk, D. Muller, J. Abecassis, et B. Wasylyk. Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. *Oncogene*, 27, 2008. (Cited page 48.)

S. Robin, A. Bar-Hen, J-J. Daudin, et L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–5493, 2007. URL http://EconPapers.repec.org/RePEc:eee:csdana:v:51:y:2007:i:12:p:5483-5493. (Cited page 37.)

M. D. Robinson et A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010. URL http://dx.doi.org/10.1186/gb-2010-11-3-r25. (Cited pages 110 and 118.)

T. Schweder et E. Spjotvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982. URL http://biomet.oxfordjournals.org/content/69/3/493.abstract. (Cited page 37.)

M. Scutari et K. Strimmer. Introduction to Graphical Modelling. *ArXiv e-prints*, Mai 2011. (Cited page 85.)

R. Simon, MD. Radmacher, K. Dobbin, et LM. McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14–18, 2003. URL http://jnci.oxfordjournals.org. (Cited page 41.)

GK. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004. ISSN 1544-6115. URL http://dx.doi.org/10.2202/1544-6115.1027. (Cited pages 2 and 44.)

X. Solé, N. Bonifaci, N. López-Bigas, A. Berenguer, P. Hernández, O. Reina, C. A. Maxwell, H. Aguilar, A. Urruticoechea, S. de Sanjosé, F. Comellas, G. Capellá, V. Moreno, et M. A. A. A. Pujana. Biological convergence of cancer signatures. *PloS one*, 4(2):e4544+, Février 2009. ISSN 1932-6203. URL http://dx.doi.org/10.1371/journal.pone.0004544. (Cited page 58.)

J. Soulier, E. Clappier, JM. Cayuela, A. Regnault, M. Garcia-Peydro, H. Dombret, A. Baruchel, ML. Toribio, et F. Sigaux. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, 106(1):274–286, 2005. URL http://bloodjournal.hematologylibrary.

org/cgi/content/abstract/bloodjournal;106/1/274. (Cited pages 48 and 49.)

S. Srivastava et L. Chen. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Res*, 38(17):e170, Sep 2010. URL http://dx.doi.org/10.1093/nar/gkq670. (Cited page 117.)

P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, L. R. Yates, E. Papaemmanuil, D. Beare, A. Butler, A. Cheverton, J. Gamble, J. Hinton, M. Jia, A. Jayakumar, D. Jones, C. Latimer, K. Wai Lau, S. McLaren, D. J. McBride, A. Menzies, L. Mudie, K. Raine, R. Rad, M. S. Chapman, J. Teague, D. Easton, A. Langerod, Oslo Breast Cancer Consortium (OSBREAC), M. T. M. Lee, C-Y. Shen, B. T. K. Tee, B. Wong Huimin, A. Broeks, A. C. Vargas, G. Turashvili, J. Martens, A. Fatima, P. Miron, S-F. Chin, G. Thomas, S. Boyault, O. Mariani, S. R. Lakhani, M. van de Vijver, L. van 't Veer, J. Foekens, C. Desmedt, C. Sotiriou, A. Tutt, C. Caldas, J. S. Reis-Filho, S. A. J. R. Aparicio, A. Vincent Salomon, A-L. Borresen-Dale, A. L. Richardson, P. J. Campbell, P. A. Futreal, et M. R. Stratton. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–404, Jun 2012. URL http://dx.doi.org/10.1038/nature11017. (Cited page 102.)

W. G. Stetler-Stevenson, S. Aznavoorian, et L. A. Liotta. Tumor cell interactions with the extracellular matrix during invasion and metastasis. *Annu Rev Cell Biol*, 9:541–573, 1993. URL http://dx.doi.org/10.1146/annurev.cb.09.110193.002545. (Cited page 102.)

J. D. Storey. A direct approach to false discovery rate. *Journal of the Royal Statistical Society: Series B*, 64(3):479–498, 2001. (Cited page 37.)

T. Strub, S. Giuliano, T. Ye, C. Bonet, C. Keime, D. Kobi, S. Le Gras, M. Cormont, R. Ballotti, C. Bertolotto, et I. Davidson. Essential role of microphthalmia transcription factor for dna replication, mitosis and genomic stability in melanoma. *Oncogene*, 30(20):2319–2332, May 2011. URL http://dx.doi.org/10.1038/onc.2010.612. (Cited pages 20 and 111.)

A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, et J. B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16):6062–6067, Apr 2004. URL http://dx.doi.org/10.1073/pnas.0400782101. (Cited page 111.)

S. Tarazona, Fe. Garcia-Alcalde, J. Dopazo, A. Ferrer, et A. Conesa. Differential expression in rna-seq: A matter of depth. *Genome Research*, 21 (12):2213–2223, 2011. URL http://genome.cshlp.org/content/21/12/2213.abstract. (Cited page 118.)

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. (Cited pages 15 and 82.)

C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, et L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5), 2010. (Cited page 24.)

VG. Tusher, R. Tibshirani, et G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001. ISSN 0027-8424. URL `http://dx.doi.org/10.1073/pnas.091062498`. (Cited pages 42 and 45.)

J-P. Vert et Y. Yamanishi. Supervised graph inference. *Advances in Neural Information Processing Systems*, pages 1433–1440, 2005. URL `http://eprints.pascal-network.org/archive/00001405/`. (Cited page 86.)

L. Wang, Z. Feng, X. Wang, X. Wang, et X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics (Oxford, England)*, 26(1):136–138, Janvier 2010. ISSN 1367-4811. URL `http://dx.doi.org/10.1093/bioinformatics/btp612`. (Cited page 117.)

P. H. Westfall et St. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 1 édition, Janvier 1993. ISBN 0471557617. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471557617`. (Cited page 36.)

J. Whittaker. *Graphical Models in Applied Multivariate Statistics (Wiley Series in Probability & Statistics)*. John Wiley & Sons, March 1990. ISBN 0471917508. URL `http://www.worldcat.org/isbn/0471917508`. (Cited page 75.)

A. Wille, P. Zimmermann, E. Vranova, A. Furholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, et P. Buhlmann. Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology*, 5(11): R92+, 2004. ISSN 1465-6906. URL `http://dx.doi.org/10.1186/gb-2004-5-11-r92`. (Cited page 85.)

G. W. Wright et R. M. Simon. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19(18):2448–2455, Dec 2003. (Cited pages 43 and 48.)

B. Wu. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics*, 21(8):1565–1571, 2005. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/8/1565`. (Cited page 55.)

T. D. Wu et S. Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

URL http://bioinformatics.oxfordjournals.org/content/26/7/873.abstract. (Cited page 24.)

Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, et F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004. URL http://www.tandfonline.com/doi/abs/10.1198/016214504000000683. (Cited page 18.)

Y. Yamanishi, J-P. Vert, et M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20 Suppl 1:i363–i370, Aug 2004. URL http://dx.doi.org/10.1093/bioinformatics/bth910. (Cited page 86.)

J. X. Yu, A. M. Sieuwerts, Y. Zhang, J. W. M. Martens, M. Smid, J. G. M. Klijn, Y. Wang, et J. A. Foekens. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, 7:182, 2007. URL http://dx.doi.org/10.1186/1471-2407-7-182. (Cited page 58.)

M. Yuan et Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006. (Cited page 84.)

S. Zhang. A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics*, 8(1):230, 2007. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/8/230. (Cited page 56.)

L. Zhou et DM. Rocke. An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics*, 21(21):3983–3989, 2005. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/21/3983. (Cited page 41.)

Ce document a été préparé à l'aide du logiciel de composition typographique LaTeX et s'inspire de la feuille de style proposée par J. Chiquet[1].

[1] http://stat.genopole.cnrs.fr/members/jchiquet/welcome