Statistical methods and software for the analysis of transcriptomic data

Andrea Rau

UNIVERSITÉ D'ÉVRY-VAL-D'ESSONNE

École doctorale "du Génome aux Organismes"

Habilitation à diriger les recherches

Spécialité: mathématiques appliquées

présentée par:
Andrea RAU

Statistical methods and software for the analysis of transcriptomic data

---

Mémoire présenté et soutenu publiquement le 26 septembre 2017 à Évry
devant le jury composé de

M. Franck PICARD	CNRS	(Président du jury)
Mme Anne-Laure BOULESTEIX	Ludwig-Maximilians-Universität München	(Rapportrice)
M. David CAUSEUR	Agrocampus Ouest	(Rapporteur)
Mme Nathalie VILLA-VIALANEIX	INRA	(Rapportrice)
M. Christophe AMBROISE	Université d'Évry-Val-d'Essonne	(Examinateur)
M. Stéphane ROBIN	AgroParisTech, INRA	(Examinateur)

Contents

Chapter 1

Introduction

1.1 Personal background

I began my scientific education at Saint Olaf College in Northfield, Minnesota in 2001-2015 with my coursework for a Mathematics major. In my junior year, I had an "aha" moment when I discovered the field of biostatistics through a one-month practicum project, during which I had the opportunity to work with data collected from the National Bone Marrow Donor Program – I had found a rewarding way to analyze and make sense of important real life problems! I quickly filled out the remainder of my bachelor's degree with statistics courses, and went on to get a Master's degree in Applied Statistics and Ph.D. in Statistics at Purdue University in West Lafayette, Indiana. My Ph.D. work, co-supervised by Rebecca W. Doerge at Purdue and Florence Jaffrézic and Jean-Louis Foulley at the French National Institute for Agricultural Research (INRA), focused on the inference of gene regulatory networks from time-course microarray data, and provided me with my first research experience in France: two six-month stays in Jouy en Josas to work at the *Station de génétique quantitative et appliquée* (SGQA) at INRA.

After my Ph.D., I obtained a one-year post-doctoral position at Inria to work on co-expression analyses of RNA-seq data with Gilles Celeux, Marie-Laure Martin-Magniette, and Cathy Maugis-Rabusseau. This was not an easy topic to address, as RNA-seq technology was still in its early days at that time, and it took us some time to fully understand the characteristics of the data and identify the most appropriate modeling strategy. However, we persevered in our research, and extensions to this work are now an active area of interest for me – a good lesson that research can take us in unplanned directions (and that sometimes it is a good idea to abandon research ideas that are going nowhere...!).

Since October 2011, I have worked as a Research Scientist (*chargée de recherche*) in the *Génétique animale et biologie intégrative* (GABI) research unit of INRA in Jouy en Josas. As a member of the biostatistics group in the Populations, Statistics, and Genome (PSGen) research team, I have had the opportunity to benefit from a rich, varied, and collaborative research environment. Throughout my career at INRA, collaborations with biologists, bioinformaticians, and fellow statisticians, both within my research unit and beyond, have provided a rich source of biologically meaningful questions to orient my research towards the development of sound, practical, and useful statistical tools to answer biologically meaningful questions. In particular, the analysis of genomic and transcriptomic data has been a rich source of inspiration for statistical methodological research to identify robust and appropriate analysis tools in the presence of the so-called "curse of dimensionality."

In this manuscript, I will focus on my research activity from 2011-2017. The manuscript is organized as follows: in the remainder of this chapter, I provide some brief thoughts on developing user-friendly software, as well as the notation used throughout the text. The second chapter is dedicated to contributions I made concerning the differential analysis of RNA-seq data, in particular methods to filter weakly expressed genes and to jointly analyze data from multiple related studies. The third chapter focuses on co-expression analyses of

RNA-seq data using finite mixture models. The fourth chapter presents some contributions for the inference of gene regulatory networks from RNA-seq or intervention expression data. Finally, the last chapter discusses some research projects I plan to develop in the future.

1.2 Implementing user-friendly software

Throughout my work, I have strived to develop and maintain open-source software packages implementing our proposed statistical methods in order to facilitate as much as possible the use of our approaches. Software implementation in the R programming language can take several forms: at its most minimal, as raw source code; better, as a structured R package; best, as a fully documented R package with reproducible vignettes, examples, and unit tests, and even tutorials, FAQs, and dedicated web pages. Writing and maintaining useable and user-friendly software is of course a time-consuming endeavor; I do not have any formal training in software engineering, nor do I have a team of software engineers to help design, implement, test, document and maintain the packages I have developed in my research. However, I have found that package development and maintenance (done to the best of my abilities!) has led to wider use of our proposed methods, as well as valuable interactions with and feedback from users.

The majority of my methodological developments have included corresponding R packages hosted on CRAN or GitHub, and I have focused particular energy on `HTSFilter` and `coseq`, two R software packages included in the Bioconductor[1] project. The constraints imposed by Bioconductor ensure that included packages make use of best practices to enable reproducible research and use and fit into the existing infrastructure of classes and methods defined for common genomic data types; Bioconductor maintainers also commit to long-term user support through the Bioconductor support site. To illustrate the use and interoperability of these packages, I have included some relevant R code in the examples in Sections 2.2.3 and 3.3.3. These are intended to be brief examples with code snippets; users should see the appropriate vignettes[2] for full and reproducible examples with each package.

Finally, the packages I have written build upon the extraordinary work provided by other open-source software developers, who are far too numerous to name individually. I am particularly indebted to the work of Hadley Wickham (in particular, the suite of packages contained in the `tidyverse`, including the `ggplot2` (Wickham, 2009) package, which is used several times throughout this work to produce graphics), the Rmixmod team (Lebret, 2015), the R core team (R Development Core Team, 2009), and the Bioconductor core team (Gentleman et al., 2004).

1.3 Notation

Throughout this manuscript, I will make use of the following unified notation unless otherwise noted. Let y_{ij} represent the observed raw read count and $\tilde{y}_{ij}$ the corresponding normalized read count (e.g., after scaling raw counts by library size) for gene i in sample j, with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, q\}$. We denote the full vector of read counts and normalized read counts in a given sample as $\mathbf{y}_j$ and $\tilde{\mathbf{y}}_j$, respectively. The vector $\mathbf{y}_i$ denotes the expression of gene i $(i = 1, \ldots, n)$ across the q samples. Let $\mathcal{C}(j) \in \{1, \ldots, d\}$ represent the experimental condition of sample j; in the context of differential analyses, the number

[1] Note that Bioconductor packages are peer-reviewed, and must meet a checklist of standards of functionality, documentation, and interoperability.

[2] `vignette("HTSFilter")` and `vignette("coseq")`

of conditions d is often equal to 2, whereas co-expression analyses are more often performed for a larger number of experimental conditions. Finally, we use dot notation to indicate summations in various directions, e.g., $y_{\cdot j} = \sum_i y_{ij}$ and $y_{i\cdot} = \sum_j y_{ij}$, and so on.

Chapter 2

Differential analysis of RNA-seq data

In recent years, next-generation high-throughput sequencing (HTS) technology has become an essential tool for genomic and transcriptomic studies. By quantifying and comparing transcriptomes among different types of tissues, developmental stages, or experimental conditions, researchers have gained a deeper understanding of how changes in transcriptional activity reflect specific cell types and contribute to phenotypic differences. In particular, the use of HTS technology to directly sequence reverse-transcribed RNA molecules (complementary DNA; cDNA), known as RNA sequencing (RNA-seq), has revolutionized the study of gene expression by opening the door to a wide range of novel applications. RNA-seq allows for high coverage of the genome, and enables detection of weakly expressed genes and quantification of gene expression without prior knowledge of the genome (e.g. for non-model species). Unlike microarray data, which are continuous, RNA-seq data represent highly heterogeneous counts for genomic regions of interest (typically genes), and often exhibit zero-inflation and a large amount of overdispersion among biological replicates. As such, a great deal of methodological research (e.g., Anders and Huber, 2010; Robinson et al., 2010; Dillies, 2013) has recently focused on appropriate normalization and analysis techniques that are adapted to the characteristics of RNA-seq data; see Oshlack et al. (2010) for a review of RNA-seq technology and analysis procedures.

Although a variety of different protocols exist for high-throughput sequencing studies, the same broad pre-processing steps are followed. Namely, after sequencing fragmented reverse-transcribed transcripts (reads), bioinformatic tools are used to perform quality control and remove adapters and low-quality sequences. Next, if an appropriate genome sequence reference is available, reads are mapped to the genome or transcriptome; otherwise, *de novo* assembly may be used. After alignment or assembly, read coverage for a given biological entity (e.g., a gene or an exon) is subsequently calculated. The quantification of gene expression in RNA-seq data remains an active area of research, and in this work, we focus on measures of digital gene expression (counts). These count-based measures are discrete, nonnegative, and highly skewed, with a very large dynamic range, often covering several orders of magnitude. In addition, sequencing depth (i.e., the library size) and coverage vary between experiments, and read counts are known to be correlated with gene length (see Figure 2.1, Oshlack and Wakefield, 2009; Łabaj, 2011). For these reasons, methods previously proposed for microarray data (which tend to make use of Gaussian distributions after normalization, background correction, and $\log$-transformation) are not typically well-suited to RNA-seq data without some modification.

In this chapter, we focus on two contributions for differential analyses of RNA-seq data: (1) a data-driven filtering criterion to flag and remove genes with weak signal; and (2) a p-value combination approach for differential meta-analyses of multi-study RNA-seq data; we thus begin the chapter with a brief overview of RNA-seq differential analyses.

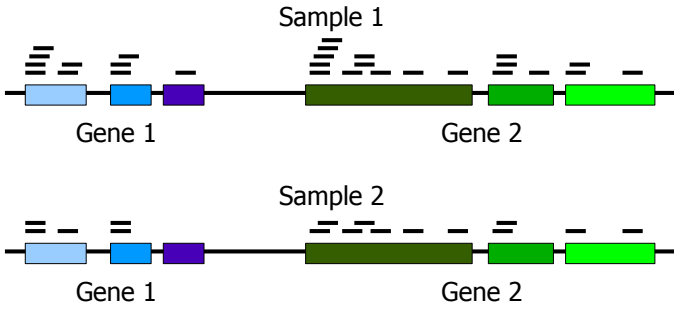

FIGURE 2.1: Schematic representation of alignment of sequenced reads (black bars) from two samples on a reference genome consisting of two genes (blue and green), each made up of three exons. Sample 1 has a larger library size than sample 2, leading to higher overall counts; gene 2 is longer than gene 1, also leading to larger counts.

2.1 Overview of RNA-seq differential analyses

For both microarray and RNA-seq data, it has been shown that normalization is an essential step in the analysis of gene expression; a variety of sources of systematic variation have been reported in RNA-seq data, most notably between-sample differences such as library size (i.e. sequencing depth). Sample-specific normalization factors for RNA-seq data account for the fact that the number of reads expected to map to a particular gene depends not only on its own expression level, but also (1) on the total number of mapped reads (also referred to as library size) in the sample, and (2) on the overall composition of the RNA population being sampled (Figure 2.1). Gene- and sample-specific normalization factors have also been proposed to account for biases due to GC content (Risso, 2011). Although a number of normalization approaches to treat RNA-seq data have emerged in the literature, initially there was no clear consensus on the appropriate normalization method to be used or the impact of a chosen method on the downstream analysis. To address this, the members of the Statomique Consortium[1] conducted a comprehensive comparison of seven proposed normalization methods for RNA-seq data using a variety of real and simulated datasets involving different species and experimental designs (Dillies, 2013). Based on this study, we found the median ratio (Love et al., 2014) and trimmed mean of M-values (TMM; Robinson and Oshlack, 2010) methods to be robust and effective. Without loss of generality, we note $\mathbf{t} = (t_j)$ as the scaling normalization factors for raw library sizes calculated using the TMM normalization method; $\ell_j = y_{.j}t_j$ is then the corresponding normalized library size for sample j, and

$$m_j = \frac{\ell_j}{\sum_{t=1}^{q} \ell_t / q} \tag{2.1}$$

is the associated normalization scaling factor by which raw counts y_{ij} are divided to obtain normalized counts:

$$\tilde{y}_{ij} = y_{ij}/m_j.$$

As with gene expression data arising from microarrays, RNA-seq data are often used to conduct differential analyses. In recent years, several approaches for gene-by-gene tests using gene-level count data have been proposed, with the most popular (including `DESeq2`

[1]The Statomique Consortium is made up of over forty statisticians and biostatisticians involved in high throughput transcriptome analysis from a variety of institutions, including INRA, the Pasteur Institute, the Curie Institute, Inria, and AgroParisTech.

and `edgeR`) making use of negative binomial distributions to account for the overdispersion (i.e., variance larger than the mean) typically observed among biological replicates for a given gene (Robinson et al., 2010; Love et al., 2014). Under these approaches, the count for gene i in sample j is assumed to follow a negative binomal distribution $y_{ij} \sim \mathcal{NB}(\mu_{ij}\phi_i)$ with mean μ_{ij} and variance $\sigma^2_{ij} = \mu_{ij} + \phi_i\mu^2_{ij}$, where

$$\begin{aligned} \mu_{ij} &= q_{ij}m_j \\ \log(q_{ij}) &= \mathbf{X}_j\boldsymbol{\beta}_i, \end{aligned} \tag{2.2}$$

and $\mathbf{X}_j$ represents the design matrix for sample j, $\boldsymbol{\beta}_i$ the vector of coefficients for gene i, and m_j the library size scaling factor for sample j. `DESeq2` and `edgeR` differ in the manner in which model parameters are estimated, but both make use of empirical Bayesian shrinkage approaches to share information among genes in order to provide more robust estimators for ϕ_i for small sample sizes.

In simple two-group experimental designs where $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1})$, the null hypothesis $H_{0i} : \beta_{i1} = 0$ may be tested using an exact conditioned test. In particular, if y_{iA} and y_{iB} represent the sum of normalized counts for gene i in condition A and B, an exact test can be constructed similar to Fisher's exact test for contingency tables, replacing hypergeometric probabilities with negative binomial probabilities:

$$p_i = \frac{\sum\limits_{a+b=y_{i\cdot}} \quad p(a,b)}{\sum\limits_{a+b=y_{i\cdot}} p(a,b)}, \tag{2.3}$$
$$\scriptstyle{p(a,b)\leq p(y_{iA},y_{iB})}$$

where under the null hypothesis it is assumed that $p(a,b) = \Pr(Y_{iA} = a)\Pr(Y_{iB} = b)$ using the negative binomial distribution described in Equation (2.2). In the more recent versions of `DESeq2` and `edgeR`, the Wald test statistic is now instead commonly used, with

$$W_{ir} = \frac{\hat{\beta}_{ir}}{\mathrm{SE}(\hat{\beta}_{ir})} \sim \mathcal{N}(0,1), \tag{2.4}$$

and where $\mathrm{SE}(\cdot)$ denotes the standard error. Because a large number of hypothesis tests are performed for gene-by-gene differential analyses, the obtained p-values must be adjusted to address the fact that many truly null hypotheses will produce small p-values simply by chance; to address this multiple testing problem, several well-established procedures have been proposed to adjust p-values in order to control various measures of experiment-wide false positives, such as the false discovery rate (FDR). Although such procedures may be used to control the number of false positives that are detected, they are often at the expense of the power of an experiment to detect truly differentially expressed (DE) genes, particularly as the number of genes in a typical RNA-seq dataset may be in the thousands or tens of thousands.

2.2 Data-driven filtering of RNA-seq data

This section corresponds to the following published article:

Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 29(17): 2146-2152.

This work is the result of the M2 and Ph.D. work of Mélina Gallopin, co-supervised by Florence Jaffrézic, Gilles Celeux, and myself.

2.2.1 Background

Several authors in the microarray literature have suggested the use of data filters in order to identify and remove genes which appear to generate an uninformative signal and have no or little chance of showing significant evidence of differential expression; only hypotheses corresponding to genes that pass the filter are subsequently tested, which in turn tempers the correction needed to adjust for multiple testing. In recent work, Bourgon et al. (2010) advocated for the use of *independent data filtering*, in which the filter and subsequent test statistic pairs are marginally independent under the null hypothesis and the dependence structure among tests remains largely unchanged pre- and post-filter, ensuring that post-filter p-values are indeed true p-values. For such an independent filter to be effective, it must be positively correlated with the test statistic under the alternative hypothesis; indeed, it is this correlation that leads to an increase in detection power after filtering. In addition, Bourgon *et al.* demonstrated that non-independent filters for which dependence exists between the filter and test statistic (e.g., making use of condition labels to filter genes with average expression in at least one condition less than a given threshold), can in some cases lead to a loss of control of experiment-wide error rates.

In practice, ad hoc filtering techniques are regularly used to moderate this correction by removing genes with low signal, with little attention paid to their impact on downstream analyses. Several ad hoc data filters for RNA-seq data have been used in recent years, including filtering genes with a total read count smaller than a given threshold (Sultan et al., 2008) and filtering genes with at least one zero count in each experimental condition (Bottomly et al., 2011); however, selecting an arbitrary threshold value to filter genes in this way does not account for the overall sequencing depth or variability of a given experiment. One exception to these ad hoc filters is the work of Ramsköld (2009), in which a comparison between expression levels of exonic and intergenic regions was used to find a threshold for detectable expression above background in various human and mouse tissues, where expression was estimated as Reads Per Kilobase per Million mapped reads (RPKM) (Mortazavi, 2008). The threshold of 0.3 RPKM identified in this work has in turn been applied to several other studies (e.g., Łabaj, 2011; Cánovas, 2010; Sam et al., 2011). However, although filters for read counts are routinely used in practice, little attention is typically paid to the choice of the type of filter or threshold used or its impact on the downstream analysis.

2.2.2 Jaccard index filtering threshold

To begin, we consider two broad categories of filters for RNA-seq data, based on the filtering criterion used: mean-based filters and maximum-based filters. Although variance-based filters are routinely used for microarray data (Bourgon et al., 2010), they have not been applied to RNA-seq data; this is likely due to the small number of replicates available in most

TABLE 2.1: Definition of the constants used to calculate the Jaccard similarity index for a pair of samples j and j' and a given threshold s. The constant a represents the number of genes with normalized counts greater than s in both samples j and j', and so on.

		Sample j	
		Normalized counts $> s$	Normalized counts $\leq s$
Sample j'	Normalized counts $> s$	a	b
	Normalized counts $\leq s$	c	d

RNA-seq datasets (and thus, the difficulty in obtaining accurate estimates of per-gene variances) and the fact that the variance is assumed to be a function of the mean under a negative binomial model.

- In *mean-based filters*, genes with mean normalized counts across all samples less than or equal to a pre-specified cutoff are filtered from the analysis. Some authors (Sultan et al., 2008) have also proposed filtering genes with a total read count less than or equal to a given threshold s; we note that this is equivalent to mean-based filters for threshold s divided by the number of samples.

- In *maximum-based filters*, genes with maximum normalized counts across all samples less than or equal to a pre-specified threshold are filtered from the analysis. A generalization of the maximum-based filter has also been proposed in the `edgeR` analysis pipeline (Robinson et al., 2010) based on counts per million (CPM), calculated as the raw counts divided by the library sizes and multiplied by one million. Genes with a CPM value less than a given cutoff (e.g., 1 or 100) in more samples (ignoring condition labels) than the size of the smallest group are subsequently filtered from the analysis.

Regardless of the type of filter used, a biologically pertinent cutoff (or alternatively, number of genes to be filtered) must be chosen; in practice, arbitrary thresholds are routinely used with little or no discussion of their impact on the downstream analysis. To address this issue, we propose a data-based choice for the threshold to be used in maximum-based filters. The main idea underlying this choice is to identify the threshold that maximizes the filtering similarity among replicates, that is, one where most genes tend to either have normalized counts less than or equal to the cutoff in all samples (i.e., filtered genes) or greater than the cutoff in all samples (i.e., non-filtered genes).

We first define a *similarity index* between a pair of replicates within the same condition $\{(\mathbf{y}_j, \mathbf{y}_{j'}) : \mathcal{C}(j) = \mathcal{C}(j')\}$ after binarizing the data for a fixed cutoff s (1 if $y_{ij} > s$ and 0 otherwise). We note that a variety of similarity indices have been proposed since the early 1900s; however, in a comparison among a set of similarity indices we found the Jaccard index (Jaccard, 1901) to be simple, natural, and easy to interpret for the anslysis of high-throughput sequencing data. This index is defined as follows:

$$J_s(\mathbf{y}_j, \mathbf{y}_{j'}) = \frac{a}{a+b+c} \tag{2.5}$$

where a, b, and c are defined in Table 2.1. We note that $J_s(\mathbf{y}_j, \mathbf{y}_{j'})$ takes on values from 0 (dissimilar) to 1 (similar). Because multiple replicates and/or conditions are typically available in HTS experiments, we extend the definition of the pairwise Jaccard index in Equation (2.5) to a global Jaccard index by averaging the indices calculated over all pairs in each

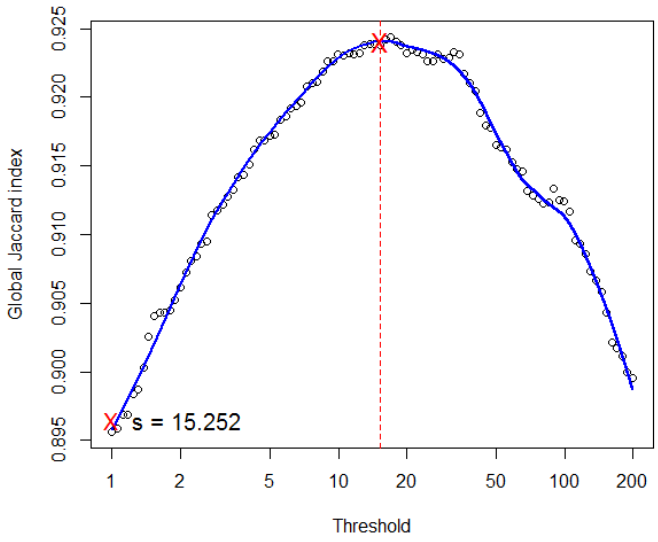

FIGURE 2.2: Global Jaccard index for the Bottomly et al. (2011) data calculated for a variety of threshold values for normalized counts, with a loess curve (blue line) superposed and data-driven threshold value (red cross and red dotted line) equal to $s^\star = 15.252$.

condition:

$$J_s^\star(\mathbf{y}) = \text{mean}\left\{J_s(\mathbf{y}_j, \mathbf{y}_{j'}) : j < j' \text{ and } \mathcal{C}(j) = \mathcal{C}(j')\right\}. \tag{2.6}$$

Using the global Jaccard index defined in Equation (2.6) as a measure of similarity, we now wish to identify the cutoff $s^\star$ for normalized counts that corresponds to the greatest similarity possible among replicates, that is, the value of s corresponding to the maximum value of the global Jaccard index:

$$s^\star = \underset{s}{\text{argmax}}\, J_s^\star(\mathbf{y}). \tag{2.7}$$

In practice, for the calculation of the data-based global filtering threshold in Equation (2.7), we calculate the value of the global Jaccard index in Equation (2.6) for a fixed set of threshold values and fit a loess curve (Cleveland, 1979) through the set of points; the value of $s^\star$ is subsequently set to be the maximum of these fitted values (see Figure 2.2).

Once the data-driven filter threshold for normalized counts $s^\star$ has been identified, the subsequent steps to be taken may change for different applications. To perform an analysis of differential expression between two experimental conditions, we propose using this threshold $s^\star$ in a maximum-based filter, as defined above; we refer to this technique as the *Jaccard filter*.

2.2.3 Data application with `HTSFilter`

The proposed Jaccard filter is implemented in our R/Bioconductor package `HTSFilter`. Using `HTSFilter`, we applied our proposed Jaccard index filter to an RNA-seq dataset from Bottomly et al. (2011) focused on differential striatal expression between inbred mouse strains C57BL/6J (ten biological replicates) and DBA/2J (eleven biological replicates). Raw read counts and phenotype tables may be obtained from the ReCount online resource (Frazee et al., 2011).

In its simplest form, the Jaccard filter may be applied to a `matrix` or `data.frame` containing the raw RNA-seq counts:

```
> library(HTSFilter)
> counts <- exprs(bottomly.eset)
> counts <- counts[rowSums(counts) > 0,]
> conds <- pData(bottomly.eset)$strain
> filter_counts <- HTSFilter(counts, conds=conds)
>
> dim(counts)
[1] 13932    21
> dim(filter_counts$filteredData)
[1] 9049   21
> filter_counts$s
[1] 15.252
```

In the above example, we note that these data, which originally contained expression counts for 13932 genes (with at least one nonzero count) in 21 samples, have been filtered down to a total of 9049 genes in 21 samples, based on the identified data-based filtering threshold of 15.252 (i.e., genes with a maximum normalized count less than this threshold in all samples were filtered from the analysis). The plot shown in Figure 2.2 is automatically generated by a call to the `HTSFilter` function.

In practice, however, the Jaccard filter is most useful if applied directly within a differential analysis pipeline; for this purpose, the negative binomial models implemented in `DESeq2` (Love et al., 2014) and `edgeR` (Robinson et al., 2010) are two popular choices. To illustrate how `HTSFilter` can be inserted into the `edgeR` pipeline, we make use of the following code:

```
> library(edgeR)
> d <- DGEList(counts=counts, group=conds)
> d <- calcNormFactors(d)
> d <- estimateDisp(d)
> fit_nofilter <- exactTest(d)
> fit_filter <- HTSFilter(fit_nofilter, d)$filteredData
> dim(fit_nofilter)
[1] 13932     3
> dim(fit_filter)
[1] 9049    3
```

The effect of the Jaccard filter on the histogram of raw p-values may be seen in Figure 2.3A. Note that the example above makes use of the exact test defined in Equation (2.3); recent versions of `edgeR` now also include a quasi-likelihood F and likelihood ratio tests that tend to be less affected by the discretization of p-values for small counts that contribute to the peak near 1 for the unfiltered analysis.

It is also of interest to consider the effect of each filter on the number of DE genes identified at various levels of expression; in Figure 2.3B, we note that `HTSFilter` leads to more discoveries at all but very weak levels of expression (i.e., mean expression less than 10). A large number of the missed discoveries for the Jaccard filter at very low levels of expression correspond to genes with zero read counts in one condition and a small number of read counts in the other; for example, in the Bottomly et al. (2011) data 49.7% of the 177 missed discoveries among genes with mean expression less than 10 had per-condition means

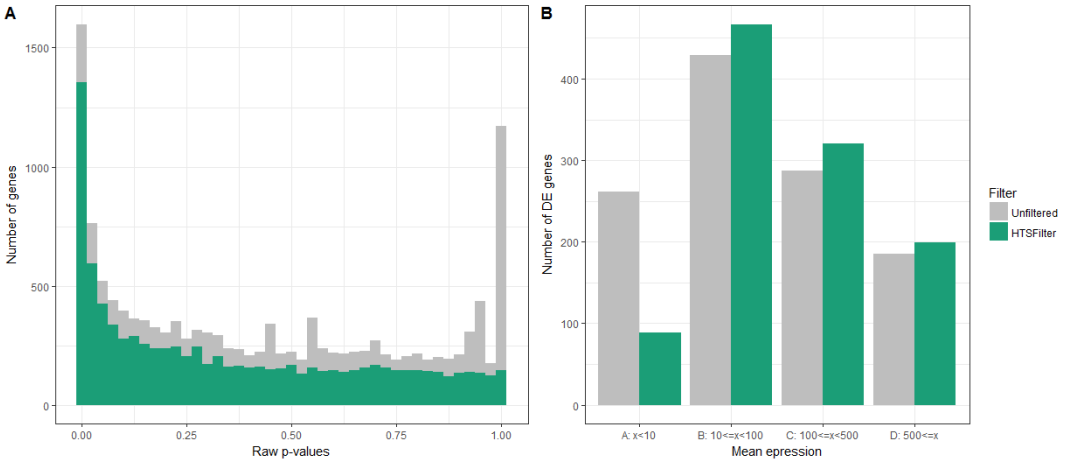

FIGURE 2.3: (left) Histogram of raw p-values from a differential analysis of the Bottomly et al. (2011) RNA-seq data. The histogram in grey in the background represents the raw p-values from a differential analysis using unfiltered data; the histogram in color in the foreground represents the raw p-values from a differential analysis of the data filtered with `HTSFilter`. (right) Number of DE genes detected in the Bottomly et al. (2011) RNA-seq data using unfiltered or `HTSFilter` filtered data, categorized by mean expression.

less than 1 in one of the two conditions and less than 5 in the other. It thus seems reasonable to remove these genes from consideration from the final differential analysis result.

2.2.4 Conclusions and discussion

Data filtering has proven to be of great practical importance for the differential analysis of high-throughput microarray and RNA-seq data by identifying and removing genes with uninformative signal prior to testing. In recent years, many ad hoc procedures have been used to filter RNA-seq data, such as filtering genes with a total or mean normalized read count less than a specified threshold. However, despite its impact on the downstream analyses, clear recommendations concerning the choice of filtering technique are not often provided.

In this work, we proposed a method to calculate a data-driven and non pre-fixed filtering threshold value for normalized counts from replicated RNA-seq data, based on the global Jaccard similarity index. In particular, our proposed filtering technique was found to flag and remove from the analysis a large number of genes with little or no chance of showing evidence of differential expression, and therefore to increase detection power at moderate to high levels of expression through a moderation of the correction for multiple testing. We emphasize that the data-driven threshold value may vary greatly among RNA-seq experiments due to differences in sequencing depth and intra-condition variability (see Figure 2.4). These differences in filtering threshold among experiments are due to both sequencing depth and variability within the data; in particular, experiments with greater sequencing depth will tend to have higher filtering thresholds, and those with greater variability will tend to have lower filtering thresholds. It is worth noting that maximum-based filters are not independent filters as described by Bourgon et al. (2010); in particular, for extremely large filtering thresholds, maximum-based filters do not guarantee control of the Type I error rate if p-values are computed using the pre-filter null distribution. For the threshold values typically used in practice (e.g., based on a quantile or using the global Jaccard index), this is usually not a concern.

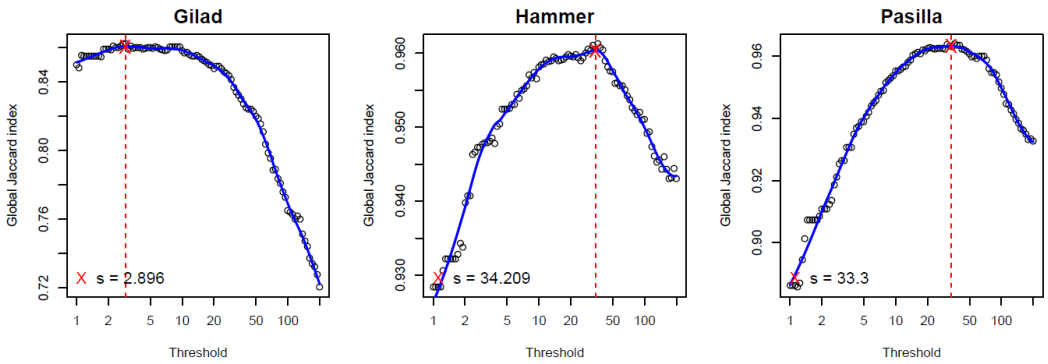

FIGURE 2.4: Global Jaccard index for three RNA-seq datasets calculated for a variety of threshold values, with a loess curve (blue line) superposed and optimal threshold value (red cross and red dotted line). (left) Liver RNA-seq from three male and three female humans Blekhman (2010). (middle) RNA-seq from the L4 dorsal root ganglion in rats with chronic neuropathic pain for two distinct protocols and two time points following spinal nerve ligation, with two replicates for each group (Hammer et al., 2010). (right) RNA-seq data arising from a study of the effect of RNAi knockdown of the Pasilla gene on the *Drosophila melanogaster transcriptome*, with three replicates of the knockdown and four of the untreated control (Brooks, 2011).

In practice there may be some question about the appropriate point in the analysis pipeline to apply data filters: Should normalized data first be filtered, then normalization factors re-estimated and the model fit (i.e., mean and dispersion parameters estimated)? Should normalization factors and model parameters be estimated based on the full data, and the data filtered only at the end of the analysis pipeline? The difference between the two options is nontrivial, particularly as the differential analysis approaches implemented in the `DESeq2` and `edgeR` packages both borrow information across genes (whether all or only those passing the filter) to obtain per-gene parameter estimates. In this work, we present results based on the application of filters applied as late in the pipeline as possible, i.e., after library size and dispersion parameter estimation.

2.3 Meta-analysis of RNA-seq data from related studies

This section corresponds to the following published article:

Rau, A., Marot, G. and Jaffrézic, F. (2014) Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, 15:91.

2.3.1 Background

As RNA-seq experiments remain relatively expensive, typical datasets tend to contain only a few biological replicates, and therefore analyses to detect differential expression between two experimental conditions tend to lack detection power. However, as the costs of such experiments continue to decrease, additional independent experiments may be conducted under the same experimental conditions, suggesting a future need for methods able to jointly analyze data from multiple independent studies. In particular, such methods must be able to appropriately account for the biological and technical variability among samples within a given study as well as for the additional variability due to study-specific effects. Such

inter-study variability may arise due to technical differences among studies (e.g., sample preparation, library protocols, batch effects) as well as additional biological variability.

Several methods have been proposed to analyze microarray data arising from multiple independent but related studies; these meta-analysis techniques have the advantage of increasing the available sample size by integrating related datasets, subsequently increasing the power to detect differential expression. Such meta-analyses include, for example, methods to combine p-values (Marot et al., 2009), estimate and combine effect sizes (Choi, 2003), and rank genes within each study (Breitling, 2004). In many cases the meta-analysis techniques previously used for microarray data are not directly applicable for RNA-seq data. In particular, differential analyses of microarray data, whether for one or multiple studies, typically make use of a standard or moderated t-test (Smyth, 2004; Jaffrézic, 2007), as such data are continuous and may be roughly approximated by a Gaussian distribution after log-transformation. On the other hand, the growing body of work concerning the differential analysis of RNA-seq data has primarily focused on the use of negative binomial models (Love et al., 2014; Robinson et al., 2010) in order to account for their highly dispersed and discrete nature. Under these models, the calculation and interpretation of effect sizes is not straightforward. In this section, we thus present two p-value combination methods for the integrated analysis of RNA-seq data arising from multiple related studies.

2.3.2 p-value combination for multi-study RNA-seq data

For the differential meta-analysis of gene expression arising from multiple studies $s \in \{1, \ldots, S\}$, we begin by conducting per-study differential analyses as described in the introduction to Chapter 2, for example using the `DESeq2` pipeline (Love et al., 2014). In the case of a simple two-group comparison, per-gene and per-study p-values p_{is} are typically calculated using the conditioned exact test in Equation (2.3); in more complex experimental designs, pairwise differential expression is now more often tested using the Wald test statistic in Equation (2.4). After obtaining these vectors of raw p-values for each study, we consider two possible approaches to combine them: the inverse normal and the Fisher combination methods, both of which assume that each vector of p-values is uniformly distributed under the null hypothesis.

- **Inverse normal method.** For each gene i, we define

$$N_i = \sum_{s=1}^{S} w_s \Phi^{-1}(1 - p_{is}) \tag{2.8}$$

  where p_{is} corresponds to the raw p-value obtained for gene i in a differential analysis for study s, Φ the cumulative distribution function of the standard normal distribution, and w_s a set of weights (Stouffer, 1949; Liptak, 1958). We propose the use of study-specific weights w_s, as described by Marot and Mayer (2009):

$$w_s = \sqrt{\frac{q_s}{\sum_\ell q_\ell}},$$

  where q_s is the total number of biological replicates in study s. This allows studies with large numbers of biological replicates to be attributed a larger weight than smaller studies. Other weights may also be defined by the user depending on the quality of the data in each study, if this information is available.

  Under the null hypothesis, the test statistic N_i in Equation (2.8) follows a $\mathcal{N}(0, 1)$ distribution. A unilateral test on the right-hand tail of the distribution may then be performed, and classical procedures for the correction of multiple testing may subsequently be applied to control the false discovery rate at a desired level α.

- **Fisher combination method.** For the Fisher combination method (Fisher, 1932), the test statistic for each gene i may be defined as

$$F_i = -2\sum_{s=1}^{S}\ln(p_{is}), \tag{2.9}$$

where p_{is} is as before. Under the null hypothesis, the test statistic F_i in Equation (2.9) follows a χ^2 distribution with $2S$ degrees of freedom. As with the inverse normal p-value combination method, classical procedures for the correction of multiple testing may be applied to the combined p-values.

The implementation of the previously described p-value combination techniques requires two additional considerations to be taken into account when dealing with RNA-seq data. First, a crucial underlying assumption for the statistics defined in Equations (2.8) and (2.9) is that p-values for all genes arising from the per-study differential analyses are uniformly distributed under the null hypothesis. This assumption is, however, not always satisfied for RNA-seq data; in particular, a peak is often observed for p-values close to 1 due to the discretization of p-values for very low counts. To circumvent this first difficulty, we first filter weakly expressed genes in each study, using the `HTSFilter` R/Bioconductor package described in Section 2.2.3. As will be seen in the following, this approach appears to effectively filter those genes contributing to a peak of large p-values, resulting in p-values that are roughly uniformly distributed under the null hypothesis (see Figure 2.3A for an example).

Second, unlike microarray data, under- and over-expressed genes are analyzed together for RNA-seq data when the conditioned exact test in Equation (2.3) is used. As such, some care must be taken to identify genes exhibiting conflicting expression patterns (i.e., under-expression when comparing one condition to another in one study, and over-expression for the same comparison in another study). In the case of microarray data, Marot et al. (2009) suggested the use of one-tailed p-values for each study to avoid directional conflicts; as the inverse normal combination method was used in their work, the combined statistic thus follows a normal distribution, which is symmetric. Because under- and over-expressed genes may be found in the left and right tail, respectively, of the corresponding normal distribution, it is thus possible to use a two-tailed test to simultaneously study over and under-expressed genes. Note that Pearson (1934) and Owen (2009) proposed another alternative to handle conflicting differential expression if the Fisher combination method is instead used. However, in the case of RNA-seq data, the use of the exact test in Equation (2.3) does not enable the separation of over- and under-expressed genes in distribution tails; in such cases it is not possible to use the approaches proposed Marot et al. (2009) or Owen (2009). We thus suggest that either (1) a one-sided p-value be used with the Wald test statistic in Equation (2.4) and use one of the approaches proposed Marot et al. (2009) or Owen (2009); or (2) genes exhibiting differential expression conflicts among studies be identified post hoc and removed from the final list of differentially expressed genes.

The p-value combination approaches detailed above are implemented in the R package `metaRNASeq`, freely available on CRAN.

2.3.3 Conclusions and discussion

We compared the p-value combination techniques, a negative binomial GLM with fixed study effect, and the intersection of individual differential analyses on real and simulated data. Unsurprisingly, the latter approach is overly conservative, as only genes with adjusted p-values less than the desired significance threshold in all studies are identified as differentially expressed. Accounting for study effects (whether through the GLM with study effect or

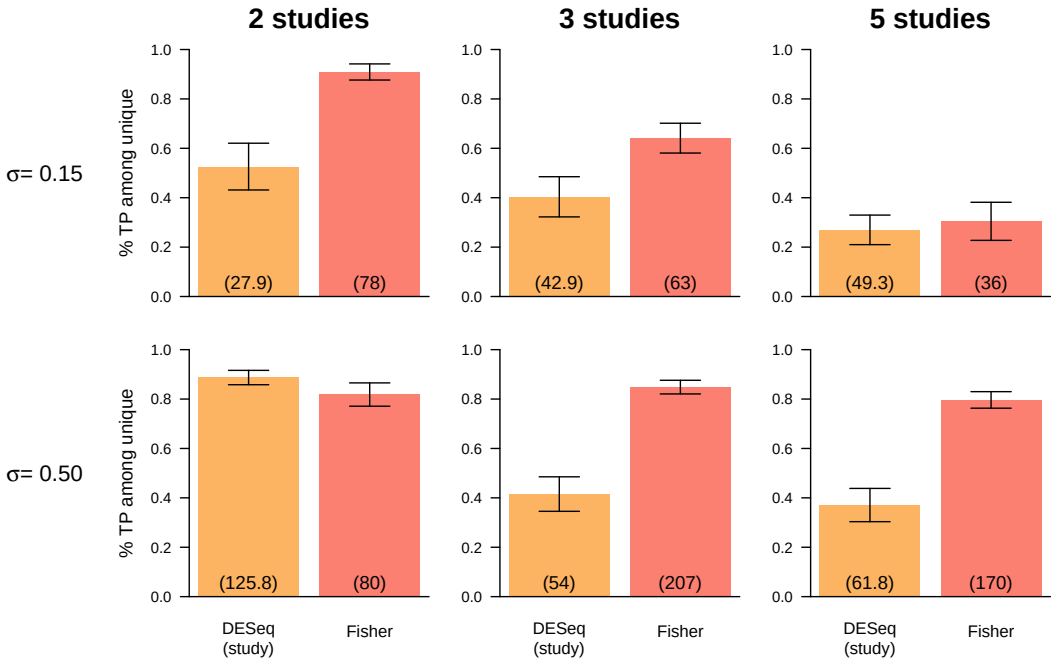

FIGURE 2.5: Proportion of true positives among unique discoveries for negative binomial GLM with a fixed study effect (orange) and Fisher p-value combination (red). Columns (from left to right) correspond to simulations with 2, 3, and 5 studies, and rows (from top to bottom) correspond to simulations with low ($\sigma = 0.15$) and high ($\sigma = 0.5$) inter-study variability. Error bars represent one standard deviation, and numbers in parentheses represent the mean total number of unique discoveries for each method.

the p-value combination approaches) considerably increases detection power; in simulations with low inter-study variability and/or a small number of independent studies (e.g., 2), these approaches had similar detection power (see Rau et al. (2014) for details). However, for increasing inter-study variability and number of studies, the gains in performance in terms of AUC, sensitivity, and proportion of true positives among uniquely identified genes for the meta-analysis techniques are more marked (see Figure 2.5).

The methods presented here are intended for the analysis of data in which all experimental conditions under consideration are included in every study, thus avoiding problems due to the confounding of condition and study effects. As with all meta-analyses, the p-value combination techniques presented here must overcome differences in experimental objectives, design, and populations of interest, as well as differences in sequencing technology, library preparation, and laboratory-specific effects. In order to be biologically relevant, the p-value combination methods rely on the fact that the same test statistics, or in the case of RNA-seq data conditioned tests, are used to obtain p-values for each study. An important challenge for the future will be to propose methods able to jointly analyze related heterogeneous data, such as microarray and RNA-seq data, or other kinds of genomic data.

Chapter 3

Co-expression analysis of RNA-seq data

Identifying biological entities that share similar profiles across several treatment conditions, such as co-expressed genes, may help identify groups of genes that are involved in the same biological processes (Eisen, 1998; Jiang et al., 2004). By identifying clusters of co-expressed genes, we thus aim both to identify co-regulated genes and to characterize potential biological functions for orphan genes (namely, those whose biological function is unknown). It is worth noting that the concept of *gene co-expression* is alternatively used to refer to two broad types of analyses (D'haeseleer et al., 2000): 1) clustering gene expression patterns to explore shared function and co-regulation (our focus in this chapter); and 2) network inference, which aims to construct a model of the network of regulatory interactions between genes (our focus in Chapter 4). Although a variety of methods have been developed for co-expression analyses in microarray data (i.e. the identification of groups of genes that share the same behavior over a set of experimental conditions), for the time being little has been proposed to study co-expression from RNA-seq data.

In the following chapter, we make use of probabilistic clustering models, where the objects to be classified (genes) are considered to be a sample of a random vector and a clustering of the data is obtained by analyzing the density of this vector (McLachlan, 2004; Yeung, 2001); we thus begin the chapter with a brief overview of finite mixture models.

3.1 Overview of finite mixture models

In the context of model-based clustering, the data $\mathbf{y}$ are assumed to be sampled from a finite mixture density of K random variables, each with parameterized density $f_k(\mathbf{y}_i; \boldsymbol{\theta}_k)$, $k = 1, \ldots, K$, where the mixture parameters $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ are all assumed to be distinct. The density of $\mathbf{y}$ may thus be written as

$$f(\mathbf{y}; K, \boldsymbol{\Psi}_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k), \tag{3.1}$$

where $\boldsymbol{\Psi}_K = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ are the parameters of the mixture model, and $(\pi_1, \ldots, \pi_K)$ are the mixing proportions with $\pi_k \in (0, 1)$ for all k, $\sum_{k=1}^{K} \pi_k = 1$.

For parameter estimation, the mixture model in Equation (3.1) may be thought of as an incomplete data structure model where $\mathbf{z}$ is the $(n \times K)$ matrix of unknown mixture labels, with $z_{ik} = 1$ if gene i is from group k and 0 otherwise. Note that this matrix defines a partition of the genes. Using the mixture labels z, the completed density of $\mathbf{y}$ may be written as follows:

$$f(\mathbf{y}, \mathbf{z}; K, \boldsymbol{\Psi}_K) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k))^{z_{ki}}.$$

The maximum likelihood estimate $\hat{\mathbf{\Psi}}_K$ of the mixture parameters is estimated using the Expectation-Maximization algorithm (Dempster et al., 1977). After initializing the parameters $\mathbf{\Psi}_K^{(0)}$ and $\mathbf{z}^{(0)}$, the E-step at iteration b corresponds to computing the conditional probability that an observation i arises from the $k\text{th}$ component for the current value of the mixture parameters:

$$\tau_{ik}^{(b)} = \tau_{ik}\left(\mathbf{\Psi}^{(b)}\right) = \frac{\pi_k^{(b)} f_k(\mathbf{y}_i; \boldsymbol{\theta}_k^{(b)})}{\sum_{m=1}^{K} \pi_m^{(b)} f_m(\mathbf{y}_i; \boldsymbol{\theta}_m^{(b)})}. \tag{3.2}$$

Then, in the M-step the mixture parameter estimates are updated to maximize the expected value of the completed likelihood, which leads to weighting the observation i for group k with the conditional probability $\tau_{ik}^{(b)}$. Thus, at iteration b of the algorithm,

$$\pi_k^{(b+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(b)}, \tag{3.3}$$

and the M-step update for $\boldsymbol{\theta}_k^{(b+1)}$ depends on the specific family of models f_k.

One important task in model-based clustering is the choice of an appropriate model, most notably the revelant number of clusters K. To this end, a standard model selection criterion is the Bayesian Information Criterion (BIC; Schwarz, 1978):

$$\text{BIC}(K) = -\log f(\mathbf{y}; K, \hat{\mathbf{\Psi}}_K) + \frac{\nu_K}{2}\log(n), \tag{3.4}$$

where $\hat{\mathbf{\Psi}}_K$ is the maximum likelihood estimator of the mixture parameters and ν_K the number of free parameters in the model with K components. This criterion is an asymptotic approximation of the logarithm of the integrated likelihood:

$$f(\mathbf{y}; K) = \int_{\mathbf{\Psi}_K} f(\mathbf{y}; K, \mathbf{\Psi}_K)\pi(\mathbf{\Psi}_K)d\mathbf{\Psi}_K,$$

where $\pi(\mathbf{\Psi}_K)$ is a weakly informative prior distribution on $\mathbf{\Psi}_K$.

An alternative to the BIC is the Integrated Completed Likelihood (ICL) criterion (Biernacki, 2000):

$$\text{ICL}(K) = \text{BIC}(K) + \text{Entropy}(K), \tag{3.5}$$

where $\text{Entropy}(K)$ is the estimated mean clustering entropy:

$$\text{Entropy}(K) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\hat{\mathbf{\Psi}}_K)\log\tau_{ik}(\hat{\mathbf{\Psi}}_K) \geq 0. \tag{3.6}$$

Note that the ICL is a BIC-like approximation of the logarithm of the completed integrated likelihood:

$$f(\mathbf{y}, \mathbf{z}; K) = \int_{\mathbf{\Psi}_K} f(\mathbf{y}, \mathbf{z}; K, \mathbf{\Psi}_K)\pi(\mathbf{\Psi}_K)d\mathbf{\Psi}_K.$$

Because of the additional entropy term defined in Equation (3.6), the ICL favors models that lead to data partitions with the greatest evidence in terms of classification.

A different approach to model selection is the use of the *slope heuristics* (Birgé and Massart, 2001; Birgé and Massart, 2007), which is a data-driven method to calibrate a penalized criterion known up to a multiplicative constant. Briefly, in our context the penalty is assumed to be proportional to the number of free parameters ν_K, such that $\text{pen}(K) \propto \kappa\nu_K$; we note

that this assumption may be verified in practice. The penalty is calibrated using the *data-driven slope estimation* (DDSE) procedure available in the `capushe` R package (Baudry et al., 2012). This procedure directly estimates the slope of the expected linear relationship of the log-likelihood with respect to the model dimension for the most complex models (here, models with large K). Denoting the estimated slope $\hat{\kappa}$, in our context the slope heuristics consists of setting the penalty to be $2\hat{\kappa}\nu_K$, yielding the following penalized criterion:

$$\text{SH}(K) = -\log f(\mathbf{y}; K, \hat{\boldsymbol{\Psi}}_K) + 2\hat{\kappa}\nu_K. \tag{3.7}$$

For more details, see Baudry et al. (2012). For all of the criteria defined in Equations (3.4)-(3.7), the number of selected clusters $\hat{K}$ corresponds to the value of K minimizing the penalized criterion. Finally, based on $\hat{\boldsymbol{\Psi}}_{\hat{K}}$, each observation i is assigned to the component maximizing the conditional probability $\hat{t}_{ik}$ using the so-called MAP rule: for each $i = 1, \ldots, n$ and each $k = 1, \ldots, K$,

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \tau_{ik}\left(\hat{\boldsymbol{\Psi}}_{\hat{K}}\right) > \tau_{i\ell}\left(\hat{\boldsymbol{\Psi}}_{\hat{K}}\right) \; \forall \ell \neq k \\ 0 & \text{otherwise,} \end{cases}$$

where $\tau_{ik}(\hat{\boldsymbol{\Psi}}_{\hat{K}})$ is as defined in Equation (3.2).

In the remainder of this chapter, we focus on three developments for model-based clustering of RNA-seq data: (1) directly modeling raw gene counts with a Poisson mixture model; (2) modeling transformed gene profiles with a Gaussian mixture model; and (3) using functional annotation information to guide model selection.

3.2 Clustering raw RNA-seq counts

This section corresponds to the following published article:

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., Celeux, G. (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31(9): 1420-1427.

3.2.1 Background and motivation

One of the first questions that must be addressed when seeking to cluster raw RNA-seq counts is to precisely define the end goal. In particular, RNA-seq data are characterized by large differences in scale between genes (due to differences in the level or rate of transcription between genes as well as to differences in the length of the coding region between genes). In this work, rather than clustering together genes with similar absolute expression (e.g., strongly expressed genes versus weakly expressed genes), we instead focus on clustering *relative* expression across experiments.

To illustrate this, consider Figure 3.1, in which we plot normalized counts $\tilde{y}_{ij}$, log-transformed normalized counts $\log(\tilde{y}_{ij})$, and normalized expression profiles $\tilde{y}_{ij}/\tilde{y}_{i\cdot}$ for a subset of genes from the mouse RNA-seq data studied by Fietz (2012) (see Section 3.3.3 for a description of these data). In particular, we consider ten representative genes from four distinct groups: non-differentially expressed (NDE) genes (Group 1); and genes expressed only in the last, first, or second experimental conditions (Group 2, 3, 4). It may clearly be seen that the large differences in magnitude that are dominant for normalized counts (Figure 3.1A) are greatly reduced by a log-transformation (Figure 3.1B), although a certain amount of spread

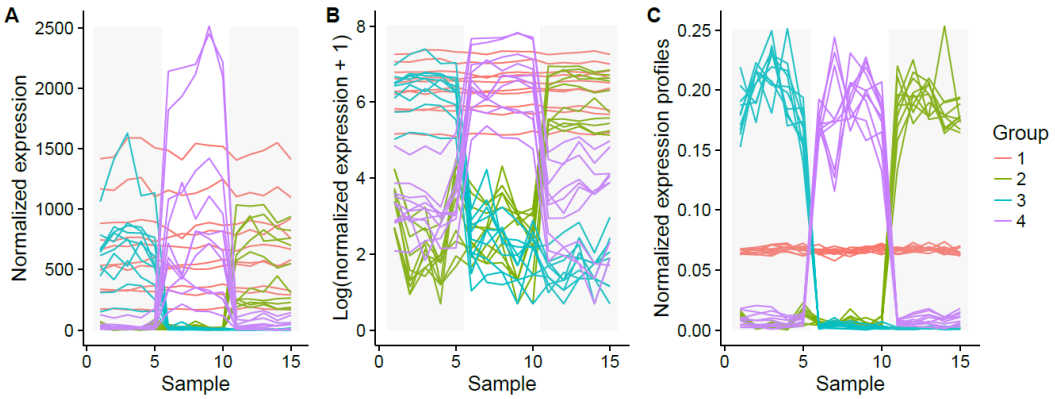

FIGURE 3.1: Normalized counts (A), log normalized counts + 1 (B), and normalized expression profiles (C) for a subset of the Fietz (2012) mouse RNA-seq data. The subset of genes include non-differentially expressed (NDE) genes across all samples (Group 1); genes expressed only in the last experimental condition (samples 11 to 15, Group 2); genes expressed only in the first experimental condition (samples 1 to 5, Group 3); and genes expressed only in the second experimental condition (samples 6 to 10, Group 4). Transparent grey boxes delimit the replicates in each of the three experimental groups.

remains between very highly and weakly expressed genes. This spread is notably reduced by considering the normalized expression profiles (Figure 3.1C). This example is thus instructive in illustrating the importance in co-expression analyses of considering a measure that is independent of the absolute expression level of the genes, as is the case for the normalized profiles, when relative expression patterns are of interest.

In the following section, we consider a model parameterization to focus on these relative expression patterns; in Section 3.3 we will revisit the question of modeling the normalized expression profiles.

3.2.2 Poisson mixture models for RNA-seq counts

We first focus on the use of Poisson loglinear models to cluster count-based RNA-seq expression profiles; however, rather than using such a model to define a distance metric to be used in a K-means (Cai, 2004) or hierarchical clustering (Si, 2014) algorithm, we make use of finite mixtures of Poisson loglinear models. This framework has the advantage of directly modeling the raw gene counts and providing a straightforward procedure for parameter estimation and model selection, as well as a per-gene conditional probability of belonging to each cluster.

Although a multivariate version of the Poisson distribution does exist (Karlis, 2003), it is difficult to implement, particularly for data with high dimensionality. For this reason, in this work we assume the samples are conditionally independent given the components:

$$f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik}) = \prod_{j=1}^{q} \mathcal{P}(y_{ij}; \mu_{ijk}),$$

where $\mathcal{P}(\cdot)$ denotes the standard Poisson probability mass function and $\boldsymbol{\theta}_{ik} = \{\mu_{ijk}\}_j$. We note that although the assumption of conditional independence of components is quite strong, it is commonly employed to analyze multivariate categorical data; for instance, the latent class model is a reference model in model-based cluster analysis of categorical data (McCutcheon, 1987). When this conditional independence assumption is not expected to hold, in

practice it leads to a larger number of clusters and a more complex mixture model that is still able to adequately fit the data.

Each mean μ_{ijk} is further parameterized by

$$\mu_{ijk} = w_i m_j \lambda_{\mathcal{C}(j)k} \tag{3.8}$$

where $w_i = y_{i\cdot}$ corresponds to the overall expression level of observation i (e.g., weakly to strongly expressed) as well as a proxy for gene length, and m_j represents the rescaled normalized library size for sample j, such that $\sum_j m_j = 1$. These normalization factors take into account the fact that the number of reads expected to map to a particular gene depends not only on its expression level, but also on the library size (overall number of mapped reads) and the overall composition of the RNA population being sampled (Dillies, 2013). We note that $\{m_j\}_j$ are estimated from the data prior to fitting the model (see the introduction to Chapter 2 for more details), and like the overall expression levels w_i, they are subsequently considered to be fixed in the Poisson mixture model. Finally, the unknown parameter vector $\boldsymbol{\lambda}_k = (\lambda_{1k}, \ldots, \lambda_{dk})$ corresponds to the clustering parameters that define the profiles of the genes in cluster k across all biological conditions. Thus,

$$\tilde{\lambda}_{ck} = \lambda_{ck} \sum_{j:\mathcal{C}(j)=c} m_j$$

can be interpreted as the proportion of reads that are attributed to condition c in cluster k, after accounting for differences due to library size; this proportion is shared among the replicates of condition c according to their respective library sizes $\{m_j\}_{j:\mathcal{C}(j)=c}$.

To estimate the mixture parameters $\boldsymbol{\Psi}_K = (\boldsymbol{\pi}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$ by computing the maximum likelihood estimate (MLE), an Expectation-Maximization (EM) algorithm is used (Dempster et al., 1977) as described in Equations (3.2)-(3.3). To complete the M-step for the Poisson mixture model, we have

$$\lambda_{ck}^{(b+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(b)} \sum_{j:\mathcal{C}(j)=c} y_{ij}}{m_j \sum_{i=1}^{n} \tau_{ik}^{(b)} y_{i\cdot}},$$

since $w_i = y_{i\cdot}$.

Particular care must be taken with the intialization of parameters for the Poisson mixture model, particularly for large K. As such, we propose the use of a hybrid splitting-small EM initialization that combines the strategies proposed by Papastamoulis et al. (2016) and

Biernacki et al. (2003), which proceeds as follows:

for $K \leftarrow 2$ **to** `Kmax` **do**
- Calculate per-class entropy $e_k = -\sum_{i \in k} \log \hat{t}_{ik}^{K-1}$ for model with $(K-1)$ clusters
- Select cluster $k^\star = \text{argmax}_k \, e_k$ to be split
- **for** $i \leftarrow 1$ **to** `init.runs` **do**
  - Randomly split the observations in cluster $k^\star$ into two clusters
  - Calculate corresponding $\boldsymbol{\lambda}^{(0,i),K}$ and $\boldsymbol{\pi}^{(0,i),K}$
  - Update values of $\boldsymbol{\lambda}^{(0,i),K}$ and $\boldsymbol{\pi}^{(0,i),K}$ via EM algorithm with `init.iter` iterations
  - Calculate the log-likelihood $L^{(i),K} = L(\hat{\boldsymbol{\lambda}}^{(0,i),K}, \hat{\boldsymbol{\pi}}^{(0,i),K})$
- **end**
- Let $i^\star = \text{argmax}_i \, L^{(i),K}$. Fix new initial values $\boldsymbol{\lambda}^{(0),K} = \hat{\boldsymbol{\lambda}}^{(0,i^\star),K}$ and $\boldsymbol{\pi}^{(0),K} = \hat{\boldsymbol{\pi}}^{(0,i^\star),K}$.
- **for** $i \leftarrow 1$ **to** `iter` **do**
  - Update values of $\boldsymbol{\lambda}^{(i),K}$ and $\boldsymbol{\pi}^{(i),K}$ via EM algorithm
  - **if** $L^{(i),K} - L^{(i-1),K} <$ `cutoff` **then** stop
- **end**
- **output:** $\hat{\boldsymbol{\lambda}}^K$ and $\hat{\boldsymbol{\pi}}^K$ for the model with K clusters.

end

Our proposed clustering procedure based on a Poisson mixture model is implemented in the R package `HTSCluster`, freely available on CRAN.

3.2.3 Data application

We illustrate the use of the proposed Poisson mixture model on data arising from the modENCODE project, which aimed to provide functional annotation of the *Drosophila melanogaster* genome. Graveley (2011) characterized the expression dynamics over 27 distinct stages of development during the life cycle of the fly using RNA-seq. In this work, we focus on a subset of these data from 12 embryonic samples that were collected at two-hour intervals for 24 hours, with one biological replicate for each time-point. The phenotype tables and raw read counts for the 13,164 genes with at least one non-zero count among the 12 time-points were obtained from the ReCount online resource (Frazee et al., 2011).

Over three independent runs, we used the `HTSCluster` package with default settings and the splitting small-EM initialization strategy to fit a sequence of Poisson mixture models with $K = 1, \ldots, 60$ clusters; for each number of clusters, the model corresponding to the largest log-likelihood among the three runs was retained. To ensure that the collection of models considered is large enough to apply the slope heuristics model selection, one additional set of Poisson mixture models was fit for $K = 65, \ldots, 95$ (in steps of 5) and $K = 100, \ldots, 130$ (in steps of 10). Using the slope heuristics, the number of clusters was determined to be $\hat{K} = 48$.

Visualizing the results of a co-expression analysis for RNA-seq data can be somewhat complicated by the extremely large dynamic range of gene counts and the fact that more highly expressed genes tend to exhibit greater variability (though much smaller coefficients of variation) than weakly expressed genes. For the purposes of co-expression, rather than directly visualizing the raw counts themselves, we propose the use of either line plots of the normalized expression profiles (Figure 3.2, top) or an alternative visualization of the overall behavior of each cluster (Figure 3.2, bottom). In the latter plot, bar widths correspond to the estimated proportion of genes in each cluster ($\hat{\pi}_k$), and the proportion of reads attributed to each developmental time-point in each cluster $\tilde{\lambda}_{ck}$ are represented by the colored segments within each bar. The advantage of such a visualization is that it enables a straightforward

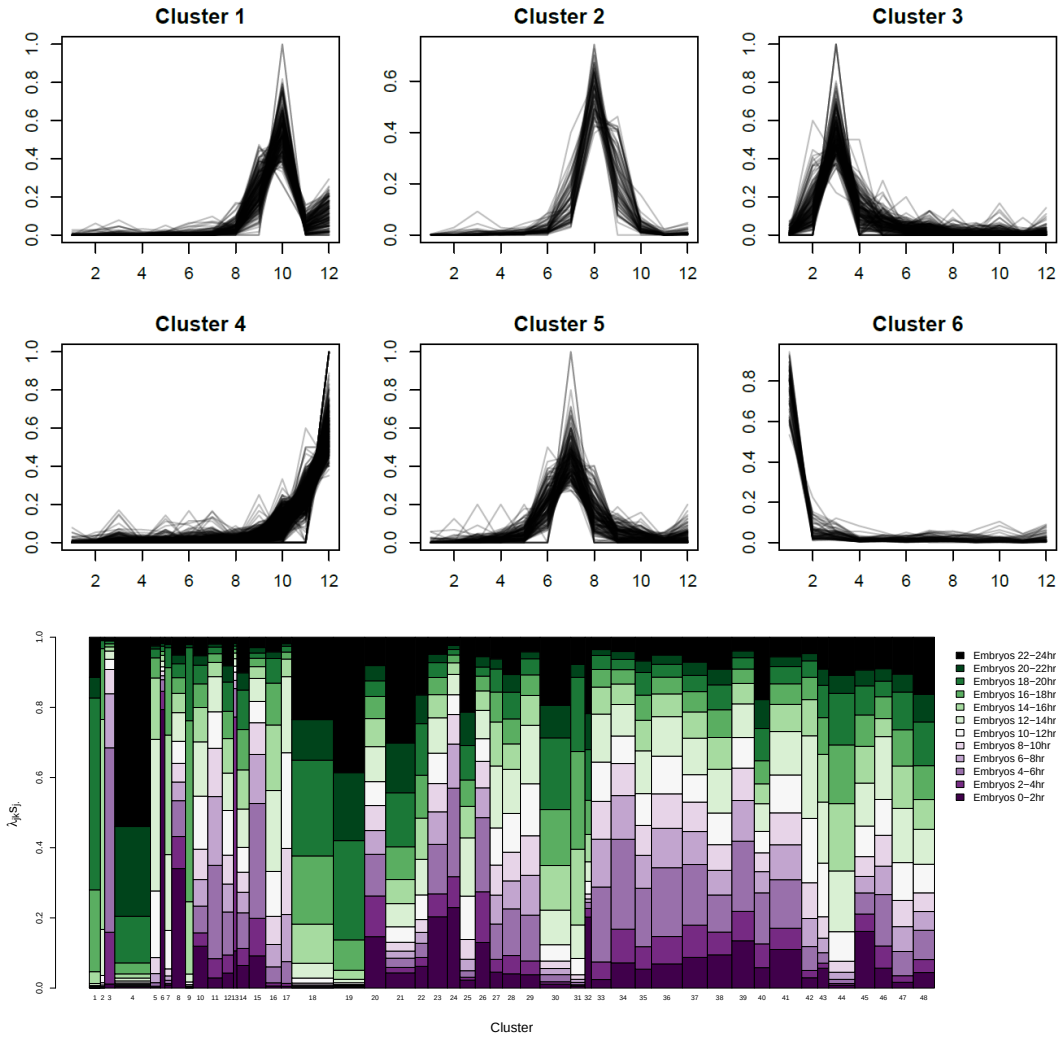

FIGURE 3.2: (top) Line plots of time versus normalized expression profile for the genes assigned to the first six clusters. (bottom) Visualization of overall cluster behavior for the *D. melanogaster* developmental data. For each cluster, bar plots of $\hat{\lambda}_{ck} \sum_{\mathcal{C}(j)=c} m_j$ are drawn for each developmental time-point, where the width of each bar corresponds to the estimated proportion $\hat{\pi}_k$.

comparison of typical gene profiles among clusters. For instance, it can be seen that clusters characterized by higher relative expression in the early embryonic stages, such as Clusters 6 and 13 (composed of 70 and 60 genes, respectively) tend to be much smaller than those with higher relative expression in later stages, e.g., Clusters 4, 18, 19, and 21 (composed of 567, 680, 485, and 475 genes, respectively).

3.2.4 Conclusions and discussion

In this work, we have proposed a method and associated R package `HTSCluster` to cluster count-based DGE profiles based on a Poisson mixture model that enables the use of a rigorous framework for parameter estimation (through the EM algorithm) and model selection (through the slope heuristics). The model is parameterized to account for several characteristics of RNA-seq data, including: (1) a set of normalization factors (m_j) to account for systematic differences in library size among biological replicates, (2) a per-gene offset parameter (w_i) to account for differences among genes due to overall expression level,

and (3) a condition-specific cluster effect ($\lambda_{C(j)k}$). As the marginal sums of each gene are fixed in the model, variations in expression among experimental conditions may be modeled throughout the extremely large dynamic range typical of RNA-seq data. In particular, this parameterization enables a straightforward interpretation of the model, as $\tilde{\lambda}_{ck}$ corresponds to the proportion of reads attributed to condition c in cluster k. However, the processing time and memory requirements of `HTSCluster` reflect the fact that parameter estimation must be performed over a large set of models to enable model selection; one run of `HTSCluster` (version 2.0.4) took about about 2 hours with 1800 MB of memory for the fly developmental data[1].

Finally, we have applied this method to a set of miRNA-seq data from divergently selected chickens produced in the PSGen and GIS (Genome, Immunity, and Health) teams in the GABI research unit (Endale Ahanda, 2014) to identify groups of stress-responsive circulating extra-cellular microRNAs in plasma that exhibited similar patterns across lines and feeding conditions. More recently, we also used this same method to identify co-expression modules from crop and wild tomato plants (Sauvage et al., 2017).

3.3 Clustering transformed RNA-seq profiles

This section corresponds to the following published and submitted articles:

Rau, A. and Maugis-Rabusseau, C. (2017) Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*, bbw128. doi: 10.1093/bib/bbw128.

Godichon-Baggioni, A., Maugis-Rabusseau, C. and Rau, A. (2017) Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. arXiv:1704.06150.

The latter is the result of the post-doctoral work of Antoine Godichon-Baggioni, co-supervised by Cathy Maugis-Rabusseau and myself.

3.3.1 Background and motivation

The work described in the previous section was a primary motivation leading to the ANR-JCJC grant "Mixture-based procedures for the statistical analysis of RNA-seq data" (MixStat-Seq; 2014-2108, coordinated by Cathy Maugis-Rabusseau, INSA/IMT Toulouse), in which I am a work package leader. In particular, although our proposed Poisson mixture model has the advantage of directly modeling the count nature and variable library sizes of RNA-seq data, it has several serious limitations: (1) the assumption of conditional independence among samples, given the clustering group, is likely to be unrealistic for the vast majority of RNA-seq datasets; (2) per-cluster correlation structures cannot be included in the model; and (3) the Poisson distribution is likely overly restrictive, as it imposes an assumption of equal means and variances. In addition, classical asymptotic model selection criteria such as the BIC and ICL were often observed to have poor behavior for the Poisson mixture model; to deal with this, the method described in the previous section instead used a non-asymptotic penalized model selection criterion calibrated by the slope heuristics. This approach requires a collection of mixture models to be fit for a very wide range of cluster numbers K; for large

[1] All analyses were run on a Dell Latitude E6530 quad-core 2.70 GHz Intel(R) Core(TM) with 10GB RAM, running a 64-bit version of Windows 7 Professional.

K, this can imply significant computational time as well as practical difficulties for parameter initalization and estimation.

To address the aforementioned limitations of the Poisson mixture model, in this work we investigate appropriate transformations to facilitate the use of Gaussian mixture models for RNA-seq co-expression analysis. This strategy has the notable advantage of enabling the estimation of per-cluster correlation structures, as well as drawing on the extensive theoretical justifications of Gaussian mixture models (McLachlan and Peel, 2000). Law (2014) employed a related strategy for the differential analyses of RNA-seq data by transforming data, estimating precision weights for each feature, and using the `limma` empirical Bayes analysis pipeline (Smyth, 2004). The identification of an "appropriate" transformation for RNA-seq co-expression is not necessarily straightforward, and depends strongly on the desired interpretability of the resulting clusters as well as the model assumptions.

3.3.2 Gaussian mixture models for transformed RNA-seq profiles

Several data transformations have been suggested for RNA-seq data, most often in the context of exploratory or differential analyses. These include a $\log$ transformation (where a small constant is typically added to read counts to avoid 0's), a variance-stabilizing transformation (VST; Tibshirani, 1988; Huber, 2003; Anders and Huber, 2010), moderated log counts per million (CPM; Law, 2014), and a regularized log-transformation (rlog; Love et al., 2014). These transformations were proposed with the objective of rendering the data homoskedastic (in the case of the VST or rlog) or to reduce the orders of magnitude spanned by untransformed RNA-seq data. Rather than making use of these transformations, we propose calculating the normalized expression *profiles* for each feature, that is, the proportion of normalized reads observed for gene i with respect to the total observed for gene i across all samples:

$$p_{ij} = \frac{\tilde{y}_{ij} + 1}{\sum_{\ell} \tilde{y}_{i\ell} + 1},$$

where a constant of 1 is added to the numerator and denominator due to the presence of 0 counts. As before, the interest of these normalized expression profiles for co-expression analysis is illustrated in Figure 3.1. We note that using these normalized expression profiles for co-expression analysis is somewhat analagous to the parametrization of our previous Poisson mixture model defined in Equation (3.8), where the $\boldsymbol{\lambda}_k$ parameters could be interpreted as the proportion of counts (weighted by relative library sizes) attributed to each experimental condition for each gene assigned to cluster k.

Transformations for normalized expression profiles

It is important to note that the profile for gene i, $\mathbf{p}_i = (p_{ij})$, represents compositional data (Aitchison, 1986), as it is a q-tuple of nonnegative numbers whose sum is 1 that can be represented in the simplex of q parts:

$$\mathcal{S}^q := \left\{ \mathbf{p}_i = (\mathbf{p}_{i1}, ..., \mathbf{p}_{iq}) \in \mathbb{R}^q | \sum_{j=1}^{q} p_{ij} = 1, p_{ij} > 0, \forall i, j \right\}.$$

This means that the vector of values $\mathbf{p}_i$ are linearly dependent, which imposes constraints on the covariance matrices Σ_k that can be problematic for the general Gaussian mixture model (and indeed for most standard statistical approaches).

For this reason, we consider two separate strategies:

1. **Apply a general transformation to break the unit sum constraint**. In particular, we focus on the logit and the arcsine (also referred to as the arcsine square root, or angular) transformations:

$$g_{\text{arcsin}}(p_{ij}) = \arcsin\left(\sqrt{p_{ij}}\right) \in [0, \pi/2], \text{ and} \tag{3.9}$$

$$g_{\text{logit}}(p_{ij}) = \log_2\left(\frac{p_{ij}}{1-p_{ij}}\right) \in (-\infty, \infty). \tag{3.10}$$

   Over a broad range of intermediate values of the proportions, the logit and arcsin transformations are roughly linearly related to one another. However, although both transformations tend to pull out the ends of the distribution of p_{ij} values, this effect is more marked for the logit transformation, meaning that it is more affected by smaller differences at the ends of the scale.

2. **Apply a compositional data transformation.** The centered log ratio (CLR) is commonly used for compositional data (Aitchison, 1986), and is defined as $CLR : \mathcal{S}^q \longrightarrow \mathbb{R}^d$ for all $\mathbf{p}_i \in \mathcal{S}^q$ by

$$\text{CLR}(\mathbf{p}_i) = \left(\ln\left(\frac{p_{i1}}{g(\mathbf{p}_i)}\right), \ldots, \ln\left(\frac{p_{iq}}{g(\mathbf{p}_i)}\right)\right), \tag{3.11}$$

   where $g(\mathbf{p}_i)$ is the geometric mean of $\mathbf{p}_i$. In this case, the transformed values belong to the hyperplane of $\mathbb{R}^d$ with normal vector $(1, \ldots, 1)$. Two other commonly used compositional data transformations, the additive log ratio (ALR), and isometric log ratio (ILR), yielded similar results to the CLR and are not discussed further here.

   For data of moderate dimension, when a large number of coordinates have very small proportions, the CLR transformation tends to be quite sensitive to small fluctuations close to zero. This can have a strong undesired effect on clustering results when a small number of observations have highly-specific profiles (e.g., for genes with condition-specific expression). To account for this phenomenon by giving more importance to coordinates with large relative values, we also proposed a novel extension of the CLR for compositional data called the Log Centered Log Ratio (LCLR). For all $\mathbf{p}_i \in \mathcal{S}^q$, the LCLR is defined by $\text{LCLR}(\mathbf{p}_i) = (\text{LCLR}(p_{i1}), \ldots, \text{LCLR}(p_{iq}))$, where for all j,

$$\text{LCLR}(p_{ij}) = \begin{cases} -\left[\ln\left(1 - \ln\left[p_{ij}/g(\mathbf{p}_i)\right]\right)\right]^2 & \text{if } p_{ij}/g(\mathbf{p}_i) \leq 1, \\ \left(\ln\left[p_{ij}/g(\mathbf{p}_i)\right]\right)^2 & \text{otherwise,} \end{cases} \tag{3.12}$$

   and $g(\mathbf{p}_i)$ is as before. The additional log term when $\frac{p_{ij}}{g(\mathbf{p}_i)} \leq 1$ accords less importance in the transformation to samples with relatively weak proportions, while the squared term facilitates the concentration of profiles close to the center of the simplex $\left(\frac{1}{q}, \ldots, \frac{1}{q}\right)$.

Given these two transformation strategies, we now turn our attention to the clustering model.

Gaussian mixture models and the K-means algorithm

We consider a collection of Gaussian mixtures, defined as $(\mathcal{S}_m)_{m\in\mathcal{M}} = (\mathcal{S}_{(K,v)})_{(K,v)\in\mathcal{M}}$, where

$$\mathcal{S}_{(K,v)} = \left\{ f\left(.|\mathbf{\Psi}_{(K,v)}\right) = \sum_{k=1}^{K} \pi_{k,v}\phi\left(.|\mu_k, \Sigma_{k,v}\right) \right\}, \tag{3.13}$$

with $\phi(.|\mu_k, \Sigma_{k,v})$ denoting the q-dimensional Gaussian density with mean μ_k and covariance matrix $\Sigma_{k,v}$. The index v denotes one of the Gaussian mixture shapes obtained by constraining one or more of the parameters in the following decomposition of each mixture component variance matrix:

$$\Sigma_k = \gamma_k D_k' A_k D_k, \tag{3.14}$$

where $\gamma_k = |\Sigma_k|^{1/q}$, D_k is the eigenvector matrix of Σ_k, and A_k is the diagonal matrix of normalized eigenvalues of Σ_k. Various constraints on these parameters respectively control the volume, orientation, and shape of the k^{th} cluster (Celeux and Govaert, 1995); by additionally allowing the proportions π_k to vary according to cluster or be equal for all clusters, we may define a collection of 28 parsimonious and interpretable mixture models. Without loss of generality, for simplicity of notation we will consider here only the most general model form, with variable proportions, volume, orientation, and shape (referred to as the $[p_k L_k C_k]$ in `Rmixmod`); as such, the model collection is defined solely over a range of numbers of clusters, $(\mathcal{S}_K)_{K\in\mathcal{M}}$. The parameters of each model $\mathcal{S}_K$ in the collection defined in (3.13) may be estimated using a standard EM algorithm (Dempster et al., 1977). After solving the density estimation problem, for each model in the collection f is estimated by $\hat{f}_K = f(.|\hat{\mathbf{\Psi}}_K)$, and model selection may be performed using the BIC, ICL, or slope heuristics defined in Equations (3.4)-(3.7).

In addition to the Gaussian mixture model described in Equation (3.13), we will also consider the K-means algorithm (MacQueen, 1967) as an easily computable and fast alternative. Briefly, for a set of q dimensional points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, let $\mathcal{P}^{(K)} = \{P_k, k = 1, \ldots, K\}$ be a partition of the n observations into K clusters, and $\boldsymbol{\mu}_k$ be the mean of the cluster P_k:

$$\boldsymbol{\mu}_k := \frac{1}{|P_k|} \sum_{i \in P_k} \mathbf{x}_i,$$

where $|P_k|$ is the cardinality of cluster k. Using the usual Euclidean norm $\|.\|_2$, the aim of K-means is to minimize the sum of squared errors (SSE), defined for each set of clusters $\mathcal{P}^{(K)}$ by

$$\text{SSE}\left(\mathcal{P}^{(K)}\right) := \sum_{k=1}^{K} \sum_{i \in P_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2,$$

with $i \in P_k$ if $\|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2 = \min_{k'=1,\ldots,K} \|\mathbf{x}_i - \boldsymbol{\mu}_{k'}\|_2$. Note that there is in fact a strict equivalence between the K-means algorithm and a uniform spherical Gaussian mixture model with equal cluster proportions estimated using the classification EM (CEM) algorithm (Celeux and Govaert, 1992).

Finally, although rarely done in practice, penalized criteria like the BIC and ICL may also be used to select among different models or transformations, as was suggested in a different context by Thomas et al. (2008) and more recently for RNA-seq data by Gallopin (2015). This is of great interest, as it removes the need for an arbitrary choice of data transformation by using the framework of formal model selection. We illustrate this principle for the choice of number of clusters K and data transformation; in a more general case, a similar procedure could be used to additionally choose among the different forms of Gaussian mixture models described in Equation (3.14) or among different parametric forms of models. Let $g(\mathbf{x})$ represent an arbitrary monotonic transformation of a dataset $\mathbf{x}$. If the new sample $g(\mathbf{x})$ is assumed to arise from an i.i.d. Gaussian mixture density, $f(.|\mathbf{\Psi}_K)$, then the initial data $\mathbf{x}$ is an i.i.d. sample from density $f_g(.|\mathbf{\Psi}_K)$, which is a transformation of $f(.|\mathbf{\Psi}_K)$ and thus not necessarily a Gaussian mixture density. If J_g denotes the Jacobian of the transformation g and $\hat{\mathbf{\Psi}}_{(K,g)}$ the maximum likelihood estimate obtained for the model with K clusters and

transformation g, we select the pair (K, g) leading to the minimum of the corrected BIC or ICL criteria:

$$\begin{aligned} \text{BIC}^*(K,g) &= -\log f(\mathbf{y}; K, \hat{\boldsymbol{\Psi}}_{(K,g)}) - \frac{\nu_K}{2}\log(n)) - \log\left(\det(J_g)\right) \\ \text{ICL}^*(K,g) &= \text{BIC}^*(K,g) + \text{Entropy}(K). \end{aligned} \tag{3.15}$$

where $\text{Entropy}(K)$ is as defined in Equation (3.6). Note that in these expressions, the number of parameters ν_K does not depend on the transformation g. In order to compare the arcsine and logit transformations, we must thus calculate the log determinant of each transformation:

$$\log\left(\det(J_{\text{arcsin}})\right) = -nq\ln(2) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q}\log\left(p_{ij}(1-p_{ij})\right),$$

$$\log\left(\det(J_{\text{logit}})\right) = -nq\ln[\ln(2)] - \sum_{i=1}^{n}\sum_{j=1}^{q}\log\left(p_{ij}(1-p_{ij})\right).$$

The corrected ICL criteria $\text{ICL}^*(K, \text{arcsin})$ and $\text{ICL}^*(K, \text{logit})$ can thus be directly compared to choose between the arcsine and logit transformations.

3.3.3 Data application with `coseq`

We implemented the strategies described above in the `coseq` (**co**-expression of RNA-**seq** data) package, available as part of the Bioconductor project. We illustrate the use of `coseq` using RNA-seq data from a study of the expansion of three regions of the neocortex (ventricular zone [VZ], subventricular zone [SVZ], and cortical plate [CP]) in five embryonic mice (Fietz, 2012). Raw read counts for this study were downloaded on December 23, 2015 from the Digital Expression Explorer (DEE) (Ziemann, 2015) using the associated SRA accession number SRP013825, and run information was downloaded using the SRA Run Selector.

A typical call to `coseq` to fit a Gaussian mixture model on arcsine- or logit-transformed normalized profiles takes the following form:

```
> library(coseq)
> data(fietz)
> counts <- exprs(fietz)
> conds <- pData(fietz)$tissue
> run_arcsin <- coseq(counts, K=2:20, model="Normal",
+    transformation="arcsin")
> run_logit <- coseq(counts, K=2:20, model="Normal",
+    transformation="logit")
```

where `counts` represents a $(n \times q)$ matrix or data frame of read counts for n genes in q samples, and `K=2:20` provides the desired range of numbers of clusters (here, 2 to 20). This function directly calls the `Rmixmod` R package to fit Gaussian mixture models (Lebret, 2015). For backwards compatibility with our previous method (Rau, 2015), a similar function call may be used to fit a Poisson mixture model on raw counts using the `HTSCluster` package:

```
> run_pois <- coseq(counts, conds, K=2:20, model="Poisson")
```

where a vector `conds` is additionally provided to identify the experimental condition associated with each column in `counts`. In all cases, the output of the `coseq` function is

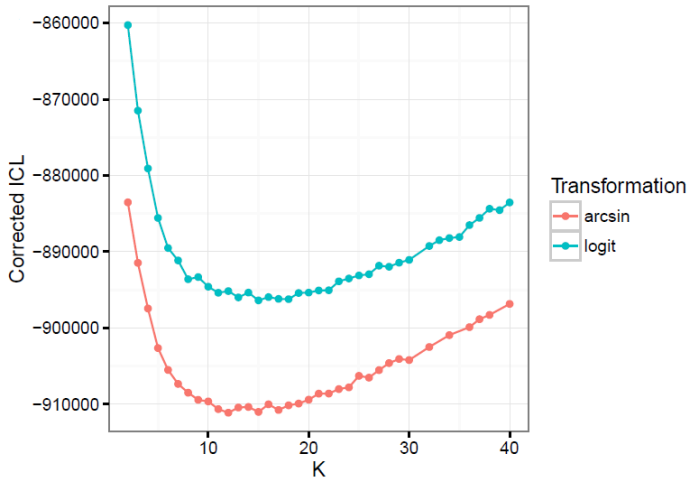

FIGURE 3.3: Corrected ICL values for the arcsine (red) and logit (blue) transformed normalized expression profiles over a range of numbers of clusters K for the Fietz (2012) mouse RNA-seq data.

an S4 object of class `coseqResults` (an extension of the `SummarizedExperiment0` Bioconductor S4 class) on which standard `plot` and `summary` functions can be directly applied; the former uses functionalities from the `ggplot2` package (Wickham, 2009). The option of parallelization via the `BiocParallel` Bioconductor package is provided, and several additional options (filtering, normalization, `Rmixmod` options) are available.

For these data, the models selected using the ICL are $\hat{K} = 12$ and $\hat{K} = 15$ for the arcsine and logit transformations, respectively. By comparing the corrected ICL defined in Equation (3.15) between these two transformations using the convenience function `compareICL`, it may be seen that in this case, the arcsine transformation is preferred (see Figure 3.3). We focus our attention on this model in the following discussion.

```
> compareICL(list(run_arcsin, run_logit))
```

A visualization of the per-cluster expression profiles and diagnostic plots can be obtained using a simple `plot` command (see Figure 3.5):

```
> plot(run_arcsin, graphs="boxplots",
+      conds=conds, average_over_conds=TRUE)
> plot(run_arcsin, order=TRUE,
+      graphs=c("probapost_boxplots","probapost_barplots"))
```

Note that the output of our `plot` function is a `ggplot2` object which can be further modified by the user (e.g., to change color schemes, add titles, change labels, etc).

One advantage of the Gaussian mixture model is that it enables an investigation of per-cluster covariance structures. It is interesting to note that although the Gaussian mixture model does not explicitly incorporate the experimental condition labels $\mathcal{C}(j)$, the estimated models include large cluster-specific correlations among replicates within each tissue (Figures 3.4A and 3.4B). In addition, cluster-specific correlation structures among regions may be clearly seen; for example, Cluster 2 is characterized by very large negative correlations between the CP and SVZ/VZ regions, while Cluster 3 instead has a strong negative correlation between the VZ and CP/SVZ regions. This strongly suggests that in these data, the assumption of conditional independence among samples assumed by the Poisson mixture model described in Rau (2015) is indeed unrealistic.

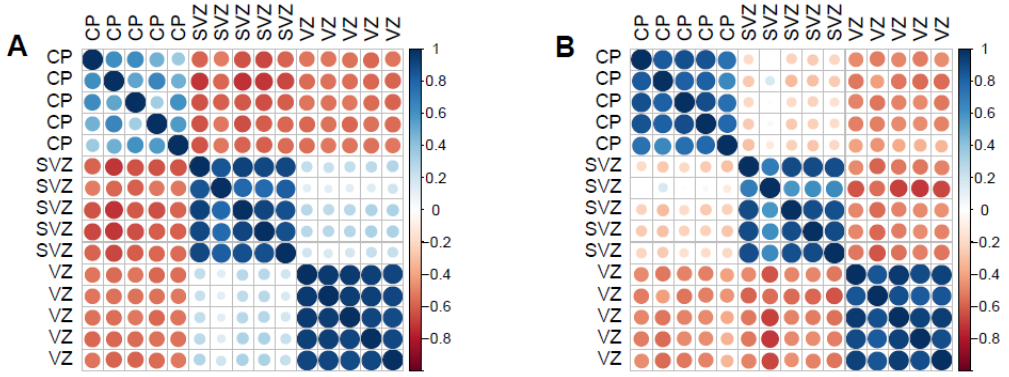

FIGURE 3.4: Per-cluster correlation matrices for clusters 2 (A) and 3 (B) from the Fietz (2012) mouse data. Dark blue and red represent correlations close to 1 and -1, respectively, and circle areas correspond to the absolute value of correlation coefficients. Correlation matrices are visualized using the `corrplot` R package.

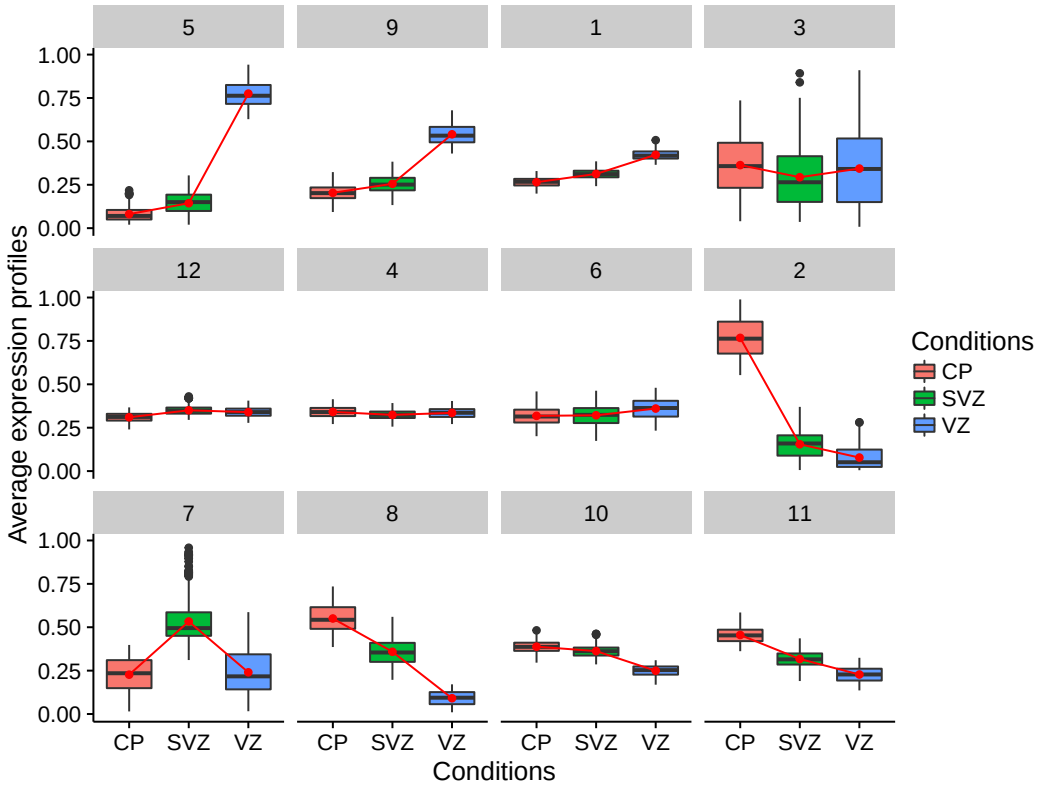

FIGURE 3.5: Per-cluster expression profiles for the Fietz (2012) data. Clusters have been sorted so that those with similar mean vectors (as measured by the Euclidean distance) are plotted next to one another. Connected red lines correspond to the mean expression profile for each group.

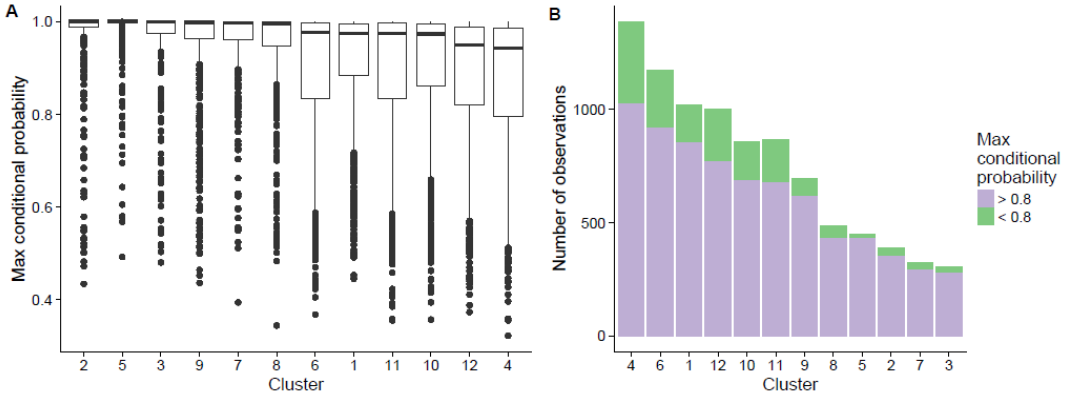

FIGURE 3.6: Evaluation of clustering quality for the Fietz (2012) mouse data. (A) Maximum conditional probabilities $\tau_{\max}(i)$ for each cluster, sorted in decreasing order by the cluster median. (B) Barplots of cluster sizes, according to $\tau_{\max}(i)$ greater than or less than 0.8, sorted according to the number of genes with $\tau_{\max}(i) > 0.8$.

An additional advantage of model-based clustering approaches is that they facilitate an evaluation of the clustering quality of the selected model by examining the maximum conditional probabilities of cluster membership for each gene $\tau_{\max}(i)$:

$$\tau_{\max}(i) = \max_{1 \leq k \leq \hat{K}} \tau_{ik}\left(\hat{\theta}_{\hat{K}}\right), \; i = 1, \ldots, n.$$

Boxplots of the maximum conditional probabilities $\tau_{\max}(i)$ per cluster for the Fietz (2012) mouse data are presented in Figure 3.6. It may be seen that across clusters, the majority of genes in both datasets have a large value (i.e., close to 1) for $\tau_{\max}(i)$; in this case, the number of genes with $\tau_{\max}(i) > 0.8$ is 7382 (82.4%). However, the boxplots also illustrate that some genes have a $\tau_{\max}(i)$ less than this threshold, in some cases as low as 0.4; this indicates that for a small number of genes, the cluster assignment is fairly ambiguous and assignment to a single cluster is questionable (the gene with the smallest $\tau_{\max}(i)$ in the Fietz (2012) mouse data had a conditional probability of 24.8%, 32.2%, 13.0% and 30.0% of belonging to clusters 1, 4, 6, and 12, respectively). In such cases, it may be prudent to focus attention on genes with highly confident cluster assignments (e.g., those with $\tau_{\max}(i) > 0.8$).

Finally, as described in the previous section, a fast and simple alternative to a Gaussian mixture model is the K-means algorithm, if per-cluster covariance matrices can be assumed to be of the form $\Sigma_k = \sigma^2 I$. In addition, in cases where highly-specific profiles may be expected (e.g., in developmental data, where some genes may be active during only a portion of developmental stages), transformations specifically tailored for compositional data, such as the CLR and LCLR in Equations (3.11)-(3.12), may be more appropriate choices:

```
> run_LCLR <- coseq(counts, K=2:20, model="kmeans",
+                   transformation="logclr")
```

As RNA-seq expression analyses are often performed on a subset of genes identified as differentially expressed, the `coseq` function can also be directly called on an `DESeqResults` S4 object or integrated with `DGELRT` S4 objects, respectively corresponding to output from the `DESeq2` (Love et al., 2014) and `edgeR` (Robinson et al., 2010) Bioconductor packages for RNA-seq differential analyses. We illustrate this using the `DESeq2` pipeline below:

```
> library(DESeq2)
> dds <- DESeqDataSetFromMatrix(counts,
+             DataFrame(group=factor(conds)), ~group)
> dds <- DESeq(dds, test="LRT", reduced = ~1)
> res <- results(dds)
> run <- coseq(dds, K=2:15, model="kmeans", alpha=0.05)
```

3.3.4 Conclusions and discussion

In this work, we addressed the choice of transformed normalized expression profiles rather than raw counts for RNA-seq co-expression analysis under the framework of Gaussian mixture models. We focused the majority of our discussion here on the use of (arcsine- or logit-transformed) normalized profiles to identify groups of co-expressed genes using Gaussian mixture models or the K-means algorithm. Gaussian mixtures in particular represent a rich, flexible, and well-characterized class of models that have been successfully implemented in a large variety of theoretical and applied research contexts. For RNA-seq data, this means that the model may directly account for per-cluster correlation structures among samples, which can be quite strong in RNA-seq data. We also illustrated the use of penalized criteria like the ICL and BIC to objectively compare results between different monotonic transformations, and potentially among different forms of Gaussian covariance matrices or among different models.

Several practical issues should be considered in co-expression analyses. First, a common question is whether genes should be screened prior to the analysis (e.g., via an upstream differential analysis or filter based on the mean expression or coefficient of variation for each gene). Such a screening step is often used in practice, as genes contributing noise but little biological signal of interest can adversely affect clustering results. A second common question pertains to whether replicates within a given experimental group should be modeled independently or summed or averaged prior to the co-expression analysis. Although technical replicates in RNA-seq data are typically summed prior to analysis, in this work we fit Gaussian mixture models on the full data including all biological replicates; subsequently to visualize clustering results, replicate profiles are summed for improved clarity of cluster profiles.

Finally, many alternative clustering strategies exist based on different algorithms (e.g., K-means and hierarchical clustering), distance measures calculated among pairs of genes (e.g., Euclidean distance, correlation, etc), and techniques for identifying the number of clusters (e.g., the Dynamic Tree Cut method for dendrograms (Langfelder and Horvath, 2008)). The difficulty of comparing clusterings arising from different approaches is well-known, and it is rarely straightforward to establish the circumstances under which a given strategy may be preferred. Following a co-expression analysis, it is notoriously difficult to validate the results of a clustering algorithm on transcriptomic data, and such results can be evaluated based on either statistical criteria (e.g., between-group and within-cluster inertia measures) or external biological criteria. In practice groups of co-expressed genes are further characterized by analyzing and integrating various resources, such as functional annotation or pathway membership information from databases like the Gene Ontology Consortium. Such functional analyses can be useful for providing interpretation and context for the identified clusters.

3.4 Annotation-based model selection

This section corresponds to the following published article:

Gallopin, M., Celeux, G., Jaffrézic, F., Rau, A. (2015) A model selection criterion for model-based clustering of annotated gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 14(5): 413-428.

This work is the result of the Ph.D. work of Mélina Gallopin, co-supervised by Florence Jaffrézic, Gilles Celeux, and myself.

3.4.1 Background and motivation

Genome annotation broadly refers to the set of meta-data associated with the coding regions in the genome, typically including the identification of the location of each gene as well as a determination of the functions related to the gene product (e.g., protein or RNA). In particular, gene annotations correspond to known functions related to the gene product, including molecular functions, biological pathways, or the cellular location of the gene products. A variety of well-known unified databases have been constructed with known functional annotations collected from bibliographic sources across species, including the Gene Ontology (GO) (Ashburner et al., 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) or the MSigDB (Molecular Signatures) databases (Liberzon, 2011). Although such databases contain a rich source of functional information about the genome in a large variety of species (e.g., *Arabidopsis thaliana*, human, rat, mouse, fly), our knowledge of functional annotations is often far from complete (Tipney and Hunter, 2010).

In practice, annotation databases are often used to perform *a posteriori* validation and interpretation of co-expressed gene clusters through tests of functional enrichment (Steuer et al., 2006). Such functional annotation may instead be directly integrated into the clustering model itself. For example, Tari et al. (2009) incorporate GO annotations as prior knowledge in a fuzzy c-means clustering. Verbanck et al. (2013) proposed a clustering approach based on a distance defined conjointly on the similarity among expression profiles and that among functional profiles. Pan (2006) and Huang and Pan (2006) proposed including gene annotation as prior information in a stratified mixture model. However, the inclusion of gene annotation directly in the model itself in this way may be questionable, particularly when they are also used to validate the gene clusters *a posteriori*. Moreover, as gene annotations tend to be incomplete, biases may be introduced if they are directly incorporated in the model, as unannotated genes (which represent those known to be unassociated with a given function as well as those of unknown function) may be erroneously separated from annotated genes.

One alternative to such approaches is to define a clustering model that accounts for external gene annotations without directly including them in the model itself. To this end model-based clustering provides a convenient framework, as it (1) allows for a large set of clustering models to be fit to the gene expression alone, and (2) facilitates the choice among this set a parsimonious model that simultaneously provides a good fit to the data and coherence with the external gene annotations. In this work, we address these points by proposing a model selection criterion that accounts for external gene annotations.

3.4.2 Integrated completed annotated likelihood model selection criterion

Baudry (2014) recently proposed an ICL-like criterion that takes advantage of the potential explicative ability of external categorical variables $\mathbf{u} = (\mathbf{u}^1, \ldots, \mathbf{u}^R)$ where $u^r_{i\ell} = 1$ indicates that the gene i is in category ℓ for the r^{th} external categorical variable and 0 otherwise. The idea is to choose a classification $\mathbf{z}$ based on $\mathbf{y}$ that is coherent with $\mathbf{u}$. Assuming that $\mathbf{y}$ and $\mathbf{u}$ are conditionally independent given $\mathbf{z}$, the Supported Integrated Completed Likelihood (SICL) criterion is an asymptotic approximation of the logarithm of the integrated completed likelihood:

$$f(\mathbf{y}, \mathbf{u}, \mathbf{z}; K) = \int f(\mathbf{y}, \mathbf{u}, \mathbf{z}; K, \boldsymbol{\Psi}_K)\pi(\boldsymbol{\Psi}_K)d\boldsymbol{\Psi}_K.$$

The SICL criterion is defined as follows:

$$\text{SICL}(K) = \text{ICL}(K) - \sum_{r=1}^{R}\sum_{\ell=1}^{U_r}\sum_{k=1}^{K} n^r_{k\ell} \log \frac{n^r_{k\ell}}{n_{k\cdot}}, \tag{3.16}$$

where U_r is the number of levels of the variable $\mathbf{u}^r$,

$$n^r_{k\ell} = \text{card}\big\{i : z_{ik} = 1 \text{ and } u^r_{i\ell} = 1\big\},$$

$n_{k\cdot} = \sum_{\ell=1}^{U_r} n^r_{k\ell}$, and $\text{ICL}(K)$ is as described in Equation (3.5). The last additional term in Equation (3.16) quantifies the strength of the link between the categorical variables $\mathbf{u}$ and the classification $\mathbf{z}$.

The objective of this work is to make use of external gene annotations to choose a model for which clusters may be meaningfully interpreted both with respect to their expression profiles and the functional properties associated with a subset of genes. Since gene annotations are binary variables (i.e., a gene is either annotated or unannotated), it may seem natural to directly use the SICL defined in Equation (3.16). However, in contrast to the situation considered by Baudry (2014), gene annotation information is often incomplete. More precisely, for each of the G annotation terms, indexed by g, the available information $\mathbf{u}^g$ is as follows:

$$u^g_i = \begin{cases} 1 & \text{if gene } i \text{ is known to be implicated in function } g, \\ 0 & \text{if gene } i \text{ is not known to be implicated in function } g. \end{cases}$$

Note that $u^g_i = 0$ can indicate that information is missing (i.e., gene i has not yet been identified for annotation g) or that gene i is known to be unrelated to annotation g. As such, $u^g_i = 0$ does not represent the null level of variable and thus represents an incomplete binary variable. For this reason, the SICL criterion is not an appropriate measure of the link between an external annotation $\mathbf{u}^g$ and a classification $\mathbf{z}$, and a specific criterion must be defined to incorporate the gene annotation information into the model selection step. To this end, we propose a novel model selection criterion as follows.

For each gene annotation $\mathbf{u}^g$ ($g = 1, \ldots, G$), we first define the random matrix $\mathbf{b}^g$ of latent variables indicating the allocation of the annotations among the K clusters:

$$b^g_{ik} = \begin{cases} 1 & \text{with probability } p^g_k \text{ if } u^g_i = 1, \\ 0 & \text{if } u^g_i = 0. \end{cases} \tag{3.17}$$

Each row of the matrix $\mathbf{b}^g$ is a random vector following a multinomial distribution with parameters u^g_i and $(p^g_1, \ldots, p^g_K)$ if $u^g_i > 0$, and is the null vector $\mathbf{0}$ if $u^g_i = 0$. Our integrated completed annotated likelihood (ICAL) criterion seeks to select the clustering model that

minimizes the negative logarithm of the integrated annotated likelihood:

$$\log f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1, \ldots, \mathbf{b}^G; K) = \log \int f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1, \ldots, \mathbf{b}^G; K, \mathbf{\Psi}_K)\pi(\mathbf{\Psi}_K)d\mathbf{\Psi}_K. \quad (3.18)$$

Assuming that $\mathbf{b}^1, \ldots, \mathbf{b}^G$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}$, the conditional distribution of each $\mathbf{b}^g$ given $\mathbf{z}$ does not depend on $\mathbf{y}$ or the mixture parameters. Thus as $f(\mathbf{b}^g \mid \mathbf{y}, \mathbf{z}; K, \theta_K) = f(\mathbf{b}^g \mid \mathbf{z}; K)$ for all g, we have

$$\begin{aligned} -\log f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1, \ldots, \mathbf{b}^G; K) = &-\log f(\mathbf{b}^1, \ldots, \mathbf{b}^G; \mathbf{z}, K) - \\ &\log \int f(\mathbf{y}, \mathbf{z}; K, \mathbf{\Psi}_K)\pi(\mathbf{\Psi}_K)d\mathbf{\Psi}_K. \end{aligned} \quad (3.19)$$

The last term in Equation (3.19) can be approximated with ICL(K) from Equation (3.5). Assuming in addition that $\mathbf{b}^1, \ldots, \mathbf{b}^G$ are independent and that gene annotations are missing at random, we can write

$$f(\mathbf{b}^1, \ldots, \mathbf{b}^G; \mathbf{z}, K) = \prod_{g=1}^{G} f(\mathbf{b}^g | \mathbf{z}, K), \quad (3.20)$$

and the first term may thus be approximated using

$$\log f(\mathbf{b}^g \mid \hat{\mathbf{z}}, K) = \sum_{k=1}^{K} n_k^g \log \frac{n_k^g}{n^g},$$

where $n^g = \text{card}\{i : u_i^g = 1\}$ and $n_k^g = \text{card}\{i : \hat{z}_{ik} = 1 \text{ and } u_i^g = 1\}$. leading to the generalized Integrated Completed Annotated Likelihood (ICAL) criterion:

$$\text{ICAL}(K) = \text{ICL}(K) - \sum_{g=1}^{G}\sum_{k=1}^{K} n_k^g \log \frac{n_k^g}{n^g}. \quad (3.21)$$

Finally, if the uncertainty associated with $u_i^g = 0$ (i.e., that gene i could either be unassociated with function g or that this information is missing) is ignored, it can be shown that our ICAL criterion can be rewritten as a function of the SICL criterion proposed by Baudry (2014):

$$\text{ICAL}(K) = \text{SICL}(K) - \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k0}^g \log n_{k0}^g + G\sum_{k=1}^{K} n_k \log n_k + \text{constant}, \quad (3.22)$$

where n_k represents the size of the cluster k. From Equation (3.22), we note that the SICL takes into account both modalities (0 and 1) of the external variables $\mathbf{u}$, while the ICAL discards the null modality (the $-\sum_{g=1}^{G}\sum_{k=1}^{K} n_{k0}^g \log n_{k0}^g$ term). Moreover, it can be seen that the ICAL penalises a large number of clusters, while the SICL does not (the $G\sum_{k=1}^{K} n_k \log n_k$ term). As such, the ICAL tends to select parsimonious models with a relatively small number of clusters, as compared to SICL. This means that the ICAL generally tends to merge clusters to group genes annotated for the same function, reducing the number of optimal clusters K with respect to the optimal number of clusters selected by ICL. SICL tends to split clusters in order to obtain clusters made up only of annotated genes, increasing the number of optimal clusters with respect to the optimal number of clusters selected by ICL.

The ICAL criterion is implemented in the R package `ICAL`, freely available on GitHub.

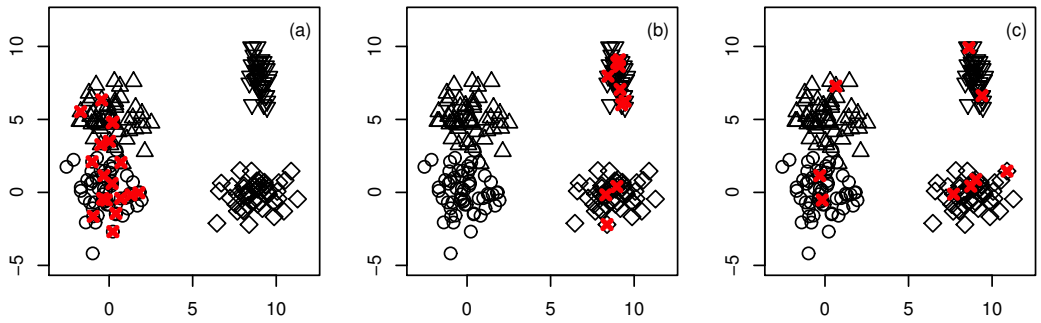

FIGURE 3.7: Illustration of a simulated dataset and three annotation patterns: associated annotation $\mathbf{u}_A$ (left), unassociated annotation $\mathbf{u}_B$ (center) and mixed annotations $\mathbf{u}_C$ (right). For each figure, the 200 observations are drawn from a mixture of Gaussian bivariate components: circles, triangles, inverted triangles and diamonds correspond to components 1, 2, 3 and 4. For each annotation type, the 20 annotated genes are represented by coloured bold crosses.

		K	1	2	3	4	5	6	7	8	9	10
		BIC			19	**81**	2					
		ICL			**54**	46						
Associated	$\mathbf{u}_A$	SICL			**53**	47						
		ICAL			**87**	13						
Unassociated	$\mathbf{u}_B$	SICL			**53**	47						
		ICAL			**53**	47						
Mixed	$\mathbf{u}_C$	SICL			49	**51**						
		ICAL			**79**	21						
Multiple	$\mathbf{u}_A, \mathbf{u}_B, \mathbf{u}_C$	SICL			48	**52**						
		ICAL			**97**	4						

TABLE 3.1: Number of simulated datasets for which each model ($K = 1, \ldots, 10$) was selected by BIC, ICL, SICL and ICAL for several external annotations over 100 independent simulated datasets. The model most commonly selected for each criterion is highlighted in boldface.

Illustration on simulated data

To illustrate this behavior, we simulated data from a mixture of four bivariate Gaussian distributions with $n = 200$ observations associated with different types of external functional annotations: $\mathbf{u}_A$, $\mathbf{u}_B$ and $\mathbf{u}_C$ (see Figure 3.7). The genes annotated for $\mathbf{u}_A$ are shared by the two closest mixture components. This annotation is designed to be *associated to the components* in the sense that it suggests the interest of merging the two clusters, as they share similar joint distributions and external annotations. The genes annotated for the second function $\mathbf{u}_B$ are shared only by the two clearly distinct components. This annotation is designed to be *unassociated with the components*: although the components share a similar function, their joint distributions are too distinct to be merged from a modeling point of view. Finally, the genes annotated for $\mathbf{u}_C$ are randomly spread over the four components, meaning the annotation is *mixed* (half associated / half unassociated).

Using the R package `Rmixmod` (Biernacki, 2006; Lebret, 2015), we estimated the parameters for models with the number of clusters K varying from 1 to 10 and subsequently performed model selection using the BIC, ICL, SICL, and ICAL to select the most appropriate number of clusters. Out of 100 simulated datasets (see Table 3.1), we found that (1) the BIC most often (81% of the datasets) selected the model with K=4 clusters, which corresponds to the true model used for simulation; (2) the ICL had some difficulty in determining whether 3 or 4 components should be preferred; (3) the SICL performed similarly to the ICL, as it tends to prefer smaller clusters containing only annotated genes (i.e., a high specificity of annotation within each cluster); (4) when relevant ($\mathbf{u}_A$), mixed ($\mathbf{u}_C$), or multiple ($\mathbf{u}_A, \mathbf{u}_B, \mathbf{u}_C$) annotations were included, the ICAL showed a strong preference for the model with $K = 3$ components, merging the close profiles that shared relevant annotations; and (5) when irrelevant ($\mathbf{u}_B$) annotations were included, the ICAL performed similarly to the ICL. This suggests that if the external information is associated to the components, even partially so, the use of the ICAL criterion improves model selection in terms of functional interpretability. If the external information is unassociated to the components, the ICAL criterion simply behaves like the ICL.

Illustration on RNA-seq data

The ICAL criterion described above was used to perform model selection for a co-expression analysis of RNA-seq data from three regions (the duodenum, the jejunum and the ileum) of the small intestine of four healthy piglets from Mach (2014). After an initial differential analysis, 4021 genes of interest were identified and normalized counts were log-transformed using the `voom` procedure from (Law, 2014). We also collected relevant annotations corresponding to the canonical pathways (CP) gene set collection from the Molecular Signatures Database (MSigDB) (Liberzon, 2011). Among the 1320 CP in the database, a total of 10 CPs of interest (Table 3.2) were found to be overrepresented in the set of differentially expressed genes (Fisher's exact test, adjusted p-value < 0.05 after Bonferroni correction) and were retained as relevant functional annotations.

After fitting Gaussian mixture models with the `Rmixmod` package (Lebret, 2015) for $K = 1, \ldots, 50$, we performed model selection with the ICL ($\hat{K} = 23$) and ICAL ($\hat{K} = 20$) criteria. Although the result of the latter is not perfectly nested in the former, in many cases the attribution of genes to clusters in the ICAL solution is a result of collapsing or partially collapsing several clusters from the ICL solution. We also examine associations between clusters and CP using Fisher's exact test for each of the selected models (see Table 3.3). The ICAL criterion yields a clustering that maximizes the number of genes annotated in each cluster for each CP while still only grouping genes that share sufficiently similar expression profiles. For example, we note that CP8 is associated with two different clusters in the ICL solution, while it is associated with a single cluster in the ICAL solution; similarly, CP10 is

CP	Name	DE	Total
1	Reactome metabolism of lipids and lipoproteins	141	480
2	Reactome transmembrane transport of small molecules	124	415
3	Reactome hemostasis	99	468
4	Reactome SLC mediated transmembrane transport	73	243
5	Reactome phospholipid metabolism	54	200
6	Reactome fatty acid triacylglycerol & ketone body metabolism	53	170
7	KEGG PPAR signaling pathway	34	71
8	KEGG ECM receptor interaction	34	86
9	Reactome transport of inorganic cations anions and amino acids oligopeptides	33	96
10	KEGG peroxisome	31	80

TABLE 3.2: Number of genes annotated for each canonical pathway (CP), among the 4021 differentially expressed (DE) genes and among the full CP gene set collection of the MSigDB database.

(a) **ICL solution**

	size	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP 8	CP9	CP10
Cluster 2	58		$\star$	$\star$	$\star$						
Cluster 5	203								$\star$		
Cluster 6	47										$\star\star$
Cluster 7	258	$\star$					$\star$				$\star$
Cluster 8	96					$\star\star$					
Cluster 10	287									$\star$	
Cluster 14	225										$\star\star$
Cluster 22	144			$\star\star$					$\star\star\star$		

(b) **ICAL solution**

	size	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP 8	CP9	CP10
Cluster 3	297									$\star$	
Cluster 5	379			$\star\star$					$\star\star\star$		
Cluster 6	156		$\star\star$	$\star$							
Cluster 7	92					$\star$					
Cluster 10	267	$\star$					$\star\star$				$\star\star$
Cluster 17	235										$\star\star$

TABLE 3.3: Table of associations between clusters and CP for the ICL solution (a) and the ICAL solution (b). Associations are detected using Fisher's exact tests: the number of stars indicates the value of the p-value ($\star$ below 0.01, $\star\star$ below 0.001, $\star\star\star$ below 0.0001).

associated with three clusters in the ICL solution and only two clusters in the ICAL solution. On the other hand, although clusters 10 and 17 in the ICAL solution both share annotations for CP10, these clusters are not collapsed into one using the proposed criterion, as their expression dynamics are too different. As such, the ICAL solution appears to enable the identification of more biologically interpretable clusters than the ICL, while still ensuring that the clustered genes share sufficiently similar expression dynamics.

3.4.3 Conclusions and discussion

In this work, we presented a novel way to incorporate functional annotations into model-based clustering of gene expression data using the ICAL criterion, which is designed to select the model that jointly maximises the goodness-of-fit to the data and the association of clusters and annotations. From a biological point of view, ICAL aims to select models with more interpretable clusters than those selected by BIC or ICL. It is important to note that the functional annotations are not directly included in the clustering model and are only used to select the best model. This approach is a good compromise between two opposite strategies: including functional annotations directly in the clustering model (Morlini, 2011) or excluding them altogether and using them only to validate clusters *a posteriori*. Since we do not include annotations in the clustering model, we detect associations between annotations and clusters with a stronger evidence than if we had included the external annotations in the clustering model. In particular, the ICAL criterion is a good way to include prior biological expertise without according it too much importance, which can provide a good balance between what can be observed in the data and what experts expect to see in the data.

Chapter 4

Inferring gene regulatory networks from expression data

4.1 Overview of gene regulatory networks

High-throughput assays like microarrays and RNA-seq may be used to study the coordinated behavior of genes during specific biological processes, such as the cell cycle or a response to an external input, often with the goal of identifying and understanding gene regulatory networks. Inference of gene networks from transcriptomic data is indeed a key aspect of systems biology that may help unravel and better understand the underlying biological regulatory mechanisms. Abstractly speaking, a gene regulatory network (GRN) can be described as the direct and indirect interactions that occur among a collection of interconnected genes (Figure 4.1, top). As these interactions regulate gene transcription and the subsequent production of functional proteins, the identification of these networks can lead to a better understanding of complex biological systems. Graphs are often used as an abstraction to visualize these networks, where nodes represent genes and edges represent interactions among the genes (Figure 4.1, bottom right).

Using high-throughput measurements of gene expression, taken over time or following experimental interventions, we aim to infer (or "reverse-engineer") the structure of GRN involved in a particular cellular process. However, these networks are generally very complicated and difficult to elucidate, particularly given the large number of genes considered (and thus, the large number of potential parameters to be estimated), the typically small number of biological replicates, the assumed sparsity of such networks, and the complexity inherent to biological networks. In this chapter, I will focus on two of our contributions to gene regulatory network (GRN) inference from expression data: (1) a hierarchical Poisson log-normal model specifically designed for inference from RNA-seq data; and (2) an approach to infer causal relationships among genes from intervention gene expression data.

4.2 Network inference for observational RNA-seq data

This section corresponds to the following published article:

Gallopin, M. Rau, A., and Jaffrézic, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. PLoS One 8(10): e77503.

This work is the result of the Ph.D. work of Mélina Gallopin, co-supervised by Florence Jaffrézic, Gilles Celeux, and myself.

Similarly to differential and co-expression analayses, it is somewhat of an open question as to whether methods developed for the inference of GRN from microarray data are

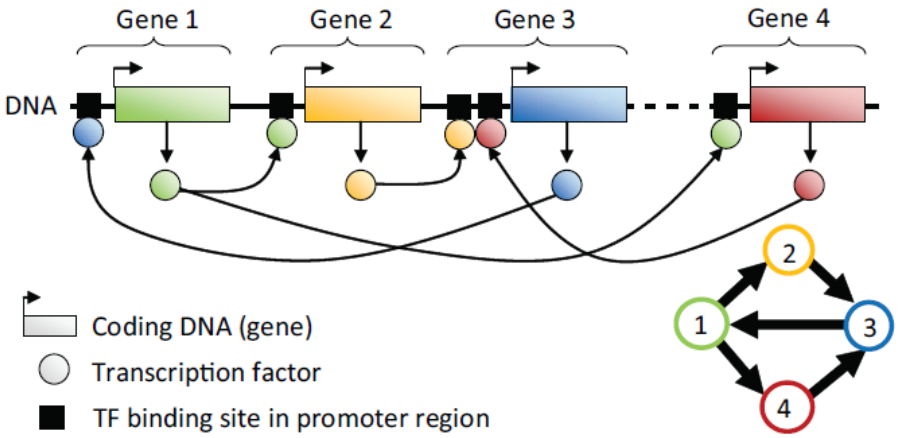

FIGURE 4.1: A schematic illustration of a simple gene regulatory network made up of four genes. Each gene is transcribed and translated into a transcription factor protein, which in turn regulates the expression of other genes in the network by binding to their respective promoter regions. The gene regulatory network may be represented using the graph in lower right corner, made up of four nodes (genes) and five edges (interactions among the genes).

appropriate for RNA-seq data. In order to identify relationships among genes from microarray data, several authors have proposed the use of co-expression networks based on Pearson correlation (Giorgi et al., 2013) or canonical correlation (Hong, 2013; Iancu, 2012), or alternatively based on Gaussian graphical models (Friedman et al., 2008; Meinshausen, 2006; Cai et al., 2012). However, at the time of this work, no specific models had been proposed for RNA-seq data.

Our goal here was thus to investigate three different strategies that could be used for this purpose: (1) apply a appropriate data transformation, using for example a Box-Cox transformation (Box and Cox, 1964), and subsequently use a Gaussian graphical model; (2) use a power transformation (Li, 2012) in conjunction with a log-linear Poisson graphical model (Allen and Liu, 2012) specifically designed to model count data; or (3) use a hierarchical log-normal Poisson graphical model specifically designed to account for overdispersed count data. In the three aforementioned strategies, lasso penalities (Tibshirani, 1996) are used to obtain to obtain a sparse respresentation of the network. We briefly describe the three approaches in the following.

- **Gaussian graphical model**.The underlying assumption of this model is that the data are normally distributed. In the case of untransformed RNA-seq data, this assumption is not valid since data counts cannot take negative values. We investigated a variety of Box-Cox transformations to lead to approximately normal data (Box and Cox, 1964), where the δ value was chosen to maximize the log-likelihood of the transformed data:

$$y_{ij} \to f(y_{ij}) = \begin{cases} \dfrac{(y_{ij}+1)^{\delta}-1}{\delta}, & \text{if } \delta \neq 0, \\ \log(y_{ij}+1), & \text{if } \delta = 0, \end{cases}$$

  where a constant of 1 has been added due to zero counts. Let $\mathbf{z}_j = (f(y_{1j}), ..., f(y_{nj}))$ be the transformed vector of expression values for n genes for the jth biological sample. We assume that $\mathbf{z}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To ensure the estimation of a sparse network (a

common assumption in this context), we consider the lasso-penalized log-likelihood:

$$L^{\lambda}(\mathbf{\Sigma^{-1}}) = -2\log(\det(\mathbf{\Sigma^{-1}})) - \text{trace}(\boldsymbol{S}\mathbf{\Sigma^{-1}}) + \lambda\|\mathbf{\Sigma^{-1}}\|_{\ell_1}, \tag{4.1}$$

where $\boldsymbol{S}$ is the empirical covariance matrix.

Many methods exist to compute the penalized maximum likelihood estimate of the $\mathbf{\Sigma}$ matrix above, including the popular `glasso` R package (Friedman et al., 2008) which makes use of a coordinate descent algorithm. The choice of the regularization parameter λ has also been extensively studied (Giraud et al., 2012). In this work, model selection for the regularization parameter λ is performed by minimizing the BIC (Schwarz, 1978). Finally, the edges of the inferred network correspond to non-zero partial correlations, i.e. the non-zero elements of matrix $\mathbf{\Sigma}^{-1}$ (Whittaker, 2009; Friedman et al., 2008).

- **Log-linear Poisson graphical model**. In a log-linear Poisson graphical model (Allen and Liu, 2012), as RNA-seq are frequently characterized by an overdispersed variance with respect to the mean (thus violating one of the assumptions of the Poisson distribution), a transformation is typically required as a first step. Allen and Liu (2012) proposed using a power transformation (Li, 2012) of the data $y_{ij} \rightarrow g(y_{ij}) = y_{ij}^{\alpha}$, with $\alpha \in]0, 1]$, where the coefficient α is chosen to maximize an adequacy criterion between the transformed data $\mathbf{y}^{\alpha}$ and a Poisson distribution.

  Let $\mathbf{z}_i = (g(y_{i1}), ..., g(y_{iq}))$ be the transformed vector of expression values for gene j in the q biological samples. It is assumed that the conditional distribution of Z_{ij} given all the other genes $\mathbf{z}_{i'(-i)} = (z_{1,j}, \ldots, z_{(i-1),j}, z_{(i+1),j}, \ldots, z_{nj})$ is a Poisson distribution $\mathcal{P}(\mu_i)$, with $\log(\mu_i)$ modeled as a linear regression on all the other genes:

$$p(Z_{ij}|\mathbf{z}_{i'(-i)}) \sim \mathcal{P}(\mu_i)$$

  with

$$\log(\mu_i) = \sum_{i' \neq i} \beta_{ii'} \tilde{z}_{i'j}.$$

  The notation $\tilde{\mathbf{z}}$ corresponds to a standardization of the log-transformed data. This standardization is a necessity since we model the mean of the gene i and not the random variable itself. An edge is present in the inferred graph if one or both parameters $\beta_{ii'}$ and $\beta_{i'i}$ are different from zero.

  To ensure sparsity of the vector $\boldsymbol{\beta}_i$, we consider the lasso-penalized log-likelihood for gene i:

$$L^{\lambda}(\boldsymbol{\beta}_i) = -2\sum_{j=1}^{q}\left[\tilde{z}_{ij}\exp\left(\sum_{i' \neq i}\beta_{ii'}\tilde{z}_{i'j}\right) - \sum_{i' \neq i}\beta_{ii'}\tilde{z}_{i'j}\right] + \lambda\|\boldsymbol{\beta_i}\|_{\ell_1} \tag{4.2}$$

  Estimation of parameters $\boldsymbol{\beta}_i$ can be obtained by a coordinate gradient algorithm as implemented in the R package `glmnet` (Friedman et al., 2010). Similarly to Allen and Liu (2012), we perform model selection for the regularization parameter using the Stability Approach to Regularization Selection criterion (StARS; Liu et al., 2010).

- **Hierarchical log-normal Poisson graphical model**. The Poisson log-linear model presented above requires a transformation of the data to account for the high dispersion. Here we propose to deal with this dispersion directly through a hierarchical log-normal Poisson model. The count expression of gene i in sample j is modeled as: $Y_{ij} \sim \mathcal{P}(\theta_{ij})$

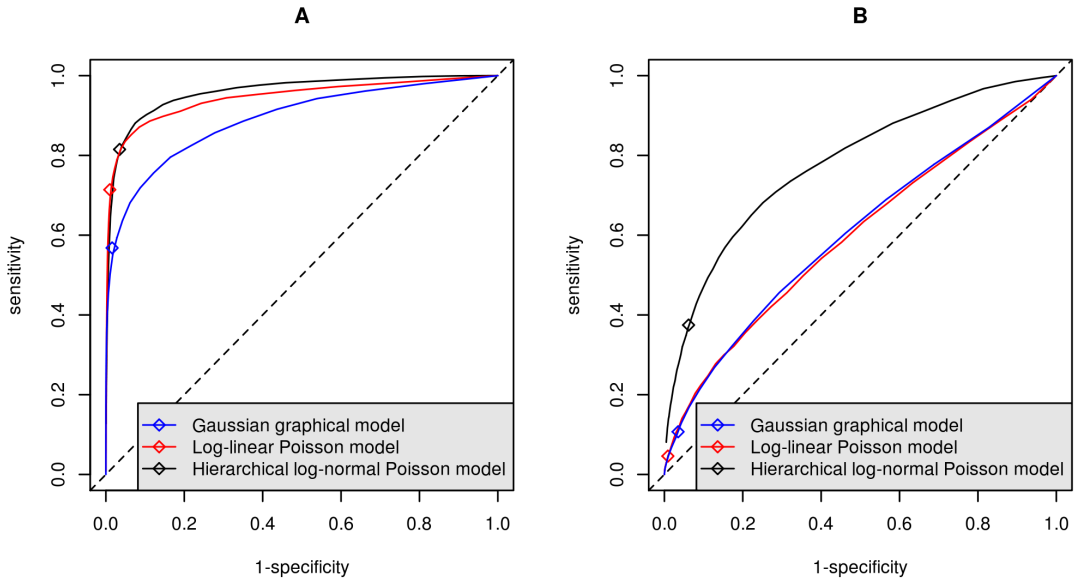

FIGURE 4.2: ROC curves, averaged over 50 simulated data sets on scale-free graphs. Results are presented for the Gaussian graphical model on log-transformed data (blue), the log-linear Poisson graphical model on power-transformed data (red) and the hierarchical log-normal Poisson model on raw data (black) on multivariate Poisson data (A) and multivariate Poisson data with inflated variance (B). The dotted black lines represent the diagonals.

with

$$\log(\theta_{ij}) = \sum_{i' \neq i} \beta_{ii'} \tilde{y}_{i'j} + \varepsilon_{ij}$$
$$\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, ..., \varepsilon_{iq}) \sim \mathcal{N}(0, \sigma_i^2 \boldsymbol{I}_q)$$

As before, the notation $\tilde{\boldsymbol{y}}$ corresponds to a standardization of the log-transformed data. Here, the vector $\mathbf{y}_i \sim \mathcal{P}(\boldsymbol{\theta_i})$ and $\boldsymbol{\theta_i}$ is itself a random variable: $\boldsymbol{\theta_i} = \mu_i \exp(\boldsymbol{\varepsilon}_i)$ with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma_i^2 \boldsymbol{I}_q)$ and $\mu_i = \exp(\sum_{i' \neq i} \beta_{ii'} \tilde{y}_{i'j})$. Note that the variance of the random variable $\mathcal{P}(\boldsymbol{\theta_i})$ is larger than its mean if σ_i^2 is positive. As previously, an edge is present in the graph between genes i and i' if one or both parameters $\beta_{ii'}$ and $\beta_{i'i}$ are different from zero.

In this model, the lasso-penalized likelihood for gene i can be written as:

$$L^{\lambda}(\boldsymbol{\beta_i}, \sigma_i) = -2 \int_{\mathbb{R}} \left(\prod_{j=1}^{q} \left[\exp(-\mu_{ij} + y_{ij} \log(\mu_{ij}) - \log(y_{ij}!) \right] \times \right. \tag{4.3}$$
$$\left. \frac{1}{(2\pi)^{q/2} \sigma_i^q} \exp\left(-\frac{1}{2\sigma_i^2} ||\boldsymbol{\varepsilon_i}||_2^2 \right) \right) d\boldsymbol{\varepsilon}_i + \lambda_i \|\boldsymbol{\beta_i}\|_{\ell_1}.$$

Estimation of parameters $\boldsymbol{\beta_i}$ and σ_i may be performed using the R function `glmmixedlasso` (Schelldorfer and Bühlmann, 2014), based on a Laplace approximation of the penalized likelihood and a coordinate descent algorithm. For model selection, we use a two-stage approach by minimizing the per-gene BIC to identify λ_i, and then averaging over genes to identify a global regularization parameter λ.

Comparisons among the three aforementioned approaches on simulated data with various

amounts of additional inter-sample variability suggested that the proposed hierarchical Poisson log-normal model exhibited better sensitivity and comparable specificity to the GGM and log-linear Poisson model for both multivariate Poisson data (Karlis and Meligkotsidou, 2005) and over-dispersed Poisson data (Figure 4.2. This suggests a benefit to methods developed specifically for RNA-seq data, although for the time being network inference for RNA-seq is limited in practice by the small number of biological replicates typically available.

4.3 Network inference for intervention gene expression data

This section corresponds to the following published article:

Rau, A., Jaffrézic, F., and Nuel, G. (2013) Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology* 7:111.

4.3.1 Background and motivation

Although Gaussian graphical models (Friedman et al., 2008) are often used to infer gene networks from observational (also referred to as wild-type or steady-state) transcriptomic data (also referred to as wild-type or steady-state expression data), they result in undirected graphs (corresponding to partial correlations among genes) that cannot highlight potential causal relationships. For this reason, a great deal of research has focused instead on the use of causal Bayesian networks for a wide variety of applications (Spirtes et al., 2001; Pearl, 2000b). Using Gaussian causal Bayesian networks (GBN) Maathuis (2010) and Maathuis et al. (2009) recently proposed a method called *Intervention-calculus when the DAG is Absent* (IDA) to predict bounds for causal effects from observational data alone. In the IDA, the PC-algorithm (Spirtes et al., 2001; Kalisch, 2012; Kalisch and Bühlmann, 2007) is first applied to find the associated completed partially directed acyclic graph (CPDAG), corresponding to the graphs belonging to the appropriate equivalence class. Following this step, bounds for total causal effects of each gene on the others are estimated using intervention calculus (Pearl, 2000a) for each directed acyclic graph (DAG) in the equivalence class.

However, if intervention experiments such as gene knock-outs or knock-downs are available, it is valuable to jointly perform causal network inference from a combination of wild-type and intervention data. One such approach was proposed by Pinna et al. (2010), based on the simple idea of calculating the deviation between observed gene expression values and the expression under each systematic intervention, where a down-ranking algorithm was applied to the initial graph to remove feed-forward edges. An improved version of that approach was also proposed Pinna (2013) to provide more accurate network inference for large-scale networks through a novel implemention of the transitive reduction step. Both methods have the dual advantages of being very fast to compute and being quite general, as they do not require any assumption of acyclicity of the graph. However, in order to evaluate all possible causal links among genes, the Pinna et al. (2010) and Pinna (2013) methods require a single replicate of observational data as well as a systematic knock-out experiment for each gene in the network.

In this work, we instead seek to identify a flexible method able to jointly infer causal relationships among genes from arbitrarily complex knock-out experiments, including partial or multiple gene knock-out experiments. Although in principle such intervention experiments could be conducted using RNA-seq technology, for the time being the majority of such interventional data have been instead collected using microarrays. With the refinement of *in vivo*

gene silencing experimental techniques such as RNA interference (RNAi) and the CRISPR-Cas9 system, intervention experiments are likely to become increasingly feasible for gene expression studies in coming years.

4.3.2 MCMC-Mallows algorithm for causal Gaussian Bayesian networks

Let $G = (V, E)$ be a graph defined by a set of vertices V and edges $E \subset (V \times V)$. Let the vertices of a graph represent p random variables $X_1, ..., X_p$. As in the approach of Maathuis et al. (2009), we consider here the framework of causal GBNs, which correspond to Bayesian networks where the nodes have a Gaussian residual distribution and edges represent linear dependencies. In this case, it also follows that the joint distribution of the network is multivariate Gaussian. In DAGs such as GBNs, we often encounter the presence of Markov equivalence classes, i.e. multiple network structures that yield the same joint distribution; in such cases, observational data alone generally cannot orient edges. For this reason, in many cases the use of intervention data can help overcome this issue, as presented below.

Following an intervention on a given node X_i, denoted $\mathsf{do}(X_i = x)$, we consider the expected value of each other gene in the network via do-calculus as shown in Theorem 3.2.2 (Adjustment for direct causes) in Pearl (2000a):

$$\mathbb{E}(X_j|\mathsf{do}(X_i = x)) = \begin{cases} \mathbb{E}(X_j) & \text{if } X_j \in \mathsf{pa}(X_i) \\ \int \mathbb{E}(X_j|x, \mathsf{pa}(X_i))\mathbb{P}(\mathsf{pa}(X_i))d\mathsf{pa}(X_i) & \text{if} X_j \notin \mathsf{pa}(X_i) \end{cases}$$

where $\mathsf{pa}(X_i)$ represents the parents of node X_i. It is important to point out that $\mathbb{P}(Y|\mathsf{do}(X = x))$ is different from the conditional probability $\mathbb{P}(Y|X = x)$. Using this framework, the total causal effects may be defined as follows:

$$\beta_{ij} = \frac{\partial}{\partial x}\mathbb{E}(X_j|\mathsf{do}(X_i = x))$$

and are equal to 0 if X_i is not an ancestor of X_j. On the other hand, the direct causal effects (i.e. the edges in the graph) are defined as:

$$\alpha_{ij} = \frac{\partial}{\partial x}\mathbb{E}(X_j|\mathsf{pa}(X_j), \mathsf{do}(X_i = x)).$$

Causal inference method

In the GBN framework, when observational data are jointly modeled with intervention data for an arbitrary subset of genes, the network follows a multivariate Gaussian distribution of dimension equal to the number of genes that had no intervention (as the expression value of the gene under intervention is fixed to a given value), and the log-likelihood value can subsequently be calculated for a proposed network.

The calculations in the following section assume that the nodes in the graph have been sorted according to an appropriate causal ordering in the graph such that if $i < j$, then X_j is not an ancestor of X_i; we note that such an ordering is possible under the assumption of acyclicity of the graph. In practice, of course, it is typically not possible to correctly order nodes in such a way without knowledge of the underlying DAG. For this reason, we aim to explore various network structures based on causal orderings, and to choose among those with the best likelihood value for an arbitrary set of observational and intervention data. The Metropolis-Hastings algorithm (Metropolis, 1953; Hastings, 1970), through the use of a proposal distribution for causal orderings, allows such an exploration to take place and to approach a local maximum of the likelihood.

Let p be the number of nodes in the graph, G the DAG structure and $\mathbf{W}$ the matrix containing the values for all edges. The nodes are assumed to have been sorted by parental order for G and $\mathbf{W}$, i.e. if $i < j$, then X_j is not an ancestor of X_i. This sorting is possible under the assumption of acyclicity and may not necessarily be unique. Under this ordering, $\mathbf{W}$ is an upper triangular matrix and thus nilpotent. In the GBN framework, it is assumed that each node of G has a residual Gaussian distribution, independently from the rest of the network. Let us consider $X_{\mathcal{I}}$ with $\mathcal{I} = \{1, \ldots, p\}$, a set of p Gaussian random variables defined by:

$$X_j = m_j + \sum_{i \in \text{pa}(j)} w_{i,j} X_i + \varepsilon_j \quad \text{with} \quad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2). \tag{4.4}$$

We assume that the ε_j are independent, and that $i \in \text{pa}(j) \Rightarrow i < j$ (this assumption is equivalent to assuming that the directed graph obtained using the parental relationships is acyclic). Given the parental structure of the graph, $w_{i,j}$ may only be nonzero on the edge set, $(i,j) \in \mathcal{E} = \{i \in \text{pa}(j), j \in \mathcal{I}\}$.

Let us now consider the matrix form of Equation (4.4):

$$\mathbf{X} = \mathbf{m} + \mathbf{X}\mathbf{W} + \boldsymbol{\varepsilon}$$

where $\mathbf{X} = (X_1, \ldots, X_p)$, $\mathbf{m} = (m_1, \ldots, m_p)$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)$ are row-vectors of dimension p, and $\mathbf{W} = (w_{i,j})_{1 \leqslant i,j \leqslant p}$ is a p-dimensional square matrix. By recursively applying this formula and taking advantage of the nilpotence of matrix $\mathbf{W}$, we obtain:

$$\mathbf{X} = \mathbf{m}\mathbf{L} + \boldsymbol{\varepsilon}\mathbf{L}$$

where $\mathbf{L} = (\mathbf{I} - \mathbf{W})^{-1} = \mathbf{I} + \mathbf{W} + \ldots + \mathbf{W}^{p-1}$. This proves that the model defined in Equation (4.4) is equivalent to $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\boldsymbol{\mu} = \mathbf{m}\mathbf{L} \quad \text{and} \quad \boldsymbol{\Sigma} = \mathbf{L}^T \text{diag}(\boldsymbol{\sigma}^2)\mathbf{L} = \sum_{j \in \mathcal{I}} \sigma_j^2 \mathbf{L}^T \mathbf{e}_j^T \mathbf{e}_j \mathbf{L}$$

where $\mathbf{e}_j$ is a p-dimensional null row-vector except for its j^{th} term which is equal to 1, and where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)$ is a row-vector of dimension p.

The log-likelihood of the model given N observations $\mathbf{x}^k = (x_1^k, \ldots, x_p^k)$ $(1 \leqslant k \leqslant N)$ is then:

$$\ell(\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W}) = -\frac{Np}{2}\log(2\pi) - N\sum_{j \in \mathcal{I}} \log(\sigma_j) - \frac{1}{2}\sum_{k=1}^{N}\sum_{j \in \mathcal{I}} \frac{1}{\sigma_j^2}(x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2.$$

To see this, let us define $\mathbf{A}_k = (\mathbf{x}^k - \mathbf{m}\mathbf{L})\boldsymbol{\Sigma}^{-1}(\mathbf{x}^k - \mathbf{m}\mathbf{L})^T$ for all k. Since $\boldsymbol{\Sigma}^{-1} = (\mathbf{I} - \mathbf{W})\text{diag}(1/\boldsymbol{\sigma}^2)(\mathbf{I} - \mathbf{W})^T$ we get:

$$\begin{aligned} \mathbf{A}_k &= \sum_{j \in \mathcal{I}} \frac{1}{\sigma_j^2}(\mathbf{x}^k(\mathbf{I} - \mathbf{W}) - \mathbf{m})\mathbf{e}_j^T \mathbf{e}_j (\mathbf{x}^k(\mathbf{I} - \mathbf{W}) - \mathbf{m})^T \\ &= \sum_{j \in \mathcal{I}} \frac{1}{\sigma_j^2}(x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2. \end{aligned}$$

Analytical formulae can be obtained for the derivatives with respect to parameters $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W})$.

The likelihood presented above only takes into account observational data. Let us now consider the case of an arbitrary mixture of observational and intervention data. We assume that we perform an intervention on a subset $\mathcal{J} \subset \mathcal{I} = \{1, \ldots, p\}$ of variables by artificially

fixing the level of the corresponding variables to a value (typically 0 in the case of knock-out experiments): $\mathsf{do}(X_{\mathcal{J}} = x_{\mathcal{J}})$. The model is then obtained by assuming that all $w_{i,j} = 0$ for $(i,j) \in \mathcal{E}$ and $j \in \mathcal{J}$; we denote the corresponding matrix $\mathbf{W}_{\mathcal{J}}$. We also assume that the variables X_j for $j \in \mathcal{J}$ are fully deterministic. As before, the resulting model is hence Gaussian: $X_{\mathcal{I}}|\mathsf{do}(X_{\mathcal{J}} = x_{\mathcal{J}}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{J}}(x_{\mathcal{J}}), \boldsymbol{\Sigma}_{\mathcal{J}})$ with

$$\boldsymbol{\mu}_{\mathcal{J}}(x_{\mathcal{J}}) = \boldsymbol{\nu}_{\mathcal{J}}(x_{\mathcal{J}})\mathbf{L}_{\mathcal{J}}, \quad \boldsymbol{\Sigma}_{\mathcal{J}} = \sum_{j \notin \mathcal{J}} \sigma_j^2 \mathbf{L}_{\mathcal{J}}^T \mathbf{e}_j^T \mathbf{e}_j \mathbf{L}_{\mathcal{J}},$$

where

$$\boldsymbol{\nu}_{\mathcal{J}}(x_{\mathcal{J}}) e_j^T = \begin{cases} x_j & \text{if } j \in \mathcal{J} \\ m_j & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{L}_{\mathcal{J}} = (\mathbf{I} - \mathbf{W}_{\mathcal{J}})^{-1} = \mathbf{I} + \mathbf{W}_{\mathcal{J}} + \ldots + \mathbf{W}_{\mathcal{J}}^{p-1}.$$

For the likelihood calculation, we consider N data generated under $x^k = (x_1^k, \ldots, x_p^k)$ $(1 \leqslant k \leqslant N)$ with intervention on $\mathcal{J}_k$ (where $\mathcal{J}_k = \emptyset$ means no intervention). We denote by $\mathcal{K}_j = \{k, j \notin \mathcal{J}_k\}$, and by $N_j = |\mathcal{K}_j|$ its cardinal. The log-likelihood of the model can then be written as:

$$\ell(\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W}) = -\frac{\log(2\pi)}{2} \sum_j N_j - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_{j \notin \mathcal{J}_k} \frac{1}{\sigma_j^2} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2. \tag{4.5}$$

This is mainly due to the fact that for any intervention set $\mathcal{J}$ we have $\mathbf{W}_{\mathcal{J}} \mathbf{e}_j^T = \mathbf{W} \mathbf{e}_j^T$ for all $j \notin \mathcal{J}$. Considering the derivative with respect to m_j for all j such that $N_j > 0$, we obtain:

$$m_j = \frac{1}{N_j} \sum_{k \in \mathcal{K}_j} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T)$$

which can be plugged into the likelihood expression to get:

$$\tilde{\ell}(\boldsymbol{\sigma}, \mathbf{W}) = -\frac{\log(2\pi)}{2} \sum_j N_j - \sum_j N_j \log(\sigma_j) - \frac{1}{2} \sum_k \sum_{j \notin \mathcal{J}_k} \frac{1}{\sigma_j^2} (y_j^{k,j} - \mathbf{y}^{k,j} \mathbf{W} \mathbf{e}_j^T)^2$$

where for (k,j) such that $j \notin \mathcal{J}_k$ we have:

$$\mathbf{y}^{k,j} = \mathbf{x}^k - \frac{1}{N_j} \sum_{k' \in \mathcal{K}_j} \mathbf{x}^{k'}$$

and $\mathbf{W}$ can be estimated by solving the following linear system:

$$\sum_{i', (i',j) \in \mathcal{E}} w_{i',j} \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_{i'}^{k,j} = \sum_{k \in \mathcal{K}_j} y_i^{k,j} y_j^{k,j} \quad \text{for all } (i,j) \in \mathcal{E}. \tag{4.6}$$

Note that the system might be degenerate if the intervention design gives no insight on some parameters. It is hence finally possible to obtain $\boldsymbol{\sigma}$ through:

$$\sigma_j^2 = \frac{1}{N_j} \sum_{k \in \mathcal{K}_j} (y_j^{k,j} - \mathbf{y}^{k,j} \mathbf{W} \mathbf{e}_j^T)^2.$$

MCMC algorithm with Mallows proposal model

The Metropolis-Hastings algorithm (Metropolis, 1953; Hastings, 1970) is a random walk over Ω, the parameter space of the model. It relies on an instrumental probability distribution Q which defines the transition from position X_t to a new position X. The probability of moving from state X_t to the new state X is defined by:

$$P(X_{t+1} = X|X_t) = \min\left\{\frac{\pi(X)Q(X_t, X)}{\pi(X_t)Q(X, X_t)}, 1\right\} \tag{4.7}$$

where $\pi(X)$ is the likelihood function.

In order to propose a new causal node ordering $\mathcal{O}^\star$ from the previous ordering $\mathcal{O}$, we propose to make use of the Mallows model (Mallows, 1957). Briefly, under this model, the density of a proposed causal ordering is defined as follows:

$$\begin{aligned} P(\mathcal{O}^\star) &= P(\mathcal{O}^\star|\mathcal{O}, \phi) \\ &= \frac{1}{Z}\phi^{d(\mathcal{O}^\star, \mathcal{O})} \end{aligned}$$

where $\phi \in (0, 1]$ is a fixed temperature parameter, Z is a normalizing constant, and $d(\cdot, \cdot)$ is a dissimilarity measure between $\mathcal{O}$ and $\mathcal{O}^\star$ based on the number of pairwise ranking disagreements. In addition, we remark that as the temperature parameter ϕ approaches zero, the Mallows model approaches a uniform distribution over all causal orderings, and if $\phi = 1$, the model corresponds to a dirac distribution on the reference ordering $\mathcal{O}$. In the following, we will use a reparameterization of the temperature coefficient ϕ such that $\phi = \exp(-1/\eta)$, with $\eta > 0$. Due to the symmetry of d, it is clear that $P(\mathcal{O}^\star|\mathcal{O}, \phi) = P(\mathcal{O}|\mathcal{O}^\star, \phi)$, which allows a simplification of the Q terms in the acceptance ratio in Equation (4.7). In practice, ϕ is a parameter that must be tuned by the user to obtain an acceptance rate near 30 to 40% (Roberts et al., 1997).

Proposals for causal node orderings using the aforementioned Mallows model may be obtained by sampling using a repeated insertion model as described in Doignon et al. (2004). Based on this new proposal for the node ordering $\mathcal{O}^\star$, maximum likelihood estimators may be calculated for the model parameters $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\sigma}, \mathbf{W})$ using the likelihood described in Equation (4.5). Subsequently, the Metropolis-Hastings ratio may be calculated and used to determine whether the proposed causal node ordering is accepted or rejected.

4.3.3 Conclusions and discussion

In simulation studies, we explored the posterior distribution of causal node orderings using our proposed MCMC-Mallows GBN model (50,000 iterations, with a burn-in of 5000 iterations and thinning interval of 50 iterations) when data consisted of (1) a mixed setting with wild-type samples and one knock-out per gene; (2) a partial knock-out design with wild-type samples and one knock-out for a subset of genes; (3) a multiple knock-out design, with wild-type samples, one knock-out per gene, and five double knock-outs (i.e., samples in which two genes were simultaneously inactivated). Comparisons with the Pinna et al. (2010) and Maathuis et al. (2009) approaches using various criteria (area under the ROC curve, area under the precision-recall curve, mean squared error) for both total and direct causal effects indicated that in settings with only partial (rather than systematic) knock-outs, the MCMC-Mallows GBN approach was better able to leverage the intervention data to provide satisfactory estimates of causal effects. Additionally, our simulations demonstrated that multiple knock-out designs contributed valuable additional information for causal network inference beyond single knock-outs; we therefore anticipate that the need for methods able to accommodate complex intervention designs will only increase as such data become more common.

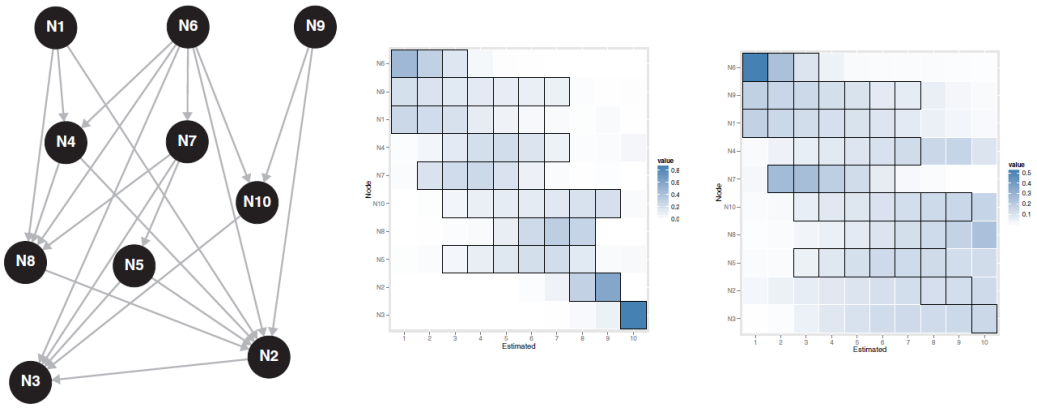

FIGURE 4.3: (left) Graph structure, with ten nodes and 21 edges, used in simulation study. Posterior distribution of node orders from the MCMC-Mallows approach from the simulation setting with complete single knock-outs (middle) and partial single knock-outs (right). Node labels are included on the vertical axis, estimated positions within causal orderings along the horizontal axis, and the intensity of color of each square corresponds to the average proportion of iterations in which a given node was placed in a given position.

As the MCMC-Mallows approach explores the posterior distribution of causal node orderings, it is also of interest to visualize this posterior distribution (Figure 4.3 middle and right). In these plots, node labels are included on the vertical axis, and estimated positions within orderings along the horizontal axis. Potential orderings for each node within the true graph are highlighted with black outlines (note that the node ordering is not unique for the DAG considered here); as an example, node N6 could be placed in the first, second, or third position, while node N3 could only be placed in the tenth position in the true graph. The intensity of colors within each box represents the average proportion of iterations in which a node was placed in a particular order. To follow our example, in the mixed setting (center of Figure 4.3), on average node N6 was most often placed in the first position, and occasionally positioned second or third, while node N3 was nearly always placed in the last position. As expected, the node orders were most accurately estimated when a systematic knock-out design was considered (with one knock-out for each gene) than for a partial knock-out design, but pertinent information can still be extracted from the latter.

The novelty of the MCMC-Mallows approach, and the primary contribution of this work, lies in its flexibility to model arbitrary single, multiple, and partial knock-out designs as well as in the focus on exploring the posterior distribution of causal orderings of nodes rather than of the directed acyclic graph itself. In its present form, the proposed algorithm is not applicable to large-scale networks made up of several hundreds of nodes. Due to the curse of dimensionality, the size of the search space of causal node orderings explodes in dimension as the number of nodes increases, meaning that alternative MCMC samplers, such as parallel tempering, may be better suited to such situations. In addition, the resolution of the linear system in Equation (4.6) needed for the likelihood calculation has complexity $O(p^6)$ when no sparsity constraints are included for matrix $\mathbf{W}$. As such, the generalization of the proposed algorithm to a $p >> n$ situation will require the addition of a ridge or Lasso penalty, as recently proposed by Fu and Zhou (2013), as well as a modification of the proposal distribution and sampling strategy. The current algorithm is fully compatible with such extensions.

Finally, during the Master's and Ph.D. work of Gilles Monneret (co-supervised by Florence Jaffrézic, Grégory Nuel, and myself) we have developed several extensions to this

initial work, including the use of pairwise gene ordering preferences rather than the Mallows model (Nuel et al., 2013) and a ridge penalty for high-dimensional networks (Monneret, 2015). Futher extension have also been motivated by two successful interdisciplinary collaborations with biologists in the GABI research unit funded by INRA Animal Genetics internal department grants. The first, entitled "Causality" (2014; coordinated by Florence Jaffrézic and Tatiana Zerjal) focused on transcriptomic data produced for wild-type and dwarf chickens, the latter of which have a naturally-occurring functional knock-out of the growth hormone receptor. To address the fact that a single gene was inactivated in this experiment, we developed a marginal causal estimation approach based on the framework of Gaussian directed acyclic graphs (Monneret, 2017) to identify genes with a causal downstream relationship to the growth hormone receptor. Although this approach performs very similarly in practice to a classical differential analysis, it has the advantage of providing a formal causal interpretation. More recently, a second project entitled "COSI-net: Using COmbinatorial gene Silencing and Inactivation to infer gene NETworks" (2016; coordinated by myself in collaboration with Jean-Luc Vilotte and Katayoun Moazami-Goudarzi (MoDiT team in GABI, INRA) was also successfully funded. This project provided a new set of data collected in double knock-out and RNAi knock-down mice for the PrnP and Shadoo prion-encoding genes; statistical exploration and analysis of these rich data is currently ongoing.

Chapter 5

Future projects

The methods presented in this manuscript primarily deal with transcriptomic data measured using either RNA-seq or microarray technology. High-throughput technologies now enable deep and multi-faceted studies of the biological variability of living organisms at a variety of levels in addition to the transcriptome, including the proteome, metabolome, and epigenome, as well as copy number variations, single nucleotide polymorphisms, and chromatin accessibility. Each of these data types provides a different, partial, but complementary view of the genome. Despite the increasing availability of these various data sources and expanding databases of genome annotations, our understanding of the function of the genome and its relationship to phenotypic and/or physiological characteristics is far from complete. Identifying an appropriate way to simultaneously exploit and model this large accumulation of hetereogeneous 'omics data collected on the same individuals remains a major obstacle and an important area of current biostatistical research.

Multi-omics integration has in fact become a bit of a buzz word in the past couple of years, and this admittedly vague and ill-defined term encompasses a broad range of topics and can mean widely different things to researchers from the fields of biostatistics, bioinformatics, and biology (as well as to researchers within each of those fields!). As such, one of the major challenges in addressing multi-omics data integration is the need to clearly define the biological questions of interest; once this is done, the statistical challenges associated with such analysis are numerous (e.g., simultaneous modeling of continuous and count data, large number of variables with a limited number of biological replicates, preprocessing steps) and often (but not always!) require the development of new statistical methodologies.

My current and future research projects will seek to pose and address some well-defined questions concerning multi-omics data integration. I will detail a few of them in this chapter.

5.1 Integrated clustering of gene expression and methylation data

Our recent work (Rau and Maugis-Rabusseau, 2017) has convinced us that using normalized expression profiles (rather than raw counts) is an appropriate strategy for RNA-seq co-expression analyses. In collaboration with Cathy Maugis-Rabusseau and Antoine Godichon-Baggioni, we have several extensions on which we would like to follow-up when relevant multi-omics data are available. For the time being, the methods proposed to perform integrated clustering of multi-omics data have primarily focused on grouping together individuals (e.g., to identify groups of patients exhibiting a molecular structure for a subtype of cancer), for example using a joint latent variable model (Shen et al., 2009; Mo, 2013). Our continued goal in this work is instead to continue our focus on clustering biological entities, such as genes. We envisage several possibilities:

- In our previous work on model selection using functional annotations (Gallopin, 2015), we focused on modeling gene expression alone and using the additional (partially missing and categorical) information to guide model selection. In somewhat related work,

our first idea is to cluster normalized gene expression profiles (for example, using a K-means algorithm as described in Section 3.3.2) for a fixed large number of clusters K, and subsequently aggregate clusters based on a distance measure that integrates a secondary set of relevant 'omics data (e.g., methylation data). Such a distance measure could be defined, for example, by adapting the weightd consensus clustering measure in the multi-view K-means algorithm (Cai et al., 2013).

- A similar approach would be to instead fix a very small number of clusters K for a K-means clustering of normalized gene expression profiles, and instead use the secondary set of relevant 'omics data to split clusters, in an analgous way as that described above.

- Finally, the multi-view K-means algorithm (Cai et al., 2013) itself could be used to directly and jointly cluster gene expression and secondary 'omics data; in this case, several questions must be addressed, including the impact of transformations on each data type, the selection of an appropriate number of clusters, the integration of qualitative data, and how best to deal with missing values (e.g., genes for which expression data is available but methylation data is not). Another interesting extension would be to identify with the multi-view K-means algorithm could incorporate cluster- and block-specific weights, thus allowing data sources to have different weights in different clusters.

5.2 Exploring molecular drivers of gene expression

In August-September 2016, I had the opportunity to be a Visiting Scholar at the University of Wisconsin-Milwaukee to work with Paul L. Auer (University of Wisconsin-Milwaukee). In our continued collaboration, we are working on an exploratory analysis of pan-cancer gene regulation using a rich and varied set of semi-public data from a project called The Cancer Genome Atlas (TCGA). In particular, transcriptomic, epigenomic, genomic, proteomic, and clinical data were collected for several thousands of patients with one of over thirty different tumor types.

Using a linear mixed model, we are currently in the process of analyzing to what extent variability in gene expression can be explained by methylation, copy number alterations, somatic mutations, genetic heritability, and transcription factor and miRNA expression, as well as how these patterns are conserved across cancer types or subtypes. We have developed an interactive R/Shiny web application, entitled "EDGE cancer dashboard: Exploring Drivers of Gene Expression in cancer genomes", to facilitate graphical exploration of our results. We are also working with a specialist in breast cancer genomics, Mike Flister (Medical College of Wisconsin), to experimentally validate genes of interest that are highlighted by our pan-cancer integrative approach. To give one example, the non-receptor protein tyrosine phosphatase (PTPN14) that regulates many breast cancer pathways has been implicated in breast cancer growth and metastasis; however somatic mutations and copy number variants of PTPN14 do not appear to be prevalent in breast cancer, and the transcriptional regulators of PTPN14 are unknown. Initial exploration with our interactive tool confirmed previous results from ChIP-seq data collected in the ENCODE project, suggesting that the transcription factors FOXA1 and GATA3 are important molecular drivers of PTPN14 expression in breast cancer. We anticipate that this work will be submitted for publication in the coming months, and will lead to futher developments in methodological research to contribute understanding of the regulatory landscape of cancer.

Longer-term extensions to this work are also expected. In particular, we plan to extend the analysis results currently presented in the web application to include information about relevant clinical/survival characteristics of patients in order to associate promising molecular

drivers of gene expression with clinical outcomes. We anticipate that this could be done, for example, by providing Kaplan-Meier plots based on extremes of gene expression levels. In addition, similar approaches could be undertaken for the analysis of other large-scale genomic projects, including the Breast Cancer Risk after Diagnostic Gene Sequencing (BRIDGES), B-CAST and Trans-Omics for Precision Medicine (TOPMed) projects. Finally, it is obviously of great interest to perform the molecular decomposition of gene expression variation in species of agricultural importance (e.g., cattle, chickens, pigs). For the time being, the primary obstacle to such an extension is simply the lack of publicly-available large-scale multi-omics data collected on the same individuals, but it is likely that such data will become increasingly available in the coming years.

5.3 Joint modeling of chromatin accessibility and gene expression data

I have recently started participating in the analysis work group for the FR-AgENCODE pilot project, which is part of the Functional Annotation of Animal Genomes (FAANG) international consortium (The FAANG Consortium, 2015) and aims to improve the functional annotation of livestock species (cattle, chicken, goat, pig) through the production and analysis of high-throughput data. In particular, for each of these species, RNA-seq and chromatin accessibility (*Assay for Transposase-Accessible Chromatin with high throughput sequencing*, referred to as ATAC-seq) data were collected for two males and two females in each of three tissues (liver and two types of lymphocytes, CD3+CD4+ and CD3+CD8+). In addition, the proximity of genomic loci in three-dimensional space, known as the chromosome conformation, was also measured in the liver cells of each individual using Hi-C technology. In ATAC-seq data, a peak-calling bioinformatic step is required (as accessible regions of chromatin are not necessarily situated within the coding regions of genes), and each called peak is then associated with a count of the number of sequenced fragments. In Hi-C data, following read mapping the data must be binned and bias-corrected (i.e., balanced) to ensure that the sum of every row/column in the matrix is equivalent. It should be noted that these pre-processing steps as well as the appropriate normalization for ATAC-seq and Hi-C data are active ongoing areas of methodological research.

Following standard differential analyses between tissues and sexes of the RNA-seq and ATAC-seq data individually, it is of primary interest to understand the shared variability of these two sources of information. We are currently in the first steps of exploring the use of multivariate exploratory techniques (e.g., multiple factor analysis, sparse partial least squares) to jointly investigate these two sources of information. To identify potential distal gene enhancers, we also plan to explore the use of a lasso penalized regression to predict the expression of each gene with respect to chromatin accessibility of peaks in a large window around the gene; one interesting extension of this model would be to make use of weighted lasso regression to incorporate into the aforementioned model the spatial proximity of ea chgene-peak pair, as measured by Hi-C.

Bibliography

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Allen, G. I. and Z. Liu (2012). "A log-linear graphical model for inferring genetic networks from high-throughput sequencing data." *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data". *Genome Biology* 11.R106, pp. 1–28.

Ashburner, M. et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature Genetics* 25.1, pp. 25–9.

Baudry, J.-P., C. Maugis, and B. Michel (2012). "Slope heuristics: overview and implementation". *Statistics and Computing* 22, pp. 455–470.

Baudry, J.-P.and others (2014). "Enhancing the selection of a model-based clustering with external categorical variables." *Advances in Data Analysis and Classification* 1.1, pp. 1–20.

Biernacki, C. et al. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7, pp. 719–725.

Biernacki, C., G. Celeux, and G. Govaert (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models". *Computational Statistics and Data Analysis* 41.1, pp. 561–575.

Biernacki, C. et al. (2006). "Model-based cluster and discriminant analysis with the MIXMOD software". *Computational Statistics and Data Analysis* 51.2, pp. 587–600.

Birgé, L. and P. Massart (2001). "Gaussian model selection". *Journal of the European Mathematical Society* 3, pp. 203–268.

— (2007). "Minimal penalties for Gaussian model selection". *Probability Theory and Related Fields* 138, pp. 33–73.

Blekhman, R. et al. (2010). "Sex-specific and lineage-specific alternative splicing in primates". *Genome Research* 20.2, pp. 180–189.

Bottomly, D. et al. (2011). "Evaluating gene expression in C57BL/GJ and DBA/2J mouse striatum using RNA-seq and microarrays". *PLoS One* 6.3, e17820.

Bourgon, R., R. Gentleman, and W. Huber (2010). "Independent filtering increases detection power for high-throughput experiments". *PNAS* 107.21, pp. 9546–9551.

Box, G. E. P. and D. R. Cox (1964). "An analysis of transformations". *Journal of the Royal Statistical Society, Series B (Methodological)* 26.2, pp. 211–252.

Breitling, R. et al. (2004). "Rank products: A simple yet powerful new method to detect differential regulated genes in replicated microarray experiments". *FEBS Letters* 573, pp. 83–92.

Brooks, A.N. et al. (2011). "Conservation of an RNA regulatory map between Drosophila and mammals". *Genome Research* 21.2, pp. 193–202.

Cai, L. et al. (2004). "Clustering analysis of SAGE data using a Poisson approach". *Genome Biology* 5, R51.

Cai, X., F. Nie, and H. Huang (2013). "Multi-view K-measn clustering on big data". *IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 2598–2604.

Cai, Y., B. Fendler, and G.S. Atwal (2012). "Utilizing RNA-seq data for cancer network inference". *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*.

Cánovas, A. et al. (2010). "SNP discovery in the bovine milk transcriptome using RNA-seq technology". *Mammalian Genome* 21, pp. 592–598.

Celeux, G. and G. Govaert (1992). "A classification EM algorithm for clustering and two stochastic versions". *Computational Statistics and Data Analysis* 14.3, pp. 315–332.

— (1995). "Gaussian parsimonious clustering models". *Pattern Recognition* 28.5, pp. 781 –793.

Choi, J. K. et al. (2003). "Combining multiple microarray studies and model interstudy variation". *Bioinformatics* 19.Suppl 1, pp. 84–90.

Cleveland, W.S. (1979). "Robust locally weighted regression and smoothing scatterplots". *Journal of the American Statistical Association* 74.368, pp. 829–836.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society, Series B (Methodological)* 39.1, pp. 1–38.

D'haeseleer, P., S. Liang, and R. Somogyi (2000). "Genetic network inference: from co-expression clustering to reverse engineering". *Bioinformatics* 16.8, pp. 707–726.

Dillies, M.-A. et al. (2013). "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". *Briefings in Bioinformatics* 14.6, pp. 671–683.

Doignon, J.P., A. Pekec, and M. Regenwetter (2004). "The repeated insertion model for rankings: Missing link between two subset choice models". *Psychometrika* 69.1, pp. 33–54.

Eisen, M. B. et al. (1998). "Cluster analysis and display of genome-wide expression patterns". *PNAS* 95.25, pp. 14863–14868.

Endale Ahanda, M.-L. et al. (2014). "Impact of the genetic background on the composition of the chicken plasma miRNome in response to a stress". *PLoS ONE* 9.12, e114598.

Fietz, S. A. et al. (2012). "Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal". *PNAS* 109.29, pp. 11836–11841.

Fisher, R. A. (1932). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Frazee, A. C., B. Langmead, and J. T. Leek (2011). "ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets". *BMC Bioinformatics* 12.

Friedman, J; T. Hastie, and R. Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". *Biostatistics* 9.3, pp. 432–441.

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". *Journal of Statistical Software* 33, pp. 1–22.

Fu, F. and Q. Zhou (2013). "Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent". *Journal of the American Statistical Association* 108.501.

Gallopin, M. (2015). "Classification et inférence de réseaux pour les données RNA-seq". PhD thesis. Université Paris-Saclay.

Gentleman, R. C. et al. (2004). "Bioconductor: Open software development for comptuational biology and bioinformatics." *Genome Biology* 5.R80.

Giorgi, F. M., C. Del Fabbro, and F. Licausi (2013). "Comparative study of RNA-seq and microarray-derived coexpression networks in *Arabidopsis thaliana*". *Bioinformatics* 29, pp. 717–724.

Giraud, C., S. Huet, and N. Verzelen (2012). "Graph selection with GGMselect". *Statistical Applications in Genetics and Molecular Biology* 11.3, pp. 1544–6115.

Graveley, B. R. et al. (2011). "The development transcriptome of *Drosophila melanogaster*". *Nature* 471, pp. 473–479.

Hammer, P. et al. (2010). "mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain". *Genome Research* 20.6, pp. 847–860.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* 57.1, pp. 97–109.

Hong, S. et al. (2013). "Canonical correlation analysis for RNA-seq co-expression networks". *Nucleic Acids Research* 41.e95.

Huang, D. and W. Pan (2006). "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data". *Bioinformatics* 22.10, pp. 1259–1268.

Huber, W. et al. (2003). "Parameter estimation for the calibration and variance stabilization of microarray data". *Statistical Applications in Genetics and Molecular Biology* 2.1, Article 3.

Iancu, O. D. et al. (2012). "Utilizing RNA-seq data for de novo coexpression network inference". *Bioinformatics* 28, pp. 1592–1597.

Jaccard, P. (1901). "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, pp. 547–549.

Jaffrézic, F. et al. (2007). "A structural mixed model for variances in differential gene expression studies". *Genetics Resesearch* 89, pp. 19–25.

Jiang, D., C. Tang, and A. Zhang (2004). "Cluster analysis for gene expression data: A survey". *IEEE Transactions on Knowledge and Data Engineering* 16.11, pp. 1370–1386.

Kalisch, M. and P. Bühlmann (2007). "Estimating high-dimensional directed acyclic graphs with the PC-algorithm". *J. Mach. Learn. Res.* 8, pp. 613–636.

Kalisch, M. et al. (2012). "Causal inference using graphical models with the R package pcalg". *Journal of Statistical Software* 47.11, pp. 1–26.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic Acids Research* 28.1, pp. 27–30.

Karlis, D. (2003). "An EM algorithm for multivariate Poisson distribution and related models". *Journal of Applied Statistics* 30.1, pp. 63–77.

Karlis, D. and L. Meligkotsidou (2005). "Multivariate Poisson regression with covariance structure". *Statistics and Computing* 15, pp. 255–265.

Łabaj, P. P. et al. (2011). "Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling". *Bioinformatics* 27.ISMB, pp. i383–i391.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". *BMC Bioinformatics* 9.559.

Law, C. et al. (2014). "voom: precision weights unlock linear model analysis tools for RNA-seq read counts". *Genome Biology* 15.R29.

Lebret, R. et al. (2015). "Rmixmod: The R Package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library". *Journal of Statistical Software* 67.6, pp. 1–29.

Li, J. et al. (2012). "Normalization, testing, and false discovery rate estimation for RNA-sequencing data". *Biostatistics* 13.3, pp. 523–538.

Liberzon, A. et al. (2011). "Molecular signatures database (MSigDB) 3.0." *Bioinformatics* 27.12, pp. 1739–40.

Liptak, T. (1958). "On the combination of independent tests". *Magyar Tudomanyos. Akademia Matematikai Kutato Intezetenek Kozlemenyei* 3, pp. 171–197.

Liu, H., K. Roeder, and L. Wasserman (2010). "Stability approach to regularization selection criterion". *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pp. 1432–1440.

Love, M. I., W. Huber, and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome Biology* 15.550.

Maathuis, M. H. et al. (2010). "Predicting causal effects in large-scale systems from observational data". *Nature Methods* 7.4, pp. 247–248.

Maathuis, M.H., M. Kalisch, and P. Bühlmann (2009). "Estimating high-dimensional intervention effects from observational data". *Annals of Statistics* 37, pp. 3133–3164.

Mach, N. et al. (2014). "Extensive Expression Differences along Porcine Small Intestine Evidenced by Transcriptome Sequencing". *PLoS ONE* 9.2, pp. 1–12.

MacQueen, J. B. (1967). "Some methods for classification and analysis of multivariate observations". *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. 1. Berkeley. University of California Press, pp. 281–297.

Mallows, C. L. (1957). "Non-null ranking models". *Biometrika* 44, pp. 114–130.

Marot, G. and C.-D. Mayer (2009). "Sequential analysis for microarray data based on sensitivity and meta-analysis". *Stat Appl Genet Mol Biol* 8.Article 3.

Marot, Guillemette, Jean-Louis Foulley, Claus-Dieter Mayer, and Florence Jaffrézic (2009). "Moderated effect size and P-value combinations for microarray meta-analyses". *Bioinformatics* 25.20, pp. 2692–2699.

McCutcheon, A. C. (1987). *Latent Class Analysis*. Beverly Hills: Sage Publications.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley-Interscience.

McLachlan, G. et al. (2004). *Analyzing Microarray Gene Expression Data*. Wiley-Interscience.

Meinshausen N.and Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the Lasso". *Annals of Statistics* 34.3, pp. 1436–1462.

Metropolis, N. et al. (1953). "Equations of state calculations by fast computing machines". *Journal of Chemical Physics* 21.6, pp. 1087–1092.

Mo, Q. et al. (2013). "Pattern discovery and cancer gene identification in integrated cancer genomic data". *PNAS* 110.11, pp. 4245–4250.

Monneret, G. et al. (2015). "Estimation d'effets causaux dans les réseaux de régulation génique: vers la grande dimension". *Revue d'intelligence artificielle* 29.2, pp. 205–227.

— (2017). "Identification of marginal causal relationships in gene networks from observational and interventional expression data". *PLoS One* 12, e0171142.

Morlini, I. (2011). "A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model." *Advances in Data Analysis and Classification* 6.1, pp. 5–28.

Mortazavi, A. et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nature Methods* 5.7, pp. 621–628.

Nuel, G., A. Rau, and F. Jaffrézic (2013). "Using pairwise ordering preferences to estimate causal effects in gene expression from a mixture of observational and intervention experiments". *Quality Technology and Quantitative Management* 11.1, pp. 23–37.

Oshlack, A. and M. J. Wakefield (2009). "Transcript length bias in RNA-seq data confounds systems biology". *Biology Direct* 4.14.

Oshlack, A., M.D. Robinson, and M. D. Young (2010). "From RNA-seq reads to differential expression results". *Genome Biology* 11.220.

Owen, A. B. (2009). "Karl Pearson's meta-analysis revisited". *Annals of Statistics* 37.6B, pp. 3867–3892.

Pan, W. (2006). "Incorporating gene functions as priors in model-based clustering of microarray gene expression data". *Bioinformatics* 22.7, pp. 795–801.

Papastamoulis, P., M.-L. Martin-Magniette, and C. Maugis-Rabusseau (2016). "On the estimation of mixtures of Poisson regression models with large numbers of components". *Computational Statistics and Data Analysis* 93, pp. 919–106.

Pearl, J (2000a). *Causality: Models, Reasoning and Inference*. New York, NY, USA: Cambridge University Press.

Pearl, J. (2000b). "The logic of counterfactuals in causal inference". *Journal of the American Statistical Association* 95, pp. 428–435.

Pearson, K. (1934). "On a new method of determining 'goodness of fit'". *Biometrika* 26, pp. 425–442.

Pinna, A., N. Soranzo, and A. de la Fuente (2010). "From knockouts to networks: establishing direct cause-effect relationships through graph analysis". *PLoS ONE* 10.5, e12912.

Pinna, A. et al. (2013). "Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation". *BMC Systems Biology* 7, p. 73.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`.

Ramsköld, D. et al. (2009). "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data". *PLoS Computational Biology* 5.12.

Rau, A. and C. Maugis-Rabusseau (2017). "Transformation and model choice for RNA-seq co-expression analysis". *Briefings in Bioinformatics*. DOI: `doi:http://dx.doi.org/10.1101/065607.`.

Rau, A., G. Marot, and F. Jaffrézic (2014). "Differential meta-analysis of RNA-seq data". *BMC Bioinformatics* 15.91.

Rau, A. et al. (2015). "Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models". *Bioinformatics* 31.9, pp. 1420–1427.

Risso, D. et al. (2011). "GC-content normalization for RNA-seq data". *BMC Bioinformatics* 12.480.

Roberts, G.O., A. Gelman, and W.R. Gilks (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". *Annals of Applied Probability* 7, pp. 110–120.

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

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26, pp. 139–140.

Sam, L.T. et al. (2011). "A comparison of single molecule and amplification based sequencing of cancer transcriptomes". *PLoS One* 6.3, e17305.

Sauvage, C. et al. (2017). "Domestication rewired gene expression and nucleotide diversity patterns in tomato". *The Plant Journal* doi:10.1111/tpj.13592.

Schelldorfer, J. and P. Bühlmann (2014). "GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using L1-penalization". *Journal of Computational and Graphical Statistics* 23.2, pp. 460–477.

Schwarz, G. (1978). "Estimating the dimension of a model". *The Annals of Statistics* 6.2, pp. 461–464.

Shen, R., A. B. Olshen, and M. Ladanyi (2009). "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". *Bioinformatics* 25.22, pp. 2906–2912.

Si, Y. et al. (2014). "Model-based clustering for RNA-seq data". *Bioinformatics* 30.2, pp. 197–205.

Smyth, G. K. (2004). "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments". *Statistical Applications in Genetics and Molecular Biology* 1.3, pp. 1–26.

Spirtes, P., C. Glymour, and R. Scheines (2001). *Causation, Prediction, and Search*. Second. Cambridge, MA, USA: The MIT Press.

Steuer, R., P. Humburg, and J. Selbig (2006). "Validation and functional annotation of expression-based clusters based on gene ontology". *BMC Bioinformatics* 7, p. 380.

Stouffer, S.A. et al. (1949). *The American soldier. Adjustment during Army life*. Princeton, NJ: Princeton University Press.

Sultan, M. et al. (2008). "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome". *Science* 15.5891, pp. 956–60.

Tari, L., C. Baral, and S. Kim (2009). "Fuzzy c-means clustering with prior biological knowledge". *Journal of Biomedical Informatics* 42.1, pp. 74–81.

The FAANG Consortium et al. (2015). "Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project". *Genome Biology* 16.57.

Thomas, I., P. Frankhauser, and C. Biernacki (2008). "The fractal morphology of the built-up landscape". *Landscape of Urban Plan* 84.2, pp. 99–115.

Tibshirani, R. (1988). "Estimating transformations for regression via additivity and variance stabilization". *Journal of the American Statistical Association* 83, pp. 394–405.

— (1996). "Regression shrinkage and selection via the Lasso". *Journal of the Royal Statistical Society, Series B (Methodological)* 58.1, pp. 267–288.

Tipney, H. and L. Hunter (2010). "An introduction to effective use of enrichment analysis software". *Human Genomics* 4.3, p. 202.

Verbanck, M., S. Lê, and J. Pagès (2013). "A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data". *BMC Bioinformatics* 14.42.

Whittaker, J. (2009). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: `http://ggplot2.org`.

Yeung, K. Y. et al. (2001). "Model-based clustering and data transformations for gene expression data". *Bioinformatics* 17.10, pp. 977–987.

Ziemann, M. et al. (2015). "Digital Expression Explorer: A user-friendly repository of uniformly processed RNA-seq data". *ComBio2015*. Vol. POS-TUE-099. Melbourne. DOI: `10.13140/RG.2.1.1707.5926`.

Appendix A

Students supervised or co-supervised

Current Ph.D. and Master's students

- **Gilles Monneret** (Ph.D., co-supervision with F. Jaffrézic and G. Nuel)
  "Estimation of causal effects in gene networks from observational & intervention data"
  Université Pierre et Marie Curie (2014-2017, defense expected Fall 2017)

- **Raphaël Momal-Leisenring** (M2 intern)
  "Integrative statistical analysis of multi-omics data"
  École Nationale de la Statistique et de l'Analyse de l'Information (April-September 2017)

Alumni

- **Frédéric Juehl** (M2 intern, co-supervison with T. Zerjal)
  "Impact of heat stress on liver and blood transcriptomes of laying hens"
  Agrocampus Ouest (January-June 2017)

- **Manuel Revilla Sanchez** (3-month Ph.D. Erasmus+ Learning Mobility,
  co-supervison with J. Estelle and Y. Ramayo Caldas)
  "An integrative gene network analysis of the genetic determination of pig fatty acid composition"
  Universitat Autònoma de Barcelona (September-December 2016)

- **Babacar Ciss** (M2 intern, co-supervision with E. Sellem)
  "Constructing predictive models for ovine production data"
  Université de Pau (April-September 2016)

- **Dr. Mélina Gallopin** (Ph.D., co-supervision with G. Celeux and F. Jaffrézic)
  "Classification and network inference for RNA-seq data"
  Université Paris-Saclay (2012-2015)

- **Audrey Hulot** (M1 intern)
  "Incorporating a priori biological knowledge into gene network inference from observational and intervention gene expression data"
  École Nationale de la Statistique et de l'Analyse de l'Information (June-August 2015)

- **Meriem Benabbas** (M1 intern)
  "Identifying differentially expressed genes from RNA-seq data using mixtures of generalized linear models"
  Université Paris Descartes (June-August 2015)

- **Mélina Gallopin** (M2 intern, co-supervision with F. Jaffrézic and G. Celeux)
  "Gene network inference from RNA-seq data"
  Institut de Statistique de l'Université de Paris (April-September 2012)

- **Rémi Bancal** (2012, M2 intern, co-supervision with F. Jaffrézic and G. Nuel)
  "Gene network estimation by adaptive knockout experiments "
  Institut de Statistique de l'Université de Paris (April-September 2012)

Appendix B

List of publications

Books

[B1] Albert, I., Ancelet, S., David, O., Denis, J.-B., Makowski, D., Parent, É., **Rau, A.**, and Soubeyrand, S. (2015). *Initiation à la statistique bayésienne: Bases théoriques et applications en alimentation, environnmenet, épidémiologie et génétique*: Éditions Ellipses, collection références sciences.

Peer-reviewed articles

Statistical Methods

[A1] Monneret, G., Jaffrézic, F., **Rau, A.**, Zerjal, T. and Nuel, G. (2017) Identification of marginal causal relationships in gene networks from observational and interventional expression data. *PLoS One* 12(3): e0171142.

[A2] **Rau, A.** and Maugis-Rabusseau, C. (2017) Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*, doi: http://dx.doi.org/10.1101/065607.

[A3] Rigaill, G., Balzergue, S., Brunaud, V., Blondet, E., **Rau, A.**, Rogier, O., Caius, J., Maugis-Rabusseau, C., Soubigou-Taconnat, L., Aubourg, S., Lurin, C., Martin-Magniette, M.-L., and Delannoy, E. (2016) Synthetic datasets for the identification of key ingredients for RNA-seq differential analysis. *Briefings in Bioinformatics*, doi: https://doi.org/10.1093/bib/bbw092.

[A4] Gallopin, M., Celeux, G., Jaffrézic, F., **Rau, A.** (2015) A model selection criterion for model-based clustering of annotated gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 14(5): 413-428.

[A5] Monnert, G., Jaffrézic, F., **Rau, A.**, Nuel, G. (2015). Estimation d'effets causaux dans les réseaux de régulation génique: vers la grande dimension. *Revue d'intelligence artificielle*, 29(2): 205-227.

[A6] **Rau, A.**, Maugis-Rabusseau, C., Martin-Magniette, M.-L., Celeux, G. (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31(9): 1420-1427.

[A7] **Rau, A.**, Marot, G., and Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* 15:91.

[A8] Nuel, G., **Rau, A.**, and Jaffrézic, F. (2014) Using pairwise ordering preferences to estimate causal effects in gene expression from a mixture of observational and intervention experiments. *Quality Technology and Quantitative Management* 11(1):23-37.

[A9] **Rau, A.**, Jaffrézic, F., and Nuel, G. (2013) Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology* 7:111.

[A10] Gallopin, M. **Rau, A.**, and Jaffrézic, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS One* 8(10): e77503.

[A11] **Rau, A.**, Gallopin, M., Celeux, G., and Jaffrézic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 29(17): 2146-2152.

[A12] Dillies, M.-A.*, **Rau, A.***, Aubert, J.*, Hennequet-Antier, C.*, Jeanmougin, M.*, Servant, N.*, Keime, C.*, Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Charif, D., Le Crom, S.*, Guedj, M.*, and Jaffrézic, F*. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* 14(6) : 671-683.

* These authors contributed equally to this work.

[A13] **Rau, A.**, Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2012). Reverse engineering gene regulatory networks using approximate Bayesian computation. *Statistics and Computing*, 22: 1257-1271.

[A14] **Rau, A.**, Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2010). An empirical Bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology*: Vol. 9: Iss. 1, Article 9.

Statistical Applications

[A15] Sauvage, C., **Rau, A.**, Aichholz, C., Chaoeuf, J., Sarah, G., Ruiz, M., Santoni, S., Causse, M., David, J., and Glémin, S. (2017) Domestication rewired gene expression and nucleotide diversity pattern in the tomato. *The Plant Journal* (to appear).

[A16] Endale Ahanda, M.-L., Zerjal, T., Dhorne-Pollet, S., **Rau, A.**, Cooksey, A., and Giuffra, E. (2014) Impact of the genetic background on the composition of the chicken plasma miRNome in response to a stress. *PLoS One*, 9(12): e114598.

[A17] Brenault, P., Lefevre, L. **Rau, A.**, Laloë, D., Pisoni, G., Moroni, P., Bevilacquia, C. and Martin, P. (2014) Contribution of mammary epithelial cells to the immune response during early stages of a bacterial infection to *Staphylococcus aureus*. *Veterinary Research* 45:16.

[A18] Furth, A., Mandrekar, S., Tan, A. **Rau, A.**, Felten, S., Ames, M. Adjei, A. Erlichman, C. and Reid, J. (2008). A limited sample model to predict area under the drug concentration curve for 17-(allylamino)-17-demethoxygeldanamycin and its active metabolite 17-(amino)-17-demethoxygeldanomycin. *Cancer Chemotherapy Pharmacology* 61(1): 39-45.

Peer-reviewed conference proceedings

[CP1] **Rau, A.**, Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2010). Approximate Bayesian approaches for reverse engineering biological networks. *Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture*. Manhattan, Kansas.

Book chapters

[BC1] Martin-Magniette, M.-L., Maugis-Rabusseau, C., **Rau, A.** (2016) Clustering of co-expressed genes. In: Choix et agrégation de modèles: Journée d'Etudes Statistiques (to appear).

Pre-prints, technical reports, and submitted articles

[PP1] Godichon-Baggioni, A., Maugis-Rabusseau, C. and **Rau, A.** (2017) Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data (submitted). *arXiv*:1704.06150

[PP2] Nuel, G., **Rau, A.**, and Jaffrézic, F. (2013). Joint likelihood calculation for intervention and observational data from a Gaussian Bayesian network. *arXiv preprint* arXiv:1305.0709.

[PP3] **Rau, A.**, Celeux, G., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2011). Clustering high-throughput sequencing data with Poisson mixture models. *Inria Research Report* 7786.

R packages

[R1] `coseq`: Co-expression analysis of sequencing data
Available on Bioconductor at `https://bioconductor.org/packages/coseq`

[R2] `ICAL`: Model selection for model based clustering of annotated data
Available on Github at `https://github.com/Gallopin/ICAL`

[R3] `metaRNASeq`: Meta-analysis of RNA-seq data
Available on CRAN at `http://cran.r-project.org/web/packages/metaRNASeq`

[R4] `HTSFilter`: Filter for replicated high-throughput sequencing data
Available on Bioconductor at `www.bioconductor.org/packages/HTSFilter`

[R5] `HTSCluster`: Clustering high throughput sequencing data
Available on CRAN at `http://cran.r-project.org/web/packages/HTSCluster`

[R6] `ebdbNet`: Empirical Bayes estimation for Dynamic Bayesian Networks
Available on CRAN at `http://cran.r-project.org/web/packages/ebdbNet`

Theses

[T1] **Rau, A.** (2010). Reverse engineering gene networks using genomic time-course data. Ph.D. thesis, Purdue University (West Lafayette, Indiana, USA).

Acknowledgements

I would first like to warmly thank David Causeur, Anne-Laure Boulesteix, and Nathalie Villa-Vialaneix for agreeing to be the *rapporteurs* for this manuscript, as well as Stéphane Robin, Franck Picard, and Christophe Ambroise for agreeing to be members of my evaluation committee. I am extremely thankful to all of you for your time and energy.

Thanks also to all of my colleagues in the GABI research unit at INRA, our director, Claire Rogel-Gaillard, and particularly the members of the (extended) PSGen team for their collaboration, support, discussions over coffee, and friendship – you have made it a real pleasure for me to come to work every day. A special thank you goes to Florence Jaffrézic for having guided me from my earliest research days as a new Ph.D. student through my first six years as a research scientist at INRA and beyond, and for teaching me the valuable lesson of how to be a "closer". Thanks to Denis Laloë, our fearless leader of PSGen, a constant source of support and good humour, and an enthusiastic user of R who graciously initiated me into the world of multivariate statistics. Thanks also to Tatiana Zerjal for many deep discussions on ANOVA sums of squares and microarray probe summarization techniques, but more importantly for your support and friendship. I would also like to acknowledge our support staff at GABI, in particular Yvelise Fricot, Alexandra Vincent, Nathalie Lenoir, and Sylvie Manga-Akoa, who help make the administrative side of research run smoothly, and to my other collaborators in Jouy en Josas, including Jordi Estelle, Yuliaxis Ramayo Caldas, and Eli Sellem.

The work I have accomplished during this first phase of my research career was made possible thanks only to my students, fellow co-supervisors, and collaborators. An extra special thank you goes to Mélina Gallopin as my first Ph.D. student – it is very gratifying to see you successfully set off on your own independent research career. To Gilles Monneret, I wish you all the best in this last home stretch for your Ph.D. work – you're almost there! Thanks also to all of my past and current Master's interns and visiting students: Manuel Revilla Sanchez, Babacar Ciss, Audrey Hulot, Meriem Benabbas, Rémi Bancal, Fréderic Jehl, and Raphaëlle Momal-Leisenring. I am very appreciative that Gilles Celeux, Marie-Laure Martin-Magniette, and Cathy Maugis-Rabusseau introduced me to the world of mixture models, and allowed me to introduce them to the world of RNA-seq data. Thanks also to Grégory Nuel, Guillemette Marot, Christopher Sauvage, Sandrine Laguarrigue, Antoine Godichon-Baggioni, and the BioBayes team (Isabelle Albert, Sophie Ancelet, Olivier David, Jean-Baptiste Denis, David Makowski, Éric Parent, and Samuel Soubeyrand). Thanks to Paul L. Auer and the Zilber School of Public Health for graciously hosting me as a Visiting Scholar for 6 weeks in 2016 at the University of Wisconsin-Milwaukee. The Statomique group has provided a consistent source of rich discussions and ideas, especially Marie-Agnès Dillies, Julie Aubert, and Christelle Henequet-Antier. I am also grateful to Brigitte Gelein for providing me the opportunity to teach the statistical genomics class for the biostatistics students at Ensai.

I would also like to say a word of thanks to Rebecca W. Doerge and Jean-Louis Foulley for having been supportive and committed advisors to me during my Ph.D. work – I am doing my best to live up to your example.

To my family and friends, near and far – none of this would have been possible without your love and encouragement. Thanks especially to my brother Paul, my sister Kris, and my new brother Peter for your support and visits over the past few years. To my parents, thanks for the continued brain waves to psych me up – they even work across the ocean! None of this would have been possible without the love and unconditional support of Grégoire – thank you for believing in me. And last but not least, to my silly, wiggly, wonderful Elise – I love you to the moon and back.