Statistical Modeling and Inference for Genomics

The work in this thesis was financially supported by the Centre for Medical Systems Biology (CMSB), established by the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO), and carried out under the auspices of the VU University Amsterdam.

The printing of this thesis was kindly supported by the Centre for Medical Systems Biology (CMSB) and the VU University Amsterdam.

VU University *Amsterdam*

CENTRE FOR
Medical Systems Biology

ISBN: 978-90-9028117-9

Document prepared with LaTeX 2ε and typeset by pdfTeX (charter font)
Printed by: Mostert & Van Onderen (www.drukkerijmostert.nl)

VRIJE UNIVERSITEIT

Statistical Modeling and Inference for Genomics

Data Integration, Shrinkage and Network Reconstruction

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op vrijdag 28 maart 2014 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Gwenaël Gérard René Leday

geboren te Rennes, Frankrijk

promotoren: prof.dr. A.W. van der Vaart
prof.dr.ir. M.A. van de Wiel

Contents

CHAPTER 1

Introduction

This thesis develops statistical models and inference procedures to address biological questions that arise from the analysis of microarray and sequencing data. Statistical analysis of such data is a difficult task, mainly because there are relatively few observations on many features. To deal with the scarcity of data regularization is often needed to produce good statistical estimators. Another difficulty is that the biological process under study can be complex, so care must be taken in choosing a (family of) statistical model(s). The following chapters treat the subjects of *integration* of different molecular data types for statistical modeling, *shrinkage* for the borrowing of strength in the analysis of high-dimensional data and *network reconstruction* to decipher dependencies between functional biological units (e.g. genes). This introduction offers the general reader a small detour through some aspects of biology and experimental data generation to approach the other chapters with more ease. It concludes with an outline of the thesis.

1.1 General background. The face of biology has changed tremendously with the emergence of technologies that allow the parallel measurement of thousands of biological sequences (such as DNA, RNA or protein sequences). These high-throughput technologies (such as microarrays and next-generation sequencing) have produced massive amounts of data that have proven valuable to researchers in understanding molecular and cellular processes. For example, in cancer research many new dysfunctions have been pinpointed as drivers for 'what goes wrong' in tumour cells and as potential targets for molecular therapy. Collected data have also helped in defining molecular signatures and characterizing the heterogeneity of cancer tissues, which are important for diagnosis and prognosis. These biotechnologies have revolutionized research in many disciplines and promise to further our understanding of complex diseases.

The large amounts of data produced by these technologies have complicated their organization and utilization. These problems have motivated a surge of interest in fields such as computer science and statistics where many new developments have occurred. Interestingly, the two disciplines have become particularly intertwined. Computer scientists are increasingly led to develop new tools that employ statistical techniques and statisticians are increasingly concerned with the computational aspects of their methods. This interplay has spawned new interdisciplinary fields such as bioinformatics, which organizes and exploits information regarding biological sequences and molecules. Similarly, statistical genomics has emerged as a new and rapidly expanding field that develops statistical procedures to answer research questions that

arise from the analysis of microarray and sequencing data. The emergence of new biotechnologies has brought new areas of research that will likely expand in the future with advancement of science and technology.

Statistical methods are constantly challenged by the analysis of genomic data. This is mainly because high-throughput technologies are diverse and fast-evolving. Data types, processing and analysis may greatly differ between technologies and are expected to change with new technological developments. A prime example of such change is the current transition from microarray-based to sequencing-based experiments. These two types of experiments result in completely different data types. In statistical words, in the former case data are (after preprocessing) continuous and conveniently assumed to be approximately Gaussian whereas in the latter case data are discrete and usually taken to follow a Poisson or negative binomial distribution. It is still unclear which discrete distribution is most appropriate, how to model overdispersion or even whether zero-inflation should be accounted for. This may actually depend on the technology used. The surge of sequence data brings new challenges to the statistician for whom stepping out the 'normal' world can be delicate. This is especially true when answering difficult problems such as network reconstruction. Methodologies for such data are currently under active research.

The present thesis intends to contribute to the development of statistical methodologies for the analysis of microarray and sequencing data. Particularly, it focuses on differential gene expression analysis, the integration of DNA copy number and gene expression data, and gene network analysis. To understand the specific nature of data being used, we briefly give an account of molecular biology and technologies that generate them. Readers interested in a more complete introduction to the field molecular biology are referred to Hunter (2009) and Strachan and Read (2010).

1.2 Cell biology. The genetic information contained in the cell is responsible for the incredible diversity of functions that it can carry out. This information is encoded in the DNA, a double-helix shaped molecule, and reflected in the order in which a set of only four nucleotide bases (adenine (A), thymine (T), guanine (G), and cytosine (C)) appear (see Figure 1.1). The two strands of the DNA helix are complementary, which means that the knowledge of one can be used to determine the other. For example, if the linear sequence of nucleotide bases on one strand is 'AGACTG', its complementary sequence on the opposite strand will be 'TCTGAC'. This is because nucleotide bases go hand in hand with each other: base A always 'pairs' with base T, and base C with G. This property is very useful in practice. For example, it is commonly used in laboratory to amplify DNA (Strachan and Read, 2010).

The information contained in the DNA is essentially the same in all cells of an individual. Yet, they all carry out very different functions. Mechanisms yielding to these functions are very complex and not yet fully understood. The most simplified picture of these mechanisms is provided by the *central dogma* of biology (Crick, 1958, Figure 1.2), which stipulates that pieces of DNA are first transcribed into a molecule called messenger RNA (mRNA) that vehicle the information outside the nucleus in the cytoplasm, to be translated into proteins, the final products that carry out most

functions in the cell. As opposed to DNA, mRNA is a single-stranded molecule which has a uracil (U) nucleotide in place of T and which linear sequence represents disjoint sets of nucleotide triplets (codons) that encode for specific amino acids. The sequence of amino acids defines in turn the protein and its structure. The cell function is mainly determined by the types of proteins that are at work.

High-throughput techniques such as microarray help in understanding the cell biology by collecting information on molecular levels such DNA or RNA. We next discuss these technologies and the data they generate.

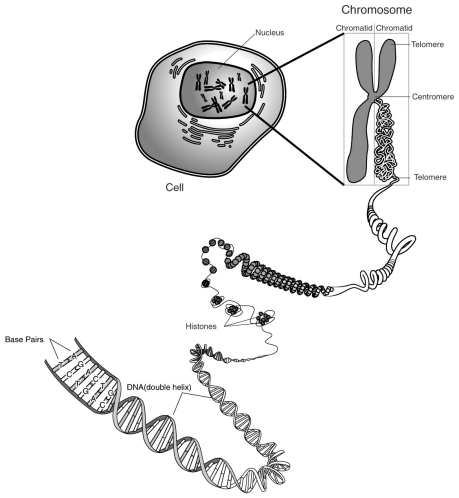


FIGURE 1.1. *Schematic representation of the cell, its nucleus and the chromosomes it contains. The figure illustrates how the DNA is organized and highly packed into chromosomes. Image courtesy of the National Human Genome Research Institute.*

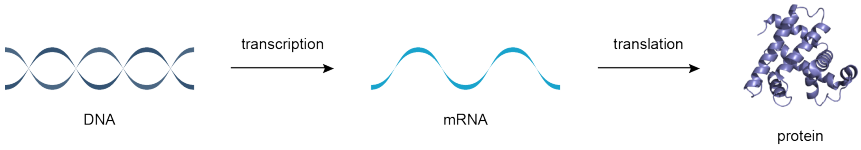
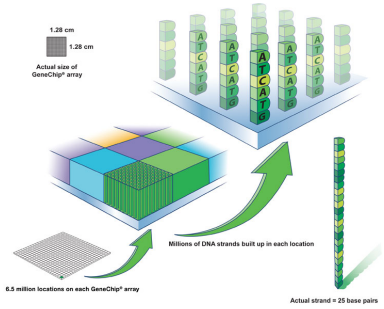


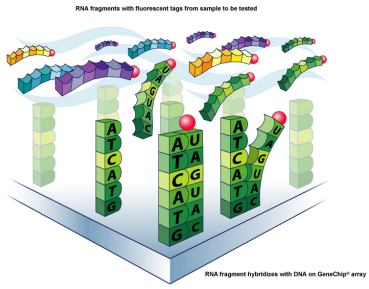
FIGURE 1.2. *Illustration of the central dogma of molecular biology. Image courtesy of atdbio (www.atdbio.com).*

1.3 Experimental data. A microarray is a measurement device that can quantify in parallel tens of thousands of predefined biological sequences (such as DNA or mRNA sequences) that have been immobilized on a solid support (see Figure 1.3(a)). These sequences are usually chosen so they correspond to known functional units. In gene expression arrays known sequences of genes are used. Then, the abundance of mRNA transcripts that have been measured for a specific sequence is used to quantify the level of 'expression' for the corresponding gene.

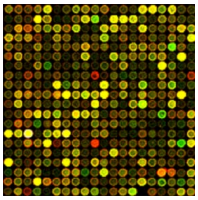
The microarray experiment consists in fixing the mRNA sequences present in a sample on the surface of the microarray chip. Briefly, RNA is first isolated from the tissue sample and labelled with a fluorescent dye. In case of more than one sample (e.g. treatment versus control or normal versus cancer), sequences are distinguished by different fluorescent dyes. Using appropriate conditions (such as temperature), the sample(s) are then hybridized on the microarray chip (see Figure 1.3(b)). After washing off non-hybridized strands, the amounts of labelled hybridized strands are quantified by optical fluorescence intensities (Figure 1.3(c)) using laser scanning. The $\log_2$-values of intensities, as measured under the dye spectrum, are taken as final measures of the levels of expression for each element on the array.



(a) Array



(b) Hybridization



(c) Optical intensities

FIGURE 1.3. *Illustration of the main steps of a microarray experiment: (a) array preparation, (b) hybridization and (c) quantification of optical fluorescence intensities. Image courtesy of Affymetrix (www.affymetrix.com).*

Before being used for analysis, microarray data need to be pre-processed in order to remove experimental artifacts that may have been introduced during the experiment. Array intensities also need to be adjusted so that meaningful biological comparisons can be made. There exists many approaches to within- and between-array normalization (Quackenbush, 2002).

Next generation sequencing is quickly replacing microarrays as a promising technique to probe different molecular levels of the cell. The technology presents various advantages over microarrays that make it attractive ('t Hoen et al., 2008, Wang et al., 2009): it does not rely on the existing knowledge of genomic sequences, has a better dynamic range of expression levels (good sensitivity at the lower end of the spectrum and no upper limit for quantification) and allows to look at coding and non-coding RNA, splicing and allele specific expression.

The experiment typically consists in isolating RNA from the tissue sample and converting it into complementary DNA (cDNA) fragments. These are then sequenced using high-throughput DNA sequencing methods, which produce counts of short sequences or reads. These reads are finally mapped to a reference genome and summarized. Different (non-Sanger-based) sequencing strategies can be adopted. See Metzker (2010) for a description of those. The final output of the experiment produces count data in the form of aligned read-counts. As for microarray, these counts need to be normalized before analysis (Robinson and Oshlack, 2010).

1.4 Outline of the thesis. The remainder of the thesis is divided into five chapters that we briefly describe here.

Chapter 2 presents a flexible class of models to decipher how DNA copy number abnormalities in cancer cells alter the mRNA gene expression level. This class of models aims to reflect the biological mechanism operating between these two molecular levels and help in identifying relevant markers. We motivate the use of piecewise linear regression splines with biologically motivated constraints on parameters to model associations. Because model estimation and selection is difficult in this context, the chapter provides methodology for testing the effect of DNA on mRNA, identifying the appropriate model and obtaining uniform confidence bands that incorporates model uncertainty. Using two real data sets, it is illustrated that flexible models may bring more insight in the interaction between the two markers.

Chapter 3 presents the `R` package `PLRS`, which implements the statistical framework introduced in chapter 2. The method is illustrated on an additional real data set from The Cancer Genome Atlas (TCGA). On such a data set, the need for flexible models is particularly pronounced.

Chapter 4 develops a Bayesian method for differential gene expression analysis using next generation sequencing data. The method is particularly useful for its large flexibility of the likelihood count model and its ability to handle complex designs while accommodating multi-parameter shrinkage. An empirical Bayes procedure for estimating parameters of priors is introduced and different types of (non-) parametric priors are discussed along with Bayesian corrections for multiplicity. The chapter and its

appendix present various model- and data-based simulations that validate the performance of the approach in detecting true differences. In particular, compared to other methods, results are shown to be more reproducible on real data.

Chapter 5 studies differences in gene expression between brain regions in aged human. In this work, the contribution lies in the differential expression analysis using CAGE data.

Finally, chapter 6 studies network reconstruction using a computationally attractive Bayesian structural equation model (SEM). It is argued that regularization by means of Gaussian priors coupled with *a posteriori* edge selection is a simple and attractive alternative to sparse priors. A novelty of the approach is the use of shrinkage priors that borrow information across equations. In simulations, it is demonstrated that the empirical Bayes procedure of chapter 4 is appropriate in this context and that shrinkage priors can substantially improve graph structure recovery. The Bayesian SEM is also shown to outperform popular sparse methods in various settings.

We now dwell on the relations between the chapters.

Chapter 2 and 3 focus on delineating the direct (in cis) transcriptional effects of copy number aberrations. As a natural extension, Chapter 6 originally aimed at introducing a Bayesian SEM for the joint estimation of direct and indirect (in trans) transcriptional effects of copy number aberrations, hence resulting in network reconstruction when incorporating genetic perturbations. Due to time considerations, chapter 6 focuses on network reconstruction only using some of the methodology developed in chapter 4. The incorporation of perturbations such as copy number aberrations was left for future research.

The work in chapter 4 was motivated by the complexity of the experimental design in chapter 5 and the lack of appropriate methods in statistical literature. Hence both works have been conducted in parallel, which explains the difference between the methodology developed in chapter 4 and the one used in chapter 5.

In all, this thesis presents diverse topics and approaches, which reflects the broadness of the statistical genomics field.

CHAPTER 2

Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines

DNA copy number and mRNA expression are widely used data types in cancer studies, which combined provide more insight than separately. Whereas in existing literature the form of the relationship between these two types of markers is fixed a-priori, in this paper we model their association. We employ piecewise linear regression splines (PLRS), which combine good interpretation with sufficient flexibility to identify any plausible type of relationship. The specification of the model leads to estimation and model selection in a constrained, nonstandard setting. We provide methodology for testing the effect of DNA on mRNA and choosing the appropriate model. Furthermore, we present a novel approach to obtain reliable confidence bands for constrained PLRS, which incorporates model uncertainty. The procedures are applied to colorectal and breast cancer data. Common assumptions are found to be potentially misleading for biologically relevant genes. More flexible models may bring more insight in the interaction between the two markers.

This chapter was published as:
Leday, G.G.R., van der Vaart, A.W., van Wieringen, W.N., and van de Wiel, M.A. (2013). Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines. *Ann. Appl. Stat.*, 7(2):823-845.

2.1 Introduction. The genetic material of the human cancer cells often exhibits abnormalities, of which DNA copy number aberrations are a prime example. These aberrations comprise gains and losses of chromosome pieces that are highly variable in size. Thereby, all or parts of a chromosome may have more or less than the two copies received from the parents. Abnormal DNA copy numbers (different from two) may alter expression levels of mRNA transcripts (encoding for functional proteins) that map to the aberration's genomic location. Apart from being concordant (copy number tends to correlate positively with expression level), the form of this association is not established and may even vary per gene. In this paper we use high-throughput data available for tissue-specific samples from unrelated patients to study the relationship between copy number (DNA) and gene expression (mRNA). We employ a wide class of interpretable models to reflect the biological mechanism operating between these two molecular levels and identify relevant markers that may serve as therapeutic targets.

DNA copy number aberrations are often measured by array comparative genomic hybridization (aCGH) (Pinkel and Albertson, 2005). This measuring device is similar to expression microarrays, which measure expression levels of thousands of genes simultaneously but interrogate DNA rather than RNA. Thereby, both profiling experiments produce a continuous value for every element/probe on the array: a $\log_2$-value of optical fluorescence intensity. Although experiments appear similar, types of information differ and so are their subsequent treatment. To understand the specific nature of these data we include a description of their processing.

Normalization of mRNA expression profiles (Quackenbush, 2002) consists in removing experimental artifacts (such as array differences, means, scales) and yields, for every gene on each array, a continuous value (normalized $\log_2$-value) which represents the amount of the gene's transcript present in the sample. Preprocessing of copy number/aCGH profiles aims to characterize the genomic instability of each tumor sample and show deleted/duplicated pieces of chromosomes. Three successive steps (illustrated in Figure 2.1) are typically executed to recover the aberration states of all probes (van de Wiel et al., 2011). Through these steps, the size, genomic position and type of copy number aberrations are determined for all samples. First preprocessing step, the *normalization* of $\log_2$-values removes technical or biological artifacts (such as tumor sample contamination, GC content) and makes the data comparable across samples. Next *segmentation* partitions the genome of each sample into segments of

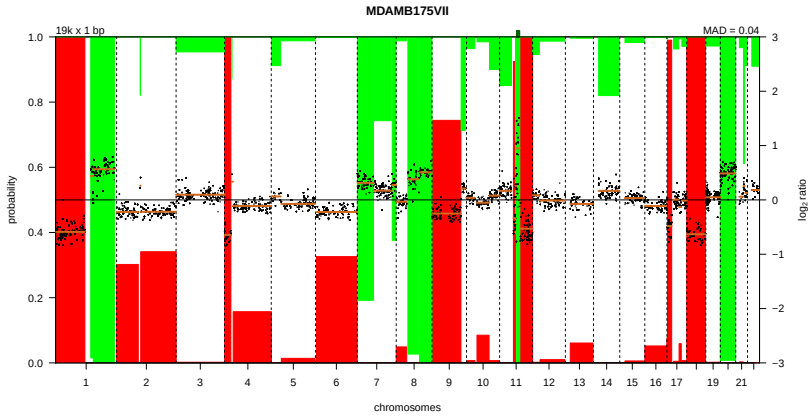
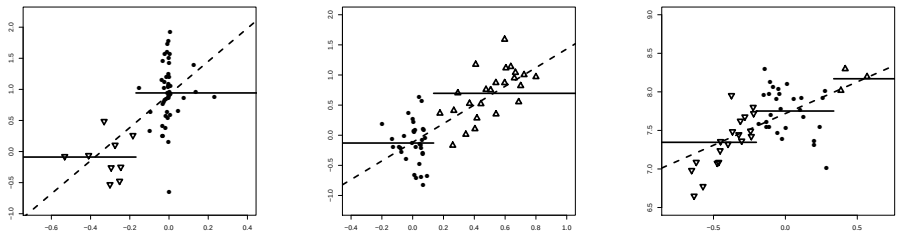


FIGURE 2.1. *Plot of a copy number/aCGH profile from the breast cancer data set (Neve et al., 2006) showing the different preprocessing steps. Probes on the array are genomically ordered on the x-axis (only the chromosome number is displayed). Black dots and orange segments indicate the normalized and segmented $\log_2$-values (right y-axis), respectively. Bars represent "loss" (red) and "gain" (green and reversed) membership probabilities (left y-axis). Amplifications are indicated by tick marks on the top axis.*

constant $\log_2$-values. These segments are considered a smoothed (and thus de-noised) version of their normalized counterparts. Segmentation is motivated by the biological breakpoint process on the DNA that may cause differential copy number between

neighbouring locations. Finally *calling* assigns an aberration state to each segment. Probabilistic calling, usually based on mixture models, results in a probability distribution over a set of ordered possible types of genomic aberrations (which we will refer to as states), typically comprising "loss" ($<$ 2 copies), "normal" ($=$ 2 copies), "gain" (3-4 copies) and "amplification" ($>$ 4 copies). A state is attributed to each probe using a classification rule on the membership probabilities. Non-probabilistic calling directly assigns states to segmented values, e.g. by using a threshold. Note that larger segmented values almost always correspond to larger or equal called copy number (see Figure 2.1). All in all, the three steps of the preprocessing procedure provide distinct, but strongly related, data sets: 1) the normalized, 2) segmented and 3) called aCGH data. While most down-stream analyses use either segmented or called data, we use them jointly.

Current methodology for integrative genomic studies assumes rather than explores the mathematical form of the relationship between copy number and expression level. The relationship is said to be either linear or stepwise (see examples in Figure 2.2). A linear relationship is often assumed in combination with segmented aCGH data. For instance, the strength of the DNA-mRNA association is measured by a (modified) correlation coefficient (Lee et al., 2008, Lipson et al., 2004, Salari et al., 2010, Schäfer et al., 2009). Alternatively, a linear regression approach is entertained (Asimit et al., 2011, Gu et al., 2008, Menezes et al., 2009). Recently published multivariate methods (Jörnsten et al., 2011, Peng et al., 2010b, Soneson et al., 2010, van Wieringen et al., 2010) also assume linearity. A piecewise DNA-mRNA relationship is considered when using the called aCGH data for integrative analysis. van Wieringen and van de Wiel (2009) and Bicciato et al. (2009) have proposed stepwise methods.



(a) MUTS1 (b) C20orf24 (c) PITPNA

FIGURE 2.2. *Illustration of the association between DNA and mRNA for three genes in the breast cancer data set (Neve et al., 2006) used in this study. Segmented copy number is on x-axis while gene expression is on y-axis. Symbols indicate the different states, namely loss ($\bigtriangledown$), normal ($\bigcirc$) and gain ($\bigtriangleup$). The dashed and "continuous" lines give the fitted linear and stepwise model, respectively.*

In this paper we develop model selection for piecewise linear regression splines (PLRS) to decipher how DNA copy number abnormalities alter the mRNA gene expression level. In addition, we propose a statistical test that accounts for model uncer-

tainty in the PLRS context to detect those genes that drive important shifts. The PLRS framework encompasses the linear and stepwise relationships, but provides flexibility, while maintaining good interpretability. In particular, it accommodates *differential* DNA-mRNA relationships across states. This is biologically plausible, because the cell has various post-transcriptional mechanisms to undo the effects of DNA aberrations. For a given gene, the efficacy of such mechanisms is likely to differ between gains and losses. E.g. a gain can directly be compensated by regulatory mechanisms that cause mRNA degradation, such as methylation. On the other hand, a complete loss of both DNA copies (which is more rare than partial loss) cannot be compensated at all.

Segmented and called data are incorporated into the analysis, and biologically motivated constraints are imposed on the model parameters. As this makes model selection and inference nonstandard, we provide methodology for testing the effect of DNA on mRNA within the context of PLRS and for selecting the appropriate model. We also present a novel and computationally inexpensive method for obtaining uniform confidence bands. We apply the proposed methodology to colorectal and breast cancer data sets, where we identify many genes exhibiting non-standard behavior.

2.2 Methods. We model the association between DNA copy number and mRNA expression by piecewise linear regression splines (PLRS), with biologically motivated constraints on the coefficients. In this section we address model selection and describe a modified Akaike criterion in this context. Further we present a method for determining uniform confidence bands, along with a statistical test for the effect of copy number on mRNA expression.

2.2.1 Model. Consider gene expression and aCGH profiling of n independent tumor samples where for a given gene $\{y_i, x_i, s_i\}_{i=1}^n$ are available, with y_i being the normalized mRNA expression ($\log_2$ scale), x_i the segmented copy number ($\log_2$ scale) and s_i the copy number state ("loss", "normal", "gain" and "amplification", coded by -1, 0, 1 and 2) value of the ith observation, respectively. Then, the "full" model with S states (or parts) takes the form:

$$y_i = f_\alpha(x_i;\theta) + \epsilon_i = \theta_0 + \theta_1 x_i + \sum_{j=1}^{S-1}\sum_{d=0}^{1} \theta_{j,d} \left(x_i - \alpha_j\right)_+^d + \epsilon_i. \tag{2.1}$$

Here $\theta = \{\theta_0, \theta_1, \theta_{1,0}, \ldots, \theta_{S-1,0}, \theta_{1,1}, \ldots, \theta_{S-1,1}\}$ is a vector of $2 \times S$ unknown parameters, the ϵ_i are independent random variables each normally distributed with mean 0 and variance σ^2, and $\{\alpha_j\}$ are $S-1$ known *knots*. The quantity $(a)_+^d$ represents the positive part $\max(a, 0)$ of a raised to the power d. The number of aberration states S varies across genes. In this study no more than four different aberration states are considered ($S \leq 4$). Below, for the purpose of discussing model (2.1) we consider the general case $S = 4$.

Knots $\{\alpha_j\}$ are obtained using data from the *calling* preprocessing step. Depending on the type of calling, two possibilities present themselves. First, consider non-

probabilistic calling which renders states $\{s_i\}_{i=1}^n$. Then, α_j is taken to be the midpoint of the interval between segmented values x_i belonging to consecutive states (method I). This makes the (natural) supposition that the calling values respect the ordering of the segmented values x_i, and should be reasonably precise if the between-state intervals are small, which is typical (see Figure 2.2). Second, consider probabilistic calling, which renders membership (or call) probabilities: $(p_{i,-1}, p_{i,0}, p_{i,1}, p_{i,2})$. These reflect the plausibility of the segmented value x_i to belong to the states $s_i \in \{-1,0,1,2\}$ (van de Wiel et al. (2007)). Then for $j \in \{1,2,3\}$, we estimate α_j (method II) by

$$\hat{\alpha}_j = \arg\max_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} p_{i,j(i,\alpha)}, \qquad j(i,\alpha) = \begin{cases} j-2 & \text{if} \quad x_i \leq \alpha \\ j-1 & \text{if} \quad x_i > \alpha \end{cases}. \tag{2.2}$$

For instance, α_2 is the knot between states 0 and 1. To determine its position we select for each sample its plausibility $p_{i,0}$ of belonging to state 0 (when $x_i \leq \alpha_2$) or $p_{i,1}$ of belonging to state 1 (when $x_i > \alpha_2$), and add over all samples. We select α_2 to maximize the sum. The maximum may not be unique but described by a small interval; in such a case, we use the corresponding midpoint. This method may be preferable as it accounts for the uncertainty of the calling states. The two methods taken here use data as provided by available *calling* algorithms. Proposed models for this preprocessing step typically depend on data from *all* samples, which stabilizes the estimation of α_j. Furthermore, knots are to be interpreted as boundaries between the (ordered) states $\{-1,0,1,2\}$, which gives us strong a priori knowledge as to their placing (see Figure 2.2). Together, these two arguments support our approach to consider knots in model (2.1) as being known. In Appendix A.2, a simulation shows that standard deviations of $\hat{\alpha}_j$ are indeed very small.

Model (2.1) contains seven basis functions besides the intercept θ_0 and hence is quite flexible. Our approach is to select appropriate basis functions ($2^7 = 128$ possible models) and estimate the parameters. The basis functions of degree zero $x \mapsto (x - \alpha)_+^0$ model discontinuities, and hence allow for a different effect of copy number on expression for each state.

This framework is a natural fundament to test meaningful hypotheses. For example, the hypothesis that for a given state there is an effect of copy number on mRNA can be expressed in terms of a linear function of the parameters being zero ($\sum_j \theta_{j,1} = 0$); a difference between the effects of two adjacent states corresponds to knot deletion. The submodel consisting of piecewise constant functions (without the functions $x \mapsto x$ and $x \mapsto (x-\alpha)_+^1$) allows testing the difference in expression between states based on discrete genomic information.

To increase biological plausibility, aid interpretation and increase the stability of estimation we impose a set of linear constraints on the parameters. As it is generally believed that direct causal effects of DNA on mRNA should be positive, we constrain all slopes to be non-negative. More exactly, we constrain the slope corresponding to the "normal" state to be non-negative ($\theta_1 + \theta_{1,1} \geq 0$), while others are forced to be at least equal to the latter (implied by $\theta_{1,1} \leq 0$ for losses, $\theta_{2,1} \geq 0$ for gains and $\theta_{2,1} + \theta_{3,1} \geq 0$ for amplifications). For the same reason we constrain jumps $\theta_{j,0}$ from state to state to be non-negative. Note that the restrictions adopted here force the

slope of the "normal" state to be small or null and make the natural assumption that a normal copy number is not expected to affect (at least severely) gene expression.

The maximum likelihood estimator of the unknown vector of coefficients θ solves the following convex optimization problem:

$$\underset{\theta}{\text{minimize}} \ \left(y-\mathbf{X}\theta\right)^{\mathrm{T}}\left(y-\mathbf{X}\theta\right) \qquad \text{subject to} \ \ \mathbf{C}\theta \geq 0. \tag{2.3}$$

This can be solved by quadratic programming (Boyd and Vandenberghe, 2004). The vector $y = \{y_1, \ldots, y_n\}$ denotes the expression signature of a given gene and $\mathbf{X}$ the associated matrix of covariates designed according to (2.1). The full row-rank matrix $\mathbf{C}$ expresses the constraints that are imposed on the parameters. For the 4-state full model we define $\mathbf{C}$ as the matrix in:

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_{1,0} \\ \theta_{1,1} \\ \theta_{2,0} \\ \theta_{2,1} \\ \theta_{3,0} \\ \theta_{3,1} \end{pmatrix} \geq 0. \tag{2.4}$$

2.2.2 Model selection. Given R competing statistical models, with log-likelihoods $\mathcal{L}_r(\theta_r)$, based on a $k_r \times 1$ parameter vector θ_r and with corresponding maximum likelihood estimators (MLE) $\hat{\theta}_r$, the *Akaike information criterion* (AIC) selects as best the model that minimizes

$$\mathrm{AIC}_r = -\mathcal{L}_r(\hat{\theta}_r) + k_r, \ \forall r \in \{1, \ldots, \mathrm{R}\}. \tag{2.5}$$

This information criterion consists of two parts: the negative maximized log-likelihood, which measures the lack of model fit, and a penalty for model complexity. Although AIC has found wide application, it is less suitable for models that include parameter constraints, as in our situation. It can be adapted as follows.

The original motivation for the criterion (Akaike, 1973) is to choose the model that minimizes the Kullback-Leibler (KL) divergence to the true distribution of the data. Indeed, the criterion AIC_r is (under some conditions) an asymptotically unbiased estimator of this KL divergence. The likelihood at a given parameter is an unbiased estimate of the KL divergence at this parameter, but evaluating it at the maximum likelihood estimator introduces a bias caused by "using the data twice", which is compensated by the penalty k_r (Bozdogan, 1987). In the constrained case (i.e., subject to $\mathbf{C}\theta \geq 0$) we can follow the same motivation, but must account for a different behaviour of the maximum likelihood estimator and the resulting bias. Intuitively, the penalty adjusts for an expected increase in the maximized log-likelihood when variables are added to the model, which is less likely under constraints. The likelihood of violation of the constraints must be taken into account.

Hughes and King (2003) adapted the AIC criterion using the asymptotic distribution of the Wald test statistic. In the constrained situation this statistic is not distributed as a chi-squared random variable anymore, but as a probability weighted mixture of chi-squared random variables (see Chernoff (1954), Gouriéroux et al. (1982), Kodde and Palm (1986), or van der Vaart (1998, Theorem 16.7)). It is of the form (partially inequality constrained Wald statistic):

$$\sum_{h=0}^{p_r} w(p_r,h)\chi^2(k_r-p_r+h), \tag{2.6}$$

where p_r is the number of inequality constraints and $w(p_r,h)$ are weights summing to one, which can be interpreted as the probabilities under the null hypothesis that the constrained maximum likelihood estimator $\widetilde{\theta}_r$ satisfies h out of p_r constraints.

Hughes and King (2003) propose to use the one-sided AIC (OSAIC) which is an asymptotically unbiased estimator of the KL divergence in the presence of one-sided information:

$$\text{OSAIC}_r = -\mathcal{L}_r(\widetilde{\theta}_r) + \sum_{h=0}^{p_r} w(p_r,h)(k_r-p_r+h). \tag{2.7}$$

Calculating the weights is a combinatorial problem, which aims to determine the probability that the vector $\widetilde{\theta}_r$ lies in any face of dimension h (Grömping, 2010, Kudô, 1963, Shapiro, 1988). This can be computationally intensive as the number of variables, k_r, increases (Grömping, 2010). However, in this study the largest model has eight free parameters (because $S \leq 4$). Therefore, the model selection procedure is still very fast (a couple of seconds).

2.2.3 Testing. To evaluate the effect of DNA copy number on expression, we test the hypothesis $H_0 : \mathbf{C}\theta = 0$ against the alternative $H_1 : \mathbf{C}\theta \neq 0, \mathbf{C}\theta \geq 0$, i.e. we test that all inequality constraints are satisfied as equalities against the possibility that at least one of them is strict. From (2.4) we observe that all parameters except the intercept θ_0 are subject to inequality constraints, and that the null hypothesis reduces the model to the intercept.

We employ the likelihood ratio statistic $\text{LR} = 2(\mathcal{L}_1 - \mathcal{L}_0)$, where $\mathcal{L}_0$ and $\mathcal{L}_1$ are the maximized log-likelihood under the null and alternative hypotheses, respectively. The test rejects the null hypothesis for large values of:

$$\min_{\mathbf{C}\theta \geq 0} (y-\mathbf{X}\theta)^{\mathrm{T}}(y-\mathbf{X}\theta) - \min_{\mathbf{C}\theta=0} (y-\mathbf{X}\theta)^{\mathrm{T}}(y-\mathbf{X}\theta). \tag{2.8}$$

This can be shown (Robertson et al., 1988) to be equivalent to rejecting for large values of

$$\overline{\chi}^2 = (\widetilde{\theta} - \widetilde{\theta}_{=})^{\mathrm{T}} \Sigma_{\mathbf{X}}^{-1} (\widetilde{\theta} - \widetilde{\theta}_{=}), \tag{2.9}$$

where $\widetilde{\theta}$ and $\widetilde{\theta}_{=}$ are the maximum likelihood estimators under the inequality and the equality constraints, respectively, and $\Sigma_{\mathbf{X}} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ is the covariance matrix

of the unconstrained least squares estimator. For known error variance σ^2 the chi-bar-squared statistic $\overline{\chi}^2$ may be employed with null distribution approximated by a weighted mixture of χ^2 distributions (Chernoff, 1954, Gouriéroux et al., 1982). As σ^2 is typically unknown, we use instead the so-called *E-bar-squared statistic* (Grömping, 2010, Robertson et al., 1988, Shapiro, 1988, Silvapulle and Sen, 2005)

$$\overline{E}^2 = \frac{(\widetilde{\theta}-\widetilde{\theta}_=)^T\Omega_X^{-1}(\widetilde{\theta}-\widetilde{\theta}_=)}{(\widetilde{\theta}-\widetilde{\theta}_=)^T\Omega_X^{-1}(\widetilde{\theta}-\widetilde{\theta}_=)+(y-X\hat{\theta})^T(y-X\hat{\theta})}. \tag{2.10}$$

Here $\Omega_X = X^TX$. The null distribution of this statistic is a weighted mixture of Beta distributions of the form

$$\sum_{h=0}^{p} w(p,h)\mathcal{B}(h/2,(n-p)/2), \tag{2.11}$$

where p is the number of parameters, and $\mathcal{B}(a,b)$ refers to a beta distribution with shape parameters a and b. The mixing weights are the same as in (2.6) (applied to the full model); unknown parameters are estimated by their MLEs.

Further details on these test statistics can be found in Robertson et al. (1988), Shapiro (1988), Silvapulle and Sen (2005).

2.2.4 Confidence bands. Confidence bands (CBs) for the (spline) function $\mathbf{x} \mapsto f_\alpha(\mathbf{x};\theta)$ in Equation (2.1) should take both the model selection procedure (see Buckland et al. (1997)) and the constraints into account.

Initially we implemented a bootstrap procedure (Grömping, 2010), accounting for model uncertainty along the lines of Burnham and Anderson (2002), who propose the construction of so-called unconditional confidence intervals where only the selected model is considered for each bootstrap sample. Unfortunately, simulated coverage probabilities were below (and sometimes far below, e.g. 0.6 instead of 0.95) the nominal level, probably due to the presence of the inequality constraints in our model (Andrews, 2000). We therefore developed an "exact" alternative based on the E-bar-squared statistic (2.10), using semidefinite programming to achieve computational efficiency. A simulation study reported in Section 2.3.2 shows that this approach yields accurate uniform CBs.

2.2.4.1 Problem formulation. We start by the construction of a joint confidence region for all parameters θ in the full model, including the intercept θ_0, by inverting the likelihood ratio test described previously. Analogously to Equation (2.10), define

$$\overline{E}^2(\theta) = \frac{(\widetilde{\theta}-\theta)^T\Omega_X^{-1}(\widetilde{\theta}-\theta)}{(\widetilde{\theta}-\theta)^T\Omega_X^{-1}(\widetilde{\theta}-\theta)+(y-X\hat{\theta})^T(y-X\hat{\theta})}.$$

Then a $(1-\alpha)$% confidence region $\mathcal{R}$ for θ is

$$\mathcal{R} = \{\theta : \overline{E}^2(\theta) \le \mathcal{Q}_{1-\alpha}, C\theta \ge 0\}, \tag{2.12}$$

where $\mathcal{Q}_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the beta mixture distribution in (2.11). Here we increment the first parameter of the Beta distributions to $(h+1)/2$, because presently we include the intercept as a parameter, whereas before it was free under the null hypothesis. Interval estimation based on inversion of a likelihood ratio statistic is known to possess good properties (Arnold and Shavelle, 1998, Brown et al., 2003, Meeker and Escobar, 1995).

Given the confidence region $\mathcal{R}$ we compute a confidence band by determining for each $\mathbf{x}$ the minimum and maximum values $f_\alpha(\mathbf{x};\theta) = \mathbf{x}^{\mathrm{T}}\theta$. This means determining:

$$\inf_{\theta\in\mathcal{R}} \mathbf{x}^{\mathrm{T}}\theta \quad \text{and} \quad \sup_{\theta\in\mathcal{R}} \mathbf{x}^{\mathrm{T}}\theta.$$

Thus a simple linear function must be minimized (or maximized) subject to linear and ellipsoidal inequality constraints. In the following section, we show that this (convex) problem can be solved efficiently by semidefinite programming.

2.2.4.2 Semidefinite programming. A semidefinite program (Vandenberghe and Boyd, 1996) is concerned with the minimization of a linear objective function under the constraint that a linear combination of symmetric matrices is positive semidefinite:

$$\underset{y\in\mathbb{R}^m}{\text{minimize}}\ b^{\mathrm{T}}y \qquad \text{subject to}\ \ \mathrm{F}(y) = \mathrm{F}_0 + \sum_{i=1}^{m} y_i \mathrm{F}_i \succeq 0. \tag{2.13}$$

The vector $b \in \mathbb{R}^m$ and the symmetric $(n\times n)$ matrices $\mathrm{F}_0,\ldots,\mathrm{F}_m$ are fixed, and the expression $\mathrm{F}(y)\succeq 0$ means that the matrix $\mathrm{F}(y)$ is positive semidefinite (that is, $z^{\mathrm{T}}\mathrm{F}(y)z \geq 0$, $\forall z \in \mathbb{R}^n$). Because a linear matrix inequality constraint $\mathrm{F}(y)\succeq 0$ is convex, the program can be solved efficiently using interior-point methods (Vandenberghe and Boyd, 1996).

We may express the optimization problem of the previous section as a semidefinite program, based on two equivalences given by Vandenberghe and Boyd (1996) and provided below.

EQUIVALENCE 1: A linear inequality constraint $\mathrm{A}x + b \geq 0$, where $\mathrm{A} = \begin{bmatrix} a_1 \cdots a_k \end{bmatrix}$ and $x \in \mathbb{R}^n$, is equivalent to the following linear matrix inequality (LMI):

$$\mathrm{F}(x) = \mathrm{F}_0 + \sum_{i=1}^{k} x_i \mathrm{F}_i \succeq 0,$$

where $\mathrm{F}_0 = \mathrm{diag}(b)$, $\mathrm{F}_i = \mathrm{diag}(a_i)$, $i = 1,\cdots,k$. $\mathrm{diag}(v)$ represents the diagonal matrix with the vector v on its diagonal.

EQUIVALENCE 2: A convex quadratic constraint $(\mathrm{A}x+b)^{\mathrm{T}}(\mathrm{A}x+b) - c^{\mathrm{T}}x - d \leq 0$, where $\mathrm{A} = [a_1 \cdots a_k]$ and $x \in \mathbb{R}^n$, is equivalent to the following LMI:

$$\mathrm{F}(x) = \mathrm{F}_0 + \sum_{i=1}^{k} x_i \mathrm{F}_i \succeq 0,$$

where

$$\mathrm{F}_0 = \begin{pmatrix} \mathrm{I} & b \\ b^{\mathrm{T}} & d \end{pmatrix}, \mathrm{F}_i = \begin{pmatrix} 0 & a_i \\ a_i^{\mathrm{T}} & c_i \end{pmatrix}, i = 1, \cdots, k.$$

Note that multiple LMIs can be expressed as a single one using block diagonal matrices (VanAntwerp, 2000).

For convenience, we replace the ellipsoidal constraint $\overline{\mathrm{E}}^2(\theta) \leq \mathcal{Q}_{1-\alpha}$ by $(\mathrm{M}\theta - \mathrm{M}\widetilde{\theta})^{\mathrm{T}}(\mathrm{M}\theta - \mathrm{M}\widetilde{\theta}) \leq \lambda$, where $\lambda = (y - \mathbf{X}\hat{\theta})^{\mathrm{T}}(y - \mathbf{X}\hat{\theta})\mathcal{Q}_{1-\alpha}/(1 - \mathcal{Q}_{1-\alpha})$ and $\Omega_{\mathbf{X}}^{-1} = \mathrm{M}^{\mathrm{T}}\mathrm{M}$. Given this, the semidefinite program is

$$\underset{\theta}{\text{minimize}}\ \mathbf{x}^{\mathrm{T}}\theta \qquad \text{subject to}\ \ \mathrm{F}(\theta) = \mathrm{F}_0 + \sum_{i=1}^{p} \theta_i \mathrm{F}_i \succeq 0, \tag{2.14}$$

where

$$\mathrm{F}_0 = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{F}_0^{(2)} \end{pmatrix},\ \mathrm{F}_i = \begin{pmatrix} \mathrm{F}_i^{(1)} & 0 \\ 0 & \mathrm{F}_i^{(2)} \end{pmatrix}, i = 1, \cdots, p,$$

with the submatrices defined as:

$$\mathrm{F}_i^{(1)} = \mathrm{diag}(c_i),\quad \mathrm{F}_0^{(2)} = \begin{pmatrix} \mathrm{I} & -\mathrm{M}\widetilde{\theta} \\ (-\mathrm{M}\widetilde{\theta})^{\mathrm{T}} & \lambda \end{pmatrix} \text{ and } \mathrm{F}_i^{(2)} = \begin{pmatrix} 0 & m_i \\ m_i^{\mathrm{T}} & 0 \end{pmatrix}.$$

Here m_i and c_i denote the ith column vector of the matrices M and $\mathbf{C}$ (the matrix of linear restrictions expressed in (2.3)), respectively.

The optimization procedure needs to be repeated twice in order to determine the lower and upper bound on $\mathbf{x}^{\mathrm{T}}\theta$. Even though this must next be repeated for every new instance $\mathbf{x}$ to obtain a confidence band, the overall procedure is fast. For instance, for 100 new instances computation on a 2.66GHz Intel quad-core took less than 12s (without parallel computing).

2.3 Simulation. We conducted simulation experiments to: 1) determine the accuracy of estimates provided by PLRS (Section 2.3.1); 2) examine the coverage probabilities of the method proposed in Section 2.2.4 (Section 2.3.2); and 3) evaluate the performance of the PLRS screening test in detecting associations of various functional forms (Section 2.3.3).

2.3.1 Point estimation. The simulation study examined the accuracy of the estimates obtained by fitting piecewise splines or a simple linear model. For simplicity, we consider a two-state model (normal and gain) and the knot was fixed to 0.5. Data were generated according to:

- model 1: $y = 1 + a_2(x - 0.5)_+^1, \quad a_2 \in \{0, 0.5, 1, 2, 5\}$
- model 2: $y = 1 + 0.5x + (a_2 - 0.5)(x - 0.5)_+^1, \quad a_2 \in \{0, 0.5, 1, 2, 5\}$

The first state (normal) has no or little effect on expression. The linear function is contained in both models, and is found for $a_2 = 0$ and $a_2 = 0.5$, respectively. We generated errors from a normal distribution $\mathcal{N}(0, \sigma^2)$ where $\sigma \in \{0.1, 0.25, 0.5, 0.75, 1\}$. This resulted in 25 cases for each of the two models (5 values of a_2 times 5 values of σ). The sample size was set to 80, and the 80 values of x were generated from a uniform distribution $\mathcal{U}(0,1)$.

We were interested in comparing the precision of the estimates of the slope a_2 when fitting a linear or a piecewise linear model (the latter with a single knot placed at 0.5; 4 parameters). For each of the 25 cases we repeated the simulation experiment 1000 times, and computed the estimator of the slope for both models. Table 2.1 reports the empirical squared bias and variance over the 1000 repetitions.

		Model 1		Model 2	
σ	a_2	linear	piecewise	linear	piecewise
0.1	0	**0.000 (0.001)**	**0.001 (0.002)**	0.059 (0.001)	0.059 (0.001)
	0.5	0.070 (0.002)	0.001 (0.008)	**0.000 (0.001)**	**0.001 (0.003)**
	1	0.278 (0.002)	0.000 (0.007)	0.059 (0.001)	0.000 (0.007)
	2	1.116 (0.002)	0.001 (0.007)	0.532 (0.001)	0.000 (0.007)
	5	6.992 (0.002)	0.001 (0.008)	4.807 (0.001)	0.000 (0.007)
0.25	0	**0.002 (0.004)**	**0.007 (0.012)**	0.060 (0.008)	0.063 (0.009)
	0.5	0.070 (0.011)	0.002 (0.039)	**0.000 (0.009)**	**0.005 (0.019)**
	1	0.282 (0.011)	0.004 (0.045)	0.058 (0.008)	0.000 (0.036)
	2	1.114 (0.011)	0.003 (0.045)	0.545 (0.008)	0.000 (0.041)
	5	6.962 (0.011)	0.003 (0.042)	4.782 (0.008)	0.000 (0.046)
0.5	0	**0.007 (0.015)**	**0.023 (0.045)**	0.063 (0.027)	0.085 (0.040)
	0.5	0.065 (0.035)	0.001 (0.110)	**0.000 (0.032)**	**0.020 (0.072)**
	1	0.272 (0.045)	0.006 (0.155)	0.059 (0.032)	0.003 (0.107)
	2	1.101 (0.041)	0.016 (0.179)	0.537 (0.036)	0.000 (0.152)
	5	6.933 (0.040)	0.014 (0.175)	4.822 (0.035)	0.000 (0.149)
0.75	0	**0.015 (0.033)**	**0.047 (0.090)**	0.075 (0.053)	0.124 (0.097)
	0.5	0.050 (0.060)	0.000 (0.193)	**0.000 (0.066)**	**0.030 (0.146)**
	1	0.270 (0.081)	0.008 (0.271)	0.055 (0.070)	0.006 (0.180)
	2	1.124 (0.094)	0.027 (0.339)	0.521 (0.075)	0.000 (0.289)
	5	6.908 (0.103)	0.022 (0.393)	4.857 (0.073)	0.004 (0.320)
1	0	**0.028 (0.061)**	**0.090 (0.166)**	0.087 (0.083)	0.154 (0.147)
	0.5	0.041 (0.090)	0.003 (0.264)	**0.001 (0.115)**	**0.055 (0.235)**
	1	0.266 (0.140)	0.006 (0.442)	0.050 (0.122)	0.014 (0.293)
	2	1.107 (0.168)	0.032 (0.600)	0.562 (0.131)	0.002 (0.448)
	5	6.947 (0.160)	0.052 (0.707)	4.806 (0.129)	0.012 (0.544)

TABLE 2.1. *Squared bias and variance (in parentheses) of the slope estimates of the linear and piecewise spline models as a function of the true slope a_2, noise σ and model. In bold: setting for which the true model is linear.*

Not surprisingly the piecewise model can capture the relationship well in all cases: the squared bias is small, and the variance never unduly large. On the other hand, the estimate of the slope given by the linear model is strongly biased for larger values of the slope a_2. As expected, the variance of the PLRS estimate is usually somewhat larger than that of the linear model estimate. However, this difference is much less prominent than for the squared bias. When the data generating process is linear, i.e. when $a_2 = 0$ in model 1 and $a_2 = 0.5$ in model 2, the difference between the estimates from the linear and PLRS models is smaller than in the other cases.

The study suggests that, when estimating or testing the effect of DNA copy number on mRNA expression, there is potentially more to loose than to gain (due to misspecification versus overspecification of the model) by applying the linear instead of the piecewise linear spline model.

2.3.2 Uniform CBs. To study the coverage probabilities of the method proposed in Section 2.2.4 we simulated data according to the model $y = 1 + (x - 0.5)_+^0 + (x - 0.5)_+^1$, with x-values drawn from a uniform distribution $\mathcal{U}(0,1)$. Gaussian errors of standard deviation $\sigma \in \{0.5, 1\}$, and three sample sizes $n \in \{20, 40, 80\}$. For a given data set we computed the confidence band on a grid of 10 equidistant values, for two different significance levels $\alpha \in \{0.05, 0.1\}$, and checked whether the 10 corresponding values of the function in the display fall simultaneously into the estimated confidence band. (For computational reasons the simulation was limited to 10 values; we believe that using the continuous range would not have altered the findings.) Table 2.2 shows the empirical coverage probabilities over 10,000 data sets for each situation.

The simulated coverage probabilities are close to their corresponding nominal values. Even though the coverage procedure is motivated by asymptotic approximations, this is true even when the sample size is small, in agreement with previous literature on likelihood-based interval estimation.

	$\sigma = 0.5$		$\sigma = 1$	
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
$n = 20$	0.953	0.898	0.968	0.922
$n = 40$	0.952	0.883	0.967	0.926
$n = 80$	0.939	0.863	0.960	0.915

TABLE 2.2. *Simulated coverage probability for different sample sizes, noise levels and significance levels.*

2.3.3 PLRS screening test. We evaluated the performance of the PLRS testing procedure in detecting associations of various functional shapes. PLRS was compared to the LM test (see Section 2.4.2), Spearman's correlation test and the test proposed

by van Wieringen and van de Wiel (2009). In Appendix A.4.4, Figures A.3 to A.12 show partial ROC curves (sensitivity versus type I error α, where $\alpha \leq 0.2$) and partial AUC. Details are provided in Appendix A.4. Here, we summarize the results.

The PLRS test yielded good performance in detecting various types of associations. It achieved the highest AUC in 68 out of the 90 simulation cases (against 23 for LM). When the true effect is linear PLRS performed reasonably well. In other cases, it always produced a high, if not the highest, AUC. In particular, PLRS presented a clear advantage over others in detecting partial effects on gene expression, i.e. when only one abnormal state (among others) affects expression. In all, results suggest that PLRS accommodates well both continuous and discrete genomic information and, unlike others, is able to detect various types of association.

2.4 Application. The proposed framework was applied to two data sets. The first data set (Carvalho et al. (2009); available at ncbi.nlm.nih.gov/geo; accession number GSE8067), consists of copy number and gene expression values for 57 samples of colorectal cancer tissue. These were generated with BAC/PAC and Human Release 2.0 oligonucleotide arrays, respectively. Normalization is as in Carvalho et al. (2009). aCGH data were segmented with the CBS algorithm of Olshen et al. (2004) and discretized with CGHcall (van de Wiel et al., 2007). Matching of mRNA and aCGH features was based on minimizing the distance between the midpoints of the genomic locations of the array elements. The final data set comprises 25,869 matched features. The second data set (Neve et al. (2006); available from Bioconductor) consists of copy number number and expression data for 50 samples (cell lines) of breast cancer, profiled with OncoBAC and Affymetrix HG-U133A arrays. Preprocessing of mRNA expression is described in Neve et al. (2006). aCGH data were segmented and called as above. The resulting data set contains 19,224 matched features. For the colorectal and breast cancer data sets, knots of the PLRS model were estimated using method I and II, respectively.

We first present some global results on model selection, and next consider testing the association between DNA and mRNA. Finally some relevant relationships are illustrated.

2.4.1 Model selection with the OSAIC procedure. Table 2.3 reports the number of genes for which our procedure (column OSAIC) selects a certain type of model, for both data sets. Clearly both the piecewise linear model and the piecewise level model are selected a large number of times. Different procedures such as AIC and BIC, $\text{BIC}_r = -2 \cdot \mathcal{L}_r(\widetilde{\theta}_r) + \log(n) \cdot k_r$, which put bigger penalities on larger models (too large given the constraints), still often prefer piecewise splines. This gives strong evidence on the inadequacy of both the simple linear and piecewise constant models for many genes. In Appendix A.1, an overlap comparison of the three procedures shows differences induced by the different penalty functions.

	Carvalho et al. (2009)			Neve et al. (2006)		
Type of model	**OSAIC**	AIC	BIC	**OSAIC**	AIC	BIC
Intercept	**14720**	18083	21700	**5081**	6968	9379
Simple linear	**4916**	3674	2043	**5262**	6689	6345
Piecewise level	**2667**	1977	992	**2761**	2477	1608
Piecewise linear	**3566**	2135	1134	**6120**	3090	1892

TABLE 2.3. *The number of times a model is selected by type of model, by three model selection procedures, for the two data sets*

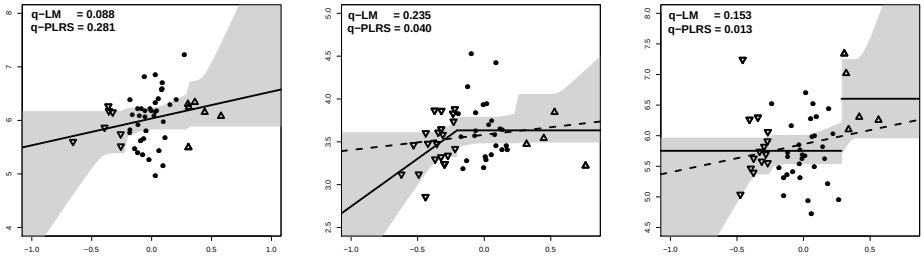
2.4.2 Testing the effect of DNA on mRNA. The hypothesis that DNA copy number has no effect on mRNA expression corresponds to model (2.1) with only the intercept parameter θ_0 nonzero. We tested this as the null model both versus the full model (2.1) (test "PLRS") and versus the linear submodel (test "LM"), with the purpose to compare these two screening models in their effectiveness to detect an association. A third possibility would be to test the null model versus the model selected by the OSAIC procedure. However, because this would naively suggest that the form of the relationship is known a priori, we did not pursue this option. For the PLRS test a minimum number of five observations (the default being three) per state was imposed.

Table 2.4 gives the number of associations with a q-value below 0.1 (based on the Benjamini and Hochberg (1995) FDR). The LM test is seen to detect slightly more associations as being significant than the PLRS test. This may be a consequence of the fact that the linear model involves fewer parameters. However, closer inspection shows that the sets of detected genes are not nested, and the PLRS test is able to detect biologically meaningful genes that are not detected by the LM test. To illustrate, three DNA-mRNA relationships are plotted in Figure 2.3. The first corresponds to an association detected as significant with the LM test, but not with the PLRS test. Reciprocally, the last two associations (genes PDE3B and CLIP1) are detected with the PLRS test but not with the LM test. The figure shows that the PLRS test is able to detect relationships for which an effect is present for only a few samples (but at least five). Identifying the last two genes may be more important than the first, as they are more interesting potential targets for studying individual effects.

H_0	H_a	Carvalho et al. (2009)	Neve et al. (2006)
intercept	linear	1726	9783
intercept	full	1554	9105

TABLE 2.4. *Number of associations with an estimated false discovery rate below 0.1 for different model comparisons.*

The first gene in Figure 2.3 also illustrates that the testing procedures may differ considerably in q-values, even though the estimated regression function found by the



(a) CDK9 (b) PDE3B (c) CLIP1

FIGURE 2.3. *Association between DNA and mRNA for different genes in the breast cancer data set (Neve et al., 2006). Segmented copy number is on x-axis while gene expression is on y-axis. Symbols indicate the different states, namely loss (▽), normal (◯) and gain (△). Grey surfaces correspond to 95% uniform CBs. The top left values correspond to q-values of test LM and PLRS, respectively. The dashed line gives the fitted LM model; the "continuous" spline is the fitted PLRS model.*

two models is the same. This is partly explained by the difference in complexity between the alternative models. However, we note that q-values for a single gene are not directly comparable, since they also depend on p-values of other genes. In Appendix A.3, we provide, for selected genes, p- and q-values for the different types of test.

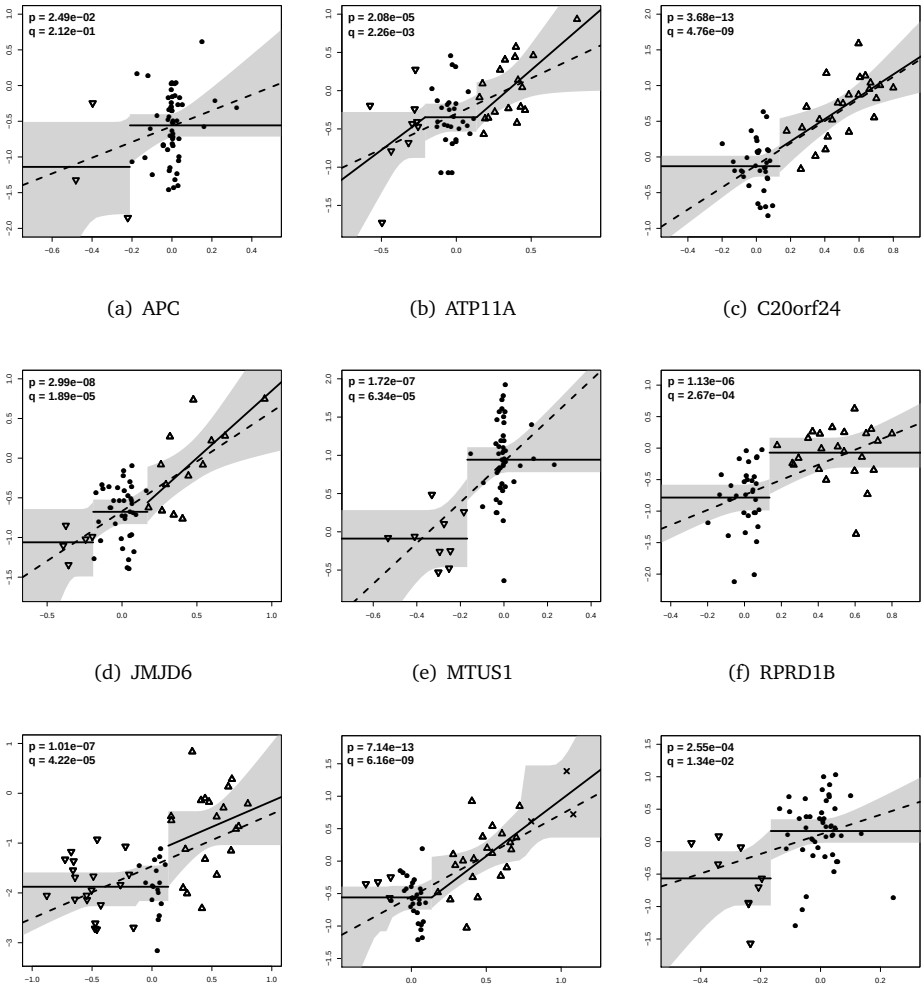
2.4.3 Results for selected genes. In this section we show the estimated relationships for selected genes. The selection is based on the Cancer Gene Census list[1] and on our observation that some associations are atypical. Also we show results for genes C20orf24, TCFL5 and TH1L, which were reported in Carvalho et al. (2009) as important for colorectal cancer progression.

Figures 2.4 and 2.5 show nine DNA-mRNA associations for each of the two data sets. Each plot displays the fit of the linear model and of the PLRS model chosen by the OSAIC criterion. Uniform 95% confidence bands (that account for model selection uncertainty) are also plotted. (Some curious shapes result from the fact that pointwise variation bursts near the boundaries and around knots.)

Both figures show a diverse set of forms of associations. Fitted models with jumps reveal that discrete copy number states can, by themselves, explain variation in expression. This is even more true when a piecewise level relationship is identified (as for gene APC and MTUS1 in Figure 2.4). More generally, piecewise linear models capture effects that differ for losses, gains and/or amplifications. Statistically speaking, this has the advantage of giving more accurate estimates of slope(s), as is clearly observed for genes ATMIN, PITPNA and PTEN in Figure 2.5. Having a better estimator, we may expect a better test. From a biological point of view, the ability to distinguish effects

[1] available at www.sanger.ac.uk/genetics/CGP/Census/

between states may help the detection of onco and tumor-suppressor genes. Moreover, genes for which these effects concern only a few samples may also be interesting to biologists for studying individual effects.

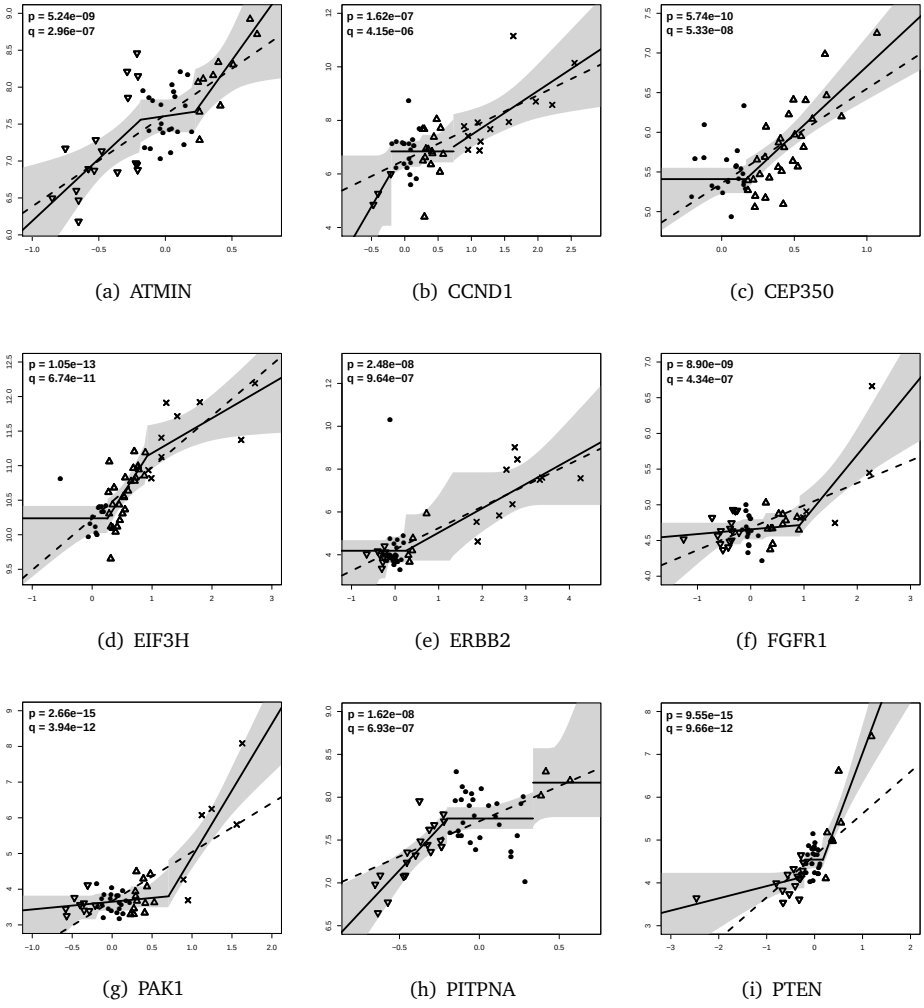


(a) APC (b) ATP11A (c) C20orf24

(d) JMJD6 (e) MTUS1 (f) RPRD1B

(g) TCFL5 (h) TH1L (i) TP53

FIGURE 2.4. *Association between DNA and mRNA for different genes in the colorectal cancer data set. Segmented copy number is on the x-axis while gene expression is on the y-axis. States are indicated by different symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform CBs. In all cases the piecewise linear model is preferred to the simple linear one (dashed line). The top left values correspond to the p- and q-values of the PLRS test.*



(a) ATMIN (b) CCND1 (c) CEP350

(d) EIF3H (e) ERBB2 (f) FGFR1

(g) PAK1 (h) PITPNA (i) PTEN

FIGURE 2.5. *Association between DNA and mRNA for different genes in the breast cancer data set. Segmented copy number is on the x-axis while gene expression is on the y-axis. States are indicated by different symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform CBs. In all cases the piecewise linear model is preferred to the simple linear one (dashed line). The top left values correspond to the p- and q-values of the PLRS test.*

The simple linear model is observed to be a tight template for modeling. As a matter of fact, it is potentially misleading when the relationship really depends on the underlying copy number state. This happens to be the case for known cancer genes (see FGFR1, PAK1 and PTEN in Figure 2.5). As a result, when testing the effect of DNA

on mRNA with the LM and PLRS tests (see Section 2.4.2), one may obtain a considerable difference between the p-values, and hence q-values (see Appendix A.3). For this reason the proposed framework may improve the detection of (highly) significant associations and their ranking.

Finally, we dwell on the notion of effect in itself. The notion of "association" is broad, and can be expressed both by an intercept and a slope. This can imply a clear difference in interpretation with respect to the linear model. Consider the simple example of gene MTUS1 in Figure 2.4, where a piecewise level model is preferable. Here intuition clearly tells us that one is more interested in assessing the difference in expression level between samples presenting loss and normal aberrations than an overall trend. Therefore, a linear model may focus on the wrong quantity of interest, whereas the PLRS procedure may yield meaningful interpretation.

We concentrated on comparing our results with those of the linear model. However, it is clear from Figures 2.4 and 2.5 that also the other alternative, the piecewise level model (which allows only horizontal lines per state), is often not adequate (see TH1L and PITPNA).

2.5 Conclusion. We proposed a statistical framework for the integrative analysis of DNA copy number and mRNA expression, which incorporates segmented and called aCGH data. By using discrete aCGH data we improved model flexibility and interpretability. The form of the relationship is allowed to vary per gene. Model interpretation is ameliorated with biologically motivated constraints on the parameters. This complicates the statistical procedures for identifying and inferring the relationship between the markers, but we provided methods for model selection, interval estimation and testing the strength of the association. We applied the methodology to two real data sets. Many (reported) genes exhibited interesting behavior.

A novelty of this work is the combined use of segmented and called aCGH data. Which of the two data types is more suitable is a matter of debate in the aCGH community, and may depend on the type of downstream analysis (van Wieringen et al., 2007). Our method provides a compromise that uses both characteristics of the data.

The form of association between copy number and expression in breast cancer is also explored in the recent paper Solvang et al. (2011) (which we received after completion of this paper). This interesting paper distinguishes (only) between linear and quadratic types of effect, and uses (only) two types of aberrations, without distinguishing gains from amplifications. The interpretation of the coefficients in our model seems to be simpler.

The proposed methodology is also applicable to the joint analysis of copy number and microRNA expression. This class of non-coding RNA was shown to play an important role in tumor development. Our method may be particularly suitable for these data, because microRNA transcripts are often expressed in part of the samples only.

Next generation sequencing data will impose new challenges, which will be taken up in future work. This type of data provides higher resolution than microarrays, while reducing biases, in particular at the lower end of the spectrum. Because expression levels are measured as counts rather than intensities, the distribution of the re-

sponse variable cannot be assumed to be Gaussian, and hence a different noise model is needed.

In short, we provide methodology for statistical inference and model selection in the framework of constrained PLRS, and showed that this is able to reveal interesting DNA-mRNA relationships for cancer genes. The method is implemented in R and available as a package from Bioconductor (as of version 2.12; http://bioconductor.org).

CHAPTER 3

PLRS: a flexible tool for the joint analysis of DNA copy number and mRNA expression data

DNA copy number and mRNA expression are commonly used data types in cancer studies. Available software for integrative analysis arbitrarily fixes the parametric form of the association between the two molecular levels and hence offers no opportunities for modeling it. We present a new tool for flexible modeling of this association. PLRS employs a wide class of interpretable models including popular ones and incorporates prior biological knowledge. It is capable to identify the gene-specific type of relationship between gene copy number and mRNA expression. Moreover, it tests the strength of the association and provides confidence intervals. We illustrate PLRS using glioblastoma data from The Cancer Genome Atlas (TCGA). PLRS is implemented as an R package and available from Bioconductor (as of version 2.12; http://bioconductor.org).

This chapter was published as:

Leday, G.G.R. and van de Wiel, M.A. (2013). PLRS: a flexible tool for the joint analysis of DNA copy number and mRNA expression data. *Bioinformatics*, 29(8): 1081-1082.

3.1 Introduction. DNA copy number aberrations are characteristics of the cancer cell. These aberrations are gains and losses of chromosomal DNA, which may alter expression levels of mRNA transcripts. The identification of genes for which an abnormal copy number affects gene expression is important in cancer studies, as these genes are likely to be relevant for tumorigenesis. Here, we present a new tool for exploratory and confirmatory analysis of such effects.

For a given gene, copy number and mRNA expression are generally believed to be *concordant*. The exact form of the association is usually not established. In fact, the shape is likely to differ between genes because of the presence of different (post-) transcriptional regulatory mechanisms. Tools that investigate the interaction between the two molecular levels assist in better understanding of regulatory mechanisms.

Numerous software packages have been proposed for joint analysis of copy number and gene expression data (Chari et al., 2008, Lê Cao et al., 2009, Lee and Kim, 2009, Louhimo and Hautaniemi, 2011, Salari et al., 2010, van Wieringen et al., 2006). However, most of these fix the association between DNA and RNA a priori, typically a linear or piecewise constant one. Hence these approaches do not permit investigation or identification of the shape of the association. Recently, the need for more subtle models has been highlighted (Leday et al., 2013, Nemes et al., 2012, Solvang et al.,

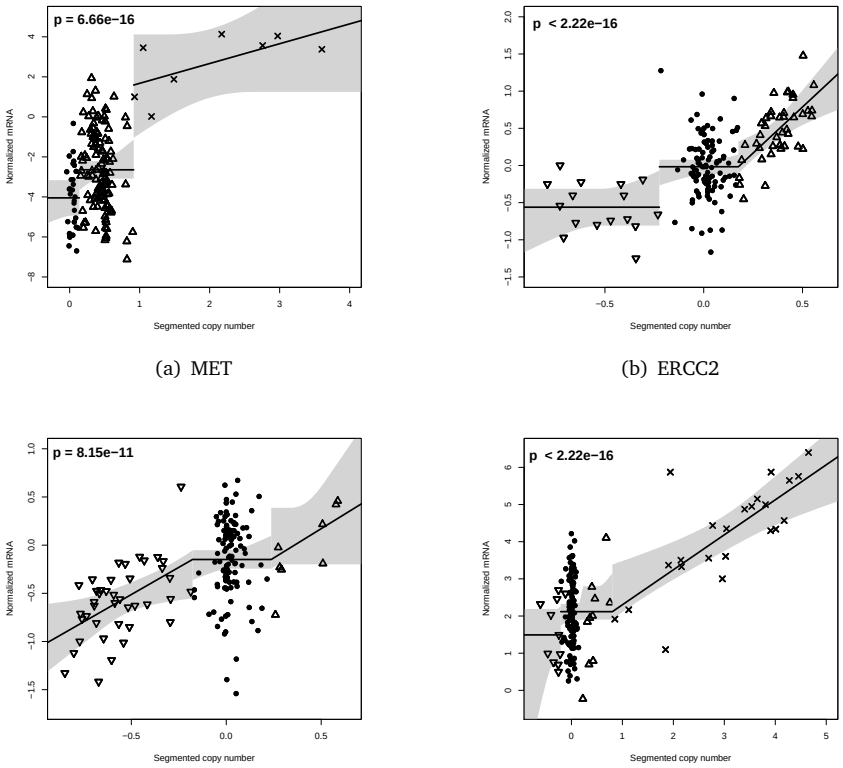
2011) to reflect the biological mechanisms between the two molecular levels. Here, we describe the R package PLRS that implements the framework recently proposed by Leday et al., 2013. PLRS uses piecewise linear regression splines, which allow multiple linear lines, and are a wide class of interpretable models including the linear and piecewise constant ones. It enforces concordance by restricting relevant model parameters. In addition, PLRS tests the strength of the overall association, identifies its functional shape, and provides confidence intervals for the estimated curve. We illustrate PLRS using a data set from 160 glioblastoma samples obtained from TCGA.

3.2 Model. PLRS models *cis*-relationships between copy number and mRNA expression by piecewise linear regression splines (Leday et al., 2013). The relevance of this class of models is multifold. Unlike other methods, PLRS combines copy number data from various steps of the preprocessing, namely the *segmented* and *called* data (van de Wiel et al., 2011). Segmented data are continuous ($\log_2$-values) and provide the (relative) amount of DNA copies (gene dosage) whereas called data represent discrete states associated with the various types of copy number aberration; the biological literature commonly distinguishes four of these: "loss" (< 2 copies of genomic DNA), "normal" ($= 2$ copies), "gain" (3-4 copies) and "amplification" (> 4 copies). Second, PLRS allows the effect of DNA on mRNA to differ across types of aberrations. This is biologically plausible: the efficacy of mechanisms that compensate for genomic aberrations may differ between losses, gains and amplifications. Third, good interpretability is ensured by the piecewise linearity of the model and a set of restrictions on the parameters. For example, copy number is concordant with gene expression and "normal" copy number cannot severely alter gene expression.

In this context, the R package PLRS implements various statistical procedures to detect which and how gene copy number abnormalities alter the gene expression level. Identification of the functional form of the association is achieved by model selection, which automatically merges copy number states when their association with mRNA expression can be captured with one regression line. Simultaneous confidence intervals on the selected curve are provided for more detailed description. Finally, a statistical test evaluates the significance of the overall association by testing the null hypothesis: copy number does not affect mRNA expression, leading to a single horizontal line.

3.3 Results. We applied PLRS to a data set of 160 glioblastoma tumor samples obtained from TCGA (http://cancergenome.nih.gov/; Verhaak et al., 2010) for which copy number (Agilent CGH Microarray 244A) and mRNA expression (Agilent 244K platform) were available. We found that for many known cancer genes, the expression level is strongly associated with DNA aberrations (cf. Supplementary Material). Figure 3.1 depicts the DNA-mRNA association for four genes, including known cancer genes MET, ERCC2 and AGAP2. Clearly, relationships are different and demonstrate that the flexibility of the PLRS model allows new insights in the association. For gene MET, we observe that the effect of amplifications extends that of gains more than proportionally. For ERCC2, the expression level of samples with loss and normal copy

number differ in average and expression increases linearly with dosage. Amplifications of gene AGAP2 have a strong effect on mRNA expression, whereas gains have none. The effect as defined by `PLRS` is broad and expressed by both an intercept and a slope for each copy number aberration state. The variety of models resulting from `PLRS` contrasts with most other methods, which impose a unique parametric form to all genes. Our method lets the data decide what is most appropriate. As a consequence `PLRS` has more power than other standard methods for detecting relatively large effects occurring in small subgroups of samples (Leday et al., 2013). Note that other non-linear techniques, e.g. based on mutual information, can be competitive but less interpretable.



(a) MET (b) ERCC2

(c) C10orf104 (d) AGAP2

FIGURE 3.1. *DNA-mRNA associations for four genes in the TCGA data set. X-axis: Gene dosage (segmented values), y-axis: mRNA gene expression. Copy number states are indicated by symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform confidence bands. The top left value corresponds to the p-value of the PLRS test.*

3.4 Conclusion. PLRS is a tool for flexible modeling of the association between DNA copy number and mRNA expression. We demonstrated its potential to reveal interesting relationships. It is particularly useful for a) a detailed understanding of the relationship between DNA copy number and mRNA expression; and b) powerful detection of copy number induced sample subgroup specific effects, thereby acknowledging heterogeneity of many cancers. The software can also be used for studying the effect of DNA copy number on microRNA expression.

CHAPTER 4

Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors

Next generation sequencing is quickly replacing microarrays as a technique to probe different molecular levels of the cell, such as DNA or RNA. The technology provides higher resolution, while reducing bias. RNA sequencing results in counts of RNA strands. This type of data imposes new statistical challenges. We present a novel, generic approach to model and analyze such data. Our approach aims at large flexibility of the likelihood (count) model and the regression model alike. Hence, a variety of count models is supported, such as the popular negative binomial model, which accounts for overdispersion. In addition, complex, non-balanced designs and random effects are accommodated. Like some other methods, our method provides shrinkage of dispersion-related parameters. However, we extend it by enabling joint shrinkage of parameters, including those for which inference is desired. We argue that this is essential for Bayesian multiplicity correction. Shrinkage is effectuated by empirically estimating priors. We discuss several parametric (mixture) and nonparametric priors and develop procedures to estimate (parameters of) those. Inference is provided by means of local and Bayesian False Discovery Rates. We illustrate our method on several simulations and two data sets, also to compare it with other methods. Model- and data-based simulations show substantial improvements in sensitivity at given specificity. The data motivate use of the zero-inflated negative binomial as a powerful alternative to the negative binomial, which results in higher detection rates for low-count data. Finally, compared to other methods, the results from small sample subsets validate better on their large sample complements, illustrating the importance of the type of shrinkage.

This chapter was published as:
Van De Wiel, M.A., Leday, G.G.R., Pardo, L.M., Rue, H., Van Der Vaart, A.W. and Van Wieringen, W.N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113-128.

4.1 Introduction. Technology to obtain digital expression data by sequencing (of parts) of the transcriptome is quickly replacing microarray technology. The promises are multi-fold including better coverage of the genome, higher resolution, less background noise and better dynamic range in particular at the low end of the spectrum. RNA sequencing technologies differ a lot in coverage and targets, but they have in common that the resulting data comprise of counts rather than (approximately) Gaussian data. Therefore, RNA sequencing data require different analysis methodology than RNA microarray data. Methodology for analyzing RNA sequencing data is rapidly

expanding. Methods differ in terms of the count model, application of shrinkage, flexibility of the designs and type of inference. Below we discuss these issues.

While there is no consensus on what type of count model fits best to RNA sequencing data, most methods focus on one specific model, even though the best count model may depend on the technology used. The negative binomial (NB; i.e. Poisson-Gamma) seems most popular (Anders and Huber, 2010, Hardcastle and Kelly, 2010, Robinson and Smyth, 2007), but other generalizations of the Poisson, usually allowing for overdispersion, are used as well (see e.g. Auer and Doerge, 2011). While we focus on generalizing the NB model to allow for zero-inflation, our framework also facilitates other types of overdispersion, e.g. Poisson-Gaussian.

The common unit of measurement in RNA sequencing data is tags: identified strands of consecutive RNA bases. Alternatively, clusters of neighboring tags are considered. These clusters may represent many different genomic features such as promotor regions, transcripts or exons. We generally refer to these as 'features'. The number of features measured is enormous, which creates an opportunity to shrink parameters. This is useful, because RNA sequencing is still expensive, and hence sample sizes are often small. Several methods for shrinking variance-related parameters are available, such as: parametric (Robinson and Smyth, 2007), empirical Bayes (Hardcastle and Kelly, 2010) and nonparametric (Anders and Huber, 2010). These methods consider shrinkage of one parameter. In many designs, it may be desirable to shrink multiple parameters. Our method provides such joint shrinkage.

Recently, Oshlack et al. (2010) noticed that "no general methods have been proposed for the analysis of more complex designs, such as paired samples or time course experiments, in the context of RNAseq data". Hence, they extended their initial approach (Robinson and Smyth, 2007) to multifactorial (McCarthy et al., 2012) and GLM (Oshlack et al., 2010) settings. While these settings provide much more flexibility, they do not allow for inclusion of random effects. Our method, presented in a GLM setting, does allow for random effects.

In terms of inference, most methods focus on generating p-values, to which standard multiple testing corrections can be applied. Bayesian methods are also available (Hardcastle and Kelly, 2010, Jiang and Wong, 2009), but without discussion of multiplicity corrections. We include estimation of local and Bayesian False Discovery Rate to account for multiplicity.

In short, we develop a framework satisfying the following criteria: 1) Allows for flexibility on the count model used; 2) Provides shrinkage of multiple parameters; 3) Allows for flexible study designs, including random effects; 4) Addresses the multiplicity problem; 5) Is reasonably fast.

Integrated nested Laplace approximations (INLA) for latent Gaussian models (Rue et al., 2009) provide the means to satisfy criteria 1, 3 and 5: it covers a large variety of Bayesian additive models, and the efficient use of numerical methods for sparse matrices and of nested Laplace approximations avoid MCMC. However, it relies on marginal models, so does not directly allow for estimating parameters that depend on all features, as needed for shrinkage and multiplicity corrections (criteria 2 and 4). We extend it using empirical Bayes-type shrinkage, which amounts to estimating (multiple) priors. Algorithms to fit the prior(s) are presented. The method allows for flexible

priors such as parametric mixture priors and nonparametric priors. Simulations illustrate that the estimation procedures perform well on a variety of designs and priors. To deal with multiplicity we shrink posteriors towards the null-domain and discuss how local and Bayesian False Discovery Rates are estimated. While we implement our methods in the context of INLA, they apply to any approach that provides marginal posteriors.

Finally, we discuss two specific data sets, and the potential of the zero-inflated negative binomial (ZI-NB) as a powerful alternative to the NB model. The two data sets illustrate two different aspects of our method: capable of handling complex designs and superior validation of small sample results by large sample ones in comparison with other methods.

4.2 Setting. We focus on the (Bayesian) Generalized Linear Model setting. Since we assume $p > n$, we denote variables (features; data rows) by $i = 1,\ldots,p$ and samples (data columns) by $j = 1,\ldots,n$. Then,

$$Y_{ij} =^d F_{\mu_{ij},\boldsymbol{\gamma}_i} \qquad \mu_{ij} = g^{-1}(\eta_{ij}) \qquad \eta_{ij} = \beta_{i0} + \sum_{k=1}^{K} \beta_{ik} x_{jk}, \tag{4.1}$$

where μ_{ij} represents the mean of distribution function F, g a link function, x_{jk} is the value of the k^{th} covariate for sample j and $\boldsymbol{\gamma}_i = (\gamma_{i1},\ldots,\gamma_{iG})$ are parameters not included in the regression on η_{ij}, possibly used for modeling overdispersion or zero-inflation. For RNA sequencing data, F often represents (a generalization of) the Poisson distribution, such as the Poisson-Gamma (Negative Binomial; NB) model or a zero-inflated version thereof (motivated in Section 4.7.2). We allow for Gaussian random effects in the regression part. In a two-group setting, inference usually focuses on one coefficient, say β_{i1}, but very general regression settings are possible.

Parameters at the lowest hierarchical level are endowed with priors. Our method allows for multiple, informative priors, which are estimated rather than assumed. To select parameters with an informative rather than vague prior, the following considerations guided us. First, for overdispersion or random effects parameters ($\phi_i = \gamma_{ig}$ and τ_i^2, respectively) we often use an informative prior to effectuate shrinkage of dispersion-related parameters, leading to more stable estimates. Second, an informative prior is applied to the main parameter of interest to accommodate multiplicity. Next, we discuss the estimation of those priors.

Denote a parameter corresponding to an informative prior by θ_i (e.g. $\theta_i = \beta_{i1}$) and denote the parametric prior of θ_i by $\pi_{\boldsymbol{\alpha}}(\theta)$ for $i = 1,\ldots,p$, where vector $\boldsymbol{\alpha}$ consists of the unknown hyper-parameters (parameters of priors). The parametric form of $\pi_{\boldsymbol{\alpha}}(\theta)$ depends on the type of parameter. E.g. $N(\mu,\sigma^2)$ gives $\boldsymbol{\alpha} = (\mu,\sigma)$. Furthermore, denote the collection of all unknown hyper-parameter vectors by A, so $\boldsymbol{\alpha} \in A$. INLA allows fitting model (4.1) for a fixed value of A, but does not facilitate estimation of the elements of A itself. We first explain how $\boldsymbol{\alpha}$ is estimated for parametric priors complying with INLA, before discussing alternative priors. The methods below are illustrated by an example in Appendix C.17.

4.3 Estimation of priors.

4.3.1 Joint estimation of hyper-parameters. We propose an empirical Bayes-type approach to estimate the elements of A. We first focus on one $\boldsymbol{\alpha} \in \mathrm{A}$ and assume $\mathrm{A}^{-} = \mathrm{A} \setminus \{\boldsymbol{\alpha}\}$ to be known. Assume a common prior $\pi_{\boldsymbol{\alpha}}(\theta)$ for all θ_i and denote the posterior density of θ_i given data $\mathrm{Y}_i = (\mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{in})$ and A by $\pi_{\mathrm{A}}(\theta|\mathrm{Y}_i)$. Hence, the posterior may also depend on hyper-parameters in A other than $\boldsymbol{\alpha}$. Assume $\mathrm{Y}_i, i = 1, \ldots, p$, to be independent samples from density $f_{\mathrm{A}}(y)$. Both $f_{\mathrm{A}}(y)$ and $\pi_{\mathrm{A}}(\theta|\mathrm{Y}_i = y)$ may depend on i through different models or covariates, but for clarity we drop the index. Then,

(4.2)
$$\pi_{\boldsymbol{\alpha}}(\theta) = \int \pi_{\mathrm{A}}(\theta|y) f_{\mathrm{A}}(y) d\mu(y) \approx \pi_{\mathrm{A}}^{\mathrm{Emp}}(\theta) = \frac{1}{p}\sum_{i=1}^{p} \pi_{\mathrm{A}}(\theta|\mathrm{Y}_i) = \frac{1}{p}\sum_{i=1}^{p} \pi_{\{\boldsymbol{\alpha}\} \cup \mathrm{A}^{-}}(\theta|\mathrm{Y}_i),$$

Hence estimation of $\boldsymbol{\alpha}$ can be implemented using software like INLA that computes *marginal* posteriors under given models for the priors and the data, by substituting the posteriors in the right side of (4.2), and finding the value of $\boldsymbol{\alpha}$ for which (4.2) holds. If $|\mathrm{A}| > 1$ and A^{-} is not known, then (4.2) becomes a system of equations with respect to the elements of A. We propose an iterative algorithm to find all $\boldsymbol{\alpha} \in \mathrm{A}$. Appendix C.1, provides the approximate equivalence of (4.2) to conventional empirical Bayes, i.e. maximization of the marginal likelihood. The crucial difference is that the method based on (4.2) depends solely on marginal posteriors, whereas direct maximization of the marginal likelihood with respect to A depends on joint posteriors.

To find $\boldsymbol{\alpha}$, we apply an 'EM-like' procedure: initialize all $\boldsymbol{\alpha} \in \mathrm{A}$, compute posteriors given the current values, re-estimate all $\boldsymbol{\alpha}$ and iterate. We first discuss re-estimation of one $\boldsymbol{\alpha}$. Let $\mathrm{A}^{(\ell)}$ be the current estimate of A. Then, the new maximum likelihood-based estimate $\boldsymbol{\alpha}^{\mathrm{ML},(\ell+1)}$ is:

(4.3)
$$\boldsymbol{\alpha}^{\mathrm{ML},(\ell+1)} = \mathrm{argmax}_{\boldsymbol{\alpha}}\ \mathcal{L}(z_{\mathrm{A}^{(\ell)}}; \boldsymbol{\alpha}),$$

where $\mathcal{L}(z_{\mathrm{A}^{(\ell)}}; \boldsymbol{\alpha}) = \sum_{s=1}^{\mathrm{S}} \log(\pi_{\boldsymbol{\alpha}}(z_{s,\mathrm{A}^{(\ell)}}))$ is the log-likelihood of the prior at $z_{\mathrm{A}^{(\ell)}}$: a large set of S independent samples from $\pi_{\mathrm{A}^{(\ell)}}^{\mathrm{Emp}}(\theta)$. Hence, ML is used atypically, because $z_{s,\mathrm{A}^{(\ell)}}$ is not an observation. Instead, it serves approximating an empirical mixture by a specific parametric form.

Re-estimation of hyper-parameters, $\boldsymbol{\alpha} \in \mathrm{A}$, is performed separately for each prior. The marginal posterior of a parameter, however, may depend on priors of others. Therefore, joint re-estimation of posteriors is required, which is accommodated by INLA. Let B be the number of informative priors and $\boldsymbol{\alpha}_b^{(\ell)}$ the b^{th} element of $\mathrm{A}^{(\ell)}$, consisting of the estimates at iteration ℓ. Then, the *iterative joint procedure* to estimate all $\boldsymbol{\alpha}_b \in \mathrm{A}$ is:

1. Initiate $\ell = 0$ and $\boldsymbol{\alpha}_b^{(0)}$, $b = 1,\ldots,\mathrm{B}$;

2. Apply INLA to estimate posteriors $\pi_{\mathrm{A}^\ell}(\theta|\mathrm{Y}_i)$ of parameters and construct an empirical estimate $\hat{\pi}_{\boldsymbol{\alpha}}(\theta)$ of $\pi_{\boldsymbol{\alpha}}(\theta)$ by averaging over posteriors of parameters;

3. Draw from $\hat{\pi}_{\boldsymbol{\alpha}}(\theta)$ and use MLE to obtain $\boldsymbol{\alpha}_b^{(\ell+1)}$, $b = 1,\ldots,\mathrm{B}$;

4. Reiterate from step 2 until convergence.

Next, we extend the algorithm above.

4.3.2 Refinement of marginal posteriors under an alternative prior. The numerical approximations in INLA allow efficient integration of the full posterior to obtain marginal posteriors. However, the above iterative procedure requires use of parametric priors that comply with INLA (or other full Bayes methodology). Often, one particular central parameter of interest exists. In small sample settings, its prior may have a considerable effect on the posterior and hence on inference. So, it may be desirable to refine its marginal posterior by using a more suitable or flexible prior. Next,we show how to refine a marginal posterior, as obtained from the iterative joint procedure, when changing one particular prior while leaving the others unchanged.

Let $\pi_{\boldsymbol{\alpha}_b^*}(\theta)$ and $\pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i)$ be the prior and posterior of θ_i, given hyper-parameters A^*, where θ_i corresponds to the b^{th} component of A^*, $\boldsymbol{\alpha}_b^*$. Moreover, the elements of A^* result from the joint iterative procedure, except for $\boldsymbol{\alpha}_b^*$ which may be chosen differently, as discussed below. Write $\mathrm{A}^*_{-b} = \mathrm{A}^* \setminus \boldsymbol{\alpha}_b^*$. Under a new prior $\pi'(\theta)$ the following provides re-estimation of the posterior:

$$\pi'_{\mathrm{A}^*_{-b}}(\theta|\mathrm{Y}_i) \propto \pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i)\frac{\pi'(\theta)}{\pi_{\boldsymbol{\alpha}_b^*}(\theta)}. \tag{4.4}$$

The proportionality constant is computed by normalization using integration. Numerically, (4.4) may be problematic when $\pi_{\boldsymbol{\alpha}_b^*}(\theta)$ is narrow. Therefore, we advise to compute posteriors $\pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i)$ under a wider prior than the one resulting from the iterative joint procedure. In our experience, a prior with sd 2 to 5 times as large works well, with very similar results in this range.

Equation (4.4) is the core of the *iterative marginal procedure*:

1. Initiate $\ell = 0$ and $\pi'(\theta) = \pi'^{\ell}(\theta) = \pi'^{0}(\theta)$;

2. Apply (4.4) to compute posterior $\pi'^{\ell}_{\mathrm{A}^*_{-b}}(\theta|\mathrm{Y}_i)$;

3. Estimate the new prior $\pi'^{\ell+1}(\theta)$;

4. Reiterate from step 2 until convergence.

We recommend initiating $\pi'(\theta)$ by $\pi_{\boldsymbol{\alpha}^*_b}(\theta)$ (and skip step 2 once, because the posteriors are known). Step 3 requires estimation of the new prior. We first cover nonparametric priors.

Nonparametric priors provide maximal flexibility and adaptivity, an advantage for the main parameter of interest, due to the consequences for inference. Although the empirical mixture of current posteriors, $\pi^{\prime\ell,\text{Emp}}_{\text{A}^*_{-b}}(\theta)$, defined analogously to (4.2), could directly be used as an estimate of $\pi'^{\ell+1}(\theta)$ in the iterative marginal procedure, imposing some degree of smoothness seems reasonable. We use straightforward Gaussian kernel density estimation on a large sample from this mixture. We offer two alternatives with increasing stability, in particular of the tails: estimation under the restrictions of unimodality and log-concavity (Lutz and Rufibach, 2011).

Parametric mixture priors that allow a point-mass may be useful to model non-differential effects. The above iterative marginal procedure can be used to estimate the mixture hyper-parameters by fitting these to a sample from the empirical mixture of current posteriors using an EM-algorithm. However, we provide a computationally more efficient method in Appendix C.2, which explicitly maximizes the marginal likelihood: the direct maximization procedure.

Combination of the iterative joint and the marginal refinement procedures provides marginal posteriors of a parameter of interest under a flexible prior while respecting dependencies on other parameters. The iterative algorithms need to be applied to a limited subset of features only, which saves considerable computing time. Appendix C.3 contains details on efficiency and convergence.

4.4 Inference, parametric priors, and multiplicity. The large number of features implies that one needs to account for multiplicity when inference is desired. We first assume a one parameter, one-sided interval null-hypothesis setting before discussing extensions to two-sided inference and multiple comparisons. The hypotheses are:

$$\text{(4.5)} \qquad \text{H}_{0i}: \beta_i \leq \Delta \ \text{ (Null)}; \qquad \text{H}_{1i}: \beta_i > \Delta \ \text{ (Alternative)},$$

with parameter of interest $\beta_i = \beta_{i1}$ and Δ set a priori. Moreover, define $\pi_{0i} = \text{P}(\text{H}_{0i}|\text{Y}_i)$ and $\pi_{1i} = \text{P}(\text{H}_{1i}|\text{Y}_i) = 1 - \pi_{0i}$. Typically, those features for which $\pi_{0i} \leq t$, for small t, are of interest. Note that β_i may also be a contrast, e.g. to detect monotonic time trends, see Appendix C.6.

4.4.1 Parametric priors. Scott and Berger (2006) extensively motivate the use of (generally informative) priors to account for multiplicity in Bayesian inference. The choice of the type of prior, nonparametric or parametric (and its form), is important. In (4.5), there is no principal reason to use $\Delta = 0$. Positive values may be useful to avoid detecting statistically 'significant', but small, non-relevant effects.

Priors and posteriors that have positive mass on $\Delta = 0$ are of interest, because these reflect a believe in true non-differential effects. We turn to parametric priors

in this setting. Natural extensions of the Gaussian prior are the Dirac-Gaussian prior (Lönnstedt and Speed, 2002) and the Gaussian-Dirac-Gaussian mixture prior (Lewin et al., 2007):

$$\pi(\beta) = p_0\delta_0 + (1-p_0)\mathcal{N}(\beta;0,\tau^2) \tag{4.6}$$

$$\pi(\beta) = p_{-1}\mathcal{N}(\beta;\mu_{-1},\tau^2_{-1}) + p_0\delta_0 + p_1\mathcal{N}(\beta;\mu_1,\tau^2_1), \tag{4.7}$$

where δ_0 is the dirac mass on 0 and $\mathcal{N}(\beta;\mu,\tau^2)$ denotes the Gaussian density with parameters (μ,τ^2), $p_0 = 1-p_{-1}-p_1$ and $\mu_{-1} < 0$ and $\mu_1 > 0$. In addition, we provide implementation of the Gamma-Dirac-reverse Gamma mixture (Lewin et al., 2007) and Dirac-central Laplace mixture priors. Priors for the precision of random effects are discussed in Appendix C.4.

4.4.2 Local fdr and BFDR. Use of informative priors accounts for multiplicity in the sense that posteriors of β_i's are typically more concentrated around zero than with flat priors. As such one may directly use the posterior probabilities π_{0i} for inference. In fact, $\text{lfdr}_i = \pi_{0i} = \text{P}(\text{H}_{0i}|\text{Y}_i)$ is a version of the local false discovery rate (lfdr, Efron et al. (2001)), based on conditioning on the data instead of on a statistic. Then, it is clear that use of an (estimated) informative prior on β_i is crucial, because $\pi_{0i} = \text{P}(\text{H}_{0i}|\text{Y}_i) = \text{P}_0/(\text{P}_0+\text{P}_1)$, with $\text{P}_0 = \int_{-\infty}^{\Delta}\text{P}(\text{Y}_i|\beta_i=\beta)\pi(\beta)d\beta$ and $\text{P}_1 = \int_{\Delta}^{\infty}\text{P}(\text{Y}_i|\beta_i=\beta)\pi(\beta)d\beta$. Hence, when sample size is small, π_{0i} may depend strongly on $\pi(\beta)$.

Alternatively, Lewin et al. (2007), Ventrucci et al. (2011) suggest use of the Bayesian False Discovery Rate (BFDR). Let $d_i(t) = \text{I}_{\{\pi_{0i}<t\}} = \text{I}_{\{\pi_{1i}\geq 1-t\}}$. Then, denoting $\text{I}_{\{\text{H}_{0i}\}}$ by H_i:

$$\text{BFDR}(t) = \text{E}\left[\frac{\sum_{i=1}^p \text{H}_i d_i(t)}{\sum_{i=1}^p d_i(t)}\middle|\sum_{i=1}^p d_i(t), \text{Y}\right] = \frac{\sum_{i=1}^p \pi_{0i} d_i(t)}{\sum_{i=1}^p d_i(t)}. \tag{4.8}$$

An estimator of $\text{BFDR}(t)$ is obtained by replacing π_{0i} and $d_i(t)$ by their estimators. In practical settings, $\text{BFDR}(t)$ is used analogously to FDR: t is tuned such that $\text{BFDR}(t)$ is below a pre-specified level. Observe from (4.8) that $\text{BFDR}(t) = \text{E}[\text{lfdr}_i|\text{lfdr}_i < t]$.

For two-sided inference we could replace β_i by $|\beta_i|$. However, directly applying $\text{lfdr}_i(t)$ and $\text{BFDR}(t)$ may lead to counterintuitive results: π_{0i} may be small due to non-negligible posterior mass of β_i on both sides of the $(-\Delta,\Delta)$ interval. Therefore, we develop alternative two-sided versions, $\text{lfdr}_i^{\text{II}}(t)$ and $\text{BFDR}^{\text{II}}(t)$. In addition, we introduce $\text{lfdr}_i^{\cup}(t)$ and $\text{BFDR}^{\cup}(t)$, which are multiple comparison counterparts of $\text{lfdr}_i(t)$ and $\text{BFDR}(t)$ (see Appendix C.5).

4.5 Modeling RNA sequencing data: zero-inflation and overdispersion. As illustrated in Section 4.7.2, accounting for zero-inflation may be useful. We use the following parametrization: the density of $\text{Y}_{ij} =^d \text{NB}(\mu_{ij},\phi_i)$ is $g(y_{ij}) = \binom{y_{ij}+n_i-1}{n_i-1}p_{ij}^{n_i}(1-p_{ij})^{y_{ij}}$, with $n_i = 1/\nu_i = \exp(-\phi_i)$. This implies $\text{E}(\text{Y}_{ij}) = \mu_{ij} = (1-p_{ij})/(\nu_i p_{ij})$ and $\text{V}(\text{Y}_{ij}) = \mu_{ij}(1+\mu_{ij}\nu_i)$. For $\phi_i \to -\infty \equiv \nu_i \to 0 \equiv n_i \to \infty$ and μ_{ij} constant, the above

density converges to a Poisson with mean μ_{ij}. For modeling zero-inflation, let h be the ZI-NB(μ_{ij}, w_{0i}, ϕ_i) density. Then,

$$h(y_{ij}) = w_{0i}\delta_0 + (1 - w_{0i})g(y_{ij}). \tag{4.9}$$

The regression involves only the second component of (4.9) by log-linking μ_{ij} to covariates $x_{j1}, \ldots, x_{jK}$. An alternative parametrization attributes all mass on 0 to the point mass and uses a conditional negative binomial in the second part. Then, the zeros have no impact on the regression parameters, whereas with (4.9) the zeros have an impact up to the extent that the negative binomial accounts for it, which implies a smoother transition from non-zeros to zeros.

Our approach allows parametric (mixture) priors on w_{0i} (see Appendix C.8 for discussion) and on $\nu_i = \exp(\phi_i)$, e.g. a mixture of a dirac mass on zero and a log-Normal distribution, which may be useful for ν_i (see Appendix C.7 and Appendix Figure C.13 for ϕ_i).

4.6 Simulation results.

4.6.1 Accuracy of estimation. We performed extensive simulations to validate our estimation procedures. Appendix C.9, provides the details on four cases. Here, we summarize the results. All cases are based on the NB model. Case 1 is a two-group comparison (sample size: 2×8) with mixture priors on both the group-related parameter β_{i1} and overdispersion parameter ν_i. Case 2 is a multiple comparison (sample size: 5×5) with a mixture prior on the pairwise differences and a Gaussian prior on $\phi_i = \log(\nu_i)$. Case 3 is a two-group comparison (sample size: 2×8) with either a t_4 prior or a shifted $\Gamma(2,1)$ prior on β_{i1} and a Gaussian prior on ϕ_i. Here, we non-parametrically estimate the prior of β_{i1}. Finally, Case 4 is a two-group comparison (sample size: 2×9) including a random effect with 6 levels. Three priors are estimated: Gaussian priors on β_{i1} and ϕ_i and a Γ-prior on log-precision of the random effect. This case is challenging, because the latter two priors both model dispersion. In all cases, the priors are very accurately estimated, both in terms of parameter values and Kolmogorov-Smirnov distance to the truth. Case 1 was also used to evaluate the accuracy of BFDR. Appendix Figure C.4 shows that it is slightly conservative w.r.t. FDR, but rather accurate for this case. More discussion on BFDR vs FDR is provided by Ventrucci et al., 2011.

4.6.2 Comparison with other methods. We compare our method, which we term 'ShrinkSeq', with: baySeq (Hardcastle and Kelly, 2010), NOISeq (Tarazona et al., 2011), DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010). We study the effect of 1) a Dirac-Gaussian mixture for overdispersion; 2) shrinkage of the parameter of interest; 3) many zeros in the data for a two-group setting; 4) many zeros in the data for a time-course setting. The first two simulations are based on the NB model, the latter two are data-based, hence unbiased with respect to any of the methods. Details are provided in Appendix C.10. Appendix Figures C.5 to C.8 show

partial ROC curves, restricted to specificity larger than 80%. At specificity equal to 95% ShrinkSeq NP, which uses a nonparametric prior for the parameter(s) of interest, has sensitivity 2 - 50%, 10 - 30%, >40% and > 15% higher than that of the others in the aforementioned scenarios. Finally, Appendix Figure C.5 shows that a parametric prior on this parameter (ShrinkSeq P) may further improve the ROC curve and Appendix Figure C.8 illustrates the ability of linear contrasts to better detect monotonic time trends (ShrinkSeq monotone).

4.7 Data analysis. Below we illustrate our methods on two data sets. The first corresponds to a fairly complex design and small sample size. We discuss the need to include random effects and illustrate the effects of accounting for zero-inflation and shrinkage of multiple parameters. Other methods provide some of these features as well, but the combination is not covered. The second is a simple two-group comparison, the large sample size of which we utilize for sample splitting to compare methods.

4.7.1 CAGE data. The data were generated by Pardo et al. (2013) to profile transcription start sites and promoter regions from aged human brain. Twenty-five libraries from RNA isolated from 5 brain regions (caudate nucleus, frontal lobe, hippocampus, putamen and temporal lobe) from 7 donors were prepared using CAGE methodology. The design is unbalanced, because some individuals lack a measurement for one or more brain regions and it includes two batches (see Table 4.1). Some normalization methods attempt to remove batch effects, but, to guard against feature-specific effects, we opt to include 'batch' in the model. To account for variation on the individual level, individuals enter the model as levels of a random factor.

More details on the CAGE methodology and preprocessing of these data, including tag clustering and filtering, are given in Appendix C.11. A set of 10.000 features (here, tag clusters that represent promotor regions) is used for illustration of our approach.

Individual / Brain region	1	2	3	4	5	6	7
1		△	△	□		△	△
2	□	□	△	□	□		
3	□	□	△	△	□		
4	□		△	□		△	△
5	□	□	△	□	□		

TABLE 4.1. *Design of the CAGE experiment. A symbol indicates that a sample from the concerning individual and brain region is present in the study, measured either in batch 1 (square) or batch 2. (triangle)*

4.7.2 Including Zero-Inflation. Anders and Huber (2010) observe that in the NB setting, the overdispersion parameter ϕ_i and the mean μ_{ij} are related. This makes univariate shrinkage of the overdispersion suboptimal. Anders and Huber (2010) solve this by using a nonparametric regression curve that locally estimates the relationship between the mean and the variance. They use the curve estimate as the final estimate of the feature's variance. We prefer to incorporate the feature's own variability and therefore provide shrinkage of the feature's dispersion towards the curve estimate (NB+ model; Appendix C.12). Here, another alternative is motivated: high overdispersion for low-count features could be caused by not accounting for 'zero-inflation' in the NB model. Figure 4.1(a) shows that for the NB model with a Gaussian prior on ϕ_i, a strong residual trend is indeed apparent: low-count features generally correspond to high overdispersion. However, when accounting for zero-inflation, this residual trend disappears, as illustrated in Figure 4.1(b).

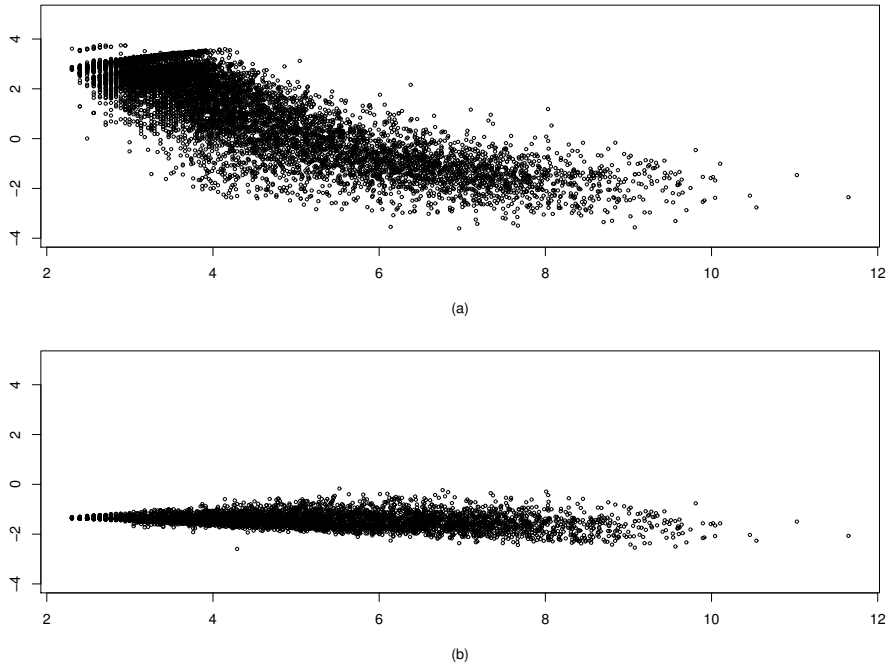


FIGURE 4.1. *Residual trend for overdispersion in NB (a) and ZI-NB (b) models after shrinking ϕ_i to a common Gaussian prior. X-axis: log-mean count for feature i, Y-axis: Posterior mean of overdispersion parameter ϕ_i*

4.7.3 Model and fitting strategies. Let G, B and I code for 'group' (brain region), 'batch' and 'individual', respectively. Regression parameters and the relevant columns of the design matrix x are coded accordingly. Moreover, let $\eta_{ij} = \log(\mu_{ij})$. Then, our model as used for joint estimation of priors and posteriors is:

$$
\begin{aligned}
Y_{ij} &=^d \text{ZI-NB}(\mu_{ij}, w_{0i}, \phi_i) \\
\nu_i &=^d q_0 \delta_0 + (1-q_0)\ell\mathcal{N}(\mu, \sigma^2) \\
\eta_{ij} &= \beta_{i0} + \sum_{\ell=1}^{5} \beta_{i\ell}^{\mathrm{G}} x_{j\ell}^{\mathrm{G}} + \beta_i^{\mathrm{B}} x_j^{\mathrm{B}} + \sum_{m=1}^{7} \beta_{im}^{\mathrm{I}} x_{jm}^{\mathrm{I}} \\
\beta_{i0} &=^d \beta_i^{\mathrm{B}} =^d \text{logit}(w_{0i}) =^d \mathcal{N}(0, 100) \\
\beta_{i\ell}^{\mathrm{G}} &=^d \mathcal{N}(0, (\tau_i^{\mathrm{G}})^2), \quad \text{for } \ell \geq 2 \\
\beta_{im}^{\mathrm{I}} &=^d \mathcal{N}(0, (\tau_i^{\mathrm{I}})^2) \\
(\tau_i^{\mathrm{I}})^{-2} &=^d \Gamma(\alpha_1, \alpha_2)
\end{aligned}
\tag{4.10}
$$

where $\beta_{i1}^{\mathrm{G}} = 0$. The set of hyper-parameters is $\mathrm{A} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3\} = \{\tau_i^{\mathrm{G}}, (q_0, \mu, \sigma), (\alpha_1, \alpha_2)\}$. The model is fitted using the iterative joint procedure, which provides estimates of all hyper-parameters and posteriors of all other parameters. In addition, posteriors of the contrasts of interest, $\beta'_{ik\ell} = \beta_{ik}^{\mathrm{G}} - \beta_{i\ell}^{\mathrm{G}}$, are computed. Finally, the marginal posteriors of these contrasts are refined using parametric mixture priors and nonparametric priors.

Here, we present the results of the analysis with a nonparametric prior as estimated by the iterative marginal procedure. The results from parametric mixture priors are discussed in Appendix C.13. We used all three options for fitting a nonparametric prior: unrestricted, unimodal and log-concave kernel densities. The latter two are superior in terms of stability of the tails. Results for these priors are very similar in terms of marginal likelihood, with the log-concave one somewhat smoother in the tails and more symmetric. Hence, we show the results for this one.

For parameters w_{0i}, β_{i0} and β_i^{B} we use vague priors instead of informative ones. Partly because of computational efficiency, but also because an informative prior is not likely to render a large advantage (these parameters are rather feature-specific; see Appendix C.8).

In short, the complete procedure is: 1) jointly shrink $\beta_{i\ell}^{\mathrm{G}}$, ϕ_i and τ_i^{I} by estimating A using the iterative joint procedure; 2) fit the model for all features using the shrunken parameters, which requires (see Appendix C.7): a) fitting the ZI-NB model and the zero-inflated Poisson (with overdispersion $\nu_i = 0$); b) combining the two posteriors for each parameter into one posterior; 3) shrink the group-related contrasts $\beta'_{ik\ell}$ to a common nonparametric prior using the iterative marginal procedure; and 4) compute posteriors and false discovery rates for the contrasts.

4.7.4 Results. Estimates of the hyper-parameters for ν_i are: $\hat{q}_0 = 0.057$, $\hat{\mu} = -1.29$ and $\hat{\sigma} = 1.07$. Appendix C.14 shows the strong stabilizing effect of shrinkage on

the stability of the estimate of ν_i, as demonstrated by others in different settings (Anders and Huber, 2010, Robinson and Smyth, 2007). Estimates of the hyper-parameters of the random effects parameter $(\tau_i^{\mathrm{I}})^{-2}$ are $\alpha_1 = 12.7, \alpha_2 = 1.01$ (shape and rate), implying $\mathrm{E}(\tau_i^{\mathrm{I}}) = 0.29$ and $\mathrm{sd}(\tau_i^{\mathrm{I}}) = 0.04$, hence a rather tight Gaussian prior on β_{im}^{I}. This aids in producing more stable results for these and other parameters. Finally, $\hat{\tau}_i^{\mathrm{G}} = 0.27$, which implies rather narrow $\mathrm{N}(0, (\hat{\tau}_i^{\mathrm{G}})^2)$ and $\mathrm{N}(0, 2(\hat{\tau}_i^{\mathrm{G}})^2)$ priors for contrasts $\beta'_{i1\ell}$ (because $\beta_{i1}^{\mathrm{G}} = 0$) and $\beta'_{ik\ell}, \ell > k > 1$, respectively. As discussed below equation (4.4), we use much wider central Gaussian priors, namely those with 10-fold variances, to initialize the iterative marginal procedure for estimating the non-parametric, log-concave prior.

The log-concave prior on $\beta'_{ik\ell}$ converges well (Appendix Figure C.15). Appendix Figure C.14 displays its final shape, which is somewhat more heavy-tailed than the corresponding Gaussian density. Its stabilizing effect on estimates of $\beta'_{ik\ell}$ compared to a vague prior is discussed in Appendix C.14.

We computed $\mathrm{BFDR}^{\mathrm{II}}(t)$ and $\mathrm{BFDR}^{\cup}(t)$ (see Appendix 4.8). Table 4.2 displays the number of detections for $\Delta = 0.1, 0.25, 0.5$ and $\mathrm{BFDR}^{\mathrm{II}}(t) \leq \mathrm{BFDR}_{\max} = 0.05, 0.10$. Observe that the comparison group 1 (caudate nucleus) vs 4 (putamen) renders relatively few detections. This is reasonable, given the underlying ontological and functional 'proximity' of striatal regions (Roth et al., 2006). Likewise, the two cortical regions, namely frontal and temporal (groups 2 and 5, respectively), are relatively similar in terms of differential expression.

Δ	$\mathrm{BFDR}_{\max}$	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5	Joint
10	0.05	47	51	8	48	52	76	38	57	74	84	294
10	0.10	123	157	47	156	156	181	120	175	195	208	787
25	0.05	9	6	2	13	11	29	5	7	14	24	74
25	0.10	29	28	3	33	32	55	19	30	51	55	193
50	0.05	3	0	0	3	1	8	0	2	2	3	15
50	0.10	4	3	0	6	4	17	0	2	5	10	34

TABLE 4.2. *Number of detections out of 10000 in the data for all 10 comparisons, using* $\Delta = 0.10, 0.25, 0.50$ *(corresponding fold changes: 1.1, 1.3, 1.6),* $\mathrm{BFDR}^{\mathrm{II}}(t)$ *for pairwise comparisons and* $\mathrm{BFDR}^{\cup}(t)$ *for multiple comparisons (Joint).* $\mathrm{BFDR}_{max} = 0.05, 0.10$ *are the thresholds used for both criteria. The comparison involving groups i and j is denoted by i-j.*

The complexity of the design complicates comparison of our ZI-NB model with other methods, mostly because these do not yet allow for inclusion of random effects. However, within our setting we do compare the ZI-NB model with the NB (using Gaussian shrinkage for ϕ_i; in the spirit of edgeR) and NB+ model (curvature shrinkage; in the spirit of DESeq). Assuming that in all three settings $\mathrm{BFDR}^{\mathrm{II}}(t)$ is correctly estimated, it is reasonable to compare the number of detected contrasts at a fixed thresh-

	ZI-NB	NB+	NB
$\mu_{ik\ell}^{\max} \leq 12.6$	41	1	0
$12.6 < \mu_{ik\ell}^{\max} \leq 37.6$	127	13	28
$37.6 < \mu_{ik\ell}^{\max} \leq 82.4$	54	41	20
$82.4 < \mu_{ik\ell}^{\max} \leq 299.4$	79	65	33
$299.4 < \mu_{ik\ell}^{\max}$	34	29	19
Sum	335	149	100

TABLE 4.3. *Number of detected differential contrasts with $BFDR^{II}(t) \leq BFDR_{max} = 0.1$ and $\Delta = 0.25$ using the ZI-NB, NB+ and NB models, where the posteriors are based on the nonparametric prior for contrasts $\beta'_{ik\ell}$ (see Appendix Figure C.14). Rows 2-6 represent very low-count, low-count, medium-count, high-count and very high-count contrasts, where $\mu_{ik\ell}^{\max} = \max(\mu_{ik}, \mu_{i\ell})$, with μ_{ih}: mean count for feature i and group h. Here, 12.6, 37.6, 82.4 and 299.4 are the 80%, 90%, 95% and 99% empirical quantiles of the vector containing all values of $\mu_{ik\ell}^{\max}$, respectively.*

old for $BFDR^{II}(t)$, which we report in Table 4.3. For this data set, both the ZI-NB and NB+ model detect more than the NB model, probably due to the improved modeling of overdispersion. The NB+ model gives fairly similar results to those of the ZI-NB model for the very high-count contrasts, detects somewhat less medium and high-count contrasts and detects much less low and very low-count contrasts. The latter is probably due to relatively high overdispersion estimates for features corresponding to those contrasts when a curvature estimate is used. Use of lfdr instead of BFDR leads to the same conclusions (see Appendix Table C.6). Appendix Table C.7 shows five illustrative contrasts, ranging from very low to very high counts. Note that the non-parametric prior, rather concentrated around 0, has a strong 'shrinkage-towards-zero' effect on the posteriors of the contrasts, which is desirable in this multiplicity context.

4.7.5 HapMap RNA-seq data. The second data set contains exon read counts for 60 samples of Caucasian (Montgomery et al., 2010) and 69 samples of Nigerian (Pickrell et al., 2010) origin. Appendix C.15 discusses preprocessing and analysis of these HapMap RNA-seq data. The large sample size and plain two-group design facilitate further comparison of ShrinkSeq with DESeq, edgeR and baySeq. NOISeq is excluded, because it did not render any detections at a 0.1 significance cut-off.

We first perform a balanced split: the last Nigerian sample is removed, and the remaining 60 vs. 68 are split into two halves of 30 vs. 34. These are used to study reproducibility. Appendix Figure C.16 shows the results: ShrinkSeq shows the highest Spearman correlation between halves, but the other methods are close. The results per half correlate well between methods (Appendix Table C.8), in particular for ShrinkSeq, DESeq and edgeR. The similar performances are likely due to the relatively modest effect of (the different types of) shrinkage for these fairly large sample sizes.

Next, we perform several unbalanced splits: the data set is split into a small part (8 vs 8) and a large complementary one (52 vs 61). Splitting is repeated four times to account for variability of the small part results. The small part mimics a realistically sized two-group discovery study, whereas the large part serves as validation. We study

to what extent detections in the small part are validated in the large part, and what proportion of detections in the large part is also detected in the small part. To this end we define the false self-validation rate (FSVR): the rate of detections in the small part that are not validated in the large part by the same method; and the self-detection rate (SDR): the rate of detections in the large part that are already detected in the small part by the same method. Detections are defined by a (B)FDR cut-off equal to 0.1.

Table 4.4 shows the results. DESeq and baySeq seem fairly conservative, leading to low FSVR, but also to much lower SDR than edgeR and ShrinkSeq. This is reflected in the number of detections in the small data sets (range), for DESeq: 60 - 1021; baySeq: 456 - 1232; edgeR: 578 - 2131 and ShrinkSeq: 1414 - 4033. For 3 out of 4 splits ShrinkSeq's FSVR is smaller than 0.1.

Comparison of the results in Table 4.4 is somewhat disturbed by the different concepts of (B)FDR used by the four methods to define a detection cut-off. Therefore, we provide a comparison which uses the same benchmark for all 4 methods and depends only on the ranking of the small set results. The common benchmark set includes all features that are detected by at least *3 out of 4 methods* in the *large* part. Then, for a set of features ranked highest by a *given* method in the *small* part the false validation rate (FVR) is defined as the proportion of features in this set not present in the benchmark set. Figure 4.2 illustrates the consistently superior performance of ShrinkSeq: its FVR is uniformly lower than that of the others when selecting the top 1 - 20% features from the small part. If one selects the 10% highest-ranked features from the small part by ShrinkSeq, the FVR is 1.3 - 1.7, 2.5 - 4.8 and 2.0 - 2.9 times smaller than the corresponding FVRs of DESeq, baySeq and edgeR, respectively. Hence, for such small sample sizes (8 vs 8), the type of shrinkage and handling of zeros clearly has an effect.

	DESeq		edgeR		ShrinkSeq		baySeq	
Split	FSVR	SDR	FSVR	SDR	FSVR	SDR	FSVR	SDR
1	0.014	0.037	0.070	0.191	0.092	0.355	0.043	0.082
2	0.081	0.245	0.150	0.340	0.134	0.518	0.076	0.183
3	0.081	0.029	0.153	0.115	0.078	0.203	0.136	0.060
4	0.033	0.015	0.109	0.094	0.057	0.187	0.057	0.067

TABLE 4.4. *False self-validation rate (FSVR) and self-detection rate (SDR) for 4 methods in 4 unbalanced splits.*

4.8 Discussion. Our method may be regarded as a hybrid full Bayes-empirical Bayes method, because we estimate some priors while leaving others vague. In essence it is empirical Bayes: priors of crucial parameters are estimated, where we allow for arbitrary parametric and nonparametric priors.

We introduced parametric priors that allow point mass on 0. However, like Lewin et al. (2007), we noted that data often prefer a smoother density close to 0. It is

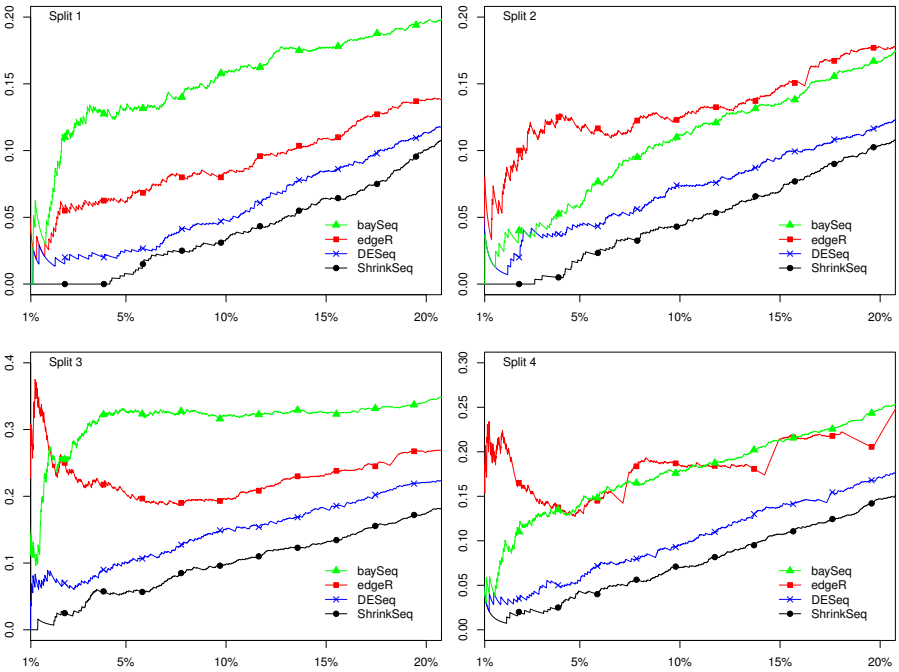


FIGURE 4.2. *Validation of small data set results by large data set ones for four random splits of the Montgomery-Pickrell data set into two parts containing 16 (8+8) and 113 (52+61) complementary samples. X-axis: percentage most differential features as selected separately by the four individual methods from the small part. Y-axis: False Validation Rate (FVR), defined as the rate of detections in the small part not validated by at least 3 out of 4 methods in the large part.*

often unclear whether 'true zero effects' exist. Still, it may be interesting to extend our method such that sparsity is enforced, e.g. by stronger penalization of non-zero effects than that effectuated by the marginal likelihood (which prefers concentrated priors) or by alternative parametric priors.

Shrinkage of priors may also be implemented by an MCMC-based full Bayesian approach, which would impose hyper-priors on the hyper-parameters. Such an approach would have the advantage of providing joint posteriors. However, nonparametric priors are not accommodated. Moreover, MCMC may perform poorly on latent Gaussian models, which include our models (Rue et al., 2009), and is computationally unpractical in (very) high-dimensional settings. Finally, implementing MCMC often requires considerable effort for a given study, while our software builds on INLA to easily handle many different designs and count models.

The ZI-NB model treats zeros differently from positive counts. Zeros have an impact on the regression (and hence on inference), but only to the extent that they fit to the NB. Zeros may be 'true zeros': the feature is really absent, but may also reflect failure to read such a feature. This reading failure is a technical artefact, hence inde-

pendent of biological conditions, which supports condition-independent zero-inflation as in our model. Also, if one includes condition-dependent zero-inflation one needs to integrate inference of those parameters with that of the regression ones, which is not trivial in terms of implementation and interpretation.

Other methods only provide shrinkage of dispersion parameters, not of regression parameters. In a frequentistic setting the latter is not required for multiplicity correction. However, the non-shrunken estimates for the most 'significant' features are biased, due to selection (Crager, 2010). Hence, for correct quantification of effect sizes for those features, shrinkage of regression parameters is important.

We foresee several extensions, both in terms of application and methodology. We aim to apply our approach to other high-dimensional count data, such as proteomics data. In addition, it is straightforward to include feature-specific covariates in the regression, such as DNA copy number variation to (partly) explain RNA counts. From the methodology viewpoint, multivariate priors and posteriors are of interest to accommodate dependencies between parameters and allow simultaneous inference on parameters. INLA includes latent models, which are useful to model spatial or other structural dependencies. We aim to apply these to account for known structures, in particular genomic position of the feature. Such dependencies have been explored by Hu et al. (2012), however not in a shrinkage context and by use of MCMC. As stated by McCarthy et al. (2012), analysis methods for differential RNAseq-based gene expression can also be applied to isoforms, once these have been identified. However, incorporating identification uncertainty and modeling the interdependency between isoforms of the same gene may lead to more efficient inference.

Our method seems promising for detecting differential features across the entire spectrum, including the lower counts. This is useful, because potential new targets may hide in the lower part of the spectrum when microarray technology failed to detect these due to higher background signal ('t Hoen et al., 2008). The novelty of our method mostly lies in the combination of several aspects relevant to the analysis of RNA sequencing data: large applicability (by allowing flexible designs and random effects), enhanced power and reproducibility (due to incorporating zero-inflation and shrinkage of dispersion parameters) and multiplicity-corrected inference (using shrinkage of inference parameters). Hence, it provides a comprehensive analysis of RNA sequence data in many settings.

CHAPTER 5

Regional differences in gene expression and promoter usage in aged human brains

To characterize the promoterome of caudate and putamen regions (striatum), frontal and temporal cortices, and hippocampi from aged human brains, we used high-throughput cap analysis of gene expression to profile the transcription start sites and to quantify the differences in gene expression across the 5 brain regions. We also analyzed the extent to which methylation influenced the observed expression profiles. We sequenced more than 71 million cap analysis of gene expression tags corresponding to 70,202 promoter regions and 16,888 genes. More than 7000 transcripts were differentially expressed, mainly because of differential alternative promoter usage. Unexpectedly, 7% of differentially expressed genes were neurodevelopmental transcription factors. Functional pathway analysis on the differentially expressed genes revealed an overrepresentation of several signaling pathways (e.g., fibroblast growth factor and wnt signaling) in hippocampus and striatum. We also found that although 73% of methylation signals mapped within genes, the influence of methylation on the expression profile was small. Our study underscores alternative promoter usage as an important mechanism for determining the regional differences in gene expression at old age.

This chapter was published as:
Pardo, L.M., Rizzu, P., Francescatto, M., Vitezic, M., Leday, G.G.R., Sanchez, J.S., Khamis, A., Takahashi, H., van de Berg, W.D.J., Medvedeva, Y.A., van de Wiel, M.A., Daub, C.O., Carninci, P., Heutink, P. (2013). Regional differences in gene expression and promoter usage in aged human brains. *Neurobiol. Aging*, 34(7):1825-1836.

5.1 Introduction. The brain is the most complex organ of the human body, and this complexity is a major landmark of human evolution (Konopka and Geschwind, 2010). The brain can be divided into different functional and anatomic regions that are established during development and maintained throughout life. The mechanisms that regulate normal brain function and differentiation are controlled by both genetic (Johnson et al., 2009b) and epigenetic factors (Miller and Sweatt, 2007), and alterations in these mechanisms can lead to neurodegenerative diseases (Abdolmaleky et al., 2005). There have been tremendous advances in our understanding of the molecular mechanisms involved in brain function, and the regional differences in these functions are beginning to be understood (Khaitovich et al., 2004, Roth et al., 2006). Less is known about the genetic mechanisms that are responsible for establishing and maintaining these differences throughout development, adulthood, and aging. In-

sights into these mechanisms are required to understand the differential susceptibility of distinct brain regions to neuronal insults (Double et al., 2010). For example, the genes for which mutations have been characterized in Alzheimer's disease (AD) (Joachim et al., 1989, Shen et al., 1997) and Parkinson's disease (PD) (Bandopadhyay et al., 2004) are often ubiquitously expressed whereas the observed pathology is restricted to specific brain regions and specific cell types (Double et al., 2010). Dissection of the molecular basis of this selective vulnerability will be pivotal to our understanding of disease pathogenesis and the development of specific therapies.

Much of our current insight into the molecular basis of brain function results from detailed studies of single genes or molecular mechanisms often initiated by the identification of genetic mutations (Hardy and Selkoe, 2002). However, unbiased approaches, where large numbers of genes are assessed simultaneously, are expected to be more powerful to dissect the genetic mechanisms controlling brain function. Large-scale analysis of gene expression in brain was pioneered by microarray experiments (Khaitovich et al., 2004). In recent years, high-throughput sequence-based technologies have been developed to analyze the mammalian transcriptome in more detail and at greater depth (Sandelin et al., 2007). These technologies have been decisive to uncover a complex picture of the mammalian transcriptome (FANTOM Consortium, 2005) and to identify new mechanisms of gene regulation and control of gene expression in brain (Kang et al., 2011, Tollervey et al., 2011). Among sequence-based technologies, tagbased approaches such as cap analysis of gene expression (CAGE) have been used to comprehensively profile the transcription start sites (TSSs) and the promoter regions (Takahashi et al., 2012). CAGE is a cap-trapping-based method that profiles 50 capped transcripts of both coding and noncoding RNA classes and has been pivotal in the discovery of alternatively regulated TSSs and novel regulatory elements (Carninci et al., 2006, Valen et al., 2009).

To understand how different promoters and control elements of genes establish and maintain region-specific expression patterns, we used CAGE in combination with massive parallel sequencing to profile TSSs of brain regions in 7 aged healthy individuals, at a genome-wide scale. We selected 5 samples of caudate nuclei, putamen, frontal and temporal cortices, and hippocampus, which are specifically vulnerable in the most prevalent neurodegenerative disorders (Double et al., 1996). First, we characterized the transcriptome of aged human brain and evaluated the extent of alternative promoter usage. Second, we quantified differences in gene expression and promoter usage across 5 brain regions. Finally, we analyzed the extent to which methylation influenced the observed expression profiles.

5.2 Methods.

5.2.1 Brain specimens. The postmortem brain tissues were obtained from the Netherlands Brain Bank (Amsterdam, The Netherlands). The donors were aged subjects (age range: 70-91 years) without clinical signs of neurodegenerative or psychiatric disorders. All brains were neuropathologically evaluated by an experienced neuropathologist and classified for neurofibrillary tangles stage 0-VI (Alafuzoff et al.,

2008), amyloid-beta plaques score 0-C, and Braak α-synuclein stage 0-VI using the staging protocols of Brain Net Europe and Braak (Alafuzoff et al., 2009a,b, Braak et al., 2006). The dissection of the caudate, putamen, hippocampus, middle frontal gyrus (F2), and middle temporal gyrus regions was performed from snap frozen human brain sections. Tissue was stored at $-80°C$ until further processing. Pathologic examination of the brain specimens showed changes consistent with the age of the individuals. The age at death, cause of death, and postmortem delay until dissection are provided in Appendix Table D.1.

5.2.2 CAGE library preparation. Total RNA was extracted and purified from tissues using the Trizol tissue kit according to the instructions provided by the manufacturer (Invitrogen). RNA quality per library was assessed using the RNA integrity number with the Agilent Total RNA Nano kit (Table 5.1). The standard CAGE protocol (Kodzius et al., 2006) was adapted for sequencing on an Illumina platform. A thorough description of the protocol to prepare CAGE libraries and to sequence CAGE tags is presented in (Takahashi et al., 2012). Briefly, complementary DNA (cDNA) was synthesized from total RNA using random primers, and this process was carried out at high temperature in the presence of trehalose and sorbitol to extend cDNA synthesis through GC-rich regions in 5′ untranslated regions (UTRs). The 5′ ends of messenger RNA within RNA-DNA hybrids were selected by the cap-trapper method (Kodzius et al., 2006) and ligated to a linker so that an EcoP15I recognition site was placed adjacent to the start of the cDNA, corresponding to the 5′ end of the original messenger RNA. This linker was used to prime second-strand cDNA synthesis. Subsequent EcoP15I digestion released the 25- to 27-base pair (bp) CAGE tags. After ligation of a second linker, CAGE tags were polymerase chain reaction amplified, purified, and sequenced on the Illumina Genome Analyzer GLXII platform (Takahashi et al., 2012). The data have been submitted to the Gene Expression Omnibus (GEO) public repository (GSE43472).

5.2.3 DNA methylation microarrays. DNA isolation and purification to detect methylation was carried out following standard protocols (online Supplementary data, Methods[1]). Genome-wide amplified input and output samples were sent to Roche NimbleGen where they were hybridized to DNA Methylation 2.1 Million Deluxe Promoter Arrays. The arrays have a mean probe spacing of 99 bp and median probe spacing of 100 bp. Each array has more than 2.1 million probes distributed in the following manner (1) promoter regions from 7250 bases upstream of each TSS to 3250 bases downstream; (2) micro RNA (miRNA) genes, starting from 15 kbp upstream of the mature gene product to its 3′ end; (3) CpG islands; and (4) ENCODE regions. Probes were chosen from the hg18 tiling database. Therefore, the probes targeted mainly annotated promoter regions and CpG islands.

[1] all online supplementary materials are available at www.neurobiologyofaging.org/article/S0197-4580(13)00023-7/addOns

Individual	Region	Batch[a]	RIN	Tag counts[b]	Unique counts[c]	Mapping rate[d]	Ribosome mapping[e]
1	Caudate	1	7.60	1,988,794	935,084	0.86	0.06
1	Frontal	1	7.00	3,453,682	1,531,751	0.87	0.05
1	Hippocampus	1	6.50	2,022,640	979,162	0.81	0.09
1	Putamen	1	7.70	3,814,753	1,627,659	0.83	0.07
1	Temporal	1	6.30	4,333,255	1,937,270	0.82	0.07
2	Hippocampus	1	6.50	1,682,943	310,481	0.84	0.07
2	Caudate	2	7.20	1,663,688	362,468	0.72	0.09
2	Frontal	2	6.90	1,745,155	801,757	0.82	0.04
2	Putamen	2	6.50	1,216,441	274,776	0.70	0.11
2	Temporal	2	6.80	936,396	259,968	0.75	0.10
3	Frontal	2	7.10	2,111,277	505,207	0.78	0.07
3	Hippocampus	2	8.80	1,785,386	413,336	0.82	0.04
3	Temporal	2	6.80	1,103,935	255,621	0.84	0.04
4	Temporal	2	5.90	1,199,974	356,84	0.71	0.13
4	Frontal	2	6.50	2,035,347	472,327	0.74	0.11
4	Hippocampus	2	6.40	1,251,589	335,644	0.73	0.11
4	Putamen	2	6.50	2,541,166	516,842	0.73	0.12
5	Caudate	1	7.90	3,096,524	1,144,105	0.88	0.06
5	Putamen	1	6.60	4,029,122	1,541,543	0.83	0.08
6	Caudate	1	7.40	3,587,220	1,296,765	0.88	0.05
6	Putamen	1	6.30	2,085,385	795,569	0.87	0.07
7	Caudate	1	6.80	4,875,578	1,625,317	0.86	0.06
7	Frontal	2	6.20	2,324,932	407,993	0.73	0.11
7	Hippocampus	2	6.20	3,158,604	597,669	0.78	0.03
7	Temporal	2	6.20	1,104,711	241,508	0.70	0.16

TABLE 5.1. *Description of the tag counts per region/sample. (a) Refers to 2 main batch effects corresponding to different period of times in which the cap analysis of gene expression libraries were prepared; (b) Refers to the total tag counts after removal of sequencing artifacts; (c) Refers to the tag counts that map to single positions in the genome unique regions; (d) Refers to proportion of tags that mapped to less than 10 positions; (e) Refers to the proportion of tags that mapped to ribosomal DNA.*

5.2.4 Bioinformatics and statistical analysis.

5.2.4.1 CAGE data. Primary quality control analysis included the removal of linker and barcode sequences as well as other sequencing artifacts to obtain raw CAGE tags of approximately 27 bps. Next, raw CAGE tags were mapped to the human genome (hg18 build) using Nexalign (http://genome.gsc.riken.jp/osc/english/dataresource/) allowing for 1 mismatch and 1 indel. The above steps were carried out with scripts and software (see Lassmann et al. (2009)) developed at the RIKEN. Following previous approaches to analyze promoter activity based on CAGE data, we grouped raw CAGE tags into CAGE clusters using a clustering pipeline from Omics Science Center bioinformatics at the RIKEN (De Hoon et al., 2010). In brief, the CAGE tags that mapped to the same position in the human genome and were on the same strand were considered CAGE Transcription Start Sites (CTSSs) (level 1 [L1]). For tags that mapped to multiple positions in the genome, a rescuing approach was applied (Faulkner et al., 2008). L1 CAGE tags were clustered into level 2 tag clusters (L2 TCs) if they over-

lapped within 20 bps and were on the same strand. L2 TCs were grouped into level 3 (L3) TCs if they overlapped within a region of 400 bps and were on the same strand. For clarity, a CTSS marks the first nucleotide that is transcribed into RNA and is considered a putative TSS, whereas a L3 TC encompasses the region that is shared between proximal TSSs (online Supplementary data, Methods) (Sandelin et al., 2007). After cluster analysis, we obtained 6,735,699 CTSSs (L1 clusters). To increase the probability of capturing genuine promoter regions, we only selected L3 TCs that were present in at least 2 CAGE libraries and with a minimum count of 5 tags per million (TPM) (De Hoon et al., 2010) in at least 1 library; for example, only CTSS present at $\geq$5 TPM in one library and $\geq$1 TPM in another were included. For all downstream analysis, we used the L3 TCs. Unless stated otherwise, TCs refer to the L3 TCs.

Next, we annotated TCs to human genes by mapping the coordinates of the TCs to all available transcripts from GENCODE version 3d. To do this, we downloaded all GENCODE transcripts from the UCSC genome browser[2] (hg18 build; University of California, Santa Cruz, CA [UCSC]) at different levels of validation. Custom Perl scripts and BEDTools (Quinlan and Hall, 2010) were used to map the coordinates of the TCs to genomic regions corresponding to specific transcriptional units (Carninci et al., 2006). TCs that did not map to a specific gene were considered intergenic. Further, we divided the TCs into mutually exclusive classes according to the gene region they mapped to. TCs that mapped to a $5'$ UTR or -300/+100 bps of a known TSS (core promoter region) were labeled as canonical. The remaining noncanonical TCs were labeled as $5'$ UTR antisense, $3'$ UTR, $3'$ UTR antisense, intronic, exonic, intronic antisense, and exonic antisense.

We classified the genes to which the TCs mapped to according to the following Biotypes: protein-coding gene (if it had an open reading frame), long noncoding RNA (lncRNA), miRNA, pseudogene, processed transcript (no open reading frame, but transcribed and not classified into any other category), and other ncRNAs using the definitions from GENCODE[3] (Harrow et al., 2006).

Differential gene expression and promoter usage derived from CAGE data across 5 brain regions. To obtain an overview of the expression (count) profile of the CAGE libraries, we first tested for differential expression across brain regions and subsequently identified patterns of differences between these regions by means of hierarchical clustering. We focused on autosomal TCs with a minimum of 9 tag counts per TC because this is the minimum number of counts needed to get reliable estimate of expression (Robinson et al., 2010). We built a model that takes into account both biological and technical variations, as we found that tag expression was subject to batch effects. The model assumes that CAGE tag counts follow a negative binomial distribution (NB), which is standard for modeling read/tag counts.

Mathematically, for tag i and sample j: $y_{ij} =^d \text{NB}(\mu_{ij}, \phi_i)$, where μ_{ij} and ϕ_i respectively denote the mean and dispersion parameters of the NB distribution, and

[2] http://genome.ucsc.edu/cgi-bin/hgTables
[3] http://www.gencodegenes.org/gencode_biotypes.html

$\log(\mu_{ij}) = \beta_{i0} + \sum_{l=1}^{5} \beta_{il}^{G} x_{jl}^{G} + \beta_{i}^{B} x_{j}^{B} + \sum_{m=1}^{7} \beta_{im}^{I} x_{jm}^{I}$, with G, B and I coding for 'group' (brain region), 'batch' and 'individual'. We decided to include a covariate 'batch' with two levels, as we observed that the expression data were clustered into two main batches (corresponding to different periods of time when the libraries were prepared). To ensure identifiability of the model, we imposed the following parameter constraints: $\sum_{l=1}^{5} \beta_{il}^{G} = 0$ and $\sum_{m=1}^{7} \beta_{im}^{I} = 0$. To our knowledge and at the time the analysis was done, no current methodology or package for differential analysis of tag expression allows extra covariates such as individual effects to be included. Therefore we built on the glm function of R package MASS and used shrunken tagwise estimates of the dispersion. We proceeded as follows. 1) We normalized raw tag counts using the quantile-adjusted method of Robinson and Smyth (2008). 2) Using these pseudo-counts, we obtained tagwise estimates of the dispersion parameter with the empirical bayes strategy of Robinson and Smyth (2007) and implemented in `R` package `edgeR` (version 1.6.15) (Robinson et al., 2010). 3) We fitted two models per tag using shrunk dispersion estimates: a full model described above, which includes three factors and a null model where the factor brain region is discarded. 4) The fit of the two models is compared via a likelihood ratio test (LRT) and the p-value obtained using the chi-squared approximation. Because the model fitting is not reliable with lowly expressed tags, we also calculated p-values via parametric bootstrapping. 5) Chi-squared based and bootstrapped p-values are corrected for multiplicity (Benjamini and Hochberg, 1995). Finally, a tag is considered differentially expressed (DE) across regions if both adjusted p-values are smaller than 0.1. This is very conservative but also more reliable as it avoids dependence on a single procedure.

To identify differentially expressed TCs (DETC) showing similar differences among (a subset of) groups, we carried out hierarchical clustering (with Euclidean distance) based on the coefficients of brain regions, which are lower than 3 in absolute value. This was carried out with the R function hclust from package stats (with default agglomeration method). We chose the partition that maximized the average silhouette index width.

Functional enrichment analysis was subsequently done on clusters (modules) of DETCs using the PANTHER version 7.0 database. All further functional pathway analyses were carried out using this database. We first looked for overrepresentation of 146 functional pathways in the dataset of non- DE TCs (Thomas et al., 2003). Next, we took the most significantly overrepresented pathways in the group of non-DE TCs as a reference to test for an overrepresentation in the group of DE TCs. The most significant functional pathways were selected after correction for multiple testing (Benjamini and Hochberg, 1995) using a p-value$\leq$0.05. The p-values of the pathways we tested in the group of DE TCs are presented in a heatmap in Appendix Figure D.9.

5.2.4.2 Methylation data. The $\log_2$ ratio of the probe intensity in the experimental sample against control DNA was determined. The $\log_2$ methylation signals were converted into methylation peaks (MPs) using the NimbleGen software (Roche) with default parameters (online Supplementary data, Methods). Further, we removed MPs that mapped to X and Y chromosomes as well as those that overlapped with

centromeres, telomeres, and segmental duplications. MPs overlapping with regions in which more than 1 segment was detected for a single sample were also removed. Next, we selected consensus MPs that were shared in a minimum of 2 samples. For this, we used the plink software version 7 (Purcell et al., 2007) and identified shared methylated "segments" with the command: plink file segment group. Next, we used BEDTools (Quinlan and Hall, 2010) to map the MPs to annotated human genes (hg18) using GENCODE version 3d at different levels of annotation. We also mapped the MPs to CpG islands downloaded from UCSC browser (Fujita et al., 2011). Details of the experimental protocol and the downstream analysis are presented in online Supplementary data.

Differential methylation analysis. We modeled the $\log_2$ ratios of the probe intensities taking both biological and technical variations into account and assuming that the ratios followed a normal distribution. Brain group (here we used the caudate as reference group), batch, and individual factors were covariates. We fitted 2 models per methylation probe: a full model, which included all 3 covariates and a null model where the factor brain group was discarded. We tested for differences in the models using a one-way analysis of variance, implemented in R version 13, and adjusted for multiple testing using the Bonferroni correction. Differentially methylated peaks (DMPs) were defined as differentially methylated probes occurring at a minimum overlap of 300 bps (R script provided by K. Lo at Roche, k.lo@roche.com).

Correlation between methylation signals and expression. First, we calculated the average methylation for every MP, adjusting for both biological and technical variations as mentioned previously. Next, we overlapped the genomic coordinates of the MPs with the genomic coordinates of the TCs (-1500/+500) using BEDTools (Quinlan and Hall, 2010) and estimated the Spearman correlation between the average mean intensity of methylation and the average expression of the overlapping TCs (geometric mean).

To test whether the expression of individual TCs were affected by methylation, we used the same statistical framework that we used to identify DETCs but included the methylation covariate as the variable of interest. Briefly, for each TC, we fitted 2 models. A full model with brain group, batch, and methylation as covariates, and a null model where methylation was removed. Because of the small number of MPs overlapping TCs, we could not fit the individual covariate. Significant differences were calculated as above.

5.3 Results. Appendix Figure D.1 shows a schema of the main steps of experimental procedure and the data analysis we carried out in this study. We prepared 25 CAGE libraries from total RNA isolated from the caudate nuclei, putamen, frontal and temporal lobes, and hippocampus from the 7 donors. In total, we sequenced 72

million CAGE tags (1-2 million per library approximately) in 5 sequencing rounds. Table 5.1 summarizes the tag count and mapping rate per library after quality control (online Supplementary data, methods). The final set of L3 TCs that were available for analysis numbered 70,202.

5.3.1 Features of brain transcriptome of aged individuals. We mapped the TCs to 16,888 human genes from the GENCODE database (Raney et al., 2011). Figure 5.1a shows that 31.2% of TCs mapped to the 5′ UTR or promoter regions of previously annotated transcripts (canonical TCs), whereas the remaining 68.9% mapped to other regions including introns, exons, and 3′ UTRs (noncanonical TCs). In addition, 13.6% of TCs did not map to any known transcript and were considered intergenic. Of these TCs, 559 (6%) mapped to lncRNAs (Jia et al., 2010) (online Supplementary Table 2). Although canonical TCs represented less than one-third of all TCs (Figure 5.1a), their expression was high and accounted for most of the overall TC expression. In contrast, the expression of most noncanonical TCs was low (Appendix Figure D.2).

Of all the expressed genes, 14,479 (87%) had canonical TCs (online Supplementary data, Data set 1). As shown in Figure 5.1b, 90% of these genes encode proteins. The remaining 10% consist of ncRNA, of which annotated pseudogenes account for 33%. We compared the list of genes that were expressed in our data set with those from RNASeq data from brain and other tissues (Ramsköld et al., 2009). We found an overlap of 77% (Appendix Figure D.3a). Genes expressed in brain according to (Ramsköld et al., 2009) that were not present in our CAGE data set included both mitochondrial (e.g., MT-ATP6, MT-ND3, and MT-CO2) and ribosomal genes. In contrast, there was a larger proportion of ncRNA in our brain CAGE data set (24% more compared with RNASeq, Appendix Figure D.3b), with a particular enrichment for pseudogenes and lncRNAs.

We looked at the expression profile of 1909 highly expressed genes with canonical TCs (90th percentile of the log geometric mean of expression distribution; online Supplementary Table 3) in more detail. This group included genes involved in brain aging (e.g., GPAFP, SPARCL1, and B2M, Starkey et al., 2012), calcium homeostasis (CALM1e3), neurodegeneration (CLU and PICALM, Mengel-From et al. (2011)), and oxidative stress (e.g., PTGD2, CA11 and SOD1, Pareek et al. (2011)). We carried out functional enrichment analysis using PANTHER version 7.0 Mi et al. (2010), Thomas et al. (2003) on the group of highly expressed genes. Although many genes could not be classified, the most significant molecular pathways identified included the ubiquitin-proteasome pathway, synaptic transmission pathway, Huntington's disease, and PD (Appendix Figure D.4). The overrepresentation of the PD pathway was mediated through genes encoding components of the ubiquitin-proteasome pathway (e.g., PSMA1 and PSMA2), heat shock proteins (e.g., HSPA2 and HSPA5), cell cycle components (e.g., SEPT2, SEPT4, and SEPT5), and synaptic genes (e.g., SNCA) among others. This shows that genes, for which mutations and/or variants that have been associated with PD, are components of cellular pathways that are highly expressed in the cortical and subcortical brain regions.

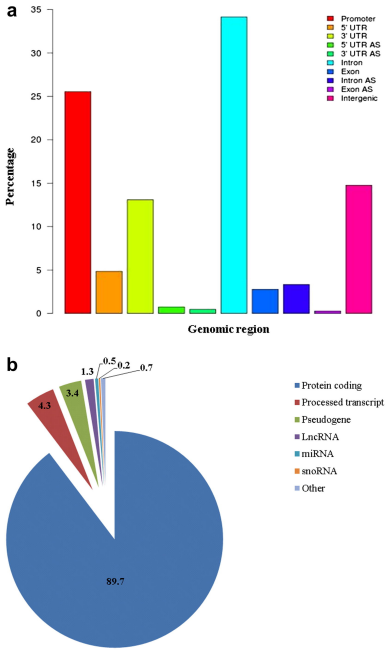


FIGURE 5.1. *Annotation of level 3 (L3) cap analysis of gene expression (CAGE) tag clusters (TCs) to human genes. (a) Barplot showing the percentage of TCs (y-axis) that map to different gene regions: promoters, 5′ untranslated regions (UTRs), 3′ UTRs, antisense, introns, exons, antisense introns, antisense exons, antisense 5′ UTR, antisense 3′ UTR, and outside genes (intergenic). Promoter regions were defined as -300/100 base pairs relative to the 5′ UTR. We defined canonical TCs those that mapped to promoters or 5′ UTRs. The TCs that mapped to other regions were classified as noncanonical. The proportion of canonical TCs represents one-third of all TCs we identified. (b) Distribution of biotype classes for genes with canonical L3 CAGE TCs. Pie chart showing the percentage of genes with at least 1 canonical TC classified by biotype class: proteincoding genes (gene with open reading frame), long noncoding RNAs (lncRNAS), pseudogenes, micro RNAs (miRNA), small nucleolar RNAs (snoRNA), and processed transcripts (no open reading frame but transcribed and not classified into any other category).*

5.3.2 Extent of alternative promoter usage in brain transcriptome. We defined alternative TCs (ATCs) as those that mapped to the same gene but were separated by a distance of >300 bp. TCs that were unique for a single gene were defined as "dominant TC" (DTC). Compared with DTCs, ATCs were mostly noncanonical and at least 34% of them mapped to introns.

In our data, 60% of genes (10,205 of 16,888 expressed genes) used ATCs (mean 5, range of 2-356, Figure 5.2). Most genes with ATC had at least 1 canonical TC. We noted that the number of ATC per genewas above 10 for 10% of the genes (Figure 5.2). Because some of the genes were quite large, we used linear regression to model the number of ATC per gene (for genes with at least 16 ATCs-5% of the genes with large number of ATCs) against gene size. We found a correlation of about 0.3 (R= 0.28, p-value <2.2e16). This shows that gene length does not account to a large extent for

the excess of ATC in genes. Outlier genes included KCNIP4, PCDH9, CADM2, BAI3, NRG3, LSAMP, NRXN1, LRRTM4, and FGF14, each with at least 100 ATCs. Functional enrichment analysis on genes with more than 16 ATCs (469 genes) showed an over-representation of glutamate receptor signaling and synaptic plasticity although most of the genes remained unclassified.

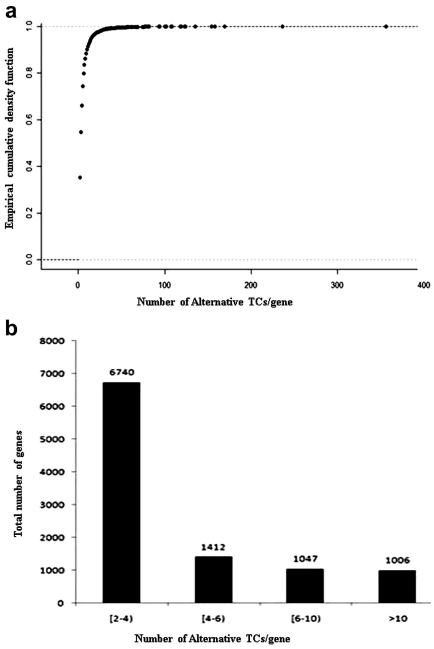


FIGURE 5.2. *Distribution of the number of alternative tag clusters (ATCs) per gene. (a) The empirical cumulative distribution (y-axis) of the number of ATCs per gene (x-axis) and (b) number of ATCs per bin category. The number of genes with 2 or more ATCs is shown at the top of every bin category.*

5.3.3 Regional differences in TC expression across brain regions. To identify signatures of gene expression across different brain regions, we sought CAGE clusters that were differentially expressed in one or more of the brain regions. We modeled the expression of the TCs using the number of counts and tested for significant differences in expression because of "regional effects" (see Section 2).We identified 7412 DETCs. Of these, 6037 were ATC of genes with a main canonical promoter. We identify neither any major differences in biotype between the differentially and nondifferentially expressed groups nor an excess of antisense TCs.

Figure 5.3 presents the results of the hierarchical clustering for the 7412 DETCs. We identified 3 main branches: one connecting the striatal regions (caudate and putamen), one connecting the cortical regions (frontal and temporal), and a third that

separated the hippocampal region from the other 2 groups. Figure 5.3 also shows that the TCs were grouped into different clusters. We separated the DETCs into nonoverlapping modules (groups of TC that were differentially expressed in one or more regions) and identified 29 modules (Table 5.2). The largest module (M13) was characterized by small differences in expression across regions, and no regionwas clearly separated from the rest (Appendix Figure D.5). The other modules were characterized by more obvious differences in the average counts in 1 or 2 brain regions relative to the others (Appendix Figures D.6-D.8). These included M18 (lower expression in striatum vs. cortical regions and hippocampus), M4 (decreased expression in the caudate nucleus vs. the rest), M27 (lower expression in hippocampus), and M2 (increased expression in the cortex).

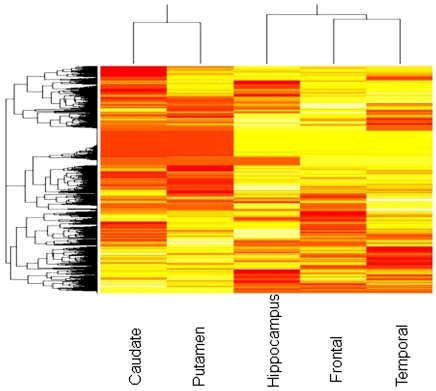


FIGURE 5.3. *Unsupervised clustering of differentially expressed tag clusters (DETC). The graph depicts the unsupervised clustering of the β coefficients of the factor "region" derived from the statistical analysis of differences in expression because of regional effects (see Section 2). The dendrogram at the top shows that basal ganglia cluster together and that frontal and temporal cortices cluster together. The dendrogram at the left of the graph was used to split the DETCs into functional modules (see Results).*

We evaluated whether specific signaling, metabolic, and disease pathways were enriched in the differentially expressed modules with at least 100 TCs.We used all the genes that were expressed in our data and that could be annotated in PANTHER version 7.0 as a reference set (online Supplementary Table 4a shows the pathways that were significantly enriched in the reference group). Compared with the reference group, few pathways were enriched in the set of differentially expressed modules (Appendix Figure D.9). The most significant pathway was the fibroblast growth factor (FGF) signaling pathway in M27 (lower expression in hippocampus) (p- value <0.0005). Several genes from the FGF pathway were differentially expressed, including FGF12, FGF14, RASA1, MAPK6, MAPK10, PPP2R2B, and PPA2. All these genes had a main canonical TC that was uniformly expressed across brain regions and an ATC showing reduced expression in hippocampus. Other significant pathways (p-value <0.005) included platelet-derived growth factor signaling in M6 (lower expression in

caudate compared with all other regions); synaptic trafficking in M2 (higher expression in cortex than in striatum and hippocampus) and M27 (lower expression in hippocampus); and glutamate receptor type I (metabotropic glutamate receptor group I [mGluRI]), Wnt signaling, and Huntington's disease pathways in M18 (lower expression in striatum compared with cortex and hippocampus). These significant pathways mediate many cell functions including proliferation, differentiation, and survival (Goldbeter and Pourquié, 2008, Moon et al., 2004, Peng et al., 2010a). A list of enriched pathways per module and genes with TC in each of the pathways is presented in online Supplementary Tables 4a and b, respectively.

Module ID	Nb TCs	Caudate	Putamen	Hippo-campus	Frontal	Temporal	% DETCs
13	3190						0.43
18	1063						0.14
6	683						0.09
27	295						0.04
2	273						0.04
20	256						0.04
23	170						0.02
4	163						0.02
19	164						0.02
10	155						0.02
16	119						0.02
8	116						0.02
11	107						0.01
3	107						0.01
9	91						0.01
7	74						0.01
22	51						0.01
1	41						0.01
25	45						0.01
17	43						0.01
5	38						0.00
12	39						0.00
26	38						0.00
15	28						0.00
28	12						0.00
21	18						0.00
29	12						0.00
24	11						0.00
14	10						0.00

TABLE 5.2. *Number of DE clusters identified for differentially expressed tag clusters (DETCs). Dark gray represents higher expression relative to other regions. Light gray represents lower expression relative to other regions. Black represents similar expression profile for all regions.*

5.3.4 Expression of neurodevelopmental transcription factors. To investigate whether differential promoter usage across brain regions can be explained by differences in the manner in which they are regulated, we searched for transcription factors (TFs) that were differentially expressed across the 5 regions. We mapped all DETCs to a manually curated list of TFs (Vaquerizas et al., 2009).We identified 519 DETCs that mapped to 320 TF genes, although only 20% mapped to the promoter or 5′ UTR region (online Supplementary Table 5). The DETF with the highest expression included those involved in neuronal postmitotic differentiation and laminar integrity in the cortex (e.g., TBR1, Bedogni et al. (2010), NR2F1, Naka et al. (2008), NEUROD1, NEUROD2, BHLHE22, and MEF2C) and neuronal plasticity (e.g., NR4A1) (Table 5.3 presents the top 20 most highly expressed TF per module). Most of the DETCs that mapped to TF were ATCs. One exception was a DTC that mapped to the promoter region of the KLF5 gene and was differentially expressed in M27. KLF5 has been shown to regulate survival and apoptosis through the regulation of MAPK kinase pathway. Other TFs that are module specific are presented in online Supplementary Table 5.

TC ID	Start	End	TF	Module	Mean (geometric)
L3_chr2_+_161981068	161980893	161981527	TBR1 (tbx family)	13	28.72
L3_chr5_+_92946017	92945793	92946068	NR2F1(COUP-tf1)	13	6.66
L3_chr12_+_50731491	50731420	50731653	NR4A1	13	6.04
L3_chr7_+_39092007	39091721	39092121	POU6F2	13	5.52
L3_chr8_+_65655474	65655301	65655790	BHLHE22	13	5.04
L3_chr19_-_41561943	41561901	41561975	ZFP14	13	4.98
L3_chr5_-_88155431	88155327	88155565	MEF2C	13	4.96
L3_chr1_-_925340	925274	925452	HES4	13	4.74
L3_chr2_-_182253487	182253446	182253729	NEUROD1	13	4.27
L3_chr2_-_242205564	242205419	242205632	THAP4	13	3.99
L3_chr17_-_35017699	35017598	35017742	NEUROD2	13	3.95
L3_chr3_+_69871321	69871264	69871369	MITF	13	3.94
L3_chr13_+_72531139	72531098	72531259	KLF5	27	3.85
L3_chr4_+_146623601	146623337	146623645	SMAD1	13	3.75
L3_chr9_-_37455447	37455266	37455461	ZBTB5	13	3.75
L3_chr13_-_73606569	73606482	73606578	KLF12	13	3.59
L3_chr1_+_13977672	13977542	13977772	PRDM2	13	3.45
L3_chr19_+_60846825	60846530	60846828	ZNF581	13	3.40
L3_chr7_+_38984037	38983927	38984054	POU6F2	13	3.38
L3_chr2_+_45022343	45022302	45022747	SIX3	3	3.37

TABLE 5.3. *List of 20 most highly differentially expressed TF*

To identify specific TFs that were coexpressed with (and possibly regulate) the DETCs, we screened proximal (-300/+100 bp) and distal (-1500/+500 bp) promoter sequences of all TCs for transcriptional factor binding sites (TFBS) using remote dependency models (see online Supplementary data, Methods). Overall, we identified 3

classes of TFBS that were significantly overrepresented in the promoter regions of the DETC, namely, BPTF (FAC1), the TBX family, and CUX1 (CDP). These TFs stand out as regulators during neurodevelopment including dendritogenesis (CUX1) (Cubelos et al., 2010), cortical formation (Tbr1-TBX) (Bedogni et al., 2010), and neurite outgrowth (BPTF) (Rhodes et al., 2003). On the other hand, we found that 15 classes of TFBS were significantly underrepresented including E2F, EGR (KROX), the Sp family (Sp1 and Sp3), Elk1, ATF6, CREB1, and MYC, and KLF5. These TFs are known to be involved in apoptosis (E2F and KLF5) and synaptic plasticity (EGR1-2, CREB, KLF5, and Elk).

We also screened every module separately. We identified significant over/underrepresentation of TFBS in 19 out of the 29 modules (online Supplementary Table 6a and b). The TBX binding site was overrepresented in most of the modules, whereas the BPTF binding site was significantly overrepresented in M13 and M27. Other TFBS were overrepresented although they did not reach statistical significance (online Supplementary Table 6a and b).

5.3.5 Methylation in the brain transcriptome of aged individuals. DNA methylation at CpG nucleotides is another crucial mechanism for the regulation of gene expression (Jones, 2012). To investigate to what extent the patterns of expression in our data correlated with methylation, we analyzed methylation signatures in all 25 samples. After quality control and filtering, we obtained 551,178 MPs distributed and 95,715 of these were shared by at least 2 samples (of the 25 samples) and were used for downstream methylation analysis. We first assessed how many annotated genes from GENCODE were methylated and found that 73% of all methylation signals mapped within genes (Figure 5.4), 43% to introns, 27% to exons, and 25% to promoter regions. We also looked at the proportion of methylation signals that occurred within CpG islands. We found that only 6% of methylated regions mapped within CpG islands. Of the promoters that mapped within CpG islands (45% of total), only 38% were methylated. Our data show that most of the methylated genomic regions occur in gene bodies and outside CpG islands (the list of MPs we used for the analysis is available on request).

5.3.5.1 Regional differences in methylation across brain regions. To identify DMP, specific for specific brain regions, we modeled the MPs using a linear model for regional effects, adjusting for both individual and possible methylation batch effects. Using this approach, we identified 13,423 DMPs, and of these 75.9% were mapped within gene bodies. Genes that were differentially methylated included NRXN1, ITPR1, MADD, CNTNAP1, SRR, GABBR1, INPP5A, HTR1D, DLGAP1, and TIAM2, which have been previously shown as methylated (Iwamoto et al., 2011) and that we found differentially methylated in frontal cortex. We also compared the list of DMPs with MPs derived from Davies et al. (2012), where differences in methylation across several brain areas (mainly cortex and cerebellum) and blood were reported. We found that at least

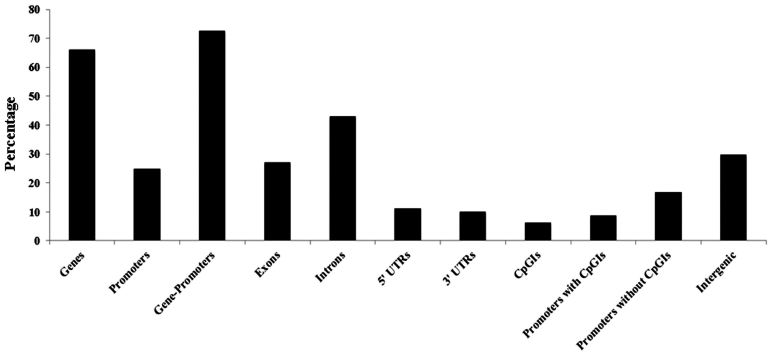


FIGURE 5.4. *Percentage of methylation peaks mapping to different gene regions, within and outside CpG islands*

39% of the DMPs overlapped with these from Davies et al. (2012). Moreover, several genes that we found differentially methylated showed also differences in methylation between cerebellum and cortex (e.g., AACS, ADCY5, EPHB4, GALNT9, and GRM4) and between brain and blood (e.g., CCDC85A, PCDH9, PDE4D, and PPP2R2B). This analysis shows that as much as 39% of methylated regions in brain (as identified by 2 different approaches) exhibit differences in their methylation profile in the brain regions we analyzed. The list of DMPs that we identified and that overlapped with MPs from Davies et al. (2012) is presented in online Supplementary data, Data set 2).

5.3.5.2 Correlation between MPs and expression. To analyze the correlation between expression and methylation in our data, we first overlapped the genomic coordinates of both data sets considering promoter regions from -1500 to +500 bp relative to the most highly expressed TSS. We found that only 9% of all TCs overlapped with at least 1 MP. Overall, there was no significant correlation between methylation and expression (Spearman correlation: r= -0.05), most likely because of the large variation in the methylation of TCs with very low counts (Appendix Figure D.10). We also analyzed the correlation between methylation and expression for protein-coding genes and noncoding genes separately (the number of ncRNA genes that overlapped with the MPs was too small to be analyzed independently) and did not observe any difference in their correlation coefficients (Spearman correlation of -0.06 and -0.07 for ncRNAs and proteincoding genes, respectively). Therefore, we tested for significant differences in expression because of "methylation effects" at individual TCs adjusting the expression for brain region and batch covariates. For this analysis, we only considered MPs that were present in at least 5 libraries. After correcting for multiple testing, we identified 312 TCs (5%) with differences in expression because of methylation effects. Of these, 34 TCs also exhibited differences in expression per region. Therefore, the differential expression because of regional effects we observed earlier was not driven by differences in methylation to a large extent. Genes with differences in ex-

pression per region because of methylation status included CDK10, NRN1, PYCARD, TIMP3, and UCP2. For these genes, promoter methylation has previously shown to regulate expression (Gloss et al., 2012, Konishi et al., 2011).

5.3.6 Correlation between MPs and expression. We compared the methylation status of differentially and non- DETFs. We did not identify any significant difference in the proportion of methylated TFs between the 2 groups (7% and 10% for differentially and non-DETFs, respectively). However, there was a significantly higher proportion of MPs mapping to the $3'$ UTR regions in the DETFs (50% vs.15%, Fisher $p = 0.0001$), whereas in the group of non-DETFs, most of the MPs mapped to the canonical promoter region (11% vs. 42%, Fisher $p = 0.0058$).

We also analyzed whether methylation could affect the expression of TCs by binding to their TFBS, presumably by modifying the spatial structure of binding sites (Choy et al., 2010). We screened the TFBS identified previously for overlaps with differentially MPs and found 304 TFBS in such locations (details of the statistical analysis are provided in online Supplementary data, Methods). Out of all these TFBS, we only selected those, which overlap with differentially MPs showing a negative correlation between expression and methylation. Because of low number of high confident TFBS predictions made by RDM, we only identified a few TFs having several binding sites in such locations, namely, E2F group, Sp1:Sp3 complex, AP2alphaA, FAC1, and NHLH1 (for details, see online Supplementary data, Methods and Table 7). This coincided with the underrepresentation of predicted TFBS for certain TFBS including E2F and Sp1:Sp3 observed earlier (online Supplementary Table 6a and b), suggesting that the corresponding TFs even being nondifferentially expressed may be involved in regulation of differential expression.

5.4 Discussion. In this study, we used CAGE in combination with massive parallel sequencing to profile transcription initiation across 5 different brain regions of aged, nondemented individuals and evaluated the extent of region specificity in alternative promoter usage and expression. At a sequencing depth of 1e2 million CAGE tags per library, we found that 40% of all GENCODE genes were expressed in brain. This estimate is probably conservative because it has been shown that deeper sequencing is needed to identify rare functional transcripts (Mercer et al., 2012). In addition, we annotated 6% of intergenic TCs to 559 lncRNAs that had previously only been predicted in silico.

We found that 77% of the genes with canonical TCs in our data set overlap with another brain transcriptome data set derived from RNA-Seq methodology (Ramsköld et al., 2009). Comparing the 2 data sets reveals that CAGE detects more ncRNA transcripts (e.g., lncRNAs and pseudogenes) whereas the proportion of proteincoding genes was higher with RNA-Seq. These differences could be the result of differences in sequencing depth or due to marked differences in the experimental design of both approaches. Indeed, although CAGE and RNA-Seq can be used to quantify the amount of gene expression and that there is a high correlation of gene expression between these

2 approaches (0.57, see Dong et al. (2012)), RNA-Seq libraries are commonly enriched for poly A+ transcripts (Mortazavi et al., 2008) of which protein-coding genes are an abundant class. In contrast, CAGE method captures capped RNA transcripts of both poly A+ and poly A- classes (Carninci, 2007). This also may explain why some genes that appeared highly expressed in brain in the RNA-Seq data set were not identified with CAGE including mitochondrial and ribosomal genes because they are uncapped and, therefore, not well covered by the CAGE approach.

Recent studies show that ncRNAs regulate gene expression in brain and play a role in the development and in the onset of neurologic diseases (Schonrock et al., 2012). Most research has focused on deciphering the functional role of lncRNAs and miRNAs, but other classes of ncRNAs may also be important. We found that more than 4.7% of the total RNA pool (and 24% of the ncRNA) consisted of annotated pseudogenes. The contribution of this ncRNA class to the transcriptome is currently unknown, with estimates ranging from 5% (Frith et al., 2006), which is consistent with our data, to 20% (Pink et al., 2011). Although the functional impact of ncRNA classes was not assessed in this study, our findings demonstrate that pseudogene expression is a pervasive feature of the transcriptome in aged brain.

We found expression patterns consistent with aging, including high expression of GPAFP (Starkey et al., 2012) and SPARCL1, which are markers of gliosis, and high expression of genes involved in protection against oxidative stress and amyloid aggregation. This group includes CLU, the gene for clusterin, an extracellular chaperone that maintains stressed proteins in a soluble state, thereby preventing their precipitation (Poon et al., 2002). Clusterin colocalizes with amyloid plaques and neurofibrillary tangles, and it has been suggested that it protects neurons from aggregate-induced damage (Yerbury et al., 2007). The ubiquitin-proteosome pathway was overrepresented in the group of highly expressed genes. This pathway has been shown to be downregulated in disorders such as AD and PD (Dennissen et al., 2012), and this decrease correlates with a failure of neurons to remove toxic protein aggregates. In this regard, it is important to stress that despite some pathologic findings consistent with aging, none of the 7 donors used for this study showed any overt AD or PD pathology (Braak tangle stages $\leq$3 and Braak α-synuclein stage 0-IV; online Supplementary Table 1). These results suggest that increased expression of genes involved in the ubiquitin-proteosome pathway and neuroprotection (e.g., CLU) may help to protect against overt protein aggregation in aged healthy individuals.

It has been recently shown that alternative promoter usage and alternative splicing can explain differences in gene expression across brain regions (Pal et al., 2011, Tollervey et al., 2011). Our data support the role of alternative promoters in causing expression differences between brain regions.We found that 81% of the DETCs were putative alternative TSS of genes with a main promoter that was similarly expressed in all the regions analyzed. This shows that the major transcripts were more often uniformly expressed whereas alternative transcripts were more likely to be region specific. Alternative promoters can alter the expression of a main transcript by competing for the cell's transcription machinery (Davuluri et al., 2008) or by antagonizing the effects of the main transcript (Tschan et al., 2003). For example, we found a DETC in M18 (online Supplementary Table 5) mapping to the promoter region of a short iso-

form of DMTF1, which has been shown to antagonize the effects of the main DMTF1 transcript in myeloid lines (Tschan et al., 2003). Whether the expression of the shorter isoform leads to the same changes observed in other cells cannot be ascertained here, but it suggests an interesting mechanism by which alternative promoter usage might lead to differences in expression.

In our data, most of the ATCs that were differentially expressed were located in noncanonical gene regions (Figure 5.1a), particularly in introns. Although there is evidence that CAGE tags can also mark post-transcriptional events (Mercer et al., 2010), we provide several lines of evidence indicating that a proportion of transcription is initiated from noncanonical gene regions. First, we only included CAGE clusters present in at least 2 biological replicas, which makes it unlikely that a tag identified twice is the result of an artifact. Second, we found that at least one-third of noncanonical TCs overlapped with other signatures of promoter activity derived from H3K4me3 histone marks (data not shown). In addition, we confirmed with RACE the existence of capped products for 4 putative alternative TSSs in the CNP, RTN4, NRG3, and AUTS2 genes (online Supplementary data, Results), which may represent novel isoforms for those genes. Indeed, we confirmed experimentally the presence of an alternative TSS in the intronic region of AUTS2, which is associated with a shorter transcript that was previously only in silico predicted. Our results indicate that at least one-third of alternative TSS map to intronic gene regions.

Several growth factor signaling pathways have been implicated in the alterations that render neuronal cell populations susceptible to neurodegeneration. Our data showed that the FGF, epidermal growth factor (EGF), insulin growth factor (IGF), and platelet-derived growth factor pathways were overrepresented in several differentially expressed modules (Appendix Figures D.5 to D.8 and online supplementary Table 4a). Common to these pathways is the mitogen-activated protein kinase (MAPK) cascade that has a broad range of effects on cellular function including survival and differentiation (Thomas and Huganir, 2004). The FGF signaling pathway was the most significantly overrepresented pathway in module M27, where a reduced expression in hippocampus was observed. The hippocampal region is a primary target of the neurodegenerative changes that lead to cognitive impartment and AD. Several mechanisms have been suggested to lead to hippocampal dysfunction, including decreased neuronal plasticity and increased calcium toxicity. The FGF pathway can influence neural plasticity through several mechanisms including MAPK/ERK activation (Thomas and Huganir, 2004), and its expression was reduced in the hippocampus relative to other regions. These findings suggest that the FGF pathway could be an important target for pharmacologic treatments to combat neurodegeneration.

The caudate and putamen regions (striatum), which are components of the cortical-subcortical circuits of motor functions, are particularly susceptible to neurodegeneration in disorders such as Huntington's disease and PD (DeLong and T., 2007). Interestingly, functional enrichment analysis based on several DETCs showed that genes encoding components of the Wnt signaling pathway and the mGluRI were significantly overrepresented in the modules where coexpression in the striatal regions was observed (M18; Table 5.2). Both Wnt signaling and mGluRI have been implicated in the development or progression of PD (Johnson et al., 2009a, L'Episcopo et al., 2011).

Moreover, mGluRI modulates neurotransmission throughout the basal ganglia, and its deregulation can contribute to neuronal damage (Johnson et al., 2009a). Our results suggest that in the absence of a clear genetic risk, pathways other than those associated with classical mutations are important determinants of the regional vulnerability in the aging brain.

We investigated whether differences in expression could be attributed to differential TF expression. We found that 7% of TFs were differentially expressed, and many of these have been shown to be involved in the neurodevelopment, which is unexpected given that neurons are postmitotic cells. The TFBS analysis also showed an overrepresentation binding sites for TFs involved in neurodevelopment. There are few explanations for this finding including a bias in the literature toward functional annotation of neurodevelopmental TFs. Another plausible explanation is that, as the brain ages, these genes may become derepressed because of, for example, damage in their promoter regions. Although we did not find decreased methylation in the group of DETFs, we found decreased methylation in the promoter region of this group and increased methylation in the 3′ UTRs. Methylation marks at both ends of transcriptional units could affect the expression of the group of DETFs (Jones, 2012).

Our analysis of methylation indicated that most of the methylation signals in our samples mapped to gene bodies and outside CpG islands. This is consistent with recent evidence that in brain most methylation signals occur within gene bodies, most likely in association with alternative promoters (Maunakea et al., 2010). However, we did not find an overall correlation between methylation and TC expression. Several factors could account for the lack of correlation. For example, batch effects were evident in the CAGE data set. In addition, only 9% of the methylated regions colocalized with a TC, which means that most of the expression in our data remained uninvestigated. The lack of overlap between the MPs and the CAGE clusters could also be because of the fact that the arrays we used to profile methylation were biased toward annotated promoters and CpG islands, whereas our CAGE clusters mapped to a large extent to noncanonical regions. Last, as a result of the small sample size, most of MPs were identified in less than 5 samples and were removed from the statistical analyses. Despite this drawback, we identified several gene-associated TCs that were affected by methylation, some of which were already documented (Iwamoto et al., 2011).

Our study is far from being comprehensive because of our small sample size and the limited number of brain regions analyzed. In addition, because of the diverse cellular composition of the brain, one might argue that the expression we observed is not exclusive to neuronal populations, although neurons and glia cells represent most of the cellular pool in human brain. A separate issue is that most of our bioinformatics analysis used public databases, which are still incomplete. For example, many protein-coding genes that we found differentially expressed could not be assigned to any functional pathway because of a lack of annotation. Therefore, inferences about functional pathways are based on a limited number of genes. Nonetheless, our data set provides an important addition to existing data on spatial expression patterns in brain.

In summary, our study shows that despite the absence of neuropathologic hallmarks of neurodegenerative disease, genetic signatures related to neurodegeneration

were already present in brain regions that are highly vulnerable to neurologic disorders.We showed that differences in transcription initiation and hence gene expression between brain regions are partly explained by alternative promoter usage and that specific signaling pathways are affected by the differential patterns in gene expression that we observed. Our data are a starting point to investigate regional susceptibility to brain aging and neurodegeneration.

CHAPTER 6

Graphical modeling using structural equation models with shrinkage priors

We study the problem of recovering an undirected graph structure using a system of (nodewise) regressions or a structural equation model (SEM). Adopting a Bayesian approach, we argue that regularization by means of Gaussian (ridge) priors coupled with *a posteriori* edge selection is a simple and attractive alternative to sparse priors. Model simplicity facilitates the use of *shrinkage priors*, which depend on all regression equations. This type of prior creates the opportunity to borrow information across equations and improves inference when the number of features is not small, which is typical in modern data sets. In this chapter, we present a computationally attractive Bayesian SEM with shrinkage priors and an empirical Bayes procedure to estimate parameters of those. We show that such priors may substantially improve graph structure recovery with SEMs. In simulations, we also demonstrate that the approach can outperform popular (sparse) methods.

6.1 Introduction. Gaussian graphical models are an important tool to describe the dependence structure among multiple variables. In recent years, these models have attracted much attention due to the emergence of complex and high-dimensional data sets, genomics data being a prime example. Among the many approaches that have been advanced, structural equation models (SEMs) have been found to be particularly valuable. In this chapter we develop a computationally attractive Bayesian SEM that uses shrinkage priors to borrow information across equations and take advantage of the dimensionality of the problem. We show that such priors may substantially improve graph structure recovery.

Gaussian graphical modeling aims to characterize the conditional dependencies between random variables as measured by partial correlations. Consider the p-dimensional gaussian random vector $\mathrm{Y} = \{\mathrm{Y}_1, \cdots, \mathrm{Y}_p\} \sim \mathcal{N}(0_p, \Omega_p^{-1})$ with positive definite precision matrix $\Omega_p = (\omega_{ij})_{i,j\in\mathcal{I}}$, $\mathcal{I} = \{1, \cdots, p\}$. The statistical distribution of Y defines a graphical model over the undirected graph $\mathcal{G} = \{\mathcal{I}, \mathcal{E}\}$ of conditional dependencies between the nodes indexed by $\mathcal{I}$. The edge set $\mathcal{E}$ is determined by Ω_p such that edge $(i,j) \in \mathcal{E}$ if $\omega_{ij} \neq 0$. This is because partial correlations may be expressed in terms of the elements of Ω_p. Precisely, $\mathrm{corr}(\mathrm{Y}_i, \mathrm{Y}_j | \mathrm{Y}_{\mathcal{I}\setminus\{i,j\}}) = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$, $\forall\ i \neq j$. This property is at the center of most modern inference techniques that focus on estimating or only recovering the support of the precision matrix. Below, we discuss penalized precision estimation and SEMs. For the sake of brevity we only cover the frequentist perspective, which is most popular in practice and used as a benchmark in our simulations.

Penalized (inverse) covariance estimation amounts to maximizing the penalized log-likelihood $\ell(\Omega_p) = \log|\Omega_p| - tr(S\Omega_p) - \lambda J(\Omega_p)$, where S is the sample covariance estimate, J a penalty function and λ an unknown scalar parameter. The penalty J may serve two purposes: (1) improve the quality of estimation and, (2) discriminate zero from non-zero entries in Ω_p. Because it simultaneously achieves (1) and (2), the ℓ_1-norm penalty (or versions thereof) is a popular choice (Friedman et al., 2008). Alternatively, a ridge-type penalty (Ledoit and Wolf, 2004, Warton, 2008) may be used in combination with a selection procedure (Schäfer and Strimmer, 2005). An obvious critical issue is the determination of the penalty parameter λ. Various solutions, usually based on resampling or cross-validation, have been proposed towards the selection of an 'optimal' value (Foygel and Drton, 2010, Gao et al., 2012, Giraud, 2008, Lian, 2011, Meinshausen and Bühlmann, 2010, Yuan and Lin, 2007).

SEMs are a powerful approach to graphical modeling. It consists in modeling the full conditional distribution of each node and results in the following system of nodewise regressions:

$$Y_i = \sum_{j \in \mathcal{I} \setminus i} Y_j \beta_{i,j} + \epsilon_i,\ i \in \mathcal{I}, \tag{6.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. By (repeatedly) partitioning the covariance Ω_p^{-1} and using properties of the mutivariate normal distribution, it can be shown that regression parameters in (6.1) are functions of the elements of Ω_p. Precisely, $\beta_{i,j} = -\omega_{ij}/\omega_{ii}$. This indicates that the problem of identifying zero entries in Ω_p can be recast into a variable selection problem in p Gaussian regression models. This approach to graphical modeling has been popularized by Meinshausen and Bühlmann (2006) who introduced an ℓ_1-penalty to each regression problem. Other penalties are also used (Krämer et al., 2009). A drawback of model (6.1) is that it misses the symmetry $\omega_{ij} = \omega_{ji}$ in Ω_p, so estimation may lack efficiency. Peng et al. (2009) showed how to work directly on partial correlations to overcome this. Alternatively, Meinshausen and Bühlmann (2006) proposed a *post-symmetrization* step with an 'AND' rule: edge $(i,j) \in \mathcal{E}$ if $\beta_{i,j} \neq 0$ and $\beta_{j,i} \neq 0$. Despite the symmetry problem, graph structure recovery based on (6.1) performs well and is widely used in practice. SEM (6.1) appears to also be a good modeling framework. For example, regularization may be node-specific and hence flexible; additional covariates are, in principle, easily accounted for; and extensions to non-Gaussian data are possible (Yang et al., 2012). It seems more difficult to achieve these goals in penalized precision estimation.

In this chapter, we adopt a Bayesian approach to Gaussian graphical modeling using SEMs. We propose to use (6.1) in combination with Gaussian priors to effectuate regularization. Then, variable selection is carried out using simple thresholding rules on posteriors. A novelty of our approach is the use of shrinkage priors that allow borrowing of information over equations. These priors have been successfully used in statistical genomics, e.g. in differential gene expression analysis (Van de Wiel et al., 2013). Here, we show that shrinkage priors can also substantially improve graph structure recovery with SEMs. To estimate parameters of priors, we employ the empirical Bayes procedure of Van de Wiel et al. (2013). The proposed Bayesian SEM is computationally attractive: the estimation procedure is coherent and complete, so does not

rely on resampling or cross-validation to estimate the regularization parameter(s). In simulations, we show that the method can outperform popular (sparse) methods.

6.2 Methods. In this section we present a Bayesian SEM for graphical modeling along with an empirical Bayes procedure to estimate parameters of priors. Finally, a selection procedure for inferring the edge set $\mathcal{E}$ is introduced.

6.2.1 Model. Suppose the n by p matrix $\mathcal{Y} = [\mathbf{y}_1 \dots \mathbf{y}_p]$ are n independent realizations from Y. Then, we adopt the following Bayesian SEM for graphical modeling:

$$
\begin{aligned}
\mathbf{y}_i &= \sum_{j \in \mathcal{I} \setminus i} \mathbf{y}_j \beta_{i,j} + \boldsymbol{\epsilon}_i, \quad i \in \mathcal{I} \\
\beta_{i,j} &\sim \mathcal{N}\left(0, \tau_i^2\right) \\
\boldsymbol{\epsilon}_i &\sim \mathcal{N}\left(0, \sigma_i^2 \mathrm{I}_n\right) \\
\tau_i^{-2} &\sim \mathcal{G}\left(a_1, b_1\right) \\
\sigma_i^{-2} &\sim \mathcal{G}\left(a_2, b_2\right)
\end{aligned}
\tag{6.2}
$$

Here $\mathcal{G}(a, b)$ denotes a gamma distribution with shape and rate parameters a and b, respectively.

In model (6.2), parameters at the lowest hierarchical level are endowed with Gaussian priors. Specifically, a ridge-type prior is imposed on the (high-dimensional) vector of regression parameters in each equation. This prior encourages coefficients to be shrunken towards zero, placing an equal weight on each of them. Conjugate gamma priors are additionnally placed on precisions. These are shrinkage priors that are shared by all regression equations. Their roles are distinct in that they allow the borrowing of information, on one hand, for precisions of regression parameters and, on the other hand, for the error precisions. It seems natural here to employ shrinkage priors for both types of precisions as they both have a prominent role in the regularization. The estimation of such priors is discussed in next section.

Gaussian regularization in model (6.2) confers the method a certain computational advantage over complex sparse priors. This is because approximations to posteriors are readily available (Ormerod and Wand, 2010, Rajagopalan and Broemeling, 1983, Rue et al., 2009), whereas sparse priors often require MCMC. Moreover, Gaussian priors allow an SVD decomposition of the high-dimensional components (West, 2003). This is computationally advantageous when $p > n$ and results in greater numerical stability. The resulting orthogonality of the components may also better accommodate multi-parameter shrinkage than the original setting. The SVD results can then be back-transformed to the original space (at least in our setting with approximately Gaussian posteriors; see Section 6.2.3).

To recover the graph structure, variable selection is required. The Gaussian ridge prior we use shrinks the vector of parameter towards the null vector, however, it does

not explicitly project it on to a lower dimensional space. Hence, the regularization has no intrinsic variable selection property and a separate procedure is needed. In Section 6.2.4 we discuss the use of simple thresholding rules on posteriors.

6.2.2 Estimation of hyperparameters. To estimate the set of hyperparameter vectors $\theta = \{(a_1, b_1), (a_2, b_2)\}$, we adopt the empirical Bayesian approach of Van de Wiel et al. (2013). This consists in estimating each element θ_k of θ by the value for which the approximation

$$\pi_{\theta_k}(\alpha_i) \approx \frac{1}{p} \sum_{i=1}^{p} \pi_{\theta_k}(\alpha_i | \mathbf{y}_i) \tag{6.3}$$

is most accurate. Here α_i is the parameter to which the prior (that depends on parameter vector θ_k, $k \in \{1, ..., \text{card}(\theta)\}$) applies, e.g. $\alpha_i = \tau_i^{-2}$ or $\alpha_i = \sigma_i^{-2}$. Van de Wiel et al. (2013) showed that (6.3) is an approximate solution to the likelihood equations that ensure maximization of the marginal likelihood (conventional empirical Bayes). Conveniently, equation (6.3) only requires marginal posteriors. In next section we discuss their approximation.

The problem of estimating θ is solved iteratively by an EM-type algorithm, which is sketched as follows:

1. Initiate $m = 0$ and $\theta_k^{(0)}$, $\forall k$
2. Approximate posteriors $\pi_{\theta_k}(\alpha_i | \mathbf{y}_i)$
3. Generate independent samples from $\pi_{\theta_k}(\alpha_i | \mathbf{y}_i)$
4. Obtain $\theta_k^{(m+1)}$ by best approximation of parametric prior $\pi_{\theta_k}(\alpha_i)$ to the empirical mixture of posteriors
5. Repeat steps 2 to 4 until convergence

In step 4, the approximation by a parametric prior (in our case a gamma distribution) may be achieved by maximum likelihood or the method of moments. Note, however, that these are being used in an unconventional manner since samples obtained in step 3 are not observations. Convergence can, e.g., be monitored using the parametric priors $\pi_{\theta_k}(\alpha_i)$ by means of a Kolmogorov-Smirnov-type metric (Van de Wiel et al., 2013) or the (log) marginal likelihoods. The computational cost of the algorithm is low provided approximations in step 2 are fast. These are discussed in the following section.

In high-dimensional cases ($n > p$), we found that joint estimation of priors is a difficult task because the prior on τ_i tends to dominate the prior on σ_i, which is pushed towards large precisions. To overcome this, we propose to first scale the prior on σ_i by estimating it (and only it) using the intercept SEM. Next, we estimate the prior on τ_i conditionally on the prior on σ_i with the procedure described above. This strategy

connects to the idea that the prior on regression parameters should be defined conditionally on noise variances (Park and Casella, 2008). In simulation 6.3.2, we adopt this strategy when $n \leq p$.

6.2.3 Approximations of posteriors. In Bayesian SEM (6.2), the posteriors of interest are $\pi(\boldsymbol{\beta}_i|\mathbf{y}_i)$ (with $\boldsymbol{\beta}_i = (\beta_{i,1},\ldots,\beta_{i,i-1},\beta_{i,i+1},\ldots,\beta_{i,p})^{\mathrm{T}}$) for inferring the graph structure, and $\pi_{\theta_1}(\tau_i^{-2}|\mathbf{y}_i)$ and $\pi_{\theta_2}(\sigma_i^{-2}|\mathbf{y}_i)$ for empirical Bayesian estimation of prior parameters. Here, we describe variational approximations to these posteriors. We apply a variational approximation because it is computationally efficient and provides an analytic expression for the lower bound on the log-marginal likelihood, which is useful for assessing model fit (see Section 6.2.4) and monitoring convergence of the above algorithm. However, other approximations are possible (Rajagopalan and Broemeling, 1983, Rue et al., 2009). In the following, we drop node index i and parameters of priors θ_k for clarity reasons.

We consider variational approximations that estimate posteriors by the product density that minimizes the Kullback-Leibler (KL) divergence to the true posterior. These are attractive because our model uses conjugate priors, and therefore, given a factorization form, the optimal product density is convenient to determine (Bishop, 2006, Winn and Bishop, 2005).

Consider the product density q that, we assume, factorizes as

$$q(\boldsymbol{\beta},\tau^{-2},\sigma^{-2}) = q_1(\boldsymbol{\beta})q_2(\tau^{-2})q_3(\sigma^{-2}), \tag{6.4}$$

then the product density that minimizes the KL divergence to $\pi(\boldsymbol{\beta},\tau^{-2},\sigma^{-2}|\mathbf{y})$ is (Ormerod and Wand, 2010):

$$\begin{aligned} q^*(\boldsymbol{\beta},\tau^2,\sigma^2;\mathbf{y}) &= q_1^*(\boldsymbol{\beta};\mathbf{y})q_2^*(\tau^{-2};\mathbf{y})q_3^*(\sigma^{-2};\mathbf{y}) \\ q_1^*(\boldsymbol{\beta};\mathbf{y}) &=^d \mathcal{N}\left(\boldsymbol{\beta}^*,\Sigma^*\right) \\ q_2^*(\tau^{-2};\mathbf{y}) &=^d \mathcal{G}\left(a_1^*,b_1^*\right) \\ q_3^*(\sigma^{-2};\mathbf{y}) &=^d \mathcal{G}\left(a_2^*,b_2^*\right) \end{aligned} \tag{6.5}$$

where

$$\begin{aligned} \Sigma^* &= \left(\frac{a_2^*}{b_2^*}\mathrm{X}^{\mathrm{T}}\mathrm{X} + \frac{a_1^*}{b_1^*}\mathbb{I}_p\right)^{-1}, \\ \boldsymbol{\beta}^* &= \left(\frac{a_2^*}{b_2^*}\right)\Sigma^*\mathrm{X}^{\mathrm{T}}\mathbf{y}, \\ a_1^* &= a_1 + p/2, \\ b_1^* &= b_1 + 0.5\left(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{\beta}^* + \mathrm{tr}(\Sigma^*)\right), \\ a_2^* &= a_2 + n/2, \text{ and} \\ b_2^* &= b_2 + 0.5\left((\mathbf{y}-\mathrm{X}\boldsymbol{\beta}^*)^{\mathrm{T}}(\mathbf{y}-\mathrm{X}\boldsymbol{\beta}^*) + \mathrm{tr}(\mathrm{X}^{\mathrm{T}}\mathrm{X}\Sigma^*)\right). \end{aligned}$$

Here X refers to the appropriate design matrix, which for the i^{th} equation equals $\mathcal{Y}_{-i} = [\mathbf{y}_1 \dots \mathbf{y}_{i-1}\mathbf{y}_{i+1} \dots \mathbf{y}_n]$.

The joint density q^* maximizes the lower bound $\mathcal{L}$ on the log-marginal likelihood $\log \pi(\mathbf{y})$:

$$\begin{aligned}\mathcal{L} = {} & 0.5p - (n/2)\log(2\pi) + (1/2)\log|\Sigma^*| + a_2 \log b_2 - a_2^* \log b_2^* + \\ & \log\Gamma(a_2^*) - \log\Gamma(a_2) + a_1 \log b_1 - a_1^* \log b_1^* + \log\Gamma(a_1^*) - \log\Gamma(a_1).\end{aligned} \tag{6.6}$$

The optimal densities in (6.5) have the property that their parameters depend on each other, apart from a_1^* and a_2^* that are deterministic. This motivates the following algorithm to estimate Σ^*, $\boldsymbol{\beta}^*$, b_1^* and b_2^* (Ormerod and Wand, 2010):

1. Initiate b_1^* and b_2^*
2. Update Σ^*, $\boldsymbol{\beta}^*$, b_1^*, b_2^* and $\mathcal{L}$ in that order using their above expressions
3. Repeat step 2 until convergence of $\mathcal{L}$

Upon convergence, (marginal) posteriors $\pi(\boldsymbol{\beta}|\mathbf{y})$, $\pi(\tau^{-2}|\mathbf{y})$ and $\pi(\sigma^{-2}|\mathbf{y})$ are approximated by $q_1^*(\boldsymbol{\beta})$, $q_2^*(\tau^{-2})$ and $q_3^*(\sigma^{-2})$, respectively.

This algorithm intervenes as the maximization step of the EM algorithm (step 2) in Section 6.2.2. In this case, for computational efficiency, we only perform one iteration. Note that the empirical Bayesian estimation procedure connects to other variational EM algorithms in literature (Corduneanu and Bishop, 2001).

In high-dimensional setting, it is preferable to use an SVD decomposition of $\mathrm{X} = \mathrm{UDV}^{\mathrm{T}} = \mathrm{FV}^{\mathrm{T}}$ and work with the n by n matrix F. Then, the posterior $\pi(\boldsymbol{\beta}|\mathbf{y})$ is approximated by a linear transformation of the Gaussian density $q_1^*(\boldsymbol{\upsilon})$, where $\boldsymbol{\upsilon}$ is the parameter vector in the new space. Note that for the purpose of variable selection we are interested in the posterior expectation $\mathrm{E}_\pi\left[\boldsymbol{\beta}|\mathbf{y}\right]$ and standard deviation $sd_\pi\left[\boldsymbol{\beta}|\mathbf{y}\right]$ of $\boldsymbol{\beta}$, hence the (high-dimensional) posterior covariance needs not to be determined in its entirety, only its diagonal.

6.2.4 Selection of edges. To recover the graph structure, variable selection is required in SEM (6.2). We here describe a model-based approach, which connects to the recent selection procedure of Bondell and Reich (2012). However, we use marginal likelihood to compare the fit of models with well-informed priors.

Let us denote $\mathrm{E}_\pi\left[\beta_{i,j}|\mathbf{y}\right]$ and $sd_\pi\left[\beta_{i,j}|\mathbf{y}\right]$ the posterior expectation and standard deviation of $\beta_{i,j}$, and define $\kappa_{i,j} = |\mathrm{E}_\pi\left[\beta_{i,j}|\mathbf{y}\right]|/sd_\pi\left[\beta_{i,j}|\mathbf{y}\right]$, $\forall i,j \in \mathcal{I}$ with $i \neq j$. Our method is to rank edges according to $\bar{\kappa}_{i,j} = (\kappa_{i,j} + \kappa_{j,i})/2$ and operate forward selection. Denoting m_k^{E} the log-marginal likelihood of regression equation k (as measured by the variational lower bound $\mathcal{L}_k$) in the SEM determined by edge set E, then forward selection is accomplished by the following iterative scheme:

1. Initiate $\ell = 0$, $\Gamma^{(0)} = \mathcal{I}^2$ and $\mathsf{E}^{(0)} = \emptyset$
2. Determine $(r,s) = \underset{k,l}{\operatorname{argmax}}(\bar{\kappa}_{k,l})$, $\forall (k,l) \in \Gamma^{(\ell)}$
3. Only if $(m_r^{\mathsf{E}^{(\ell)} \cup \{(r,s)\}} + m_s^{\mathsf{E}^{(\ell)} \cup \{(r,s)\}}) > (m_r^{\mathsf{E}^{(\ell)}} + m_s^{\mathsf{E}^{(\ell)}})$ update $\Gamma^{(\ell+1)} = \Gamma^{(\ell+1)} \setminus \{(r,s)\}$, $\mathsf{E}^{(\ell+1)} = \mathsf{E}^{(\ell)} \cup \{(r,s)\}$, $\ell = \ell + 1$ and go back to previous step

Note that $m_k^{\emptyset}$ is determined by the SEM with intercepts only. We finally estimate $\mathcal{E}$ by the last update of E.

Our method performs simultaneous selection of (the two) parameters that are associated with each edge. We found this is a good compromise which acts a smooth transition between the post-symmetrization procedures based on 'AND' or 'OR' rule (Meinshausen and Bühlmann, 2006), which are found to be either too conservative or liberal.

6.3 Model-based simulations.

6.3.1 Accuracy of hyperparameter estimates. The simulation study examined the accuracy of hyperparameter estimates obtained with the procedure described in Section 6.2.2. To do so, we use the following simulation model:

$$
\begin{aligned}
\mathrm{Y}_j &= \mathrm{X}\beta_j + \epsilon_j, \quad j = 1, \ldots, q \\
\beta_j &\sim \mathcal{N}(0, \tau_j^2 \mathrm{I}_p) \\
\epsilon_j &\sim \mathcal{N}(0, \sigma_j^2 \mathrm{I}_n) \\
\tau_j^{-2} &\sim \mathcal{G}(a_1, 1) \\
\sigma_j^{-2} &\sim \mathcal{G}(5, 1)
\end{aligned}
\tag{6.7}
$$

The n by p design matrix X was generated from a normal distribution $\mathcal{N}(0, \Phi)$ where Φ has entries $\phi_{k,l} = 0.7^{|k-l|}$, $\forall k,l \in \{1,\ldots,p\}$. We consider a medium-dimensional model and set $n = 100$ and $p = q = 50$. To study the accuracy of estimation when priors are rather informative, we take hyperparameter $a_1 \in \{5, 10, 50, 100\}$. In other words, different amounts of regularization are considered. For each case, precisions σ_j^{-2}, τ_j^{-2} and β_j were generated according to their prior and subsequently considered fixed. Our simulation experiment consisted in generating $\forall j \in \{1,\ldots,q\}$ errors from $\mathcal{N}(0, \sigma_j^2 \mathrm{I}_n)$ and obtaining estimates of a_1, b_1, a_2 and b_2 using our procedure. This was repeated 100 times. Table 6.1 reports the mean and standard deviations of hyperparameter estimates.

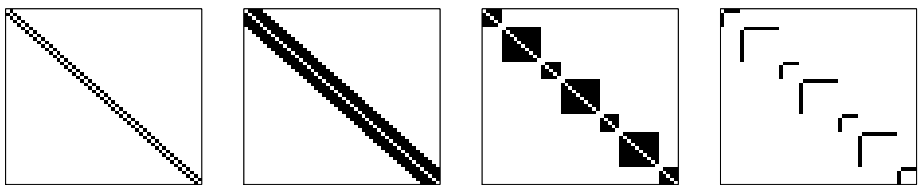
Clearly, estimates in Table 6.1 are in average close to the true values of parameters. Hence, estimation is accurate. We also observe that when a_1 increases the variance of $\hat{a}_1$ becomes larger. This reflects a loss of efficiency when the prior on precisions of regression parameters is more informative. In all, the present experiment confirms that the estimation procedure of Van de Wiel et al. (2013) is appropriate for the type

of model under study.

	Estimates			
a_1	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$
5	4.735 (0.059)	1.039 (0.011)	5.037 (0.177)	0.993 (0.027)
10	9.584 (0.285)	1.043 (0.011)	5.101 (0.198)	1.034 (0.029)
50	52.769 (4.359)	1.046 (0.005)	5.722 (0.253)	1.059 (0.027)
100	100.454 (16.285)	1.027 (0.003)	5.638 (0.247)	1.023 (0.026)

TABLE 6.1. *Mean and standard deviation (in parentheses) of hyperparameter estimates*

6.3.2 Graph structure recovery. We are interested in evaluating the performance of the Bayesian SEM in recovering an undirected graph structure and comparing it to various popular approaches. To do so, we generate 100 samples from a multivariate normal distribution with vector mean 0 and $d \times d$ covariance matrix Σ_d^{-1}, where $d \in \{20, 50, 100, 200\}$. Different patterns of non-zero entries are considered in the precision matrix $\Sigma_d = (\sigma_{ij})_{i,j \in \{1,\dots,d\}}$, corresponding to different graph structures. We consider popular *band*, *cluster* and *hub* structures (Zhao et al., 2012), which we illustrate in Figure 6.1.



(a) Band 1 (b) Band 2 (c) Cluster (d) Hub

FIGURE 6.1. *Structures considered for the precision matrix Σ_{50} in our simulation. Black and white dots represent non-zero and zero entries, respectively. Only off-diagonal elements are displayed. For precision matrices with block-diagonal structures (clusters and hubs), block sizes were set to 5 and 10 regardless of the dimension d. Hence, sparsity increases with dimension. Band 1 and 2 are diagonal matrices with bandwidth 1 and 4, respectively.*

We generate Σ_d as follows. We start with the null matrix and generate non-zero entries independently from $\mathcal{N}(0.5, 0.1^2)$. Then, to make the matrix positive definite we increment its diagonal by $|c| + 0.1$, where c is the lowest eigenvalue. Finally, the matrix is subsequently scaled so to have unit diagonal. The simulation of sparse partial correlations matrix is a difficult task mainly because the scaling step disrupts the range

of generated values (Krämer et al., 2009). Entries are then greatly shrunk, which renders graph reconstruction harder. In our simulation setting, we found that generating entries according to $\mathcal{N}(0.5, 0.1^2)$ was giving reasonable ranges for partial correlations. In Appendix E.1, we provide statistical summaries on generated partial correlations.

We compared the Bayesian SEM with shrinkage priors ($\text{SEM}_{\text{B\_SHRINK}}$) to the SEM with ℓ_1-penalty (SEM_{L}) (Meinshausen and Bühlmann, 2006), the Graphical Lasso (GL_λ) (Friedman et al., 2008) and GeneNet (Schaefer et al., 2006) that uses a ridge-type penalty with *a posteriori* edge selection. Additionally, we included $\text{SEM}_{\text{L\_STAB}}$ and GL_{STAB}, the improved versions of SEM_{L} and GL_λ (respectively) based on stability selection (Meinshausen and Bühlmann, 2010). In cases where $d \in \{20, 50\}$ (i.e, $n > p$), for the purpose of comparison with $\text{SEM}_{\text{B\_SHRINK}}$, we also considered the Bayesian SEM (SEM_{B}) with non-informative priors $\mathcal{G}(0.001, 0.001)$ on τ_i^{-2} and σ_i^{-2}.

Briefly, graph selection is as follows: for SEM_{L} and GL_λ we use BIC for selecting the optimal regularization parameter(s) and a threshold along with stability selection; for GeneNet a cut-off on FDRs is taken; and for the Bayesian SEMs we use the approach described in Section 6.2.4. In Appendix E.2, we provide for each method more details as to how edge ranking is obtained and how edges are selected.

To evaluate the performance of methods in recovering the graph structure we show partial ROC curves (which depict the true positive rate (TPR) as a function of the false positive rate (FPR), when FPR< 0.2) and average TPRs and FPRs on selected structures as these provide complementary pictures on accuracies. The former gives information on edge ranking (although it does not compare it to a true ranking) while the other evaluates edge selection, which is often the most difficult problem. Results are reported in Figures 6.2 to 6.5 and Tables 6.2 to 6.5. We now discuss these.

The relatively less sparse cases $d \in \{20, 50\}$ (i.e. $n > p$) are informative, because they allow the comparison of SEM_{B} and $\text{SEM}_{\text{B\_SHRINK}}$, and observing the effect of shrinkage. The simulation results show that shrinkage priors can substantially improve structure recovery with the Bayesian SEM. Indeed, partial ROC curves in Figures 6.2 and 6.3 indicate that for most structures shrinkage improves edge ranking. This is particularly true with cluster structures. Shrinkage seems also beneficial for identifying the graph. Tables 6.2 and 6.3 show that $\text{SEM}_{\text{B\_SHRINK}}$ identifies more edges that are true while keeping a low error rate. For hub structures, the shrinkage does not appear to improve edge ranking and selected graphs seem to include too many edges, higher TPR is achieved at the price of a higher FPR. This may suggest that the parametric form of the shrinkage prior is not appropriate or flexible enough to accommodate the particular structure of hubs.

The low-dimensional case where $d = 20$ shows that sparse methods can achieve poor performance when the network is smaller and less sparse. In addition, it seems that $\text{SEM}_{\text{B\_SHRINK}}$ is preferable to GeneNet for most structures.

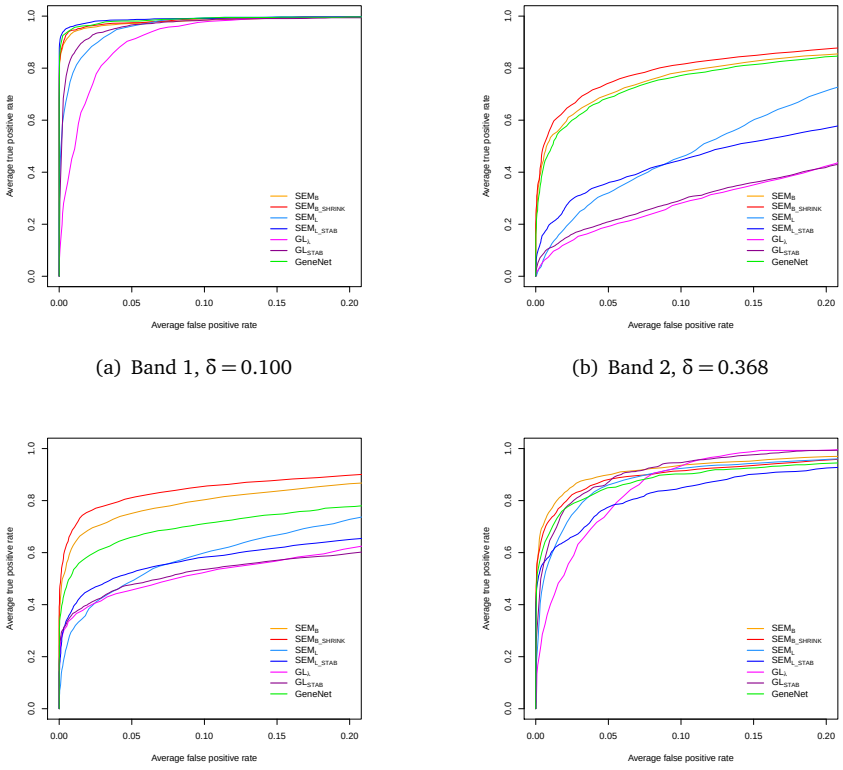
When the true precision matrix ($d \in \{20, 50\}$) is tridiagonal (band 1) most methods perform well. However, when the bandwidth is larger results are rather idiosyncratic. The graphical lasso seem to perform worse both for the ranking of edges and selection of the graph. SEM_{L} is shown to be better. The Bayesian SEM with shrinkage performs best but the selection of edges seems too conservative when $d = 50$. GeneNet shows

good performance. Only when $d = 20$ the selection of edges is difficult.

In high-dimensional settings ($d \in \{100, 200\}$), simulation results suggest that our approach can recover sparse structures. It compares well with other methods for edge ranking. Edge selection seem more difficult since TPRs are often somewhat lower than others. However, it is interesting to see that $\text{SEM}_{\text{B\_SHRINK}}$ usually achieves a fairly low FPR (as opposed to the BIC-based methods). GeneNet has an even lower FPR, but usually also a lower TPR. SEM_{L} performs well for edge ranking but BIC results in high FPRs and too many selected edges. Stability selection improves performances, especially for identifying the graph. The graphical lasso is shown to be inferior to others (for all structures) for edge ranking and BIC can results in too many edges. Again, stability selection enhances performance, sometimes quite dramatically. For example, GL_λ performs worse on hub structures but, combined with stability selection, it does best for edge ranking and graph selection.

In all, none of the methods performed uniformly better than others. The Bayesian SEM is shown to perform well on a variety of structures and in high-dimension situations where it is comparable to other sparse methods. In low- and medium-dimensional problems, it can clearly outperform others, including resampling-based methods. The use of shrinkage priors can substantially improve graph structure recovery. However, we observed that the model meets difficulties in the presence of hub structures. For the SEM with ℓ_1 penalty and the graphical lasso, simulation results suggest that they are best combined with stability selection. Only then, good performance is achieved for both ranking and selection. Graph selection based on the BIC criterion may not be appropriate. Finally, edge ranking with GeneNet is often good, however, edge selection is rather conservative.

Note that, from a computational perspective, it seems most fair to compare $\text{SEM}_{\text{B\_SHRINK}}$ with SEM_{L}, GL_λ and GeneNet only, because these together with our approach do not require cross-validation or resampling. $\text{SEM}_{\text{B\_SHRINK}}$ often outperforms those even for the sparser settings.



(a) Band 1, $\delta = 0.100$

(b) Band 2, $\delta = 0.368$

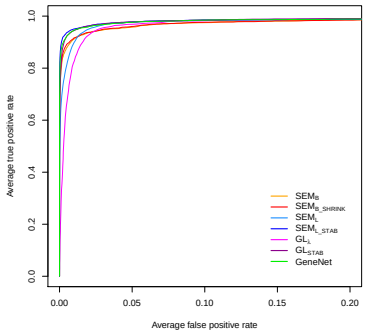
(c) Cluster, $\delta = 0.342$

(d) Hub, $\delta = 0.089$

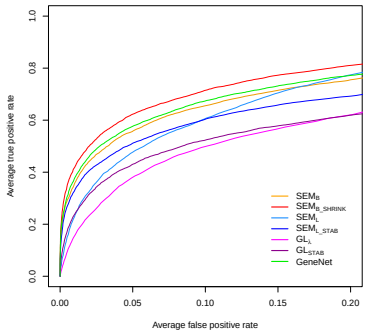
FIGURE 6.2. *Partial ROC-curves for the different graph structures when $d = 20$. Each plot depicts the average true positive rate (y-axis) as a function of the average false positive rate (x-axis) over 50 repetitions. For each network structure we indicate the density of true edges δ.*

	Band 1		Band 2		Cluster		Hub	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
SEM_B	0.937	0.010	0.185	0.003	0.411	0.006	0.807	0.015
SEM_{B_SHRINK}	0.965	0.032	0.265	0.007	0.590	0.013	0.881	0.064
SEM_{L_BIC}	0.958	0.019	0.618	0.142	0.502	0.069	0.952	0.022
SEM_{L_STAB}	0.185	0.000	0.180	0.008	0.144	0.000	0.093	0.000
GL_{BIC}	0.998	0.282	0.905	0.622	0.563	0.147	0.996	0.351
GL_{STAB}	0.933	0.027	0.210	0.050	0.264	0.001	0.909	0.061
Genenet	0.903	0.007	0.030	0.001	0.312	0.001	0.662	0.012

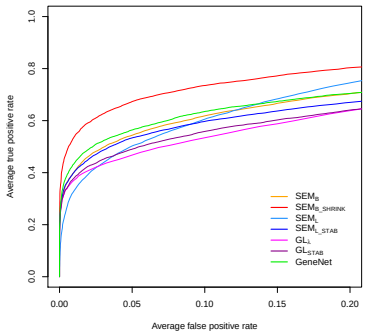
TABLE 6.2. *Average true and false positive rates of selected graphs over 50 repetitions, for each method in case $d = 20$.*



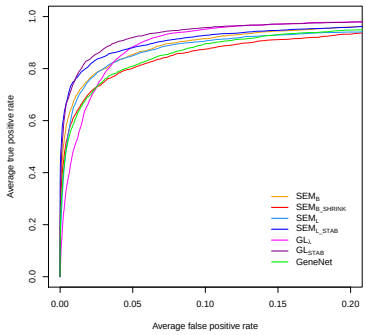
(a) Band 1, $\delta = 0.040$



(b) Band 2, $\delta = 0.155$



(c) Cluster, $\delta = 0.143$

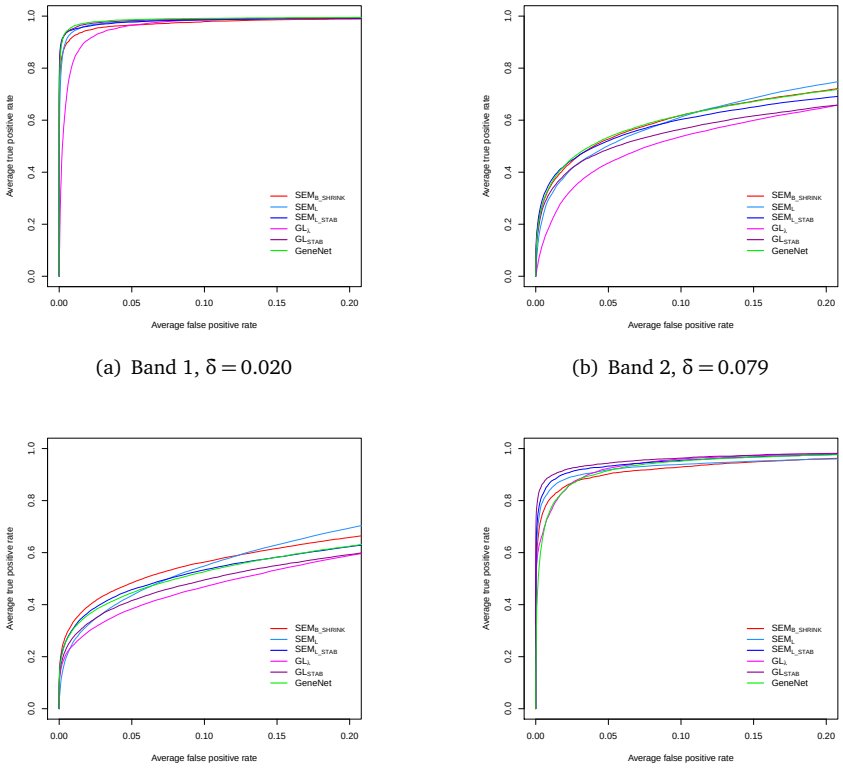


(d) Hub, $\delta = 0.035$

FIGURE 6.3. *Partial ROC-curves for the different graph structures when $d = 50$. Each plot depicts the average true positive rate (y-axis) as a function of the average false positive rate (x-axis) over 50 repetitions. For each network structure we indicate the density of true edges δ.*

	Band 1		Band 2		Cluster		Hub	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
SEM_B	0.800	0.001	0.186	0.002	0.296	0.002	0.480	0.002
SEM_{B_SHRINK}	0.915	0.012	0.187	0.002	0.413	0.005	0.703	0.024
SEM_{L_BIC}	0.925	0.007	0.397	0.032	0.302	0.009	0.768	0.007
SEM_{L_STAB}	0.893	0.001	0.304	0.008	0.287	0.001	0.635	0.003
GL_{BIC}	0.978	0.084	0.555	0.142	0.438	0.034	0.926	0.076
GL_{STAB}	0.929	0.007	0.353	0.028	0.280	0.001	0.773	0.013
Genenet	0.910	0.006	0.237	0.002	0.353	0.004	0.475	0.005

TABLE 6.3. *Average true and false positive rates of selected graphs over 50 repetitions, for each method in case $d = 50$.*



(a) Band 1, $\delta = 0.020$

(b) Band 2, $\delta = 0.079$

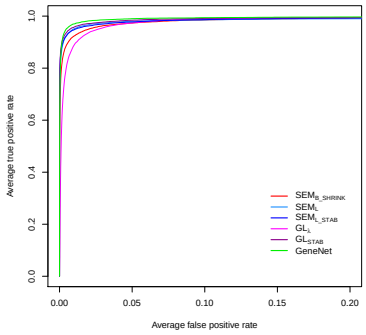
(c) Cluster, $\delta = 0.071$

(d) Hub, $\delta = 0.017$

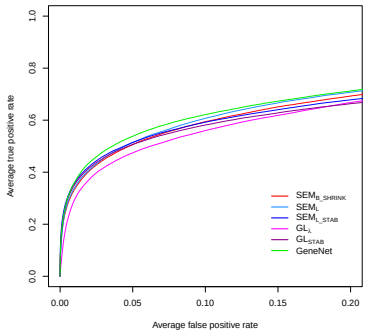
FIGURE 6.4. *Partial ROC-curves for the different graph structures when $d = 100$. Each plot depicts the average true positive rate (y-axis) as a function of the average false positive rate (x-axis) over 50 repetitions. For each network structure we indicate the density of true edges δ.*

	Band 1		Band 2		Cluster		Hub	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
SEM_{B_SHRINK}	0.886	0.005	0.211	0.003	0.301	0.007	0.730	0.005
SEM_{L_BIC}	0.915	0.143	0.516	0.288	0.479	0.322	0.791	0.117
SEM_{L_STAB}	0.928	0.004	0.331	0.007	0.264	0.005	0.827	0.005
GL_{BIC}	0.949	0.038	0.399	0.039	0.230	0.007	0.859	0.024
GL_{STAB}	0.940	0.006	0.337	0.012	0.203	0.003	0.851	0.003
Genenet	0.939	0.005	0.208	0.002	0.174	0.001	0.653	0.005

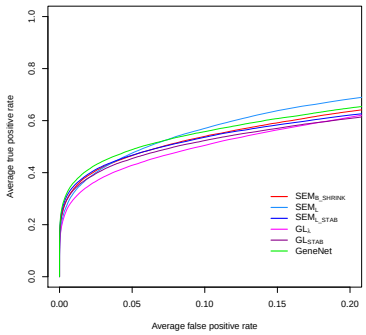
TABLE 6.4. *Average true and false positive rates of selected graphs over 50 repetitions, for each method in case $d = 100$.*



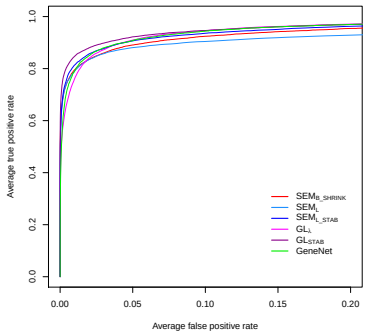
(a) Band 1, $\delta = 0.010$



(b) Band 2, $\delta = 0.040$



(c) Cluster, $\delta = 0.035$



(d) Hub, $\delta = 0.009$

FIGURE 6.5. *Partial ROC-curves for the different graph structures when $d = 200$. Each plot depicts the average true positive rate (y-axis) as a function of the average false positive rate (x-axis) over 50 repetitions. For each network structure we indicate the density of true edges δ.*

	Band 1		Band 2		Cluster		Hub	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
SEM_{B_SHRINK}	0.876	0.005	0.269	0.005	0.283	0.005	0.750	0.006
SEM_{L_BIC}	0.933	0.094	0.592	0.201	0.584	0.214	0.852	0.120
SEM_{L_STAB}	0.919	0.004	0.310	0.006	0.301	0.004	0.769	0.005
GL_{BIC}	0.908	0.013	0.317	0.012	0.264	0.006	0.765	0.010
GL_{STAB}	0.965	0.014	0.404	0.018	0.351	0.013	0.862	0.014
Genenet	0.929	0.003	0.217	0.002	0.222	0.001	0.623	0.002

TABLE 6.5. *Average true and false positive rates of selected graphs over 50 repetitions, for each method in case $d = 200$.*

6.4 Conclusion. In this chapter, we proposed a Bayesian SEM with shrinkage priors for recovering the structure of an undirected (Gaussian) graph. The model employs Gaussian priors to impose regularization. Because these are not sparse, a selection procedure was presented that uses posterior thresholding to infer the graph structure. The computational efficiency of the Bayesian SEM is accomplished by fast variational approximations of posteriors and the help of SVD decompositions. The empirical Bayesian estimation of prior parameters, which connects to a variational EM algorithm, is also fast. In simulations, we have shown that the proposed approach is competitive with popular sparse methods in high-dimensional cases and often superior in low- and medium-dimensional ones. Our approach is computationally competitive with frequentist methods. To illustrate this, in all simulations we have undertaken the computing time of the Bayesian SEM was always less than a minute (and often less than half a minute). This was found to be comparable to the frequentist approaches included in the simulation study, provided the tuning procedure of regularization parameters is accounted for. An important practical advantage of our approach is that the estimation procedure is coherent and complete, so does not rely on tuning, resampling or cross-validation to estimate the regularization parameter(s). This is in particular encouraging for extending the method to settings with multiple types of high-dimensional covariates which would require different amounts of shrinkage. For methods based on resampling or cross-validation this may become overly computationally burdensome.

A novelty of our work is the use of shrinkage priors that allow sharing of information across regression equations. To our knowledge, few works go in that direction. For example, Yuan et al. (2012) borrow information about the regularizing parameters corresponding to ℓ_1-penalties by combining local and global searches. In Bayesian SEMs for graphical modeling, the focus is often on studying the equivalence in between the SEM and a proper joint distribution (Dobra et al., 2004, Geiger and Heckerman, 2002). We are not aware of any previous works using shrinkage priors. In this chapter, we have shown that these could improve graph structure recovery.

The proposed method is particularly suitable for gene network reconstruction using expression data. This type of network aims at providing a picture of regulatory mechanisms that act between genes. In practice, the interest often lies in a relatively small subset of genes that are known to be functionally linked (e.g. a pathway). In this context, the Bayesian SEM may be more appropriate than others, because such a gene set is usually of moderate dimension and, hence, due to the functional link the corresponding network is likely to be relatively less sparse. Therefore, strong dependencies between genes are more likely to occur and this may be in favor of ridge-type regularization. In addition, the coherence in functionality may render shrinkage to be beneficial for parameter estimation in the SEM.

We have focused on only recovering the support of the precision matrix. However, it is also possible to obtain an estimate of it. An immediate approach is to use the graph structure provided by the Bayesian SEM as a prior for precision estimation. In a Bayesian setting, this happens naturally and literature refers to this step as *parameter learning* (Scutari, 2013). For estimation, versions of the conjugate Wishart distribution, such as the G-Wishart (Dobra et al., 2011, Wang and Li, 2012), are com-

putationally attractive. Other estimation strategies have been proposed outside the Bayesian paradigm. These are usually based on the idea of thresholding. See, for example, Zhou et al. (2011) and Yuan (2010).

We foresee several extensions. Motivated by simulations, which show a deterioration of performance using shrinkage priors in the presence of hub structures, more flexible shrinkage priors could be considered. For example, the particular form of hub structures may suggests that a gamma mixture prior for precisions of regression parameters may be more appropriate. SEMs being appropriate for directed networks, it would be interesting to investigate types of shrinkage priors that are suitable in this context. For example, it may be more appropriate to shrink differently in- and outgoing edges. Extension to non-Gaussian data is possible. In this context, however, it may be desirable to adopt a flexible likelihood model and other types of posterior approximations may be considered (Rue et al., 2009).

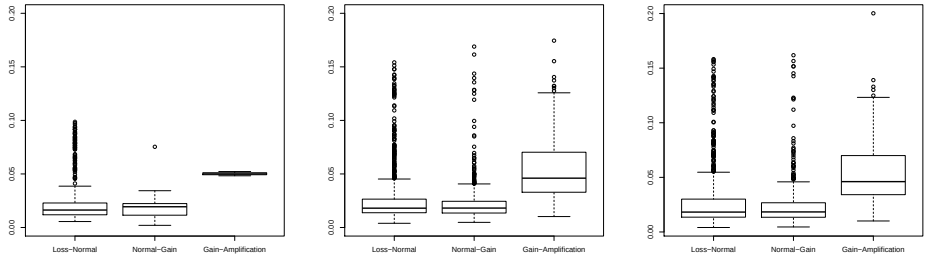
APPENDIX A

A.1 Overlap of model selection procedures.

	Carvalho et al. (2009)				Neve et al. (2006)			
	AIC				**AIC**			
OSAIC	INT	LIN	PLE	PLI	INT	LIN	PLE	PLI
INT	14720	0	0	0	5081	0	0	0
LIN	1563	3352	0	0	504	4758	0	0
PLE	897	19	1751	0	770	58	1933	0
PLI	903	303	227	2134	613	1873	550	3084
	BIC				**BIC**			
OSAIC	INT	LIN	PLE	PLI	INT	LIN	PLE	PLI
INT	14720	0	0	0	5081	0	0	0
LIN	3232	1683	0	0	1334	3928	0	0
PLE	1860	21	786	0	1623	79	1059	0
PLI	1888	339	208	1132	1341	2338	554	1887
	BIC				**BIC**			
AIC	INT	LIN	PLE	PLI	INT	LIN	PLE	PLI
INT	18083	0	0	0	6968	0	0	0
LIN	1699	1975	0	0	889	5800	0	0
PLE	982	16	979	0	860	72	1545	0
PLI	936	52	17	1130	662	473	71	1884

TABLE A.1. *Pairwise overlap comparison of model selection procedures for the two data sets. The number of times a model is selected by the type of model (INT=intercept, LIN=linear, PLE=piecewise level and PLI=piecewise linear) is displayed.*

A.2 Simulation: precision of knots. We conducted a small simulation to determine the standard deviation of knots $\hat{\alpha}_j$, $j = 1,2,3$. We randomly selected 1000 genes in both data sets. For each gene we sampled with replacement the copy number data and re-estimated knots using method I (both data sets) and II (only for the breast cancer data set as call probabilites were not available for the other) ; this was repeated 1000 times. Figure A.1 displays the distribution of the standard deviation of the three knots over the 1000 genes in each data set using the different methods. All boxplots show that standard deviations of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are very small (mostly below 0.05). Standard deviations of $\hat{\alpha}_3$ are larger (probably because amplifications may be a rare event and only concern few individuals) but still very small (mostly below 0.1). Figures A.1(b) and A.1(c) suggest the difference between method I and II for determining knots is negligible in the breast cancer data set.



(a) (b) (c)

FIGURE A.1. *Boxplots displaying the distribution of the standard deviation of $\hat{\alpha}_1$ (knot between states -1 and 0), $\hat{\alpha}_2$ (knot between states 0 and 1) and $\hat{\alpha}_2$ (knot between states 1 and 2) for (a) the colorectal data set (1000 genes) using method I, (b) the breast cancer data (1000 genes) using method I (c) and II.*

A.3 Testing results.

	Linear		OSAIC		Full	
	p	q	p	q	p	q
APC	2.49e-02	2.04e-01	2.26e-02	6.38e-02	2.49e-02	2.12e-01
ATP11A	7.34e-06	9.79e-04	5.88e-06	2.98e-04	2.08e-05	2.26e-03
C20orf24	1.71e-12	2.21e-08	3.06e-13	1.14e-09	3.68e-13	4.76e-09
JMJD6	5.44e-09	4.85e-06	1.78e-08	4.31e-06	2.99e-08	1.89e-05
MTUS1	6.83e-07	1.77e-04	6.38e-08	1.06e-05	1.72e-07	6.34e-05
RPRD1B	3.18e-06	5.45e-04	5.17e-07	5.02e-05	1.13e-06	2.67e-04
TCFL5	6.49e-06	8.88e-04	1.75e-08	4.31e-06	1.01e-07	4.22e-05
TH1L	1.06e-10	3.25e-07	2.72e-13	1.14e-09	7.14e-13	6.16e-09
TP53	6.54e-03	9.87e-02	9.42e-05	2.25e-03	2.55e-04	1.34e-02
ATMIN	1.12e-09	6.45e-08	1.13e-09	4.56e-08	5.24e-09	2.96e-07
CCND1	1.91e-08	5.71e-07	3.56e-08	6.88e-07	1.62e-07	4.15e-06
CEP350	8.55e-08	1.93e-06	3.07e-10	1.69e-08	5.74e-10	5.33e-08
EIF3H	1.70e-12	4.88e-10	8.22e-15	3.75e-12	1.05e-13	6.74e-11
ERBB2	4.46e-10	3.18e-08	4.34e-10	2.15e-08	2.48e-08	9.64e-07
FGFR1	1.62e-06	2.03e-05	3.99e-10	2.02e-08	8.90e-09	4.34e-07
PAK1	1.15e-10	1.21e-08	<2.2e-16	<2.2e-16	2.66e-15	3.94e-12
PITPNA	1.85e-06	2.25e-05	8.40e-10	3.66e-08	1.62e-08	6.93e-07
PTEN	7.27e-09	2.64e-07	9.10e-15	4.02e-12	9.55e-15	9.66e-12

TABLE A.2. *P and q-values of the test when under the alternative hypothesis H_a the linear, OSAIC-selected and the full models are successively considered. The top and bottom parts correspond respectively to the selected genes from the colorectal and breast data.*

A.4 Simulation: PLRS as screening test. The simulation study aimed to determine and compare the performance of the PLRS testing procedure in detecting associations of various functional shapes. The simulation is based on the recent work of Louhimo et al. (2012), given some modifications.

A.4.1 Simulation settings. We were interested in simulating relationships of different parametric forms. To do so, we used real DNA copy number data (Neve et al., 2006) of 80 randomly selected genes. As we aimed to build different models based on segmented and called copy number data we selected genes which presented various types of copy number aberration affecting samples in different proportions (see Table A.3). We selected 40 genes with only two aberration states (20 with loss and normal; 20 with normal and gain) and 40 others with three states (20 with loss, normal and gain; 20 with normal, gain and amplification).

We built upon the copy number data different types of association with gene expression. Let's define $f(x,s)$ the gene expression as a function of segmented and called DNA copy number x and s, respectively; then, we considered the following parametric forms for f:

- linear (LIN):
  $f_{\text{LIN}}(x,s) = ax+2$, where $a \in \{0.4, 0.6, 0.8\}$
- piecewise linear with equal slopes (PLES):
  $f_{\text{PLES}}(x,s) = ax1_{\{s\neq 0\}}+2$, where $a \in \{0.4, 0.6, 0.8\}$
- two-state stepwise (STEP2):
  $$f_{\text{STEP2}}(x,s) = \begin{cases} 2-j & \text{if } s=-1 \\ 2 & \text{if } s=0 \\ 2+j & \text{if } s=1 \end{cases}$$
  where $s \in \{-1,0\}$ or $s \in \{0,1\}$ and $j \in \{0.2, 0.3, 0.4\}$
- three-state stepwise (STEP3):
  $$f_{\text{STEP3}}(x,s) = \begin{cases} 2-j_1 & \text{if } s=-1 \\ 2 & \text{if } s=0 \\ 2+j_2 & \text{if } s=1 \end{cases}$$
  where $s \in \{-1,0,1\}$, $j_1 \in \{0, 0.2, 0.4\}$ and $j_2 \in \{0.2, 0.3, 0.4\}$
- two-state piecewise linear with unequal slopes (PLUS2):
  $$f_{\text{PLUS2}}(x,s) = \begin{cases} 2+ax & \text{if } s=-1 \\ 2 & \text{if } s=0 \\ 2+ax & \text{if } s=1 \end{cases}$$
  where $s \in \{-1,0\}$ or $s \in \{0,1\}$ and $a \in \{0.4, 0.6, 0.8\}$

- three-state piecewise linear with unequal slopes (PLUS3):

$$f_{\text{PLUS3}}(x,s) = \begin{cases} 2+a_1x & \text{if } s=-1 \\ 2 & \text{if } s=0 \\ 2+a_2x & \text{if } s=1 \end{cases}$$

where $s \in \{-1,0,1\}$, $a_1 \in \{0, 0.2, 0.4\}$ and $a_2 \in \{0.4, 0.6, 0.8\}$

Above, we distinguished between two and three-state relationships as we noticed clear differences in results depending on the complexity of the relationship. To make the simulation realistic, the values of parameters $\{a, j, j_1, j_2, a_1, a_2\}$ were motivated by data. For example, we have observed that for most of the genes in the data set the average difference in expression between samples presenting a loss and a normal copy number was less than 0.4. This turned out to be similar between states gain and normal. Hence our choice of values $\{0.2, 0.3, 0.4\}$ for jumps j, j_1 and j_2. Similarly, we found that $\{0.4, 0.6, 0.8\}$ were reasonable values for slopes a, a_1 and a_2. Together the above functions and the parameter values yielded 30 different types of associations. Note that apart from the linear function (f_{LIN}), a normal copy number is not allowed to affect gene expression.

We also considered several nonlinear associations such as the cubic, sigmoid and mixture of sigmoids. However, results appeared to be highly variable for each method and highly dependent on the parametric form chosen. Since there is no biological motivation to prefer either form, they were not included in the present simulation.

We compared the PLRS testing procedure with the LM test, the one-sided Spearman's rank correlation test and the nonparametric test based on the stepwise model proposed by van Wieringen and van de Wiel (2009), which is one of the top-ranked methods according to Louhimo et al. (2012). For the latter test we used the R function *cisEffectTest()* from the Bioconductor package *sigaR* (version 1.1.1) and $nperm = 1000$ permutations. Subsequently, we will refer to this test as "sigaR".

We generated errors from a Gaussian distribution with mean zero and standard deviation $\sigma \in \{0.2, 0.35, 0.5\}$. This resulted in 9 cases for f_{LIN}, f_{PLES}, f_{STEP2} and f_{PLUS2}; and 27 cases for f_{STEP3} and f_{PLUS3}. In all, for each of the 90 cases we repeated the simulation 100 times for each of the 80 genes.

To evaluate the different screening tests we show partial ROC-curves (which depict the true positive rate as function of the type I error cutoff α when $\alpha \leq 0.2$) as for the purpose of testing only small cutoffs are relevant (Dodd and Pepe, 2003). We also report the relative AUC, $\text{rAUC}_{0.2} = \text{pAUC}_{0.2}/(0.2^2/2)$, where $\text{pAUC}_{0.2}$ is the partial AUC when $\alpha \leq 0.2$ and $0.2^2/2$ represents the expected pAUC for an uninformative test.

A.4.2 Results. Not surprisingly, when the true association only depends on segmented copy number data and is linear (f_{LIN}; Figure A.3) the LM test is most appropriate and sigaR the least. PLRS, although slightly inferior, performs similarly to the Spearman test. Linearity, however, assumes the effect to be identical over the entire range of copy number values. In fact, this is unlikely to be true since typically a non-negligible proportion of samples have a normal copy number (see Table A.3), which

ID	L	N	ID	N	G	ID	L	N	G	ID	N	G	A
1	5	45	**21**	21	29	**41**	12	34	4	**61**	22	27	1
2	4	46	**22**	22	28	**42**	7	33	10	**62**	22	27	1
3	6	44	**23**	22	28	**43**	7	38	5	**63**	21	28	1
4	7	43	**24**	42	8	**44**	17	32	1	**64**	21	28	1
5	8	42	**25**	44	6	**45**	20	28	2	**65**	21	28	1
6	9	41	**26**	44	6	**46**	3	34	13	**66**	20	29	1
7	18	32	**27**	43	7	**47**	2	35	13	**67**	20	28	2
8	17	33	**28**	41	9	**48**	2	33	15	**68**	19	29	2
9	24	26	**29**	41	9	**49**	3	28	19	**69**	20	29	1
10	24	26	**30**	40	10	**50**	4	28	18	**70**	20	29	1
11	23	27	**31**	40	10	**51**	9	30	11	**71**	18	31	1
12	23	27	**32**	39	11	**52**	9	28	13	**72**	19	30	1
13	26	24	**33**	39	11	**53**	23	21	6	**73**	19	30	1
14	27	23	**34**	37	13	**54**	17	28	5	**74**	9	31	10
15	28	22	**35**	37	13	**55**	15	29	6	**75**	9	31	10
16	26	24	**36**	38	12	**56**	2	38	10	**76**	8	30	12
17	20	30	**37**	44	6	**57**	15	25	10	**77**	6	31	13
18	19	31	**38**	44	6	**58**	19	28	3	**78**	14	34	2
19	14	36	**39**	24	26	**59**	34	13	3	**79**	14	34	2
20	15	35	**40**	24	26	**60**	6	23	21	**80**	14	33	3

TABLE A.3. *Selected genes in the breast cancer data set (Neve et al., 2006). Displayed is the number of samples by copy number aberration state (L=loss, N=normal, G=gain and A=amplification) by gene ID (in bold).*

is not expected to shift the gene expression. If we now consider the true association to be linear for abnormal copy numbers only (f_{PLES}; Figure A.4), the PLRS procedure is slightly superior to LM and $\text{rAUC}_{0.2}^{(\text{PLRS})} > \text{rAUC}_{0.2}^{(\text{LM})}$ in all cases. Also, sigaR seems preferable to Spearman. Note that PLRS and sigaR tend to suffer most from noise.

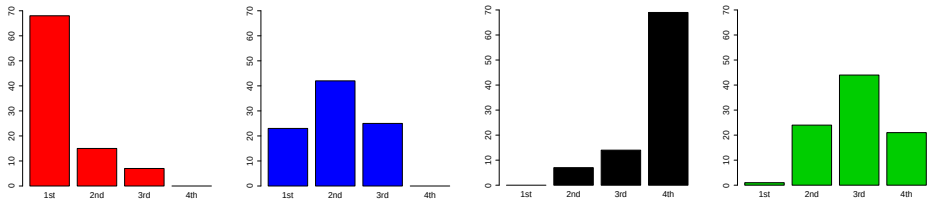
Assuming that the true association only depends on discrete genomic information and is stepwise (f_{STEP2} and f_{STEP3}), the PLRS test is generally preferable. Consider the situation where genes present only two types of DNA copy number aberration (f_{STEP2}; Figure A.5); then, PLRS and sigaR are clearly superior to others, with PLRS being slightly better ($\text{rAUC}_{0.2}^{(\text{PLRS})} \geq \text{rAUC}_{0.2}^{(\text{SIGAR})}$ in all cases). Now consider genes with three aberration states, the PLRS procedure clearly outperforms others when $j_1 = 0$ (f_{STEP3}; Figure A.6), i.e. when only one of the two abnormal states alter gene expression (partial effect). Surprisingly, in this situation sigaR performs rather badly. This may be a consequence of the uncertainty of the method as to choose correctly which groups to compare (sigaR compares expression of samples either with loss and normal, or with normal and gain). In the situation where both abnormal states affect expression, i.e. when $j_1 = 0.2$ and $j_1 = 0.4$ (f_{STEP3}; Figure A.7 and A.8), it appears that all methods are rather equivalent in detecting associations (though PLRS and LM tend to be slightly superior). The presence of two jumps $j_1 > 0$ and $j_2 > 0$ may help in detecting non-zero slopes and monotonic increasing trends, and improve the true positive rate of the LM and Spearman tests.

Finally, when the true association is piecewise linear, results are similar to those obtained when it is assumed to be stepwise: the PLRS test is preferable in general.

Consider two-state associations (f_{PLES2}; Figure A.9), PLRS outperforms others. Also, consider genes with three aberration states where only one is affecting expression (f_{PLES2} and $a_1 = 0$; Figure A.10), PLRS is clearly superior. In other cases where both abnormal states alter expression (f_{PLES2}, $a_1 > 0$ and $a_2 > 0$; Figure A.10), methods are rather comparable (although PLRS and LM are slightly superior).

A.4.3 Conclusion. In the present simulation study, none of the methods performed uniformly better (or worse) than others. In general, the Spearman test was inferior to others but as expected, it behaved well when strong noise is present. The LM test was superior when the association between DNA and mRNA was assumed to be linear. When this assumption was relaxed for normal copy number (only), it turned out to be slightly inferior to the PLRS test. In other cases, when the relationship is assumed to be stepwise or piecewise linear, it performed well only when the two abnormal states were affecting expression (strong effect). Otherwise, it is clearly inferior. The test proposed by van Wieringen and van de Wiel (2009) (sigaR) was clearly inferior when the true association was linear and piecewise linear with equal slopes. For stepwise and other piecewise linear associations it showed good performance. However, for three-state relationships, when the effect on expression is partial (only one of the two abnormal states alter gene expression), it had a poor performance. Finally, the PLRS test yielded good performance in detecting associations of various functional forms. Indeed, it achieved the highest $\text{rAUC}_{0.2}$ in 68 out of the 90 simulation cases (against 23 for the LM test; see Figures A.2). It was shown to perform reasonably well when the true effect is linear and often better otherwise. It offered a clear advantage over others when the effect is partial, i.e. when only one of the two abnormal states alter gene expression. This suggests that PLRS accommodates well both continuous and discrete genomic information.

The simulation results confirm that the PLRS procedure is superior to others in identifying various types of relationships and hence is flexible.

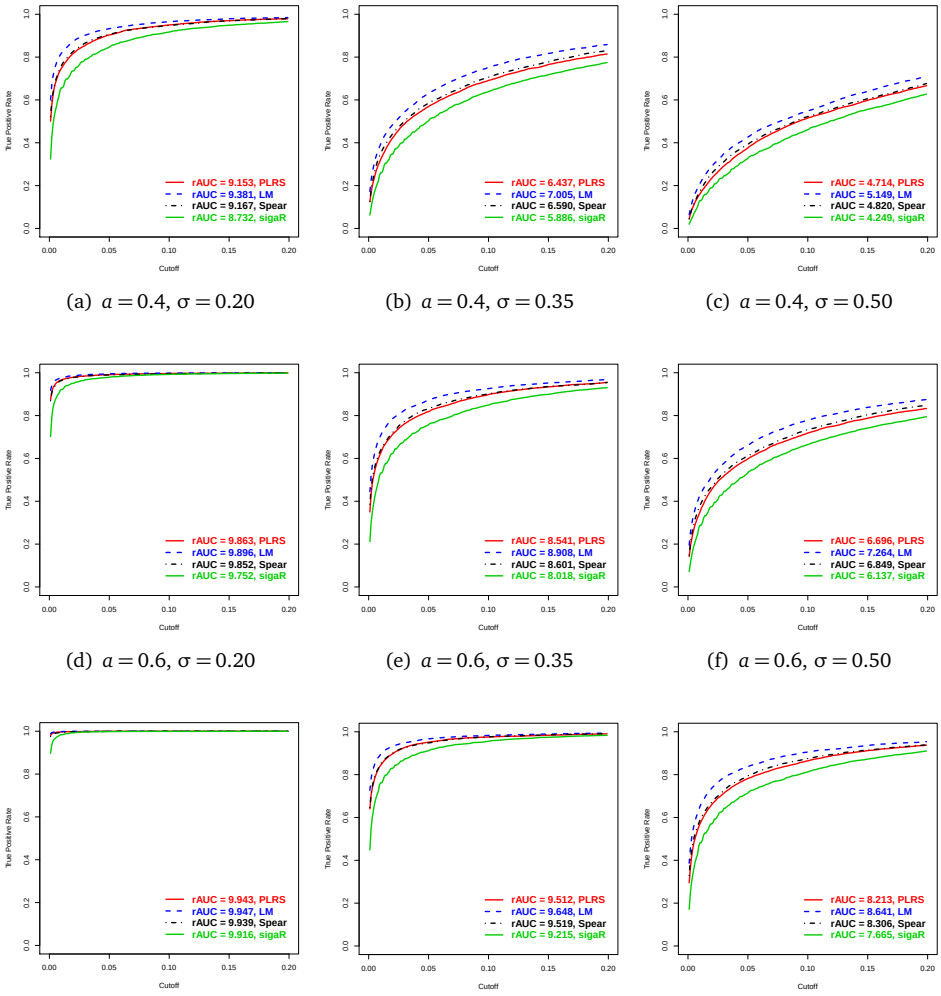


(a) PLRS (b) LM (c) Spearman (d) sigaR

FIGURE A.2. *Distribution of ranks based on $\text{rAUC}_{0.2}$ over the 90 simulation cases for each test.*

A.4.4 Partial ROC curves.

A.4.4.1 Partial ROC curves for f_{LIN}.



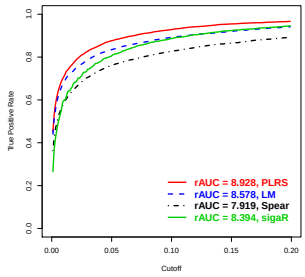
(a) $a = 0.4$, $\sigma = 0.20$ (b) $a = 0.4$, $\sigma = 0.35$ (c) $a = 0.4$, $\sigma = 0.50$

(d) $a = 0.6$, $\sigma = 0.20$ (e) $a = 0.6$, $\sigma = 0.35$ (f) $a = 0.6$, $\sigma = 0.50$

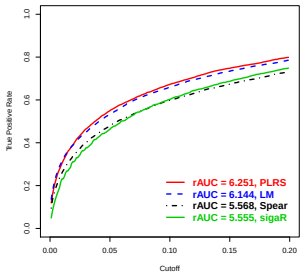
(g) $a = 0.8$, $\sigma = 0.20$ (h) $a = 0.8$, $\sigma = 0.35$ (i) $a = 0.8$, $\sigma = 0.50$

FIGURE A.3. *Partial ROC-curves when the true association is linear (f_{LIN}). For each plot, the mean true positive rate (y-axis) over the 80 genes is displayed as a function of the cut-off $c \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

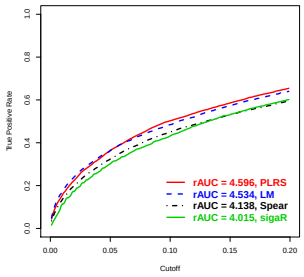
A.4.4.2 Partial ROC curves for f_{PLES}.



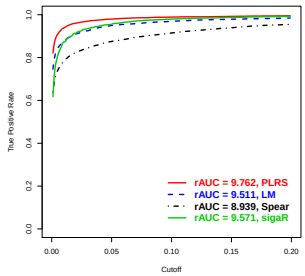
(a) $a = 0.4$, $\sigma = 0.20$



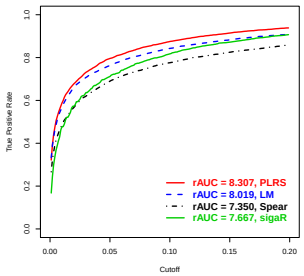
(b) $a = 0.4$, $\sigma = 0.35$



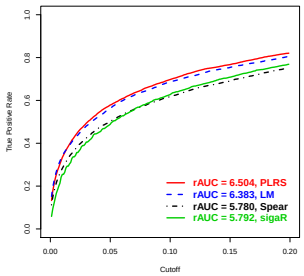
(c) $a = 0.4$, $\sigma = 0.50$



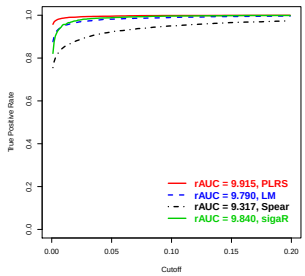
(d) $a = 0.6$, $\sigma = 0.20$



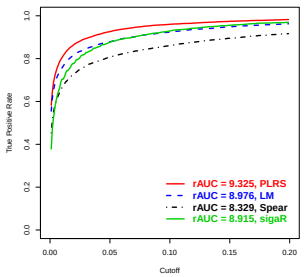
(e) $a = 0.6$, $\sigma = 0.35$



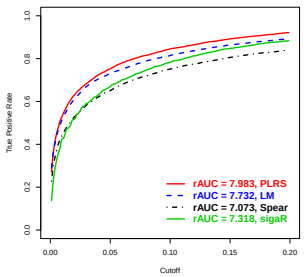
(f) $a = 0.6$, $\sigma = 0.50$



(g) $a = 0.8$, $\sigma = 0.20$



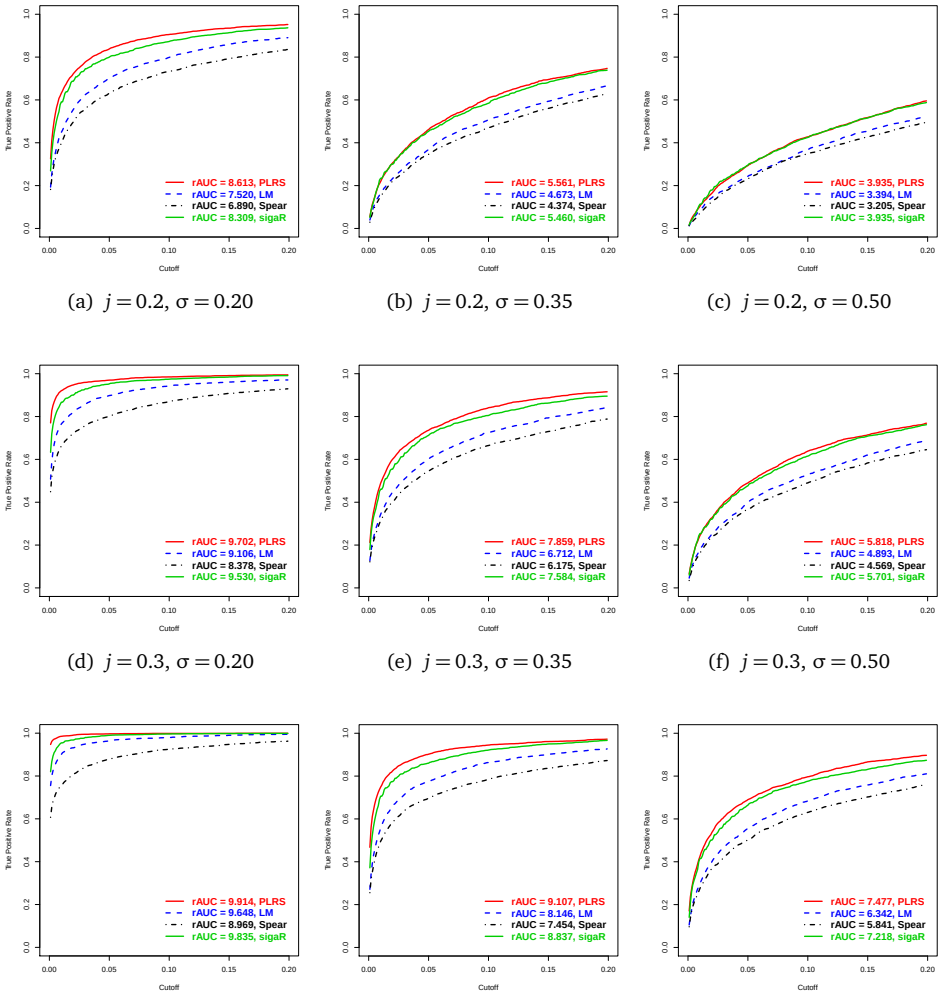
(h) $a = 0.8$, $\sigma = 0.35$



(i) $a = 0.8$, $\sigma = 0.50$

FIGURE A.4. *Partial ROC-curves when the true association is piecewise linear with equal slopes (f_{PLES}). For each plot, the mean true positive rate (y-axis) over the 80 genes is displayed as a function of the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.3 Partial ROC curves for f_{STEP2}.



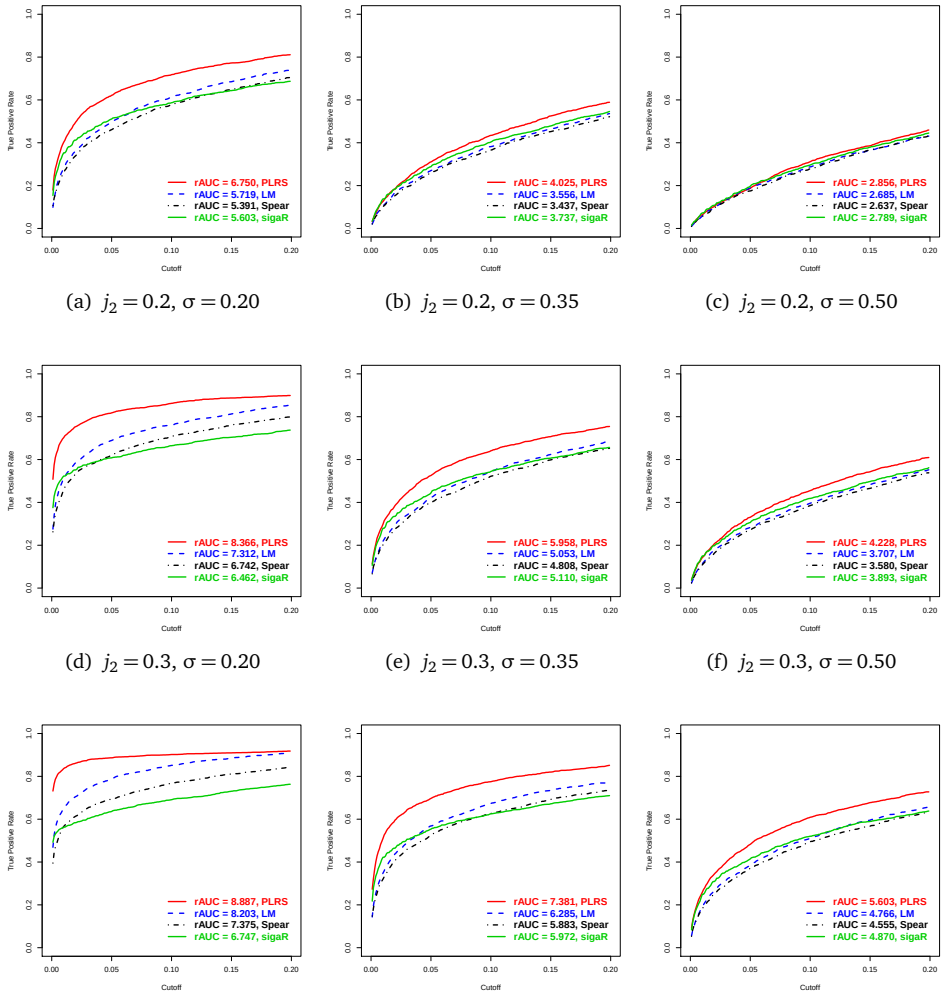
(a) $j = 0.2$, $\sigma = 0.20$ (b) $j = 0.2$, $\sigma = 0.35$ (c) $j = 0.2$, $\sigma = 0.50$

(d) $j = 0.3$, $\sigma = 0.20$ (e) $j = 0.3$, $\sigma = 0.35$ (f) $j = 0.3$, $\sigma = 0.50$

(g) $j = 0.4$, $\sigma = 0.20$ (h) $j = 0.4$, $\sigma = 0.35$ (i) $j = 0.4$, $\sigma = 0.50$

FIGURE A.5. *Partial ROC-curves when the true association is stepwise with two states (f_{STEP2}). For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.4 Partial ROC curves for f_{STEP3} when $j_1 = 0$.



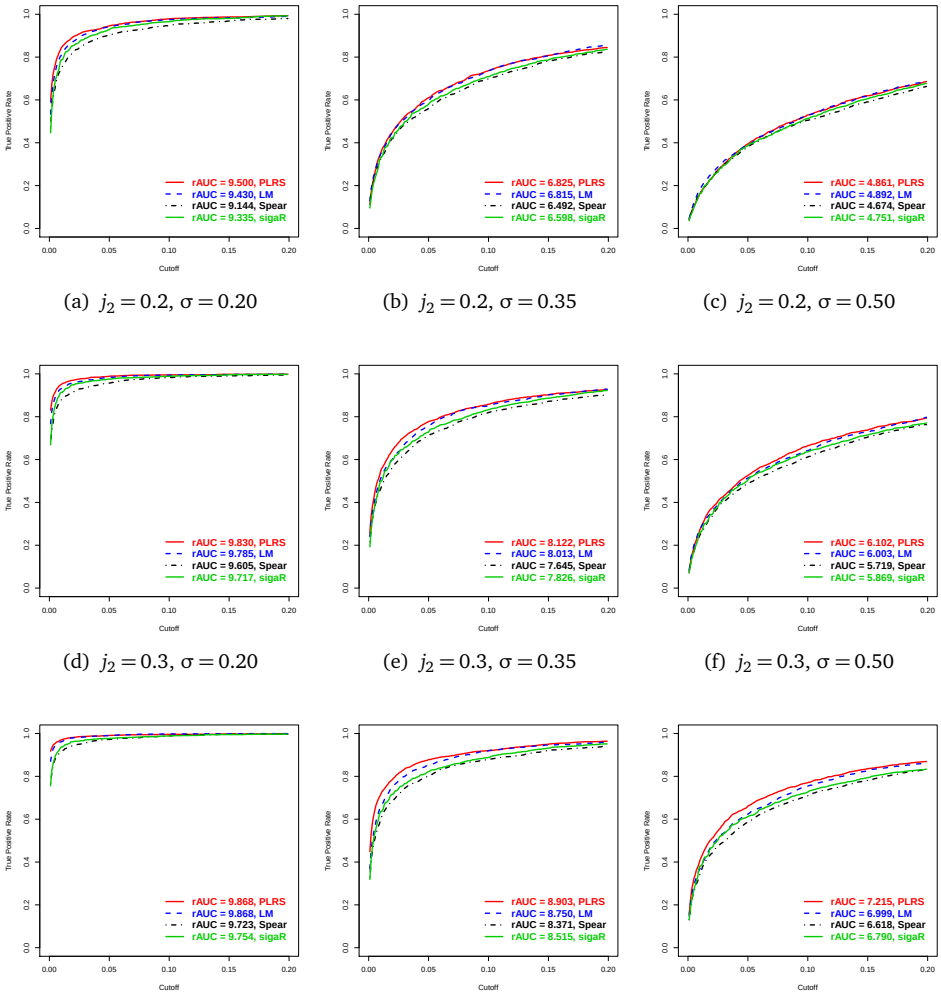
(a) $j_2 = 0.2$, $\sigma = 0.20$ (b) $j_2 = 0.2$, $\sigma = 0.35$ (c) $j_2 = 0.2$, $\sigma = 0.50$

(d) $j_2 = 0.3$, $\sigma = 0.20$ (e) $j_2 = 0.3$, $\sigma = 0.35$ (f) $j_2 = 0.3$, $\sigma = 0.50$

(g) $j_2 = 0.4$, $\sigma = 0.20$ (h) $j_2 = 0.4$, $\sigma = 0.35$ (i) $j_2 = 0.4$, $\sigma = 0.50$

FIGURE A.6. *Partial ROC-curves when the true association is stepwise with three states (f_{STEP3}) and $j_1 = 0$. For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.5 Partial ROC curves for f_{STEP3} when $j_1 = 0.2$.



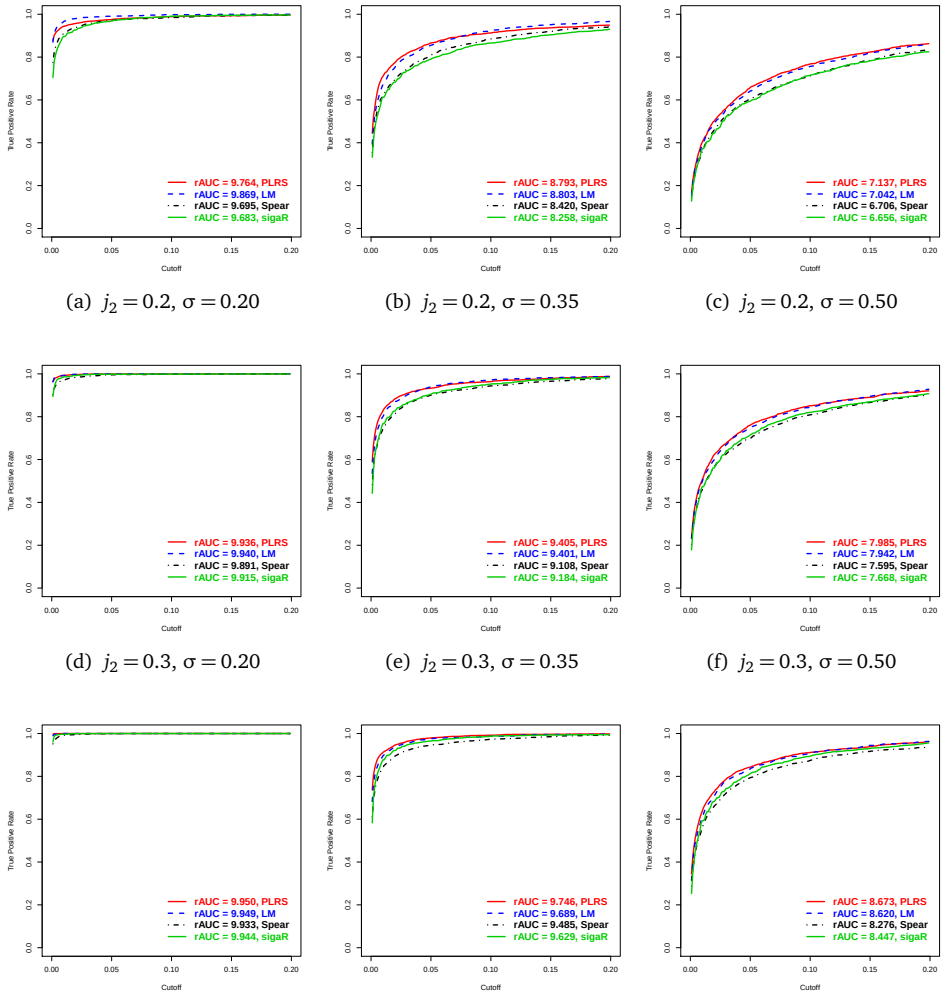
(a) $j_2 = 0.2$, $\sigma = 0.20$ (b) $j_2 = 0.2$, $\sigma = 0.35$ (c) $j_2 = 0.2$, $\sigma = 0.50$

(d) $j_2 = 0.3$, $\sigma = 0.20$ (e) $j_2 = 0.3$, $\sigma = 0.35$ (f) $j_2 = 0.3$, $\sigma = 0.50$

(g) $j_2 = 0.4$, $\sigma = 0.20$ (h) $j_2 = 0.4$, $\sigma = 0.35$ (i) $j_2 = 0.4$, $\sigma = 0.50$

FIGURE A.7. *Partial ROC-curves when the true association is stepwise with three (f_{STEP3}) and $j_1 = 0.2$. For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.6 Partial ROC curves for f_{STEP3} when $j_1 = 0.4$.



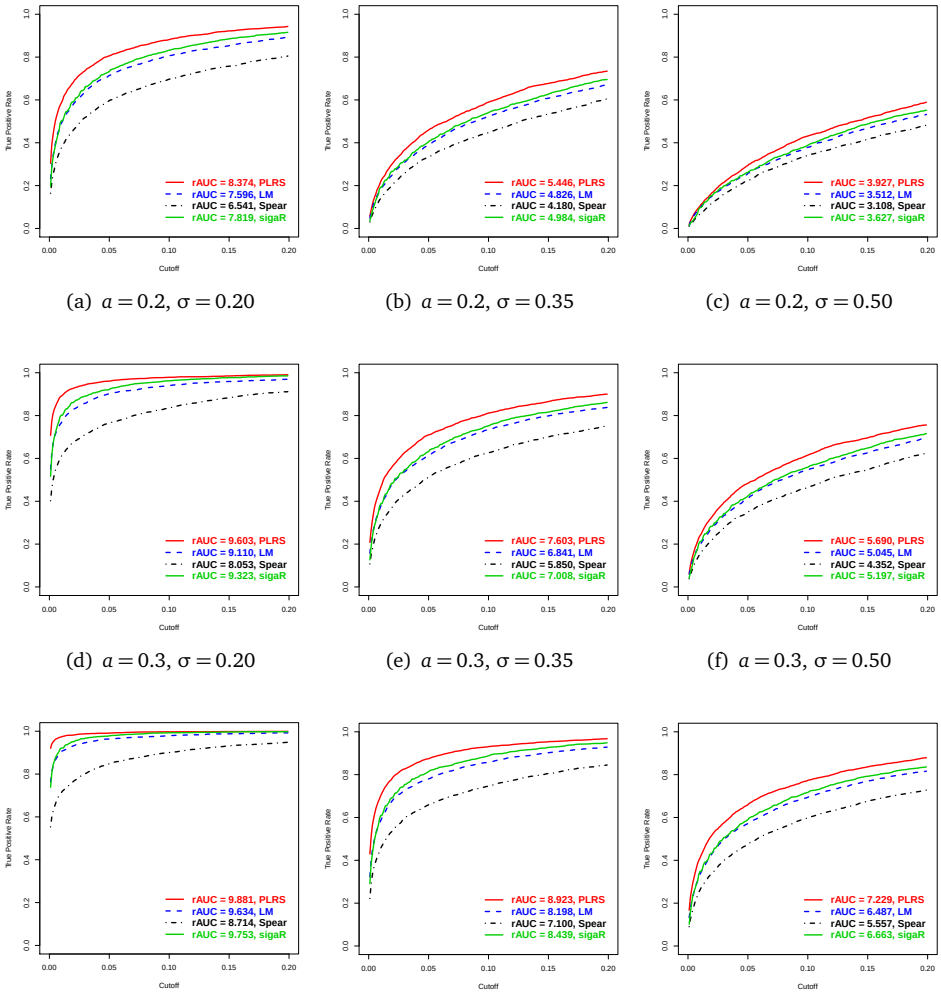
(a) $j_2 = 0.2$, $\sigma = 0.20$ (b) $j_2 = 0.2$, $\sigma = 0.35$ (c) $j_2 = 0.2$, $\sigma = 0.50$

(d) $j_2 = 0.3$, $\sigma = 0.20$ (e) $j_2 = 0.3$, $\sigma = 0.35$ (f) $j_2 = 0.3$, $\sigma = 0.50$

(g) $j_2 = 0.4$, $\sigma = 0.20$ (h) $j_2 = 0.4$, $\sigma = 0.35$ (i) $j_2 = 0.4$, $\sigma = 0.50$

FIGURE A.8. *Partial ROC-curves when the true association is stepwise with three states (f_{STEP3}) and $j_1 = 0.4$. For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.7 Partial ROC curves for f_{PLUS2}.



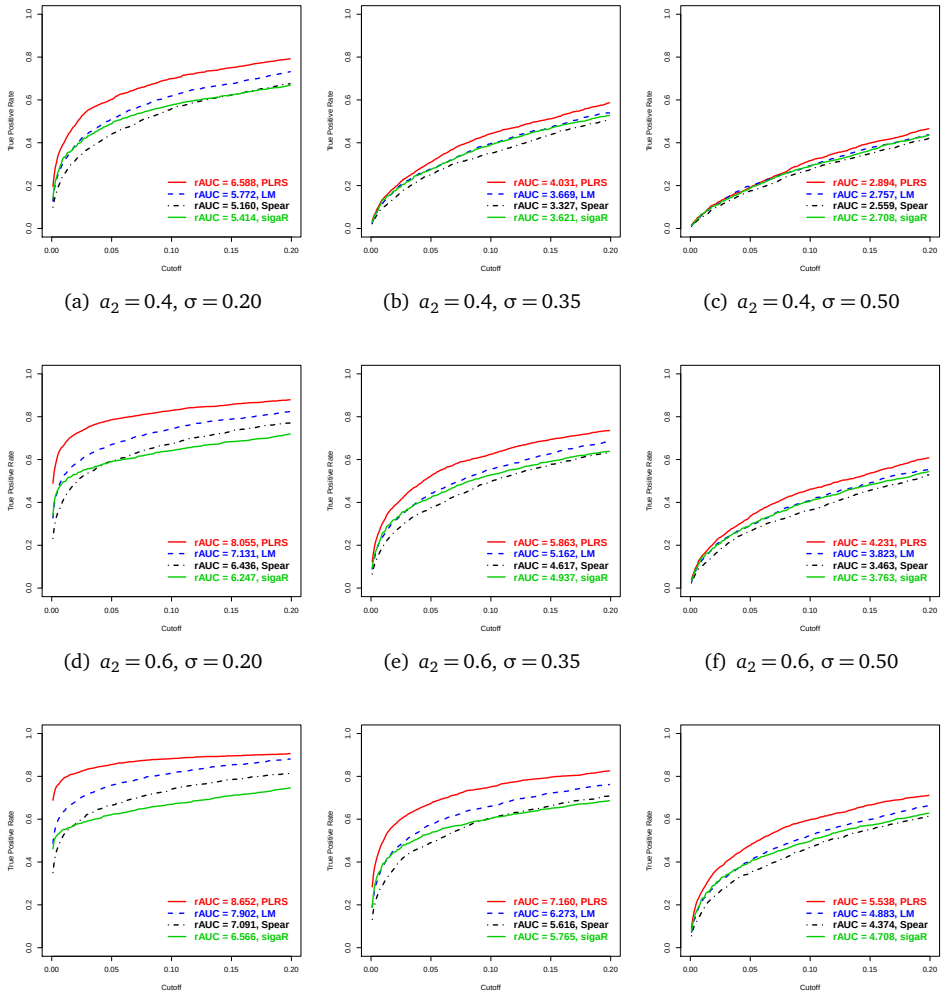
(a) $a = 0.2$, $\sigma = 0.20$ (b) $a = 0.2$, $\sigma = 0.35$ (c) $a = 0.2$, $\sigma = 0.50$

(d) $a = 0.3$, $\sigma = 0.20$ (e) $a = 0.3$, $\sigma = 0.35$ (f) $a = 0.3$, $\sigma = 0.50$

(g) $a = 0.4$, $\sigma = 0.20$ (h) $a = 0.4$, $\sigma = 0.35$ (i) $a = 0.4$, $\sigma = 0.50$

FIGURE A.9. *Partial ROC-curves when the true association is piecewise linear with unequal slopes and with two states (f_{PLUS2}). For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.8 Partial ROC curves for f_{PLUS3} when $a_1 = 0$.



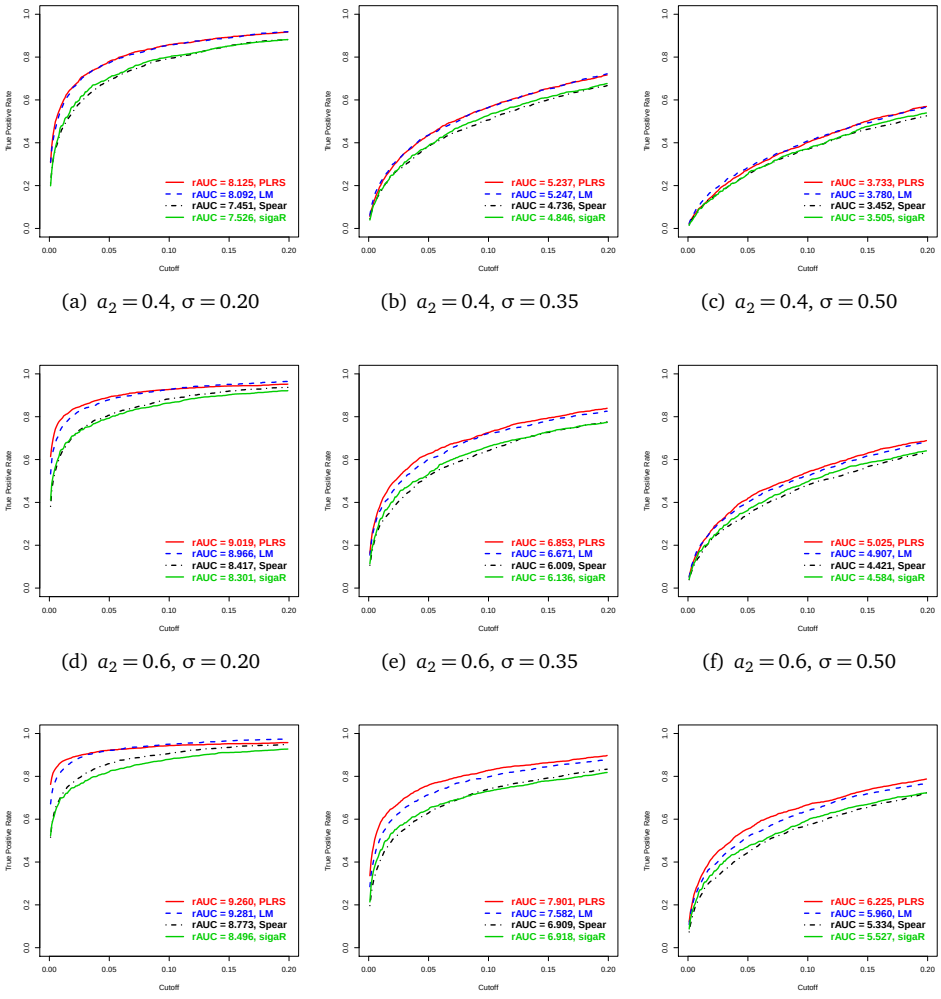
(a) $a_2 = 0.4$, $\sigma = 0.20$ (b) $a_2 = 0.4$, $\sigma = 0.35$ (c) $a_2 = 0.4$, $\sigma = 0.50$

(d) $a_2 = 0.6$, $\sigma = 0.20$ (e) $a_2 = 0.6$, $\sigma = 0.35$ (f) $a_2 = 0.6$, $\sigma = 0.50$

(g) $a_2 = 0.8$, $\sigma = 0.20$ (h) $a_2 = 0.8$, $\sigma = 0.35$ (i) $a_2 = 0.8$, $\sigma = 0.50$

FIGURE A.10. *Partial ROC-curves when the true association is piecewise linear with unequal slopes, with three states (f_{PLUS3}) and $a_1 = 0$. For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.9 Partial ROC curves for f_{PLUS3} when $a_1 = 0.2$.



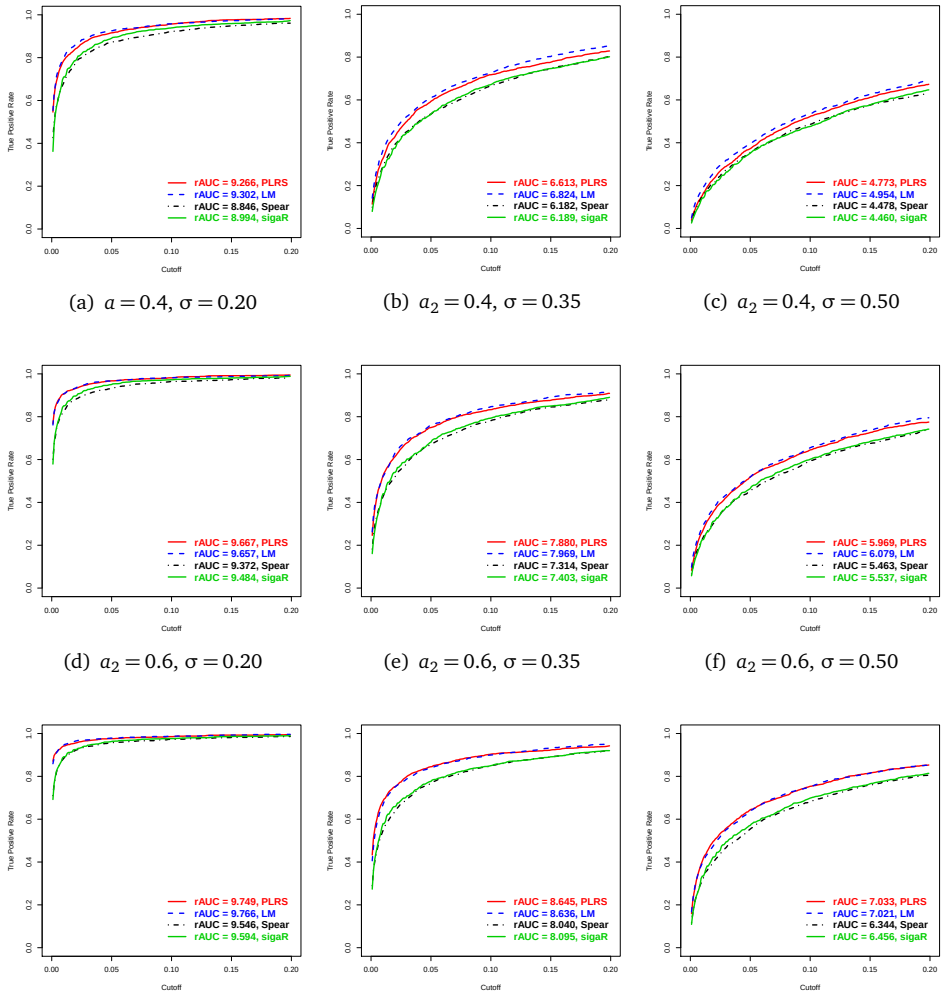
(a) $a_2 = 0.4$, $\sigma = 0.20$ (b) $a_2 = 0.4$, $\sigma = 0.35$ (c) $a_2 = 0.4$, $\sigma = 0.50$

(d) $a_2 = 0.6$, $\sigma = 0.20$ (e) $a_2 = 0.6$, $\sigma = 0.35$ (f) $a_2 = 0.6$, $\sigma = 0.50$

(g) $a_2 = 0.8$, $\sigma = 0.20$ (h) $a_2 = 0.8$, $\sigma = 0.35$ (i) $a_2 = 0.8$, $\sigma = 0.50$

FIGURE A.11. *Partial ROC-curves when the true association is piecewise linear with unequal slopes, with three states (f_{PLUS3}) and $a_1 = 0.2$. For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

A.4.4.10 Partial ROC curves for f_{PLUS3} when $a_1 = 0.4$.



(a) $a = 0.4$, $\sigma = 0.20$ (b) $a_2 = 0.4$, $\sigma = 0.35$ (c) $a_2 = 0.4$, $\sigma = 0.50$

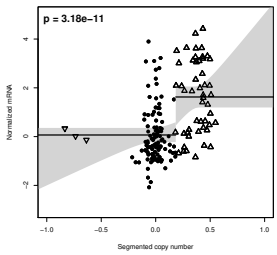
(d) $a_2 = 0.6$, $\sigma = 0.20$ (e) $a_2 = 0.6$, $\sigma = 0.35$ (f) $a_2 = 0.6$, $\sigma = 0.50$

(g) $a_2 = 0.8$, $\sigma = 0.20$ (h) $a_2 = 0.8$, $\sigma = 0.35$ (i) $a_2 = 0.8$, $\sigma = 0.50$

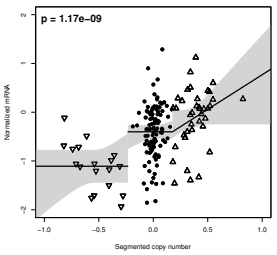
FIGURE A.12. *Partial ROC-curves when the true association is piecewise linear with unequal slopes, with three states (f_{PLUS3}) and $a_1 = 0.4$. For each plot, the mean true positive rate (y-axis) over the 40 genes is displayed as a function a the cut-off $\alpha \leq 0.2$ (x-axis). Each plot is function of the true slope a (rows) and noise σ (columns). ROC curves of the PLRS (red), LM (blue), Spearman (black) and sigaR (green) tests are displayed.*

APPENDIX B

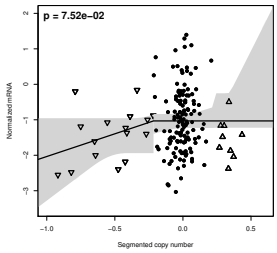
B.1 The `PLRS` screening procedure. We applied `PLRS` to the TCGA Glioblastoma data set and created a list of genes for which the association between DNA copy number and mRNA expression was found to be significant with type I error cutoff $\alpha = 0.1$ on adjusted p-values (Benjamini and Hochberg, 1995). This list was then compared with the top 300 cancer genes as provided by the Gene Ranker TCGA GBM 6000 (http://cbio.mskcc.org/tcga-generanker/). We found that 71% of the cancer candidate genes were detected by the PLRS testing procedure, hence providing evidence that for most of these genes copy number aberrations induces differential expression. Below, figures B.1 to B.3 display 16 DNA-mRNA associations for such genes. Figure B.4 reports cancer genes for which the effect of copy number aberrations on expression appears more uniform accross samples and are not detected by the test.



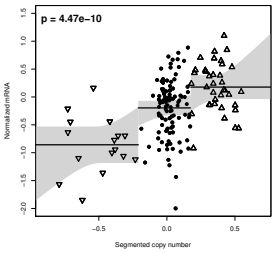
(a) GNAS



(b) PPP2R1A

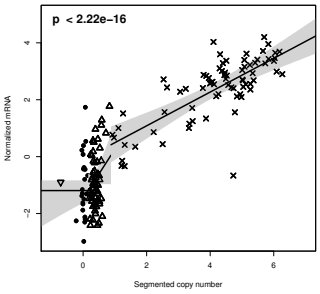


(c) PRKCZ

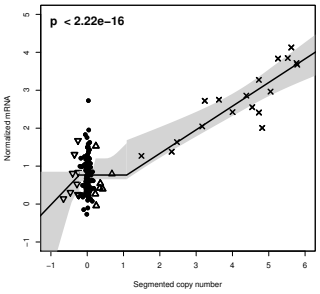


(d) BAX

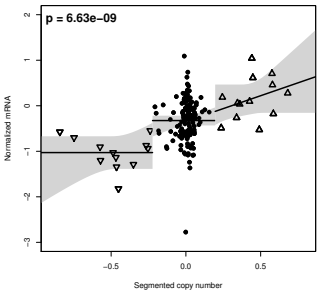
FIGURE B.1. *Association between DNA and mRNA for different genes in the TCGA Glioblastoma data set. Segmented copy number is on the x-axis while gene expression is on the y-axis. States are indicated by different symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform CBs. The top left values correspond to the p-values of the PLRS test.*



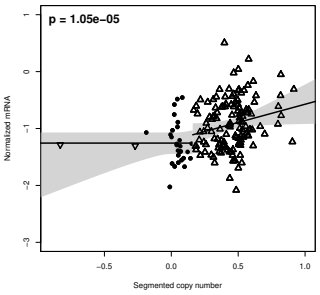
(a) EGFR



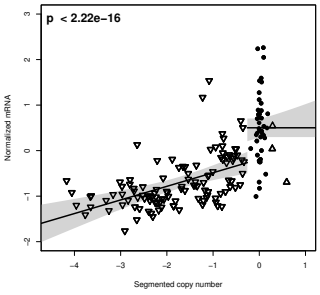
(b) MDM2



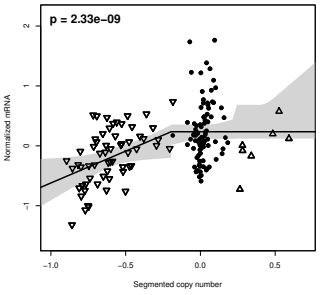
(c) KRAS



(d) BRAF

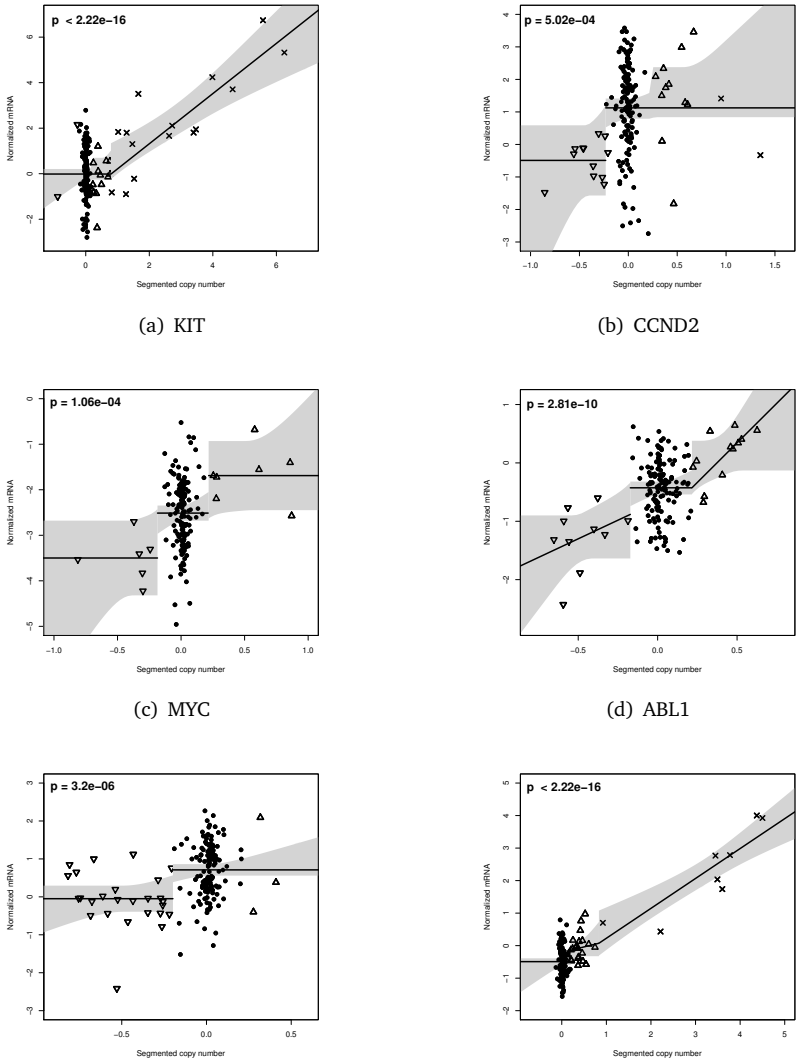


(e) CDKN2B



(f) JAK2

FIGURE B.2. *Association between DNA and mRNA for different genes in the TCGA Glioblastoma data set. Segmented copy number is on the x-axis while gene expression is on the y-axis. States are indicated by different symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform CBs. The top left values correspond to the p-values of the PLRS test.*

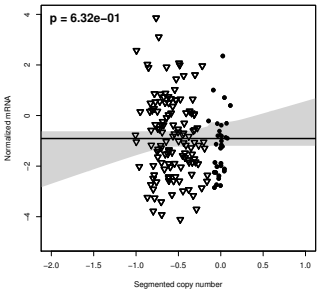


(a) KIT (b) CCND2

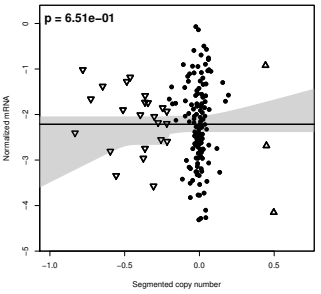
(c) MYC (d) ABL1

(e) FOXO3 (f) MDM4

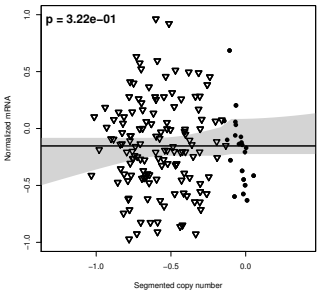
FIGURE B.3. *Association between DNA and mRNA for different genes in the TCGA Glioblastoma data set. Segmented copy number is on the x-axis while gene expression is on the y-axis. States are indicated by different symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform CBs. The top left values correspond to the p-values of the PLRS test.*



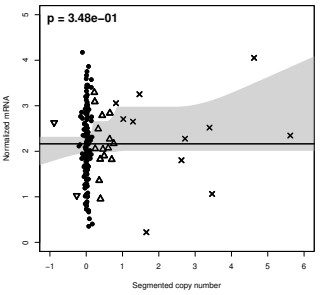
(a) RET



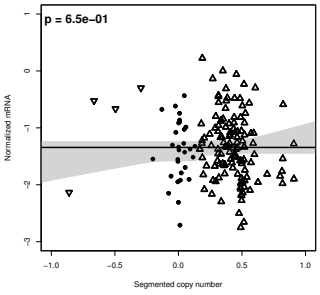
(b) CHEK1



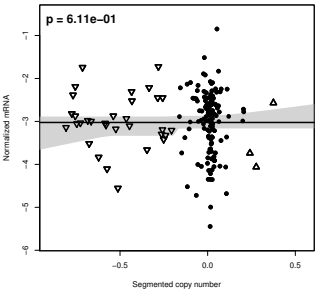
(c) FGFR2



(d) KDR



(e) EPHB6



(f) ESR1

FIGURE B.4. *Association between DNA and mRNA for different genes in the TCGA Glioblastoma data set. Segmented copy number is on the x-axis while gene expression is on the y-axis. States are indicated by different symbols: loss (▽), normal (◯), gain (△) and amplification (×). Grey surfaces correspond to 95% uniform CBs. The top left values correspond to the p-values of the PLRS test.*

APPENDIX C

C.1 Approximate equivalence of iterative procedures with maximum marginal likelihood. Given latent variables (or parameters) $\theta_1, \theta_2, \ldots, \theta_p$ let $Y_1, \ldots, Y_p$ be independent observations such that the marginal distribution of Y_i depends on θ_i only. Hence, first assume a situation where the interest is in one prior only, parametrized with hyper-parameters $\boldsymbol{\alpha}$. Now allow the density of Y_i given $\theta_i = \theta$ to depend on i, but drop the index for clarity reasons: $y \mapsto f(y|\theta_i = \theta) = f_i(y|\theta_i = \theta)$ (relative to some measure μ). Furthermore, suppose that $\theta_1, \ldots, \theta_p$ are independent. Then $Y_1, \ldots, Y_p$ are (unconditionally) independent, and Y_i has density

$$y \mapsto f_{\boldsymbol{\alpha}}(y) = \int f(y|\theta)\, d\Pi_{\boldsymbol{\alpha}}(\theta).$$

Furthermore, the vectors $(\theta_1, Y_1), \ldots, (\theta_p, Y_p)$ are independent and the marginal distribution $\Pi_{\boldsymbol{\alpha}}$ of θ_i can be disintegrated as

$$\Pi_{\boldsymbol{\alpha}}(\theta) = \int \Pi_{\boldsymbol{\alpha}}(\theta|y) f_{\boldsymbol{\alpha}}(y)\, d\mu(y),$$

for $\Pi_{\boldsymbol{\alpha}}(\cdot|y)$: the conditional distribution of θ_i given $Y_i = y$.

The right side is the expectation of $\Pi_{\boldsymbol{\alpha}}(\theta|Y_i)$ relative to Y_i, for every i, and hence the average of $\Pi_{\boldsymbol{\alpha}}(\theta|Y_i)$ over the data is a reasonable estimate of $\Pi_{\boldsymbol{\alpha}}(\theta)$. In other words, we expect

$$\Pi_{\boldsymbol{\alpha}}(\cdot) \approx \frac{1}{p} \sum_{i=1}^{p} \Pi_{\boldsymbol{\alpha}}(\cdot|Y_i). \tag{C.1}$$

Next we may estimate the unknown $\boldsymbol{\alpha}$ by the value for which this approximation is most accurate, which is what our iterative procedures attempt. An alternative is conventional empirical Bayes, which consists of maximizing the marginal likelihood:

$$\boldsymbol{\alpha} \mapsto \prod_{i=1}^{p} f_{\boldsymbol{\alpha}}(Y_i). \tag{C.2}$$

The equations (C.1) and (C.2) turn out to be approximately equivalent, depending on how $\approx$ in (C.1) is made precise. To see this rewrite (C.1) in terms of densities as

$$\pi_{\boldsymbol{\alpha}}(\theta) \approx \frac{1}{p} \sum_{i=1}^{p} \pi_{\boldsymbol{\alpha}}(\theta|Y_i) = \frac{1}{p} \sum_{i=1}^{p} \frac{f(Y_i|\theta)\pi_{\boldsymbol{\alpha}}(\theta)}{f_{\boldsymbol{\alpha}}(Y_i)}.$$

Cancel $\pi_{\boldsymbol{\alpha}}$ left and right to rewrite this as

$$1 \approx \frac{1}{p} \sum_{i=1}^{p} \frac{f(Y_i|\theta)}{f_{\boldsymbol{\alpha}}(Y_i)}. \tag{C.3}$$

On the other hand, in view of (C.2) the empirical Bayes estimator for $\boldsymbol{\alpha}$ solves the likelihood equation

$$0 = \frac{1}{p}\sum_{i=1}^{p} \frac{\dot{f}_{\boldsymbol{\alpha}}(\mathrm{Y}_i)}{f_{\boldsymbol{\alpha}}(\mathrm{Y}_i)} = \int \frac{1}{p}\sum_{i=1}^{p} \frac{f(\mathrm{Y}_i|\theta)}{f_{\boldsymbol{\alpha}}(\mathrm{Y}_i)}\, \dot{\pi}_{\boldsymbol{\alpha}}(\theta)\, d\theta. \tag{C.4}$$

If (C.3) would hold exactly and identically in θ, then the right side would vanish, because $\int \dot{\pi}_{\boldsymbol{\alpha}}(\theta)\, d\theta = 0$. Thus we can view (C.3), and hence (C.1), as a method to find an approximate solution to the likelihood equation.

The above equations do not change much in the multi-prior situation, assuming independent priors. Denote the collection of all unknown $\boldsymbol{\alpha}$ by A, so each element of A corresponds to one (type of) parameter (e.g. one would correspond to the prior of regression parameter $\beta_{i,1}$ and another to that of overdispersion parameter ϕ_i). In the equations above, the posteriors $\pi_{\boldsymbol{\alpha}}(\theta|\mathrm{Y}_i)$ should be replaced by $\pi_{\mathrm{A}}(\theta|\mathrm{Y}_i)$, the marginal likelihoods $f_{\boldsymbol{\alpha}}(\mathrm{Y}_i)$ by $f_{\mathrm{A}}(\mathrm{Y}_i)$ and the conditional likelihoods $f(\mathrm{Y}_i|\theta)$ by $f_{\mathrm{A}\setminus\boldsymbol{\alpha}}(\mathrm{Y}_i|\theta)$. Furthermore, (C.4) becomes a set of likelihood equations, for which the right sides also vanish when (C.3) (with the above replacements) holds.

C.2 Proof of marginal likelihood maximization. A computationally attractive alternative is the following. Let $\pi'_{\boldsymbol{\alpha}'}(\theta)$ be an arbitrary parametric prior with hyper-parameters $\boldsymbol{\alpha}'$ and let $f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i) = f_{\boldsymbol{\alpha}',\mathrm{A}^*_{-b}}(\mathrm{Y}_i)$ be the marginal likelihood given the prior for θ_i and the hyper-parameters of the other priors. Finally, let $f_{\boldsymbol{\alpha}'}(\mathrm{Y}) = \prod_{i=1}^{p} f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i)$ be the product marginal likelihood, then $f_{\boldsymbol{\alpha}'}(\mathrm{Y})$ is maximized for

$$\tilde{\boldsymbol{\alpha}} = \mathrm{argmax}_{\boldsymbol{\alpha}'} \sum_{i=1}^{p} \log\left[\int \pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i)\frac{\pi'_{\boldsymbol{\alpha}'}(\theta)}{\pi_{\boldsymbol{\alpha}^*_b}(\theta)}\, d\theta\right]. \tag{C.5}$$

Proof

Let $\pi_{\boldsymbol{\alpha}^*_b}(\theta), \pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i), f(\mathrm{Y}_i|\theta) = f_{\mathrm{A}^*}(\mathrm{Y}_i|\theta)$ and $f(\mathrm{Y}_i) = f_{\mathrm{A}^*}(\mathrm{Y}_i)$ respectively be the prior, marginal posterior, conditional likelihood and marginal likelihood as used in or resulting from the iterative joint procedure with hyper-parameters A^*, where we assume the bth component of A^* to correspond with the common prior of $\theta_i, i = 1, \ldots, p$. Likewise, a new parametric prior for θ_i that replaces $\pi_{\boldsymbol{\alpha}^*_b}(\theta)$ is denoted by $\pi'_{\boldsymbol{\alpha}'}(\theta)$. Finally, let $f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i|\theta) = f_{\boldsymbol{\alpha}',\mathrm{A}^*_{-b}}(\mathrm{Y}_i|\theta)$ be the conditional likelihood, resulting from applying $\pi'_{\boldsymbol{\alpha}'}(\theta)$ instead of $\pi_{\boldsymbol{\alpha}^*}(\theta)$ while keeping the priors for the other parameters unchanged.

Then,

$$f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i) = \int f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i|\theta)\pi'_{\boldsymbol{\alpha}'}(\theta)\, d\theta = \int f(\mathrm{Y}_i|\theta)\pi'_{\boldsymbol{\alpha}'}(\theta)\, d\theta,$$

because $f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i|\theta)$ does not depend $\boldsymbol{\alpha}'$. Apply Bayes' rule to obtain

$$f_{\boldsymbol{\alpha}'}(\mathrm{Y}_i) = f(\mathrm{Y}_i)\int \pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i)\frac{\pi'_{\boldsymbol{\alpha}'}(\theta)}{\pi_{\boldsymbol{\alpha}^*_b}(\theta)}\, d\theta.$$

Then $f_{\boldsymbol{\alpha}'}(\mathrm{Y})$ is maximized at

$$\tilde{\boldsymbol{\alpha}} = \mathrm{argmax}_{\boldsymbol{\alpha}'} \sum_{i=1}^{p} \log \left[\int \pi_{\mathrm{A}^*}(\theta|\mathrm{Y}_i) \frac{\pi'_{\boldsymbol{\alpha}'}(\theta)}{\pi_{\boldsymbol{\alpha}^*_b}(\theta)} d\theta \right].$$

While the iterative marginal procedure easily applies to any parametric and non-parametric prior, the direct maximization procedure requires a dedicated maximization procedure for each type of prior. However, in particular for priors with few hyper-parameters it is computationally superior. Note that (C.5) also provides estimation of the total marginal log-likelihood for any prior.

C.3 Computational efficiency and convergence. The number of features, p, can be enormous. This might hamper practical application of the iterative procedures, in particular the iterative joint one: if the algorithm would be applied to all features and would require L iterations, the total number of INLA calls is proportional to Lp. Fortunately, the following heuristic decreases the required total number of INLA calls tremendously, and is therefore computationally much more efficient: initialize $t = 1$ and $p^{(1)} <<< p$, randomly select $p^{(t)}$ features, run the above algorithm and use the final $\boldsymbol{\alpha}$-estimate as an initial value for a new loop using $p^{(t+1)} > p^{(t)}$ features. We propose to use the trajectory $\mathrm{p} = (p^{(1)}, \ldots, p^{(\mathrm{T})}) = (100, 200, 500, 1000, 2000, 5000)$. In fact, we monitor the convergence in $p^{(t)}$ using (C.6). In practice, we observe the algorithms usually stops at $p^{(t)} \approx 2000$. This implies that for p very large (say in the order 10^5 to 10^6) the computational cost for Bayesian shrinkage by estimating the prior is relatively low with respect to that of the final fit to all features.

Our iterative methods require to monitor convergence of the estimates of the priors. We propose to do so by considering a Kolmogorov-Smirnov-type metric:

$$\mathrm{KS}_{\boldsymbol{\alpha},\ell} = \max_x |\mathrm{F}_{\boldsymbol{\alpha},\ell}(x) - \mathrm{F}_{\boldsymbol{\alpha},\ell-1}(x)|, \tag{C.6}$$

where $\mathrm{F}_{\boldsymbol{\alpha},\ell}$ is the distribution function corresponding to the prior of $\boldsymbol{\alpha}$ fitted at iteration ℓ. Then the algorithm stops when $\mathrm{KS}_{\boldsymbol{\alpha},\ell} \leq \nu$, where ν is a user-defined threshold, e.g. $\nu = 0.005$. If multiple parameters are shrunken, the latter inequality should hold for each fitted prior. Monitoring changes in hyper-parameters directly may be less suitable, because the impact of a change in hyper-parameters may depend on other hyper-parameters. E.g. in a Gaussian mixture changes of the mean and standard deviation hyper-parameters for a given component have very little impact on the shape of the distribution when the corresponding mixture proportion is very small.

For the iterative marginal procedure, we also monitor the marginal likelihood, because we know our iterative procedures are approximately equivalent to maximal marginal likelihood (Appendix C.1). The marginal likelihood should initially increase while iterating, then level off and 'wiggle' around the maximum when converging. The algorithm is stopped when k (e.g. $k = 2$) consecutive estimates of the marginal likelihood are smaller than the maximum observed so far.

C.4 Priors for random effects. Our method accommodates inference for and shrinkage of random effects. Suppose we have $\beta_{ik} =^d \mathrm{N}(0, \tau_i^2)$ for $k = 1, \ldots, \mathrm{K}$. Then, shrinkage focuses on τ_i^2, or equivalently on precision, τ_i^{-2}. The conjugate prior for precision is the Gamma distribution, $\Gamma(\alpha_1, \alpha_2)$, where α_1 and α_2 are the shape and rate hyper-parameters. We also allow for fitting a nonparametric prior, which is preferable when inference is desired using an interval null-hypothesis. However, the parametric option is particularly useful when one is not interested in inference for τ_i^2 (as is often the case in studies with few replicates), but would still like to shrink τ_i^2 in combination with other (e.g. fixed or dispersion) parameters using the iterative joint procedure. When the prior of τ_i^2 is concentrated, shrinkage can be beneficial for better disentanglement of parameter effects in non-balanced designs. Moreover, a concentrated prior close to zero effectively acts as intrinsic model selection, which may render more effective inference for the other parameters.

C.5 BFDR and lfdr for two-sided inference and multiple comparisons. Below we detail the modifications of lfdr and BFDR, as introduced in Section 4.4.2, for two-sided inference and multiple comparisons.

Two-sided lfdr and BFDR. For two-sided testing we may reformulate Equation 4.5 as $\mathrm{H}_{0i}^{\mathrm{II}} : |\beta_i| \leq \Delta$. Directly applying the above definitions for lfdr and BFDR to $\mathrm{H}_i^{\mathrm{II}}$ may, however, lead to counterintuitive results when posteriors are wide: π_{0i} may be small due to non-negligible posterior masses on both $\beta_i = \beta^- < -\Delta$ and $\beta_i = \beta^+ > \Delta$, in particular for $\Delta \approx 0$. One prefers not to select such cases as being 'significant'. Therefore, we adjust the procedure as follows. For lfdr, we simply define lfdr_i^- using $\mathrm{H}_{0i}^- = \mathrm{H}_{0i}$ as in Equation 4.5, and analogously lfdr_i^+ using $\mathrm{H}_{0i}^+ : \beta_i \geq -\Delta$ instead. Then, define the two-sided version:

$$\mathrm{lfdr}_i^{\mathrm{II}} := \min(\mathrm{lfdr}_i^-, \mathrm{lfdr}_i^+) \geq \mathrm{P}(|\beta_i| \leq \Delta | \mathrm{Y}_i) = \mathrm{P}(\mathrm{H}_{0i}^{\mathrm{II}} | \mathrm{Y}_i).$$

So, $\mathrm{lfdr}_i^{\mathrm{II}} \leq \alpha$ provides interpretability while also guaranteeing $\mathrm{P}(\mathrm{H}_{0i}^{\mathrm{II}} | \mathrm{Y}_i) \leq \alpha$.

To define a two-sided version of $\mathrm{BFDR}(t)$ we simply use the aforementioned correspondence between lfdr and BFDR, the latter being a conditional mean of the first. Moreover, let $d_i^{\mathrm{II}}(t) = \max(d_i^-(t), d_i^+(t))$ where definitions of $d_i^+(t)$ and $d_i^-(t)$ (and also π_{0i}^+ and π_{0i}^-) are analogous to those of $d_i(t)$ (π_{0i}), replacing H_i by H_i^+ or H_i^-, respectively. Then, analogous to Equation 4.5, we have

$$\mathrm{BFDR}^{\mathrm{II}}(t) = \mathrm{E}[\mathrm{lfdr}_i^{\mathrm{II}} | \mathrm{lfdr}_i^{\mathrm{II}} < t] = \frac{\sum_{i=1}^p \mathrm{lfdr}_i^{\mathrm{II}} d_i^{\mathrm{II}}(t)}{\sum_{i=1}^p d_i^{\mathrm{II}}(t)}. \tag{C.7}$$

$\mathrm{BFDR}^{\mathrm{II}}(t)$ is interpreted like $\mathrm{BFDR}(t)$ and avoids unwanted detections due to wide posteriors.

Multiple comparisons. In some studies, more than two groups are to be compared with each other. Assume w.l.o.g. that $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iL})$ denote the parameters in the regression model corresponding to the L groups. Methods that approximate marginal posteriors, like INLA, generally do not return joint posterior intervals of (functions of) parameters, which restricts the use of omnibus L-group comparisons. INLA does provide an approximation of the marginal likelihood, so for each feature i a Bayes' factor can be computed for the full model versus the model with $\boldsymbol{\beta}_i = \mathbf{0}$. This can be useful as a ranking criterion. However, it does not render an inference statement. Also, in our experience, results from omnibus comparisons often immediately lead to the next question about the relevant pair-wise differences. Then, in a lfdr paradigm those pairwise differences are the relevant discoveries. The INLA software is able to provide posteriors of linear combinations. Hence, we focus on inference for pair-wise differences by computing the marginal null-probabilities $\pi_{0ik\ell} = P(\beta'_{ik\ell} \leq \Delta | Y_i)$ for each pair-wise difference:

$$\beta'_{ik\ell} = \beta_{ik} - \beta_{i\ell}, \tag{C.8}$$

where $k \neq \ell, k = 1, \ldots, L$. In a one versus many comparison situation, the control group ℓ should be fixed to L. The choice of priors for $\beta'_{ik\ell}$ is the same as for β_i in the two-group situation. One may opt to use the same prior for all pairs (k, ℓ), or fit a different one for each pair.

In case one does desire a summary per data row (feature), we provide the following bound for lfdr. The null-hypothesis for all pairwise comparisons is: $H^{\cup}_{0i}$: $\max_{k<\ell} |\beta'_{ik\ell}| \leq \Delta$. Denote the vector containing all absolute differences by $|\vec{\beta}'_i| = (|\beta'_{ik\ell}|)_{k<\ell}$. Similarly, define the vector of all contrasts $\vec{\beta}'_i = (\beta'_{ik\ell})_{k\neq\ell}$. Analogous definitions apply to $\vec{\pi}_{0i}$ and $\vec{1}$. Then, we have

$$\begin{aligned} \text{lfdr}^{\cup}_i = P(H^{\cup}_{0i} | Y_i) &= P(\max(|\vec{\beta}'_i|) \leq \Delta | Y_i) = 1 - P(\max(\vec{\beta}'_i) > \Delta | Y_i) \\ &\leq 1 - \max(\vec{1} - \vec{\pi}_{0i}) = \min(\vec{\pi}_{0i}), \end{aligned}$$

which is just the minimum over all pairwise lfdr's. Finally, define a discovery on the feature-level as any feature for which at least one contrast does not obey the null and a false discovery as any feature for which all nulls of the rejected contrasts are true. Then, extension of the BFDR to this context, is analogous to $\text{BFDR}^{\text{II}}(t)$: $\text{BFDR}^{\cup}(t) = E[\text{lfdr}^{\cup}_i | \text{lfdr}^{\cup}_i \leq t]$, which can be interpreted as the mean ratio of total posterior mass on the joint null (summed over the features detected at threshold t) and the number of detections at threshold t. Substitution of the above upper bound for $\text{lfdr}^{\cup}_i$ provides an upper bound for $\text{BFDR}^{\cup}(t)$.

C.6 Monotonic time trends.

Time-course experiments enable one to determine trends over time. These can be performed on related or non-related samples. From a design perspective, the first is often preferable over the latter, but sometimes not practical (e.g. when animals have to be sacrificed to retrieve a sample). In this section, we discuss how easily our method and software enable efficient analysis of time-course data.

First of all, when individuals correspond to multiple measurements over time, this is trivially accommodated by our software by including a random effect on the individual level, as we did in our data example in Section 4.7.3. Other software for analysing RNAseq data does not easily allow for such inclusion.

Second, while (unstructured) multi-group analysis can be applied to detect overall time effects, researchers are often interested in monotonic effects. Of course, unstructured multi-group analysis is suboptimal for such cases. Here, we propose a simple solution. More complex solutions are feasible in the context of INLA, but would usually require changing the regression model. Define a contrast of parameters that is targeted to detect such trends:

$$C_i = \sum_{j<k} (\beta_{ik} - \beta_{ij}), \tag{C.9}$$

where β_{ij} is the parameter for feature i and time point j. Alternatively, pairwise comparisons can be weighted unequally depending on expected effects and spacing over time. Our approach allows for shrinking C_i, after which the posterior of C_i is used for inference. Note that inference for contrasts is also available in edgeR (Robinson et al., 2010), but not in most other methods.

C.7 Inclusion of a mixture prior in the iterative joint procedure. When multiple mixture priors are desired, or a combination of a mixture prior and a nonparametric prior, only one of these is accommodated by the marginal refinement procedures. Below we assume a situation as in simulation Case 1 (Section C.9), where the prior of the regression parameter is estimated by one of the two refinement procedures, and the mixture for ν_i needs to be estimated within the iterative joint procedure. Extension of the latter procedure is explained here.

The iterative joint procedure requires use of specific parametric priors that comply with INLA. INLA does not allow mixture priors, but the following provides a solution. Introduce the latent variable $G_i = 0, 1$ for ν_i corresponding to the point mass on 0, and non-negative log-Gaussian component, respectively. Then, the posterior of any parameter θ_i (either ν_i or another one, e.g. β_i), $\pi(\theta_i|Y_i)$, depends on $\pi(\theta_i|Y_i, G_i)$ and the likelihoods $\pi(Y_i|G_i)$. As opposed to their unconditional counterparts, these conditional results can directly be obtained from INLA for every G_i, because $G_i = 1$ implies a Gaussian prior on $\phi_i = \log(\nu_i)$, whereas $G_i = 0$ implies a model without ν_i, hence a Poisson model instead of NB. The desired posterior equals:

$$\pi(\theta_i|Y_i) = \sum_{g=0}^{1} \pi(\theta_i|Y_i, G_i = g)\pi(G_i = g|Y_i), \tag{C.10}$$

with

$$\pi(G_i = g|Y_i) = \frac{\pi(Y_i|G_i = g)P(G_i = g)}{\sum_{h=0}^{1} \pi(Y_i|G_i = h)P(G_i = h)} = \frac{\pi(Y_i|G_i = g)q_g}{\sum_{h=0}^{1} \pi(Y_i|G_i = h)q_h}, \tag{C.11}$$

where $q_1 = 1 - q_0$, with q_0 the current estimate of the prior mixture proportion. Note that the procedure above comes at a computational price, because it requires two INLA fits for each feature i.

C.8 Shrinkage of ϕ_i versus shrinkage of w_{0i}. In the ZI-NB model both overdispersion parameter ϕ_i and zero-inflation parameter w_{0i} have the ability to overdisperse the Poisson distribution, albeit via very different mechanisms. However, for some features the posterior means of either $\nu_i = \exp(\phi_i)$ or w_{0i} are zero or extremely close to zero. This is natural: for low-count features, posterior mean $\hat{\nu}_i \approx 0$ may result when w_{0i} sufficiently accounts for the increased variability with respect to the Poisson distribution, while for the high-count features, $\hat{w}_{0i} \approx 0$ simply renders the best fit, because no or very few zeros occur for these features. Shrinkage of ϕ_i and/or w_{0i} is mostly of interest when it robustifies the posterior intervals of the regression parameter(s) of interest, say β_{i1}. We noticed that the effect of shrinking w_{0i} is very minor (standard deviations of β_{i1} altered by less than 1%), whereas shrinkage of ϕ_i generally had much more effect. The small effect of w_{0i} shrinkage may be explained by the intercept of the regression model, which can partly 'repair' the bias introduced by the shrunken estimator of w_{0i}. Therefore, we propose to use flat priors on w_{0i}, although the software accommodates informative priors as well.

C.9 Simulation results: accuracy of estimation. The following simulations are used to determine the accuracy of the estimation methods for a number of situations. In addition, we perform a comparison with edgeR (Robinson et al., 2010) for a case to which both methods are applicable and assess the accuracy of $\text{BFDR}^{\text{II}}(t)$ as an estimate of FDR.

We do not include zero-inflation here, simply because it has little impact on the shrinkage results of the other parameters. Or, to put it differently, the results for data that include a moderate amount of zero-inflation are extremely similar to those from a somewhat larger (in terms of sample size) data set without zero-inflation, when the same simulation hyper-parameters are used. For all simulations, the likelihood and link part of the model are given by:

$$\begin{aligned} Y_{ij} &=^d \text{NB}(\mu_{ij}, \phi_i), \\ \log(\mu_{ij}) &= \eta_{ij}, \end{aligned} \tag{C.12}$$

Below we illustrate the results for a variety of specifications for the regression and prior parts of the model. The fitting strategy always starts with iterative joint estimation of all hyper-parameters, possibly followed by refining the prior for one central parameter of interest using the refinement procedure. In all cases the regression intercept is endowed with a flat $\text{N}(0, (10)^2)$ prior.

Case 1: Two groups, Gaussian-Dirac-Gaussian mixture prior on fixed effects, Dirac-log-Gaussian prior on dispersion. Model (C.12) is further extended by:

$$\begin{aligned}\eta_{ij} &= \beta_{i0} + \beta_{i1} x_j \\ \nu_i &= \exp(\phi_i) =^d q_0 \delta_0 + (1-q_0)\ell\mathrm{N}(\mu, \sigma^2) \\ \beta_{i1} &=^d p_{-1}\mathrm{N}(\mu_{-1}, \tau_{-1}^2) + p_0\delta_0 + p_1\mathrm{N}(\mu_1, \tau_1^2),\end{aligned}$$

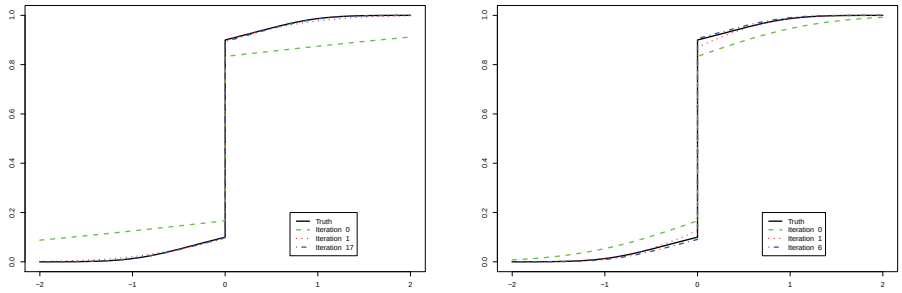
where ℓN denotes the log-Normal distribution, sample size $n = 2 \times 8 = 16, p = 10000$ and $x_j = 0$ for $j \leq n/2$ and $x_j = 1$, otherwise. The mixture prior for ν_i implies a Poisson - NB mixture for Y_{ij} (C.12). True and estimated simulation hyper-parameters are listed in Table C.1. Kolmogorov-Smirnov (KS) distances between the true and estimated distributions of ϕ_i and β_{i1} are given in Table C.2. Figure C.1(a) displays the true and estimated distribution function of β_{i1}.

	$\log(\phi_i)$			β_i					$\log(\tau_i^{-2})$
Case	q_0	μ	σ^2	p_0	(p_{-1}, p_1)	(μ_{-1}, μ_1)	(τ_{-1}^2, τ_1^2)	$(\sigma')^2$	(α_1, α_2)
1, True	.30	-.50	.25	.80	(.10,.10)	(-.50,.50)	(.20,.20)	-	-
1, Est.	.30	-.47	.25	.80	(.09,.11)	(-.53,.44)	(.21,.24)	-	-
2, True	-	-1.50	.25	.80	(.10,.10)	(-.50,.50)	(.20,.20)	-	-
2, Est.	-	-1.51	.23	.79	(.10,.10)	(-.47,.48)	(.22,.20)	-	-
3, True	-	-.50	.25	$t(4)$-dist.					-
3, Est.	-	-.50	.26	Emp.					-
4, True	-	-1.50	.25	-	-	-	-	2.00	(5.00, 1.00)
4, Est.	-	-1.48	.25	-	-	-	-	1.91	(5.51, 1.01)

TABLE C.1. *True hyper-parameter values and their estimates*

Case	KS_ϕ	KS_β	KS_τ
1	0.006	0.007	-
2	0.003	0.011	-
3	0.005	0.005	0.076
4	0.009	0.006	-

TABLE C.2. *Kolmogorov-Smirnov (KS) distance between the estimated and true distribution functions*



(a) β_{i1}: Dirac-Gaussian mixture (b) $\beta_{i\ell}$: Dirac-Gaussian mixture

FIGURE C.1. *Results for Case 1 (a) and Case 2 (b): Estimated and true distribution functions*

Case 2: Multiple groups, Gaussian-Dirac-Gaussian mixture prior on fixed effects, Gaussian prior on log-dispersion. Model (C.12) is further extended by:

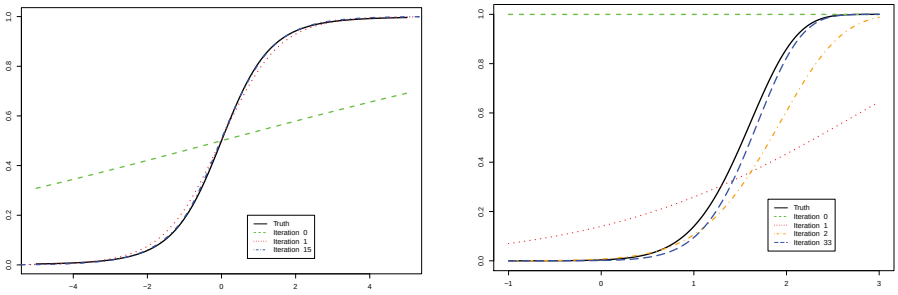
$$\begin{aligned}\eta_{ij} &= \beta_{i0} + \sum_{\ell=1}^{L} \beta_{i\ell} x_{j\ell} \\ \phi_i &=^d \mathrm{N}(\mu, \sigma^2) \\ \beta_{i\ell} &=^d p_{-1}\mathrm{N}(\mu_{-1}, \tau_{-1}^2) + p_0\delta_0 + p_1\mathrm{N}(\mu_1, \tau_1^2).\end{aligned}$$

Here, $n = 25, p = 5000$ and the number of groups $L = 5$, which means that only five measurements per group are available. True and estimated simulation hyper-parameters are listed in Table C.1. KS distances between the true and estimated distributions of ϕ_i and $\beta_{i\ell}$ are given in Table C.2. Figure C.1(b) displays the true and estimated distribution function of $\beta_{i\ell}$.

Case 3: Two groups, t_4-prior on fixed effects, Gaussian prior on log-dispersion. Model (C.12) is further extended by:

$$\begin{aligned}\eta_{ij} &= \beta_{i0} + \beta_{i1} x_j \\ \phi_i &=^d \mathrm{N}(\mu, \sigma^2) \\ \beta_{i1} &=^d t_4,\end{aligned}$$

where sample size $n = 2 \times 8 = 16, p = 5000$, $x_j = 0$ for $j \leq n/2$ and $x_j = 1$, otherwise. Here, we applied the iterative marginal procedure with log-concave nonparametric prior for β_{i1}. True and estimated simulation hyper-parameters are listed in Table C.1. KS distances between the true and estimated distributions of ϕ_i and β_{i1} are given in Table C.2. Figure C.2(a) displays the true and estimated distribution function of β_{i1}.



(a) β_{i1}: $t(4)$-distribution (b) $\log(\tau_i^{-2})$: log-$\Gamma(5,1)$-distribution

FIGURE C.2. *Results for Case 3 (a) and Case 4 (b): Estimated and true distribution functions*

Case 4: Two groups, Gaussian priors on fixed effects and log-dispersion, log-Gamma prior on log-precision of random effect. Model (C.12) is further extended by:

$$\begin{aligned}
\eta_{ij} &= \beta_{i0} + \beta_{i1} x_{j1} + \sum_{\ell=1}^{L} \beta_{i\ell+1} x_{j\ell+1} \\
\phi_i &=^d N(\mu, \sigma^2) \\
\beta_{i1} &=^d N(0, (\sigma')^2) \\
\beta_{i\ell} &=^d N(0, \tau_i^2) \quad \text{for } \ell \geq 2 \\
\tau_i^{-2} &=^d \Gamma(\alpha_1, \alpha_2),
\end{aligned}$$

where sample size $n = 2 \times 9 = 18$, $p = 10000$, the number of random effect levels $L = 6$, x_{j1} as x_j before, and for $\ell \geq 2 : x_{j\ell} = 1$ for $j = 3(\ell-1)-2, \ldots, 3(\ell-1)$ and $x_{j\ell} = 0$, otherwise; hence, a random effects parameter with 6 levels and 3 observations per level. This case is particularly challenging, because it contains two types of (over-)dispersion: one on the observation level (ϕ_i), one on the random effect level (grouping only three observations). True and estimated simulation hyper-parameters are listed in Table C.1. KS distances between the true and estimated distributions of ϕ_i, β_{i1} and τ_i^2 are given in Table C.2. Figure C.2(b) displays the true and estimated distribution function of log-precision, $\log(\tau_i^{-2})$.

Results. In general, the results are very encouraging. From Tables C.1 and C.2 we observe that the hyper-parameters of the (Dirac)-Gaussian prior of $\log(\phi_i)$ are very accurately estimated in all cases. The hyper-parameter estimates of the mixture priors of β_{i1} and $\beta_{i\ell}$ are generally slightly less accurate, but the KS distances to the truth are still very small. This reflects that for mixture priors, which include many hyper-parameters, several configurations of the hyper-parameters are very close in

terms of distribution functions. Note that for case 2 the posterior of each $\beta_{i\ell}$ is based on only 2×5 measurements. From Figure C.2(a) we conclude that the nonparametric prior closely approximates the $t(4)$-distribution, which is rather heavy-tailed. For the same simulation setting, but with the $t(4)$-distribution replaced by a shifted Γ-distribution, Figure C.3 illustrates that the nonparametric prior also performs well for such a skewed distribution.

We obtain the least accurate result for the log-precision of the random effects parameter ($\log(\tau_i^{-2})$) in case 4. KS distance is notably larger than for the other hyperparameters (Table C.2), caused by a too narrow left-tail of the distribution (see Figure C.2(b) for log-precision). Apparently, the posteriors are rather insensitive to the exact shape of the left-tail of the prior, which is reasonable because this tail concerns low precision and hence high variance. Parameter estimates (Table C.1) are fairly accurate though. Given that another dispersion-related parameter is present in the model, and only three repeats per each of the six levels of the random effect are available, we believe the result is still good.

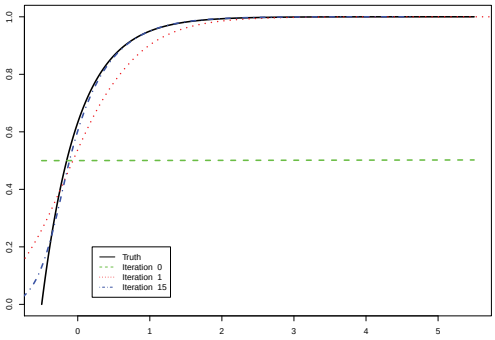


FIGURE C.3. *Simulation results for Case 3, with the t_4 distribution replaced by a shifted Γ-distribution with rate and shape equal to 2 and 1, respectively. Negative shift equals shape/rate to enforce mean equal to 0. Empirically estimated and true distribution functions*

Case 1 revisited: FDR. We used Case 1 to assess the accuracy of the FDR estimate, $\mathrm{BFDR}^{\mathrm{II}}(t)$, when $\mathrm{H}_{0i}^{-} : \beta_{i1} \leq 0$ and $\mathrm{H}_{0i}^{+} : \beta_{i1} \geq 0$. Figure C.4 illustrates that the estimate is slightly conservative, but very accurate over the entire range.

C.10 Simulation results: Comparison with other methods. We compare our method (ShrinkSeq) with four others: edgeR (Robinson et al., 2010, version 2.2.6), DESeq (Anders and Huber, 2010, version 1.6.1), baySeq (Hardcastle and Kelly, 2010, version 1.8.1) and NOISeq (Tarazona et al., 2011, R-scripts). The first three are based on the negative binomial distribution and all allow for shrinkage of overdispersion. NOIseq is a nonparametric method that may be more robust against deviations

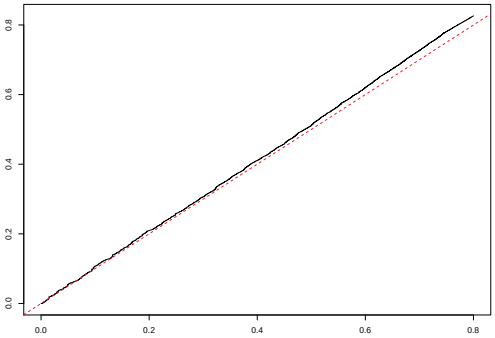


FIGURE C.4. *True versus estimated FDR (solid) and reference diagonal line (dashed). X-axis: true FDR at cut-off t, y-axis: BFDR*$^{\text{II}}(t)$

from the negative binomial. We applied defaults settings. For edgeR, dispersion was shrunken towards a common value; the results for a spline-based trend-estimate were very similar.

For all comparisons we show partial ROC-curves and partial AUC, because in a testing setting only large specificity (we use specificity > 0.8, hence False Positive Rate (FPR) ≤ 0.2) is relevant (Dodd and Pepe, 2003). Partial AUC (pAUC) is expressed in terms of relative AUC (rAUC), $\text{rAUC} = \text{pAUC}/(0.2^2 \times 0.5)$, where the denominator is the expected AUC for a non-informative decision procedure. For the frequentistic methods p-values are used to detect positives, while posterior null-probabilities are used for the Bayesian methods. For baySeq, the latter are computed by comparing models with and without the relevant parameter. For ShrinkSeq, we use $\text{lfdr}_i^{\text{II}}$ for two-sided two-group testing and $\text{lfdr}_i^{\cup}$ for comparing multiple groups (time points). Note that ROC-curves depend on the ranking of the features only, so the comparison is fairly robust against the metric used for declaring a positive.

Model-based simulation: effect of mixture prior on overdispersion. Case 1 is revisited to compare our results with the others in a setting without zero-inflation. The main methodological difference is that the others do not accommodate the mixture prior on the overdispersion. In order to create a fairer comparison, we did not generate $\nu_i = \exp(\phi_i)$ from the above Dirac-log-Gaussian mixture, but instead from a Dirac-log-Gamma mixture, with rate and shape equal to 2 and 1, respectively. This implies a fairly skewed Gamma distribution for ϕ_i (mixed with point mass on $-\infty$). Then, for the actual analysis our method assumes a Dirac-log-Gaussian prior.

In this case, the main parameter of interest is β_{i1}. For the simulated data, we applied our method with a unimodal nonparametric (NP) and a parametric (P) mixture prior (Equation 4.7) for β_{i1}. The latter renders a somewhat biased comparison, because this prior is also used to generate β_{i1} in Case 1. However, it is interesting as

a benchmark for the nonparametric setting and may illustrate a potential gain when using a parametric prior that contains a point mass.

Figure C.5 shows the partial ROC-curves for all methods: the two corresponding to our method, ShrinkSeq, are clearly higher than those for other methods, with the exception of baySeq, which performs similar. For FPR=0.05 the sensitivity of ShrinkSeq is 2% larger than that of baySeq and at least 50% larger than that of edgeR, DESeq and NOISeq. Hence, even though the sample sizes are not very small in this simulation (8 vs. 8), inclusion of a point mass in the prior on the overdispersion is relevant for sensitivity. The parametric version of ShrinkSeq outperforms the nonparametric one, but the difference is fairly small.

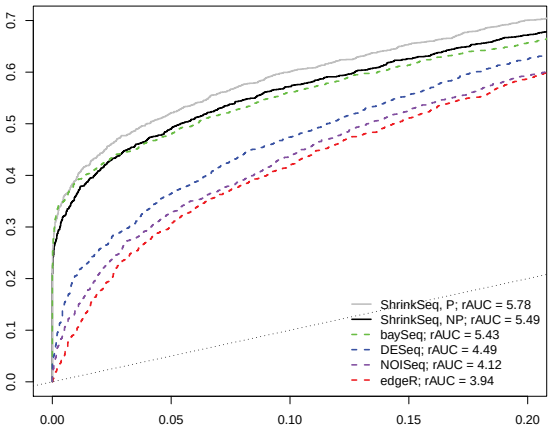


FIGURE C.5. *ROC-curves and relative AUC for β_{i1} in simulation case 1 with a Dirac-log-Gamma mixture distribution of ν_i. X-axis: 1-specificity (false positive rate), y-axis: sensitivity (true positive rate).*

Model-based simulation: effect of shrinking the parameter of interest. The second simulation is the same as the first one, except for ϕ_i being generated from a simple Gaussian distribution, hence complying with the assumptions for overdispersion shrinkage in edgeR, DESeq and baySeq. Performances are now more similar (see Figure C.6), although at FPR $= 0.05$ ShrinkSeq still detects 10-30% more than the other methods. Also, in terms of rAUC, baySeq, edgeR and NOISeq are inferior to ShrinkSeq. The better performance of ShrinkSeq illustrates that (nonparametric) shrinkage of the parameter of interest (β_{i1}), which is not included in any of the other methods, may also aid in terms of power, in addition to its beneficial effects on Bayesian multiplicity correction and prevention of selection bias (Crager, 2010).

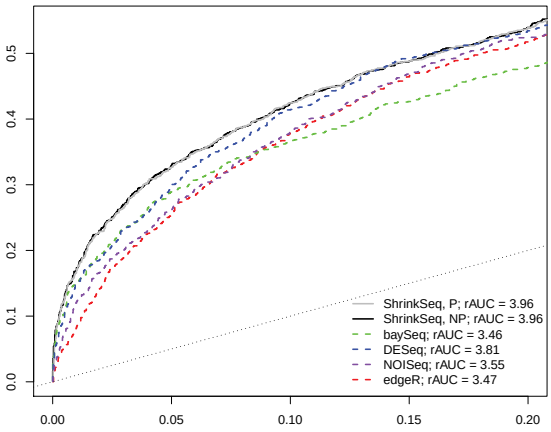


FIGURE C.6. *ROC-curves for β_{i1} in simulation case 1 with a Gaussian distribution of $\phi_i = \log(\nu_i)$: our method (solid) and edgeR (dashed). X-axis: 1-specificity (false positive rate), y-axis: sensitivity (true positive rate)*

Data-based simulation: effect of zeros. To compare our method with others in a setting where the data contains a substantial proportion of zeros we performed the following simulation. We use our data, consisting of 25 observations per feature (see Section 4.7.1), as a template. We do so to a) avoid simulating from a specific parametric setting (e.g. ZI-NB) that would a priori favor our or any other method and b) assure that our simulated data reasonably mimic real data.

The two-group simulation is set up as follows. For each data row we randomly sample 2×8 observations (hence two conditions) from the empirical feature-wise data distribution ($\hat{F}_i$, as constructed from the 25 observation), which results in two sets of observations which are drawn from the same null-distribution. Feature-wise batch effects (which can be substantial) are accounted for in the design by sampling an equal number (4) of observations from the two batches in both groups. In addition, batch is incorporated as a covariate when the software allows for it (DESeq, edgeR, ShrinkSeq). For 10% of the features a differential effect is enforced by multiplying the data of the second group by k_i, which implies a k_i-fold effect and a distribution defined by $\hat{H}_i(k_i x) = \hat{F}_i(x)$. Moreover, for $i = 1,\ldots,1000$, $\beta_{i1} = \log(k_i)$ is drawn from a $N(0,1)$ distribution ($\beta_{i1} = 0$ for $i = 1001,\ldots,10000$). To avoid any bias in the comparison we compare the other methods with ShrinkSeq NP only, which uses a nonparametric prior for β_{i1}.

Figure C.7 shows the partial ROC-curves for all methods. ShrinkSeq clearly outperforms all the other methods. For FPR=0.05, sensitivity of ShrinkSeq is at least 40% larger than that of the others. Hence, accounting for zeros is very relevant for the purpose of increasing sensitivity.

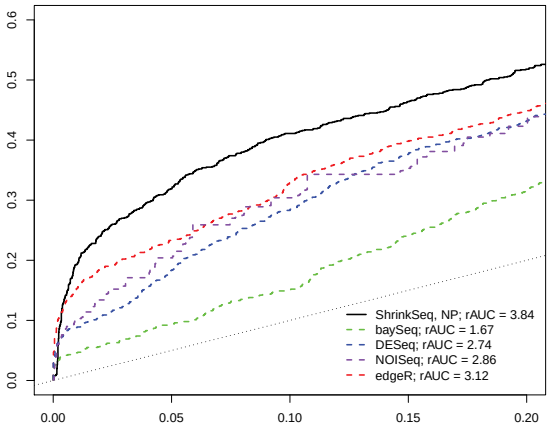


FIGURE C.7. *ROC-curves for β_{i1} in two-group data-based simulation. X-axis: 1-specificity (false positive rate), y-axis: sensitivity (true positive rate)*

Data-based simulation: monotonic time trend. This final simulation assumes a time-course design, where measurements have been taken at four different time points on non-related samples. Again, we use the data as a template to obtain a realistic and fair comparison.

The time-course simulation is set up as follows. For each data row we randomly sample 4×6 observations from the empirical feature-wise data distribution ($\hat{F}_i$, as constructed from the 25 observation), which results in two sets of observations which are drawn from the same null-distribution. Tag-wise batch effects (which can be substantial) are accounted for in the design by sampling an equal number (3) of observations from the two batches for all time points. In addition, batch is incorporated as a covariate when the software allows for it (DESeq, edgeR, ShrinkSeq). NOISeq does not accommodate this design, so is excluded from the comparison. For 10% of the features a differential effect is enforced by multiplying the data of the jth time point ($j > 1$) by k_{ij}, where $\beta_{ij} = \log(k_{ij}) \sim N(0, \sigma_j^2)$ and $(\sigma_2, \sigma_3, \sigma_4) = (0.6, 0.8, 1)$. To avoid any bias in the comparison we compare the other methods with ShrinkSeq NP only, which uses a common nonparametric prior for contrasts $\beta_{ij} - \beta_{ik}$.

Figure C.8 shows the partial ROC-curves for all methods. ShrinkSeq NP outperforms the other methods. The relative good performance of edgeR (with respect to the two-group setting, Figure C.7) illustrates the efficiency of ANOVA-type tests for multi-group differences. Still, for FPR=0.05, sensitivity of ShrinkSeq NP is around 15% better than that of edgeR and at least 50% larger than that of DESeq and baySeq.

In addition, we illustrate a potential gain in power for detecting *monotonic* trends when using a simple linear combination of parameters of the components of $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \beta_{i3}, \beta_{i4})$, $C_i = 3\beta_{i4} + \beta_{i3} - \beta_{i2} - 3\beta_{i1}$, see (C.9). This is trivial in the edgeR

and our setting and does not require any additional development or programming (se Appendix C.6). Figure C.8 displays two additional curves, ShrinkSeq monotone (using a log-concave prior on C_i) and edgeR monotone. Note that here we restrict our attention to the features that correspond to simulated monotonic trends or no differential effect at all. This does not alter the ROC curves for the other methods (apart from some small fluctuations), because these are not using the time order. In both cases we observe an improvement with respect to the same method without accounting for monotonic effects. ShrinkSeq monotone, however, clearly outperforms the others, including edgeR monotone.

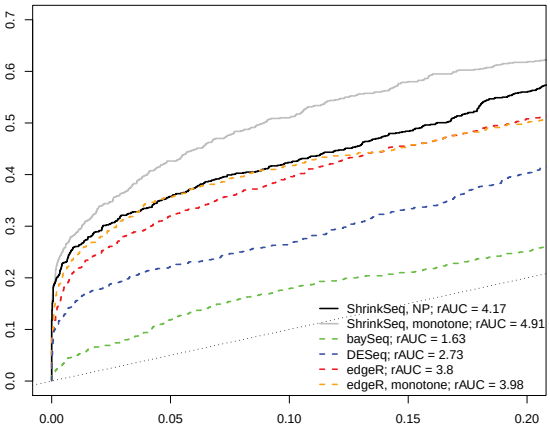


FIGURE C.8. *ROC-curves for β_i in a multi-group data-based simulation. X-axis: 1-specificity (false positive rate), y-axis: sensitivity (true positive rate)*

C.11 Preprocessing of CAGE data. The CAGE methodology (Kodzius et al., 2006) is a validated approach that has been extensively used to profile promoter activity in mice and humans. CAGE tags are 21-27 nucleotides sequence features generated from full length RNA transcripts and mark the transcription start site (TSS) and upstream promoter regions. After sequencing, the CAGE tags were mapped to human genome (Hg18 build). Further the CAGE tags were hierarchically clustered for downstream analysis. Briefly, the CAGE tags that mapped to the same genomic position and on the same strand were considered as CAGE-tag starting sites (CTS). CTSs that were on the vicinity of 21 bps and on the same strand were grouped into a CAGE cluster. Finally, CAGE clusters located within a region of 400 base pairs were grouped into a promoter. A promoter region is then a genomic region comprising a distribution of nearby TSSs that are expected to share the same transcription machinery. To increase the likelihood of identifying real TSS, Pardo et al. (manuscript in preparation) only include CAGE tags that are present (count larger than 0) in at least two libraries and that had a total

count of at least 6 tags per million. A total of 45.000 tag clusters is available. Here, a subset of 10.000 tag clusters is used for the illustration of our approach.

C.12 NB+: Embedding a trend-prior for ϕ_i into our framework. As detailed by Anders and Huber (2010), overdispersion $\phi_i = \log(\nu_i)$ relates to the abundance in a systematic way when using an NB model. They propose to estimate a nonparametric function h, such that $\phi_i = h(c_i) + \epsilon_i$, where c_i is the log of the total feature count. Among many curve fitting methods, LOESS can be used to estimate h. Some initial estimate of ϕ_i is needed to fit h. In our case we suggest to use the feature-wise posterior mean estimates given a flat prior.

Then, assuming $\hat{h} \approx h$, shrinkage is implemented by assuming

$$\phi_i - \hat{h}(c_i) =^d \mathrm{N}(0, \sigma^2),$$

where σ^2 is estimated by our iterative procedure. Then, effectively, the shrunken estimate of ϕ_i pools between the curve estimate and the feature's own overdispersion.

C.13 Results from parametric priors on contrasts in the CAGE data set. Besides the nonparametric priors, we also applied the Dirac-Gaussian and the Gaussian-Dirac-Gaussian mixture priors to the contrasts in the CAGE data set (see Equations 4.6 and 4.7). The parameters of these priors, which include a point mass on zero with mixture proportion p_0, were estimated by using the direct maximization procedure (C.5). The results were not supportive for the use of a point mass: estimates of p_0 were rather low in both cases, approximately 0.3. In fact, we also applied constrained maximization, forcing $p_0 \geq q$ for $q > 0.5$. However, the optimum was always achieved at $p_0 = q$, rendering the result too dependent of the imposed constraint. Alternative parametric priors, Gamma-Dirac-reverse Gamma and Dirac-central Laplace, did not increase the p_0 estimate. The low estimates of p_0 could be real in the sense that perhaps many small, but non-zero, effects exist in this setting. However, note also that this data set is very challenging, because the number of data points per condition is small, 5, and nuisance factors like 'batch' are present. Hence, as in deconvolution problems, it is hard to recover the prior, in particular when it contains a point mass.

C.14 Stabilizing effect of the priors in the CAGE data. We first show that shrinkage of ν_i has a large effect on the stability of the estimate of ν_i, as demonstrated by others in different settings (Anders and Huber, 2010, Robinson and Smyth, 2007). For reasons of comparison, we also fitted model 4.10 using a simple vague prior, $\log(\nu_i) = \phi_i =^d \mathrm{N}(0, (10)^2)$. We evaluated the ratios of posterior standard deviations of $\nu_i = \exp(\phi_i)$ obtained under the vague (V) and informative (I) prior (that effectuates shrinkage): $\mathrm{sd_V}(\nu_i|\mathrm{Y}_i)/\mathrm{sd_I}(\nu_i|\mathrm{Y}_i)$. Of those ratios, 74% were larger than 1, 48% were larger than 5 and 42% were larger than 25, which clearly indicates a strong stabilizing effect of dispersion shrinkage in our setting as well.

Next, we illustrate the strong stabilizing effect of the nonparametric prior with respect to a fairly vague central Gaussian prior (variance equal to 5; results are more dramatic for vaguer priors) we performed a leave-one-out analysis on 250 randomly chosen features: every sample is left out once from the data and all contrasts are re-estimated (posterior means) using 24 instead of 25 samples. Then, the stability of the resulting estimates was evaluated. Figure C.9 clearly illustrates that the range (over the 25 repeats) of the shrunken estimates is much smaller than that of the non-shrunken estimates. This is not just a scaling effect: also the ranks of each parameter estimate (w.r.t. parameter estimates for all features and contrasts) are more stable (see Figure C.10).

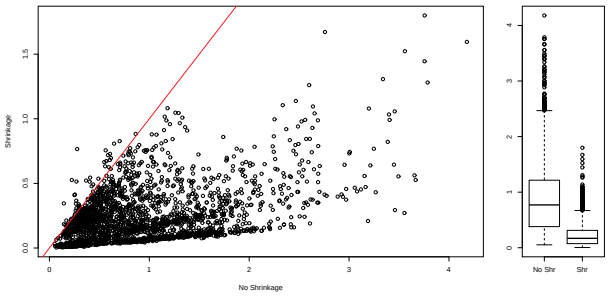


FIGURE C.9. *Stabilizing effect of shrinking the contrasts* $\beta'_{ik\ell}$. *Leave-one-out analysis was performed: for 250 randomly chosen features, each contrast was repeatedly re-estimated from data with one sample less. Each sample was left out exactly once, resulting in 25 estimates of each contrast. X-axis: range of estimate (posterior mean) computed from 25 instances when shrinkage of the contrasts was not used. Y-axis: idem but with shrinkage. Box plots show the distributions of the ranges.*

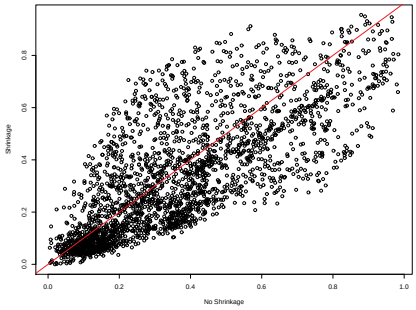


FIGURE C.10. *Stabilizing effect of shrinking the prior of contrasts* $\beta'_{ik\ell}$. *Leave-one-out analysis was performed: for 250 randomly chosen features, each contrast was repeatedly re-estimated from data with one sample less. Each sample was left out exactly once, resulting in 25 estimates of each contrast. For each left-out sample, a contrast estimate (posterior mean) was ranked with respect to the entire vector of 250×10=2500 contrast estimates. X-axis: range of rank (as assessed from 25 instances) when shrinkage of the contrasts was not used. Y-axis: idem but with shrinkage.*

C.15 Details on the analysis of the HapMap RNA-seq data set. This data set is a standard HapMap RNA-seq data set which consists of 60 samples of Caucasian (Montgomery et al., 2010) origin and 69 samples of Nigerian (Pickrell et al., 2010) origin. The data are available from the ReCount (Frazee et al., 2011) web site: http://bowtie-bio.sourceforge.net/recount/. The data was normalized using edgeR's quantile-adjusted conditional maximum likelihood (qCML) method (Robinson and Smyth, 2008).

For studying reproducibility, the data were first randomly split into two equal-size splits after removing the last Nigerian sample. To avoid analyzing features with almost only zeros, features with less than 5 non-zeros in either of the splits were filtered out, rendering 10,369 features in total. For consistency reasons, the unbalanced splits, described in 4.7.5, are applied to the same feature set.

DESeq and edgeR p-values were corrected using the Benjamini-Hochberg (1995) FDR correction. baySeq provides an internal FDR correction which we use. ShrinkSeq was used with $\text{BFDR}^{\text{II}}(t)$ (C.7) and $\Delta = 0.1$. Cut-off 0.1 was used for all four methods.

Application of ShrinkSeq is straightforward for this simple design. First, the single parameter of interest coding for the group difference, β_{i1}, and the overdispersion parameter ϕ are shrunken using the iterative joint procedure. As for the CAGE data a mixture prior was used for $\nu_i = \exp(\phi_i) =^d q_0\delta_0 + (1-q_0)\ell\text{N}(\mu,\sigma^2)$. For this data, the iterative joint procedure rendered $\hat{q}_0 = 0$ for all splits, indicating that the ZI-NB should be used for all features and highlighting automatic model selection properties of our method. Next, posteriors are computed for all features using INLA. Finally, as for the CAGE data, the posteriors of β_{i1} are updated by applying a log-concave nonparametric prior determined by the marginal iterative procedure.

C.16 Software. All the methodology discussed in this paper is implemented in R. The code and the data are available from www.few.vu.nl/˜mavdwiel. Parts of our software rely on the INLA R-package (available from www.r-inla.org, in particular fitting of the count models and the iterative joint procedure for estimating hyper-parameters. Other parts, like the marginal iterative procedures and the BFDR/lfdr implementation, apply to any software or methodology that provides numerical representations of marginal posteriors.

Computing time. As discussed in Section 4.1, one of the reasons to use INLA is its computational efficiency with respect to MCMC. Still, our method usually takes more time than p-value-based methods. However, we show in this section that, given the amount of data, computing times are reasonable, and large, realistic data sets can be processed.

Computing time depends on the complexity of the design, the count model, type of shrinkage and convergence speeds of the iterative procedures. As a first indication: the entire analysis of the presented data in Section 4.7.1 (10000 features, fairly complex design) took 2 hours on an ordinary quad-core PC, 37% (45 min. = 0.75 hour) of which was used for fitting the models on all features. Computing time scales up less

than proportionally to the number of features, because, as explained earlier, the time consumed by the iterative procedures does not increase with the number of features exceeds, say, 10000. Hence, in this data setting computing time (in hours) for $p >$ 10000 features approximately equals: $T_p = 1.25 + 0.75p/10000$. This was verified on the entire data set from which our illustration data was extracted from and indeed $T_{70000} \approx 1.25 + 0.75 \times 7 = 6.5$ hours. Table C.3 provides computation times for several values of p.

To illustrate how computing time depends on complexity of the design, we also present computing times for the data-based simulation with two groups and no additional covariates (see Appendix C.10). On 10000 features, computations were twice as fast as for the previous, more complex design: approximately one hour, 30% (18 min. $= 0.3$ hour) of which was used for fitting the model. Hence, computation time for $p > 10000$ equals $T_p = 0.7 + 0.3p/10000$. Table C.3 provides computation times for several values of p.

# features p	I	II
1×10^4	2.0	1.0
2×10^4	2.8	1.3
5×10^4	5.0	2.2
1×10^5	8.8	3.7
2×10^5	16.3	6.7
5×10^5	38.8	15.7
1×10^6	76.3	30.7

TABLE C.3. *Computation times (hours) for ShrinkSeq on a quad-core PC for two settings: complex design (I), simple two-group design (II). PC specifications: Intel® Xeon®, CPU E5530, 2.40GHz; 12GB RAM.*

Our code allows for parallel computing. Since many of the functions parallelize trivially (e.g. fitting the models for all features) computing time scales down nearly proportionally when the number of computing cores increase. A 12-core cluster, with the same clock speed as the quad-core, processes 10^6 features in just more than a day (26 hours) for the complex design (I) and approximately 10 hours for the two-group design (II).

C.17 Example. Below we provide an example of the methods for estimating priors that are presented in Section 4.3. The methods can be applied to *any* software that provides approximations of marginal posteriors for *given* priors. Our implementation is based on INLA (Rue et al., 2009), but, since our software provides wrappers that invoke INLA for our setting, only limited knowledge of INLA is required. Note also that www.r-inla.org provides many examples on INLA itself. In addition, Fong et al., 2010 present examples of INLA in a GLM setting.

For illustration purposes we use the simplest setting: the two-group comparison for the HapMap RNA-seq data, discussed in 4.7.5 and Appendix C.15. We use the results for a small subset (8 vs 8) to illustrate the effect of shrinkage. The number of features equals $p = 10,369$. This example data set is available from www.few.vu.nl/~mavdwiel, including the R-scripts. For the posteriors we focus on three example features with data:

Index	Data group 1								Data group 2							
4079	129	147	148	175	116	209	171	181	473	137	182	216	548	178	177	281
4004	114	338	698	589	1082	410	1036	161	4297	1681	1356	476	596	807	859	792
6080	0	0	0	0	0	0	2	2	21	2	4	1	2	1	7	5

TABLE C.4. *Three example features.*

Iterative joint procedure. We first use the iterative joint procedure to fit:

$$
\begin{aligned}
Y_{ij} &=^d \text{ZI-NB}(\mu_{ij}, w_{0i}, \phi_i,) \\
\log(\mu_{ij}) &= \eta_{ij} \\
\eta_{ij} &= \beta_{i0} + \beta_i x_j \\
\beta_{i0} &=^d \text{logit}(w_{0i}) =^d \text{N}(0, 100) \\
\nu_i = \exp(\phi_i) &=^d q_0 \delta_0 + (1-q_0)\ell\text{N}(\mu, \sigma^2) \\
\beta_i &=^d \text{N}(0, (\sigma')^2)
\end{aligned}
\tag{C.13}
$$

where $x_j = 0, 1$ codes for the two groups, w_{0i} is the common zero-inflation parameter, ϕ_i is the overdispersion parameter and $\beta_i, i = 1, \ldots, p$ are the main parameters of interest. Write the prior density of β_i, corresponding to cdf $\text{N}(0, (\sigma')^2)$, evaluated at β as $\psi(\beta; 0, (\sigma')^2)$. The latter two equations in (C.13) contain the shrinkage priors that we aim to estimate using our iterative algorithms. Hence, in the notation of Section 4.3, the hyper-parameters that we aim to estimate are: $A = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2\} = \{\sigma', (q_0, \mu, \sigma)\}$. For *known* A model (C.13) fits within the context of INLA, which provides marginal posteriors of β_i and the other parameters. Below we illustrate the steps of the iterative joint procedure.

STEP 1. We first initiate $\ell = 0$, $\sigma' = \boldsymbol{\alpha}_1^{(0)} = 10$ (hence a large sd), $(q_0, \mu, \sigma) = \boldsymbol{\alpha}_2^{(0)} = (0.2, 0, 10)$, and $A^{(0)} = \{\boldsymbol{\alpha}_1^{(0)}, \boldsymbol{\alpha}_2^{(0)}\}$. Let us, for illustration purposes, focus on $\boldsymbol{\alpha}_1 = \sigma'$ and $\theta = \beta$, implying for Equation 4.2: $\pi_{\boldsymbol{\alpha}}(\theta) = \pi_{\boldsymbol{\alpha}_1}(\theta) = \pi_{\sigma'}(\beta) = \psi(\beta; 0, (\sigma')^2)$ and $A^- = \{\boldsymbol{\alpha}_2\} = \{(q_0, \mu, \sigma)\}$.

STEP 2. Given $A^{(0)}$ and hence all priors, INLA is used to approximate marginal posteriors of β_i, $\pi_{A^0}(\beta|Y_i)$ for all features $i = 1, \ldots, p$. It returns this posterior in a numerical matrix format which contains the support and the posterior density at the support.

The resulting posterior density is displayed in light-blue (Iter. 0) in Figure C.12 for the three example features.

STEP 3. The empirical mixture of the posteriors of all β_i is $\pi^{\text{Emp}}_{A^{(0)}}(\beta) = \frac{1}{p}\sum_{i=1}^{p} \pi_{A^0}(\beta|Y_i)$. We sample say $S = 100.000$ observations from this empirical mixture rendering $z_{A^{(0)}}$ [see Equation 4.3] by simply collecting one or multiple samples from each posterior $\pi_{A^0}(\beta|Y_i), i = 1,\ldots,p$. Then, we use MLE on $z_{A^{(0)}}$ to re-estimate σ' under the central Gaussian prior, which results in $\boldsymbol{\alpha}_1^{(1)}$. Steps 2 and 3 are repeated for $\boldsymbol{\alpha}_2 = (q_0, \mu, \sigma)$ which results in $\boldsymbol{\alpha}_2^{(1)}$.

STEP 4. The algorithm is iterated until the criteria based on Kolmogorov-Smirnov distance, as explained in Appendix C.3, are satisfied, which is the case after 30 iterations. In that section we also explain that for the purpose of estimating the hyper-parameters smaller values of p (the number of features), $p^{(t)}$, may be used to reduce computing time. For this data, the first 14 iterations used $p^{(t)} = 100$, the next 5 on $p^{(t)} = 200$, the next 7 on $p^{(t)} = 500$ and the final 4 on $p^{(t)} = 2000$.

Table C.5 shows part of the subsequent estimates of A. The final result is $\hat{A} = \{\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2\} = \{\hat{\sigma}', (\hat{q}_0, \hat{\mu}, \hat{\sigma}) = \{0.361, (0.000, 1.579, 0.974)\}$. Because $\hat{q}_0 = 0$ the 'zero-overdispersion' component in (C.13) is not needed for this data. Figure C.11 shows the iterative priors for β_i and $\phi_i = \log(\nu_i)$ [the Normal component of it] in blue. Figure C.12 shows the iterative posteriors for the three example features. The shrinkage effect is the strongest on β_{6080}, both in terms of the location (mainly due to its own prior) and variance (mainly due to the prior of ϕ_i). The strong effect of the priors for this feature is likely due to the many zeros in group 1. This makes estimation of the log-ratio, which is what this parameter represents, imprecise, and potentially also inaccurate, when relatively vague priors are used.

Iteration	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$		
	$\hat{\sigma}'$	$\hat{q}_0$	$\hat{\mu}$	$\hat{\sigma}$
0	10.000	0.200	0.000	10.000
2	1.170	0.083	1.196	1.500
4	0.745	0.032	1.305	1.190
7	0.558	0.016	1.283	1.130
11	0.479	0.012	1.273	1.150
15	0.454	0.006	1.368	1.100
19	0.469	0.000	1.468	1.040
23	0.415	0.000	1.443	0.964
27	0.380	0.000	1.537	0.971
30	0.361	0.000	1.579	0.974

TABLE C.5. *Estimates of the hyper-parameters in* A *from the iterative joint procedure*

Iterative marginal procedure. Since β_i is our central parameter of interest assuming its prior to be central Gaussian may be too stringent. One may prefer to use a non-parametric prior, but this is not allowed by INLA, which we need to (re-)estimate the posteriors under given priors. However, INLA can be by-passed by using the iterative marginal procedure outlined in Section 4.3.2.

We now show how to use Equation 4.4 to re-compute posteriors under a log-concave prior $\pi'(\beta)$. This is essential in the iterative marginal procedure: the old prior and posteriors comply with INLA, the new ones need not to. Let $A^* = \{\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*\} = \{\sqrt{(10)}\hat{\sigma}', \hat{\boldsymbol{\alpha}}_2\} = \{1.142, (0.000, 1.579, 0.974)\}$. Hence, for reasons mentioned below Equation 4.4, first posteriors $\pi_{A^*}(\beta|Y_i)$ are computed with INLA under a prior $\pi_{\boldsymbol{\alpha}_1^*}(\beta)$, which is a central Gaussian prior density with wider support than the one resulting from the iterative joint procedure. Let us emphasize that, even though the estimate from the iterative joint procedure $\hat{\boldsymbol{\alpha}}_1 = \hat{\sigma}'$ is only indirectly used now, it is still important to first jointly estimate $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ (as we did), due to their potential interdependency. Next, we illustrate the steps of the iterative marginal procedure.

STEP 1. Initiate $\ell = 0$ and $\pi'(\beta) = \pi'^0(\beta) = \pi_{\boldsymbol{\alpha}_1^*}(\beta) = \psi(\beta; 0, 10(\hat{\sigma}')^2)$. One may also initiate $\pi'(\beta) = \pi_{\boldsymbol{\alpha}_1}(\beta) = \psi(\beta; 0, (\hat{\sigma}')^2)$ (the Gaussian prior resulting from the iterative joint procedure), but in our experience convergence occurs earlier with a 'too wide' prior than with a 'too narrow' one.

STEP 2. [This step may be skipped for $\ell = 0$, because $\pi'^0(\beta) = \pi_{\boldsymbol{\alpha}_1^*}(\beta)$] Compute the right-hand side of Equation 4.4: simply re-weigh the mass on posteriors $\pi_{A^*}(\beta|Y_i)$ by the ratio of prior masses. Then normalize the result to obtain proper posterior densities. This is done by (univariate) numerical integration.

STEP 3. As before compute the empirical mixture of the posteriors of all β_i (under prior π'^0) and sample from it. The best log-concave density (Lutz and Rufibach, 2011) is then fit to this sample, which provides a new estimate of the prior $\pi'^1(\beta)$.

STEP 4. The algorithm is iterated until the marginal likelihood criterion, as explained in Appendix C.3, is satisfied. This typically requires somewhat more iterations than the Kolmogorov-Smirnov-based criterion, but this is acceptable, because the resulting prior is the final one used for the central parameter of interest, β_i and hence should be very accurate.

The final log-concave prior is displayed in black in the left display of Figure C.11. The effect on the posteriors of the three example features is visualized in Figure C.12. The differential shrinkage effect w.r.t. the final Normal prior is the strongest on β_{6080}, for which the log-concave prior implies a much more skewed posterior.

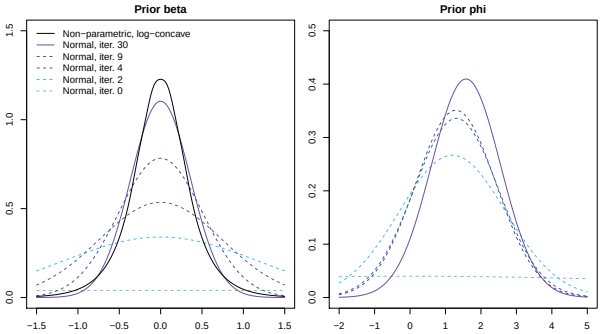


FIGURE C.11. *Estimated priors for β_i and ϕ_i [Gaussian component]*

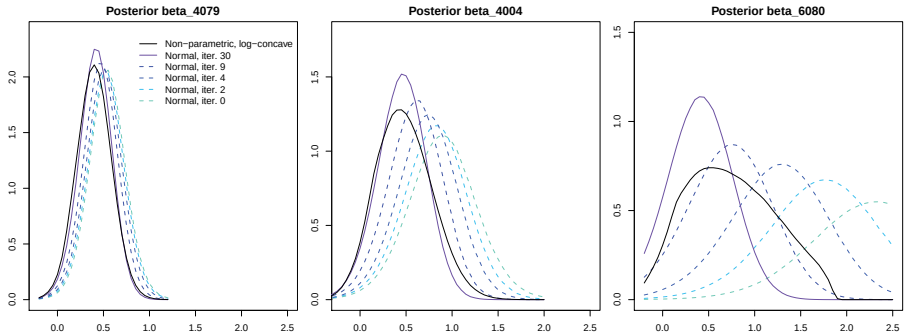


FIGURE C.12. *Posteriors of $\beta_{4079}, \beta_{4004}, \beta_{6080}$ under the corresponding priors of Figure C.11*

C.18 Additional Figures and Tables.

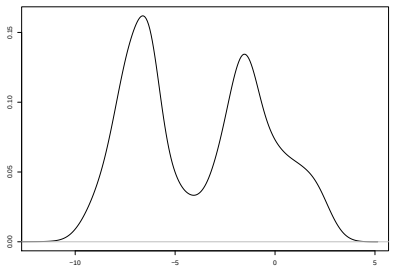


FIGURE C.13. *Motivation for use of mixture prior for overdispersion for the CAGE data. X-axis: posterior mean of ϕ_i when an (initial) vague Gaussian prior is used. Y-axis: density. Because of the log-scale, the left subpopulation corresponds to overdispersions $\nu_i = \exp(\phi_i)$ very close to 0. This motivates a Dirac-logNormal mixture prior on ν_i.*

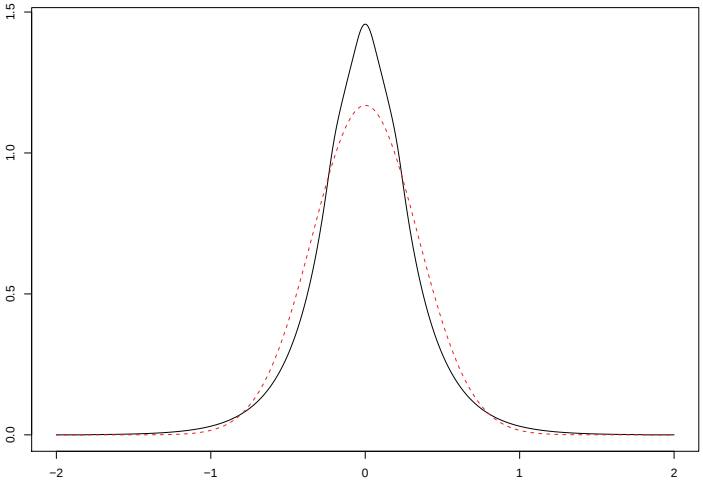


FIGURE C.14. *The nonparametric prior density of the contrasts $\beta'_{ik\ell}$, as obtained from the data (solid) and the corresponding central Gaussian density with the same mean and standard deviation (dashed).*

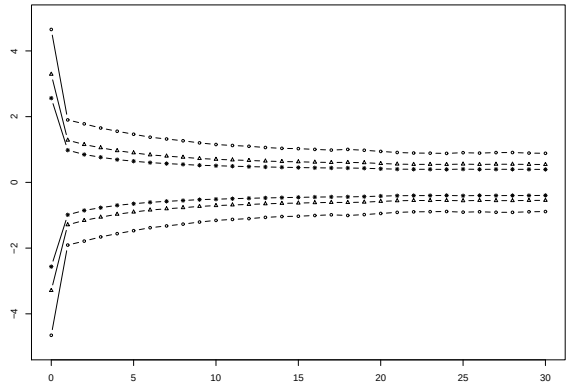


FIGURE C.15. *Convergence of the nonparametric prior of the contrasts $\beta'_{ik\ell}$ for the data. X-axis: iteration, y-axis: left and right 1% (solid), 5% (dashed), 10% (dotted) quantiles of the estimated prior. Iteration* 0 *corresponds to the initial Gaussian prior.*

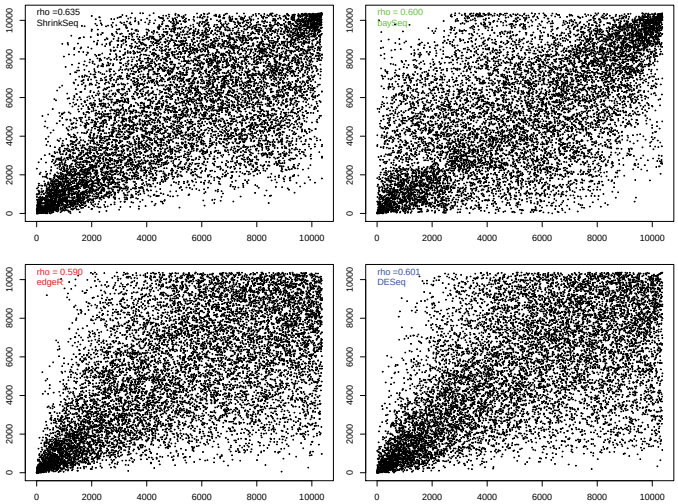


FIGURE C.16. *Correlation of the results between two halves of the Montgomery-Pickrell RNA-seq data set for four methods. X-axis: rank according to significance for the first half; Y-axis: rank according to significance for the second half. Spearman's ρ is also displayed.*

	ZI-NB	NB+	NB
$\mu_{ik\ell}^{\max} \leq 12.6$	78	1	2
$12.6 < \mu_{ik\ell}^{\max} \leq 37.6$	172	28	43
$37.6 < \mu_{ik\ell}^{\max} \leq 82.4$	82	60	40
$82.4 < \mu_{ik\ell}^{\max} \leq 299.4$	108	85	56
$299.4 < \mu_{ik\ell}^{\max}$	38	32	31
Sum	478	206	172

TABLE C.6. *Number of detected differential contrasts with $lfdr^{II}(t) \leq lfdr_{max} = 0.2$ and $\Delta = 0.25$ using the ZI-NB, NB+ and NB models, where the posteriors are based on the nonparametric prior for contrasts $\beta'_{ik\ell}$ (depicted in Figure C.14). Rows 2-6 represent very low-count, low-count, medium-count, high-count and very high-count contrasts, where $\mu_{ik\ell}^{\max} = \max(\mu_{ik}, \mu_{i\ell})$, with μ_{ih}: mean count for feature i and group h. Here, 12.6, 37.6, 82.4 and 299.4 are the 80%, 90%, 95% and 99% empirical quantiles of the vector containing all values of $\mu_{ik\ell}^{\max}$, respectively.*

Index	Data group 4	Data group 5	q_5, ZI-NB q_{10}, ZI-NB q_{20}, ZI-NB	q_5, NB+ q_{10}, NB+ q_{20}, NB+	q_5, NB q_{10}, NB q_{20}, NB
3937	2 3 0 3 5	7 2 10 6 4	-0.087 0.008 0.129	-0.261 -0.162 -0.044	-0.137 -0.039 0.081
176	0 1 4 3 9	30 17 3 27 27	-0.034 0.076 0.227	-0.118 -0.019 0.106	-0.091 0.011 0.146
3058	0 13 6 8 13	17 17 64 30 88	0.096 0.211 0.375	0.011 0.120 0.268	-0.199 -0.097 0.024
8067	80 4 26 2 3	120 253 9 204 36	0.007 0.109 0.242	0.004 0.098 0.220	-0.042 0.056 0.186
3173	221 12 79 18 17	928 869 168 466 345	0.757 0.908 1.078	0.720 0.860 1.025	0.654 0.811 0.997

TABLE C.7. *Lower 5%, 10% and 20% quantiles of the posteriors of contrast* $\beta'_{i54} = \beta^G_{i5} - \beta^G_{i4}$ *for* $i = 3937, 176, 3058, 8067, 3173$ *using the ZI-NB, NB+ and NB models, where the posteriors are based on the nonparametric prior for* $\beta'_{ik\ell}$ *(depicted in Figure C.14). The five contrasts are representatives of very low-count, low-count, medium-count, high-count and very high-count contrasts (see Table C.6). Quantile* $q_5 > \Delta$ *indicates that the contrast would be detected when using* $H_0 : \beta'_{ik\ell} < \Delta$ *and* $lfdr^+ < 0.05$*, (likewise for* q_{10} *and* q_{20}*).*

1st half	Shrink-Seq	edgeR	DE-Seq	bay-Seq	2nd half	Shrink-Seq	edgeR	DE-Seq	bay-Seq
ShrinkSeq	1.000	0.964	0.941	0.765	ShrinkSeq	1.000	0.954	0.909	0.685
edgeR	0.964	1.000	0.957	0.751	edgeR	0.954	1.000	0.922	0.660
DESeq	0.941	0.957	1.000	0.659	DESeq	0.909	0.922	1.000	0.606
baySeq	0.765	0.751	0.659	1.000	baySeq	0.685	0.660	0.606	1.000

TABLE C.8. *Spearman correlations between the results of four methods within two halves of the Montgomery-Pickrell RNA-seq data sets.*

APPENDIX D

D.1 Supplementary Figures.

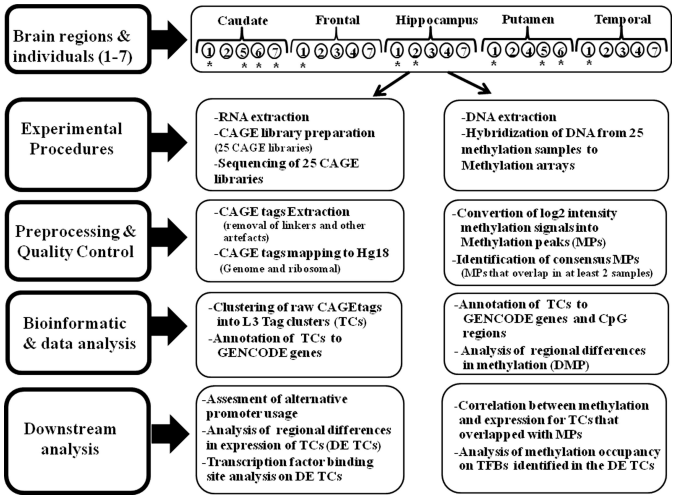


FIGURE D.1. *Schema of main experimental and data analysis procedures.*

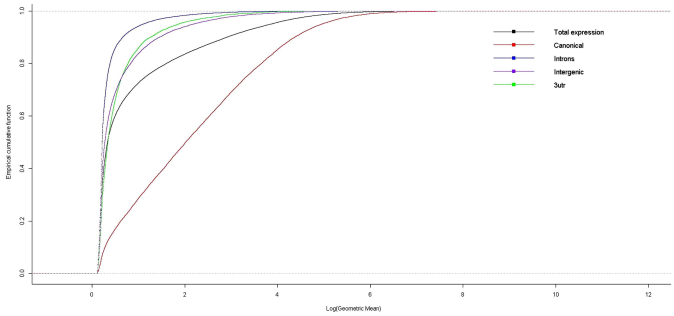


FIGURE D.2. *Empirical cumulative frequency distribution of the expression (log geometric mean) of Level 3 CAGE Tag clusters (TCs). The empirical cumulative distribution of expression (y-axis) of the logarithmic geometric mean expression (x-axis) of TCs is presented in black. The empirical cumulative function of canonical (red) and non-canonical TCs (intronic (blue), intergenic (purple) and 3'UTR (green)) is also presented. The graph shows that canonical TCs expression accounts for most of the overall expression. In contrast, most non-canonical TCs are expressed at low levels.*

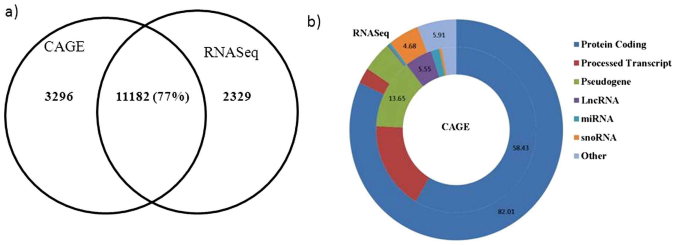


FIGURE D.3. *Comparison of genes identified by RNA-Seq and CAGE. a) Venn diagram showing the number of genes expressed in brain and identified with CAGE and/or RNASeq (Ramsköld et al., 2009); b) Biotype classes of the genes that were identified by CAGE (inner circle) or RNASeq (outer circle). A larger proportion of ncRNA classes were identified by CAGE*

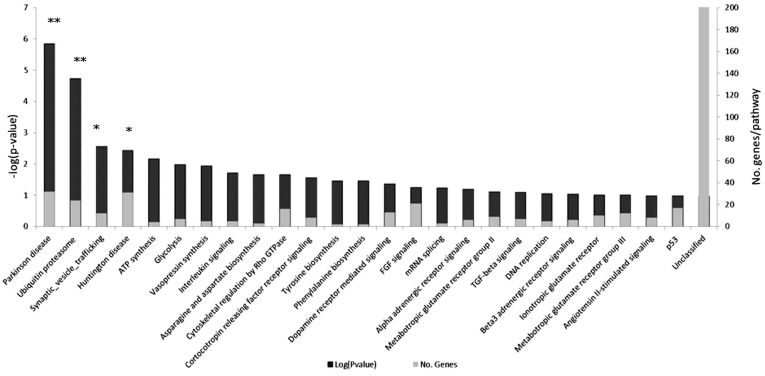


FIGURE D.4. *Functional pathway analysis of highly expressed genes with canonical Level 2 CAGE Tag clusters. Comparative graph of the functional pathways showing significant overrepresentation identified for the most highly expressed genes. The left y-axis presents* -log *p-value of the binomial test. The right y-axis presents the number of genes per pathway. Asterisks represent significant p-values (** p-value$<10^{-5}$;*p-value$<10^{-3}$)*

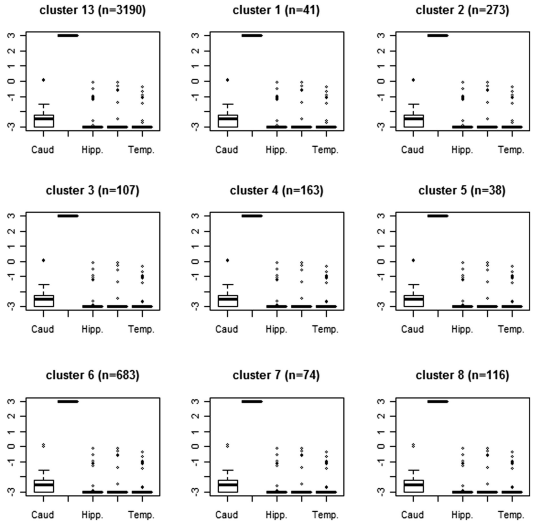


FIGURE D.5. *Boxplots of β-regression coefficients (y-axis) for every brain region (x-axis) and for each of the 29 DE modules.*

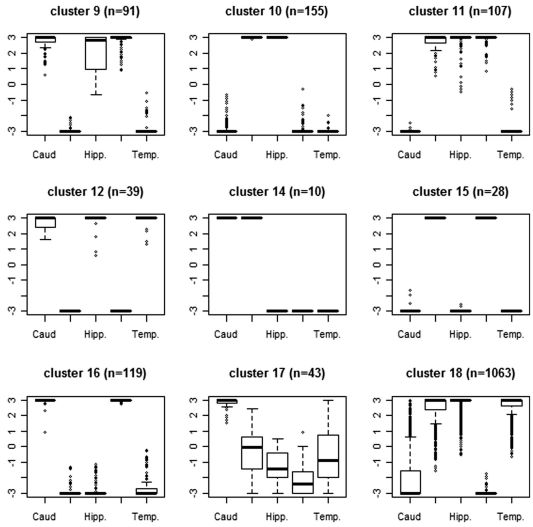


FIGURE D.6. *Boxplots of β-regression coefficients (y-axis) for every brain region (x-axis) and for each of the 29 DE modules.*

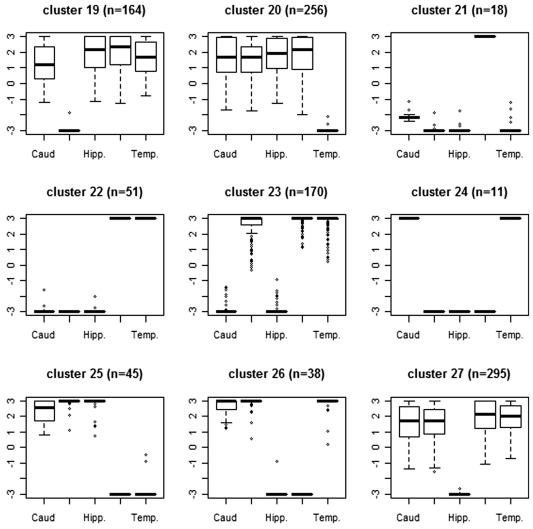


FIGURE D.7. *Boxplots of β-regression coefficients (y-axis) for every brain region (x-axis) and for each of the 29 DE modules.*

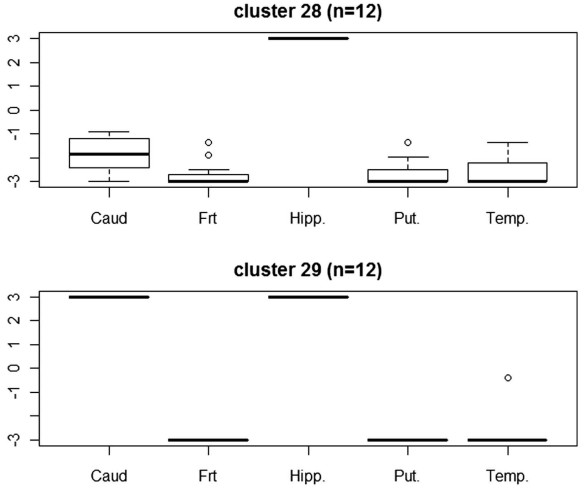


FIGURE D.8. *Boxplots of β-regression coefficients (y-axis) for every brain region (x-axis) and for each of the 29 DE modules.*

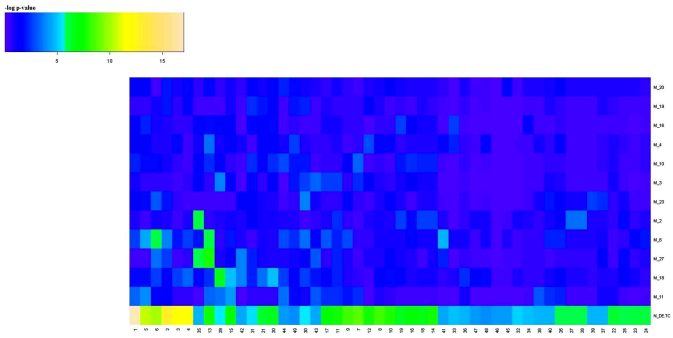


FIGURE D.9. *Heatmap of* -log *p-values of overrepresented functional pathways per differentially expressed module.*

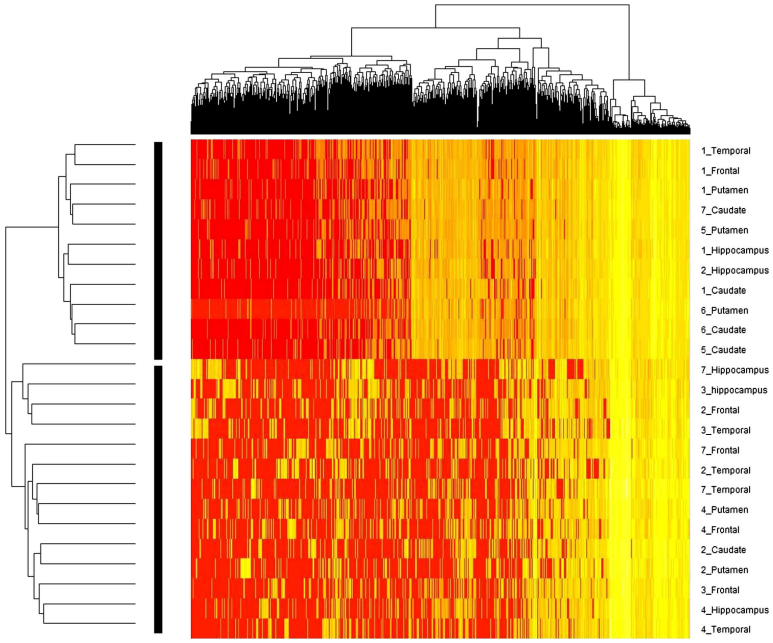


FIGURE D.10. *Heatmap of the unsupervised clustering of TC expression profiles from 25 libraries. Only data from chromosome 1 is depicted due to computer limitations.*

D.2 Supplementary Tables.

Sample id	Gender	Age	Braak staging	Braak Amyloid score	Braak Alpha syn	Cause of death	Post-mortem delay (hours)
1	M	91	1	B	1	Cardiac decompensation	08:00
2	M	87	3	A	0	Unknown	06:05
3	F	82	3	B	0	Cardiac failure	05:10
4	F	87	2	O	0	Cachexia and dehydration	07:00
5	F	89	2	0	0	Dyspnea	04:15
6	M	70	0	A	0	Pancreas carcinoma	06:55
7	F	97	1	C	0	Cachexia and dehydration	05:00

TABLE D.1. *Demographic features of brain donors.*

APPENDIX E

E.1 Generated partial correlations.

d	Graph	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$d = 20$	Band 1	0.2273	0.3773	0.4280	0.4252	0.4930	0.5368
	Band 2	0.1038	0.2727	0.3106	0.3134	0.3549	0.4762
	Cluster	0.1328	0.3710	0.4080	0.4113	0.4658	0.6274
	Hub	0.1593	0.3031	0.3539	0.3385	0.3949	0.4630
$d = 50$	Band 1	0.1381	0.3420	0.4076	0.4069	0.4652	0.5821
	Band 2	0.1082	0.2630	0.3044	0.3049	0.3475	0.5052
	Cluster	0.1627	0.3895	0.4539	0.4536	0.5153	0.6871
	Hub	0.1367	0.2371	0.2806	0.2902	0.3451	0.4440
$d = 100$	Band 1	0.1405	0.3586	0.4182	0.4140	0.4688	0.6214
	Band 2	0.1396	0.2475	0.2862	0.2846	0.3212	0.4439
	Cluster	0.1685	0.3606	0.4086	0.4135	0.4628	0.6764
	Hub	0.1378	0.2555	0.2912	0.2929	0.3364	0.4768
$d = 200$	Band 1	0.1482	0.3497	0.3939	0.3995	0.4509	0.6300
	Band 2	0.1334	0.2451	0.2825	0.2830	0.3204	0.4538
	Cluster	0.1370	0.3891	0.4528	0.4546	0.5204	0.7236
	Hub	0.1174	0.2362	0.2732	0.2786	0.3187	0.4610

TABLE E.1. *Five-point summary on absolute values of the true (non-zero) partial correlations*

E.2 Methodological details. For each method under comparison in Section 6.3.2, we here describe how the edge ranking is obtained and how the graph structure is selected.

1. **SEM$_{\text{L}}$**

   Consider a single regularization parameter λ for all regression equations, then for a given λ the graph is determined by the following heuristic: if $\beta_{i,j} \neq 0$ and $\beta_{j,i} \neq 0$, then an edge is present between nodes i and j. Ranks of edges are determined by the order in which they enter the graph when decreasing λ from $\lambda_{\max}$ to $\lambda_{\min}$, where $\lambda_{\max}$ (resp. $\lambda_{\min}$) is high (low) enough so the null (full) graph is obtained.
   Graph selection: For each regression, variables are selected with BIC (hence λ varies accross regressions) and the graph is determined with the above heuristic.

2. **SEM$_{\text{L_STAB}}$**

   For each node-wise regression, stability selection is as follows. We first choose a regularization parameter λ_0 that includes about $\lfloor \pi_0 p \rfloor$ variables, where p is

the total number of covariates and $0 \leq \pi_0 \leq 1$. The model is chosen so as to 'overfit' (the subset of true variables is believed to be smaller). Then, B random samples of size $\lfloor n/2 \rfloor$ are taken without replacement from the data set and a linear model with an ℓ_1-penalty and fixed regularization parameter λ_0 is fitted on each of them. Variable importance is reflected by the empirical probability of selection in the model. Now consider $P_{i,j}$ the probability of selection of variable j when it is regressed on variable i with all other variables, then for a fixed threshold $\min_{i,j}(P_{i,j}) \leq \pi \leq \max_{i,j}(P_{j,i})$ the graph is determined by the following heuristic: if $P_{i,j} \geq \pi$ and $P_{j,i} \geq \pi$, then an edge is present between nodes i and j. Last, ranks of edges are determined by the order in which they enter the graph when decreasing π from $\max_{i,j}(R_{i,j})$ to $\min_{i,j}(R_{i,j})$. In our simulations, we set $\pi_0 = 0.25$, $B = 200$ and used R package `glmnet` (Friedman et al., 2010). This approach to stability selection is called "pointwise control" and described in Meinshausen and Bühlmann (2010).
Graph selection: We first symmetrize the matrix P. Then, to determine the graph structure, a probability threshold π is chosen so the expected proportion of falsely selected edges is less than 10% (Meinshausen and Bühlmann, 2010).

3. **GL$_\lambda$**

   Ranks of edges are determined by the order in which they enter the graph when the regularization parameter λ of the graphical lasso is decreased from λ_{max} to λ_{min}, where λ_{max} (resp. λ_{min}) is high (low) enough so the null (full) graph is obtained.
   Graph selection: Select λ based on BIC

4. **GL$_{STAB}$**: Graphical lasso with stability selection

   We first choose a regularization parameter λ_0 that includes about $\lfloor \pi_0 \frac{p(p-1)}{2} \rfloor$ edges, where p is the total number of variables and $0 \leq \pi_0 \leq 1$. The model is chosen so as to 'overfit' (the subset of true edges is believed to be smaller). Then, B random samples of size $\lfloor n/2 \rfloor$ are taken without replacement from the data set and graphical lasso with fixed regularization parameter λ_0 is fitted on each of them. Ranks of edges are determined by their empirical probability of selection in the model over the subsamples. Here, we set $\pi_0 = 0.20$ and $B = 200$. R package `glasso` was used.
   Graph selection: To determine the graph structure, a probability threshold π is chosen so the expected proportion of falsely selected edges is less than 10% (Meinshausen and Bühlmann, 2010).

5. **GeneNet**: Shrinkage estimation and *a posteriori* node selection

   An estimate of the partial correlation matrix is first obtained (using function *ggm.estimate.pcor()* of R package GeneNet; Schaefer et al. (2006), Schäfer and Strimmer (2005)). Then, for each edge a two-sided test for the null hypothesis of no correlation is realized. Edges are ranked according to their FDR.
   Graph selection: cut-off on FDR (0.1)

References

Abdolmaleky, H. M., Cheng, K. H., Russo, A., Smith, C. L., Faraone, S. V., Wilcox, M., Shafa, R., Glatt, S. J., Nguyen, G., Ponte, J. F., Thiagalingam, S., and Tsuang, M. T. (2005). Hypermethylation of the reelin (RELN) promoter in the brain of schizophrenic patients: a preliminary report. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 134(1):60–66. *Referred to on page 47.*

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest. *Referred to on page 12.*

Alafuzoff, I., Arzberger, T., Al-Sarraj, S., Bodi, I., Bogdanovic, N., Braak, H., Bugiani, O., Del-Tredici, K., Ferrer, I., Gelpi, E., Giaccone, G., Graeber, M. B., Ince, P., Kamphorst, W., King, A., Korkolopoulou, P., Kovács, G. G., Larionov, S., Meyronet, D., Monoranu, C., Parchi, P., Patsouris, E., Roggendorf, W., Seilhean, D., Tagliavini, F., Stadelmann, C., Streichenberger, N., Thal, D. R., Wharton, S. B., and Kretzschmar, H. (2008). Staging of neurofibrillary pathology in alzheimer's disease: A study of the brainnet europe consortium. *Brain Pathol.*, 18(4):484–496. *Referred to on page 48.*

Alafuzoff, I., Ince, P., Arzberger, T., Al-Sarraj, S., Bell, J., Bodi, I., Bogdanovic, N., Bugiani, O., Ferrer, I., Gelpi, E., Gentleman, S., Giaccone, G., Ironside, J., Kavantzas, N., King, A., Korkolopoulou, P., Kovács, G., Meyronet, D., Monoranu, C., Parchi, P., Parkkinen, L., Patsouris, E., Roggendorf, W., Rozemuller, A., Stadelmann-Nessler, C., Streichenberger, N., Thal, D., and Kretzschmar, H. (2009a). Staging/typing of lewy body related a-synuclein pathology: a study of the BrainNet Europe Consortium. *Acta Neuropathol.*, 117(6):635–652. *Referred to on page 49.*

Alafuzoff, I., Thal, D., Arzberger, T., Bogdanovic, N., Al-Sarraj, S., Bodi, I., Boluda, S., Bugiani, O., Duyckaerts, C., Gelpi, E., Gentleman, S., Giaccone, G., Graeber, M., Hortobagyi, T., Höftberger, R., Ince, P., Ironside, J., Kavantzas, N., King, A., Korkolopoulou, P., Kovács, G., Meyronet, D., Monoranu, C., Nilsson, T., Parchi, P., Patsouris, E., Pikkarainen, M., Revesz, T., Rozemuller, A., Seilhean, D., Schulz-Schaeffer, W., Streichenberger, N., Wharton, S., and Kretzschmar, H. (2009b). Assessment of β-amyloid deposits in human brain: a study of the brainnet europe consortium. *Acta Neuropathol.*, 117(3):309–320. *Referred to on page 49.*

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106. *Referred to on pages 32, 38, 39, 40, 42, 113, and 119.*

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405. *Referred to on page 14.*

Arnold, B. C. and Shavelle, R. M. (1998). Joint confidence sets for the mean and variance of a normal distribution. *Amer. Statist.*, 52(2):133–140. *Referred to on page 15.*

Asimit, J. L., Andrulis, I. L., and Bull, S. B. (2011). Regression models, scan statistics and reappearance probabilities to detect regions of association between gene expression and copy number. *Stat. Med.*, 30(10):1157–1178. *Referred to on page 9.*

Auer, P. L. and Doerge, R. W. (2011). A two-stage Poisson model for testing RNA-Seq data. *Stat. Appl. Genet. Mol. Biol.*, 10:Art. 26, 28. *Referred to on page 32.*

Bandopadhyay, R., Kingsbury, A. E., Cookson, M. R., Reid, A. R., Evans, I. M., Hope, A. D., Pittman, A. M., Lashley, T., Canet-Aviles, R., Miller, D. W., McLendon, C., Strand, C., Leonard, A. J., Abou-Sleiman, P. M., Healy, D. G., Ariga, H., Wood, N. W., de Silva, R., Revesz, T., Hardy, J. A., and Lees, A. J. (2004). The expression of DJ-1 (PARK7) in normal human CNS and idiopathic Parkinson's disease. *Brain*, 127(2):420–430. *Referred to on page 48.*

Bedogni, F., Hodge, R. D., Elsen, G. E., Nelson, B. R., Daza, R. A. M., Beyer, R. P., Bammler, T. K., Rubenstein, J. L. R., and Hevner, R. F. (2010). Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. *P. Natl. Acad. Sci. USA*, 107(29):13129–13134. *Referred to on pages 59 and 60.*

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300. *Referred to on pages 20, 52, 99, and 121.*

Bicciato, S., Spinelli, R., Zampieri, M., Mangano, E., Ferrari, F., Beltrame, L., Cifola, I., Peano, C., Solari, A., and Battaglia, C. (2009). A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Res.*, 37(15):5057–5070. *Referred to on page 9.*

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York. *Referred to on page 71.*

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.*, 107(500):1610–1624. *Referred to on page 72.*

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge. *Referred to on page 12.*

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 52(3):345–370. *Referred to on page 12.*

Braak, H., Bohl, J. R., Müller, C. M., Rüb, U., de Vos, R. A., and Del Tredici, K. (2006). Stanley Fahn Lecture 2005: The staging procedure for the inclusion body pathology associated with sporadic Parkinson's disease reconsidered. *Movement Disord.*, 21(12):2042–2051. *Referred to on page 49.*

Brown, L. D., Cai, T. T., and DasGupta, A. (2003). Interval estimation in exponential families. *Statist. Sinica*, 13(1):19–49. *Referred to on page 15.*

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2):603–618. *Referred to on page 14.*

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-theoretic Approach*. Springer, New York. *Referred to on page 14.*

Carninci, P. (2007). Constructing the landscape of the mammalian transcriptome. *J. Exp. Biol.*, 210(9):1497–1506. *Referred to on page 63.*

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engstrom, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, 38(6):626–635. *Referred to on pages 48 and 51.*

Carvalho, B., Postma, C., Mongera, S., Hopmans, E., Diskin, S., van de Wiel, M. A., van Criekinge, W., Thas, O., Matthäi, A., Cuesta, M. A., Terhaar sive Droste, J. S., Craanen, M., Schröck, E., Ylstra, B., and Meijer, G. A. (2009). Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut*, 58(1):79–89. *Referred to on pages 19, 20, 21, and 83.*

Chari, R., Coe, B. P., Wedseltoft, C., Benetti, M., Wilson, I. M., Vucic, E. A., MacAulay, C., Ng, R. T., and Lam, W. L. (2008). SIGMA2: a system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes. *BMC bioinformatics*, 9:422+. *Referred to on page 27.*

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statistics*, 25:573–578. *Referred to on pages 13 and 14.*

Choy, M.-K., Movassagh, M., Goh, H.-G., Bennett, M., Down, T., and Foo, R. (2010). Genome-wide conserved consensus transcription factor binding motifs are hypermethylated. *BMC Genomics*, 11(1):519. *Referred to on page 62.*

Corduneanu, A. and Bishop, C. (2001). Variational Bayesian model selection for mixture distributions. In *Proceedings of Artificial Intelligence and Statistics*, pages 27–34. *Referred to on page 72.*

Crager, M. R. (2010). Gene identification using true discovery rate degree of association sets and estimates corrected for regression to the mean. *Stat. Med.*, 29(1):33–45. *Referred to on pages 46 and 115.*

Crick, F. (1958). On protein synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163. *Referred to on page 2.*

Cubelos, B., Sebastian-Serrano, A., Beccari, L., Calcagnotto, M. E., Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo, J. M., Bovolenta, P., Walsh, C. A., and Nieto, M. (2010). Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the corte. *Neuron*, 66(4):523–535. *Referred to on page 60.*

Davies, M. N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., Coarfa, C., Harris, R. A., Milosavljevic, A., Troakes, C., Al-Sarraj, S., Dobson, R., Schalkwyk, L. C., and Mill, J. (2012). Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome biology*, 13(6):R43+. *Referred to on pages 60 and 61.*

Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T. H.-M. (2008). The functional consequences of alternative promoter use in mammalian genomes. *Trends in Genetics*, 24(4):167–177. *Referred to on page 63.*

De Hoon, M., Bertin, N., and Chalk, A. (2010). Using cage data for quantitative expression. In Carninci, P., editor, *Cap-Analysis Gene Expression (CAGE): The Science of Decoding Gene Transcription*, pages 101–121. Pan Stanford Publishing, Yokohama. *Referred to on pages 50 and 51.*

DeLong, M. and T., W. (2007). Circuits and circuit disorders of the basal ganglia. *Arch. Neurol.*, 64(1):20–24. *Referred to on page 64.*

Dennissen, F., Kholod, N., and van Leeuwen, F. (2012). The ubiquitin proteasome system in neurodegenerative diseases: Culprit, accomplice or victim? *Prog. Neurobiol.*, 96(2):190–207. *Referred to on page 63.*

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212. *Referred to on page 81.*

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.*, 106(496):1418–1433. *Referred to on page 81.*

Dodd, L. E. and Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics*, 59(3):614–623. *Referred to on pages 86 and 114.*

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigo, R., Birney, E., and Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, 13(9):1–17. *Referred to on page 63.*

Double, K., Halliday, G., Krill, J., Harasty, J., Cullen, K., Brooks, W., Creasey, H., and Broe, G. (1996). Topography of brain atrophy during normal aging and alzheimer's disease. *Neurobiol. Aging*, 17(4):513–521. *Referred to on page 48.*

Double, K., Reyes, S., Werry, E., and Halliday, G. (2010). Selective cell death in neurodegeneration: Why are some neurons spared in vulnerable regions? *Prog. Neurobiol.*, 92(3):316–329. *Referred to on page 48.*

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160. *Referred to on page 37.*

FANTOM Consortium (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563. *Referred to on page 48.*

Faulkner, G. J., Forrest, A. R., Chalk, A. M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D. A., and Grimmond, S. M. (2008). A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by {CAGE}. *Genomics*, 91(3):281–288. *Referred to on page 50.*

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412. *Referred to on page 122.*

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 604–612. *Referred to on page 68.*

Frazee, A., Langmead, B., and Leek, J. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(1):449. *Referred to on page 121.*

Friedman, J. H., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441. *Referred to on pages 68 and 75.*

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22. *Referred to on page 138.*

Frith, M. C., Wilming, L. G., Forrest, A., Kawaji, H., Tan, S. L., Wahlestedt, C., Bajic, V. B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., and Huminiecki, L. (2006). PseudoŰMessenger RNA: Phantoms of the transcriptome. *PLoS Genet.*, 2(4):e23+. *Referred to on page 63.*

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The ucsc genome browser database: update 2011. *Nucleic Acids Res.*, 39(suppl 1):D876–D882. *Referred to on page 53.*

Gao, X., Pu, D. Q., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statist. Sinica*, 22(3):1123–1146. *Referred to on page 68.*

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.*, 30(5):1412–1440. *Referred to on page 81.*

Giraud, C. (2008). Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2:542–563. *Referred to on page 68.*

Gloss, B. S., Patterson, K. I., Barton, C. A., Gonzalez, M., Scurry, J. P., Hacker, N. F., Sutherland, R. L., OŠBrien, P. M., and Clark, S. J. (2012). Integrative genome-wide expression and promoter {DNA} methylation profiling identifies a potential novel panel of ovarian cancer epigenetic biomarkers. *Cancer Letters*, 318(1):76–85. *Referred to on page 62.*

Goldbeter, A. and Pourquié, O. (2008). Modeling the segmentation clock as a network of coupled oscillations in the notch, wnt and {FGF} signaling pathways. *J. Theor. Biol.*, 252(3):574–585. *Referred to on page 58.*

Gouriéroux, C., Holly, A., and Monfort, A. (1982). Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, 50(1):63–80. *Referred to on pages 13 and 14.*

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *J. Stat. Software*, 33(i10):1–31. *Referred to on pages 13 and 14.*

Gu, W., Choi, H., and Ghosh, D. (2008). Global associations between copy number and transcript mRNA microarray data: An empirical study. *Cancer Inform.*, 6:17–23. *Referred to on page 9.*

Hardcastle, T. and Kelly, K. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422. *Referred to on pages 32, 38, and 113.*

Hardy, J. and Selkoe, D. J. (2002). The amyloid hypothesis of alzheimer's disease: Progress and problems on the road to therapeutics. *Science*, 297(5580):353–356. *Referred to on page 48.*

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S., and Guigo, R. (2006). GENCODE: producing a reference annotation for encode. *Genome Biology*, 7(Suppl 1):1–9. *Referred to on page 51.*

Hu, M., Zhu, Y., Taylor, J. M. G., Liu, J. S., and Qin, Z. S. (2012). Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics*, 28(1):63–68. *Referred to on page 46.*

Hughes, A. W. and King, M. L. (2003). Model selection using AIC in the presence of one-sided information. *J. Statist. Plann. Inference*, 115(2):397–411. *Referred to on pages 12 and 13.*

Hunter, L. (2009). *The Processes of Life: An Introduction to Molecular Biology*. The MIT Press, Cambridge, Massachusetts. *Referred to on page 2.*

Iwamoto, K., Bundo, M., Ueda, J., Oldham, M. C., Ukai, W., Hashimoto, E., Saito, T., Geschwind, D. H., and Kato, T. (2011). Neurons show distinctive dna methylation profile and higher interindividual variations compared with non-neurons. *Genome Res.*, 21(5):688–696. *Referred to on pages 60 and 65.*

Jia, H., Osak, M., Bogu, G. K., Stanton, L. W., Johnson, R., and Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, 16(8):1478–1487. *Referred to on page 54.*

Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-seq. *Bioinformatics*, 25(8):1026–1032. *Referred to on page 32.*

Joachim, C., Mori, H., and Selkoe, D. (1989). Amyloid β-protein deposition in tissues other than brain in alzheimer's disease. *Nature*, 341:226–230. *Referred to on page 48.*

Johnson, K., Conn, P., and Niswender, C. (2009a). Glutamate receptors as therapeutic targets for parkinsonŠs disease. *CNS Neurol. Disord. Drug Targets*, 8(6):475–491. *Referred to on pages 64 and 65.*

Johnson, M. B., Kawasawa, Y. I., Mason, C. E., Krsnik, Ő., Coppola, G., Bogdanovic, D., Geschwind, D. H., Mane, S. M., State, M. W., and Łestan, N. (2009b). Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron*, 62(4):494–509. *Referred to on page 47.*

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492. *Referred to on page 65.*

Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.*, 7:486. *Referred to on page 9.*

Kang, H. J. J., Kawasawa, Y. I. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Z., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., Weinberger, D. R., Mane, S., Hyde, T. M., Huttner, A., Reimers, M., Kleinman, J. E., and Sestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489. *Referred to on page 48.*

Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigele, S., Do, H.-H., Weiss, G., Enard, W., Heissig, F., Arendt, T., Nieselt-Struwe, K., Eichler, E. E., and Pääbo, S. (2004). Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.*, 14(8):1462–1473. *Referred to on pages 47 and 48.*

Kodde, D. A. and Palm, F. C. (1986). Wald criteria for jointly testing equality and inequality restrictions. *Econometrica*, 54(5):1243–1248. *Referred to on page 13.*

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nat. Meth.*, 3(3):211–222. *Referred to on pages 49 and 118.*

Konishi, K., Watanabe, Y., Shen, L., Guo, Y., Castoro, R. J., Kondo, K., Chung, W., Ahmed, S., Jelinek, J., Boumber, Y. A., Estecio, M. R., Maegawa, S., Kondo, Y., Itoh, F., Imawari, M., Hamilton, S. R., and Issa, J.-P. J. (2011). DNA methylation profiles of primary colorectal carcinoma and matched liver metastasis. *PLoS ONE*, 6(11):e27889. *Referred to on page 62.*

Konopka, G. and Geschwind, D. H. (2010). Human brain evolution: Harnessing the genomics (r)evolution to link genes, cognition, and behavior. *Neuron*, 68(2):231–244. *Referred to on page 47.*

Krämer, N., Schafer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(1):384. *Referred to on pages 68 and 75.*

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 50:403–418. *Referred to on page 13.*

Lassmann, T., Hayashizaki, Y., and Daub, C. O. (2009). TagDustŮa program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, 25(21):2839–2840. *Referred to on page 50.*

Lê Cao, K.-A. A., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855–2856. *Referred to on page 27.*

Leday, G. G. R. and van de Wiel, M. A. (2013). PLRS: a flexible tool for the joint analysis of DNA copy number and mRNA expression data. *Bioinformatics*, 29(8):1081–1082. *Referred to on page 27.*

Leday, G. G. R., van der Vaart, A. W., van Wieringen, W. N., and van de Wiel, M. A. (2013). Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines. *Ann. Appl. Stat.*, 7(2):823–845. *Referred to on pages 7, 27, 28, and 29.*

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411. *Referred to on page 68.*

Lee, H., Kong, S. W., and Park, P. J. (2008). Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–896. *Referred to on page 9.*

Lee, M. and Kim, Y. (2009). CHESS (CgHExpreSS): a comprehensive analysis tool for the analysis of genomic alterations and their effects on the expression profile of the genome. *BMC Bioinformatics*, 10(1):424+. *Referred to on page 27.*

L'Episcopo, F., Serapide, M., Tirolo, C., Testa, N., Caniglia, S., Morale, M. C., Pluchino, S., and Marchetti, B. (2011). A Wnt1 regulated frizzled-1/β-catenin signaling pathway as a candidate regulatory circuit controlling mesencephalic dopaminergic neuron-astrocyte crosstalk: Therapeutical relevance for neuron survival and neuroprotection. *Molecular Degeneration 6*, 6(1):49–78. *Referred to on page 64.*

Lewin, A., Bochkina, N., and Richardson, S. (2007). Fully Bayesian mixture model for differential gene expression: simulations and model checks. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 36, 28. *Referred to on pages 37 and 44.*

Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *J. Statist. Plann. Inference*, 141(8):2839–2848. *Referred to on page 68.*

Lipson, D., Ben-Dor, A., Dehan, E., and Yakhini, Z. (2004). Joint analysis of DNA copy numbers and gene expression levels. In *Algorithms in bioinformatics*, volume 3240 of *Lecture Notes in Comput. Sci.*, pages 135–146. Springer, Berlin. *Referred to on page 9.*

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statist. Sinica*, 12(1):31–46. Special issue on bioinformatics. *Referred to on page 37.*

Louhimo, R. and Hautaniemi, S. (2011). CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, 27(6):887–888. *Referred to on page 27.*

Louhimo, R., Lepikhova, T., Monni, O., and Hautaniemi, S. (2012). Comparative analysis of algorithms for integration of copy number and expression data. *Nat. Meth.*, 9(4):351–355. *Referred to on pages 85 and 86.*

Lutz, D. and Rufibach, K. (2011). logcondens: Computations related to univariate log-concave density estimation. *J. Statist. Software*, 39(6):1–28. *Referred to on pages 36 and 125.*

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J., Haussler, D., Marra, M. A., Hirst, M., Wang, T., and Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257. *Referred to on page 65.*

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.*, 40(10):4288–4297. *Referred to on pages 32 and 46.*

Meeker, W. Q. and Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *Amer. Statist.*, 49(1):48–53. *Referred to on page 15.*

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462. *Referred to on pages 68, 73, and 75.*

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473. *Referred to on pages 68, 75, and 138.*

Menezes, R., Boetzer, M., Sieswerda, M., van Ommen, G. J., and Boer, J. (2009). Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*, 10(1):203+. *Referred to on page 9.*

Mengel-From, J., Christensen, K., McGue, M., and Christiansen, L. (2011). Genetic variations in the {CLU} and {PICALM} genes are associated with cognitive function in the oldest old. *Neurobiol. Aging*, 32(3):554.e7 – 554.e11. *Referred to on page 54.*

Mercer, T. R., Dinger, M. E., Bracken, C. P., Kolle, G., Szubert, J. M., Korbie, D. J., Askarian-Amiri, M. E., Gardiner, B. B., Goodall, G. J., Grimmond, S. M., and Mattick, J. S. (2010). Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.*, 20(12):1639–1650. *Referred to on page 64.*

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., and Rinn, J. L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotech.*, 30(1):99–104. *Referred to on page 62.*

Metzker, M. L. (2010). Sequencing technologies [mdash] the next generation. *Nat. Rev. Genet.*, 11(1):31–46. *Referred to on page 5.*

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, 38(suppl 1):D204–D210. *Referred to on page 54.*

Miller, C. A. and Sweatt, J. D. (2007). Covalent modification of {DNA} regulates memory formation. *Neuron*, 53(6):857–869. *Referred to on page 47.*

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464:773–777. *Referred to on pages 43 and 121.*

Moon, R. T., Kohn, A. D., Ferrari, G. V., and Kaykas, A. (2004). Wnt and beta-catenin signalling: diseases and therapies. *Nat. Rev. Genet.*, 5(9):691–701. *Referred to on page 58.*

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Meth.*, 5(7):621–628. *Referred to on page 63.*

Naka, H., Nakamura, S., Shimazaki, T., and Okano, H. (2008). Requirement for coup-tfi and ii in the temporal specification of neural stem cells in cns development. *Nat. Neurosci.*, 11:1014–1023. *Referred to on page 59.*

Nemes, S., Parris, T. Z., Danielsson, A., Kannius-Janson, M., Jonasson, J. M., Steineck, G., and Helou, K. (2012). Segmented regression, a versatile tool to analyze mRNA levels in relation to DNA copy number aberrations. *Gene Chromosome Canc*, 51(1):77–82. *Referred to on page 27.*

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P. P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W.-L. L., Stilwell, J. L., Pinkel, D., Albertson, D. G., Waldman, F. M., McCormick, F., Dickson, R. B., Johnson, M. D., Lippman, M., Ethier, S., Gazdar, A., and Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell*, 10(6):515–527. *Referred to on pages 8, 9, 19, 20, 21, 83, 85, and 87.*

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572. *Referred to on page 19.*

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *Amer. Statist.*, 64(2):140–153. *Referred to on pages 69, 71, and 72.*

Oshlack, A., Robinson, M., and Young, M. (2010). From RNA-seq reads to differential expression results. *Genome Biol*, 11(12):220. *Referred to on page 32.*

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L. C., Dahmane, N., and Davuluri, R. V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Research*, 21(8):1260–1272. *Referred to on page 63.*

Pardo, L. M., Rizzu, P., Francescatto, M., Vitezic, M., Leday, G. G. R., Sanchez, J. S., Khamis, A., Takahashi, H., van de Berg, W. D., Medvedeva, Y. A., van de Wiel, M. A., Daub, C. O., Carninci, P., and Heutink, P. (2013). Regional differences in gene expression and promoter usage in aged human brains. *Neurobiol. Aging*, 34(7):1825–1836. *Referred to on pages 39 and 47.*

Pareek, T. K., Belkadi, A., Kesavapany, S., Zaremba, A., Loh, S. L., Bai, L., Cohen, M. L., Meyer, C., Liby, K. T., Miller, R. H., Sporn, M. B., and Letterio, J. J. (2011). Triterpenoid modulation of IL-17 and Nrf-2 expression ameliorates neuroinflammation

and promotes remyelination in autoimmune encephalomyelitis. *Scientific Reports*, 1:1–11. *Referred to on page 54.*

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686. *Referred to on page 71.*

Peng, F., Yao, H., Bai, X., Zhu, X., Reiner, B. C., Beazely, M., Funa, K., Xiong, H., and Buch, S. (2010a). Platelet-derived growth factor-mediated induction of the synaptic plasticity gene arc/arg3.1. *J. Biol. Chem.*, 285(28):21615–21624. *Referred to on page 58.*

Peng, J., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 104(486):735–746. *Referred to on page 68.*

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010b). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, 4(1):53–77. *Referred to on page 9.*

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464:768–772. *Referred to on pages 43 and 121.*

Pink, R. C., Wicks, K., Caley, D. P., Punch, E. K., Jacobs, L., and Francisco Carter, D. R. (2011). Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, 17(5):792–798. *Referred to on page 63.*

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, 37(Suppl):S11–S17. *Referred to on page 8.*

Poon, S., Treweek, T. M., Wilson, M. R., Easterbrook-Smith, S. B., and Carver, J. A. (2002). Clusterin is an extracellular chaperone that specifically interacts with slowly aggregating proteins on their off-folding pathway. *{FEBS} Letters*, 513:259–266. *Referred to on page 63.*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575. *Referred to on page 53.*

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat. Genet.*, 32:496–501. *Referred to on pages 5 and 8.*

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842. *Referred to on pages 51 and 53.*

Rajagopalan, M. and Broemeling, L. (1983). Bayesian inference for the variance components in general mixed linear models. *Comm. Statist. A—Theory Methods*, 12(6):701–723. *Referred to on pages 69 and 71.*

Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, 5(12):e1000598. *Referred to on pages 54, 62, and 132.*

Raney, B. J., Cline, M. S., Rosenbloom, K. R., Dreszer, T. R., Learned, K., Barber, G. P., Meyer, L. R., Sloan, C. A., Malladi, V. S., Roskin, K. M., Suh, B. B., Hinrichs, A. S., Clawson, H., Zweig, A. S., Kirkup, V., Fujita, P. A., Rhead, B., Smith, K. E., Pohl, A., Kuhn, R. M., Karolchik, D., Haussler, D., and Kent, W. J. (2011). ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.*, 39(suppl 1):D871–D875. *Referred to on page 54.*

Rhodes, J., Lutka, F. A., Jordan-Sciutto, K. L., and Bowser, R. (2003). Altered expression and distribution of FAC1 during NGF-induced neurite outgrowth of PC12 cells. *NeuroReport*, 14(3):449–452. *Referred to on page 60.*

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester. *Referred to on pages 13 and 14.*

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3):R25. *Referred to on page 5.*

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140. *Referred to on pages 38, 51, 52, 108, 109, and 113.*

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887. *Referred to on pages 32, 42, 52, and 119.*

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332. *Referred to on pages 52 and 121.*

Roth, R. B., Hevezi, P., Lee, J., Willhite, D., Lechner, S. M., Foster, A. C., and Zlotnik, A. (2006). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, 7(2):67–80. *Referred to on pages 42 and 47.*

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):319–392. *Referred to on pages 32, 45, 69, 71, 82, and 122.*

Salari, K., Tibshirani, R., and Pollack, J. R. (2010). DR-integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics*, 26(3):414–416. *Referred to on pages 9 and 27.*

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studie. *Nat. Rev. Genet.*, 8(6):424–436. *Referred to on pages 48 and 51.*

Schaefer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News*, 6/5:50–53. *Referred to on pages 75 and 138.*

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 32, 28 pp. (electronic). *Referred to on pages 68 and 138.*

Schäfer, M., Schwender, H., Merk, S., Haferlach, C., Ickstadt, K., and Dugas, M. (2009). Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, 25(24):3228–3235. *Referred to on page 9.*

Schonrock, N., Matamales, M., Ittner, L., and Götz, J. (2012). Microrna networks surrounding app and amyloid-beta metabolism–implications for Alzheimer's disease. *Exp. Neurol.*, 235(2):447–454. *Referred to on page 63.*

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference*, 136(7):2144–2162. *Referred to on page 36.*

Scutari, M. (2013). On the prior and posterior distributions used in graphical modelling. *Bayesian Analysis*, 8(1):1–28. *Referred to on page 81.*

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.*, 56(1):49–62. *Referred to on pages 13 and 14.*

Shen, J., Bronson, R. T., Chen, D. F., Xia, W., Selkoe, D. J., and Tonegawa, S. (1997). Skeletal and {CNS} defects in presenilin-1-deficient mice. *Cell*, 89(4):629–639. *Referred to on page 48.*

Silvapulle, M. J. and Sen, P. K. (2005). *Constrained statistical inference*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. Inequality, order, and shape restrictions. *Referred to on page 14.*

Solvang, H., Lingjaerde, O. C., Frigessi, A., Borresen-Dale, A.-L., and Kristensen, V. (2011). Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics*, 12(1):197. *Referred to on pages 24 and 27.*

Soneson, C., Lilljebjorn, H., Fioretos, T., and Fontes, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11(1):191. *Referred to on page 9.*

Starkey, H., Van Kirk, C., Bixler, G., Imperio, C., Kale, V., Serfass, J., Farley, J., Yan, H., Warrington, J., Han, S., Mitschelen, M., Sonntag, W., and Freeman, W. (2012). Neuroglial expression of the MHCI pathway and PirB receptor is upregulated in the hippocampus with advanced aging. *J.Mol. Neurosci.*, 48(1):111–126. *Referred to on page 63.*

Strachan, T. and Read, A. (2010). *Human molecular genetics*. Garland Science Publishing, 4th edition. *Referred to on page 2.*

't Hoen, P. A. C., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H. A. M., de Menezes, R. X., Boer, J. M., van Ommen, G.-J. B., and den Dunnen, J. T. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.*, 36(21):e141. *Referred to on pages 5 and 46.*

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5[prime] end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protocols*, 7(3):542–561. *Referred to on pages 48 and 49.*

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12):2213–2223. *Referred to on pages 38 and 113.*

Thomas, G. M. and Huganir, R. L. (2004). MAPK cascade signalling and synaptic plasticity. *Nat. Rev. Neurosci.*, 5(3):173–83+. *Referred to on page 64.*

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.*, 13(9):2129–2141. *Referred to on pages 52 and 54.*

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., and Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neurosci.*, 14(4):452–458. *Referred to on pages 48 and 63.*

Tschan, M. P., Fischer, K. M., Fung, V. S., Pirnia, F., Borner, M. M., Fey, M. F., Tobler, A., and Torbett, B. E. (2003). Alternative splicing of the human cyclin d-binding myb-like protein (hdmp1) yields a truncated protein isoform that alters macrophage differentiation patterns. *J. Biol. Chem.*, 278(44):42750–42760. *Referred to on pages 63 and 64.*

Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T. T., Tang, M.-H. E., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A., and Carninci, P. (2009). Genome-wide detection and analysis of hippocampus core promoters using deepcage. *Genome Res.*, 19(2):255–265. *Referred to on page 48.*

van de Wiel, M. A., Kim, K. I., Vosse, S. J., van Wieringen, W. N., Wilting, S. M., and Ylstra, B. (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23(7):892–894. *Referred to on pages 11 and 19.*

Van de Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128. *Referred to on pages 31, 68, 70, and 73.*

van de Wiel, M. A., Picard, F., van Wieringen, W. N., and Ylstra, B. (2011). Preprocessing and downstream analysis of microarray DNA copy number profiles. *Briefings in Bioinformatics*, 12:10–21. *Referred to on pages 8 and 28.*

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. *Referred to on page 13.*

van Wieringen, W. N., Belien, J. A., Vosse, S. J., Achame, E. M., and Ylstra, B. (2006). ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data. *Bioinformatics*, 22(15):1919–1920. *Referred to on page 27.*

van Wieringen, W. N., Berkhof, J., and van de Wiel, M. A. (2010). A random coefficients model for regional co-expression associated with DNA copy number. *Stat. Appl. Genet. Mol. Biol.*, 9:Art. 25, 30. *Referred to on page 9.*

van Wieringen, W. N. and van de Wiel, M. A. (2009). Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 65(1):19–29. *Referred to on pages 9, 19, 86, and 88.*

van Wieringen, W. N., van de Wiel, M. A., and Ylstra, B. (2007). Normalized, segmented or called aCGH data? *Cancer Inform.*, 3:321–7. *Referred to on page 24.*

VanAntwerp, J. (2000). A tutorial on linear and bilinear matrix inequalities. *J. Process Contr.*, 10(4):363–385. *Referred to on page 16.*

Vandenberghe, L. and Boyd, S. (1996). Semidefinite programming. *SIAM Rev.*, 38(1):49–95. *Referred to on page 15.*

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, 10(4):252–263. *Referred to on page 59.*

Ventrucci, M., Scott, E. M., and Cocchi, D. (2011). Multiple testing on standardized mortality ratios: a bayesian hierarchical model for fdr estimation. *Biostatistics*, 12(1):51–67. *Referred to on pages 37 and 38.*

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., Hayes, D. N., and Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, 17(1):98–110. *Referred to on page 28.*

Wang, H. and Li, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electron. J. Stat.*, 6:168–198. *Referred to on page 81.*

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63. *Referred to on page 5.*

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Amer. Statist. Assoc.*, 103(481):340–349. *Referred to on page 68.*

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 733–742. Oxford Univ. Press, New York. *Referred to on page 69.*

Winn, J. and Bishop, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.*, 6:661–694. *Referred to on page 71.*

Yang, E., Ravikumar, P., Allen, G., and Liu, Z. (2012). Graphical models via generalized linear models. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1367–1375. *Referred to on page 68.*

Yerbury, J. J., Poon, S., Meehan, S., Thompson, B., Kumita, J. R., Dobson, C. M., and Wilson, M. R. (2007). The extracellular chaperone clusterin influences amyloid formation and toxicity by interacting with prefibrillar structures. *FASEB J.*, 21(10):2312–2322. *Referred to on page 63.*

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286. *Referred to on page 82.*

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35. *Referred to on page 68.*

Yuan, Y., Curtis, C., Caldas, C., and Markowetz, F. (2012). A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Trans Comput Biol Bioinform*, 9(4):947–954. *Referred to on page 81.*

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The `huge` package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.*, 13:1059–1062. *Referred to on page 74.*

Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.*, 12:2975–3026. *Referred to on page 82.*

Acknowledgments

I wish to express my gratitude to Aad van der Vaart and Mark van de Wiel for giving me the opportunity to work on this very interesting project. I am deeply grateful for their trust and guidance during these four years. Their scientific qualities have been an example. I particularly thank Mark for his enthusiasm in working with genomics data, which has been a constant source of motivation.

I would like to thank the many (present and former) people at the Mathematics Department at the VU who made my time enjoyable during this project: Geert, Rikkert, Beata, Bartek, Jakub, Michael, Shota, Tim, Suzanne and Mehran. I am also very grateful for the kind hospitality of the Epidemiology and Biostatistics Department at the VUmc where I was staying weekly. I wish to thank my officemates there: Askar, Nimisha, Viktor and Marloes. I particularly thank Wessel for the nice discussions in which I have learned a great deal. Thank you also for motivating *borrels* and promoting beer time to time. I very much thank Renée for inviting me to the Integration Meeting and for stimulating collaborations. I thank particularly Carel for being so easy to share with and all his kind advices. Also, thank you helping me with the *samenvatting*. Special thanks to Maryke (VU) and Anne-Marie (VUmc) for their hospitality and their constant administrative help.

I wish to thank the members of the reading committee, including Mathisca de Gunst, Stéphane Robin and Ernst Wit, for their interest in this thesis and for spending some of their valuable time on it.

I would not have started this PhD project without the early encouragement of Sue Worner. In January 2008, she warmly welcomed me at Lincoln University in New Zealand to do my Master's thesis and invited me to stay subsequently as a research assistant. She introduced me to the world of research and made me want to explore it further. I am also grateful to the many people there who in a way or another inspired me. I am sure they will recognize themselves in these few lines.

Special thank to the paramaribo crew & co: LN, David, Pedro, João, Mech, Sanne, and all others who simply hung around the flat (pensées speciales aux keush!). I really enjoyed my time there and I miss our random discussions in the kitchen as well as the improvised drinks. To all of you I say... Olive!!! Also, special thank to Micka for showing me Leiden. We miss you here!

I always had a great time seeing friends when I was going back to France or when I was welcoming them in Amsterdam. I really thank my old folks Nono, Jeff, Poussin, Morgane, Koobie, David, Rhum1 and Chanchan, but also Jamin, the Kado brothers, Antereur, Lol and Jygoon for coming over. Thanks a lot to my dear Delghust family: Marc, Jo, Nino and Lila. Thank you for welcoming me each time I was passing by.

Thanks to my lovely Andrea. I would not have made it without you. Thank you for being by my side all along. Thank you for tolerating the excessive amount of time I spent working on this thesis. Thank you for being so understanding, supportive and open. In short, thank you for being the way you are!

To my little one who brings us so much every day (apart from sleep). Your lovely curiosity for the wonders of life simply makes us happy.

Mon dernier mot de remerciements va à ma famille pour leur soutien inconditionel, depuis toujours. Je remercie ma sœur qui, sans être statisticienne, a été la seule à retenir ce que je faisais. Je remercie mes parents de m'avoir toujours soutenu, encouragé et aidé dans mes choix (parfois exotiques). Vous nous avez donné à tous les deux toutes les chances pour réussir. J'espére que vous trouverez dans la rélisation de ce travail une partie de l'aboutissement de vos efforts et de ma profonde reconnaissance.

Gwenaël G.R. Leday
Leiden, January 2014

Summary

The face of biology has changed tremendously with the emergence of technologies that allow for the parallel measurement of thousands of biological sequences (such as DNA, RNA or protein sequences). These technologies (such as microarrays and next-generation sequencing) have produced massive amounts of data that have proven invaluable to researchers in the understanding of (the causes of) complex diseases such as cancer. The surge of 'Big Data' has raised questions regarding their organization and utilization and, hence, has spurred substantive interest in fields such as computer science and statistics.

In this thesis statistical models and inference procedures are developed that address biological questions arising from the analysis of microarray and sequencing data. We particularly treat the subjects of integration of DNA copy number and gene expression data (Chapters 2 and 3), differential gene expression analysis (Chapters 4 and 5), and gene network reconstruction (Chapter 6). Chapter 1 introduces some aspects of molecular biology and experimental molecular data generation in order to aid understanding of the remaining chapters.

In Chapter 2 piecewise linear regression splines is presented as a flexible class of models to decipher how DNA copy number abnormalities in cancer cells alter messenger RNA (mRNA) gene expression levels. This class of models aims to reflect the biological mechanism operating between these two molecular levels and helps in identifying relevant disease markers. We thus utilize piecewise linear regression splines with biologically-motivated parameter constraints to model associations. A novelty in this model is the combined use of copy number data from various (standard) preprocessing steps, namely the continuous *segmented* and discrete *called* data. Because model estimation and selection is difficult in this context, the chapter provides methodology for testing the effect of DNA on mRNA, identifying the appropriate model and obtaining uniform confidence bands that incorporate model uncertainty. Using two real data sets, it is illustrated that flexible models may bring more insight in the interaction between the two molecular levels.

In Chapter 3 the `R` package `PLRS` is presented, which implements the statistical framework introduced in Chapter 2. The package is illustrated with an additional data set from The Cancer Genome Atlas (TCGA). For these data, the need for flexible models is particularly pronounced.

Chapter 4 presents a Bayesian approach to differential gene expression analysis using sequencing (count) data. The method is particularly useful for its large flexibility of the likelihood count model and its ability to handle complex designs. It also accommodates multi-parameter shrinkage for the borrowing of strength in high-dimension.

An novel empirical Bayes procedure for estimating parameters of priors is introduced and different types of (non-)parametric priors are discussed along with Bayesian corrections for multiplicity. The chapter and its appendix present various model- and data-based simulations that validate the performance of the approach in detecting true differences. In particular, compared to other methods, results are shown to be more reproducible on real data.

In Chapter 5 we study differences in gene expression between brain regions in elderly humans. In this work, our contribution lies in the differential expression analysis of cap analysis gene expression (CAGE) data.

Finally, Chapter 6 introduces a computationally attractive Bayesian structural equation model (SEM) for gene network reconstruction. We argue that regularization by means of Gaussian priors coupled with *a posteriori* edge selection is a simple and attractive alternative to sparse priors. A novelty of this work is the use of shrinkage priors that allow the borrowing of strength across regression equations. In simulations, it is demonstrated that the empirical Bayes procedure of Chapter 4 is appropriate in this context and that shrinkage priors can substantially improve graph structure recovery. The Bayesian SEM is also shown to outperform popular sparse methods in various settings.

Samenvatting

Statistische Modellering en Inferentie voor Genomica

Data Integratie, Shrinkage en Netwerk Reconstructie

De aard van biologisch onderzoek is enorm veranderd met de opkomst van technologieën die de parallelle meting van duizenden biologische sequenties (zoals DNA, RNA of proteïne sequenties) mogelijk maken. Deze technologieën (zoals microarrays en next-generation sequencing) produceren massieve hoeveelheden data welke van onschatbare waarde blijken voor het begrijpen van (de oorzaken van) complexe ziekten zoals kanker. Deze golf aan 'Big Data' heeft vragen opgeworpen over de organisatie en het gebruik van massieve datasets. Deze vragen hebben geleid tot hernieuwde, biologisch-gemotiveerde interesse in informatica en statistiek.

In dit proefschrift worden statistische modellen en inferentie-procedures ontwikkelt voor de biologische vraagstukken die voortkomen uit microarray en next-generation sequencing data. In het bijzonder komen de volgende onderwerpen aan bod: de integratie van DNA copynumbervariatie en genexpressie data (hoofdstukken 2 en 3), de analyse van differentiële genexpressie (hoofdstukken 4 en 5), de reconstructie van genexpressie netwerken (hoofdstuk 6). Hoofdstuk 1 introduceert enkele aspecten uit de moleculaire biologie en moleculaire-data productie om het begrip van de volgende hoofdstukken te vergroten.

In hoofdstuk 2 worden piecewise linear regression splines gepresenteerd als een flexibele klasse van modellen om te ontcijferen hoe DNA copynumbervariaties veranderingen teweeg brengen in de expressie van messenger RNA (mRNA) in kankercellen. Deze klasse van modellen reflecteert expliciet het biologische mechanisme onder de twee genoemde moleculaire niveaus en is behulpzaam bij de identificatie van relevante biomarkers. De implementatie van piecewise linear regression splines maakt gebruik van biologisch gemotiveerde restricties op de model-parameters. Een noviteit van dit gerestricteerde model is het gecombineerde gebruik van meerdere, uit de datapreparatie voortkomende, DNA copynumbervariatie datatypes, namelijk: continue *gesegmenteerde* alsook discrete *called* data. Hoofdstuk 2 verstrekt dan de methodologie voor: het testen van het effect van DNA copynumbervariaties op mRNA expressie, het identificeren van het best passende gerestricteerde model, het verkrijgen van uniforme confidence bands waarin modelonzekerheid is opgenomen. Aan de hand van twee echte datasets wordt geïllustreerd hoe de flexibiliteit van de voorgestelde modellen meer inzicht oplevert in de interactie tussen DNA copynumbervariatie en mRNA genexpressie.

In hoofdstuk 2 wordt het `R` pakket `PLRS` gepresenteerd, welke de methoden uit hoofdstuk 2 implementeert. Het pakket wordt geïllustreerd met additionele data van The Cancer Genome Atlas (TCGA). De behoefte aan de flexibiliteit die dit pakket biedt is bijzonder uitgesproken voor deze data.

Hoofdstuk 4 ontwikkelt een Bayesiaanse aanpak voor de analyse van differentiële genexpressie op basis van next-generation sequencing data. De methode is vooral van nut door zijn flexibele omgang met de waarschijnlijkheidsfunctie en zijn vermogen complexe designs te verwerken. Het raamwerk incorporeert ook 'shrinkage' in de zin dat de parameterschattingen worden verbeterd door het gebruik van empirisch gemotiveerde *prior* verdelingen. Verder geeft dit hoofdstuk een nieuwe empirical Bayes procedure voor het schatten van de hyperparameters van prior verdelingen en bespreekt het verschillende soorten (niet-)parametrische priors alsook Bayesiaanse multipliciteitscorrecties. Verscheidene model- en data-gebaseerde simulaties valideren de aanpak met betrekking tot de detectie van differentiële genexpressie. Een vergelijking met andere methoden op echte data laat zien dat de resultaten verkregen met de voorgestelde methode beter reproduceerbaar zijn.

In hoofdstuk 5 bestuderen we verschillen in genexpressie tussen hersengebieden bij senioren. Onze bijdrage ligt in de analyze van differentiële genexpressie op basis van cap analysis gene expression (CAGE) data.

Tenslotte introduceert hoofdstuk 6 een computationeel aantrekkelijk Bayesiaans structureel vergelijkingsmodel voor de reconstructie van genexpressie netwerken. We stellen dat regularisatie door middel van Gaussische priors in combinatie met *a posteriori* zijde-selectie een simpel en aantrekkelijk alternatief is voor het gebruik van spaarzame priors. De vernieuwing ligt hier in het gebruik van shrinkage priors die informatie 'lenen' uit de verschillende regressievergelijkingen van het structurele model. Simulaties tonen aan dat de empirical Bayes procedure van hoofdstuk 4 ook in deze context gebruikt kan worden en dat het gebruik van shrinkage priors superieure netwerk reconstructie oplevert. Simulaties tonen ook aan dat het voorgestelde Bayesiaanse structureel vergelijkingmodel in verscheidene situaties beter presteert dan populaire spaarzame methoden.

Résumé

Modélisation et Inférence Statistique pour la Génomique

Intégration de Données, Shrinkage et Reconstruction de Réseaux

Le visage de la biologie a énormément changé avec l'émergence de technologies qui permettent de mesurer en parallèle des milliers de séquences biologiques (telles que les séquences d'ADN, d'ARN ou de protéines). Ces technologies (comme par exemple les puces à ADN ou le séquençage haut débit) ont produit d'énormes quantités de données qui se sont avérées indispensables pour les chercheurs dans la compréhension de maladies complexes telles que le cancer. L'organisation et l'utilisation de ces données massives sont très tôt devenues un défi, et par conséquent, ont soulevé un intérêt certain dans des domaines tels que l'informatique et la statistique.

Dans cette thèse, nous développons des modèles et procédures d'inférence statistiques qui répondent à des questions biologiques soulevées par l'analyse de données de puces à ADN et de séquençage. Nous abordons notamment l'intégration des données du nombre de copies d'ADN et d'expression génique (chapitre 2 and 3), l'analyse différentielle de l'expression des gènes (chapitre 4 and 5) et la reconstruction de réseaux géniques (chapitre 6). Le chapitre 1 introduit certains aspects de biologie moléculaire et les différents types de données expérimentales afin d'apporter une base nécessaire à la compréhension des autres chapitres.

Au chapitre 2, la régression par splines linéaires est présentée comme une classe flexible de modèles pour décrire la façon dont le nombre de copies d'ADN dans les cellules cancéreuses modifient le niveau d'expression des gènes, c'est à dire les quantités d'ARN messagers (ARNm). Cette classe de modèles vise à refléter les mécanismes biologiques entre ces deux niveaux moléculaires et identifier les marqueurs importants de la maladie. Pour modéliser ces associations, nous utilisons donc la régression par splines linéaires et imposons des contraintes sur les paramètres pour améliorer l'interprétation biologique. La particularité principale de ce model est l'utilisation conjointe de différents types de données (standards) du nombre de copies d'ADN issues des différentes étapes de prétraitement, à savoir les données *segmentées* et *discrétisées*. Puisque l'estimation et la sélection de modèle est difficile dans ce contexte, le chapitre décrit comment tester l'effet de l'ADN sur l'ARNm, identifier le modèle le plus approprié et obtenir des intervals de confiance pour la fonction de régression tout en prenant en compte l'incertitude du modèle choisit. Sur deux jeux de données réels, nous illustrons la pertinence de ce type de model pour décrire l'intéraction entre les deux marqueurs.

Au chapitre 3, nous présentons le package `PLRS` pour `R`, qui met en œuvre le cadre statistique introduit dans le chapitre 2. La méthode est illustrée sur un jeu de données supplémentaire issue du The Cancer Genome Atlas (TCGA). Pour ces données, la nécessité d'une une classe flexible de modèles est particulièrement prononcée.

Au chapitre 4, nous développons une approche bayésienne pour l'analyse différentielle de l'expression des gènes à partir de données de séquençage (comptage). La méthode est particulièrement utile au vu de la flexibilité du modèle de comptage et de sa capacité à prendre en compte des designs expérimentaux complexes. Elle permet également la régularisation de multiple paramètres pour améliorer l'estimation en grande dimension. Nous présentons une nouvelle procédure d'estimation bayésienne empirique pour les paramètres des lois *a priori* et discutons différents types de lois (non-) paramétriques ainsi que l'approache bayésienne du problème de comparaison multiples. Le chapitre et son appendice contiennent de nombreuses simulations réalisées à partir de modèles statistiques et de données reelles. Celles-ci valident la performance de la méthode pour la détection de vraies différences. En particulier sur données réelles, la reproductibilité des trouvailles semble meilleure que les autres méthodes.

Au chapitre 5, nous étudions les différences d'expression géniques entre plusieurs régions du cerveau chez des personnes âgées. Dans ce travail, notre contribution réside dans l'analyse différentielle de l'expression génique en utilisant les données de CAGE.

Enfin, au chapitre 6, nous introduisons un modèle bayésien d'équations structurelles (MBES) pour la reconstruction de réseaux gèniques. Nous argumentons que la régularisation au moyen de lois *a priori* gaussiennes avec une sélection des arêtes *a posteriori* est une alternative simple et attrayante face aux lois dites 'sparse'. Une nouveauté de ce travail réside aussi dans l'utilisation de lois *a priori* qui permettent de mettre en commun, pour chaque équation, l'estimation de certains paramètres. Nous montrons à l'aide de simulations que la procédure bayésienne empirique du chapitre 4 est appropriée dans ce contexte et que ce type de lois *a priori* peuvent améliorer sensiblement la recouvrement de la structure du graphe. Le MBES apparait, dans divers cas, être supérieur à certaines méthodes 'sparses' couramment utilisées.