STATISTICAL METHODS FOR THE ANALYSIS OF LARGE-SCALE GENOMIC AND PROTEOMIC DATA

TEO GUO SHOU

NATIONAL UNIVERSITY OF SINGAPORE

2015

STATISTICAL METHODS FOR THE ANALYSIS OF LARGE-SCALE GENOMIC AND PROTEOMIC DATA

TEO GUO SHOU

(B.Sc.(Hons.),NTU)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2015

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Teo Guo Shou

October 2015

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Assistant Professor Choi Hyungwon for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge.

Contents

Summary

Modern high-throughput technologies have undergone continuous innovation, enabling biologists and clinical scientists to detect functional molecules such as proteins and metabolites and measure their concentrations in a scale and coverage that had hardly been imagined a decade ago. Contemporary molecular biologists are taking advantage of these technological advances to move toward the new heights: systems biology. Analysis of the resulting data in systems biology investigations, generated from the state-of-the-art omics technologies, is not trivial and it can no longer be framed in the conventional line of statistical analysis pipeline, heavily dominated by multiple hypothesis testing. Capitalizing on an array of statistical techniques to model the high-dimensional data, especially from the Bayesian literature, the thesis presents two novel statistical methods that provide efficient data analysis solution for systems biologists.

We first present a method termed mapDIA, a comprehensive statistical software suite for data preprocessing and model-based differential expression analysis for quantitative proteomics data generated by data independent acquisition (DIA) mass spectrometry (MS), a newly rising platform of choice in proteomics that can generate tandem mass spectrometry (MS/MS) data for an unbiased set of peptides. The analysis is based on a hierarchical Bayesian latent variable model with Markov random field prior with fast estimation and inference.

Next we developed a method to analyze dynamic gene expression regulation in time course transcriptomic and proteomic data sets to quantitatively dissect

the contribution of RNA-level and protein-level regulation to the variation in gene expression. The Bayesian statistical method embodies a mass action-based model for protein synthesis and degradation rates of individual genes, and allows statistical inference of change points in the stochastic process of the kinetic parameters, leading to characterization of gene expression regulation patterns in the time course profiles. A reversible-jump MCMC sampler is used to compute the posterior distribution, and the performance of the method has been evaluated using both simulated and real experimental data.

The two methodologies tackle statistical problems with distinct biological applications and data complexity, yet both approaches embody hierarchical Bayes inference of high dimensional models for data sets. Through these developments, we showcase exemplary solutions to derive biologically sensible interpretation of data in non-trivial computational problems posed in systems biology applications.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical challenges in genomic and proteomic data analysis

Differential expression analysis using -omics data, such as gene expression microarrays for mRNA transcripts, DNA copy numbers and methylation, etc, has garnered a large amount of methodologies in modern statistics literature [10, 15, 16, 24, 30, 50, 61]. During the early advances in the late 1990s, many statisticians have viewed this as multiple hypothesis testing problem and seized the opportunity to develop mathematical solutions to control the *overall* type I error to a reasonable low bound, i.e. *on average*. This is best exemplified by the array of methods to control the family wise error rates or false discovery rates [4, 18, 64], which is now considered as the most important contributions of statistics to the field of modern biology.

Another important feature in the high-throughput -omics data is the so-called $n \ll p$ problem: small sample size. The dimension of whole genome-wide data sets easily hovers over 50,000, as most technologies such as DNA-seq and RNA-seq now cover all known transcripts of genes and the resulting protein isoforms. By contrast, typical group comparisons in non-clinical, molecular biology problems are based on sample size no greater than several replicates per group. Hence as the technologies are further refined and more specific biological hypotheses are tested, far from those model "cancer" data sets

in clinical -omics featuring at least a few tens of samples, the $n \ll p$ problem only deepens. Even in these extreme examples of controlled experiments, the procedures to deal with the false discoveries in multiple testing problems become vulnerable: the base measure of statistical significance, p-value, cannot be calculated very accurately for various reasons. For example, many low abundance molecules are known to be prone to measurement inaccuracy because of masking by larger and abundant contaminants, and this hinders identification of robust null hypotheses. Hence the robustness of all procedures that adjust the raw p-values to a higher stringency level (to account for multiple testing) becomes undermined.

The contribution of Bayesian statistics to this field lies in addressing this issue with hierarchical Bayes inference [51, 57, 58, 67], or the equivalent approach in the form of empirical Bayes [18, 19, 43, 44, 46, 66]. By formulating probabilistic models on the data so that the gene specific model parameters, such as mean and variance of log expression data, follow a global distribution (i.e. prior), the estimation of parameters in an individual gene naturally incorporates the locale and variability in other genes. Although this introduces the shrinkage of estimates (e.g. effect sizes such as log fold change) towards the genome-wide mean and thus incurs bias, it renders the estimates robust against outlier observations and genuine heterogeneity in biological samples.

However, this thesis is motivated by a few additional challenges that have yet to be addressed by statistical modeling, which can be summarized: (i) the increasing complexity in the data structure and (ii) that of biological question being asked. First, the existing statistical methods make the implicit assumption that each molecule, such as messenger RNA or protein, is quantified into a single concentration measurement. However, the most widely used technologies such as next generation sequencing (NGS) and mass spectrometry (MS) produce the quantitative data summarized at multiple levels. In the case of NGS applied for mRNAs, each gene can be quantified for annotated transcripts, i.e. assembly of multiple protein-coding regions (exons). In the case of MS, proteins are detected and

quantified at the level of peptides or their fragments, which are enzymatically digested fragments of the intact proteins. In other words, these technologies generate the data at much deeper level, which can effectively be considered as *repeated measure* of target molecules.

Second, the biological questions being asked using these technologies are becoming more complex. Integrative -omics, referring to studies utilizing more than one -omics technologies to study the dynamics or association between different molecular levels, is becoming increasingly common. One such area is gene expression regulation at the transcriptional and translational levels, i.e. the kinetic changes in the regulatory parameters for mRNA synthesis and degradation, and those for protein synthesis and degradation. This is indeed an essential question in the systems biology: how do cells find new homeostatic equilibrium in response to environmental stress? To answer this question, one must measure concentration changes over a time course at both mRNA and protein levels, monitoring the input and the output in the central dogma of molecular biology. Analysis of the resulting dataset, especially phrased in terms of appropriate kinetic parameters of synthesis and degradation, is not trivial and cannot be formulated as simple hypothesis testing problem.

These two examples are merely a snippet of much bigger complexity to come in contemporary systems biology investigations. More complex biological questions, along with the quantitative data acquired with ever increasing resolution, will require robust probabilistic modeling strategies in the near future. In this context, the modeling framework offered by Bayesian methods, with its powerful sampling-based inference, is an attractive and reliable toolbox, and this thesis provides an early landscape of those applications.

1.2 Related statistical methodologies

Before we introduce our methods to address these challenges, we first review the substantial amount of literature on the application of Bayesian hierarchical modeling in high-throughput genomic data analysis, such as for multiple testing correction [18, 48] and differential expression analysis [2, 33, 34, 58, 65].

One of the first empirical Bayesian analysis of expression data was published by Newton et al. (2001) [47]. The authors worked on preprocessed, two-channel microarray data, highlighting the drawbacks of the naive fold change estimator R/G, obtained from each gene's intensity measurements R (red) and G (green) in the two color channels on a spotted cDNA microarray. The proposed empirical bayes estimate of fold change was $(R+\nu)/(G+\nu)$ where the value of the statistic ν depends on sources of variation affecting the intensity measurements and is calculated from data on the gene set. The authors showed by simulation how this estimator has improved the ranking of genes and reduced mean squared error. The authors also presented a bayesian hierarchical mixture model to address the problem of testing for significant differential expression. The simulation results seem to show the effectiveness of this bayesian hierarchical model that does not require Markov chain Monte Carlo methods. Similarly, our bayesian hierarchical model in chapter 2 will rely on carefully chosen priors to avoid Markov chain Monte Carlo sampling. The authors further extended the method to allow replicate expression profiles in multiple mRNA populations in Kendziorski et al. (2003) [35].

Efron et al. (2001) [18] proposes a nonparametric empirical bayesian mixture model to analyse differential expression. This model eliminates the problems associated with parametric modeling. This analysis makes use of permutation to estimate a null distribution, and takes advantage of the large number of genes for the nonparametric density estimate. However, it can be problematic when there is an insufficient number of replicate (microarray) plates. Further, this paper was the first to associate a gene's posterior

probabilities of equal expression to the rates of false discovery in a gene set.

Newton et al. (2004) [48] describes the dual role of posterior probabilities in the context of multiple testing for differential gene expression. Consider gene g from a large set of genes. By applying Bayesian or empirical Bayesian methods in the analysis of the gene set will yield the posterior probability of gene g being equally expressed, β_g. Genes that are evidently differentially expressed will have the smallest values of β_g and get on the list of discoveries. The duality is that a small β_g is the ticket with which gene g gets on this list; at the same time β_g is the chance that the placement of gene g on the list is a false discovery.

Consider a list of discoveries containing all genes having values β_g less than some bound κ. Given the data, the expected number of false discoveries is

$$C(\kappa) = \sum_g \underbrace{\beta_g}_{\text{error rate}} \underbrace{1[\beta_g \leq \kappa]}_{\text{discovery}}$$

since β_g is the conditional probability that placing gene g on the list creates a type I error. The false discovery rate is $\frac{C(\kappa)}{\sum_g 1[\beta_g \leq \kappa]}$, with $\sum_g 1[\beta_g \leq \kappa]$ being the size of the list. Using the direct posterior probability approach [48], $\kappa \leq 1$ is set as large as possible satisfying $\frac{C(\kappa)}{\sum_g 1[\beta_g \leq \kappa]} \leq \alpha$ so that we can obtain the largest possible list of discoveries while bounding the rate of false discoveries by α. This idea is not new but is evident in the recent and fruitful literature on FDR. In Efron et al. (2001) [18], β_g was called the local false discovery rate because it measured the conditional type I error rate for gene g; the ranking of genes by β_g and the formation of a gene list with level α FDR would give the same thing as if the Storey (2003) [64] q-value method was applied to the β_g themselves and if we formed the list of genes for which these q-values are bounded by α. The direct posterior probability approach will be adopted in the next two chapters.

Wei and Li (2007) [74] describes the use of Markov random field [36] for identifying the subnetworks that show differential expression patterns between two conditions by utilizing

the network structure information. Consider the problem of identifying genes which are differentially expressed given microarray gene expression profiling data under two experimental conditions. Let $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{im}; y_{i(m+1)}, y_{i(m+2)}, \ldots, y_{i(m+n)})$ denote the observed mRNA expression level of gene i composed of the first m replicates under condition 1 and the next n replicates under condition 2. The hypothesis test of interest is

$$H_{i0} : \mu_{1i} = \mu_{2i},$$

where μ_{ki} is the mean expression level of the ith gene under condition k. Let x_i be the state of gene i,

$$x_i = \begin{cases} 1 & \text{if gene } i \text{ is differentially expressed } (\mu_{1i} \neq \mu_{2i}) \\ 0 & \text{otherwise } (\mu_{1i} = \mu_{2i}). \end{cases}$$

Using the gamma-gamma model with Gamma(shape=α, scale=$\frac{\alpha}{\mu_i}$) as the likelihood and setting the prior of $\frac{\alpha}{\mu_i}$ as Gamma(shape=α_0, scale=ν) for gene expression data [35, 47], the conditional density for gene i is

$$\begin{aligned} f(\mathbf{y_i}|x_i = 1) =& K_1 K_2 \frac{(\prod_{j=1}^{m+n} y_{ij})^{\alpha-1}}{(\nu + y_{i.m})^{m\alpha+\alpha_0}(\nu + y_{i.n})^{n\alpha+\alpha_0}}, \\ f(\mathbf{y_i}|x_i = 0) =& K \frac{(\prod_{j=1}^{m+n} y_{ij})^{\alpha-1}}{(\nu + y_{i.m} + y_{i.n})^{(m+n)\alpha+\alpha_0}}, \end{aligned}$$

where

$$
\begin{aligned}
y_{i.m} =& \sum_{j=1}^{m} y_{ij}, \\
y_{i.n} =& \sum_{j=m+1}^{m+n} y_{ij}, \\
K_1 =& \frac{\nu^{\alpha_0}\Gamma(m\alpha+\alpha_0)}{\Gamma^m(\alpha)\Gamma(\alpha_0)}, \\
K_2 =& \frac{\nu^{\alpha_0}\Gamma(n\alpha+\alpha_0)}{\Gamma^n(\alpha)\Gamma(\alpha_0)}, \\
K =& \frac{\nu^{\alpha_0}\Gamma((m+n)\alpha+\alpha_0)}{\Gamma^{m+n}(\alpha)\Gamma(\alpha_0)}.
\end{aligned}
$$

Note that the model parameters are α, the shape parameter of the gamma distribution of the gene expression level and (α_0, ν) the shape and scale parameters of the gamma prior of $\frac{\alpha}{\mu_i}$. Using this model will require α to be fixed at some "suitable values". In Chapter 2, we will use the log normal-normal inverse gamma model as it allows for all likelihood parameters to be assigned a prior and thus avoiding the need to plug in an arbitrary value. The authors applied their model to two real datasets and demonstrated that the procedure is more sensitive in identifying the differentially expressed genes than those procedures that do not utilize the pathway structure information.

1.3 Outline

As stated earlier, this thesis consists of two chapters, each addressing statistical problems associated with different experiments.

Chapter 2 adapted from Teo *et al* [68] describe a software package mapDIA for statistical analysis of differential expression using tandem mass spectrometry (MS/MS) fragment-level quantitative data. mapDIA offers a series of tools for essential data preprocessing, including a novel retention time-based normalization method and multiple

peptide/fragment selection steps. Using the preprocessed data, mapDIA provides hierarchical model-based statistical significance analysis for multi-group comparisons under representative experimental designs.

Chapter 3 adapted from Teo *et al* [69] presents a statistical method based on a mass action-based model for protein synthesis and degradation rates of individual genes, and change points in the stochastic process of the kinetic parameters are derived to identify distinct patterns of regulation of gene expression in time course profiles. A sampling-based inference procedure using Markov chain Monte Carlo is implemented and the posterior probabilities of change points in the ratio of protein synthesis and degradation are used to control the Bayesian false discovery rate.

Chapter 2

Preprocessing and Statistical Analysis of Quantitative Proteomics Data from Data Independent Acquisition Mass Spectrometry

2.1 Significance analysis of quantitative proteomic data

The data dependent acquisition (DDA) mode of analysis is the prevailing platform of mass spectrometry-based[1] shotgun[2] proteomics[3] In the DDA mode, more abundant precursor peptide[4] ions[5] are preferentially isolated and fragmented to generate MS/MS[6] spectra.

[0]Adapted with permission from Teo *et al*, "mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry ," *J. Proteomics*, 2015; Copyright (2015) Elsevier.

[1]Mass spectrometry: A technique for determining the exact mass of every peptide present in a sample of purified protein or protein mixture.

[2]Shotgun proteomics identifies proteins from tandem mass spectra of their proteolytic peptides.

[3]Proteomics: The large-scale study of the structure and function of proteins.

[4]Proteins are long polymers of amino acids, peptides are shorter, usually fewer than 50 amino acids long.

[5]Ion: An atom or molecule carrying an electrical charge, either positive or negative.

[6]MS/MS (tandem mass spectrometry): This combines two mass spectrometers: one (MS1) for the detection and selection of precursor ions, which is followed by a second (MS2) for the analysis of fragment

These MS/MS spectra are then computationally analyzed to identify the peptides and to infer the corresponding proteins[7]. In this strategy, peptides are quantified using the intensity of the precursor peptide signal detected in the first stage of MS analysis (MS1 intensity). A well-known limitation of the DDA strategy is that precursor selection is systematically biased in favor of more abundant peptides, which results in inconsistent quantification of lower abundance peptides across multiple samples. This is particularly a problem in complex samples where the number of co-eluting species to be sequenced exceeds the duty cycle of the mass spectrometer.

An alternative mode of analysis, called data independent acquisition (DIA), has the potential to provide more consistent peptide quantification. In the currently favored DIA set-ups, the entire mass range relevant to the experimentalist is covered using a set of wide sequential windows, which allows segmented acquisition of MS/MS fragment ion spectra for an unbiased set of precursors. All precursor peptide ions within each window are co-isolated and subjected to fragmentation to produce multiplex MS/MS spectra. Although DIA had been initially proposed nearly a decade ago [49, 71], it was not until recently that significant advances in the scan speed and the accuracy of mass measurements enabled practical implementations of this strategy. One commonly used DIA strategy, SWATH-MS, was first implemented on a Qq-TOF AB SCIEX instrument using a sequence of 25 m/z-wide precursor isolation windows (see figure 2.1)[27], and related methods are now available on MS instruments from other manufacturers, including on the Thermo Fisher Q Exactive system. A variant of this strategy, called MSX, uses a stochastic selection of smaller (e.g. 4 m/z wide) precursor isolation windows and has been shown to reduce the fragment ion interference and increased precursor selectivity [20].

Because virtually every peptide ion is selected for fragmentation, DIA theoretically allows

---

ion spectra generated from selected precursor ions after collision-induced fragmentation. The information from the fragment ion spectra is used for peptide identification.

[7]Protein: Polymer built from amino acids that provides cells with their shape and structure and performs most of their activities.

more consistent peptide detection and quantification across multiple samples resulting in more complete quantitative coverage (i.e., less missing data). In addition, DIA data changes the way quantitative data are analyzed compared to the traditional quantitative

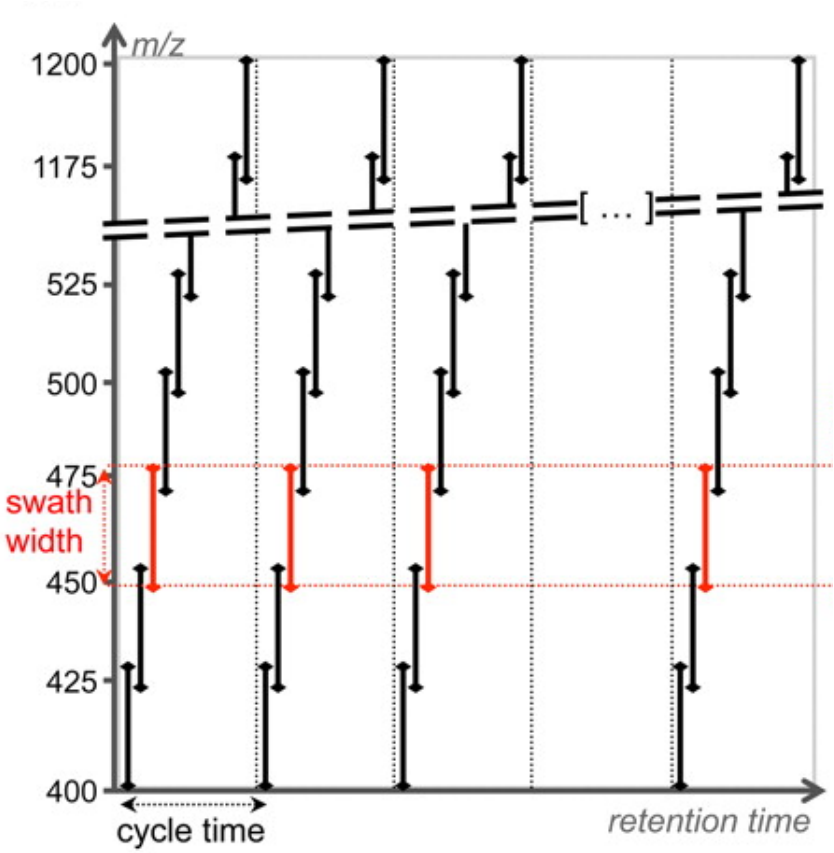

Figure 2.1: SWATH MS data-independent acquisition: the data-independent acquisition method consists of the consecutive acquisition of high resolution, accurate mass fragment ion spectra during the entire chromatographic elution (retention time) range by repeatedly stepping through 32 discrete precursor isolation windows of 25-Da width (black double arrows) across the 400-1200 m/z range. The series of isolation windows acquired for a given precursor mass range and across the LC is referred to as a "swath" (e.g., series of the red double arrows). The cycle time is defined as the time required to return to the acquisition of the same precursor isolation window. Note that the dotted line before the beginning of each cycle depicts the optional acquisition of a high resolution, accurate mass survey (MS1) scan. Adapted with permission from Gillet *et al*, " Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis," *Mol. Cell. Proteomics*, 2012; 11(6):1753-68. Copyright (2012) The American Society for Biochemistry and Molecular Biology.

DDA proteomics analysis. The volume of quantitative information in the DIA data is considerably larger than that of the DDA data, since the intensity data can be extracted not only at the peptide level from MS1 data but also at the MS/MS fragment level (MS2 quantification). The current approaches for DIA data analysis, however, often do not take full advantage of this extended (fragment level) data and instead aggregate intensities from fragments to peptide intensities or even protein intensities. At the same time, the fragment intensity data can be used as an extra layer of valuable information in the sense that fragment intensities serve as "repeated measures" of the intensity of their parent peptides. From a statistical point of view, these data immediately create the opportunity to improve statistical significance analysis from those approaches designed for MS1 peptide intensity data, since the fragment intensity data provide clues for the reliability or reproducibility of relative quantification as long as they are *on average* faithful to the quantitative level of their parent peptides across the samples being compared. In other words, there are much more data to work with to draw inferences for protein expression changes per protein basis in the DIA data.

Nevertheless, the complexity of the DIA data poses numerous challenges to its extraction and analysis. At present, the default data analysis strategy for DIA data is targeted quantification using tools such as OpenSWATH [55] and Skyline [42], which depend on spectral libraries obtained from DDA experiments. This requirement for external spectral libraries is however not absolute, and can be alleviated using, for example, the new computational workflow DIA-Umpire that enables untargeted and semi-targeted identification and quantitative extraction [70]. In either case, the MS2 DIA data may contain fragments that are shared across multiple co-eluting precursor ions within the same isolation window, creating a difficult problem for quantification. Furthermore, after data extraction for each sample, the fragment maps will not necessarily be reproducible across multiple samples due to fragment ion interference and other sources of noise, and therefore a reliable set of fragments has to be selected carefully before the statistical

analysis is performed.

These challenges have direct ramifications for statistical analysis of large DIA datasets. Figures 2.2 and 2.3 demonstrate this challenge through real examples of fragment intensity data in the 14-3-3β dynamic interactome dataset we will analyze later. In these figures, the intensity data from a time course experiment with three biological replicates were transformed into natural log scale and centered by median within each replicate in each fragment. Figure 2.2 shows the examples of reliably extracted fragment intensity data in which most fragments from these peptides are well correlated with one another and faithfully represent their parent protein abundance (unknown yet can be inferred). By contrast, Figure 2.3 shows the other side of the reality. Here, MYCBP2 and YWHAB (14-3-3β/α) contain a large number of peptides with sufficient fragment intensity data, yet they both suffer from serious lack of reproducibility within each protein across peptides. In the case of CYB5R3, the reproducibility within and across samples is fair, yet there are only two peptides to draw our statistical inference for each protein. The alarming fact is that these cases are ubiquitous in all SWATH-MS datasets we have analyzed and are unrelated to the method of extraction. Because different types of challenging cases (non reproducible peptides; too little data) are simultaneously present in a single dataset, careful post-extraction processing is rapidly becoming a necessity, especially to preclude spurious findings to percolate through the final stage of statistical significance analysis.

In this chapter, we present the first comprehensive software package specifically designed for the fragment intensity data generated in the DIA mode, which tackles the challenges in two stages: preprocessing and statistical modeling. Most existing statistical software tools for quantitative proteomics data analysis are amenable for protein or peptide intensity data, but not fragment intensity data. For example, the DANTE software package offers regression model-based analysis of peptide intensity data [52]. The MaxQuant-Perseus packages enable protein quantification via the LFQ (label-free quantification) or iBAQ (intensity-based, absolute quantification) values and subsequent statistical

analysis of these data [12]. The DIA-Umpire tool, specifically developed for DIA data, implements several approaches for selecting most reproducible fragments and peptides as part of its procedure for computing protein-level quantification. MSstats (version 2.3.4) is currently the only statistical software capable of differential expression analysis using the fragment intensity data, since it was originally written for the S/MRM (selected/multiple reaction monitoring)[8] data [8]. However, whether the regression-based framework currently implemented in MSstats is adaptive to far more complex DIA data has not been rigorously examined to date. In particular, as illustrated in Figures 2.2 and 2.3, the fragment intensities can vary significantly between (co-)isolated peptide precursors within the same protein, which may expose the regression model to erroneous quantification and resulting false discoveries more easily than the S/MRM data that uses specifically isolated transitions that have been carefully selected by the experimentalists. In light of these issues and with the number and scope of DIA studies rapidly expanding, it is therefore of great importance to evaluate the existing options and develop new tools, if necessary, which will render the statistical significance analysis of fragment-level intensity data as robust as possible.

[8]Selected reaction monitoring: This is a sensitive mass spectrometry-based method for targeted proteomics that is based on the measurement of precursor-fragment ion pairs (transitions) of proteotypic peptides.

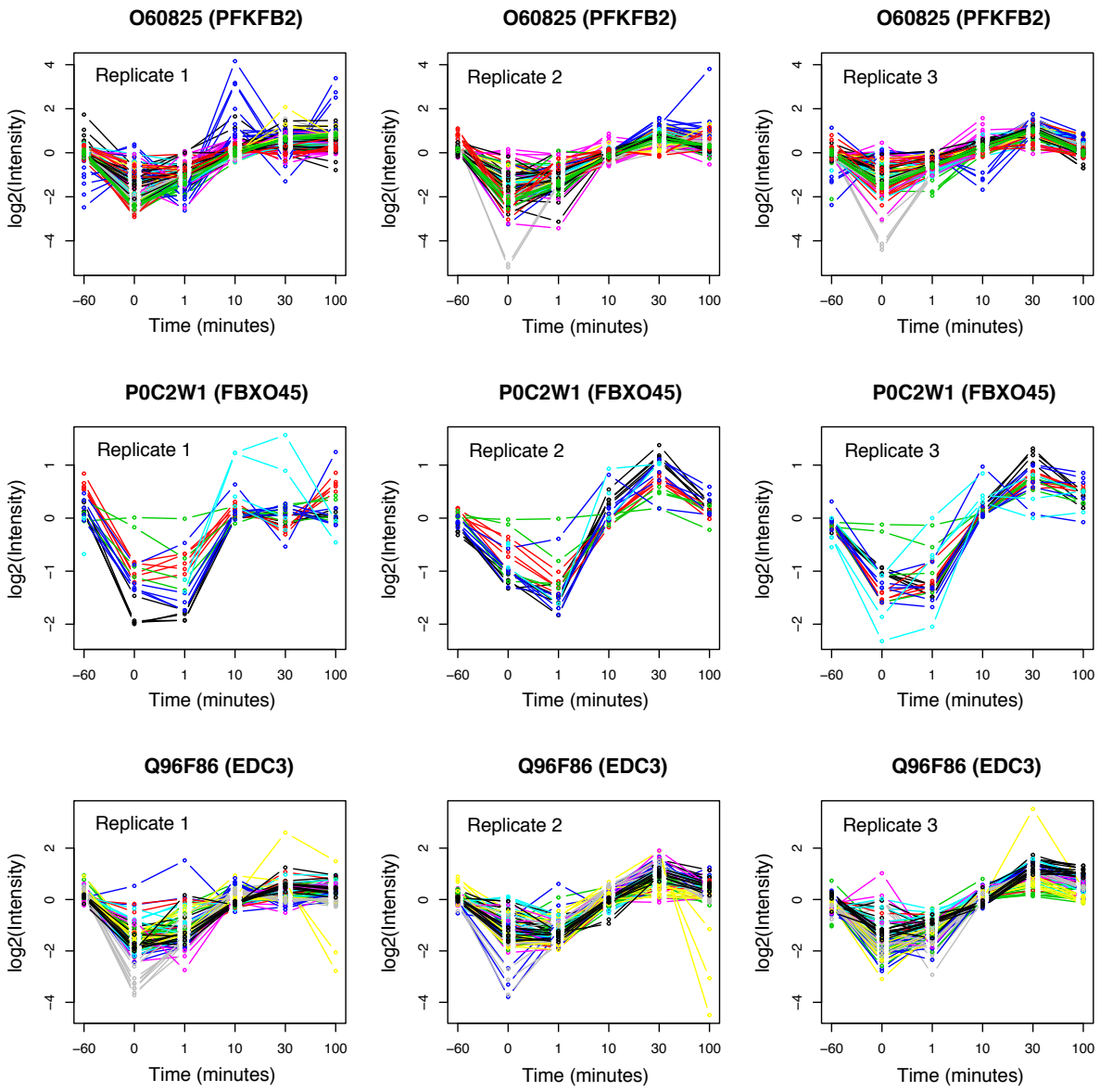

Figure 2.2: The example of three proteins in which fragment-level intensity data are highly consistent within each peptide and peptide-level abundances are highly consistent within the same protein. Each line color corresponds to a unique peptide.

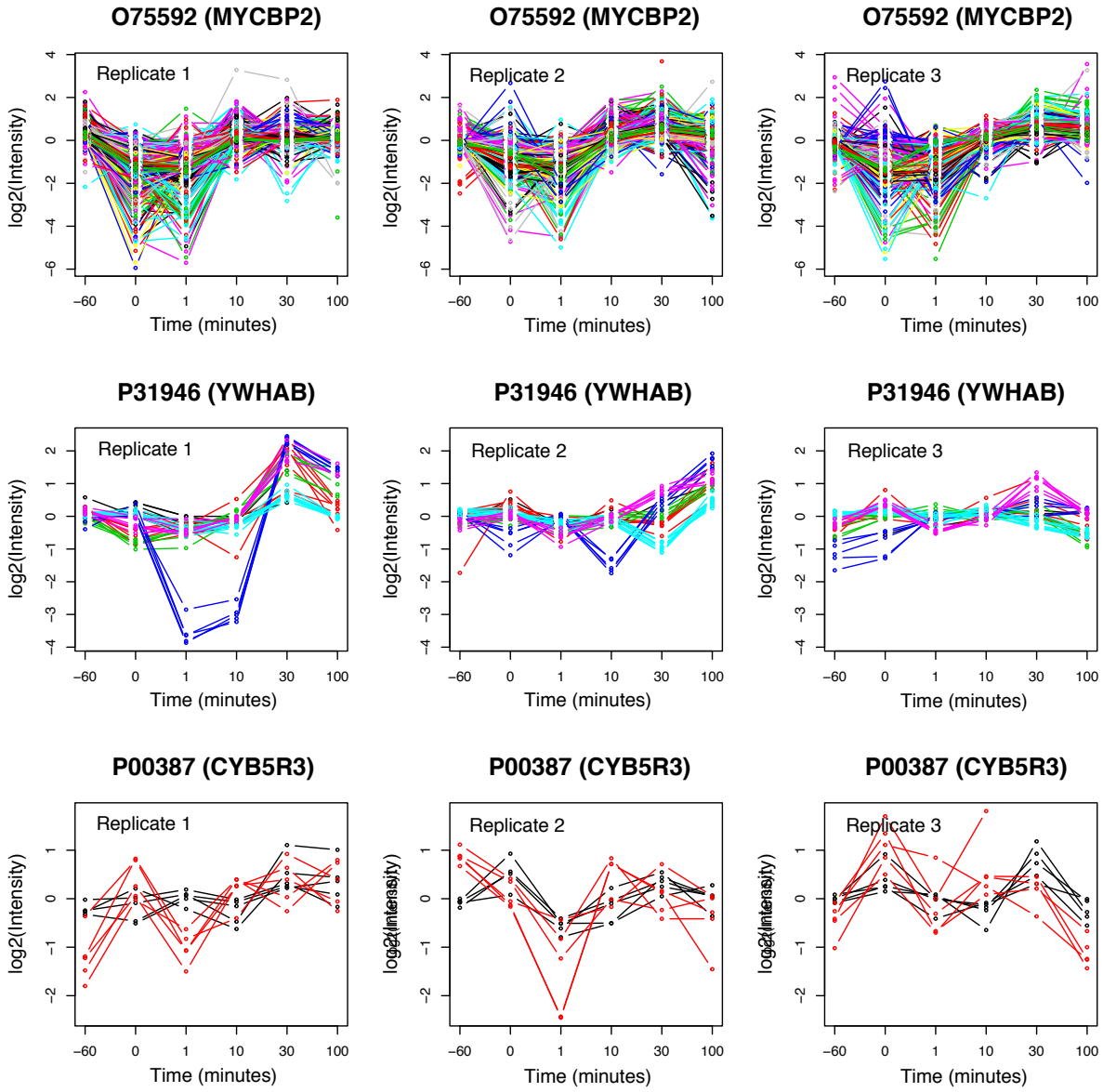

Figure 2.3: The example of three proteins in which peptide-level abundances are highly inconsistent within the same protein, with relatively faithful fragment-level intensity data. Each line color corresponds to a unique peptide.

2.2 Experimental Procedures

We first introduce key notations for the purpose of clarity. We denote the entire dataset with all fragments by $\mathbf{Y} = \{y_{fs}\}$, a $F \times S$ matrix of intensity values for F fragments in S samples across all comparison groups. For the purpose of flexible indexing in different parts of the model, we use bold fonts to indicate vectors and sub-matrices, and regular fonts to indicate scalars. Specifically, $\mathbf{y}_p$ denotes the rows in $\mathbf{Y}$ that correspond to the fragments of protein p. Likewise, $\mathbf{y}_q$ will concomitantly be used to denote the rows of $\mathbf{Y}$ for the fragments of peptide q, without specific reference to the protein, and $\mathbf{y}_f$ will denote a specific row of $\mathbf{Y}$ for the fragment f. When we need to specify a subset of fragments/peptides in specific samples or groups, we use additional superscripts $\mathbf{y}_f^g$, $\mathbf{y}_q^g$, to denote the sub-vector of $\mathbf{y}_f$ and the sub-matrix of $\mathbf{y}_q$ in the comparison group g respectively. Finally, we use $\mathcal{S}_g$ to denote the sample index set for group g.

2.2.1 Normalization

Once the fragment intensity data are extracted, the first data preprocessing in mapDIA begins with the normalization of intensity data (Figure 2.4A), where the goal is to remove systematic variations associated with easily identifiable experimental factors. A commonly used data normalization is scaling by the total intensity sum (TIS), i.e. the sum of intensities of all detected peak features in each sample. Since mapDIA takes the extracted data only, we use the sum of all reported fragment intensities as the denominator and transform the data as: $y_{fs} \to y_{fs} / \sum_{h=1}^{F} y_{hs}$. Following this transform, we multiply a constant factor to all fragment intensities so that the TIS of the normalized data is equal to the TIS of the raw data, so that the two data will be in a similar scale of values as the original data. This global normalization option is suitable when the inter-sample variation is constant for all peptides/fragments.

To accommodate the situation where the systematic variation fluctuates across the chromatographic time or retention time (RT), we developed a local normalization procedure, termed RT(δ) normalization. Let $T = (t_1, \ldots, t_F)$ denote the retention times of all F fragments in the dataset (e.g. RT of the apex of the elution profile of each fragment or its precursor). Then the RT(δ) normalization transforms the data as: $y_{fs} \to y_{fs} / \sum_{h=1}^{F} y_{hs} g_\delta(t_h - t_f)$, where $g_\delta(\cdot)$ is the normal density with mean 0 and standard deviation δ, and δ is the user-specified RT window for local normalization. Similar to the global TIS normalization, we multiply a constant to the normalized data to put them back on a comparable scale as the original data, with the exception that the scaling occurs in each fragment. Note that the the normalization factor in the RT(δ) normalization can change significantly across the RT axis, and thus the order of absolute abundance within each sample will change as a result of the normalization. Rescaling data within fragments preserves the order relatively intact.

It is crucial that the window size δ is not too small since an extremely small window

will cause the scaling factor to be dominated by the intensity of the fragment itself (or the fragments of the same peptide). On the other hand, a large δ will lead to equivalent outcome to the TIS normalization. In a typical 2-3 hour chromatography gradient, our current default choice of δ ranges from 10 to 30 minutes in proteomics applications (experiments with $\geq$2 hour gradient), which can be decided based on the visualization of total ion chromatograms of all samples on the same panel. The range of 10 to 30 minutes empirically resulted in similar and stable normalization in the datasets we have analyzed so far.

Once the data are normalized, the resulting data are log2 transformed and the intensities for each fragment will be centered by median. The median centering is performed differently depending on the experimental design (Figure 2.4A): the median is computed across all the samples for the "Independent Sample" (IS) design, whereas it is computed within each biological replicate for the "Replicate" (REP) design. See "IS design versus REP design" in section 2.2.3 below for more details.

2.2.2 Fragment filtering and selection

The next preprocessing step **(Step 2)** is a three-tiered fragment selection procedure (Figure 2.4A). Exclusion of noisy or irreproducible fragments is critical for statistical analysis because data extraction is typically performed in one sample at a time and thus not all fragments are detected and measured consistently across different samples.

(Step 2a) The first filter is for outlier fragment detection. We define "outlier fragment" as a fragment whose intensity data substantially deviates from the average normalized intensity of all other fragments within the same protein. To find these fragments, we apply row-wise median centering to the log-scale data for all fragments in each protein, compute sample standard deviation of the fragments in each sample, and tag a fragment as outlier if its intensity is outside a certain bound (default $\pm$2sd) in the sample. Note that this step removes the fragment data in each relevant sample, not across the samples at once.

(Step 2b) The second filter is for searching for the most representative fragments based on the "average cross-fragment correlation" of quantitative data. Suppose that protein p contains F_p fragments and a $F_p \times F_p$ correlation matrix is computed between all pairs of fragments, where the entry in the row a and column b is the Pearson correlation between $\mathbf{y}_a$ and $\mathbf{y}_b$ (fragments a and b). We denote the median correlation of a fragment f by m_f^p, where the median is taken over the correlations with all other fragments (excluding the self correlation), which will serve as the consistency score for the fragment. This score is stored in a score vector $m^p = (m_1^p, \ldots, m_f^p, \ldots, m_{F_p}^p)'$. After score calculation, the fragments with $m_f^p < m_*$ are removed by the user specified threshold m_*. As a result of this filter, the fragments that are correlated with the major cross-sample pattern in each protein will be retained, rendering the statistical analysis more robust than unfiltered data. In addition, the user can specify the maximum number of fragments per peptide (K) to keep the number of available fragments balanced for different peptides, where the top K fragments are selected based on average cross-fragment correlation within each

peptide. See examples and guidelines for choosing the optimal parameters in the software user manual.

(Step 2c) The final filter is to set inclusion/exclusion criteria based on the minimum number of fragments R and peptides Q available for each protein. Since our model requires repeated measurements for each peptide, at least two fragments must be available per peptide. In our experience so far, there are typically a large number of proteins that will be quantified by a single peptide, and the decision as to whether these proteins should be included or not must be made by the user and specified in the input parameter setting depending on the circumstances. The suggested default threshold values for protein and peptide-level differential expression (DE) analysis can be found in the example datasets distributed in the mapDIA package.

2.2.3 Statistical Model for Differential Expression (DE) analysis

Basic Modeling Framework

Using the preprocessed data, mapDIA proceeds to the DE analysis based on a Bayesian latent variable model with Markov random field prior, an adaptation of the model described in Wei and Li [74] with application to genomic data analysis. While our implementation automatically performs all pairwise comparisons requested by the user, here we illustrate using a comparison for two groups of samples for the clarity of explanation. We recall that the data $\mathbf{Y}$ were median centered (differently depending on the experimental design), then the probability model can be written as

$$\pi(\mathbf{Y}|\mathbf{Z}) = \prod_{p=1}^{P} \pi(\mathbf{y}_p|z_p) \tag{2.1}$$

where the observed data $\mathbf{y}_p$ for protein p is associated with the latent state z_p. $z_p = 1$ and $z_p = 0$ indicate that protein p is differentially expressed (DEd) and non-differentially (equally) expressed (EEd hereafter), respectively. Denoting the two groups in comparison by i and j,

$$\begin{aligned}
\pi(\mathbf{y}_p|z_p = z) &= \prod_{q\in\mathcal{I}_p} \int \varphi(\mathbf{y}_q^i, \mathbf{y}_q^j|z_p = z, \Theta_z)\pi(\Theta_z)d\Theta_z &(2.2)\\
&= \prod_{q\in\mathcal{I}_p} \int \prod_{f\in\mathcal{F}_{pq}} \varphi(\mathbf{y}_f^i, \mathbf{y}_f^j|z_p = z, \Theta_z)\pi(\Theta_z)d\Theta_z &(2.3)\\
&= \prod_{q\in\mathcal{I}_p} \pi(\mathbf{y}_q^i, \mathbf{y}_q^j|z_p = z). &(2.4)
\end{aligned}$$

where $\pi(\Theta_z)$ denotes the prior distribution of all model parameters for DE status z, and $\mathcal{I}_p$ and $\mathcal{F}_{pq}$ denote the peptide index for protein p and fragment index for peptide q

respectively. Here $\varphi(\cdot)$ denotes the product of all element-wise Gaussian densities, i.e.

$$\begin{aligned}
\varphi(\mathbf{y}_f^i, \mathbf{y}_f^j | z_p = 1, \Theta_1) &= \prod_{g \in \{i,j\}} \prod_{s \in \mathcal{S}_g} \frac{1}{\sigma_f \sqrt{2\pi}} \exp\left\{ -\frac{(y_{fs} - \mu_{qg})^2}{2\sigma_q^2} \right\} \\
\varphi(\mathbf{y}_f^i, \mathbf{y}_f^j | z_p = 0, \Theta_0) &= \prod_{s \in \{\mathcal{S}_i, \mathcal{S}_j\}} \frac{1}{\sigma_f \sqrt{2\pi}} \exp\left\{ -\frac{(y_{fs} - \mu_q)^2}{2\sigma_q^2} \right\}
\end{aligned}$$

where fragment f is from peptide q, $\Theta_1 = \{(\mu_{q0}, \mu_{q1}, \sigma_q^2)\}$ and $\Theta_0 = \{(\mu_q, \sigma_q^2)\}$ in protein p. The priors and closed form expression of $\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p)$ for DEd and EEd proteins is provided in the next two sections.

Prior distributions

The prior distribution for μ_{qg}, the average of the all intensities in peptide q group $g \in \{i, j\}$ and μ_q, or the average of the all intensities in peptide q group i and j, is conditional on the variance parameter σ_q^2 and is the Gaussian distribution with mean 0 and variance $(\sigma_q^2 \cdot V)$.

The hyperparameter V is set to 1000 to render this prior to be effectively subjective.

$$\begin{aligned}
\mu_{qg} | \sigma_q^2 &\sim \mathcal{N}(0, \sigma_q^2 \cdot V) \\
\mu_q | \sigma_q^2 &\sim \mathcal{N}(0, \sigma_q^2 \cdot V)
\end{aligned}$$

The prior distribution for σ_q^2, the variance of the all intensities in peptide q group i and j, is the inverse gamma distribution with hyperparameters (a, b).

The hyperparameters (a, b) is set to the method of moments estimates of the gamma distribution based on the sample variance calculated assuming equal means across the two groups, i.e.:

$$\text{first moment, } M_1 = \frac{\sum_{q\in\mathcal{Q}} s_q^2}{|\mathcal{Q}|}$$
$$\text{second moment, } M_2 = \frac{\sum_{q\in\mathcal{Q}} (s_q^2)^2}{|\mathcal{Q}|}$$
$$\text{shape parameter, } a = \frac{2\cdot M_2 - M_1^2}{M_2 - M_1^2}$$
$$\text{scale parameter, } b = \frac{M_1 \cdot M_2}{M_2 - M_1^2}$$

where s_q^2 is the sample variance for peptide q, $\mathcal{Q}$ is the set of peptides in the data and $|\mathcal{Q}|$ is the number of peptides in the data.

Closed form expression of the marginal likelihood

The closed form expression of $\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p, \Theta_z)$ for the EEd case is

$$\begin{aligned}
&\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p = 0, \Theta_0) \\
=&\int_0^\infty \int_{-\infty}^\infty \varphi(\mathbf{y}_q^i, \mathbf{y}_q^j | \mu_q, \sigma_q^2)\varphi(\mu_q | 0, \sigma_q^2 V)\Gamma^{-1}(\sigma_q^2 | a, b)\, d\mu_q d\sigma_q^2 \\
=&\frac{1}{\sqrt{(n_i+n_j)V+1}} \frac{\Gamma(a+(n_i+n_j)/2)}{\Gamma(a)} \frac{1}{(2\pi)^{(n_i+n_j)/2}} \\
&\times \frac{b^a}{\left[b + \frac{1}{2}\left(\sum_{y\in\mathbf{y}_q^i,\mathbf{y}_q^j} y^2 - (\frac{1}{n_i+n_j+1/V})(\sum_{y\in\mathbf{y}_q^i,\mathbf{y}_q^j} y)^2\right)\right]^{a+(n_i+n_j)/2}}.
\end{aligned}$$

The closed form expression of $\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p, \Theta_z)$ for the DEd case is

$$
\begin{aligned}
&\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p = 1, \Theta_1) \\
=&\int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \varphi(\mathbf{y}_q^i | \mu_{qi}, \sigma_q^2) \varphi(\mathbf{y}_q^j | \mu_{qj}, \sigma_q^2) \\
&\times \varphi(\mu_{qi} | 0, \sigma_q^2 V) \varphi(\mu_{qj} | 0, \sigma_q^2 V) \mathcal{IG}(\sigma_q^2 | a, b)\, d\mu_{qi} d\mu_{qj} d\sigma_q^2 \\
=&\frac{1}{\sqrt{n_i V + 1}} \frac{1}{\sqrt{n_j V + 1}} \frac{\Gamma(a + (n_i + n_j)/2)}{\Gamma(a)} \frac{1}{(2\pi)^{(n_i+n_j)/2}} \\
&\times \frac{b^a}{\left[b + \frac{1}{2}\left(\sum_{y \in \mathbf{y}_q^i} y^2 - (\frac{1}{n_i + 1/V})(\sum_{y \in \mathbf{y}_q^i} y)^2 + \sum_{y \in \mathbf{y}_q^j} y^2 - (\frac{1}{n_j + 1/V})(\sum_{y \in \mathbf{y}_q^j} y)^2\right)\right]^{a + (n_i+n_j)/2}},
\end{aligned}
$$

where $n_g = \sum_{y \in y_q^g} I\{\text{y observed}\}$ is the number of observed intensities in peptide q group g.

Markov random field (MRF) model, significance scores and FDR

We denote the true (unknown) state by $\mathbf{Z}_*$ and interpret this as a particular realization of the random vector $\mathbf{Z}$. Our goal is to recover the true state $\mathbf{Z}_*$ from the observed data $\mathbf{Y}$ across all comparisons,

$$\mathbf{Z}_* = \text{argmax}\ \pi(\mathbf{Z}|\mathbf{Y}) \tag{2.5}$$

where the joint distribution of $\mathbf{Z}$ is approximated by the Markov random field (MRF) model [5]

$$\pi(z_p = z|\cdot) \propto \exp\left(\gamma_z - \beta \sum_{k \in \partial p} 1\{z_k \neq z\}\right) \tag{2.6}$$

with ∂p denoting the set of neighbor proteins of protein p. Note that, if the module information is not utilized ($\beta = 0$), then the entire model will be equivalent to the mixture model treating the latent states as independent binary random variables. From the model above, we can derive the overall optimal solution $\mathbf{Z}*$ or derive the posterior probability of being DEd (with no module information) as the final protein significance score for

comparing group i and j:

$$\hat{s}_p = \pi(z_p = 1|\mathbf{y}) = \frac{e^{\hat{\gamma}_1}\pi(\mathbf{y}_p|z_p = 1)}{e^{\hat{\gamma}_1}\pi(\mathbf{y}_p|z_p = 1) + e^{\hat{\gamma}_0}\pi(\mathbf{y}_p|z_p = 0)} \tag{2.7}$$

Here $e^{\hat{\gamma}_1}/(e^{\hat{\gamma}_1} + e^{\hat{\gamma}_0})$ represents the prior probability of differential expression in the dataset, i.e. the estimate proportion of DEd proteins. In addition, we provide the posterior odds $\hat{o}_p = \pi(z_p = 1|\mathbf{y})/\pi(z_p = 0|\mathbf{y})$ as a supplemental score (in natural log scale), which is useful when further prioritization is needed among the high scoring proteins (e.g. among the proteins scoring $\hat{s}_p = 1$).

When the module information is utilized, the probability and odds scores are derived in the same manner by using the approximation

$$\hat{s}_p \approx \pi(z_p = 1|\mathbf{y}, \hat{z}_{(\Omega/p)}) \tag{2.8}$$

$$= \frac{e^{\hat{\gamma}_1 - \hat{\beta}\sum_{k\in\partial p}(1-\hat{z}_k)}\pi(\mathbf{y}_p|z_p = 1, \hat{z}_{(\Omega/p)})}{e^{\hat{\gamma}_1 - \hat{\beta}\sum_{k\in\partial p}(1-\hat{z}_k)}\pi(\mathbf{y}_p|z_p = 1, \hat{z}_{(\Omega/p)}) + e^{\hat{\gamma}_0 - \hat{\beta}\sum_{k\in\partial p}\hat{z}_k}\pi(\mathbf{y}_p|z_p = 0, \hat{z}_{(\Omega/p)})}. \tag{2.9}$$

Once the scores $\{\hat{s}_p\}$ are computed (omitting groups in the notation), the Bayesian FDR [48] is computed as

$$BFDR(s^*) = \frac{\sum_{\hat{s}_p > s^*}(1 - \hat{s}_p)}{\sum_{\hat{s}_p > s^*} 1}. \tag{2.10}$$

Estimation

The model parameters $\Phi = (\gamma, \beta)$ are estimated by the iterative conditional maximization (ICM) algorithm [5] as follows:

1. Obtain an initial estimate $\hat{\mathbf{Z}}$ of the true state $\mathbf{Z}_*$, using simple two sample t-tests.
2. Estimate Φ by the value $\hat{\Phi}$ which maximizes the pseudo-likelihood $\prod_p \pi(\{z_p\}_{ij}|\{z_{(\partial p)}\}_{ij}, \Phi)$.
3. Carry out a single cycle of ICM based on the current $\hat{\mathbf{Z}}, \hat{\theta}, \hat{\Phi}$, to obtain a new $\hat{\mathbf{Z}}$.

For $p = 1, \ldots, P$, update z_p which maximizes

$$\pi(z_p|\mathbf{y}, \hat{z}_{(\Omega/p)}) \propto \left[\prod_{q \in \mathcal{I}_p} \pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p, \hat{\theta})\right] \pi(z_p|\hat{z}_{(\partial p)}, \hat{\Phi}). \tag{2.11}$$

4. Go to step 2 until $\hat{\mathbf{Z}}$ converges.

This estimation is performed simultaneously for all pairwise comparisons specified by the user $\{(i, j)\}$ and a single set of MRF coefficients is applied to all the comparisons.

IS design versus REP design

The model derivation above is based on the independent sample comparisons (IS design, Figure 2.4D), where the samples in one group are compared to those in another group. A good example is the glycoproteomic data we present later, where 2 or 3 samples from each of 4 different prostate cancer stages are compared in a pairwise manner. In our modeling scheme, the replicate design (REP design, Figure 2.4D) refers to a situation where two or more conditions are compared within each of the biological samples. An example of REP design will be shown in the analysis of the dynamic interactome data of 14-3-3β, where the time course expression before and after a certain treatment is monitored within each of three biological replicates of an affinity purified sample. In the analogy of conventional hypothesis testing, the IS design corresponds to the t-test for two independent samples, whereas the REP design corresponds to the t-test for paired samples. mapDIA does not allow *nested* replicates in the comparisons, i.e. biological or technical replicates for individual samples when the comparison is made between groups of samples.

For modelling the data in the REP design, an obvious choice is to derive a similar model with replicate specific mean parameters and use the resulting marginal likelihood in the MRF model. However, we discovered that this leads to over-parameterization and usually performs poorly in small sample datasets. For this reason, we take the approach of

removing replicate specific averages from the data prior to modeling. Specifically, we apply median centering within each replicate separately first and analyze the data using the same model as the IS design. This adjustment efficiently removes the differences in the baseline intensity levels across different replicates and thus achieves reliable modelling of the data without the over-parameterization problem mentioned above. Note that, unless otherwise stated, replicates should be understood as biological replicates, not technical replicates (repeated MS runs over the same biological specimen), as the variability in such datasets do not represent the biological variation assumed in the variance component of the model.

2.3 Results

2.3.1 Overview of mapDIA workflow

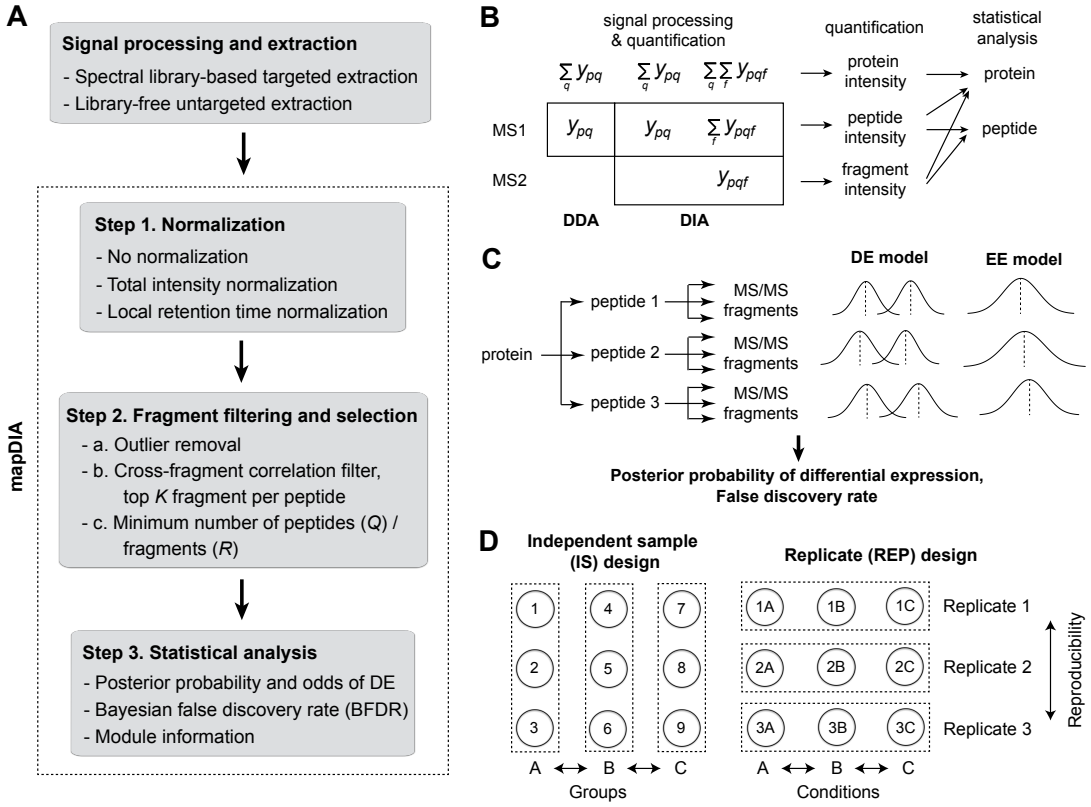

Figure 2.4: (A) The workflow of data processing and analysis using mapDIA. (B) Protein/peptide quantification possibilities using the data extracted from DDA and DIA data. (C) A conceptual diagram of the hierarchical model in mapDIA. (D) Experimental designs accommodated in mapDIA.

Our analysis framework follows a three-step workflow (Steps 1 through 3) as illustrated in Figure 2.4A. The input data should be obtained from a signal processing software that extracts peak features, either via targeted extraction of fragment intensities using spectral libraries (e.g. OpenSWATH, Skyline) [42, 55] or other novel approaches that do not rely on the spectral libraries (e.g. DIA-Umpire) [70]. The input data are further processed in two preprocessing steps by mapDIA, namely normalization (Step 1) and fragment selection (Step 2).

In the first step, mapDIA offers two optional normalization methods. One approach is the widely-used procedure of scaling the data by the total intensity sum in each sample (TIS), which essentially corrects for the variation in the total amount of samples analyzed per sample load. We developed an alternative procedure that scales intensity data by the locally weighted intensity sums on the RT axis, which is applied to each fragment in each sample separately. The latter procedure is more adaptive than the TIS-based universal normalization in the sense that temporal fluctuations in the chromatography and mass spectrometry can be adjusted [56].

The next step is fragment filtering and selection. This is a critical step since the data extraction tools operate on one sample at a time, and as a result the detection rate and quantification quality is not the same across all reported peptides and fragments. In mapDIA, there are three-tiered selection thresholds, including (a) standard deviation tolerance to define outliers, (b) minimum average cross-fragment correlation, and finally (c) minimum number of peptides and fragments required for differential expression (DE) analysis.

The last step (Step 3) is the model-based analysis for selecting differentially expressed proteins (DEPs). Although mapDIA's probability model is constructed flexibly enough to accommodate peptide and protein intensity data (Figure 2.4B), we will describe the model primarily for the analysis of fragment intensity data, uniquely reported from DIA data. mapDIA embodies a fairly sophisticated Bayesian hierarchical model for multi-group comparisons, which borrows statistical strength across all proteins in each dataset and thus confers robustness to the significance analysis, especially when the sample size is small (e.g. 3 samples per group). By contrast, the existing software package MSstats fits an independent fixed effects or random effects regression model for each protein and performs statistical significance inference using p-values with multiple testing correction [4], which heavily depends on the accurate estimation of fixed effects parameters and prediction of random effects parameters with a limited number of samples.

The structure of the probability model for individual proteins in mapDIA is illustrated in Figure 2.4C. After proper centering of the log scaled data, each fragment intensity is considered as repeated measurements of the parent peptide and is modelled by probability distributions under the differential expression (DE) scenario and equal expression (EE) scenario respectively. The posterior probability and the posterior odds of DE is reported as the significance score of the corresponding protein along with the false discovery rate estimates directly derived from the probabilities [48]. This model is constructed for two most common experimental designs, namely independent sample comparison (IS design) and within-replicate comparison (REP design) (Figure 2.4D), adding to the flexibility of our method to various kinds of experimental data.

2.3.2 Simulation Study (Default Model)

Key factors in simulation

We performed extensive simulation studies to evaluate the ability of mapDIA to identify DEPs. Although data preprocessing steps are essential components of mapDIA, the major goal of this first simulation was to evaluate the performance of the model in comparison to MSstats [8] in terms of classification of proteins into DEPs and non-DEPs and the quality of FDR estimation, without the influence of preprocessing steps that could give mapDIA an unfair advantage.

As discussed in the Introduction, we varied two factors that were deemed to be critical determinants of performance in our empirical observation over several test datasets. The first factor is what we dub as the "loyalty" of isolated precursor ions (peptides) to their parent protein, i.e. the deviation from the underlying protein abundance across the samples. The second factor is the measurement error or noise in the fragment intensities, which can be interpreted as the loyalty of the fragments to the abundance of their precursor peptides. Based on our empirical observations, the loyalty of fragments to their precursor peptides tended to be better than that of peptides to their precursor proteins. One extra factor we varied was the amount of data points per protein, which was controlled by the number of peptides per protein (n_p) and the number of fragments per peptide (n_{pq}). At a fixed value of the first two factors, we would expect that the simulation performance improve as more data are reported per peptide and per protein basis.

Data generation process

Specifically, we generated log-scaled data for two group comparison (group A and B) from the following simulation model:

$$y_{pqfj} = x_0 \mathbf{1}\{p \in \mathcal{D}, j \in \mathcal{S}_B\} + x_{pqj} + e_{pqfj} \tag{2.12}$$

for $p = 1, \ldots, 1\,500$, $q = 1, \ldots, n_p$, $f = 1, \ldots, n_{pq}$, and each group had 3 samples. Here $x_{pqj} \sim N(0, \tau^2)$ and $e_{pqfj} \sim N(0, \sigma^2)$ represent the intensity deviation of peptide q from the protein abundance in sample j and measurement error for fragments, respectively. The term x_0 corresponds to the effect size (the magnitude of DE for the protein), the set $\mathcal{D}$ is the set of DEPs, and $\mathcal{S}_B$ is the index set for samples in group B.

Figure 2.5 illustrates how these two factors affect the simulated data, where we assume a scenario of two group comparison and each group has three samples. Panels A through D correspond to $(\tau, \sigma) = (0.3, 0.3), (0.1, 0.3), (0.3, 0.2)$ and $(0.1, 0.2)$. In each panel, the log fragment intensities of each peptide were visualised by the dots of the same color, with additional lines connecting them across the samples. First, the parameter τ represents the variability between peptides, and thus the distance between the dots of different colors within each sample in the visualized data. Hence for a fixed value of measurement error σ, a small value of τ reflects good consistency between different peptides (panel B compared to A, panel D compared to C). On the other hand, the parameter σ represents the measurement error of fragment intensities, and this can be interpreted as the distance between fragments (dots) of the same color in each sample. Here for a fixed value of peptide deviation τ, a small value of σ reflects good consistency between fragments belonging to the same peptide (panels C/D compared to A/B).

In all simulation scenarios, we generated 100 datasets and averaged the results to produce the pseudo receiver operating characteristic (pROC) and FDR accuracy plots, where pROC is pseudo in the sense that "1-specificity" was replaced by the FDR in the horizontal axis. Specifically, we created 150 DEPs and 1,350 non-DEPs, where 10% of the proteins are DEPs in each simulation set. We set the effect size at $x_0 = 1$ and the noise level at $\sigma = 0.2$ and $\sigma = 0.3$, and varied τ between 0.1 and 0.3. Note that the peptide abundance deviates more from the true protein abundance as τ increases, i.e. quantification of peptides becomes less faithful to the underlying protein abundance level in each sample. In each simulation setup, we mixed proteins containing a different number of peptides

and fragments $(n_p, n_{pq}) = (2,3), (2,5), (5,5)$ per protein in equal proportions .

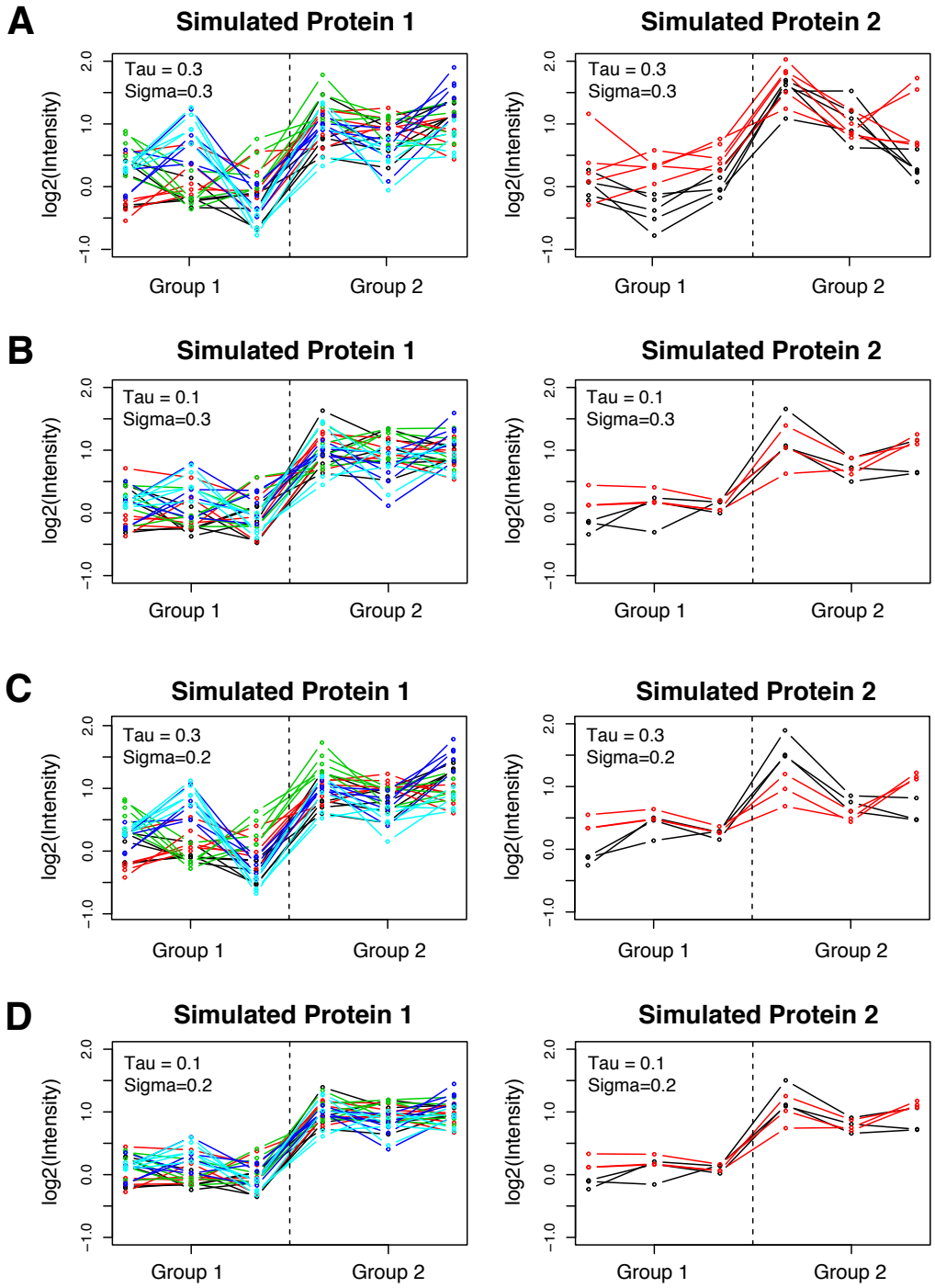

Figure 2.5: The example of two proteins across different simulation setting in terms of the peptide deviation from protein abundance τ and fragment intensity measurement error σ. (A) $(\tau, \sigma) = (0.3, 0.3)$. (B) $(\tau, \sigma) = (0.1, 0.3)$. (C) $(\tau, \sigma) = (0.3, 0.2)$. (D) $(\tau, \sigma) = (0.1, 0.2)$.

Classification performance and FDR accuracy

In all simulation settings, mapDIA and MSstats showed comparable classification performance in terms of prioritization of proteins. This is demonstrated by the overlapping pROCs for mapDIA (with no module information) and MSstats in Figures 2.6A and 2.6C. This comparable classification performance was retained even when the peptide abundance was very inconsistent with protein abundance ($\tau = 0.3$). With regard to the data volume, as expected, the classification performance improved as the intensity data were included from more peptides and fragments (data not shown).

However, the accuracy in the FDR estimates was markedly different between mapDIA and MSstats (Figures 2.6B and 2.6D, Figure 2.7). In mapDIA, the FDR estimates were highly accurate when the peptide deviation τ was below 0.2 (data for $\tau < 0.2$ not shown due to overlap), and the error began to be underestimated as τ increased above 0.2 (green and red line). Interestingly, the FDR accuracy was more dependent on the ratio of the two sources of error τ and σ than the sheer magnitude of each parameter. For a fixed level of peptide deviation $\tau = 0.2$ or $\tau = 0.3$ (green line), the FDR estimates were more heavily underestimated in the critical region (e.g. FDR < 0.1 in Figures 2.6B and 2.6D) when the ratio τ/σ was smaller. This suggests that the peptide deviation τ becomes much more influential for the error control in mapDIA when the fragment intensity measurement error is low, i.e. when the peptide deviation dominates the fragment measurement error. Note that, however, the data preprocessing steps not factored into this simulation were implemented to prevent these scenarios, since the filtering Step 2b based on the cross-fragment correlation score shall remove most fragments from such peptides with large deviation.

By contrast, MSstats showed unexpected results in terms of error control, with consistent patterns observed across 100 simulations of various settings. In MSstats, the users are expected to make the decision to model the data with fixed effects versus mixed effects

over the biological replicates and/or the MS runs (technical replicates). Since each sample is an independent MS run in our simulation, there are neither biological nor technical replicates as defined in the MSstats package. Hence we first assigned different biological replicate IDs to the samples and ran the analysis with fixed effects for biological replicates and another analysis with random effects for them (scopeOfBioRep option). We have also run the analysis with identical biological replicate IDs, but the results were similar. The option of fixed versus random effects changed the outcome dramatically. Figure 2.7 shows the FDR accuracy plots in simulations of the same setting as before, where the adjusted p-values showed poor control of the FDRs. When the model included random effects terms, the adjusted p-values were ultra conservative. On the other hand, when the model included fixed effects terms (currently the default option), the adjusted p-values were too sensitive, losing control of false discoveries. This phenomenon alarmed us to investigate this behavior carefully in all the experimental datasets, and this pattern remained consistent in those datasets as it will be shown later.

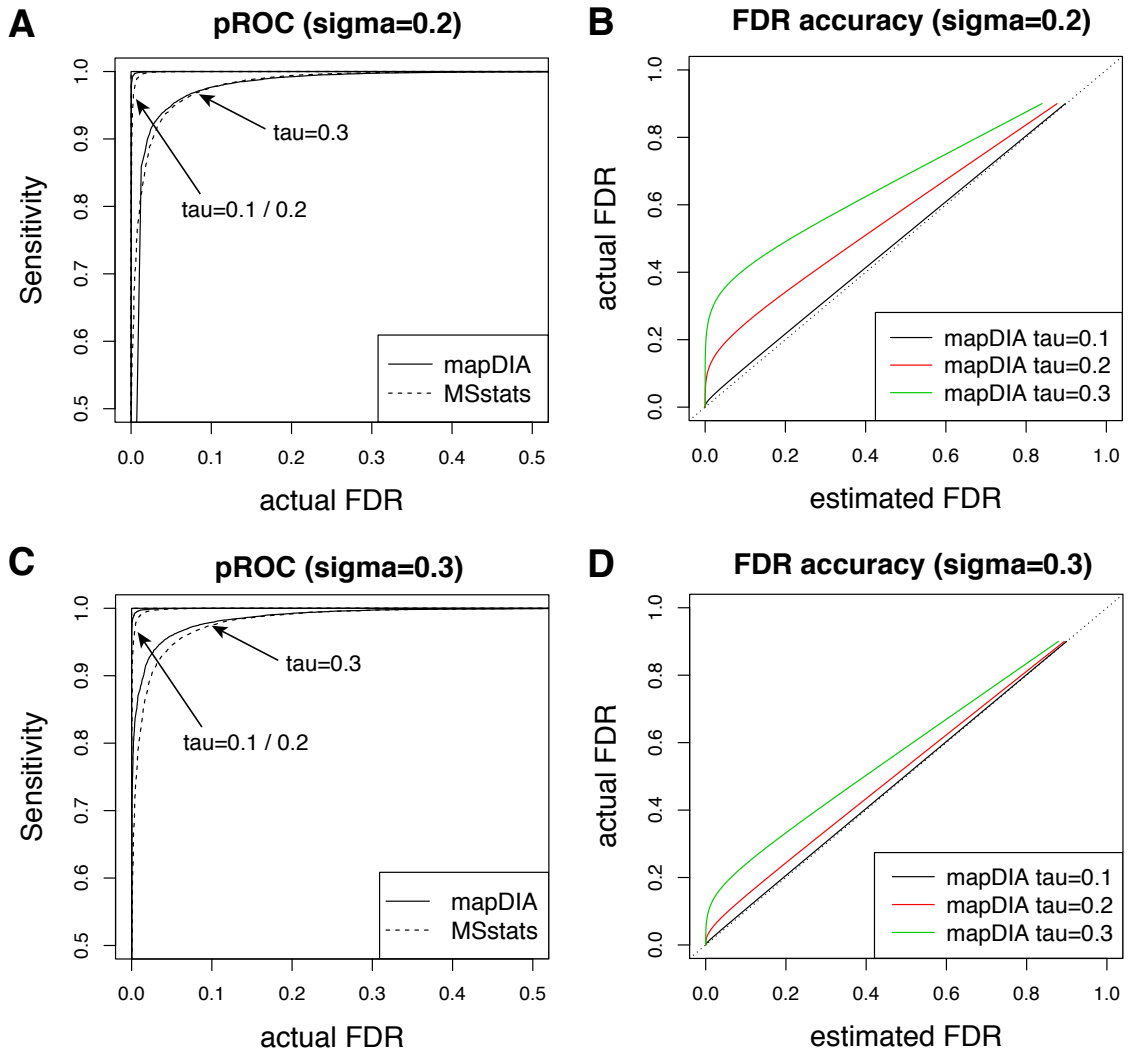

Figure 2.6: Classification performance and FDR accuracy in simulation studies. In each plot, the fragment intensity measurement error σ ("sigma") was fixed and the peptide deviation from protein abundance τ ("tau") was varied. (A, C) Sensitivity versus FDR (pseudo-ROC curve) plot and (B, D) FDR accuracy plot for mapDIA (solid) and MSstats (dashed). For each method, τ was varied between 0.1 and 0.3 at a fixed value of σ (0.2 or 0.3).

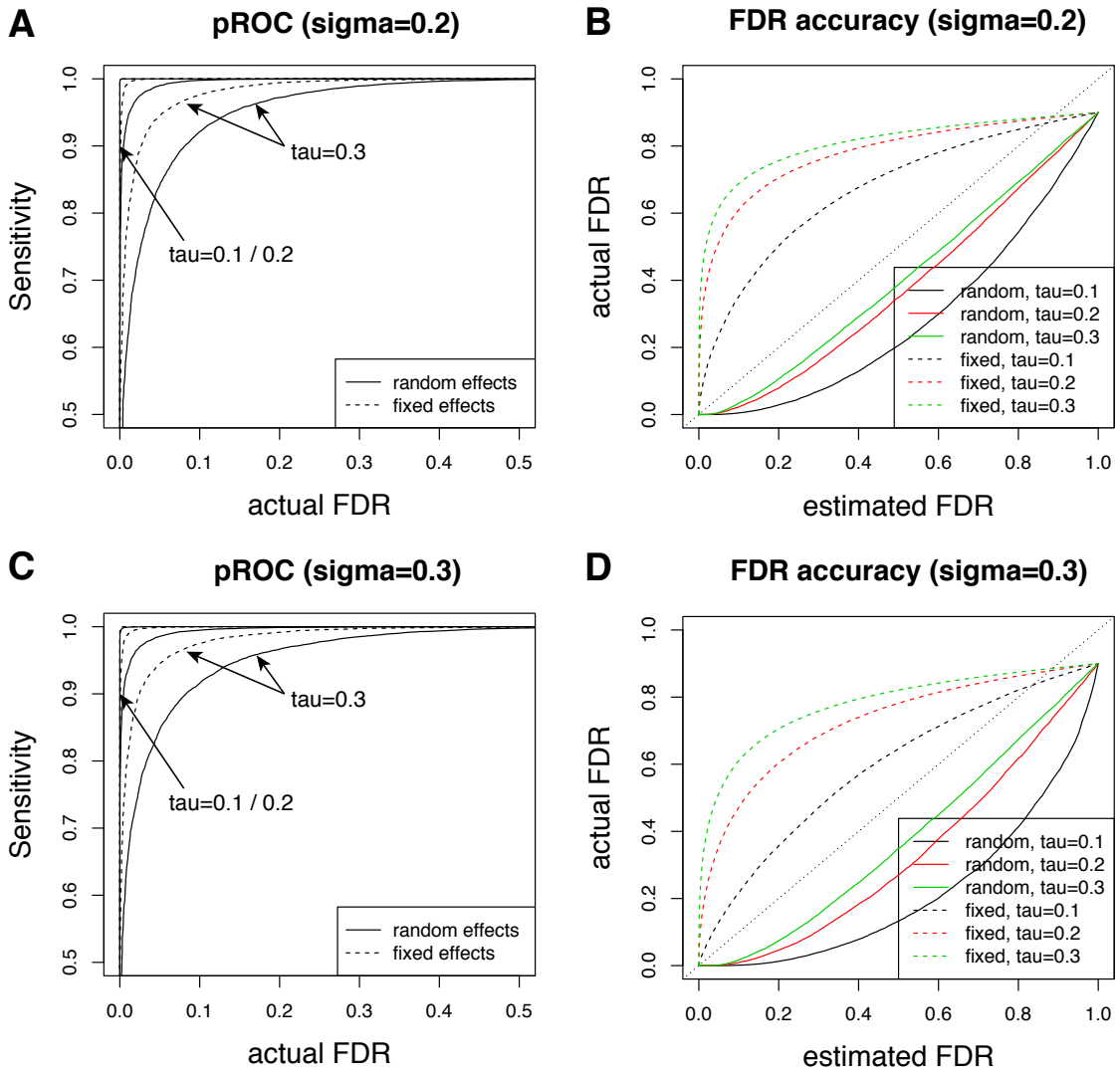

Figure 2.7: Classification performance and FDR accuracy in MSstats. In each plot, the fragment intensity measurement error σ ("sigma") was fixed and the peptide deviation from protein abundance τ ("tau") was varied. (A, C) Sensitivity versus FDR (pseudo-ROC curve) plot and (B, D) FDR accuracy plot in MSstats with fixed effects model (dashed) and random effects model (solid) for different values of τ ranging from 0.1 to 0.3 at fixed value of σ at 0.2 or 0.3.

2.3.3 Simulation Study With Module Information

We also evaluated mapDIA assuming a situation where module information is available, i.e. relational information between proteins. One example would be utilizing existing protein-protein interaction network (e.g. iRefIndex [53]) or functional modules (e.g. Gene Ontology [1] or Reactome [14]) in the analysis. Another example is to use peptide-protein membership as the module information when mapDIA is applied to score individual peptides, not proteins (shown later in the prostate cancer glycoproteomic data analysis).

In this simulation, we created the most ideal scenario where the module information can be maximized the most to demonstrate the concept for the purpose of illustration. To do this, we first created a scale-free network (Figure 2.8) using the algorithm of Herrera and Zufiria [31], and verified that the degree of connectivity follows the power law as expected in such a network ($P(k) \sim k^{-2.03}$). Next we allocated 150 DEPs in local subnetworks (see next section) so that the DEPs are network neighbors with one another. Using one realization of this network generation process, we simulated 100 datasets for DEPs and non-DEPs the same way as above, and compared the performance of mapDIA with and without the network (module) information.

As expected from such an ideal setup, the results showed that mapDIA assisted with the module information through the MRF prior brought significant improvement in the classification performance and FDR accuracy (Figure 2.9). The improvement was pronounced in proteins with few peptides and fragments, specifically for proteins with 2 peptides and 3 fragments per peptide. There are two caveats here. First, our analysis was conducted assuming that we have the complete knowledge of the underlying network/module. Second, the DEPs are often dispersed throughout the entire network in realistic datasets, i.e. not as concentrated around a subnetwork as in our simulation example. Both properties are not likely be fulfilled in real applications, and therefore it is expected the performance improvement will be more moderate than our demonstration. Surprisingly, though, there

are circumstances where such module information is indeed useful as we show later.

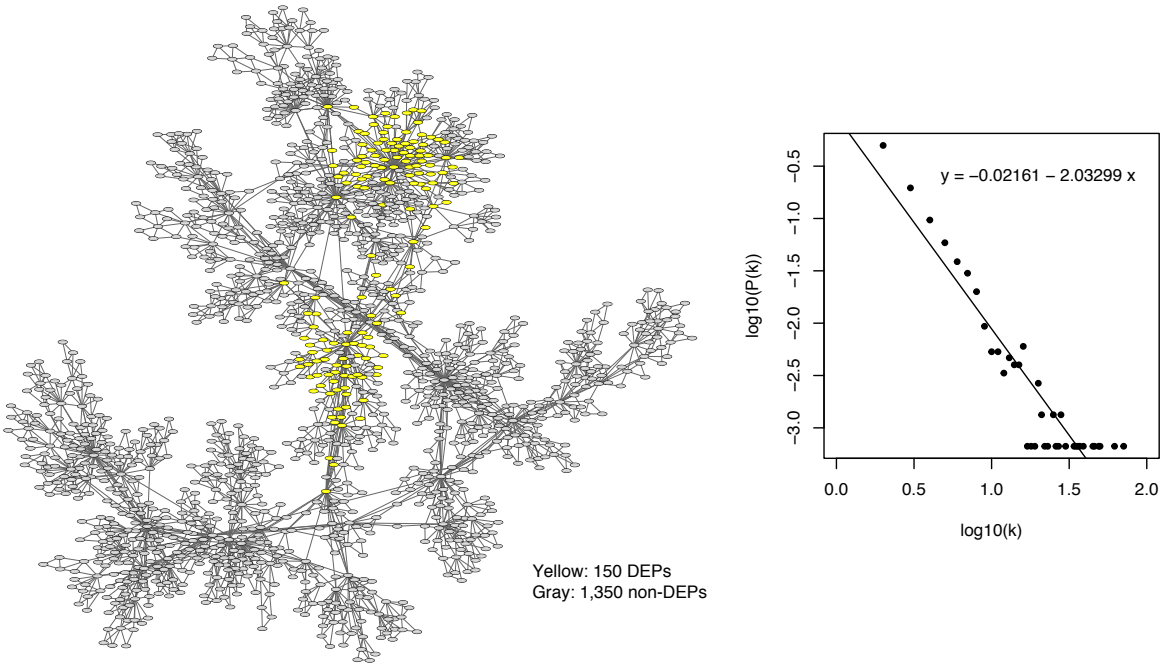

Figure 2.8: The scale-free network of 1,500 proteins with 150 DEPs concentrated in localized subnetworks (yellow).

DEPs on the scale-free network

Using the algorithm of Herrera and Zufiria [31], the generation of the 1,500 node network starts with a circular graph of 11 nodes. Most these 11 nodes are highly connected and play the role of hubs in the protein interactome. 2 neighbouring nodes from these 11 nodes were arbitrarily selected as DEPs. Next, the neighbors of these 2 nodes were also set as DEPs. This process was repeated until 150 DEPs were produced.

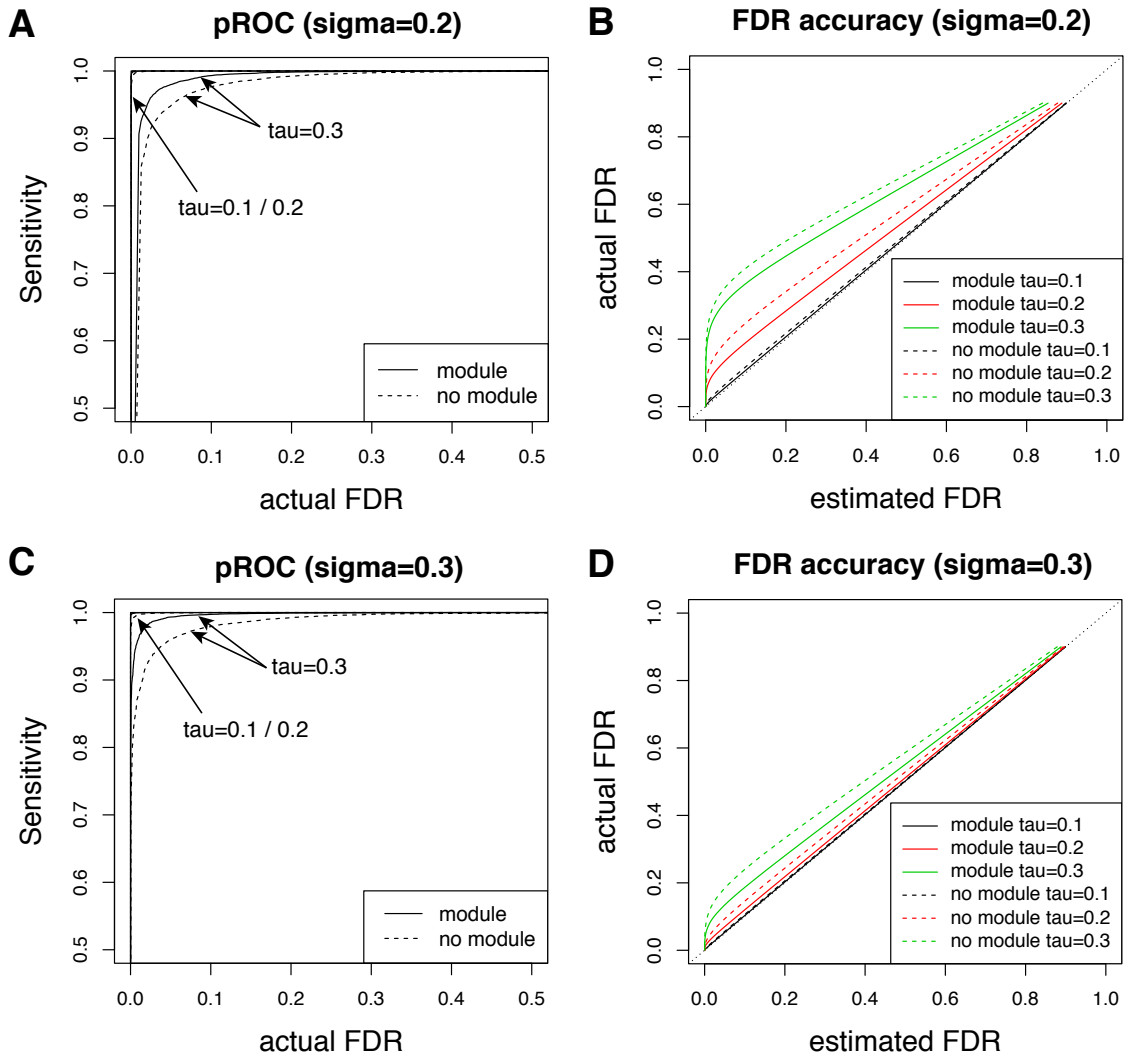

Figure 2.9: Classification performance and FDR accuracy in mapDIA. In each plot, the fragment intensity measurement error σ ("sigma") was fixed and the peptide deviation from protein abundance τ ("tau") was varied. (A, C) Sensitivity versus FDR (pseudo-ROC curve) plot and (B, D) FDR accuracy plot in mapDIA with module information (solid) and mapDIA without module information (dashed) for different values of τ ranging from 0.1 to 0.3 at fixed value of σ at 0.2 or 0.3.

2.3.4 Analysis of 14-3-3β Dynamic Interactome Data

We applied mapDIA to a recently published SWATH-MS dataset by Collins *et al* [9], who investigated the 14-3-3β interactome in IGF-stimulated HEK293 cells via affinity purification-mass spectrometry (AP-MS)[9] experiments in a time-resolved manner. The AP-MS experiments were performed in three biological replicates at six time points: the PI3K inhibitor LY294002 was added to prevent AKT activation (-60 minute) prior to IGF1 stimulation (0 minute), and the interactome was followed at four post-treatment time points (1, 10, 30, and 100 minutes) after IGF1 stimulation. GFP control purifications were also prepared in triplicates at each of three time points (-60, 0, and 30) to remove non-specific binders. The SWATH-MS data was extracted using the OpenSWATH tool based on an existing spectral library as described in [9], which produced the original data for 1,967 proteins, 16,180 peptides, and 85,545 fragments across all bait and control purifications.

We performed the data analysis similar to the original paper, where we used mapDIA (REP design, Figure 2.4D) and MSstats for the purpose of comparison. Since AP-MS experiments capture contaminants in addition to real interaction partners [7, 17, 28], we first compared the bait purification versus the control purification at each of the three time points using mapDIA, and identified proteins significantly enriched in the bait purification over controls (1% FDR) at one or more time points (648 / 1,967). Using this filtered data, we performed the DE analysis to compare protein abundance at all time points against the baseline at IGF1 stimulation (0 minute) using mapDIA and MSstats.

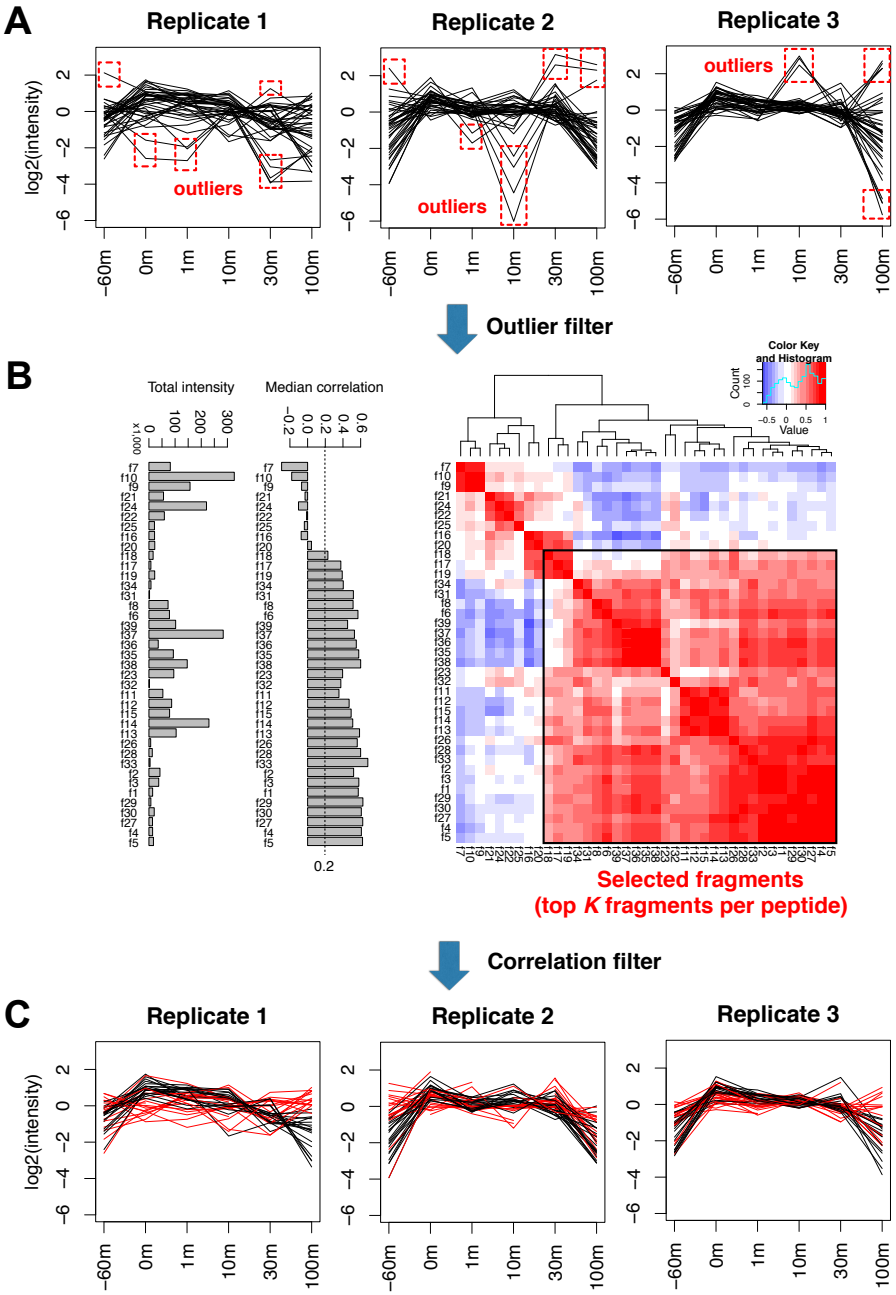

Figure 2.10: (A) The raw data for a sample protein in the 14-3-3β dataset. Each black line is the time course trajectory of fragment intensity data in each biological replicate. The first step detects outliers at each time point within each replicate. Red boxes indicate the data points outside 2 standard deviation at specific time point in each sample. The data shown are log-transformed data for all fragments with centering within each replicate. (B) After removal of outliers, the average cross-fragment correlation (using all 18 time points/samples) is used to score the fragment reliability. The dashed line 0.2 is a user-specified correlation threshold. After this filter, a user specified maximum number of fragments per peptide is selected (K). (C) The threshold 0.2 leads to removal of the fragments shown in red lines. Following this step, the proteins with at least Q peptides with minimum R fragments are kept for further analysis.

Fragment Filtering and Selection

Throughout the analysis, we applied 2 standard deviation threshold for outlier detection, average cross-fragment correlation 0.2 with maximum number of fragment per peptide $K = 5$, and at least 1 peptide per protein / 3 fragments per peptide in the fragment selection step. Since normalization of quantitative data in dynamic conditions can remove real biological signals [40], we applied no normalization procedure to this dataset. Figure 2.10A shows the outlier detection and removal step in the time course analysis, where the red boxes indicate the outliers that are removed in the subsequent analysis in specific samples.

Following this step, for each fragment, the Pearson correlation is computed with all other fragments within the same protein and the average cross-fragment correlation is reported as the consistency score for that fragment. The fragments with a cross-fragment correlation score below the minimal threshold are removed from further analysis (Figure 2.10B-2.10C).

After these two filtering steps, mapDIA analyzes the proteins with at least Q peptides containing at least R fragments for statistical analysis, where Q and R are specified by the user. As a result of this procedure, some outlier observations for 4,025 fragments were removed in at least one of the 18 samples (3 replicates, 6 time points), and 8,277 fragments (19%) from 495 of 648 proteins were removed as inconsistent fragments for the analysis. Lastly, 4,232 fragments were further removed by requiring $Q = 1$, $R = 3$ and $K = 5$, which resulted in the final dataset consisting of 31,038 fragments (6,872 peptides) in 632 proteins.

[9]Affinity purification-mass spectrometry: A method for the analysis of protein complexes that combines purification of protein complexes using affinity reagents and mass spectrometry.

Differential Expression Analysis in the REP design

Following the fragment selection step, we ran the DE analysis using mapDIA, comparing pre- and post-treatment time points (-60,1,10,30,100 min) against the time at IGF1 stimulation (0 min). Note that quantitative comparison is made at each of the 5 time points for 632 proteins (3,151 comparisons in total). In mapDIA, the estimated probability score associated with the estimated 1% FDR was $s^* = 0.825$ (no module information), and this threshold gave 1,018 significant comparisons. Here DEP refers to a protein that was DE in at least one of the five comparisons.

Figure 2.11A shows the plot of the significance scores against log2 fold change estimated from the three replicates at all five time points of comparison, showing clear separation between significant and non-significant comparisons. Here many proteins with absolute log2 fold change around 0.5 or below (raw fold change 30% increase or decrease) scored near zero probability. However, there was indeed an increasing tendency to score favorably as the number of peptides and fragments per protein increases. For example, the classification calls were very clear cut once the number of fragments per protein reached 30 or so, and the number of significant comparisons and non-significant ones were evenly distributed (577/1,370 significant comparisons were from the proteins with $\geq$30 fragments). Moreover, some comparisons were called significant at the target FDR level even with moderate average log2 fold change. These cases came from the proteins in which clear DE was observed in two biological replicates across many fragments, but not in third replicate. In the REP design, mapDIA automatically reports the inter-replicate correlation for each fragment, with which the user can identify these patterns once scoring is done.

Comparison with MSstats

In order to compare the results above with MSstats, we ran the analysis with all possible combinations of fixed effects and random effects terms for both biological replicates and

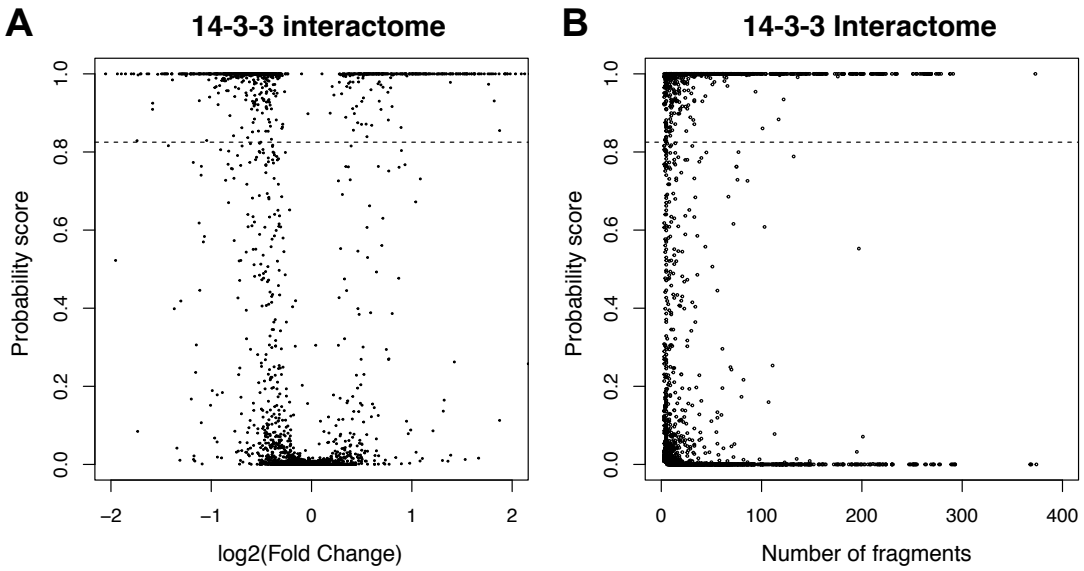

Figure 2.11: (A) Statistical significance scores against log2 fold change in the 14-3-3β interactome data. (B) Statistical significance scores against the number of fragments in each protein in the two methods in the 14-3-3β interactome data.

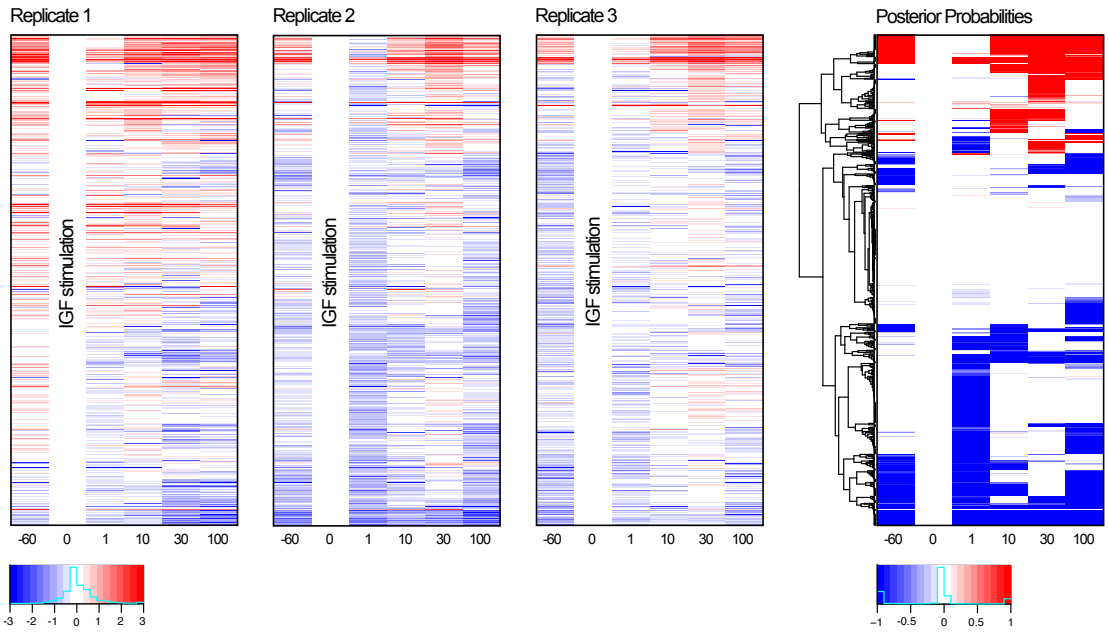

Figure 2.12: The left panel shows the log2 fold change against time 0 (at IGF simulation). The right panel shows the posterior probabilites of differential expression against time 0, with red indicating up-regulation, blue indicating down-regulation.

technical replicates, which gave us four different analysis outputs. Consistent with our experience in the simulation datasets, the output gave us two extreme results as far as the reported p-values are concerned, mostly dependent on the options for the biological replicates. When random effects were specified for biological replicates, merely 127 comparisons were found to be significant at 1% FDR threshold, whereas 2,244 comparisons (out of 3,151) were reported to be significant when fixed effects were specified (Figure 2.13A). The options for technical replicates only played a minor role, as expected from the design where the basal intensity levels for the time course for each biological replicates are expected to vary between biological replicates, rather than across the time points.

When fixed effects for biological replicates were used for comparison in MSstats, the significant comparisons from mapDIA were completely nested within the selection by MSstats (1008/1018), even though the two algorithms reported almost perfectly correlated log2 fold changes (Figure 2.13B). In fact, when we looked at the proteins called significant by MSstats but not by mapDIA, the fold change in the majority of these proteins was 40% or less (Figure 2.13C). When it came to the behavior of p-values as a function of the number of peptides and fragments, the majority of comparisons in those proteins with $\geq$30 fragments were called statistically significant by MSstats (1,154/1,370 comparisons) in the data with fixed effects for biological replicates (Figure 2.13D). Taken together, this finding suggests that the additional comparisons reported by MSstats tended to come from the pool of proteins with a large number of fragments showing moderate fold changes.

The comparisons called significant in MSstats also tended to come from the proteins with lower inter-replicate correlations (Figure 2.13E). This finding indicates that the time course profile of dynamic changes was inconsistent in the three biological replicates for these proteins. In addition, as an indirect evidence to show that mapDIA is not under-powered, we also looked at the enrichment of Akt substrates[10] (Akt1/Akt2) in the top scoring comparisons. Akt is the central kinase in the insulin-IGF1 signalling pathway

[10]Substrate: A molecule on which an enzyme acts.

which was modulated by the perturbation, and substrates of Akt are well known to bind 14-3-3 proteins at the phosphorylated site. As such, binding of Akt substrates to 14-3-3 is expected to be significantly modulated by this treatment. The substrate list was extracted from PhosphoSitePlus [32] and NetworKIN [38] and the comparisons were ordered by the log odds scores for mapDIA and adjusted p-values for MSstats. Figure 2.13F clearly shows that Akt substrates were more enriched in the comparisons prioritized by mapDIA than that of MSstats.

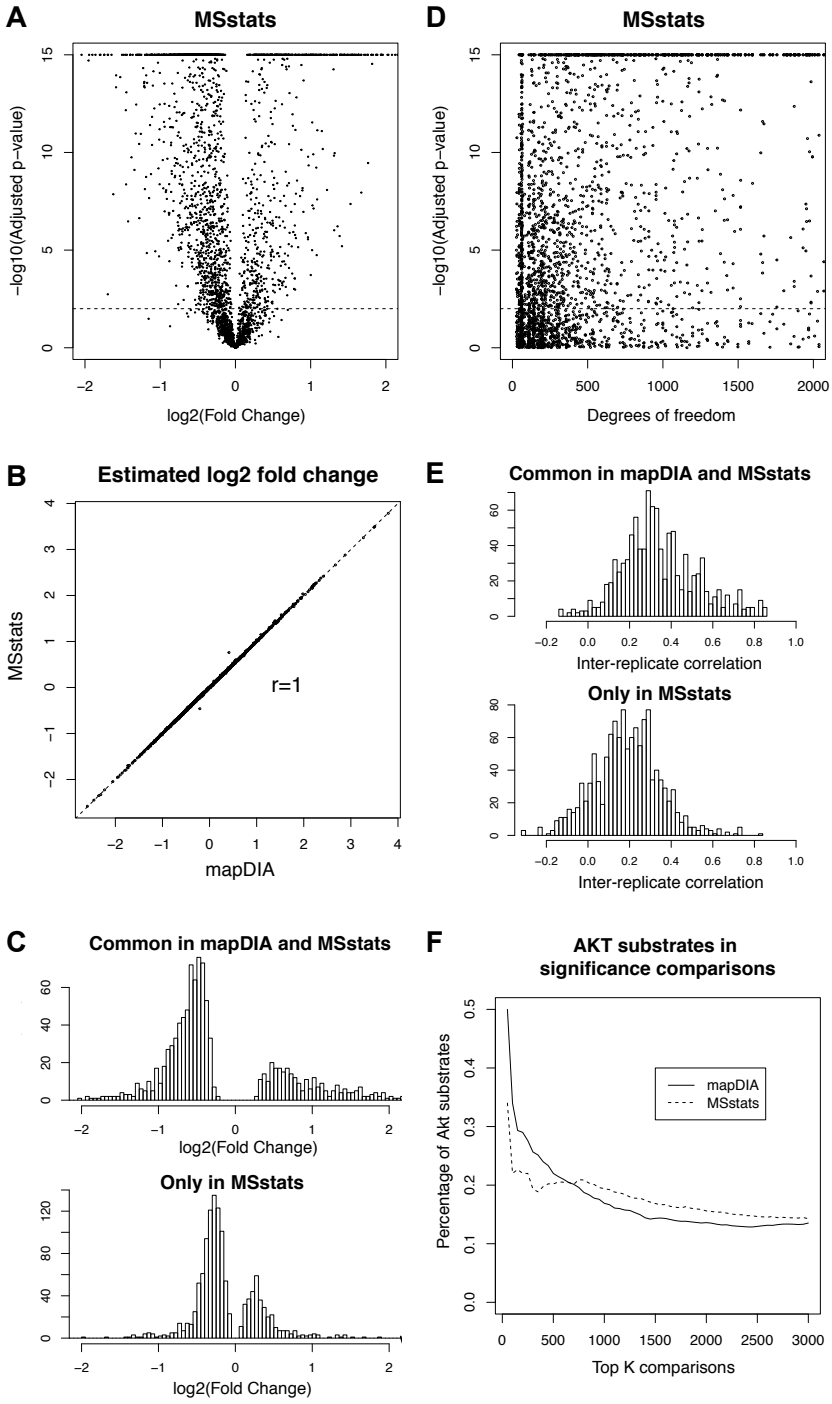

Figure 2.13: (A) Significance score versus log2 fold change in `MSstats`. (B) The reported log2 fold changes from `mapDIA` and `MSstats`. (C) Log2 fold changes for the comparisons found significant in both softwares (top) and those found significant only in `MSstats` (bottom). (D) The trend in significance scores along the number of fragments (represented by the degrees of freedom in the regression model with fixed effects). (E) Inter-replicate correlation for the comparisons found significant in both softwares versus those significant in `MSstats` only. (F) Akt substrate enrichment in the top K comparisons in `mapDIA` and `MSstats`.

2.3.5 Analysis of Prostate Cancer Glycoproteomics Data

We next analyzed a published glycoproteomics dataset of prostate cancer samples with varying tumor aggressiveness [39]. In this study, *N*-linked glycopeptides were isolated from 10 normal samples (N), 24 non-aggressive (NAG), 16 aggressive (AG) and 25 metastatic[11] (M) prostate cancer samples, and each group was pooled into 2 or 3 sample pools and analyzed by SWATH-MS (effective samples sizes are 2 N, 2 NAG, 3 AG, 3 M). We extracted the data for 302 glycoproteins[12] (2,641 peptides, 27,361 fragments) using the recently developed DIA-Umpire tool [70], and performed DE analysis using mapDIA in the independent sample (IS) design (Figure 2.4D) for all 6 pairwise comparisons between the four groups at the protein level.

Since the extracted SWATH-MS data was based on glycopeptide enrichment, there are many proteins for which only a few peptides are available. For this reason, we allowed proteins with single peptide as long as each peptide has at least 3 to 5 fragments per peptide ($Q = 1$, $R = 3$ and $K = 5$). Outliers were filtered by 2 standard deviation threshold, and the cross-fragment correlation threshold was set to 0. In addition, we required that each fragment have at least 2 non-missing values in each sample group, so that the AG and M groups can be compared even if the data was missing in one of the three samples. These parameters led to selection of 9,697 fragments in 298 glycoproteins, for which 1,258 comparisons could be made after removing proteins in specific groups with too many missing data.

Normalization

In this data, we tested all variants of normalization methods implemented in mapDIA first. According to a recent report that investigated the variation in multi-center proteomic

[11]Metastasis, or metastatic disease: The spread of cancer cells from the initial site of the tumor to form secondary tumors at other sites in the body.

[12]Glycoprotein: Any protein with one or more covalently linked oligosaccharide chains.

data [56], the major source of systematic variation in LC-MS experiments turned out to be chromatography retention time (RT), charge state, and ion suppression during the ionization in each MS run. To address such temporal variation by the RT(δ) normalization, we used Gaussian kernel weights with standard deviation of $\delta = 10$ and $\delta = 30$ minutes. If there is no such temporal or local variation, then this normalization method will lead to an equivalent outcome as the TIS normalization. When we compared the post-normalization data among no normalization, TIS-normalized, and RT(30) and RT(10) normalized data, the fragment intensity data were significantly more correlated in RT(10) normalized data between samples belonging to the same sample group (Figure 2.14), suggesting improved normalization of the data therein. We therefore decided to use the data normalized by the RT(10) normalization for further downstream analysis.

Protein DE Analysis in the IS design

We first performed protein DE analysis using mapDIA under the IS design (Figure 2.4D) and MSstats, comparing every pair of groups (up to 6 comparisons per protein). For MSstats, the aberrant behavior of p-values were consistently observed with the choice of fixed effects and random effects specification in the model, where nearly no comparisons were reported to be significant once random effect terms were used, whereas three quarters of comparisons (1,167 / 1,537) were found to be significant (see below). Since the random effects model of MSstats gave too few significant comparisons, we used the fixed effects model for comparison. We noticed that 279 / 1,537 comparisons reported from MSstats were not found in the mapDIA output due to minimal fragment requirement $(Q, R) = (1, 3)$ in the latter. Therefore we compared the two methods only for the comparisons reported from both (1,258 in total).

At the 1% FDR threshold in each method, mapDIA and MSstats reported 511 and 971 comparisons as significant respectively, and the comparisons significant in the former were again almost completely nested within those reported by the latter. In the mapDIA

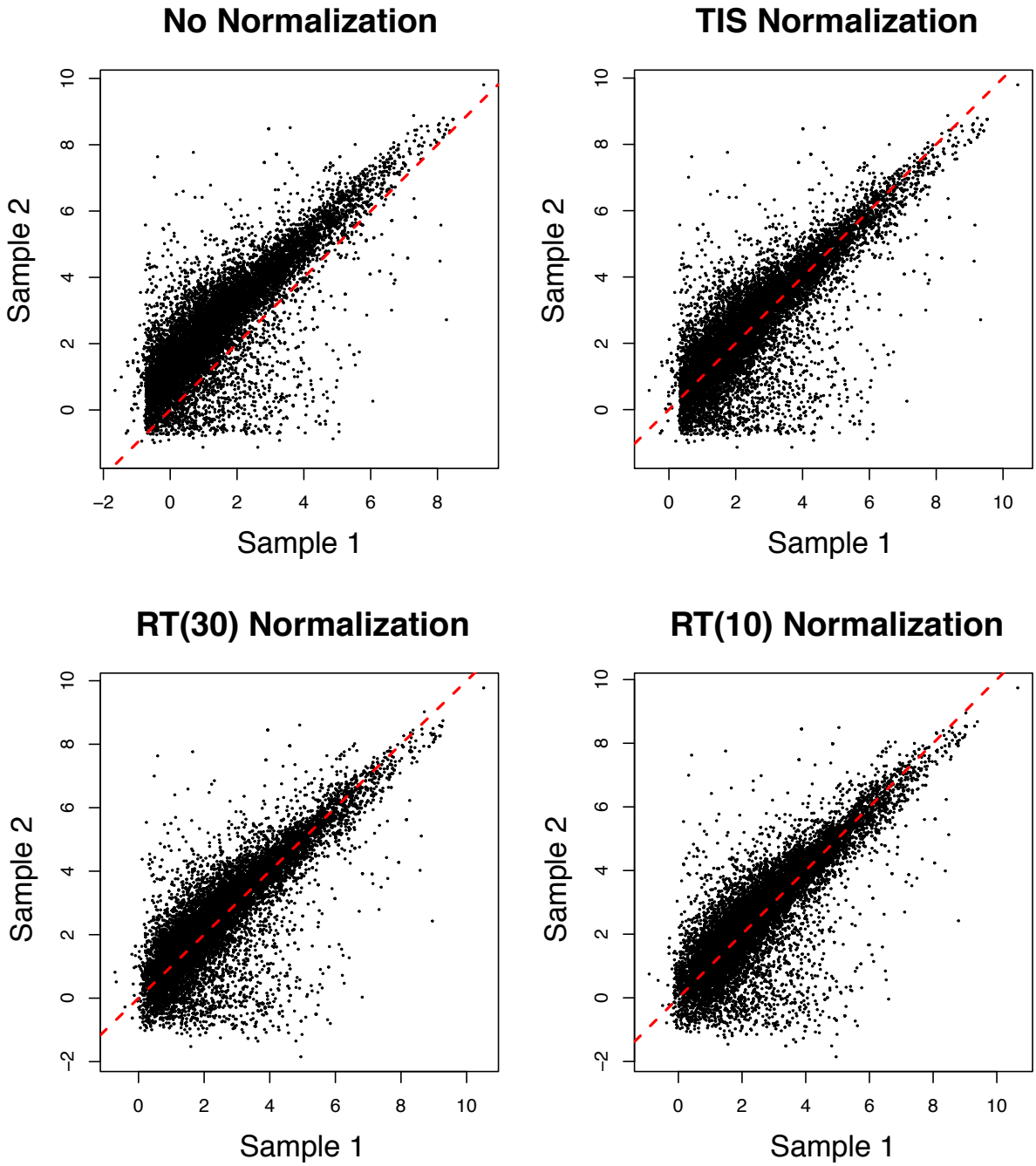

Figure 2.14: Within-group pairwise scatter plot of fragment-level intensity data using four different normalization options: no normalization, TIC normalization, RT(30) and RT(10) normalization in prostate cancer glycoproteomics data (control groups). The trend and improvement was observed in the other three groups, which are not shown due to large file sizes.

analysis, most significant glycoproteins were differentially expressed by at least 40% (absolute log2 fold change 0.5), and there was no bias in favor of the proteins with a large number of fragments per protein (Figures 2.16A and 2.16B). Examining the 511 significant DE cases, the majority came from the comparisons between cancer patients and controls (transplant donors) and the comparisons between the MET group and AG/NAG groups. When we examined the estimated fold changes reported from both methods, they were again highly correlated ($r = 0.952$, Figure 2.15), indicating that the disparity in the significant comparisons at the same FDR threshold comes from the distinct statistical approach to model the variability in the data, rather than the estimation of effect size (magnitude of change).

Peptide DE Analysis using the MRF model

For quantitative analysis for post-translational modifications, it may be of interest to perform DE analysis at the level of modification site. Hence we performed the analysis at

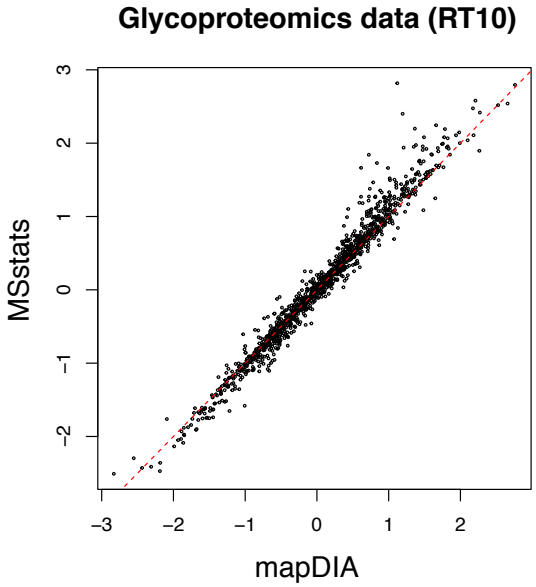

Figure 2.15: The reported log2 fold changes from `mapDIA` and `MSstats`.

the peptide level, by specifying peptide names as the protein identifiers in the mapDIA. We ran the analysis with and without the module information in the MRF model, where we set a group of peptides belonging to the same protein as a module (i.e. a parent protein is a module for the member peptides). This specification represents the hypothesis that a glycopeptide is more likely to be DE if other glycopeptides are also DE in the same protein.

After applying the same fragment filtering and selection criteria, peptide DE analysis could be performed for 6,754 comparisons. mapDIA reported 2,095 comparisons to be significant at 1% FDR threshold without the module information. Figure 2.16C shows the plot of the significance scores against the log2 fold change estimates. Unlike the previous two protein-level analyses, this plot looks similar to a typical "volcano plot" one would expect from the analysis of a typical gene or protein expression dataset where each gene or protein is quantified with a single value. This was expected because the analysis was performed at the peptide level, each containing at most 5 representative fragments in terms of cross-fragment correlation, and therefore the amount of data for each unit (protein or peptide) was much more balanced for peptide-level analysis that for protein-level analysis.

When the module information was utilized through the MRF model, 2,185 comparisons were found to be significant, where the majority (2,005) were in agreement between the two models. As expected, additional DE peptides in the model using the module information were found in the proteins containing other significant DE peptides. Figure 2.16D shows that, when we looked at the 180 additional comparisons significant in the model with group information, on average 75% of the other peptides in the same proteins were significant DE peptides. This indicates that the MRF model effectively pooled information within modules (individual proteins) to boost probability scores for peptides when other peptides in the same protein were DE and vice versa.

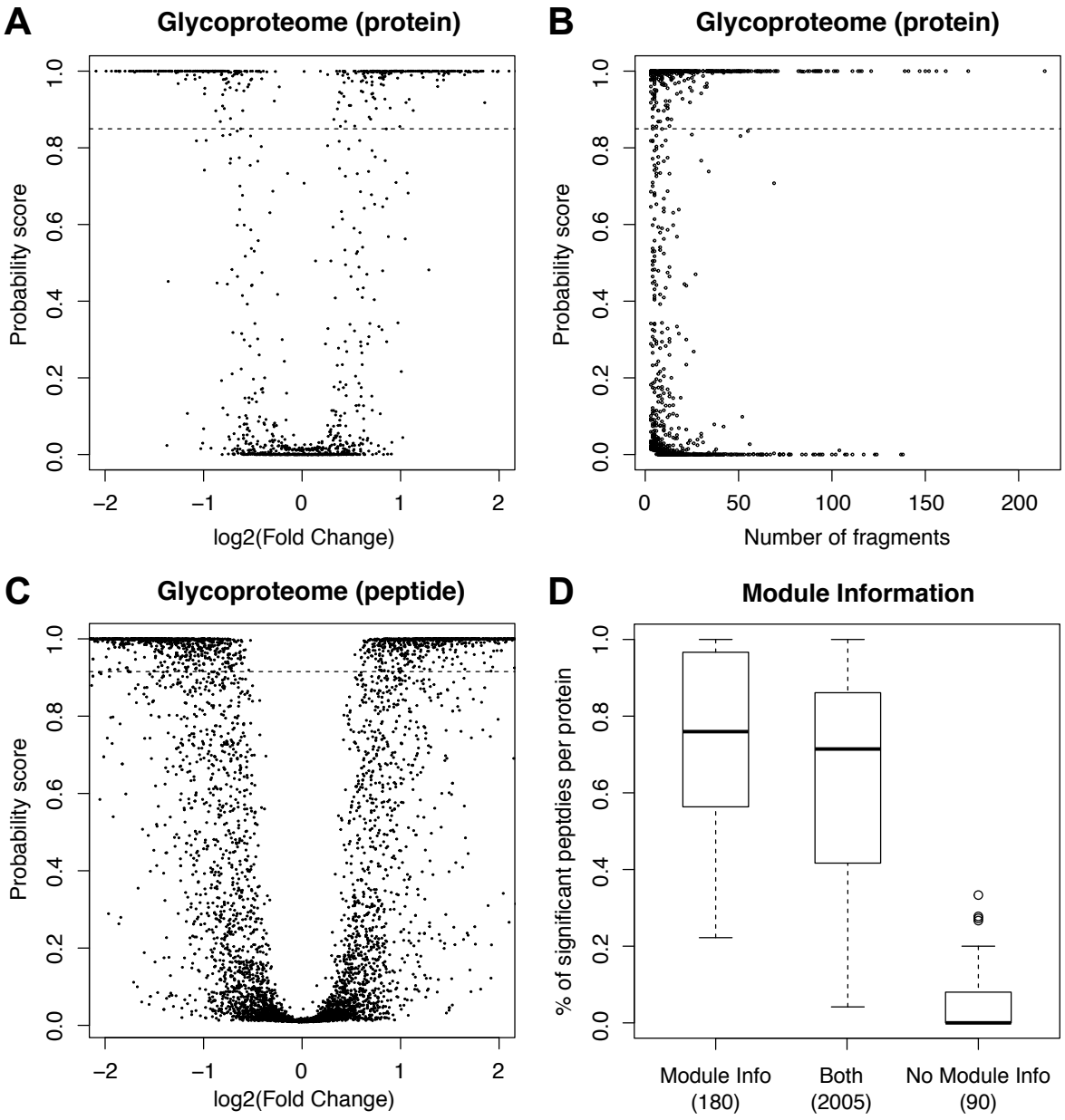

Figure 2.16: (A) Statistical significance scores against log2 fold change of glycoproteins in the prostate cancer glycoproteome data. (B) Statistical significance scores against the number of fragments in each protein in the two methods in the prostate cancer glycoproteome data. (C) Statistical significance scores against log2 fold change of glycopeptides in the prostate cancer glycoproteome data (no module information was used). (D) The proportion of other significant peptides in the same protein for the peptides significant in the model with and without module information. Here the module information is the parent protein ID for each peptide.

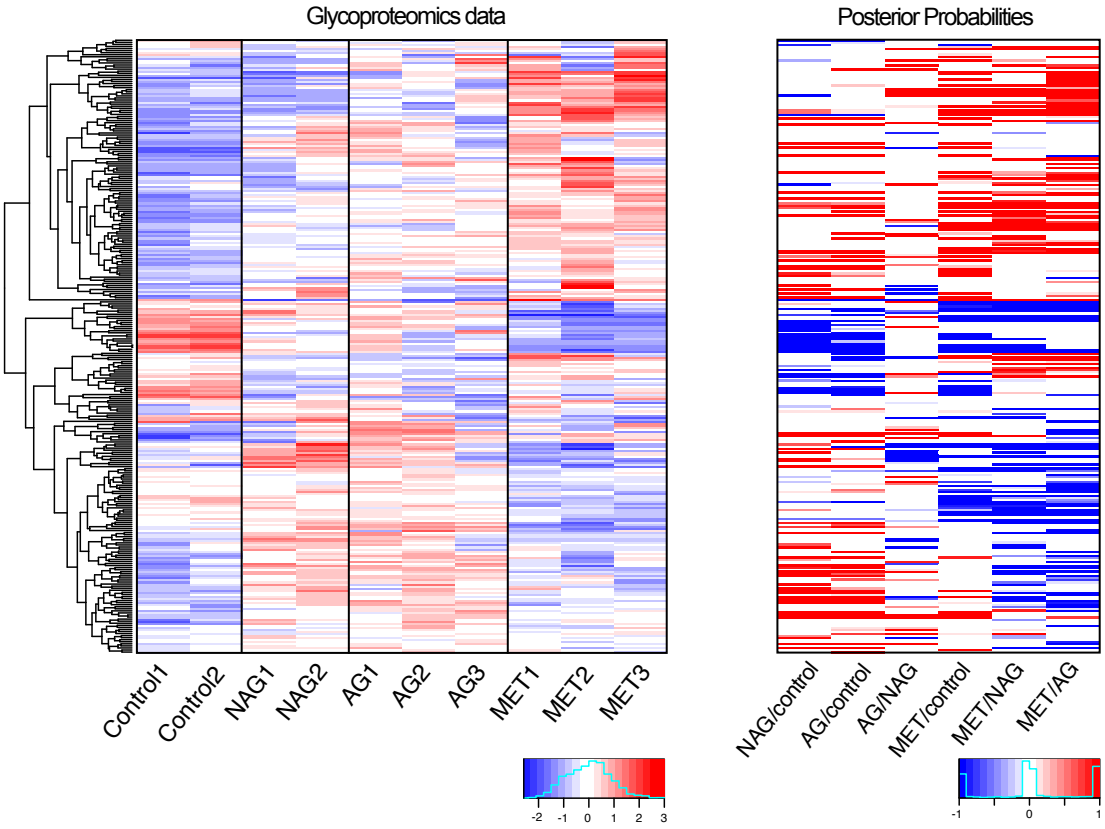

Figure 2.17: The left panel shows heatmap of the log2 intensities of the Glycoproteomics data. The right panel shows the posterior probabilities of differential expression for the various comparisons.

2.4 Discussion

In this chapter, we presented a novel software package mapDIA for statistical analysis of quantitative proteomics data generated in the DIA mode. Our data preprocessing routines include normalization methods that can remove systematic bias that is constant or temporal between the samples, and a series of fragment filtering and selection procedures to remove noisy and irreproducible fragments. The Bayesian model, previously developed for microarray data [74] and adapted here for the protein-peptide-fragment hierarchy in DIA data, allows robust control of FDR and sensitive selection of DEPs with the power of repeated measurements over multiple fragments/peptides, highlighting a unique advantage of MS/MS fragment quantification offered by the DIA mode analysis. The software is also flexible enough to accommodate different experimental designs and we believe that our method works very robust in terms of controlling false positive findings even in small sample datasets as illustrated in our two data analysis examples.

We have used MSstats as our main benchmark for comparison in this chapter. The conclusion we drew in the performance comparison warrants further investigation across a large number of datasets. We conjecture that the main reason behind the puzzling behavior of their reported p-values is related to the differences in the property of data between the integrated peak area data in SRM/MRM experiments and the fragment intensity data in generic DIA mode. In the former, the protein-peptide-transition pairing is carefully selected, which yields more reliable quantification for protein-level statistical inference. In the latter, by contrast, we discovered that peptides isolated across the swaths in SWATH-MS can deviate quite significantly from the average cross-peptide pattern of each protein and the fact that MSstats's current regression model handles sample-to-sample variation only at the protein level through fixed or random effects (even in the presence of the interaction terms in their model) may not be sufficient to account for such variability. One possible remedy is to expand the model specification with peptide-level

random effects rather than protein-level random effect terms, which does not cost the model any significant loss since the best linear unbiased predictors in linear mixed effects models do not take away large degrees of freedom for differential expression inference. However, it will be important to control the number of random effects terms, for example, by adding a variable selection step to prevent over-parametrization as the number of peptides increases [22].

An important feature of mapDIA is the three-tiered fragment selection step. In particular, the average cross-fragment correlation score can effectively remove noisy fragments in both data examples. These steps are critical because the peak data is extracted in each sample separately and thus the fragment intensity data may not be of the same quality across different samples. Moreover, even with the well-behaving fragments, we noticed that a certain degree of data reduction is crucial for reliable statistical modelling in both methods because the amount of data can be severely unbalanced for different proteins, i.e. while some proteins have hundreds of proteins from tens of peptides, others may have a single peptide with a few reliable fragments. To address this problem, we have allowed the user to control the maximum number of fragments (K) per peptide, where the most representative K fragments in terms of the average cross-fragment correlation score are chosen within each peptide. However, we also remark that excessive application of these filtering steps can lead to spurious findings, and thus our recommendation is to carefully specify the input parameters in a way that the final selection of reliable fragments preserves the underlying quantitative trends across the samples. To facilitate this monitoring process, our software automatically reports the filtering outcome at every stage as a part of the analysis output and also saves the filtered data to allow the users to visualize the data and monitor the changes as different filtering criteria are applied.

With regard to the statistical inference of DE, we formulated a hierarchical Bayesian model with the MRF prior, which enables module-oriented DE analysis. However, we discovered that this additional feature of MRF prior was impactful when there are a limited

number of quantitative data per protein (i.e. a small number of peptides and fragments) as illustrated in our simulation study. In other words, DIA mode offers more than a sufficient number of repeated measurements (fragments and peptides) for many proteins to support solid probabilistic decision for protein-level analysis ("data dominates the prior"). Nevertheless, the MRF model can be still useful when the number of observations per unit of analysis is relatively small, which occurs in two practical scenarios. First, when the quantitative data is rolled up to the protein or peptide level (summed over fragments and peptides), the model can incorporate the functional module information such as Gene Ontology and protein-protein interaction data in the DE analysis, assuming that the proteins in a common functional module are likely to behave similarly. Second, as we demonstrated in the glycoproteomic data, the model can be used for peptide DE analysis of post-translational modifications, using the peptide-protein group information as the module information. mapDIA's data input format was flexibly designed to accommodate various types of module information (see our software manual).

A frequently arising topic in the statistical analysis of label-free quantitative data is the treatment of missing data. Currently, we do not perform any missing data imputation or model-based treatment in mapDIA. We analyze the data using fragments with non-missing data in at least two samples within each comparison group in the IS design, or using fragments with no missing data in the REP design. While the existing missing data imputation methods such as the nearest neighbor-based approach are appealing, their performance has not been benchmarked using gold standard DIA datasets, and more fundamentally, it is difficult to judge whether such methods address the underlying missing mechanism in DIA data (e.g. de-convolution of co-eluting ions, data extraction parameters), which is non-random and associated with the data extraction pipeline such as the quality of DDA spectral library in targeted extraction, etc. Indeed, this problem can potentially be better addressed at the data extraction stage, where one can further quantify low abundance fragment data in the below-the-detection-limit range.

Finally, the current implementation of mapDIA requires that fragment intensity data be organized in the two-layered hierarchy, that is, protein to peptides and peptides to fragments. However, the software can be immediately applied to protein intensity and peptide intensity datasets. As mentioned earlier, for example, quantitative phosphoproteomics analysis requires significance scores at the peptide level, and the user can format the data with peptide sequences as protein and peptide identifiers, which will inform the software to compute scores for peptides. Likewise, protein DE analysis can be performed if protein intensities are provided along with protein ID specified as protein/peptide/fragment identifiers.

Overall, we believe that mapDIA is an attractive method for robust statistical analysis of DIA mode quantitative proteomics data. There are refinements that can be made in the mapDIA pipeline in the future, such as handling of technical/biological replicates in the IS design and improved control of the fragment selection step, more elaborate evaluation of the built-in normalization methods. More importantly, a comprehensive investigation of the interplay between various data extraction methods and the preprocessing steps in mapDIA will be of utmost interest, which will reveal the optimal integrated data analysis pipeline for this type of data from start to finish.

Chapter 3

A mass action-based model for gene expression regulation in dynamic systems

3.1 Study of time-dependent gene expression regulation

The process of RNA[1] synthesis (transcription) is closely related with protein synthesis (translation) according to the central dogma of molecular biology[2] [13]. Considering gene[3] expression[4] as an array of biochemical processes to produce gene products, regulation of gene expression is a highly complex mechanism with multiple access points through transcriptional, post-transcriptional, translational, and post-translational regulations. For instance, when cells encounter environmental stress, they are challenged to reprogram the transcriptome first (all messenger RNAs[5]) to confer increased viability and fitness in the

[0]Adapted with permission from Teo *et al*, "PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation," *J. Proteome Res.*, 2014; 13(1):29-37. Copyright (2014) American Chemical Society.

[1]RNA (ribonucleic acid): Molecule produced by the transcription of DNA; usually single-stranded, it is a polynucleotide composed of covalently linked ribonucleotide subunits.

[2]Central dogma of molecular biology: The principle that genetic information flows from DNA to RNA to protein.

[3]Gene: Unit of heredity containing the instructions that dictate the characteristics or phenotype of an organism.

[4]Gene expression: The process by which a gene makes a product that is useful to the cell or organism by directing the synthesis of an RNA molecule with a characteristic activity.

[5]Messenger RNA: RNA molecule that specifies the amino acid sequence of a protein.

new environment, and further adjust protein expression and additional post-translational regulations [6]. However, the dynamic relationship between the transcriptome and the proteome has remained elusive due to the lack of technology to measure protein expression at a scale comparable to gene expression and it is of great interest to investigate how much of transcriptional and translational regulation determine the fate of the final gene products [25, 73].

To achieve this aim, proteome-wide expression datasets must be generated with sufficient coverage and quantitative precision, especially in a time resolved manner. Thanks to recent advances in large-scale high-resolution mass spectrometry (MS), comprehensive quantitative proteomics datasets are now becoming available with longitudinal designs (e.g. following a treatment of interest) [11, 23, 25, 37, 59, 62, 73]. For example, a few recent studies used time course transcriptomic and proteomic datasets to monitor stress response in yeast and described the distinct roles of regulation at the level of RNAs and proteins where the variation in protein expression was only partially explained by transcription changes [37, 72]. While these results are intriguing, statistical analysis was limited to linear correlation or the analysis of variance in these studies, separately applied to RNA and protein datasets. In other words, there is no generalizable statistical method to jointly model the two datasets to objectively extract biological signals of regulation at different molecular levels.

With the emergence of these new datasets, time is now ripe to develop robust statistical methods to identify candidate genes that are regulated at the RNA and/or protein levels and to quantitatively dissect the different layers of gene expression control. Since the final protein concentration is the combined result of these processes, the key task is to construct a mathematical model of gene regulation, equipped with appropriate kinetic parameters for transcription, translation, and the respective degradation. In this framework, the synthesis and degradation rates can be inferred from the data and formally tested for significant changes, providing statistically rigorous interpretation of the regulation activities that

resulted in the observed concentration changes for each protein. In other words, we aim to convert expression data into information on the rates of concentration changes and regulation.

In this work, we propose a statistical modeling framework called Protein Expression Control Analysis (PECA) to identify genes putatively regulated at the RNA or protein levels based on parallel time course datasets of mRNAs and proteins. Adopting the kinetic mass-action model used in the simulation exercise by [37], PECA dissects the change in the protein concentration during each time interval (i.e. the period between adjacent time points) into two potential sources: the change in the concentration of mRNA transcripts and the change in the protein synthesis/degradation rate ratios. This deconvolution renders the inferred protein rate ratios specific to the regulation at the protein level. As explained later, the same model can be posited to infer the RNA rate ratios to infer RNA-level regulation, under the reasonable assumption that the DNA copy numbers[6] do not change over time. For both analyses, PECA derives the posterior probability that the rate ratio of synthesis versus degradation changed at each time point (before and after each time point), along with the associated false discovery rates (FDR) [45]. Hence this scoring framework leads to unbiased statistical framework of regulation changes at both molecular levels.

We remark that there are a few methods for analyzing time course datasets in the current statistics and bioinformatics literature [10, 50, 63, 66] . However, these methods are not suitable for the multi-omics data of our interest, especially for detecting regulation at the RNA and protein levels simultaneously. First, those methods are designed to analyze single source datasets (e.g. transcriptomics data alone) and they do not explicitly model the kinetic parameters of synthesis and degradation. Second, they are not able to account for the contribution of mRNA concentration changes when analyzing protein-level regulation.

[6]Copy-number variation: Large segment of DNA, 1000 nucleotide pairs or greater, that has been duplicated or lost in an individual genome.

Third, they perform statistical tests whether the expression has changed anywhere in the time course, not the temporal changes of regulatory parameters at specific time points and the direction of change, which is offered by PECA.

The rest of the chapter is organized as follows. We first present the statistical model and propose a straightforward estimation procedure using Markov chain Monte Carlo sampler. We will evaluate the performance of our approach with simulation studies and report the re-analysis of the yeast data by [37].

3.2 Method

3.2.1 Change-point model for gene expression regulation

Suppose that we have parallel gene and protein expression data $\boldsymbol{X} = \{x_{jit}\}$ and $\boldsymbol{Y} = \{y_{jit}\}$ for protein $i = 1, \ldots, I$ in replicates $j = 1, \ldots, N$ observed over time points $(h_0, \ldots, h_T)$. Time h_0 indicates the time point before the samples are treated or the baseline of subsequent time points. We assume that the protein expression measurements follow log normal distributions

$$y_{jit} \sim \mathcal{LN}\left(\eta_{jit}, \tau_i^2\right)$$

after proper normalization of the data. Our goal is to infer the protein synthesis rate κ_{it}^s and the degradation rate κ_{it}^d during the interval (h_t, h_{t+1}) of length $\Delta h_t = (h_{t+1} - h_t)$ for protein i. More importantly, the mean parameters are related between adjacent time points as follows:

$$\eta_{ji,t+1} = \eta_{jit} + \Delta h_t \left(x_{jit}\kappa_{it}^s - \eta_{jit}\kappa_{it}^d\right) \tag{3.1}$$

for $t = 0, 1, \ldots, T-1$. At time t, the mRNA abundance is x_t and the current protein abundance y_t, and we would expect that the protein abundance will increase or decrease by $x_t\kappa_{it}^s - \eta_t\kappa_{it}^d$. This is based on the mass-action kinetic action model, which underpins the simulation model of [37].

Equation (3.1) is a straightforward representation of time course profile of mean parameters as a simultaneous outcome of synthesis and degradation of each molecule. If the abundance of a protein is regulated by transcriptional regulation only, then we assume that the two parameters $\{(\kappa_{it}^s, \kappa_{it}^d)\}_{t=0}^{T-1}$ do not change over time. By contrast, if the protein is regulated by altering either the synthesis or the degradation rate, we assume that $\{(\kappa_{it}^s, \kappa_{it}^d)\}_{t=0}^{T-1}$

change over time. Translational regulation is a useful mechanism to react to sudden changes because transcriptional regulation of protein expression entails a lengthy chain of cellular processes, such as transport of mRNA from nucleus to ribosome, some biological functions require an immediate response via translation control of synthesis and degradation at the protein level [60]. Thus proteomic response to such an environment shock can be delivered by altering protein synthesis and degradation directly, rather than altering the concentration level of their precursors (mRNAs).

To detect the change in these rate parameters, we formulated a change point model to describe the probability distribution of $\boldsymbol{\kappa}_i^s = (\kappa_{i0}^s, \cdots, \kappa_{i,T-1}^s)$ as follows. We first note that κ_{it}^s and κ_{it}^d are always positive since they are rate parameters by definition, and thus the issue of identifiability arises. This is expected since we model the change in protein expression as the difference of two positive values, where there can be infinite number of solutions. Hence we impose the restriction $\kappa_{it}^d = 1 - \kappa_{it}^s$ for all i. This condition does not undermine the aim of this model since our interest is ultimately in the rate ratio $\kappa_{it}^s/\kappa_{it}^d$. Under this simplex constraint, it suffices to keep track of κ_{it}^s only. For protein i, let $\boldsymbol{C}_i$ and $|\boldsymbol{C}_i|$ denote the set of time points $\{t : \kappa_{i,t-1}^s \neq \kappa_{it}^s | 0, 1, \cdots, T-1\}$ and the size of the set, respectively. If the elements of $\boldsymbol{\kappa}_i^s$ remained constant across time, $\boldsymbol{C}_i$ is an empty set; if some elements of $\boldsymbol{\kappa}_i^s$ were distinct from others, $\boldsymbol{C}_i$ is the set of all intermediate time points from 1 to $T-1$ with different adjacent rates. Given a specific configuration of $\boldsymbol{C}_i$, we can re-parameterize this model by $\boldsymbol{\theta}_i = (\boldsymbol{C}_i, \{(\kappa'_{it})\}_{t=0}^{|\boldsymbol{C}_i|})$ where $\kappa'_{it} = \kappa_{it}^s/(\kappa_{it}^s + \kappa_{it}^d)$, which further reduces to $\kappa'_{it} = \kappa_{it}^s$ under the simplex constraint. We remark that this change point model resembles the well-known model of [29], but our model is simpler than his because change points can occur at the observed time points only. This is a reasonable choice since there are often a few time points in dynamic expression studies (often <10 time points), but the location of change points can be easily incorporated in the model for datasets with sufficiently dense time points.

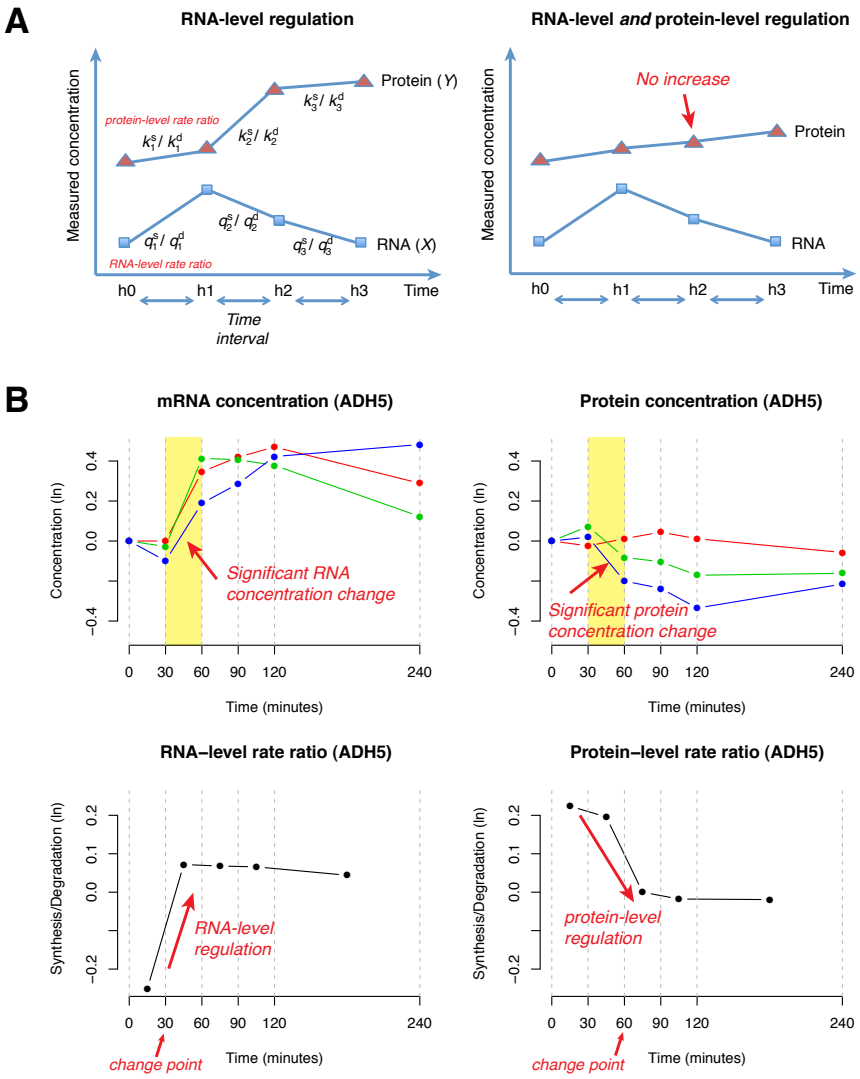

Figure 3.1: (A) Hypothetical examples of RNA-level regulation and protein-level regulation. (k_t^s, k_t^d) refer to the protein synthesis and degradation rates during the t-th time interval, respectively. (q_t^s, q_t^d) are the corresponding RNA-level rate parameters. PECA models the ratio of the two rates over the time course. The left panel illustrates the case in which protein concentration is entirely regulated at the RNA level, i.e. with no change in the protein-level kinetic parameters. The right panel shows the case in which protein concentration is regulated at both RNA and protein level, where the protein-level regulation compensates the RNA-level regulation to maintain the protein homeostasis, which is not easy to see from the concentration data alone. (B) Extraction of RNA- and protein-level regulation signals in alcohol dehydrogenase class-3 protein encoded by ADH5 gene. The upper panels of the figure show the time course of mRNA and protein concentrations of ADH5 after an osmotic shock. The lower panels show the kinetic parameters (rate ratios) during time intervals at both levels, as reported by PECA. The protein-level rate ratio is estimated accounting for the available amount of mRNA at the beginning of each time interval. The consistency of rate ratio profiles across the replicates is automatically taken into consideration in the estimation process.

3.2.2 Estimation and Inference

To estimate the model parameters, we constructed a MCMC sampler that combines standard Metropolis-Hastings updates and dimension switching updates in the form of reversible-jump MCMC [29].

First, the likelihood of the entire model is

$$(\text{likelihood}) = \prod_{i=1}^{I}\prod_{j=1}^{N}\prod_{t=0}^{T} \frac{1}{y_{jit}\tau_i\sqrt{2\pi}} \exp\left[-\frac{1}{2\tau_i^2}(\ln(y_{jit}) - \ln(\eta_{jit}))^2\right]$$

where

$$\eta_{jit} = \eta_{ji0} + \sum_{\ell=0}^{t-1} \Delta h_\ell \left(x_{ji\ell}\kappa'_{i\ell} - \eta_{ji\ell}(1-\kappa'_{i\ell})\right)$$

We specify prior distributions that are the effectively informative in our view:

$$\begin{aligned}
\eta_{ji0} &\sim \mathcal{N}(0, 100^2) \quad \text{for} \quad j = 0, \ldots, N \\
\kappa'_{i\ell} &\sim \mathcal{U}(0,1) \quad \text{for} \quad \ell = 0, \ldots, |\boldsymbol{C}_i| \\
\tau_i^{-2} &\sim \mathcal{G}(a_\tau, b_\tau)
\end{aligned}$$

for fixed $\boldsymbol{C}_i$ for all i, where $\mathcal{N}, \mathcal{U}, \mathcal{G}$ denote normal, uniform, and gamma distributions respectively. We also assume that the change point configuration $\boldsymbol{C}_i$ has the following prior:

$$\pi(\boldsymbol{C}_i) \propto \varphi^{|\boldsymbol{C}_i|}(1-\varphi)^{T-1-|\boldsymbol{C}_i|}$$

where we set $\varphi = 0.5$ assuming that nothing is known *a priori* about the chance of having a change point in any of the proteins.

To elicit the hyperprior parameters (a_τ, b_τ), we first calculate the sample variance of

the protein intensities across all time points in each replicate and plug in the maximum likelihood estimates for the shape parameter a and scale parameter b:

$$(a_\tau, b_\tau) = \arg\max_{a,b} \left[\left(\frac{b^a}{\Gamma(a)} \right)^{N \times I} \left(\prod_{i,j} v_{ij} \right)^{-\alpha-1} \exp\left(-b \sum_{i,j} \frac{1}{v_{ij}} \right) \right]$$

where v_{ij} is the sample variance for protein i replicate j.

In summary, the prior can be written as

$$\text{(prior)} \;\propto\; \prod_{i=1}^{I} \left\{ \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} (\tau_i^2)^{-a_\tau-1} e^{-\frac{b_\tau}{\tau_i^2}} \cdot \prod_j \phi\left(\frac{\eta_{ji0}}{100}\right) \cdot \varphi^{|\boldsymbol{C}_i|} (1-\varphi)^{T-1-|\boldsymbol{C}_i|} \right\}$$

where the prior for $\{\kappa'_{it}\}$ is omitted conditional on the fact that they are all on the unit interval, and ϕ denotes standard normal density.

The model parameters are updated in the following order:

$$\{\eta_{ji0}\}_{j=0}^{N} \to \tau_i^2 \to \{\kappa'_{it}\}_{t=0}^{T-1} \to \boldsymbol{C}_i$$

for all i. This whole cycle is repeated for 5,000 iterations for burn-in period and $M = 20,000$ iterations for the main iteration with thinning of 20 samples, in both simulation and data analysis sections that follow. We use hat and tilde symbols to denote current and proposal values respectively.

1. We first start with η_{ji0} by a Metropolis-Hastings step, with proposal value $\tilde{\eta}_{ji0}$ drawn from $\mathcal{N}(\hat{\eta}_{ji0}, 0.1^2)$, and compute the Metropolis-Hastings ratio to complete the update. Since this parameter is involved in the mean values at all time points, the likelihood has to be evaluated at all time points for updating each of these parameters.

2. Next, we draw the variance parameter τ_i^2 by Gibbs sampling from inverse gamma distribution $\mathcal{IG}(a_\tau + N(T+1)/2, b_\tau + \sum_{j,t}(y_{jit} - \eta_{jit})^2/2)$.

3. Next, we draw $\{\kappa'_{i\ell}\}$ for $\ell = 0, \ldots, |\boldsymbol{C}_i|$ under the fixed $\boldsymbol{C}_i$ for each protein i. We use random walk Metropolis-Hastings steps to update them, i.e. draw a proposal value $\tilde{\kappa}'_{i\ell}$ from $\mathcal{N}(\hat{\kappa}'_{i\ell}, 0.1^2)$ and accept or reject afterwards.

4. Finally, we update the change point configuration $\boldsymbol{C}_i$. There are two different moves: birth of a new change point and removal (death) of an existing change point. Since these two moves are reversible in notation, we just describe the birth move here. Suppose that $\hat{\kappa}'_{i\ell}$ covers a time period (h_t, h_{t+m}) that contains at least one observation time(s). Then we propose a birth of a new change point $h^* \in \{h_{t+1}, \ldots, h_{t+m-1}\}$ within the interval (chosen from one of the intermediate time points) and break the current rate parameter into two daughter parameters, namely $(\tilde{\kappa}'_{i\ell}, \tilde{\kappa}'_{i,\ell+1})$ where it is required to meet

$$(h^* - h_t) \cdot \text{logit}(\tilde{\kappa}'_{i\ell}) + (h_{t+m} - h^*) \cdot \text{logit}(\tilde{\kappa}'_{i,\ell+1}) = (h_{t+m} - h_t) \cdot \text{logit}(\hat{\kappa}'_{i\ell})$$

with a random perturbation such that

$$\frac{\tilde{\kappa}'_{i,\ell+1}}{1 - \tilde{\kappa}'_{i,\ell+1}} = \frac{1-u}{u} \frac{\tilde{\kappa}'_{i\ell}}{1 - \tilde{\kappa}'_{i\ell}},$$

with $u \sim \text{Uniform}(0, 1)$. Under this transformation, the Jacobian is $\frac{(\tilde{\kappa}'_{i\ell}(1-\tilde{\kappa}'_{i\ell}) + \tilde{\kappa}'_{i,\ell+1}(1-\tilde{\kappa}'_{i,\ell+1}))^2}{\tilde{\kappa}'_{i\ell}(1-\tilde{\kappa}'_{i\ell})}$ for $(\hat{\kappa}'_{i\ell}, u) \to (\tilde{\kappa}'_{i\ell}, \tilde{\kappa}'_{i,\ell+1})$. Hence the Metropolis-Hastings ratio for the birth move just equals the posterior ratio times the Jacobian since the acceptance probability of this proposal is

$$\min\{1, \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio} \times \text{Jacobian}\},$$

where the prior and proposal ratios are the ratios of Uniform distribution over unit

intervals. Then the Metropolis-Hastings ratio becomes

$$\prod_{j,t}\left[\exp\left\{-\frac{1}{2\tau_i^2}(\ln(y_{jit})-\ln(\eta_{jit}))^2\right\}\right]\frac{\varphi}{1-\varphi}\frac{(\tilde{\kappa}'_{i\ell}(1-\tilde{\kappa}'_{i\ell})+\tilde{\kappa}'_{i,\ell+1}(1-\tilde{\kappa}'_{i,\ell+1}))^2}{\hat{\kappa}'_{i\ell}(1-\hat{\kappa}'_{i\ell})}.$$

Using the samples drawn from the posterior distributions, we perform statistical inference as follows. Our main goal is to identify the time points where the protein rate ratio shifts, i.e. $p_{it} = P(\kappa^s_{it} \neq \kappa^s_{i,t+1}|X_i, Y_i)$, where X_i and Y_i denote the gene and protein expression data for protein i, respectively. This score has the nice property that it is a marginal probability computed after accounting for the data and change point configurations at all time points. Instead of seeking the *maximum a posteriori estimate* of $\boldsymbol{C}_i$, we perform our inference based on this probability. Denote the posterior samples of $\{\kappa'_{it}\}$ by $r^{(1)}_{it}, \ldots, r^{(M)}_{it}$ for each κ'_{it}. We first compute p_{it} by $\hat{p}_{it} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{r^{(m)}_{it} \neq r^{(m)}_{i,t+1}\}$. If $\hat{p}_{it} \geq p^*$ holds for at least one t, where p^* is the probability threshold, we consider protein i to be translationally regulated. To determine an optimal threshold, we compute the Bayesian false discovery rate (BFDR) as

$$BFDR(p^*) = \frac{\sum_{i,t}(1-\hat{p}_{it})\delta_{it}(p^*)}{\sum_{i,t}\delta_{it}(p^*)} \tag{3.2}$$

where $\delta_{it}(p^*) = 1\{\hat{p}_{it} \geq p^*\}$ [26, 45]. This decision rule $\delta_{it}(\cdot)$ results in the selection of specific time points where translation regulation shifted from the preceding time period. Furthermore, we can perform functional clustering [3] using this surrogate data $\{\kappa'_{it}\}$ instead of the raw expression data $\boldsymbol{Y}$ by the agglomerative hierarchical clustering [21] with the Euclidean distance metric on the matrix data $\{\kappa'_{it}\}$ for the selected proteins, ultimately identifying different groups of proteins with a similar translational regulation pattern.

3.2.3 Simulation study

We first conducted simulation studies to evaluate the sensitivity and specificity of the method. We simulated expression datasets for K transcripts (mRNAs) and proteins in parallel in single biological replicate across six different time points. Among these, we created three groups that are different in terms of the translational control mechanism, which emulated the protein expression profiles of up- and down-regulated proteins in [37]. Specifically, each of the three groups represents a different combination of transcriptional and translational regulation. Protein expression in Group 1 is regulated entirely by transcriptional regulation (gene up-regulation). Protein expression in Group 2 is translationally up-regulated by an increased rate of translation during the first time period in addition to the transcriptional up-regulation. This pattern is expected to occur in immediate shock conditions when direct translational regulation is required. Finally, Group 3 represents the case of down-regulation in both data, where down-regulation of protein expression was driven by increased degradation rates in the late time points as well as transcriptional down-regulation in the early time points.

Group	Size	μ_0	μ_1	μ_2	μ_3	μ_4	μ_5
1	500	1.00	1.25	1.20	1.10	1.00	1.00
2	500	1.00	1.25	1.20	1.10	1.10	1.10
3	500	1.00	0.75	0.80	0.90	0.90	0.90

Table 3.1: Mean parameters of gene expression data in the three groups.

Group	κ_0^s	κ_1^s	κ_2^s	κ_3^s	κ_4^s
1	1.00	1.00	1.00	1.00	1.00
2	r^*	1.00	1.00	1.00	1.00
3	1.00	1.00	r^{*-1}	r^{*-2}	r^{*-2}

Table 3.2: Protein synthesis rates in protein expression data in the three groups with fixed degradation rate $\{\kappa_t^d\} = 1$ at all time points. Essentially r^* plays the role of protein rate ratio.

Here we describe the data generation process in detail. We simulated gene expression

data reflecting the burst of up- and down- regulation of mRNAs between the first two time points, from log normal distribution with their respective mean parameters in each group as specified in Table 3.1 along the time course, and the variance parameters fixed at $\sigma^2 = 0.1$. To simulate protein expression data according to the turnover mechanism, we set the translation and degradation rates (κ^s, κ^d) as tabulated in Table 3.2, in which the protein synthesis rate changes by a factor of r^*. We fixed κ^d at 1, and thus r^* essentially represents the "scaled" rate ratio. This leads to the time-dependent mean expression values following the relationship in Equation (3.1). Using these means, we simulated protein expression data from log normal distribution, where different variance parameters τ^2 were attempted to control the signal-to-noise ratio. Based on Equation (3.1), the ratio e^{τ}/r^* can be interpreted as a variant of the coefficient of variation (CV), provided that the gene and protein expression data are properly scaled. We have evaluated the performance at different CVs, where r^* ranged from 1.5 to 2.0 and τ^2 ranged from 0.01 to 0.2. In each scenario, we looked at three different probability thresholds $p^* = 0.5, 0.6, 0.7$.

Figure 3.2 shows the results. The sensitivity, specificity, and $BFDR$ estimates in the figure were computed by averaging the results over 100 simulations of each setting. The MCMC sampler converged quickly to the posterior distribution, as illustrated in Figure 3.3 (left panel). First, any detection in Group 1 represents false positives. For Group 2, r^* increased sharply during the first time period (h_0, h_1) and thus the second time point h_1 is the true change point. Likewise for Group 3, r^* decreased from unit rate twice at h_2 and h_3. Hence any detection at these time points at Groups 2 and 3 represents true positives. To see the range of the CVs we cover in the simulation, consider the worst case scenario with $r^* = 1.5$ and $\tau = 0.2$. This means that the protein rate ratio increases by 50%, yet the standard deviation of the error is at about 22% ($e^{0.2} \approx 1.22$). In this case, the level of translational regulation signal will be masked by the noise. By contrast, in the scenario with $r^* = 2.0$ and $\tau = 0.01$, the protein rate ratio increases by 100% and the noise is ignorable (1%).

As expected, the proposed model performed very well in the scenarios with low CV ($\tau = 0.1$ or below), achieving almost perfect specificity (>97%) and good sensitivity (>80%) with increasing r^*. Interestingly, the sensitivity for down-regulation in Group 3 at h_3 (not shown) was very low compared to the sensitivity at h_2, even though the rate ratio went down by the same factor of r^*. This is possibly because the gene expression level increased at h_3 from 0.8 to 0.9 in the simulation scheme, compensating the decrease in protein turnover. Finally, the estimated $BFDR$ was not trivially small at all three thresholds, ranging from 12% to 36% across the scenarios (last panel of Figure 3.2). However, since very few false positives were detected in Group 1 in the modest signal-to-noise ratio settings (top panel), these estimates can be considered to be conservative.

Overall, our method showed good sensitivity and specificity for the scenarios with modest signals at all probability score thresholds. The result also suggests that the optimal threshold can be set as low as (~ 0.6) in the scenarios with a high signal-to-noise ratio, even if the associated $BFDR$ estimates may be greater than conventional FDR targets such as 5%. However, in the scenario where large variation in the mRNA abundance coexists with protein expression changes, a selection criterion that controls $BFDR$ reasonably low will be desired. We illustrate such a case in the next section.

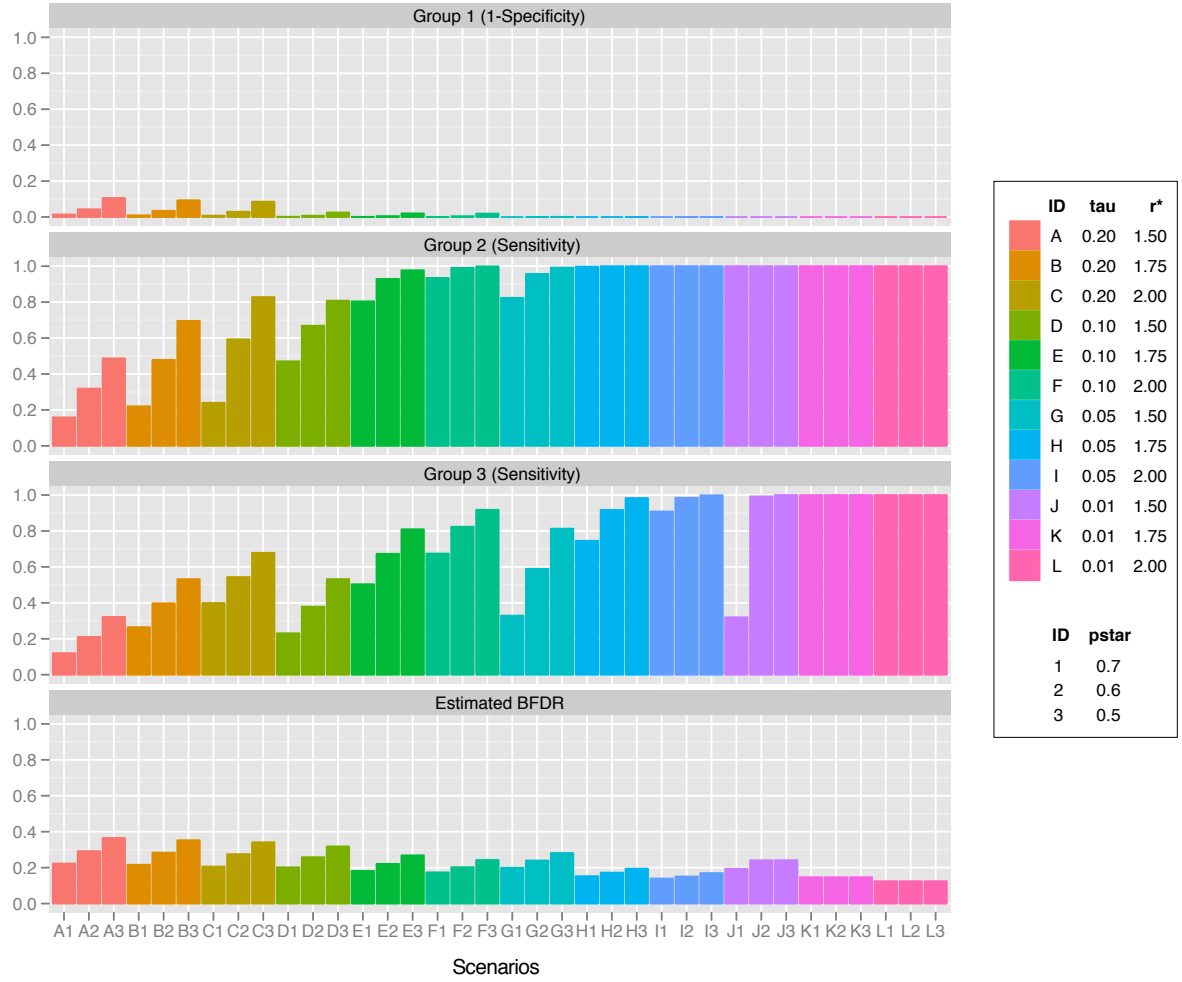

Figure 3.2: Simulation results. Proteins selected from Group 1 across all time points are false positives. Proteins selected from Group 2 at h_1 and Group 3 at (h_2, h_3) are true positives.

3.3 Application: analysis of osmotic shock in yeast

Next we reanalyzed the yeast dataset which profiled the cellular response to an osmotic shock using three biological replicates [37]. In the experiment, 0.7M NaCl was applied to budding yeast in growth medium, where the dose of salt provides a robust physiological response, but results in high viability and eventual resumption of cell growth. Samples were collected before and at 30, 60, 90, 120, and 240 minutes after treatment to capture cells acclimated to the new environment, and the samples were divided for gene and protein expression measurements using microarray and quantitative mass spectrometry respectively. For data analysis, the authors performed modified t-test to select proteins that are differentially expressed before and after the treatment, and used mRNA expression as a post-hoc analysis to show correlated changes therein. In our analysis, we first analyzed the data for protein-level regulation inference, treating microarray data as $\boldsymbol{X}$ and mass spectrometry data as $\boldsymbol{Y}$ as described in the method above. We also analyzed the data for RNA-level regulation, using microarray data as $\boldsymbol{Y}$ and a fictitious DNA copy number dataset filled with constant element 1 as $\boldsymbol{X}$ (0 on log scale). This represents the assumption that the DNA copy number remains constant along the time course in the genome, which is a realistic assumption in normal cell populations.

Gene expression was measured using a custom Nimblegen tiled microarray platform (Gene Expression Omnibus GSE23798). Instead of quantile normalization the authors used, the data was further examined for systematic shift in expression level distribution across different samples, but it was deemed that no further normalization was necessary since such normalization may remove real signals due to the burst in transcriptional regulation upon osmotic shock [41]. Protein expression was measured using an isobaric tagging and liquid chromatography and mass spectrometry on an Orbitrap Velos instrument. The mass spectrometry analysis was performed simultaneously for the samples of each biological replicate taken at different time points. This batch analysis is beneficial in time course designs since a multiplexed design can control the within-sample variation better than

other designs such as label-free protein quantitation. From this experiment, 1,999 proteins were quantified consistently across the time points in at least two replicates, and among those we analyzed 1,508 proteins with no missing data across all three replicate samples.

For model parameter estimation, we ran the MCMC for 20,000 iterations with thinning (every 10th sample) after 5,000 iterations for burn-in period. We elicited the same prior distributions used in the simulation studies, and the acceptance rates for the Metropolis-Hastings updates (13%) and reversible jump MCMC (21%) remained reasonably good (before thinning of the chain). We performed visual inspection of model fit by plotting the estimated level of protein expression $\{\eta_{jit}\}$ against the observed values and found that the fit was reasonably good. We also confirmed the convergence of the MCMC sampler to the posterior distribution by the trace plot of the log likelihood (Figure 3.3).

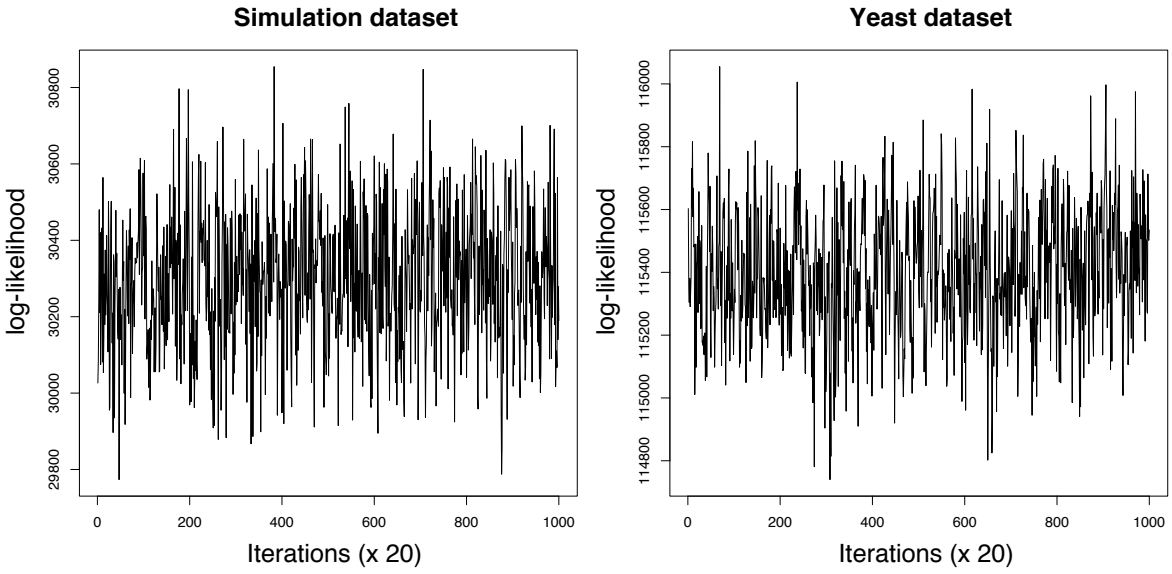

Figure 3.3: The log-likelihood trajectory of the model shows that the parameter values were drawn from the appropriate posterior distributions in the simulation data and the yeast dataset.

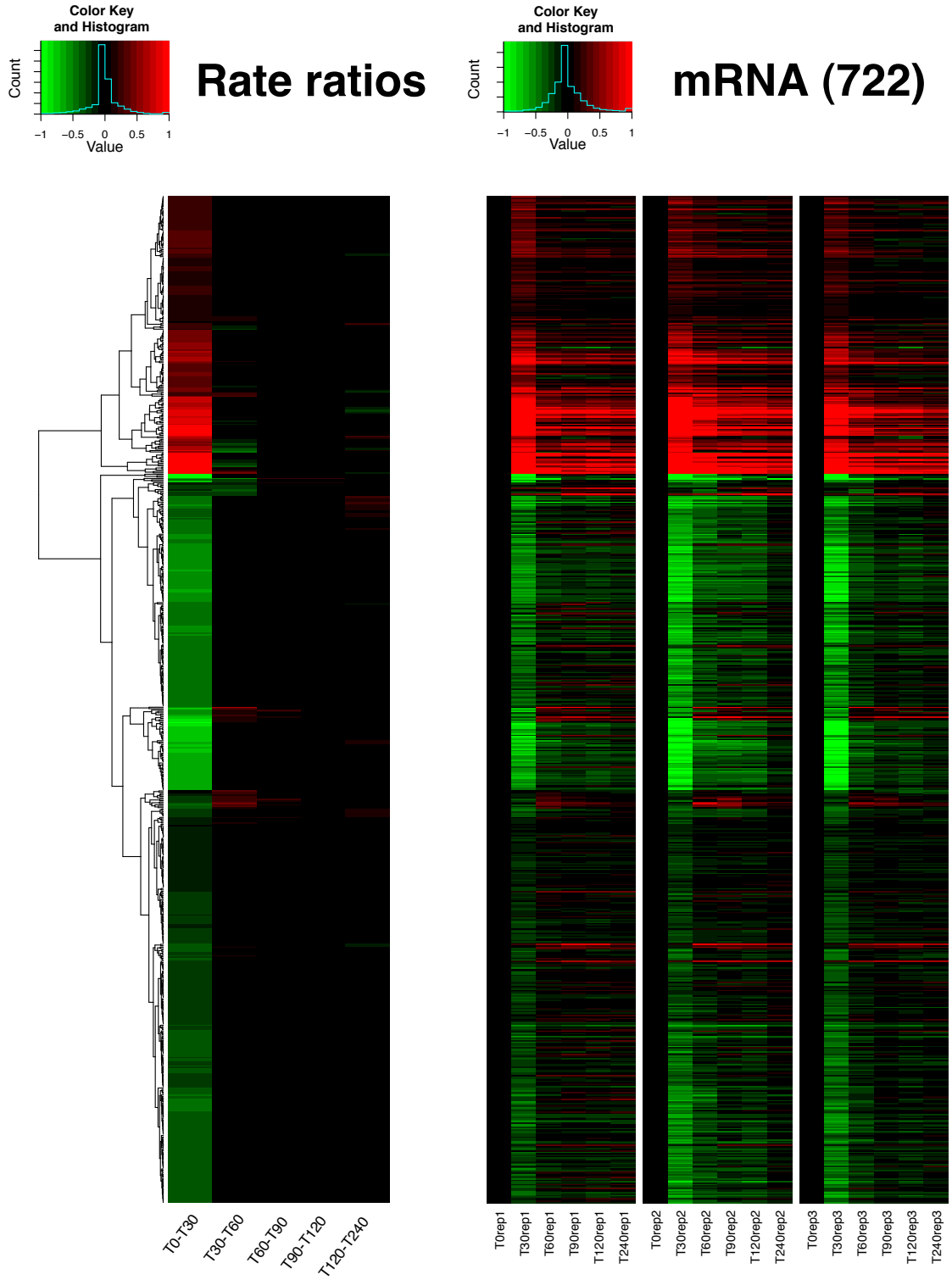

Figure 3.4: Heatmaps of the 722 stress induced and repressed proteins subject to RNA-level regulation. The left panel is the rate ratios at the RNA-level, estimated for each time interval (between two adjacent time points). The right panel is the mRNA data.

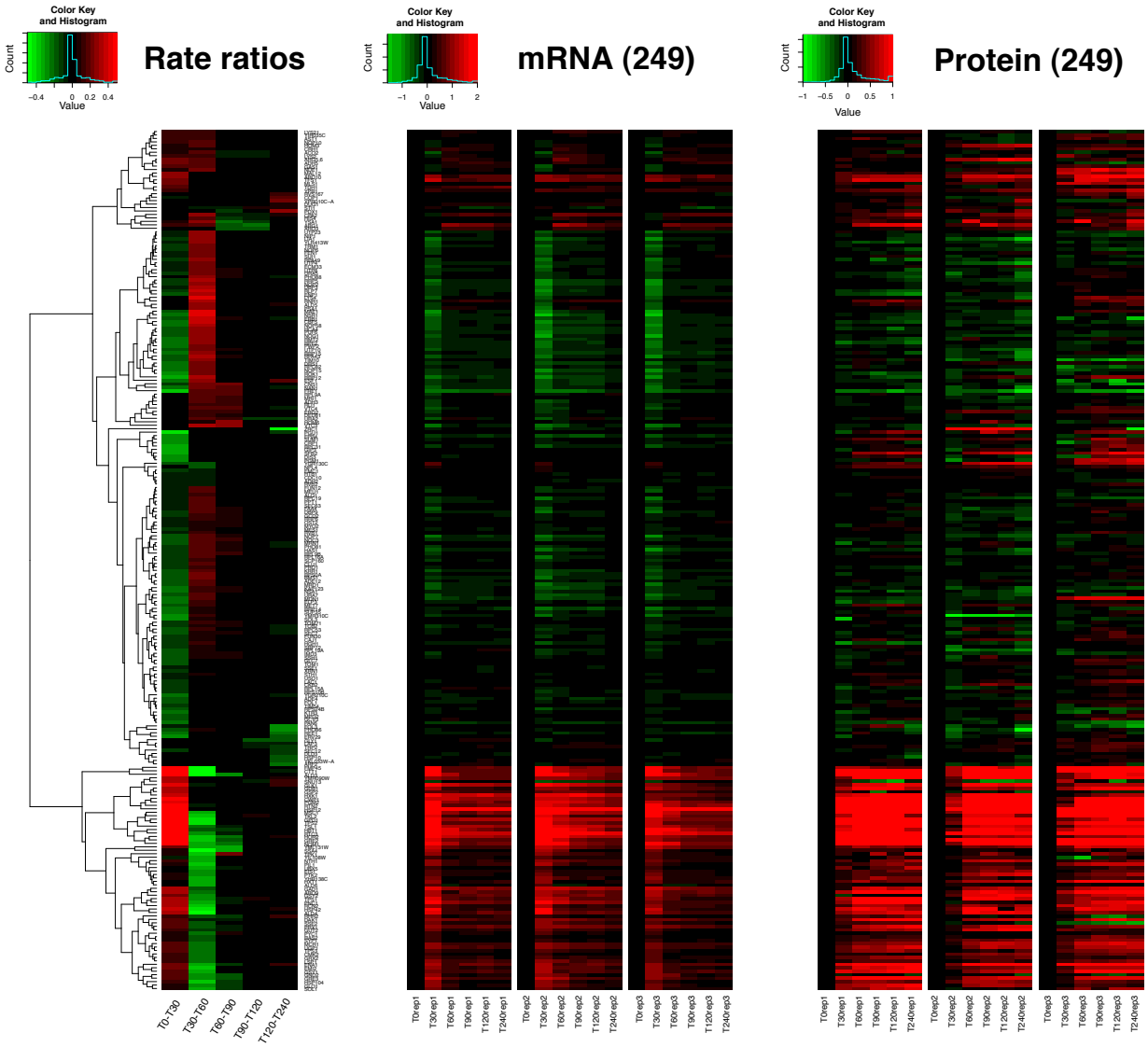

Figure 3.5: Heatmaps of the 249 stress induced and repressed proteins subject to protein-level regulation. The left panel is the rate ratios at the protein-level, estimated for each time interval (between two adjacent time points). The middle and right panels are the mRNA and protein expression data.

3.3.1 Scoring protein-level regulation changes

Using the output above, we extracted candidate genes subject to protein-level regulation. We selected 249 out of the 1,508 proteins with the posterior probability >0.8 at any of the time points as proteins putatively regulated at the protein level, controlling the FDR at 10%. To see gene function enrichment in the proteins regulated at the protein level, we selected two clusters of 68 and 131 proteins that were up and down regulated respectively at 30 minutes. All 1,508 proteins in the protein/RNA dataset served as the background list for hypergeometric test. Similar to the transcriptome analysis above, up-regulated proteins showed enrichment of stress-related functions (p-value < 0.001). By contrast, down-regulated proteins showed enrichment of the terms related to RNA processing and regulation of translation, indicating immediate shutdown of translation activities under high osmolarity (p-value < 0.001).

Similar to the RNA data, we found that the major change in the protein-level rate ratios also occurred immediately after the treatment (0∼30 min). In addition, most proteins regulated at the protein level (161/199) were also significantly regulated at the RNA level, implying that the regulation of gene expression during osmotic stress response was highly coordinated at both levels, particularly at early time points. However, as only 249 genes (17%) were regulated at the protein level while 722 genes (49%) were at the RNA-level at much more stringent FDR, one may hypothesize that transcriptional reprogramming is the dominant response to osmotic stress and ultimately protein concentrations change only by carefully selected paths of protein-level regulation.

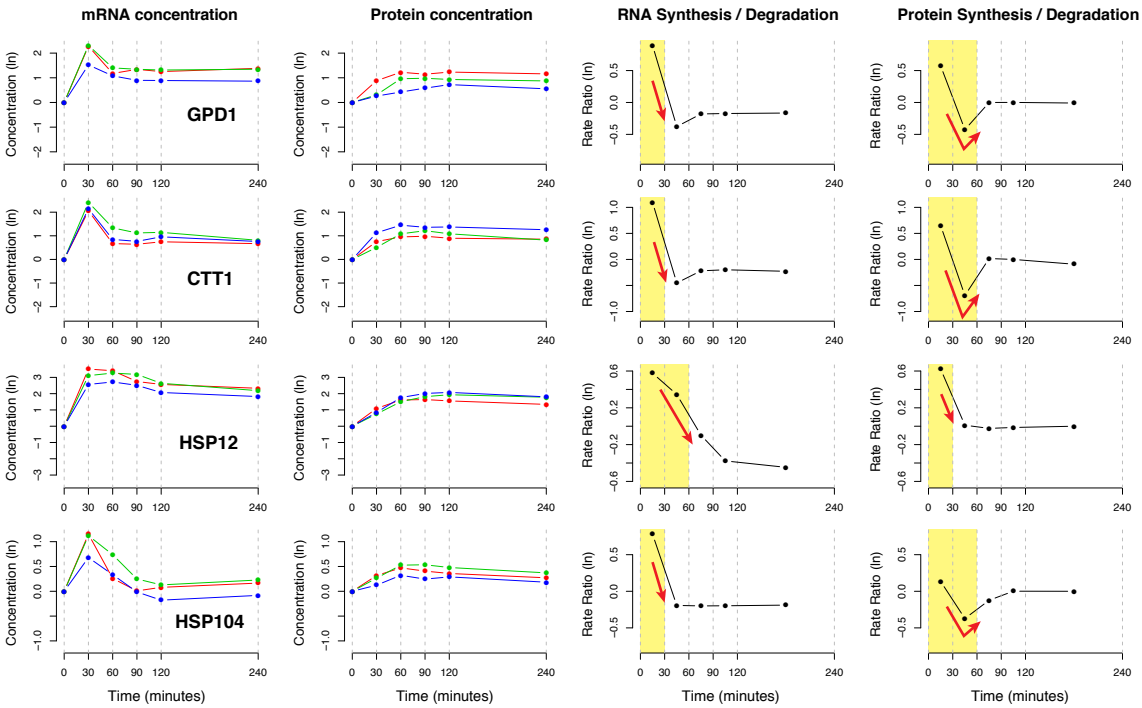

Figure 3.6: The mRNA and protein concentration data and estimated rate ratios at both levels of regulation for GPD1, CTT1, HSP12, and HSP104. These are four proteins with osmotic shock-induced expression. Blue, red, and green curves are time course data for each biological replicate. Yellow background indicates the time intervals during which the rate ratios deviated from the average range across the time course. Red arrows indicate significant regulation change at the RNA and protein level in each gene.

3.3.2 Characterizing the link between the regulatory processes

Next, we inspected the correlation between the regulatory patterns using the rate ratio profiles within the same molecules. Figure 3.6 shows the RNA and protein concentrations, and the rate ratios for four key proteins known to be up-regulated during osmotic stress response [54]: glycerol-phosphate dehydrogenase GPD1, cytosolic catalase T CTT1, heat shock proteins HSP12 and HSP104. Time intervals where the rate ratios changed significantly were indicated by yellow rectangles - illustrating that the RNA level up-regulation was most active during the first 30 minutes and subsided afterwards, with mRNA concentration recovering the stability within 60 minutes. By contrast, protein-level regulation was also the most active during the first time interval but it counterbalanced the RNA-level regulation in the opposing direction (down) during the next time interval, resulting in stabilized protein level concentrations. This pattern suggests that protein-level regulation buffered the abrupt change at the RNA level and contributed to the stable protein concentration levels.

The possible role of buffering by protein-level regulation was even more pronounced for down-regulated mRNAs, consistent with Lee et al.'s observation of less correlation between mRNA and protein concentrations for down-regulated RNAs [37]. For example, PECA provides strong evidence of protein-level regulation that resulted in stable protein concentration for four members of the large subunit of ribosome (RPL9A, RPL9B, RPL16A, RPL19A) and several subunits of RNA polymerase[7] I and III subunits (RPA43, RPA49, RPC19, RPC53, and RPC82; figures 3.7 and 3.8). In these examples, mRNA concentration decreased significantly at the 30 minutes and recovered to the pre-treatment level at 60 minutes, whereas protein concentrations hardly changed. The rate ratio profiles reported by PECA showed that there was substantial protein-level up-regulation between 30 minutes and 60 minutes to fend off the effect of reduced mRNAs during the same time interval.

[7]RNA polymerase: Enzyme that catalyzes the synthesis of an RNA molecule from a DNA template using nucleoside triphosphate precursors.

In sum, RNA-level and protein-level regulation were orchestrated together in the early response in this data, but the protein-level regulation clearly acted as the buffer to the vast transcriptome changes in this dataset. Figure 3.9 shows the correlation patterns between RNA-level regulation and protein-level regulation across the time points. Since most regulation activities occurred in the first time interval in this dataset, we focus on the first row of the figure, i.e. correlation between protein-level regulation with the RNA-level regulation during the first time interval. Consistent with the evidence from Figure 3.5, the top left panel shows that RNA and protein expression were consistently up- or down-regulated in many genes during the first time interval with positive correlation ($r = 0.51$). The negative correlations in the next two panels clearly illustrate that the RNA-level regulation of the first time interval were countered by protein-level regulation of opposite direction of the second and third time interval ($r = -0.78, -0.37$). In those intervals, the majority of the buffering effect was for the RNA-level down-regulation (countered by protein-level up-regulation), suggesting proteome-wide evidence of proteostasis through protein-level regulation. Interestingly, the positive correlations in the remaining two panels with large time lags (last two in the first row) suggest that the effect of RNA-level down-regulation takes a long time to come through at the protein concentration.

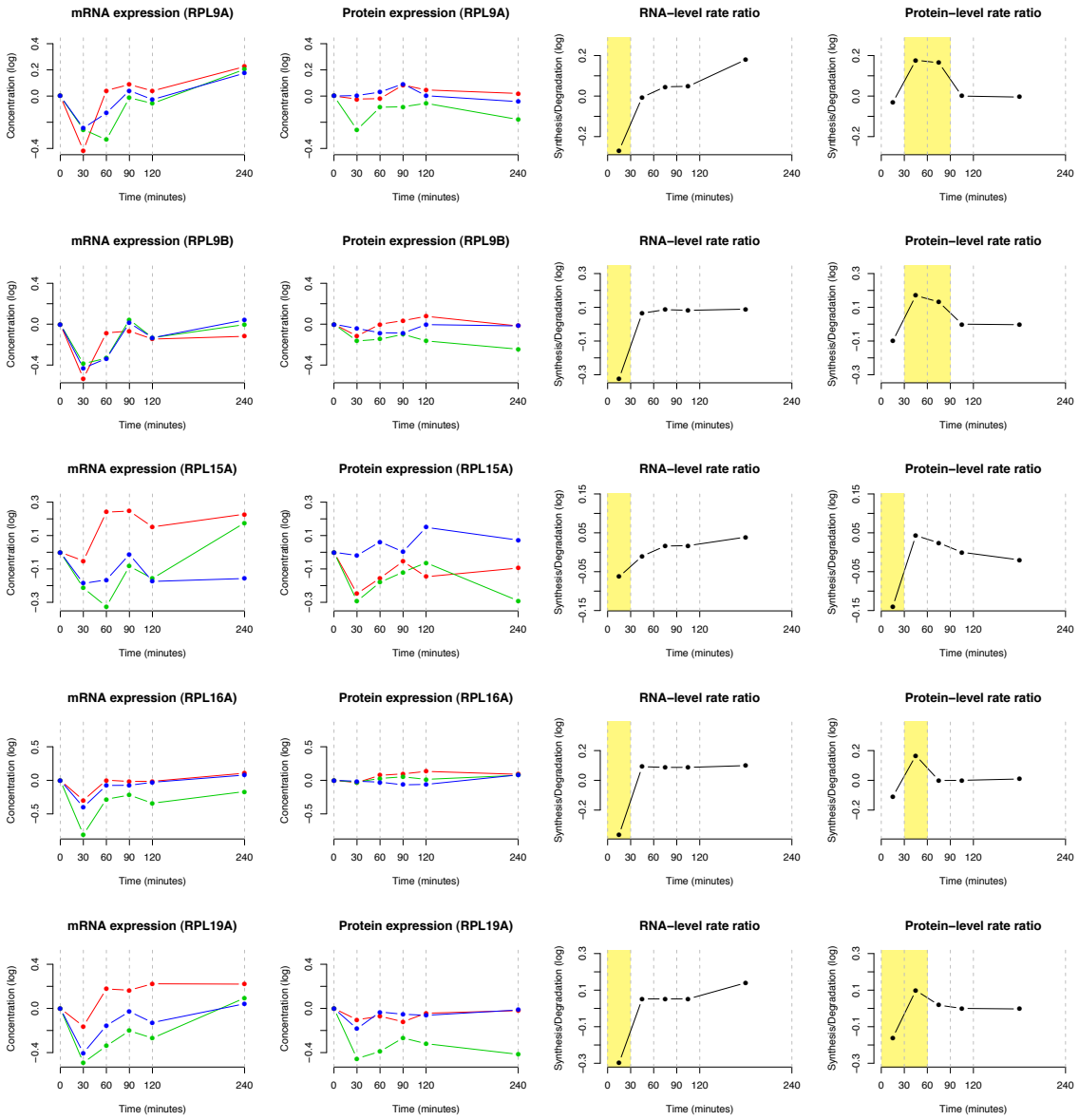

Figure 3.7: The mRNA and protein expression and estimated rate ratios at both levels of regulation for RPL9A, RPL9B, RPL16A, RPL19A, which are members of the large subunit of ribosome.

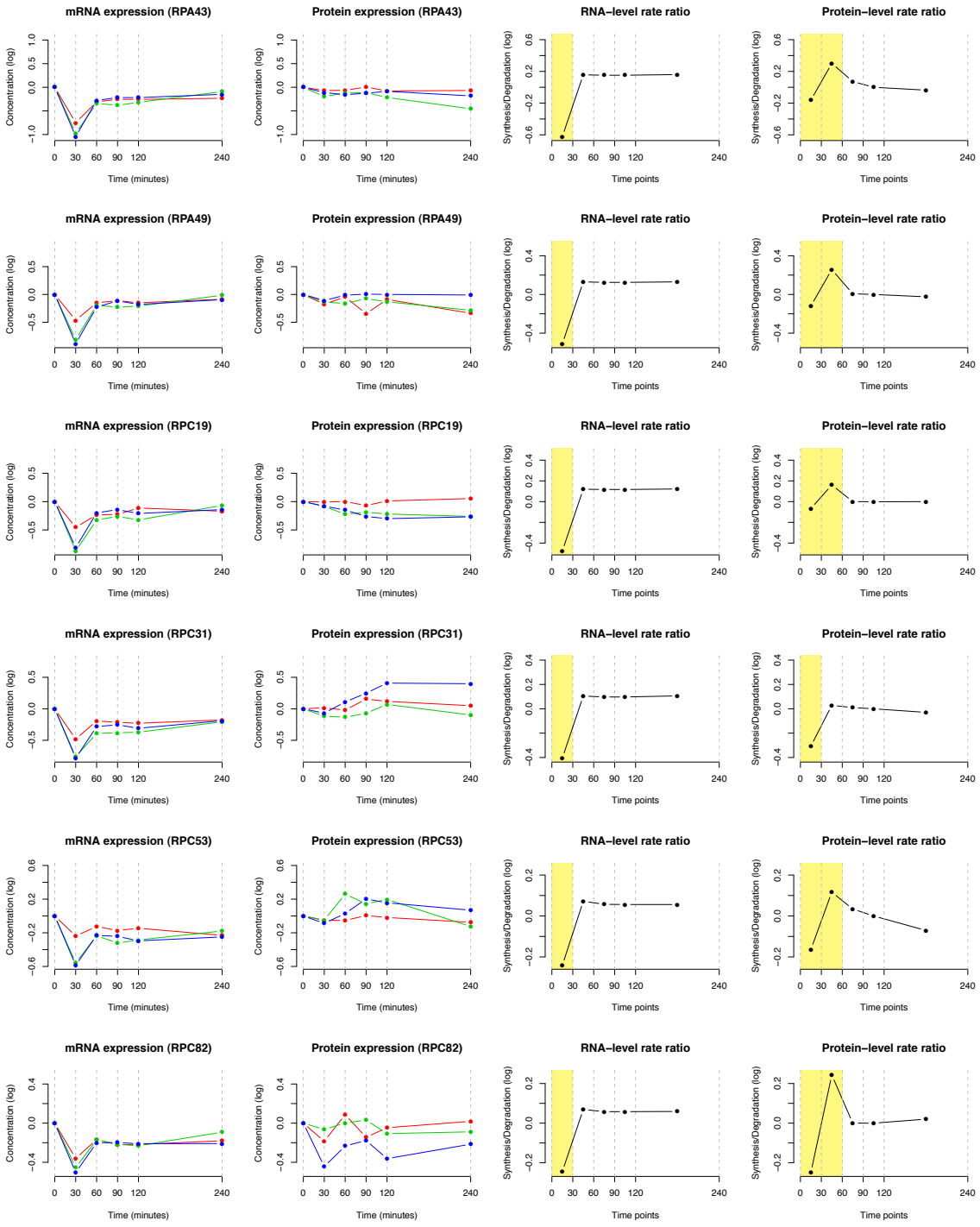

Figure 3.8: The mRNA and protein expression and estimated rate ratios at both levels of regulation for RPA43, RPA49, RPC19, RPC53, and RPC82, which are subunits of RNA polymerase I and III.

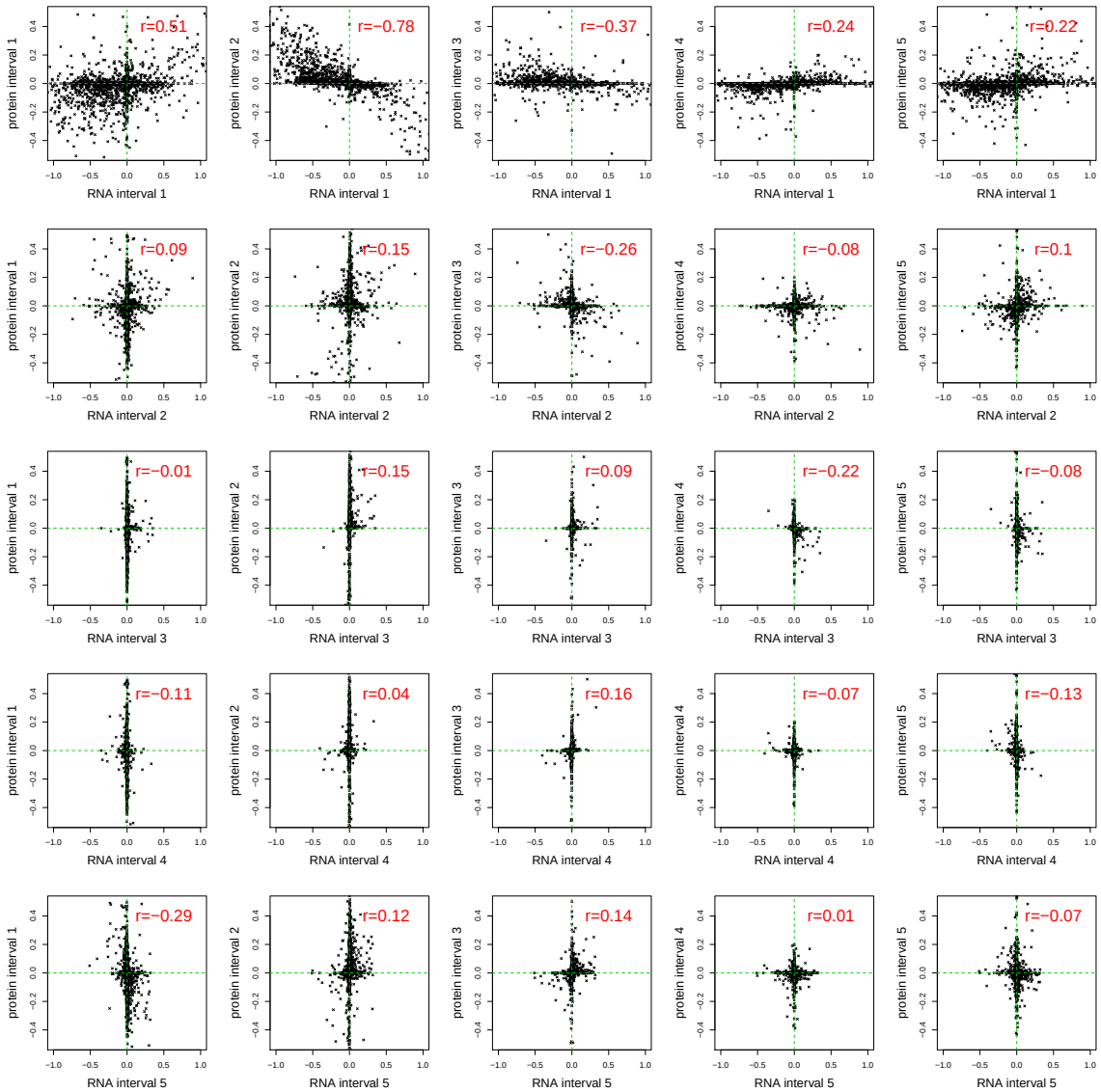

Figure 3.9: *S. cerevisiae* data with osmotic stress. The panels were arranged so that each row and column corresponds to each time point respectively. In each panel, the protein rate ratios were plotted against the RNA rate ratios (transformed by log base 2, then centered by median in each protein). The panels on diagonal positions show coupling at the same time point, whereas the panels on off-diagonal positions show buffering at different time points or time-delayed correlation.

3.4 Discussion

In this work, we proposed a statistical method to describe the patterns of gene expression regulation with respect to mass action kinetics of individual genes. Our method carries out probabilistic inference for the essential kinetic parameters of synthesis and degradation using time course data, extracting the proteins with statistically significant evidence of translational regulation. While the method has been demonstrated using paired gene and protein expression data, the kinetic relationship is also applicable to other types of paired data such as DNA copy number and gene expression data as illustrated in the analysis of osmotic shock, where one type of molecule is the precursor of the other by the central dogma. In this framework, we identify the signals of expression regulation in terms of synthesis/degradation rates instead of mean expression values, which provides biologically more interpretable results in temporal expression data.

We formulated the change point model to perform probabilistic inference and appropriate control of FDR. As illustrated in both simulation and yeast data analysis, the MCMC sampling procedure is straightforward and efficient with good mixing rates and it showed swift convergence to the stationary distribution after starting from arbitrary initial points. In the simulation study, we showed that the method is able to detect protein-level regulation activities in scenarios with reasonably modest signal-to-noise ratios. We also validated this methodology using a yeast dataset where cells were challenged to adapt to a sudden increase in osmolarity. Our method recovered a profile of translational regulation in a highly variable system where excessive gene expression changes occurred yet not all of them led to protein expression changes as expected.

A few components in the the statistical model need further improvement. First, the model specification includes the constraint $k_t^d + k_t^s = 1$ for all t, which was introduced to address the identifiability problem. In the absence of this condition, both parameters must be estimated independently at each time point, which has no unique numerical solution as

explained above. The imposed condition mirrors the assumption that the total regulation activity (synthesis and degradation) adds up to a constant at all time points, which has to be modified if estimation of absolute rates of synthesis and degradation is of interest. Second, the prior for change points $\pi(\boldsymbol{C}_i) \propto \varphi^{|\boldsymbol{C}_i|}(1-\varphi)^{T-1-|\boldsymbol{C}_i|}$ with $\varphi = 0.5$ reflects the assumption that any change point arrangement with the same number of total change points has the same prior probability. This specification can deviate from biological reality in dynamic systems during perturbations, in which expression changes are induced in the early response more often than in the late response. Such prior information can be extracted from the data itself via an Empirical Bayes approach, or careful elicitation on φ can also be an alternative remedy. We leave these aspects for future investigation.

Chapter 4

Conclusion

This thesis presented Bayesian hierarchical models for the analysis of large-scale genomic and proteomic data from high-throughput experiments in modern molecular biology. Due to the limited sample size and complex structure in the data, hierarchical Bayes can be a powerful modeling framework. The preceding chapters clearly demonstrate that, with modern computing and efficient sampling method, Bayesian inferential methods are viable for such large scale datasets. Besides hierarchical Bayes, the proposed methods integrates a collection of well known statistical techniques, including MRF model (chapter 2), iterated conditional modes (chapter 2) and reversible jump Markov chain Monte Carlo-based change point analysis (chapter 3).

In chapter 2, we described a software package mapDIA that performs essential data preprocessing, including novel retention time-based normalization method and a sequence of peptide/fragment selection steps, and more importantly, hierarchical model-based statistical significance analysis for multi-group comparisons under representative experimental designs. The advanced modeling technique also allows the user to incorporate relational information such as existing network data or protein-peptide mapping, which enables module-oriented analysis of differential expression. Using a comprehensive set of simulation datasets, we showed that mapDIA provides reliable classification of differentially expressed proteins with accurate control of the false discovery rates. Together with the analysis of two SWATH-MS datasets of 14-3-3 dynamic interaction network and prostate cancer glycoproteome the results showed the mapDIA performed better than the frequentist

based, regression framework implemented in the MSstats package.

In chapter 3, a model-based method was developed to simultaneously analyze time course transcriptomic and proteomic datasets to quantitatively dissect the contribution of RNA-level and protein-level regulation to the variation in gene expression. The statistical method is based on a mass action-based model for protein synthesis and degradation rates of individual genes, and change points in the stochastic process of the kinetic parameters are derived to identify distinct patterns of regulation of gene expression in time course profiles. A sampling-based inference procedure using MCMC was implemented and the posterior probabilities of change points in the ratio of protein synthesis and degradation are used to control the Bayesian false discovery rate.

In both the preceding chapters, the hierarchical prior enables the borrowing of strength across the genome or the proteome and carefully chosen prior that allows for an explicit form of the posterior distribution or efficient implementation of Markov chain Monte Carlo allowed parameter estimation despite the presence of a large number of parameters. In sum, these examples show that hierarchical Bayes, coupled with other complimentary statistical inferential techniques, can be a desirable model framework for use in complex biological data analysis involving high-throughput technologies in the future.

References

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

[2] Pierre Baldi and Anthony D. Long. A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

[3] Z. Bar-Joseph, A. Gitter, and I. Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, 13:552–564, 2012.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc. B*, 57(1):289–300, 1995.

[5] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statist. Soc. B*, 48:259–302, 1986.

[6] H.C. Causton, B. Ren, S.S. Koh, C.T. Harbison, E. Kanin, E.G. Jennings, T.I. Lee, H.L. True, E.S. Lander, and R.A. Young. Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, 271(23):323–337, 2001.

[7] G.I. Chen and A.-C. Gingras. Affinity-purification mass spectrometry (AP-MS) of serine/threonine phosphatases. *Methods*, 42(3):298–305, 2007.

[8] M. Choi, C.Y. Chang, T. Clough, D. Broudy, T. Killeen, B.X. MacLean, and O. Vitek.

MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, 2014.

[9] B.C. Collins, L.C. Gillet, G. Rosenberger, H.L. Röst, A. Vichalkovski, M. Gstaiger, and R. Aebersold. Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods*, 10(12):1246–1253, 2013.

[10] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talon. masigpro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9):1096–1102, 2006.

[11] B. Cox, T. Kislinger, and A. Emili. Integrating gene and protein expression data: pattern analysis and profile mining. *Methods (San Diego, Calif.)*, 35(3):303–314, March 2005.

[12] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26:1367–1372, 2008.

[13] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[14] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahaja, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39:D691–697, 2011.

[15] K-.A. Do, P. Muller, and F. Tang. A Bayesian mixture model for differential expression. *J. Roy. Statist. Soc. Series C*, 54:627–644, 2005.

[16] Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. Statistical

methods for identifying differentially expressed genes in replicated cdna microarray experiments. *STATISTICA SINICA*, 12:111–139, 2002.

[17] W.H. Dunham, M. Mullin, and A.-C. Gingras. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics*, 12(10):1576–1590, 2012.

[18] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.*, 96:1151–1160, 2001.

[19] Bradley Efron. Robbins, empirical bayes and microarrays. *Ann. Statist.*, 31(2):366–378, 2003.

[20] J.D. Egertson, A. Kuehn, G.E. Merrihew, N.W. Bateman, B.X. MacLean, Y.S. Ting, J.D. Canterbury, D.M. Marsh, M. Kellmann, V. Zabrouskov, C.C. Wu, and M.J. MacCoss. Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods*, 10:744–746, 2013.

[21] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[22] Y. Fan and R. Li. Variable selection in linear mixed effects models. *Ann. Statist.*, 40(4):2043–2068, 2012.

[23] M.L. Fournier, A. Paulson, N. Pavelka, A.L. Mosley, K. Gaudenz, W.D. Bradford, E. Glynn, Li. H., M.E. Sardiu, B. Fleharty, C. Seidel, L. Florens, and M.P. Washburn. Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggcl in cellular sensitivity to rapamycin. *Molecular & Cellular Proteomics*, 9:271–284, 2010.

[24] Alyssa C. Frazee, Sarven Sabunciyan, Kasper D. Hansen, Rafael A. Irizarry, and Jeffrey T. Leek. Differential expression analysis of rna-seq data at single-base resolution. *Biostatistics*, 15(3):413–26, 2014.

[25] E. Garre, L. Romero-Santacreu, N. De Clercq, N. Blasco-Angulo, P. Sunnerhagen, and P. Alepuz. Yeast mRNA cap-binding protein Cbc1/Sto1 is necessary for the rapid reprogramming of translation after hyperosmotic shock. *Mol. Biol. Cell*, 23(1):137–150, 2012.

[26] C. Genovese and L. Wasserman. Bayesian and frequentist multiple testing. In J.M. Bernardo, J.O. Berger, M. Bayarri, and A.P. Dawid, editors, *Bayesian Statistics 7*, chapter 7, pages 145–161. Oxford University Press, Oxford, 2003.

[27] Ludovic C. Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, 11(6), 2012.

[28] A.-C. Gingras, M. Gstaiger, B. Raught, and R. Aebersold. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell. Biol.*, 8(8):645–654, 2007.

[29] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[30] N.L. Heinecke, B.S. Pratt, T. Vaisar, and L. Becker. PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. *Bioinformatics*, 26(12):1574–1575, 2010.

[31] C. Herrera and P.J. Zufiria. Generating scale-free networks with adjustable clustering coefficient via random walks. *arXiv*, 1105.3447, 2011.

[32] P.V. Hornbeck, J.M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, 40:D261–270, 2012.

[33] Joseph G Ibrahim, Ming-Hui Chen, and Robert J Gray. Bayesian models for gene

expression with dna microarray data. *Journal of the American Statistical Association*, 97(457):88–99, 2002.

[34] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Statist.*, 33(2):730–773, 04 2005.

[35] C. M. Kendziorski, M. A. Newton, H. Lan, and M. N. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22(24):3899–3914, 2003.

[36] Ross Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. AMS, 1980.

[37] M. Lee, S. Topper, S. Hubler, J. Hose, C. Wenger, J. Coon, and A. Gasch. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology*, 7(1):514, 2011.

[38] R. Linding, L. J. Jensen, G.J. Ostheimer, M.A.T.M. van Vugt, C. Jorgensen, I.M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J.G. Park, L.D. Samson, J.R. Woodgett, R.B. Russell, P. Bork, M.B. Yaffe, and T. Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129:1415–1426, 2007.

[39] Y. Liu, J. Chen, A. Sethi, Q.K. Li, L. Chen, B.C. Collins, L.C. Gillet, B. Wollscheid, H. Zhang, and R. Aebersold. Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovery N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. *Mol. Cell. Proteomics*, 13(7):1753–1768, 2014.

[40] J. Loven, D.A. Orlando, A.A. Sigova, C.Y. Lin, P.B. Rahl, C.B. Burge, D.L. Levens, T.I. Lee, and R.A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012.

[41] J. Loven, D.A. Orlando, A.A. Sigova, C.Y. Lin, P.B. Rahl, C.B. Burge, D.L. Levens, T.I. Lee, and R.A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012.

[42] B.X. MacLean, D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, R. Kern, D.L. Tabb, D.C. Liebler, and M.J. MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, 2010.

[43] A.A. Margolin, S.E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, and T.R. Golub. Empirical Bayes analysis of quantitative proteomics experiments. *PLOS One*, 4(10):e7454, 2009.

[44] R. Martin and S.T. Tokdar. A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–39, 2011.

[45] P. Muller, G. Parmigiani, and K. Rice. FDR and Bayesian multiple comparison rules. *Johns Hopkins University Department of Biostatistics Working Paper*, page 115, 2006.

[46] O. Muralidharan. An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.*, 4(1):422–438, 2010.

[47] M. A. Newton, C. M. Kendziorski, C. S. Richmond, Frederick R. Blattner, and K. W. Tsui. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8:37–52, 2001.

[48] M.A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[49] A. Panchaud, A. Scherl, S.A. Shaffer, P.D. von Haller, H.D. Kulasekara, S.I. Miller,

and D.R. Goodlett. PAcIFIC: how to dive depper into the proteomics ocean. *Anal. Chem.*, 81(15):6481–6488, 2009.

[50] T. Park, S. G. Yi, S. Lee, S. Y. Lee, D. H. Yoo, J. I. Ahn, and Y. S. Lee. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19(6):694–703, 2003.

[51] G. Parmigiani, E.S. Garrett, R. Anbazhagan, and E. Gabrielson. A statistical fraemwork for expression-based molecular classfication in cancer. *J. Roy. Statist. Soc. Series B*, 64:717–736, 2002.

[52] A.D. Polpitiya, W.J. Qian, N. Jaitly, V.A. Petuk, J.N. Adkins, D.G. 2nd Camp, G.A. Anderson, and R.D. Smith. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*, 24(13):1556–1558, 2008.

[53] S. Razick, G. Magklaras, and I.M. Donaldson. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9:405, 2008.

[54] M. Rep, V. Reiser, U. Gartner, J.M. Thevelein, S. Hohmann, G. Ammerer, and H. Ruis. Osmotic stress-induced gene expression in saccharomyces cerevisiae requires msn1p and the novel nuclear factor hot1p. *Molecular and Cellular Biology*, 19:5474–5485, 1999.

[55] H.L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S.M. Miladinović, O.T. Schubert, W. Wolski, B.C. Collins, J. Malmström, L. Malmström, and R. Aebersold. Openswath enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, 32(3):219–223, March 2014.

[56] P.A. Rudnick, X. Wang, X. Yan, N. Sedransk, and S.E. Stein. Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. *Mol. Cell. Proteomics*, 13(5):1341–1351, 2014.

[57] Maureen Sartor, Craig Tomlinson, Scott Wesselkamper, Siva Sivaganesan, George

Leikauf, and Mario Medvedovic. Intensity-based hierarchical bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, 7(1):538, 2006.

[58] Robert B. Scharpf, HÃěkon Tjelmeland, Giovanni Parmigiani, and Andrew B. Nobel. A bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, 104(488):1295–1310, 2009.

[59] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473:337–342, 2011.

[60] N. Sonenberg and A.G. Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–745, 2009.

[61] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(1):91, 2013.

[62] B. Soufi, C.d. Kelstrup, G. Stoehr, F. Frohlich, T.C. Walther, and J.V. Olsen. Global analysis of the yeast osmotic stress response by quantitative proteomics. *Molecular Biosystems*, 5:1337–1346, 2009.

[63] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[64] John D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035, 12 2003.

[65] Y. C. Tai and T. P. Speed. A multivariate empirical bayes statistic for replicated microarray time course data. *The Annals of Statistics*, 34(5):2387–2412, 2006.

[66] Yu Chuan Tai and Terence P. Speed. A multivariate empirical bayes statistic for replicated microarray time course data. *Ann. Statist.*, 34(5):2387–2412, 10 2006.

[67] Donatello Telesca, Peter MÃijller, Giovanni Parmigiani, and Ralph S. Freedman. Modeling dependent gene expression. *Ann. Appl. Stat.*, 6(2):542–560, 06 2012.

[68] Guoshou Teo, Sinae Kim, Chih-Chiang Tsou, Ben Collins, Anne-Claude Gingras, Alexey I. Nesvizhskii, and Hyungwon Choi. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *Journal of Proteomics*, pages –, 2015.

[69] Guoshou Teo, Christine Vogel, Debashis Ghosh, Sinae Kim, and Hyungwon Choi. PECA: A novel statistical tool for deconvoluting time-dependent gene expression regulation. *Journal of Proteome Research*, 13(1):29–37, 2014. PMID: 24229407.

[70] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A.I. Nesvizhskii. DIA-Umpire: comprehensive computational framework for data independent acquisition proteomics. *Nat. Methods*, 12(3):258–64, 2015.

[71] J.D. Venable, M.Q. Dong, J. Wohlschlegel, A. Dillin, and J.R. Yates. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods*, 1(1):39–45, 2004.

[72] C. Vogel, G.M. Silva, and E.M. Marcotte. Protein expression regulation under oxidative stress. *Molecular and Cellular Proteomics*, 10(12):M111.009217, 2011.

[73] J. Warringer, M. Hult, S. Regot, F. Posas, and P. Sunnerhagen. The HOG pathway dictates the short-term translational response after hyperosmotic shock. *Mol. Biol. Cell*, 21:3080–3092, 2010.

[74] Z. Wei and H. Li. A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, 2007.