Fachhochschul-Masterstudiengang
BIOMEDIZINISCHE INFORMATIK
4232 Hagenberg, Austria

Computational analysis of post-transcriptional control of cell state transitions by RNA-binding proteins

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science in Engineering

Eingereicht von

Konstantin Krismer, BSc

Betreuer:	Michael B. Yaffe, M.D., Ph.D., MIT, Cambridge, MA, USA
Begutachter:	Andreas Heinzel, MSc

September 2015

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Hagenberg, September 9, 2015

Konstantin Krismer

Contents

Acknowledgements

I would like to thank Michael B. Yaffe and his entire lab for providing the best work environment imaginable. I hope my next experience in academia will be at least half as exciting and intellectually stimulating as my time at the Yaffe lab. After more than a year in the Yaffe lab I believe the world of academia is full of scientific adventures, fueled by exploratory spirit and intrinsic motivation. According to researchers outside the Yaffe lab it is a path of trial and tribulation, littered with bureaucratic obstacles. A path that eventually and inevitably leads to suicide. I think I will give it a shot anyway. There must be other Mikes out there.

I want to thank Ian Cannell and Brian Joughin, my informal project supervisors at MIT, and Andreas Heinzel, my supervisor at my home university. Three very different people, with different fields of expertise and different drinking habits. I highly appreciated your thought provoking conversations and ingenious insights.

Transite was built on top of Anna Gattinger's excellent work, without her, this thesis would not have been possible.

I want to express my gratitude to the Marshall Plan Foundation for helping me to finance my research, and to an even greater extent to my liberal parents, for financing the first k years of my life. I don't take it for granted.

I also feel the desire to thank my friends in Boston, Vienna and Salzburg for making my life as awesome as it is. Without them I would have finished this work three months earlier.

Abstract

Despite its crucial role in post-transcriptional regulation of gene expression, the functions of the majority of RNA-binding proteins (RBPs) are largely unknown. Unlike transcriptional control of gene expression, which has been studied extensively over the past decades, post-transcriptional regulation in general, and RBPs in particular, are comparatively poorly understood and have not until recently been the focus of large systematic studies.

In the light of this research gap, this thesis presents Transite, a novel computational method that allows cost-effective, time-effective and comprehensive analysis of the regulatory role of RBPs in various cellular processes by leveraging a wealth of preexisting gene expression data and current knowledge of RBP binding preferences. To gain insights into vastly complex processes including the DNA damage response or the immune response, the preliminary step is to calculate the change of mRNA expression levels after stimulus, i.e., the administration of DNA-damaging agents or an immune stimulus, respectively. Based on these results, Transite provides two approaches to investigate inferred mRNA stability changes due to differences in transcript abundance, Transcript Set Motif Analysis and Spectrum Motif Analysis. The former focuses on significantly upregulated and downregulated sets of transcripts and identifies RBPs whose binding sites are overrepresented among those transcripts, whereas the latter approach examines the distribution of RBP binding sites across the entire spectrum of transcripts, sorted according to their fold change.

Transite will be available as an R/Bioconductor package to ensure a seamless integration in current workflows. Additionally, a user-friendly online version will be accessible at http://transite.mit.edu.

Kurzfassung

Obwohl der posttranskriptionellen Ebene der Regulation der Genexpression eine bedeutende Rolle zukommt, ist die Funktion der Mehrheit der RNA-bindenden Proteine (RBPs) noch großteils unbekannt. Im Gegensatz zur transkriptionellen Regulation, die in den letzten Jahrzehnten ausgiebig erforscht wurde, ist das Wissen über die posttranskriptionelle Regulation im Allgemeinen und RBPs im Besonderen mangels großer systematischer Studien als vergleichsweise lückenhaft anzusehen.

Angesichts der aufgezeigten Forschungslücke wird in dieser Arbeit Transite präsentiert, eine neuartige bioinformatische Methode, die es erlaubt, die regulative Rolle der RBPs in verschiedenen zellulären Prozessen mittels vorhandener Genexpressionsdaten kosten- und zeiteffizient zu untersuchen. Um Einsichten in äußerst komplexe Prozesse wie die zelluläre DNA-Schadensantwort oder immunologische Abwehrreaktionen zu erlangen, werden initial die Veränderung der Expressionsniveaus der Transkripte nach Gabe beispielsweise einer DNA-schädigenden Substanz oder eines Antigens ermittelt. Auf diesen Ergebnissen basierend werden mit Transite zwei Ansätze angeboten, um RBP-induzierte Änderungen der mRNA-Stabilität zu untersuchen: Transcript Set Motif Analysis und Spectrum Motif Analysis. Ersterer konzentriert sich auf signifikant hoch- beziehungsweise runterregulierte Transkripte und identifiziert RBPs, deren Bindungsstellen in diesen Gruppen von Transkripten überrepräsentiert sind. Der zweitgenannte Ansatz untersucht die Verteilung der RBP-Bindungsstellen über das gesamte Transkript-Spektrum, sortiert nach Stärke der Änderung des Expressionsniveaus.

Der Funktionsumfang der Transite-Pipeline wird in zwei Formen zur Verfügung gestellt: als R/Bioconductor-Package, um die Integration in bestehende Datenanalyse-Workflows zu vereinfachen; und als benutzerfreundliche Website, erreichbar unter http://transite.mit.edu.

Chapter 1

Introduction

This thesis presents the analytical pipeline Transite, a computational method that has been developed to shed light on the post-transcriptional control of cell state transitions by RNA-binding proteins (RBPs). Leveraging newly available data from large-scale assays that identify the binding sites of a great number of RBPs, Transite generates hypotheses regarding how the change of transcript abundance levels in gene expression data can be explained by RBP-mediated mRNA stability changes. Specifically, position specific scoring matrices (see section 3.5) of RBP binding sites are used to quantify the combined binding evidence among 5' UTRs, 3' UTRs or intronic regions of *meaningful* sets of transcripts (e.g., transcripts upregulated after the administration of a DNA-damaging agent). In this way, the increased transcript abundance can be partially attributed to the stabilizing effect of certain RBPs.

Transite can be seen as the post-transcriptional counterpart to Scansite [1], which predicts (post-translational) phosphorylation sites.

1.1 Post-transcriptional regulation and RBPs

Post-transcriptional control of gene expression regulates all aspects of RNA metabolism and function, including mRNA stability, localization, silencing, splicing, transport and translation. Examples are the 3' UTR-dependent protein localization that is independent of RNA localization [2], sequence-specific downregulation or suppression of gene expression by microRNAs, or alternative splicing mediated by cis-acting RNA sequence elements present in pre-mRNAs and trans-acting RBPs [3].

Alongside microRNAs, RBPs are major post-transcriptional regulators. In general, RBPs bind to short sequence stretches in mRNAs, nascent transcripts, noncoding RNAs, and damaged DNA [4]. Their binding sites are predominantly found in evolutionary conserved regions in the 5' and 3' untranslated regions of mature mRNA [5] (see schematic in figure 1.1 for ori-

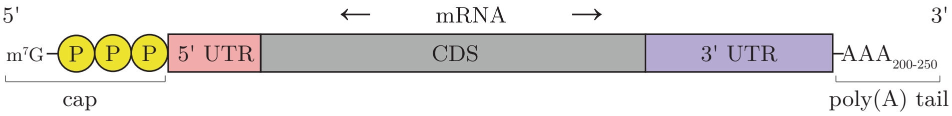


Figure 1.1: Structure of mature messenger RNA: 5' cap, 5' untranslated region, coding sequence, 3' untranslated region, poly(A) tail

entation), and to a lesser extent in intronic regions of unspliced mRNA precursors. RBPs are involved in pre-mRNA splicing, polyadenylation, mRNA stability, and translation. They are key regulatory factors in a vast number of cellular processes.

Once mRNAs are transcribed, the amount of protein produced is essentially determined by two factors, mRNA stability and translation. Both are subject to post-transcriptional regulation via RBPs. RBPs like ELAVL1 [6, 7] regulate the longevity of their mRNA targets, and as a result, the half-lives of mRNAs differ greatly in a transcript-specific manner, in eukaryotic cells up to a 100-fold [8]. This RBP-mediated regulation of transcript stability is the reason why gene expression data from microarrays or RNA-sequencing can be used to investigate RBP activity. After the activation of a stabilizing RBP, the measurable mRNA levels of its targets will rise, given that the rate of transcription does not change.

1.2 Aim of this study

The DNA damage response is traditionally considered to have two main arms, an early and rapid protein kinase-driven signaling response and a delayed transcriptional response mediated by a subset of dominant transcriptional regulators. However, growing evidence suggests that a third perhaps equally important and vastly complex response exists at the level of post-transcriptional control, through the modulation of mRNA splicing, stability and translation. [4]

The response to DNA damage is one of many cellular processes in which RBPs are assumed to play a vital part. But unlike transcriptional control, which has been studied extensively over the past decades, post-transcriptional regulation in general, and RBPs in particular, are comparatively poorly understood and have not been the focus of large systematic studies (apart from a few exceptions). This evident research gap underscores the need for efficient computational methods to elucidate the role of RBPs in various contexts.

The aim of this study is to develop a tool that helps biologists to understand how key post-transcriptional regulators (mainly RNA-binding proteins, but also microRNAs) contribute to the concerted regulation and function of cellular processes. The idea is to utilize the large body of publicly available gene expression data from microarray and RNA sequencing experiments to identify changes in mRNA expression levels upon certain stimuli (DNA-damaging agents, antigens, etc.) coupled to the subsequent identification of enriched or depleted RBP binding sites in sequence regions of those mRNAs. In this way, hypotheses can be generated regarding what RBPs interact preferentially with mRNAs that are sensitive to the aforementioned stimuli.

A brief introduction to RBPs and post-transcriptional regulation is given in chapter 1. The novel analytical pipeline and algorithms behind Transite are described in chapter 2. Data sources and existing statistical methods are described in chapter 3, and evaluations and applications of the Transite pipeline can be found in chapter 4.

Chapter 2

Transite pipeline

This chapter describes the analytical pipeline of Transite and the statistical methods that have been developed.

Sections 2.1 to 2.6 correspond to the six steps of the schematic diagram in figure 2.1. They explain the main components of the analytical pipeline of Transite, broken down into six steps. The tasks of step 1 include to retrieve gene expression data from the GEO database, generate quality control plots to visually inspect the data, and detect and—if necessary—exclude outlier samples. In step 2 the requested sequence regions (3' UTR, 5' UTR or intronic regions) of all platform genes are collected from current genome assemblies. Step 3 contains the groundwork for step 4 by computing fold changes and associated p-values between user-defined sample groups, i.e., *treatment* and *control* samples of the gene expression data. In step 4 the foreground sets are defined. Depending on the analysis type, foreground sets are either upregulated and downregulated transcripts (Transcript Set Motif Analysis), or n equally sized bins, grouping transcripts with similar fold changes (Spectrum Motif Analysis). Steps 5 and 6 contain the RNA-binding protein scoring algorithms, based on publicly available position specific scoring matrices (PSSMs). The results of step 5 help to understand which RBPs have targets that are predominantly overrepresented or underrepresented in the sets of upregulated and downregulated transcripts relative to all platform transcripts. The Spectrum Motif Analysis in step 6, in contrast, is not limited to the most upregulated or downregulated transcripts, but investigates the distribution of RBP targets across the entire spectrum of transcripts, sorted by their fold change. Both the Transcript Set Motif Analysis in step 5 and the Spectrum Motif Analysis in step 6 are available in a hexamer-based and a transcript-based version. Section 2.7 introduces methods to systematically evaluate the results of Spectrum Motif Analysis.

Section 2.8 describes Single Transcript Motif Analysis, a Transite add-on, which is not part of the usual Transite workflow.

Transite supports a number of different methods to combine and adjust p-values, only one of which is used per analysis. Since there is no universally superior method that outperforms all others, it is the user's choice to decide which one to use.

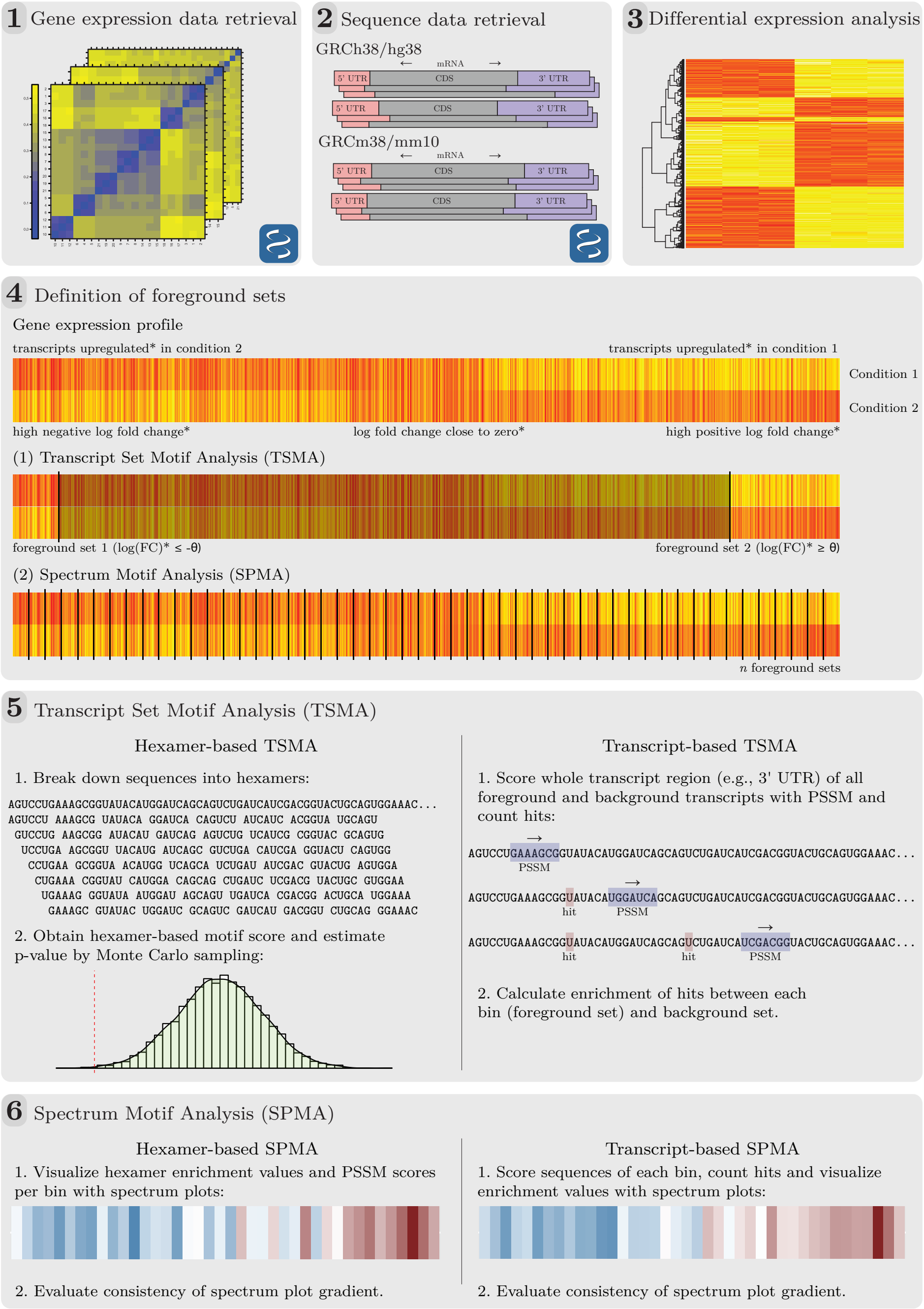

1 Gene expression data retrieval
2 Sequence data retrieval
GRCh38/hg38
mRNA
5' UTR
CDS
3' UTR
GRCm38/mm10
3 Differential expression analysis
4 Definition of foreground sets
Gene expression profile
transcripts upregulated* in condition 2
transcripts upregulated* in condition 1
Condition 1
Condition 2
high negative log fold change*
log fold change close to zero*
high positive log fold change*
(1) Transcript Set Motif Analysis (TSMA)
foreground set 1 (log(FC)* ≤ -θ)
foreground set 2 (log(FC)* ≥ θ)
(2) Spectrum Motif Analysis (SPMA)
n foreground sets
5 Transcript Set Motif Analysis (TSMA)
Hexamer-based TSMA
1. Break down sequences into hexamers:
AGUCCUGAAAGCGGUAUACAUGGAUCAGCAGUCUGAUCAUCGACGGUACUGCAGUGGAAAC...
AGUCCU AAAGCG UAUACA GGAUCA CAGUCU AUCAUC ACGGUA UGCAGU
GUCCUG AAGCGG AUACAU GAUCAG AGUCUG UCAUCG CGGUAC GCAGUG
UCCUGA AGCGGU UACAUG AUCAGC GUCUGA CAUCGA GGUACU CAGUGG
CCUGAA GCGGUA ACAUGG UCAGCA UCUGAU AUCGAC GUACUG AGUGGA
CUGAAA CGGUAU CAUGGA CAGCAG CUGAUC UCGACG UACUGC GUGGAA
UGAAAG GGUAUA AUGGAU AGCAGU UGAUCA CGACGG ACUGCA UGGAAA
GAAAGC GUAUAC UGGAUC GCAGUC GAUCAU GACGGU CUGCAG GGAAAC
2. Obtain hexamer-based motif score and estimate p-value by Monte Carlo sampling:
Transcript-based TSMA
1. Score whole transcript region (e.g., 3' UTR) of all foreground and background transcripts with PSSM and count hits:
→
AGUCCUGAAAGCGGUAUACAUGGAUCAGCAGUCUGAUCAUCGACGGUACUGCAGUGGAAAC...
PSSM
→
AGUCCUGAAAGCGGUAUACAUGGAUCAGCAGUCUGAUCAUCGACGGUACUGCAGUGGAAAC...
hit
PSSM
→
AGUCCUGAAAGCGGUAUACAUGGAUCAGCAGUCUGAUCAUCGACGGUACUGCAGUGGAAAC...
hit
hit
PSSM
2. Calculate enrichment of hits between each bin (foreground set) and background set.
6 Spectrum Motif Analysis (SPMA)
Hexamer-based SPMA
1. Visualize hexamer enrichment values and PSSM scores per bin with spectrum plots:
2. Evaluate consistency of spectrum plot gradient.
Transcript-based SPMA
1. Score sequences of each bin, count hits and visualize enrichment values with spectrum plots:
2. Evaluate consistency of spectrum plot gradient.

Figure 2.1: Transite pipeline: A schematic of the main steps; starting with preliminary tasks like gene expression data retrieval, sequence data retrieval, data preprocessing, quality control, and differential gene expression analysis in panels **(1)** to **(3)**; and foreground/background sequence assignments in panel **(4)**. (*continued on next page*)

Figure 2.1: The asterisk in panel **(4)** denotes the exchangeability of the sorting approach, where sorting the transcripts according to their fold change is one possibility. The θ in the TSMA section of the same panel represents a threshold that determines the foreground sets (usually a threshold for differential expression). The k-mer-based and transcript-based approaches of Transcript Set Motif Analysis and Spectrum Motif Analysis are briefly presented in panels **(5)** and **(6)**.

2.1 Data preprocessing

In case the analysis is based on publicly available gene expression data, the R/Bioconductor package `GEOquery` [9] is used to retrieve the data set from the Gene Expression Omnibus database [10].

The data set of a gene expression study consists of a set of samples, where each sample is a vector of gene expression values and a sample label, e.g., *treatment* or *control*. Each gene expression value is associated with a platform-specific probe identifier, which can be mapped to gene name and RefSeq identifier. The type of the gene expression values depend on the underlying experiment. Values from single channel microarrays are normalized signal count data, dual channel microarrays are normalized log ratios, and values from RNA-seq experiments are discrete counts. Formally, the data of a gene expression study can be denoted as matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, where n is the number of probes and m is the number of samples.

Unless microarray gene expression values are already transformed, $\log_2$ transformation is performed as an initial step. Furthermore, all probes, i.e., rows of $\boldsymbol{A}$, with missing values are removed.

Outlier[1] samples are detected and, if required, excluded from further analysis. Samples are identified as outliers based on (1) their Euclidean distance to other samples, and (2) the Kolmogorov-Smirnov test statistic between the distribution of intensities from the sample in question and the pooled distribution of intensities from all samples.

The R/Bioconductor package `arrayQualityMetrics` [12] is used to produce diagnostic plots (see figure 2.2) to get a first impression of how well the sample groups are reflected in the data. Biological replicates should be similar to each other, their pairwise distances are expected to be smaller than the distances between samples from different conditions. This is the case in columns (1) and (2) of the exemplary gene expression experiments shown in figure 2.2, but not in column (3). The dendrograms are obtained by sample clustering with complete linkage and $1 - \rho$ as distance metric, where ρ is the Pearson product-moment correlation coefficient.

[1] Outliers are points that are 1.5 interquartile ranges below the first quartile or above the third quartile[11]

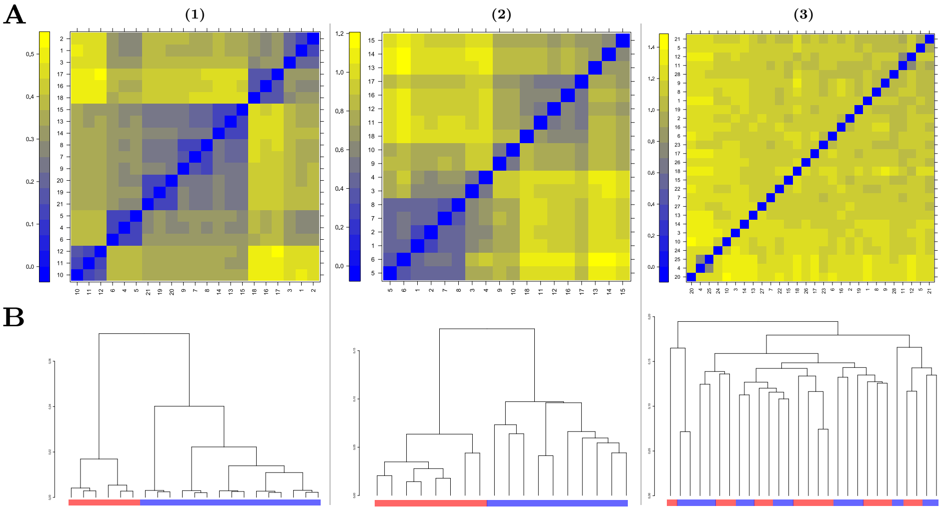


Figure 2.2: Diagnostic plots: (A) Columns (1) to (3) contain heatmaps of the ℓ_2 norm (Euclidean distance) between samples of three different studies. The heatmaps in columns (1) and (2) depict distinct groupings of samples that correspond to the sample labels *treatment* and *control*. Column (3) is an example of a study without a clear distinction between treatment and control samples with respect to sample distance. **(B)** recapitulates the visual sample groupings in the heatmaps by unsupervised sample clustering based on sample correlation. The red and blue bars below the dendrograms indicate sample labels. The dendrograms of columns (1) and (2) cluster samples according to their labels, whereas the treatment and control samples in dendrogram (3) are seemingly randomly distributed.

2.2 Sequence retrieval

After the gene expression data is preprocessed, the requested sequence regions of all transcripts of the platform are retrieved. Available sequence regions include three prime untranslated region (3’ UTR), intronic regions, and five prime untranslated region (5’ UTR).

In order to request transcript sequences from NCBI genome assemblies, probe identifiers are mapped to RefSeq identifiers by using information stored in the platform annotation files from GEO.

Probes without associated RefSeq identifiers are removed, as well as probes with invalid (e.g., obsolete) RefSeq identifiers or RefSeq identifiers of entries without annotated sequence regions.

The GRCh38/hg38 genome assembly is used for human platforms and the GRCm38/mm10 for mouse platforms.

Figure 2.3 shows histograms of the length distribution of the retrieved set of sequences. These plots are part of the final Transite analysis output and help to immediately spot serious issues in the sequence retrieval step, such as insufficient platform annotation or identifier mapping issues.

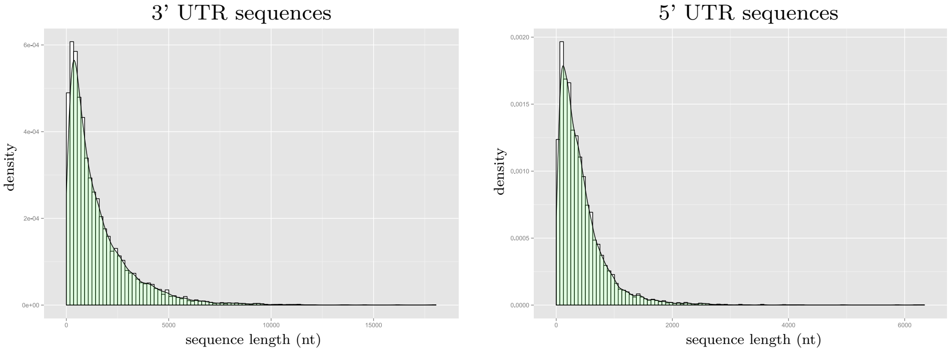


Figure 2.3: Sequence length distribution: Diagnostic plots are provided to identify incompatible platforms or platforms with incomplete annotations, which cause identifier mapping and sequence retrieval issues.

2.3 Differential gene expression analysis

The aim of this step is to rank genes in order of their differential expression. This ranking constitutes the basis of the foreground set definitions in the next step.

Differential gene expression analysis is applied in order to subdivide the transcripts of the gene expression data obtained in step 1 (described in section 2.1) into *meaningful* foreground sets.

The following section describes the differential gene expression analysis workflow using the R/Bioconductor package `limma` [13]. This is one of many ways to obtain the ranking of transcripts and subsequent steps of the Transite pipeline are agnostic to the used methodology. In fact, Transcript Set Motif Analysis only requires a nominal measure per transcript to define the foreground sets. For Spectrum Motif Analysis, at least an ordinal measure is required to rank transcripts.

The inital step towards a ranking of differential expression is to specify the sample groups, i.e., which samples belong to the *treatment* and which to the *control* group. Only two sample groups are allowed. If there are more than two (e.g., *treatment 1*, *treatment 2* and *control*), the differential expression analysis is broken into two separate analyses excluding all samples belonging to one of the treatment groups at a time. The assignment of sample groups is defined by the so-called design matrix. `limma` fits a linear model (specified by the design matrix) to each row of the $\log_2$-transformed expression value matrix $\boldsymbol{A}$ defined in section 2.1, where each row corresponds to one of n probes/transcripts, measured in m samples. The coefficients of the fitted models describe the differences between the *treatment* and *con-*

trol groups. An empirical Bayes method is used to obtain the significance and the strength of the log fold change between sample groups for each transcript [14].

The result of the differential expression analysis are two values per transcript: (1) the expression fold change between the sample groups, and (2) the p-value, which quantifies the significance of the change between the sample groups. Both can be used to define foreground sets for Transcript Set Motif Analysis and Spectrum Motif Analysis. The raw p-values are adjusted to avoid alpha error accumulation (see section 3.3). By default, the Benjamini-Hochberg procedure [15] is used.

2.4 Definition of foreground sets

The elements of the background set are the requested sequence regions (3' UTR, 5' UTR or introns) of all platform transcripts.

Foreground sets are proper subsets of the background set and their definition depends on the desired motif analysis approach. In any case, foreground and background sets define the groups of sequences relative to which the overrepresentation and underrepresentation of RBP binding evidence (also called *binding* or *target sites*, or *hits*) is investigated.

2.4.1 Foreground sets for Transcript Set Motif Analysis

When gene expression data is used, the two foreground sets for Transcript Set Motif Analysis are composed of the statistically significantly upregulated and downregulated transcripts (see figure 2.1, panel 4). Transcripts are deemed statistically significantly differentially expressed if their Benjamini-Hochberg adjusted p-value is equal to or less than 0.05. Whether a transcript belongs to the upregulated or downregulated foreground set is naturally given by the sign of their log fold change.

Various deviations of this canonical definition of foreground sets are possible. Upregulated and downregulated sets can be defined by fold change only, neglecting the p-value. It is not even necessary to use gene expression data. TSMA can be used with predefined gene sets as well, for example all (human/murine) genes associated with a certain Gene Ontology [16] term. In which case the background set would consist of all genes of human or mouse, respectively, which are annotated with at least one GO term.

2.4.2 Foreground sets for Spectrum Motif Analysis

The Spectrum Motif Analysis approach requires a number of foreground sets (usually 40), which collectively cover the entire spectrum of transcripts. The transcripts are sorted according to their fold change, signed log p-value ($\text{sign}(\log_2(\text{FC})) * (-1) * \log_2(p)$), or other user-defined ordinal or metric measures. Then this so-called transcript spectrum is subdivided into 40 bins (foreground sets) of equal width, i.e., equal number of transcripts per bin. The number of bins is somewhat arbitrary. If the number is too high, the number of transcripts per bin is low, leading to noisy spectrum plots. If the number of bins is too low, the gradient from the lowest to the highest bin is covered or evened out by intermediate transcripts. How the number of bins influences the outcome is investigated in section 4.1.1. An illustration of the subdivision of the sorted transcript spectrum is part of figure 2.1, panel 4.

2.5 Transcript Set Motif Analysis

The aim of Transcript Set Motif Analysis (TSMA) is to identify the overrepresentation and underrepresentation of potential RBP targets (binding sites) in a set (or sets) of sequences, i.e., the foreground set, relative to the entire population of sequences. The latter is called background set, which can be composed of all sequences of the genes of a microarray platform or all sequences of an organism or any other meaningful superset of the foreground sets.

Once foreground and background sets are defined, there are two approaches to analyze overrepresentation (or underrepresentation, respectively) of RBP binding sites.

2.5.1 *k*-mer-based TSMA

Before sequences are scored with the PSSMs that define RBP binding sites (see section 3.5), they are broken into k-mers, i.e., oligonucleotide sequences of k bases. And only statistically significantly enriched or depleted k-mers are then used to calculate a score for each RBP, which quantifies its target overrepresentation.

k-mer enrichment analysis

After foreground and background sets are defined, the sequences of both sets are broken into hexamers, i.e., k-mers of length 6. While Transite also supports heptamers and octamers (sequences of length 7 and 8, respectively), hexamers are recommended, since run-time increases exponentially with k and the results for heptamers and octamers mirror the ones for hexamers.

Strength of hexamer enrichment: There are 4^6 or 4096 distinct hexamers, for which the occurrences in foreground and background sets are counted. In the following, we call the vector of hexamer counts for the foreground set $\boldsymbol{f}$ and for the background set $\boldsymbol{b}$.

$$\overbrace{\begin{pmatrix} |\mathrm{AAAAAA}|_1 \\ |\mathrm{AAAAAC}|_2 \\ |\mathrm{AAAAAG}|_3 \\ \vdots \\ |\mathrm{GGUUUU}|_{4094} \\ |\mathrm{GUUUUU}|_{4095} \\ |\mathrm{UUUUUU}|_{4096} \end{pmatrix}}^{\boldsymbol{f}} \overbrace{\begin{pmatrix} |\mathrm{AAAAAA}|_1 \\ |\mathrm{AAAAAC}|_2 \\ |\mathrm{AAAAAG}|_3 \\ \vdots \\ |\mathrm{GGUUUU}|_{4094} \\ |\mathrm{GUUUUU}|_{4095} \\ |\mathrm{UUUUUU}|_{4096} \end{pmatrix}}^{\boldsymbol{b}} \tag{2.1}$$

Hexamer enrichment values are calculated as follows:

$$e_i = \frac{f_i/F}{b_i/B}, \tag{2.2}$$

where $F = \sum f_i$ and $B = \sum b_i$.

Significance of hexamer enrichment: First, a contingency table for k-mer i called C_i, where $i \in [1, 4096]$ is created.

$$C_i = \begin{pmatrix} f_i & F - f_i \\ b_i & B - b_i \end{pmatrix}. \tag{2.3}$$

Then the p-value p_i for C_i is approximated with Pearson's χ^2 test. If $p_i < 5\alpha$, p_i is replaced with the p-value obtained by Fisher's exact test for C_i. This odd procedure reduces computation time dramatically (approximately 50-fold decrease), because the computationally expensive Fisher's exact test is only used in cases, where the approximate p-value from the computationally inexpensive χ^2 test is close to the decision boundary (α). Mathematical accuracy is traded for efficiency.

Fisher's exact test is always used if at least one of the expected counts is less than five.

The p-values are subsequently adjusted to avoid alpha error accumulation (see section 3.3).

Transite uses so-called volcano plots to visualize the hexamer enrichment values (x-coordinate, log transformed) and associated p-values (y-coordinate, log transformed and multiplied by -1) of a TSMA run (see panel **(A)** of figure 2.4). The enrichment values on the x-axis are logarithmized in order to display enrichment values symetrically around zero (e.g., $|\log(0.5)| = |\log(2)|$).

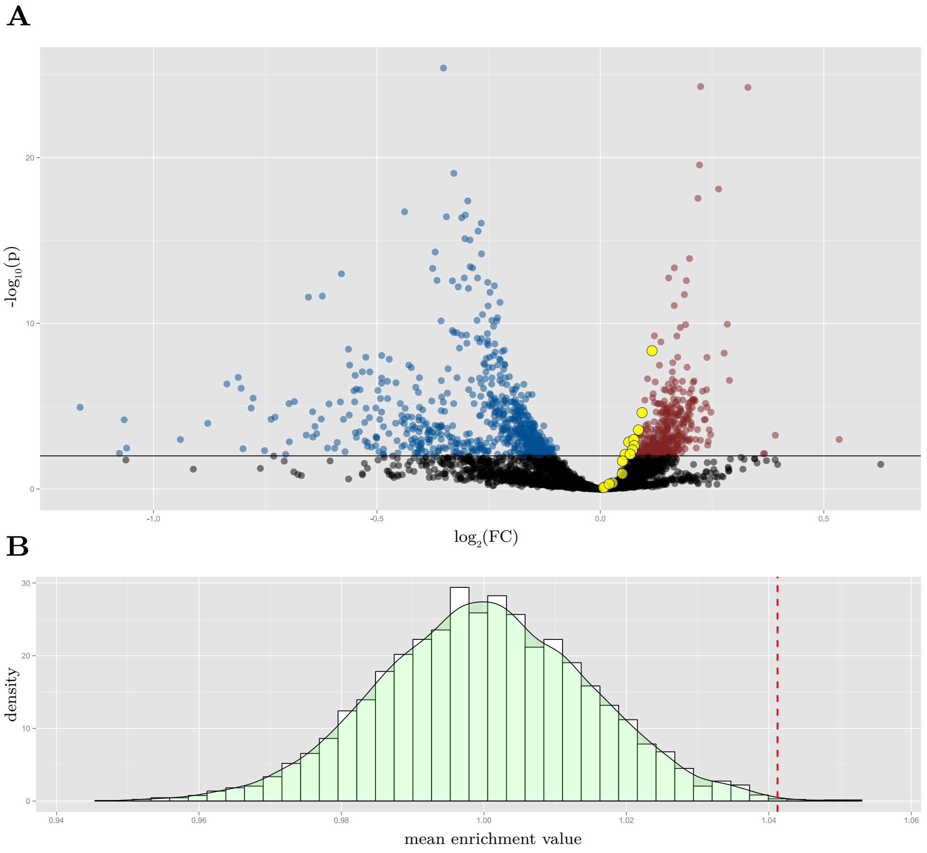


Figure 2.4: *k*-mer-based TSMA result plots: (A) Volcano plot of hexamers: black dots represent insignificant hexamers, blue dots denote significantly depleted and red dots significantly enriched hexamers. Yellow hexamers are compatible hexamers to motif j. **(B)** Histogram of the distribution of geometric mean enrichment values of compatible hexamers after Monte Carlo sampling. Red line denotes observed mean enrichment value of compatible hexamers (yellow hexamers in **(A)**).

Motif hexamer enrichment

There are two values to describe each hexamer, its enrichment value e_i and the associated adjusted p-value p_i, i.e., the significance of the enrichment.

For each sequence motif (position probability matrix in this case) in Transite there is a set of compatible hexamers (also called *motif-associated hexamers*). Hexamer i is compatible to motif j if and only if hexamer i can be aligned with motif j in a way that the probability of each position in hexamer i is greater than threshold θ. The exact value of θ is not important. For all subsequent analyses, θ was set to 0.2.

One way to describe the overrepresentation (or underrepresentation, respectively) of RNA-binding protein target sites is to provide a summary of enrichment values and enrichment p-values of the compatible hexamers of each sequence motif. An adequate summary of the enrichment values of compatible hexamers of some motif j would be their mean, and since enrichment values are ratios, the geometric mean must be used.

$$\bar{e} = exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(e_i)\right), \tag{2.4}$$

where $\bar{e}$ is the geometric mean of $\boldsymbol{e}$, the vector of enrichment values of motif-associated hexamers. The sum of logarithms is used instead of the product to avoid arithmetic underflow.

Monte Carlo tests (permutation tests) are performed to obtain an estimate of the significance of $\bar{e}$. The procedure is described in detail in section 3.6.

Panel **(B)** of figure 2.4 shows the empirical distribution of the mean of motif-associated hexamer enrichment values of an exemplary TSMA run, where the actual mean enrichment value is indicated by the red bar.

How $\boldsymbol{p}$, the vector of enrichment p-values of motif-associated hexamers, can be combined to a single value is described in section 3.2.

Motif scoring

Instead of merely looking at compatible hexamers (and summaries of their enrichments), an algorithm was developed by Anna Gattinger [17] to calculate a score, which uses the position weight matrices described in section 3.5 to quantify the degree of binding evidence among statistically significantly enriched and depleted hexamers. RNA-binding proteins with a positive score have stronger binding sites among the enriched hexamers, whereas the binding sites of RBPs with a negative score are predominantly found in the set of depleted hexamers.

Scoring algorithm: Initially, all statistically significant hexamers (enriched or depleted) are selected as input. The score s_i for motif i is calculated as follows: (1) All input hexamers are scored by the position weight matrix representation of motif i (see section 3.5 for details). (2) The binding evidence among enriched and depleted hexamers is

$$s_i = \sum_j \frac{e_j}{r(e_j)} - \sum_j \frac{d_j}{r(d_j)}, \tag{2.5}$$

where e_j is the score of the jth enriched hexamer, d_j is the score of the jth depleted hexamer, and $r(x)$ returns the rank of hexamer score x, hexamers with a score (e_i or d_i) below zero are discarded. Each hexamer's contribution to the overall score s_i is proportional to its rank. The enrichment (or depletion, respectively) of highly scored hexamers (strong binding evidence) has more weight than the one of poorly scored hexamers.

Staged Monte Carlo tests: The null distribution of motif scores (raw scores) depends on the position weight matrix, which makes it necessary to normalize scores in order to be able to compare scores from different motifs in a meaningful way. Monte Carlo tests are performed to obtain an empirical null distribution of motif scores for each motif. Scores are normalized by subtracting the mean and dividing by the standard deviation of the empirical score distribution. Furthermore, the empirical score distribution is used to obtain an estimate of the two-sided p-value of the raw score (see section 3.6). In order to significantly reduce the execution time of the Monte Carlo tests without reducing the number of permutations, the tests are implemented in a staged fashion: At first (stage 1), the null distributions of the scores of all 175 motifs in Transite are generated with only 100 permutations. Raw scores are normalized and empirical p-values are calculated based on this rough estimate. Only motifs with a score-associated p-value estimate whose lower bound of the confindence interval is less or equal to 0.01 are considered in stage 2 (see section 3.6 for details on confidence intervals of p-value estimates). In stage 2, another Monte Carlo test is performed, this time with 1000 permutations. And for the final stage of Monte Carlo tests, stage 3, only motifs with a lower bound of 0.001 of their p-value estimates are considered. Stage 3 Monte Carlo tests are performed with 5000 permutations.

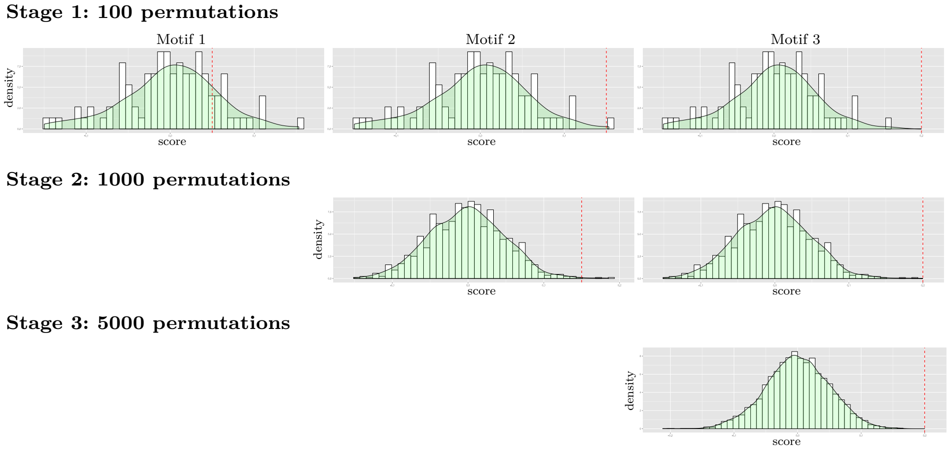


Figure 2.5: Staged Monte Carlo test: The procedure of the staged Monte Carlo test is illustrated with the aid of three generic motifs, *motif 1*, *motif 2* and *motif 3*, the histograms of their empirical score distributions and their observed raw scores (dashed red lines). The score of *motif 1* is not significant, whereas scores of motifs *2* and *3* are significant and highly significant, respectively. The raw score of *motif 1* falls in the main body of the null distribution in stage 1, thus, unlike *motif 2* and *motif 3*, *motif 1* is not considered in stages 2 and 3. Similarly, the score of *motif 2* is not significant enough to be considered in stage 3. In this way the bulk of the time-consuming permutations are spent on motifs with significant scores.

2.5.2 Transcript-based TSMA

The transcript-based approach skips the *k*-merization step and instead scores the transcript sequence as a whole with a position specific scoring matrix.

For each sequence in foreground and background sets and each sequence motif, the scoring algorithm evaluates the score for each sequence position (by applying the algorithm described in section 3.5). Positions with a relative score greater than a user-defined threshold (0.9 is usually used - 90% of the theoretical maximum of the given position weight matrix) are considered *hits*, i.e., putative binding sites.

By scoring all sequences in foreground and background sets, a hit count for each motif and each set is obtained, which is used to calculate enrichment values and associated p-values in the same way in which motif-compatible hexamer enrichment values are calculated in the *k*-mer-based approach. P-values are adjusted with one of the methods in section 3.3.

An advantage of the transcript-based approach is the possibility of detecting clusters of binding sites. This can be done by counting regions with many hits using positional hit information or by simply applying a hit count threshold per sequence, e.g., only sequences with more than some number of hits are considered. Homotypic clusters of RBP binding sites may play a similar role as clusters of transcription factors [18].

2.6 Spectrum Motif Analysis

The essential differences between TSMA and Spectrum Motif Analysis (abbreviated as SPMA) are the way how foreground sets are defined (see section 2.4) and how results are visualized. Apart from these two differences, k-mer-based and transcript-based SPMA are using the same algorithms that are described in the previous section.

SPMA helps to illuminate the relationship between RBP binding evidence and the transcript sorting criterion, e.g., fold change between *treatment* and *control* samples.

2.6.1 Transite spectrum plots

Spectrum plots are compact graphical representations of the results of a SPMA run. A spectrum plot visualizes scores and associated p-values of an RBP motif across the spectrum of transcripts (subdivided into 40 bins, i.e., foreground sets). The numerical values of the scores and p-values are represented as colors, which supports the human interpretation of spectrum plots.

There are three different types of spectrum plots: (1) k-mer enrichment spectrum plots visualize k-mer enrichment values and combined enrichment p-values of the compatible k-mers of a given motif. (2) k-mer-based motif score spectrum plots visualize motif scores (see 2.5.1) and their p-values. (3) Transcript-based hit enrichment spectrum plots depict hit enrichment values and associated p-values.

The automatic evaluation of spectrum plots is described in section 2.7.

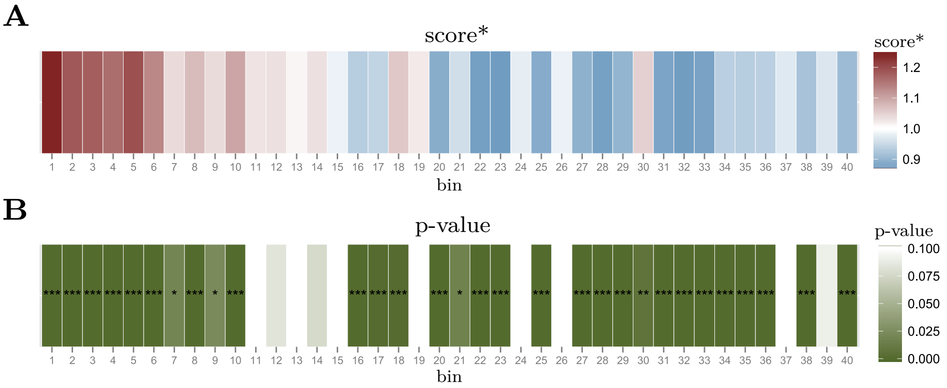


Figure 2.6: Spectrum plot: (A) Color representation of the motif scores per bin (depending on the type of spectrum plot, the *score** is k-mer enrichment value, k-mer-based motif score, or transcript-based hit enrichment value). **(B)** Color-coded associated p-values (can be one of the following: combined k-mer enrichment p-values based on Fisher's exact tests, empirical p-values obtained by Monte Carlo sampling, or hit enrichment p-value by Fisher's exact test).

2.7 SPMA: Evaluating spectrum plots

Unbiased k-mer-based or transcript-based SPMA (see section 2.6) generate one spectrum plot for each RBP motif in the motif database. With currently 175 motifs, it is imperative to provide a means to automatically identify spectrum plots that exhibit a defined, non-random pattern. The following section describes methods available in Transite for separating spectrum plots with a coordinated pattern—a pattern that might be indicative of an underlying biological process—from spectrum plots without a clear trend, which are more likely to occur by chance. An example of the former is depicted in panel A of figure 2.7, the latter in panel B of the same figure. In this thesis the labels *non-random* and *random* are used to assign spectrum plots to one of the two categories.

A spectrum consists of three vectors, $\boldsymbol{s} \in \mathbb{R}^n$, $\boldsymbol{b} \in [1, n]^n$, and $\boldsymbol{p} \in \mathbb{R}^n$. $\boldsymbol{s}$ is a vector of scores (RBP target enrichment values, mean of enrichment values of RBP-associated k-mers, or RBP scores, respectively), $\boldsymbol{b}$ contains bin numbers (n, the number of bins, usually 40), and $\boldsymbol{p}$ contains p-values. p_i is the significance of score s_i in bin b_i.

Two different methods have been developed to obtain a spectrum score x, given $\boldsymbol{s}$, $\boldsymbol{b}$, and $\boldsymbol{p}$: (1) A local consistency score, which quantifies the local noise of the gradient in the spectrum, and (2) an approach using the adjusted R^2 value of a polynomial regression model, fitted to the gradient. Moreover, the coefficient of the linear term of the polynomial model can

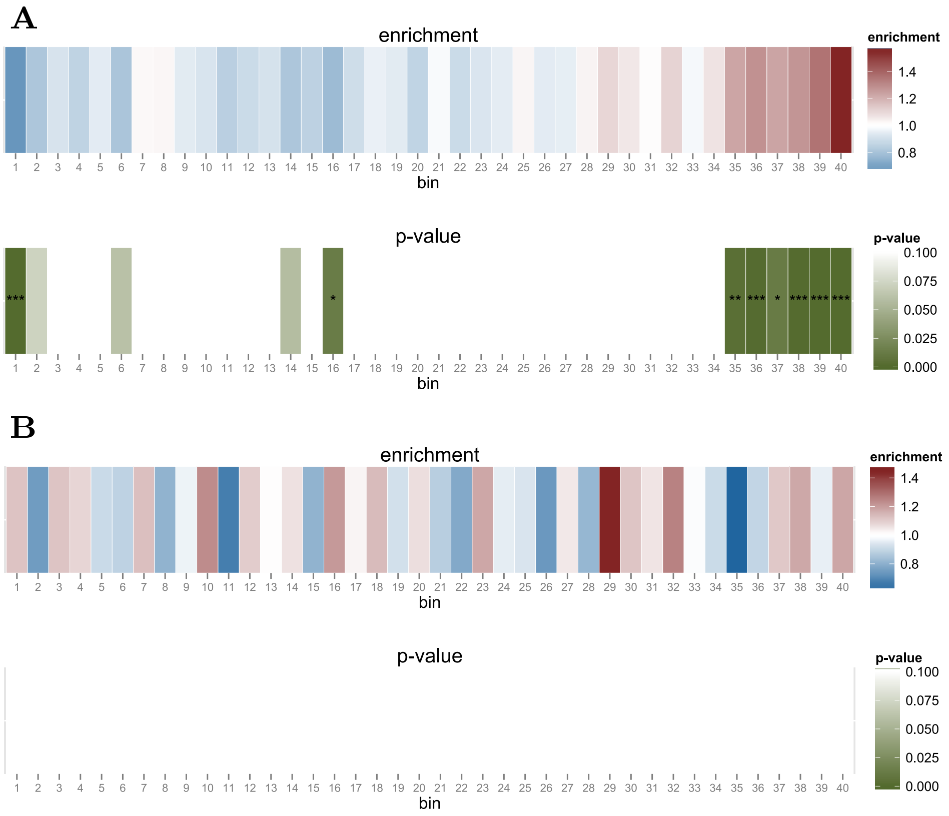


Figure 2.7: Groups of spectrum plots: (A) *Non-random* spectrum: Enrichment values form a gradient along the spectrum of ordered transcripts. **(B)** *Random* spectrum: Enrichment values do not follow a clear trend with respect to bin number.

be used to automatically distinguish between spectra with increasing and decreasing linear relationships as illustrated in figures 2.9 and 2.10.

2.7.1 Local consistency score

One way to quantify the meaningfulness of a spectrum plot is to calculate the deviance between the linear interpolation of the scores of two adjoining bins and the score of the middle bin, for each position in the spectrum. The lower the score, the more consistent the trend in the spectrum plot. Formally, the local consistency score x_c is defined as

$$x_c = \frac{1}{n} \sum_{i=1}^{n-2} \left| \frac{s_i + s_{i+2}}{2} - s_{i+1} \right|. \tag{2.6}$$

In order to obtain an estimate of the significance of a particular score x_c', Monte Carlo sampling is performed by randomly permuting the coordinates of the scores vector $\boldsymbol{s}$ and recomputing x_c. The probability estimate $\hat{p}$ is given by the lower tail version of the cumulative distribution function (see equation 3.29), where T equals x_c in equation 2.6.

2.7.2 Polynomial regression model

An alternative approach to assess the consistency of a spectrum plot is via polynomial regression. In a first step, polynomial regression models of various degrees are fitted to the data, i.e., the dependent variable $\boldsymbol{s}$ (vector of scores), and orthogonal polynomials of the independent variable $\boldsymbol{b}$ (vector of bin numbers). Secondly, the model that reflects best the true nature of the data is selected by means of the F-test. And lastly, the adjusted R^2 and the sum of squared residuals are calculated to indicate how well the model fits the data. These statistics are used as scores to rank the spectrum plots.

In general, the polynomial regression equation is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \epsilon_i, \tag{2.7}$$

where m is the degree of the polynomial (usually $m \leq 5$), and ϵ_i is the error term. The dependent variable $\boldsymbol{y}$ is the vector of scores $\boldsymbol{s}$ and $\boldsymbol{x}$ to $\boldsymbol{x}^m$ are the orthogonal polynomials of the vector of bin numbers $\boldsymbol{b}$.

Orthogonal polynomials are used in order to reduce the correlation between the different powers of $\boldsymbol{b}$ and therefore avoid multicollinearity in the model (see figure 2.8). This is important, because correlated predictors lead to unstable coefficients, i.e., the coefficients of a polynomial regression model of degree m can be greatly different from a model of degree $m + 1$.

The orthogonal polynomials of vector $\boldsymbol{b}$ are obtained by centering (subtracting the mean), QR decomposition, and subsequent normalization [19].

Given the dependent variable $\boldsymbol{y}$ and the orthogonal polynomials of $\boldsymbol{b}$ $\boldsymbol{x}$ to $\boldsymbol{x}^m$, the model coefficients $\boldsymbol{\beta}$ are chosen in a way to minimize the deviance between the actual and the predicted values characterized by equation 2.9, where $L(\text{actual value}, \text{predicted value})$ denotes the loss function.

$$\mathcal{M}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m \tag{2.8}$$

$$\mathcal{M} = \underset{\mathcal{M}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} L(y_i, \mathcal{M}(x_i)) \right) \tag{2.9}$$

Ordinary least squares is used as estimation method for the model coefficients $\boldsymbol{\beta}$. The loss function of ordinary least squares is the sum of squared residuals (SSR) and is defined as follows

$$SSR(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{2.10}$$

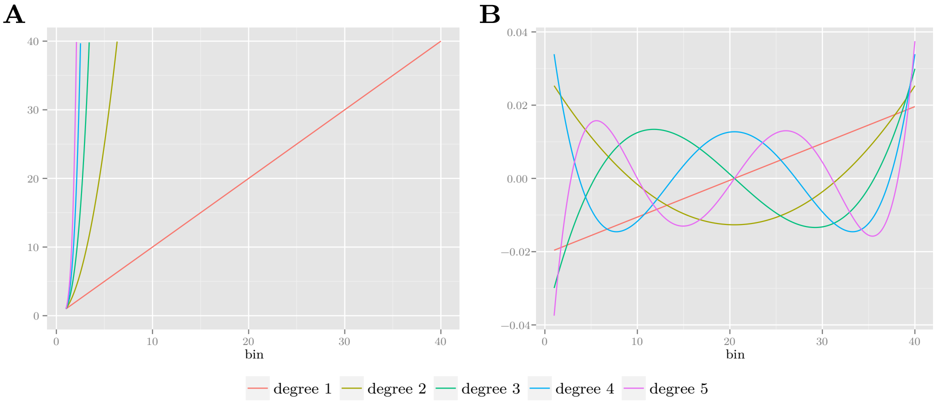


Figure 2.8: Ordinary and orthogonal polynomials: (A) The ordinary polynomials of degrees 1 to 5 are highly correlated. Moreover, polynomials of high degree can lead to floating point underflow of model coefficients. **(B)** Correlation between orthogonal polynomials is strongly reduced.

where $\boldsymbol{y}$ are the observed data and $\hat{\boldsymbol{y}}$ the model predictions.

Thus the ordinary least squares estimate of the coefficients $\hat{\boldsymbol{\beta}}$ (including the intercept $\hat{\beta}_0$) of the model $\mathcal{M}$ is defined by

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_i^j \right)^2 \right). \tag{2.11}$$

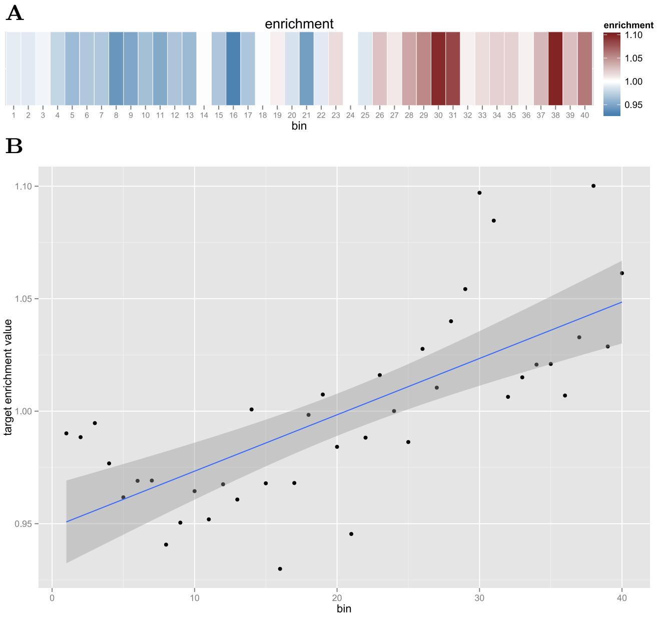


Figure 2.9: Spectrum with increasing linear relationship: (A) RBP binding evidence correlates with expression, i.e., positively regulated stability. **(B)** Linear approximation of the gradient.

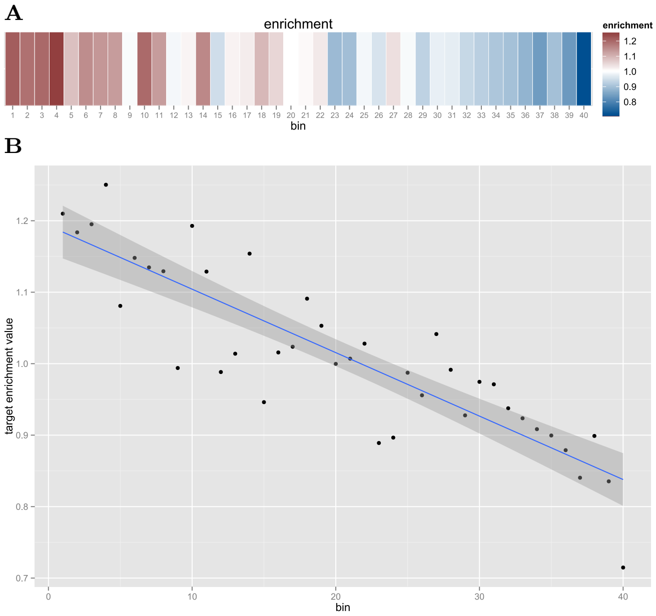


Figure 2.10: Spectrum with decreasing linear relationship: (A) RBP binding evidence anticorrelates with expression, i.e., negatively regulates stability. **(B)** Linear approximation of the gradient.

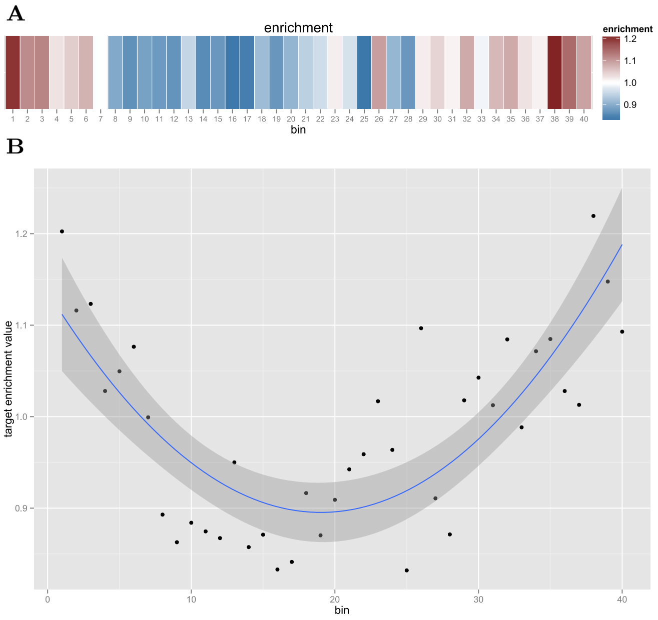


Figure 2.11: Spectrum with convex relationship: (A) RBP binding evidence increases in transcripts on either end of the spectrum. **(B)** Quadratic approximation of the gradient.

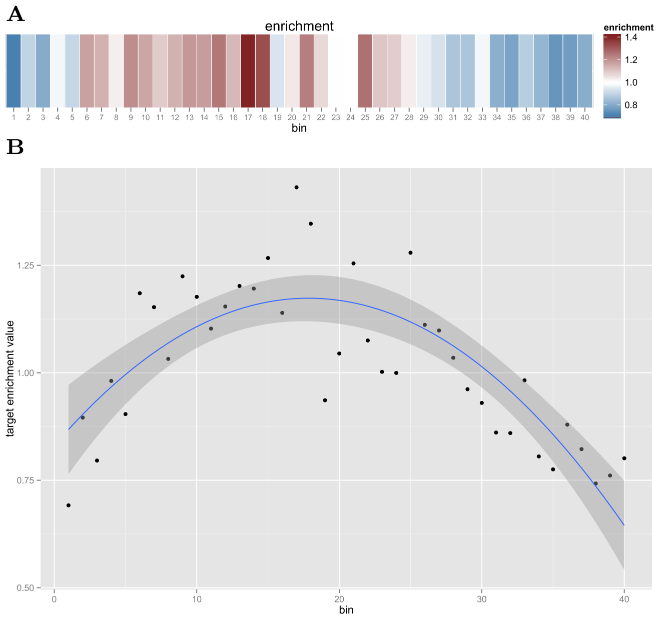


Figure 2.12: Spectrum with concave relationship: (A) RBP binding evidence decreases in transcripts on either end of the spectrum. **(B)** Quadratic approximation of the gradient.

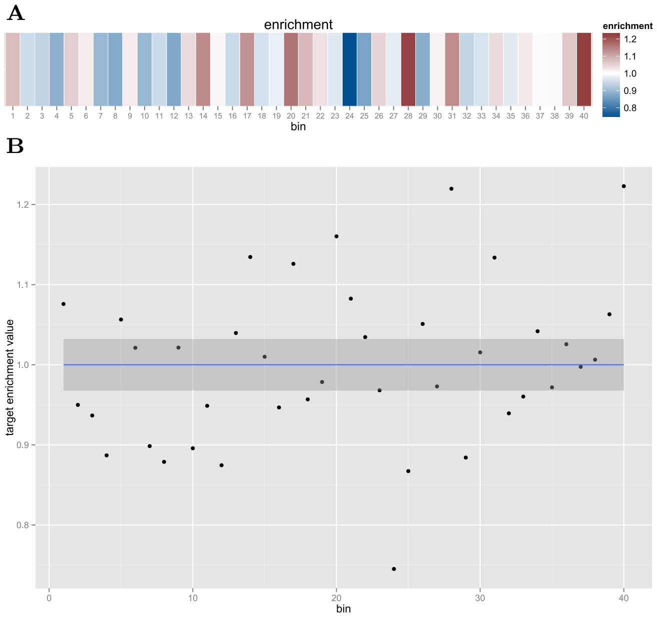


Figure 2.13: Inherently inconsistent spectrum: (A) Enrichment values do not follow a clear trend with respect to bin number. **(B)** No polynomial model decreased the *SSR* to an extent that justified the increase in complexity (degrees of freedom) compared to the null model.

Model selection via F-test

After polynomial models of various degrees have been fitted to the data, the F-test is used to select the model that best fits the data. Since the *SSR* monotonically decreases with increasing model degree (model complexity), the relative decrease of the *SSR* between the simpler model and the more complex model must outweigh the increase in model complexity between the two models. The F-test gives the probability that a relative decrease of the *SSR* between the simpler and the more complex model given their respective degrees of freedom is due to chance. A low p-value indicates that the additional degrees of freedom of the more complex model lead to a better fit of the data than would be expected after a mere increase of degrees of freedom.

The F-statistic is calculated as follows

$$F = \frac{(SSR_1 - SSR_2)/(p_2 - p_1)}{SSR_2/(n - p_2)}, \tag{2.12}$$

where SSR_i is the sum of squared residuals and p_i is the number of parameters of model i. The number of data points, i.e., bins, is denoted as n.

F is distributed according to the F-distribution with $df_1 = p_2 - p_1$ and $df_2 = n - p_2$.

Goodness-of-fit statistics

After a model has been selected, the adjusted R^2 is calculated as an additional way to evaluate the goodness of fit.

The R^2 statistic is 1 minus the ratio between the *SSR* (see equation 2.10) and the total sum of squares (*TSS*). *TSS* is given by

$$TSS = \sum_{i=1}^{n} (\bar{y} - y_i)^2, \tag{2.13}$$

where $\bar{y}$ is the mean of the observed data.

Because *SSR* decreases with every additional degree, R^2 increases. This behavior is undesirable as it favors models of high complexity which overfit the data. The adjusted R^2 corrects the R^2 statistic based on the residual degrees of freedom.

$$adjusted\ R^2 = 1 - \frac{SSR/(n-k)}{TSS/(n-1)}, \tag{2.14}$$

where n is the number of data points and k is the number of model parameters, i.e., the number of fitted coefficients.

2.8 Single Transcript Motif Analysis

For researchers interested in RBP binding sites in one specific transcript, Transite offers the Single Transcript Motif Analysis (STMA). In STMA the RBP motifs in Transite are utilized to score each position of the sequence of a single transcript. The computational representation of motifs and the scoring algorithm are described in section 3.5. Positions with a score higher than a relative threshold (e.g., 90% of the theoretical maximum score) are called hits. Hits in the transcript-based fashions of TSMA and SPMA are defined in an analogous manner.

Unfortunately, the identification of individual hits in a single transcript is an error-prone process with low specificity. This is not only true for the identification of potential RBP binding sites, but also for binding sites of transcription factors and kinases. The latter having the advantage of a bigger alphabet (20 amino acids instead of four bases). For an RBP sequence motif of length six (most motifs are six to eight nucleotides long), one would expect a perfect score (100% of theoretical maximum score) roughly every 4000 positions, because there are 4096 (4^6) different hexamers. In many motifs more than one hexamer yields a perfect score, which makes the inherent false positive rate even worse.

TSMA and SPMA are less affected by this drawback, because of the stabilizing effect of looking at thousands of sequences at once and focusing on ratios of hit occurrences (i.e., enrichments) in foreground and background sequences instead of absolute hit counts.

Positional evolutionary conservation scores are incorporated into STMA in order to discern low quality hits from high quality hits. Conservation scores are retrieved from the UCSC genome browser conservation track. Basewise conservation is based on the alignment of 7 vertebrate genomes by PhyloP [20] and PhastCons [21].

PhyloP conservation score: This method evaluates the scores for each position independently. Scores are signed -log p-values. A positive score indicates conserved, a negative score fast-evolving nucleotides.

PhastCons conservation score: A method based on a hidden Markov model that estimates the probability of each nucleotide to be part of a conserved region, thus the score ranges from 0 to 1. Unlike PhyloP, PhastCons considers the conservation of neighboring nucleotides, leading to a smoother score gradient.

Chapter 3

Materials and Methods

This chapter briefly covers existing statistical methods that are either used directly in the Transite pipeline or in the process of analyzing the results of Transite runs. The context of their application is explained in chapters 2 and 4. Furthermore, the two databases of RBP binding preferences are introduced.

3.1 Motif databases

Transite incorporates sequence motifs of RBP binding sites from two databases:

CIS-BP Catalog of Inferred Sequence Binding Preferences [22]

RBPDB a database of RNA-binding specificities [23]

Together they contribute 175 sequence motifs of varying lengths (between six and 18 nucleotides). All motifs were obtained using *in vitro* techniques for determining RNA targets. The majority of motifs were determined by either systematic evolution of ligands by exponential enrichment (SELEX) [24] or RNAcompete [25]. The RNA binding specificities of only two RBPs were obtained by electrophoretic mobility shift assays (EMSA) [26].

3.2 Methods for combining p-values

The following section describes methods to combine the significance (p-values) of enrichment values of a set of *k*-mers that are associated with an RNA-binding protein. These methods are used to obtain a single p-value for the overall significance of enriched or depleted RBP-associated *k*-mers.

In general, the methods of this section can be applied to combine the results of independent significance tests. They are commonly used in meta-analysis, where the goal is to systematically assess and integrate findings of a number of studies about a common body of research.

The problem can be specified as follows: Given a vector of n p-values $p_1, ..., p_n$, find p_c, the combined p-value of the n significance tests. Most of the methods introduced here combine the p-values in order to obtain a test statistic, which follows a known probability distribution. The general procedure can be stated as:

$$T(h, C) = \sum_{i=1}^{n} h(p_i) * C \tag{3.1}$$

The function T, which returns the test statistic t, takes two arguments. h is a function defined on the interval $[0, 1]$ that transforms the individual p-values, and C is a correction term.

3.2.1 Fisher's method

Fisher's method (1932) [27], also known as the inverse chi-square method is probably the most widely used method for combining p-values. Fisher used the fact that if p_i is uniformly distributed (which p-values are under the null hypothesis), then $-2 \log p_i$ follows a chi-square distribution with two degrees of freedom. Therefore, if p-values are transformed as follows,

$$h(p) = -2 \log p, \tag{3.2}$$

and the correction term C is neutral, i.e., equals 1, the following statement can be made about the sampling distribution of the test statistic T_f under the null hypothesis:

$$t_f \overset{H_0}{\sim} \chi^2_{2n}, \tag{3.3}$$

where n is the number of p-values.

3.2.2 Stouffer's method

Stouffer's method [28], or the inverse normal method, uses a p-value transformation function h that leads to a test statistic that follows the standard normal distribution by transforming each p-value to its corresponding normal score. The correction term scales the sum of the normal scores by the root of the number of p-values.

$$h(p) = \Phi^{-1}(1 - p) \tag{3.4}$$

$$C = \frac{1}{\sqrt{n}} \tag{3.5}$$

$$t_s \overset{H_0}{\sim} N(0, 1), \tag{3.6}$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function.

An extension of Stouffer's method with weighted p-values is called Lipták's method [29].

3.2.3 Mudholkar and George's method

The logit method by Mudholkar and George [30] uses the following transformation:

$$h(p) = -\ln(p/(1-p)) \tag{3.7}$$

When the sum of the transformed p-values is corrected in the following way:

$$C = \sqrt{\frac{3(5n+4)}{\pi^2 n(5n+2)}}, \tag{3.8}$$

the test statistic t_m is approximately t-distributed:

$$t_m \overset{H_0}{\sim} t_{5n+4} \tag{3.9}$$

3.2.4 Edgington's method

Edgington's method [31] is an additive procedure to combine p-values.

$$h(p) = p \tag{3.10}$$

The sampling distribution of the test statistic t_e under the null hypothesis is given by combinatorics:

$$Pr(t_e) = \sum_{r=0}^{\lfloor t_e \rfloor} (-1)^r \binom{n}{r} \frac{(t_e - r)^n}{n!} \tag{3.11}$$

3.2.5 Tippett's method

In Tippett's method [32] the smallest p-value is used as the test statistic t_t and the combined significance is calculated as follows:

$$Pr(t_t) = 1 - (1 - t_t)^n \tag{3.12}$$

3.3 Methods for adjusting p-values

When multiple statistical tests are performed in order to identify non-random events in a large pool of events, it is imperative to adjust either the p-values themselves or the significance level α, which is the probability of making a type I error (incorrectly rejecting the null hypothesis). Failure to do so leads to alpha error accumulation, i.e., many false positives.

Without accounting for alpha error accumulation in the hexamer enrichment step described in section 2.5.1, the enrichment values of on average 204 out of 4096 hexamers would be deemed significant between randomly chosen sets of sequences (assuming $\alpha = 0.05$). This is an immediate consequence of the number of tests (4096 in this case) and the probability of making a wrong decision per test (α).

Transite supports several methods to adjust p-values in order to avoid the multiple testing problem, all of which take a vector of p-values $\boldsymbol{p} \in \mathbb{R}^n$ and return a vector of adjusted p-values $\boldsymbol{q} \in \mathbb{R}^n$. The ith smallest or largest p-value is denoted by $p_{(i)}$, depending on whether the method belongs to the step-down (ordered from lowest to highest) or step-up (highest to lowest) group. The methods can be categorized according to the definition of type I error they control.

3.3.1 Familywise error rate controlling methods

The familywise error rate (FWER) is defined as

$$FWER = Pr(V > 0), \tag{3.13}$$

where V is the number of false positives in n tests (i.e., "the family").

Methods controlling the FWER guarantee that $FWER \leq \alpha$.

Holm's method

The adjusted p-values [33] obtained by Holm's method are defined as

$$q_{(i)} = \max_{j \leq i} \left(\min \left((n - j + 1) p_{(j)}, 1 \right) \right), \tag{3.14}$$

where $p_{(j)}$ is the jth lowest p-value and thus characterizing Holm's approach as a step-down method.

Hochberg's method

Hochberg's method is the step-up version of Holm's method ($p_{(i)}$ is highest p-value) and is uniformly more powerful [34].

$$q_{(i)} = \begin{cases} p_{(n)} & \text{for } i = n \\ \min \left(q_{(i+1)}, (n - i + 1) p_{(i)} \right) & \text{otherwise} \end{cases} \tag{3.15}$$

Bonferroni's method

Bonferroni corrected p-values [35] are given by

$$q_i = \min(p_i * n, 1). \tag{3.16}$$

It is the oldest and most conservative correction.

3.3.2 False discovery rate controlling methods

The false discovery rate (FDR) is defined as

$$FDR = E\left(\frac{V}{V+S}\right), \tag{3.17}$$

where V is the number of false positives and S the number of true positives in n tests.

Methods controlling the FDR are less conservative than the ones controlling the FWER.

Benjamini and Hochberg's method

Similar to Hochberg's method for controlling the familywise error rate, this method is defined as a step-up adjustment [36]:

$$q_{(i)} = \begin{cases} p_{(n)} & \text{for } i = n \\ \min\left(q_{(i+1)}, \frac{n}{i} p_{(i)}\right) & \text{otherwise} \end{cases} \tag{3.18}$$

Compared to the FWER controlling method, the multiplier is less conservative ($\frac{n}{i}$ to $n-i+1$), leading to smaller adjusted p-values. This method can be used if the components (i.e., p-values) of $\boldsymbol{p}$ are independent and uniformly distributed.

Benjamini and Yekutieli's method

If there are dependencies among the p-values or if independency cannot be guaranteed, Benjamini and Yekateuli's method [37] can be used instead:

$$q_{(i)} = \begin{cases} \gamma p_{(n)} & \text{for } i = n \\ \min\left(q_{(i+1)}, \gamma \frac{n}{i} p_{(i)}\right) & \text{otherwise} \end{cases} \tag{3.19}$$

where $\gamma = \sum_{i=1}^{n} \frac{1}{i}$.

3.4 Similarity coefficients for binary data

This section introduces three coefficients that quantify the similarity between two binary vectors. In this thesis they are used to compare the results of two different analysis approaches in Transite (see section 4.3).

The two binary vectors can also be represented as a contingency table of two binary attributes. The agreement (or disagreement, respectively) of the two attributes is captured in four numbers, a, b, c and d, which correspond to the cells of the four-fold table 3.1. Three commonly used similarity coefficients for binary data are defined on the basis of a, b, c and d.

The attributes *spectrum label (fold change)* and *spectrum label (p-value)* will be introduced in section 4.2 and their meaning is not important to describe the general concept of similarity coefficients.

Table 3.1: Contigency table with Transite spectrum label attributes

	spectrum label (fold change)	
spectrum label (p-value)	non-random	random
non-random	a	b
random	c	d

3.4.1 Matthews correlation coefficient

It is also known as ϕ coefficient. The *MCC* is for binary data the equivalent to the Pearson product-moment correlation coefficient for continuous data [38], and is defined as

$$x = (a+b) * (a+c) * (b+d) * (c+d) \tag{3.20}$$

$$MCC = \begin{cases} ad - bc & \text{for } x = 0 \\ \frac{ad-bc}{\sqrt{x}} & \text{otherwise} \end{cases} \tag{3.21}$$

MCC can be tested for significance using the χ^2 distribution with one degree of freedom:

$$(a+b+c+d) * MCC^2 \overset{H_0}{\sim} \chi^2_1. \tag{3.22}$$

3.4.2 Simple matching coefficient

The simple matching coefficient is the ratio of concordant labels to all labels and is naturally defined as

$$SMC = \frac{a+d}{a+b+c+d}. \tag{3.23}$$

3.4.3 Jaccard similarity coefficient

Unlike *MCC* and *SMC*, Jaccard similarity coefficient does not include the quantity d, i.e., the cases labeled as *random* by both approaches.

$$JSC = \begin{cases} 1 & \text{for } a = b = c = 0 \\ \frac{a}{a+b+c} & \text{otherwise} \end{cases} \tag{3.24}$$

3.5 Motif representations

Position specific scoring matrices (PSSM) are used to represent sequence motifs. Transite inherits PSSMs describing RBP binding sites from two sources (see section 3.1). Motif databases provide PSSMs in one of three types: Position frequency matrices (PFM), position probability matrices (PPM), or position weight matrices (PWM). Internally, Transite algorithms work exclusively with PWMs in order to make subsequent calculations more efficient.

The elements of a PFM represent absolute frequencies of each nucleotide at each position. In general, PFMs are not bound to a specific alphabet (A, C, G, and U in this case), but can also be used with other types of sequences, e.g., protein sequences. How PPMs and PWMs are derived from PFMs is stated in equations 3.26 and 3.27.

$$\overbrace{\begin{array}{cc} & \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{U} \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} & \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & x_{n,4} \end{bmatrix} \end{array}}^{\text{PFM}} \overset{f_1}{\rightsquigarrow} \overbrace{\begin{array}{cc} & \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{U} \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} & \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & y_{1,4} \\ y_{2,1} & y_{2,2} & y_{2,3} & y_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{n,3} & y_{n,4} \end{bmatrix} \end{array}}^{\text{PPM}} \overset{f_2}{\rightsquigarrow} \overbrace{\begin{array}{cc} & \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{U} \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} & \begin{bmatrix} z_{1,1} & z_{1,2} & z_{1,3} & z_{1,4} \\ z_{2,1} & z_{2,2} & z_{2,3} & z_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n,1} & z_{n,2} & z_{n,3} & z_{n,4} \end{bmatrix} \end{array}}^{\text{PWM}} \tag{3.25}$$

The conversion functions f_1 (from PFM to PPM) and f_2 (from PPM to PWM) are applied to each element of the matrix.

$$f_1(\boldsymbol{x}, i, j) = \frac{x_{i,j}}{\sum_k x_{i,k}}, \tag{3.26}$$

where $\boldsymbol{x}$ is a PFM, and i and j are its indices.

$$f_2(\boldsymbol{y}, i, j) = \log_2 \frac{y_{i,j}}{p_j} \tag{3.27}$$

where $\boldsymbol{y}$ is a PPM, and p_j is the *a priori* probability of nucleotide j. In Transite, nucleotides are assumed to be equiprobable ($\Pr(\mathrm{A}) = \Pr(\mathrm{C}) = \Pr(\mathrm{G}) = \Pr(\mathrm{U}) = 0.25$).

Laplace smoothing: Laplace smoothing (also known as additive smoothing) is applied to avoid zeros in PFMs and PPMs. Zeros might occur if the number of sequences on which the PSSM is based, is too small to contain at least one occurrence of each nucleotide per position. In this case, pseudocounts are introduced [39].

Scoring algorithm: After Laplace smoothing was applied and PFMs and PPMs were converted to PWMs, the scoring algorithm itself is trivial: In order to obtain the score of a hexamer with a PWM of length six, the elements in the PWM that correspond to the correct nucleotide per position are added up. If this score is above zero, the hexamer is more likely to be found in the binding site described by this PWM rather than in a random sequence of length six.

3.6 Monte Carlo tests

Permutation tests are a means to determine the statistical significance of a test statistic with an unknown null distribution. Since no assumptions are made about the underlying distribution of the statistic, permutation tests belong to the group of nonparametric tests. The null distribution of the statistic is obtained empirically by calculating all possible values of the statistic by rearrangement of the labels of the observations (data points). Each unique ordering of the labels is called a permutation, hence the name. Labels are categorical variables that subdivide the set of observations into groups, e.g., *treatment* and *control*. In order to build the sampling distribution of the test statistic T based on n labeled observations, T needs to be calculated for $n!$ permutations of the observation labels. The upper tail probability of the actual test statistic, i.e., the test statistic T calculated with the actual observations x, here denoted $T(x)$, is given as follows:

$$Pr(T(x)) = \sum_{y:T(y)\geq T(x)} Pr(y), \tag{3.28}$$

where y are the permuted observations.

Since the number of permutations grows factorially with the number of observations, calculating T for all permutations is infeasible even for small numbers of n. Therefore, instead of building the complete null distribution, a sample of the distribution is picked randomly to determine an estimate of the probability of $T(x)$. This process is called Monte Carlo sampling. The estimate is determined by the empirical cumulative distribution functions (lower-, upper- and two-tailed probability):

$$\hat{Pr}(T(x)) = \frac{\sum\limits_{i=1}^{n} \mathbb{1}\left(T(y_i) \leq T(x)\right) + 1}{n+1} \tag{3.29}$$

$$\hat{Pr}(T(x)) = \frac{\sum\limits_{i=1}^{n} \mathbb{1}\left(T(y_i) \geq T(x)\right) + 1}{n+1} \tag{3.30}$$

$$\hat{Pr}(T(x)) = \frac{\sum\limits_{i=1}^{n} \mathbb{1}\left(|T(y_i)| \geq |T(x)|\right) + 1}{n+1}, \tag{3.31}$$

where $\mathbb{1}$ is the indicator function and n is the sample size, i.e., the number of performed permutations.

One is added to both the numerator and the denominator to avoid p-values of zero when the actual test statistic is smaller than all of the test statistics of the permuted data [40].

A confidence interval around $\hat{p}$, i.e., $\hat{Pr}(T(x))$, can be calculated based on the cumulative probabilities of the binomial distribution. This interval is referred to as Clopper-Pearson interval [41]. The exact confidence limits c_l and c_u satisfy the following equations:

$$\sum_{i=n_1}^{n} \binom{n}{i} c_l^i (1 - c_l)^{n-i} = \alpha/2 \tag{3.32}$$

$$\sum_{i=0}^{n_1} \binom{n}{i} c_u^i (1 - c_u)^{n-i} = \alpha/2, \tag{3.33}$$

where n_1 is the number of cases where $T(y_i) \geq T(x)$ (see equation 3.30). If $n_1 = 0$, the lower confidence limit is 0, whereas if $n_1 = n$, the upper limit is 1.

Chapter 4

Results

This chapter contains evaluations of aspects of the newly developed Transite pipeline (see chapter 2), as well as its applications. The evaluation of sorting approaches for SPMA can be found in section 4.3 and the evaluation of permutation approaches for TSMA and SPMA is contained in section 4.5. Section 4.6 presents the combined results of Transite analyses of gene expression data from cisplatin-treated samples in five different cell lines. Brief comments about the development of the Transite R package and the website can be found in sections 4.7 and 4.8, respectively.

4.1 Transite configuration

Transite analysis runs can be customized in various ways. This section describes the settings with which the results of this chapter were generated.

P-value adjustment method: The Benjamini-Hochberg procedure was used to adjust (1) differential expression p-values obtained by `limma`, (2) k-mer enrichment p-values calculated by Fisher's exact tests, (3) k-mer enrichment p-values approximated by χ^2 tests, (4) empirical p-values of the geometric mean of motif-associated hexamer enrichment values obtained by Monte Carlo tests, (5) empirical p-values of motif scores obtained by Monte Carlo tests, and (6) target site enrichment p-values obtained by Fisher's exact tests. These six categories of adjusted p-values can be related to the steps of the Transite pipeline schematic in figure 2.1: P-values of category (1) belong to step three, categories (2) to (5) occur in the hexamer-based fashion of steps five (TSMA) and six (SPMA), and p-values of category (6) are generated in the transcript-based fashion of the aforementioned steps.

P-value combining method: Fisher's, Stouffer's, Tippett's, and Mudholkar and George's methods were used in parallel to calculate combined k-mer enrichment p-values of all hexamers associated with a particular motif.

No method is universally superior to all others. Their applicability depends on the pattern of evidence, that is to say how the total evidence is distributed across the individual p-values. Edgington's method was discarded due to its generally very poor power [42].

Significance threshold for differentially expressed genes: 0.05, after adjustment method was applied.

***k*-mer length:** $k = 6$. Although Transite can operate with heptamers and octamers as well, hexamers are recommended. The results of heptamer-based and octamer-based runs are similar to hexamer-based runs and do not justify the tremendous increase in run time ($\mathcal{O}(4^k)$).

Significance threshold for *k*-mer enrichment: 0.01 - after p-value adjustment method was applied.

Number of *k*-mer enrichment permutations: 5000. In case Bonferroni, the most conservative p-value adjustment method is used, 5000 permutations are still enough to get significant results with 175 motifs ($\frac{1}{5001} * 175 = 0.034993$).

Significance threshold for motif scores: 0.05 - after p-value adjustment method was applied.

Transcript sorting approach: By fold change, for general explanation see section 2.4.2, sorting approaches are compared in section 4.3.

Threshold for PWM hits: 90% of the theoretical maximum score.

Maxmimum number of hits per transcript: 5, transcripts with more than five potential binding sites contribute only five hits to the target hit enrichment calculation (see section 4.4).

Number of bins: 40, see next section.

4.1.1 Choice of bin number

In this section the influence of the number of bins on the SPMA result is examined. The bin number determines in how many foreground sets the sorted "spectrum" of transcripts is subdivided. This procedure is described in more detail in section 2.4.2.

The results of 32 transcript-based SPMA runs with various bin numbers (including 40—the Transite default bin number setting—in bold face)

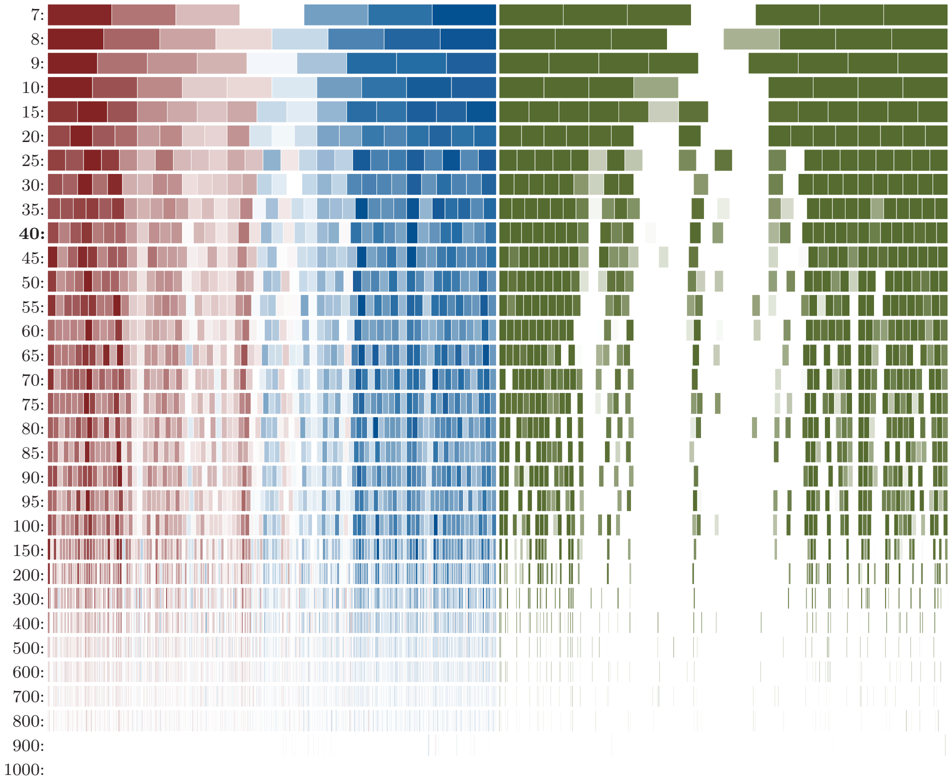


Figure 4.1: Influence of bin number on transcript-based SPMA spectra: This figure shows 32 rows of spectrum plots with bin numbers ranging from 7 to 1000. The left column contains a color representation of transcript-based SPMA target enrichment values from blue (underrepresented) to red (overrepresented) and the right column depicts their corresponding p-values, where a dark green hue indicates high significance.

are shown in figure 4.1. The results were obtained with the GEO series `GSE46493`, which serves as an exemplary data set. The spectrum plots depict enrichment values of potential targets of the RNA-binding proteins CPEB2 and CPEB3 motif `M012_0.6`. This motif was selected because its spectrum yielded the highest adjusted R^2 value in the transcript-based SPMA run with 40 bins.

In addition to the SPMA run with 40 bins, motif `M012_0.6` scored first place (out of 175) in 25 of 31 SPMA runs with varying bin numbers. It yielded the second highest adjusted R^2 in runs with 15, 30, 85 and 100 bins, the fifth highest runs with eight and nine bins, and sixth place in the SPMA run with seven bins. In general, the obtained spectra of `M012_0.6` and other motifs did not change significantly as a function of bin number. Consequently, the spectrum characteristics (adjusted R^2, consistency score, slope), on which the identification of non-random spectra is based, did not change substantially. Spectrum characteristics were especially stable between runs with bin numbers in the range of 20 to 60. This becomes apparent when spectra of the same motif but with different bin numbers are juxtaposed as in figure 4.1. The conclusion is always the same, irrespective of the number of bins: the number of potential target sites of RNA-binding proteins CPEB2 and CPEB3 in 3' UTRs of transcripts is negatively proportional to the fold change of said transcripts. Considering the specifics of the gene expression experiment at hand (GEO series `GSE46493`), one could draw the less abstract conclusion, that CPEB2/CPEB3 targets are downregulated in U2OS cells after doxorubicin treatment.

Returning to the question of the appropriate bin number, the analysis showed that there is a range of acceptable bin numbers, of which 40 is one of them. In most cases the general trend is apparent with as few as ten bins, but at least 20 bins are recommended due to difficulties arising from fitting a polynomial to too few observations, which is essential to identify *meaningful* spectra (see section 2.7.2). Since run time increases linearly with the bin number, there is little reason to choose numbers higher than 50. Another reason to discard high bin numbers is reduced interpretability. Especially spectrum plots with more than 100 bins tend to be less convenient to interpret due to their jagged gradient.

4.2 Spectrum labeling

A single run of transcript-based SPMA produces 175 spectrum plots (one for each RBP currently in Transite) and the hexamer-based approach is even more elaborate with twice as many spectrum plots (175 hexamer enrichment spectrum plots and 175 hexamer-based motif score spectrum plots). Visual inspection of each one is tedious at best, if not infeasible. Therefore, it is imperative to provide some sort of quality metric to introduce a ranking for spectrum plots, which in turn can be used to focus the user's attention on the most *interesting*, i.e., non-random, spectrum plots. The inner workings of the quality metrics are described in section 2.7. In essence, two spectrum characteristics serve as quality metrics: adjusted R^2 and the p-value of the consistency score. They can be used to introduce the necessary ranking for spectrum plots, which then allows the Transite user to visually inspect the spectrum plots of the top end of the list (to present the presumably more *interesting* ones first). If the goal is to judge the agreement between several SPMA runs, a binary label per spectrum plot (*non-random*, *random*) is used. These labels are based on whether predefined constraints are met. This way the comparison is not influenced by potentially biased visual inspection.

Spectrum plots that meet the following constraints are labeled *non-random* (and *random* otherwise):

- adjusted $R^2 \geq 0.4$,

- consistency score p-value < 0.1,
- at least 4 out of 40 bins with significant ($\alpha = 0.05$) p-values.

These thresholds are not rigorously derived with regard to any statistical property, but rather local, i.e., pertaining to this thesis, conventions similar to the (global) convention that a p-value less than 0.05 is deemed significant. They were established by visually categorizing blinded spectrum plots, i.e., spectrum plots without motif labels. The thresholds are selected to be as close as possible to the manual assignment of *non-random-random* labels.

4.3 Comparison of sorting approaches

As described in section 2.4.2, there are two predefined ways to sort transcripts in order to subdivide them into a number of bins that define the foreground sets of SPMA. (1) sort according to the transcript fold change (FC) and (2) according to the signed log p-value.

In order to investigate how the transcript sorting approach influences the result of SPMA, i.e., the specta of which RBPs are identified as non-random ("spectrum label"), transcript-based SPMA runs of 28 data sets were performed with both the fold change sorting and the p-value sorting approach. The results were subsequently compared using similarity coefficients to quantify the agreement between the spectrum labels obtained by the two sorting approaches for each data set.

If a spectrum met the constraints defined in section 4.2, it was labeled *non-random*, and *random* otherwise. On this level, the results of one SPMA run were represented by a 175-dimensional binary vector, where each component indicates the label of one of 175 spectra (one spectrum for each RBP motif in Transite).

Similarity coefficients for binary data were used to quantify the similarity between two 175-dimensional vectors, representing the results of the two sorting approaches for the same data set.

Only the results of the (deterministic) transcript-based SPMA runs were examined, because the stochasticity introduced by Monte Carlo tests in hexamer-based SPMA runs could potentially conceal the differences between the sorting approaches.

In panels A and B of figure 4.2 one can appreciate that the ratios between non-random (positive and negative slope, respectively) and random spectrum labels of the two approaches are similar, but fold change sorted SPMA deems slightly more spectra *non-random*. Among spectra with positive slope, 247 were labeled *non-random* in p-value sorted SPMA, compared to 304 with fold change sorted SPMA. Among negative slope spectra the two approaches are more similar with 292 to 311 *non-random* labels. In total, there were 4900 spectrum labels (28 data sets, 175 spectra per data set), the vast majority of which were labeled *random* by both approaches (89%

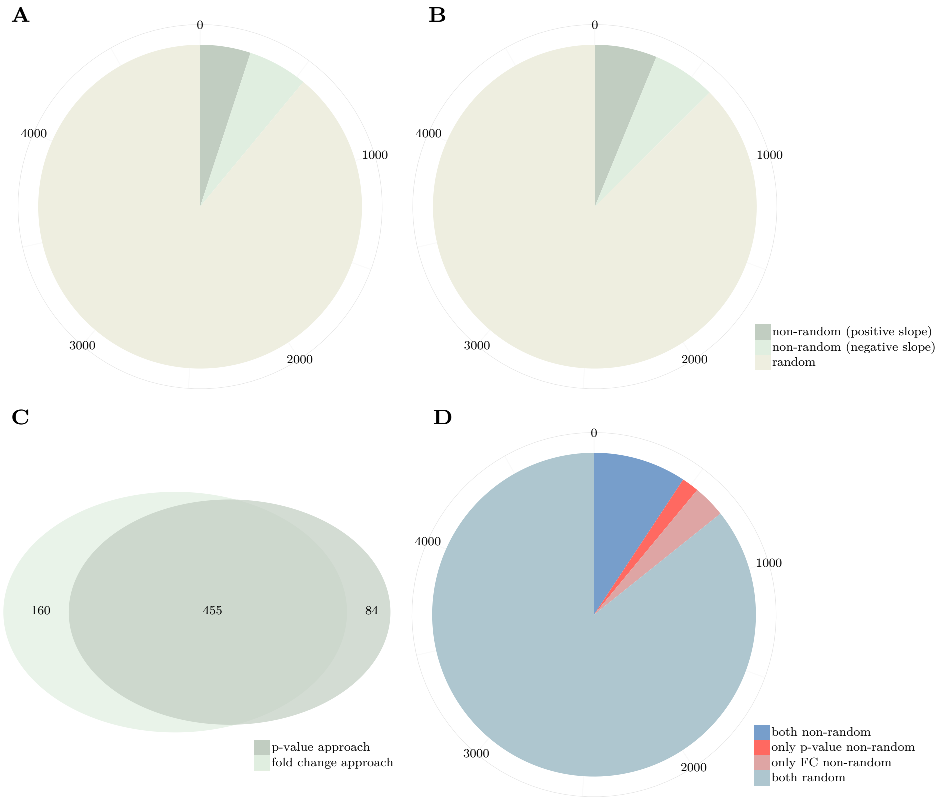


Figure 4.2: Agreement between p-value and fold change sorting approaches: The pie charts in panels **A** and **B** depict the fractions of non-random spectra with positive, non-random spectra with negative slope and random spectra for the p-value approach (panel **A**) and the fold change approach (panel **B**). **(C)** Venn diagram of spectra with non-random label. **(D)** Agreement between spectrum labels of the two approaches, blue hues indicate agreement, red hues disagreement.

of p-value sorted spectra and 87% of fold change sorted spectra).

The overall agreement between the two approaches is visualized in panels C and D of figure 4.2. Counting the label decisions of both approaches in all 4900 cases, we arrive at the contigency table 4.1, which yields the following similarity coefficients: a highly significant *MCC* of 0.76, a *SMC* of 0.95, and a *JSC* of 0.65. According to the general rule of thumb for correlation coefficients, it can be interpreted as a moderate (*JSC*) to strong (*MCC*, *JSC*) positive association between the results of p-value sorted and fold change sorted SPMA.

The heat map in figure 4.3 represents the similarity coefficients for the spectrum label assignments of all 28 data sets, separated in spectra with positive and negative slope. The slope separation is a cautionary measure,

Table 4.1: Spectrum labels of p-value and fold change sorted SPMA

	spectrum label (fold change)	
spectrum label (p-value)	non-random	random
non-random	455	84
random	160	4201

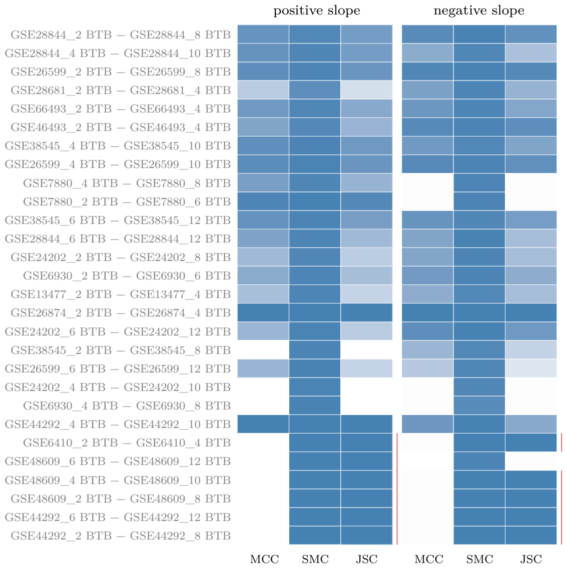


Figure 4.3: Similarity coefficients between sorting approaches: Colors represent similarity coefficients from zero (white) to one (dark blue). There are no negative coefficients. The red line to the right hand side of the heat map rows indicate SPMA runs without *non-random* labeled spectra ($a = b = c = 0$).

which prevents hidden mislabelings, i.e., the spectra of one RBP motif are labeled *non-random* in both approaches, but exhibit opposite gradients. In the case where all spectra are labeled *random* by both approaches ($a = b = c = 0$), *MCC* is zero, whereas *SMC* and *JSC* are one. Intuitively, the latter is closer to the notion of similarity that is appropriate here.

In order to get an even closer look at the differences and similarities between the two sorting approaches, the correlation between spectrum characteristics (adjusted R^2, consistency score p-value and slope) was examined. For this purpose the characteristics of two exemplary data sets—one with a strong spectrum label similarity between the sorting approaches, the other with a weaker similarity—and their correlation are depicted in figures 4.4 and 4.5. The long stretches of highly correlated adjusted R^2 and slope values are due to the fact that most spectra do not exhibit a gradient that can be

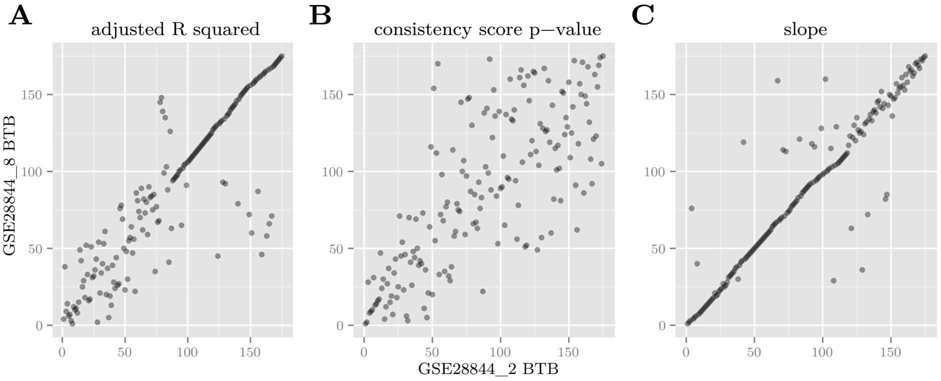


Figure 4.4: Similarity coefficients of a data set with strong spectrum label association between p-value and fold change approaches: Pearson's product-moment correlation coefficients for the three measures are 0.89 (**A**), 0.52 (**B**) and 0.93 (**C**).

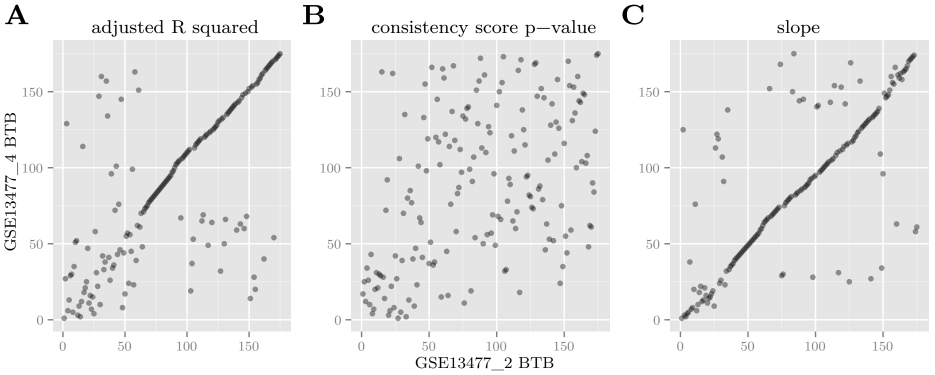


Figure 4.5: Similarity coefficients of a data set with a weak to moderate spectrum label association between p-value and fold change approaches: Pearson's product-moment correlation coefficients for the three measures are 0.68 (**A**), 0.30 (**B**) and 0.59 (**C**).

approximated with a polynomial well enough to yield statistically significant coefficients (see section 2.7.2 about model selection), leading to adjusted R^2 and slope values of zero.

In conclusion, signed p-value and fold change sorting approaches yield similar, but not equivalent SPMA results. In general, more spectra are labeled *non-random* when the latter is used. A disadvantage of the sorting by p-value is the fact that some transcripts with statistically significant (differential expression) p-values lack a large enough fold change to be biologically meaningful.

4.4 Adverse effects of transcripts with multiple hits

In transcript-based TSMA and SPMA the score for RBP binding evidence is the (putative) target site enrichment between foreground and background sets and its p-value. The procedure is explained in section 2.5.2 in detail, but in short, all target sites in all transcripts of the foreground set are added up and compared to the number of target sites of all transcripts in the background set. Fisher's exact test is then used to obtain a p-value, i.e., the significance of the observed enrichment. One of the assumptions of Fisher's exact test is the independence of the observations. In this case the assumption states that whether one target site is in the foreground set does not depend on any other target site. While this assumption holds true on a transcript level (fold changes of transcripts are assumed to be independent from each other), the target sites of a single transcript are not independent from another. If one target site of a certain transcript is in the foreground set, all other target sites of that transcript are also in the foreground set.

A solution to completely avoid the interdependence of target sites of the same transcript would be to count only one site per transcript. But according to previous studies [43], the number of binding sites per transcript is an indicator of the strength of the regulatory effect, e.g., the more ELAVL1/HuR binding sites a transcript exhibits, the more highly stabilized it is. This finding suggests to preserve the number of binding sites per transcript. In order to find a tradeoff between the number of false-positives due to interdependences between binding sites and the maximum number of recognized putative binding sites (i.e., hits) per transcript, the number of expected false-positives was obtained as a function of the maximum number of recognized hits per transcript. Transcript-based SPMA was performed on transcripts of the `GPL570` platform with random fold change values. In this case one would expect only insignificant target site enrichments. SPMA runs were carried out with *maximum number of recognized hits per transcript* values ranging from 1 to 30, and unbounded. If this number is 15 or higher, on average around 250 out of 7000 bins (40 bins per motif, 175 motifs) turn out to have significant enrichment values, which can be considered false-positives. The number of false-positives highly correlates with the number of hits per transcript. The data are displayed in figure 4.6. As a result, a maximum of 5 hits per transcripts were chosen as the ideal setting.

Important to note is, that none of the 24,500 spectra (175 spectra per run, 10 runs per setting, 14 settings) was labeled *non-random* as defined by the criteria in section 4.2. In that sense there were no false-positives, regardless of the maximum hits per transcripts setting. The false-positives at bin enrichment level did not exhibit a coherent pattern on the spectrum level.

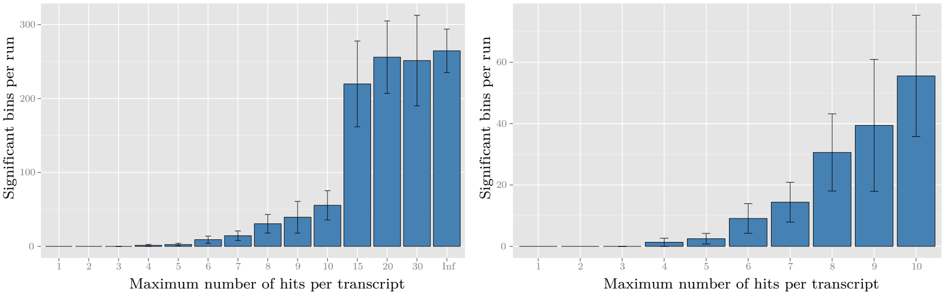


Figure 4.6: False-positives due to interdependencies: The numbers were obtained by 10 transcript-based SPMA runs per *maximum number of hits per transcript* setting with transcripts of the `GPL570` platform with random fold change values. Error bars indicate 95% confidence intervals. The right panel is a zoom-in of the left panel.

4.5 Permutation approaches for Monte Carlo tests

This section contains an evaluation of the three different permutation approaches for the Monte Carlo sampling procedure, which is used to obtain the null distribution of the geometric mean of motif-associated hexamer enrichment values. The null distribution is required to determine empirical p-values in hexamer-based TSMA and SPMA. The theoretical underpinnings of Monte Carlo tests are described in section 3.6.

To be completely independent of any preexisting pattern in real sequencing data, this analysis used simulated data, consisting of 10,000 random sequences, each 1600 nucleotides long, with equal proportions of all four nucleotides.

The sequences were divided into 100 folds, where the first fold contained the first hundred sequences, the second fold the second hundred, and so forth. The sequences were broken into hexamers and their enrichment values were calculated between each fold (foreground sets) and all folds (background set), which is analogous to the enrichment value calculations in hexamer-based SPMA. Lastly, the geometric mean of enrichment values of motif-associated hexamers were calculated for three exemplary motifs (`LC1`, `M031_0.6`, `M152_0.6`). This was done for each fold, yielding 100 enrichment means per motif.

When empirical p-values of the enrichment means were obtained by Monte Carlo tests, one would expect them to be uniformly distributed between 0 and 1—given that the sequences and the subdivision into folds were random, i.e., the null hypothesis was true.

The aim is to decide which of the three approaches comes closest to the expected uniform distribution of enrichment value p-values.

The three motifs, which are presented in the next section, were selected

Table 4.2: Multiple sequence alignment of associated hexamers of `LC1` motif compared with the alignment of random hexamers

`LC1` hexamers		random hexamers	
AUUUAU	--AUUUAU-	AACUAC	--AACUAC--
AUUUUU	-AUUUUU--	ACAGGG	--ACAGGG-
UAUUUA	-UAUUUA-	ACCUAU	--ACCUAU--
UAUUUU	UAUUUU--	AUUUCA	-AUUUCA---
UGUUUU	--UGUU-UU	CACUAU	--CACUAU--
UUAUUU	-UUAUUU--	CGUUGC	CGUUGC---
UUCUUU	-UUCUUU--	GAGGUA	---GAGGUA-
UUGUUU	-UUGUU-U-	GCAAAC	GCAAAC---
UUUAUU	--UUUAUU	GGAACU	-GGAACU---
UUUCUU	UUUCUU--	GUGUGA	---GUGUGA-
UUUGUU	-UUUGUU--	UAAUGU	--UAAUGU--
UUUUGU	UUUUGU--	UAGCCG	-UAGCCG--
UUUUUA	-UUUUUA-	UCCCUC	-UCCCUC--
UUUUUC	-UUUUUC-	UGGUGG	---UGGUGG
UUUUUG	-UUUUUG-	UUCAAU	--UUCAAU--
UUUUUU	-UUUUUU--	UUUCAU	-UUUCAU--

as to represent the span of motifs in the Transite database; unspecific (i.e., many motif-associated hexamers) U-rich motifs (`LC1`), specific U-rich motifs (`M031_0.6`), and motifs with a balanced distribution of nucleotides in their associated hexamers (`M152_0.6`). `LC1` was specifically chosen to examine how the permutation approaches perform with respect to interdependent sets of motif-associated motifs. Such interdependencies are a result of partly overlapping hexamers, as illustrated by the multiple sequence alignments in table 4.2.

4.5.1 Permutation approaches

P1: Choose *k* hexamers randomly

One way to obtain samples of the null distribution is to choose a different set as motif-associated hexamers and calculate the geometric mean of their enrichment values. For example the null distribution of enrichment means of motif `LC1` was obtained by randomly choosing 16 hexamers, calculating their geometric mean, and repeating this process 10,000 times.

$$\overbrace{\begin{pmatrix} \text{AAAAAA}_1 \\ \text{AAAAAC}_2 \\ \text{AAAAAG}_3 \\ \vdots \\ \text{GGUUUU}_{4094} \\ \text{GUUUUU}_{4095} \\ \text{UUUUUU}_{4096} \end{pmatrix}}^{\boldsymbol{x}} \overset{\text{shuffle}}{\rightsquigarrow} \overbrace{\begin{pmatrix} \text{AAAAAA}_1 \\ \text{AAAAAC}_2 \\ \text{AAAAAG}_3 \\ \vdots \\ \text{GGUUUU}_{4094} \\ \text{GUUUUU}_{4095} \\ \text{UUUUUU}_{4096} \end{pmatrix}}^{\boldsymbol{r_1}} \cdots \overbrace{\begin{pmatrix} \text{AAAAAA}_1 \\ \text{AAAAAC}_2 \\ \text{AAAAAG}_3 \\ \vdots \\ \text{GGUUUU}_{4094} \\ \text{GUUUUU}_{4095} \\ \text{UUUUUU}_{4096} \end{pmatrix}}^{\boldsymbol{r_n}} \tag{4.1}$$

where $\boldsymbol{x}$ is a vector of hexamer enrichment values with motif-associated hexamers in green, and $\boldsymbol{r_1}$ to $\boldsymbol{r_n}$ are its permutations, namely, the permutations of the *motif-associated hexamer* labels (or equivalently, permutations of the enrichment values).

P2: Shuffle nucleotides motif matrix

The second permutation approach introduces randomness by shuffling the components of the row vectors of the motif matrix. The idea is to avoid choosing k hexamers randomly, but to change the underlying motif matrix and derive a new set of hexamers in the same way the original motif-associated hexamers were derived. That way, the structure of interdependencies between the initial set of hexamers is assumed to be similar to the new set of hexamers, which were also derived from one (permuted) motif.

This process is illustrated with colored position probability matrices below.

$$\overbrace{\begin{array}{c|cccc|} & \text{A} & \text{C} & \text{G} & \text{U} \\ 1 & 0.4 & 0.3 & 0.1 & 0.2 \\ 2 & 0.1 & 0.1 & 0 & 0.7 \\ 3 & 0.1 & 0.4 & 0.1 & 0.3 \\ 4 & 0.4 & 0.4 & 0.1 & 0.1 \\ 5 & 0.1 & 0 & 0 & 0.9 \\ 6 & 0 & 0 & 0.1 & 0.9 \\ 7 & 0.3 & 0.1 & 0.1 & 0.5 \end{array}}^{\boldsymbol{x}} \overset{\text{shuffle}}{\rightsquigarrow} \overbrace{\begin{array}{c|cccc|} & \text{A} & \text{C} & \text{G} & \text{U} \\ 1 & 0.3 & 0.4 & 0.2 & 0.1 \\ 2 & 0.1 & 0.1 & 0 & 0.7 \\ 3 & 0.1 & 0.4 & 0.3 & 0.1 \\ 4 & 0.1 & 0.4 & 0.4 & 0.1 \\ 5 & 0.1 & 0.9 & 0 & 0 \\ 6 & 0.1 & 0 & 0 & 0.9 \\ 7 & 0.5 & 0.3 & 0.1 & 0.1 \end{array}}^{\boldsymbol{r_1}} \cdots \overbrace{\begin{array}{c|cccc|} & \text{A} & \text{C} & \text{G} & \text{U} \\ 1 & 0.2 & 0.4 & 0.3 & 0.1 \\ 2 & 0.7 & 0.1 & 0 & 0.1 \\ 3 & 0.4 & 0.4 & 0.3 & 0.1 \\ 4 & 0.1 & 0.4 & 0.4 & 0.1 \\ 5 & 0.9 & 0 & 0 & 0.1 \\ 6 & 0.1 & 0.9 & 0 & 0 \\ 7 & 0.5 & 0.1 & 0.1 & 0.3 \end{array}}^{\boldsymbol{r_n}} \tag{4.2}$$

P3: Permute foreground sets

The most rigorous and time-consuming approach to permute the data is to keep hexamers and motifs as they are and permute the foreground/background assignments of transcripts instead. Each permutation requires the recomputation of the hexamer enrichment values.

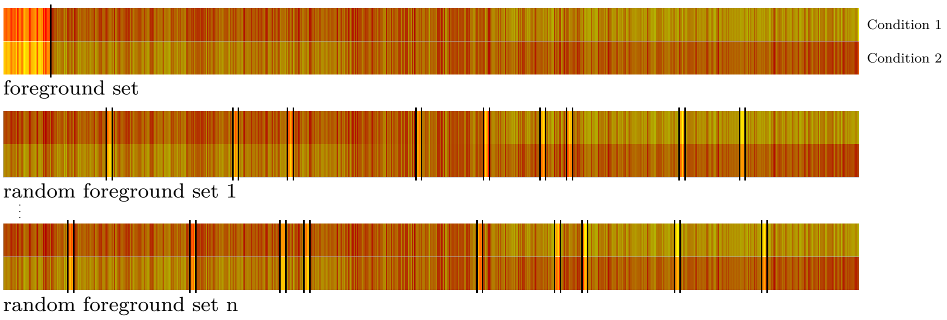


Figure 4.7: Permute foreground/background assigment: For each permutation, a new set of foreground transcripts (with equal cardinality) is selected. For example, if the actual foreground set consists of 103 transcripts with the highest fold change, one of 10,000 permutations is to choose 103 transcripts by random, use them as foreground set, and recompute hexamer enrichment values based on the new foreground/background assignment.

Figure 4.7 illustrates the process.

4.5.2 Comparison of permutation approaches

The p-value distributions for the three exemplary motifs and the three permutation approaches are displayed in figures 4.8 to 4.10. Two tests were applied to assess which of the three approaches generates uniformly distributed and overall insignificant p-values. The empirical p-value distributions were compared to a uniform distribution using the Kolmogorov-Smirnov test, where the p-value is the probability that a sample at least as extreme as the observed one was randomly drawn from a uniform distribution. And secondly, Fisher's combined p-value (see section 3.2.1) was calculated to obtain the overall significance of the empirical p-values. In case the permutation approach is valid, both tests yield insignificant p-values. Substantial deviations from the uniform distribution can also be seen by eye.

Permutation approach P1 tends to produce far too many highly significant p-values, especially when there are interdependencies between motif-associated hexamers (see panel D of figure 4.8). Permutation approach P2 may be appropriate for motifs with a balanced distribution of nucleotides in their associated hexamers, but performs poorly when used with specific U-rich motifs. Permutation approach 3 yields uniformly distributed and overall insignificant p-values (when used with random sequences), regardless of the structure of the motif.

In conclusion, permutation approach P3 outperforms the other two by far.

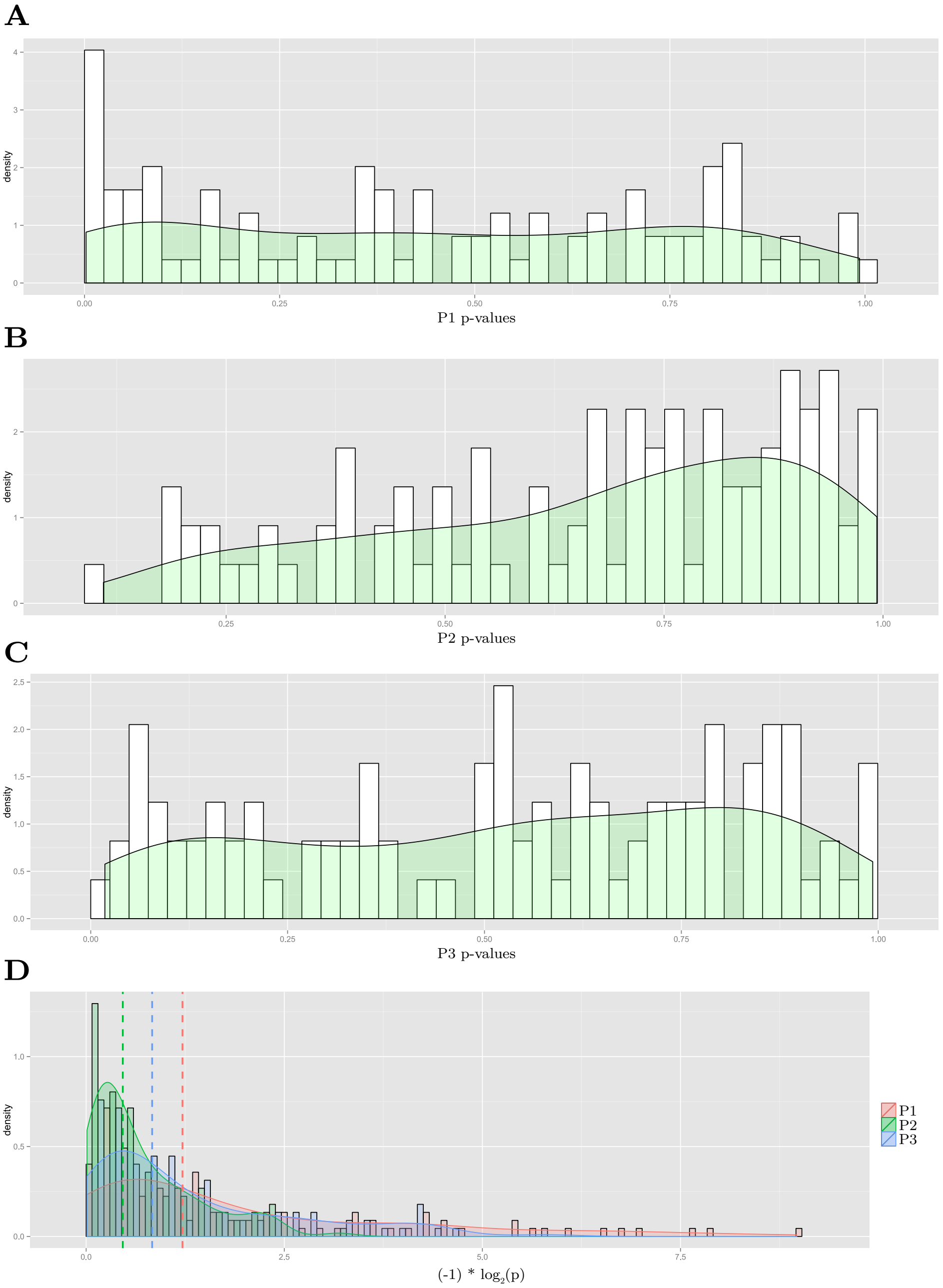


Figure 4.8: P-value distributions of motif `LC1`: Panels **A** to **C** are histograms of 100 p-values obtained by the aforementioned permutation approaches. (**A**) P-value distribution of P1, Fisher's combined p-value: 0.000344, Kolmogorov-Smirnov test with H_0 "p-values are uniformly distributed": $p = 0.041747$. (**B**) P2: Fisher's combined p-value: 1, KS test: $p < 0.000002$. (**C**) P3: Fisher's combined p-value: 0.815129, KS test: $p = 0.173931$. In panel **D** the log_2 transformed p-values of all three approaches are plotted, which facilitates the identification of very low p-values.

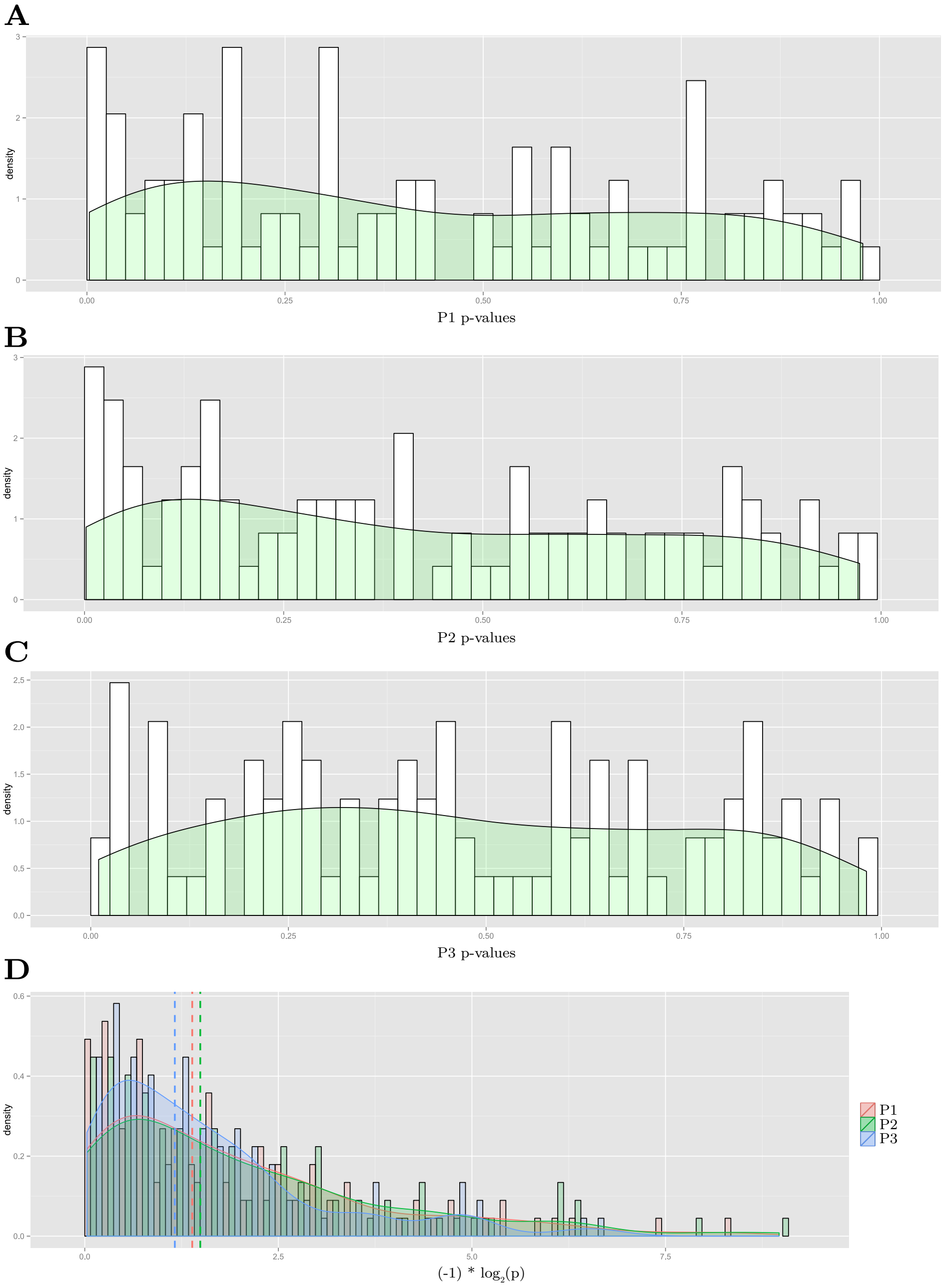


Figure 4.9: P-value distributions of motif M031_0.6: (**A**) Fisher's combined p-value: 0.001428, Kolmogorov-Smirnov test p-value: 0.023154. (**B**) Fisher's combined p-value: 0.000211, Kolmogorov-Smirnov test p-value: 0.019478. (**C**) Fisher's combined p-value: 0.215873, Kolmogorov-Smirnov test p-value: 0.443363.

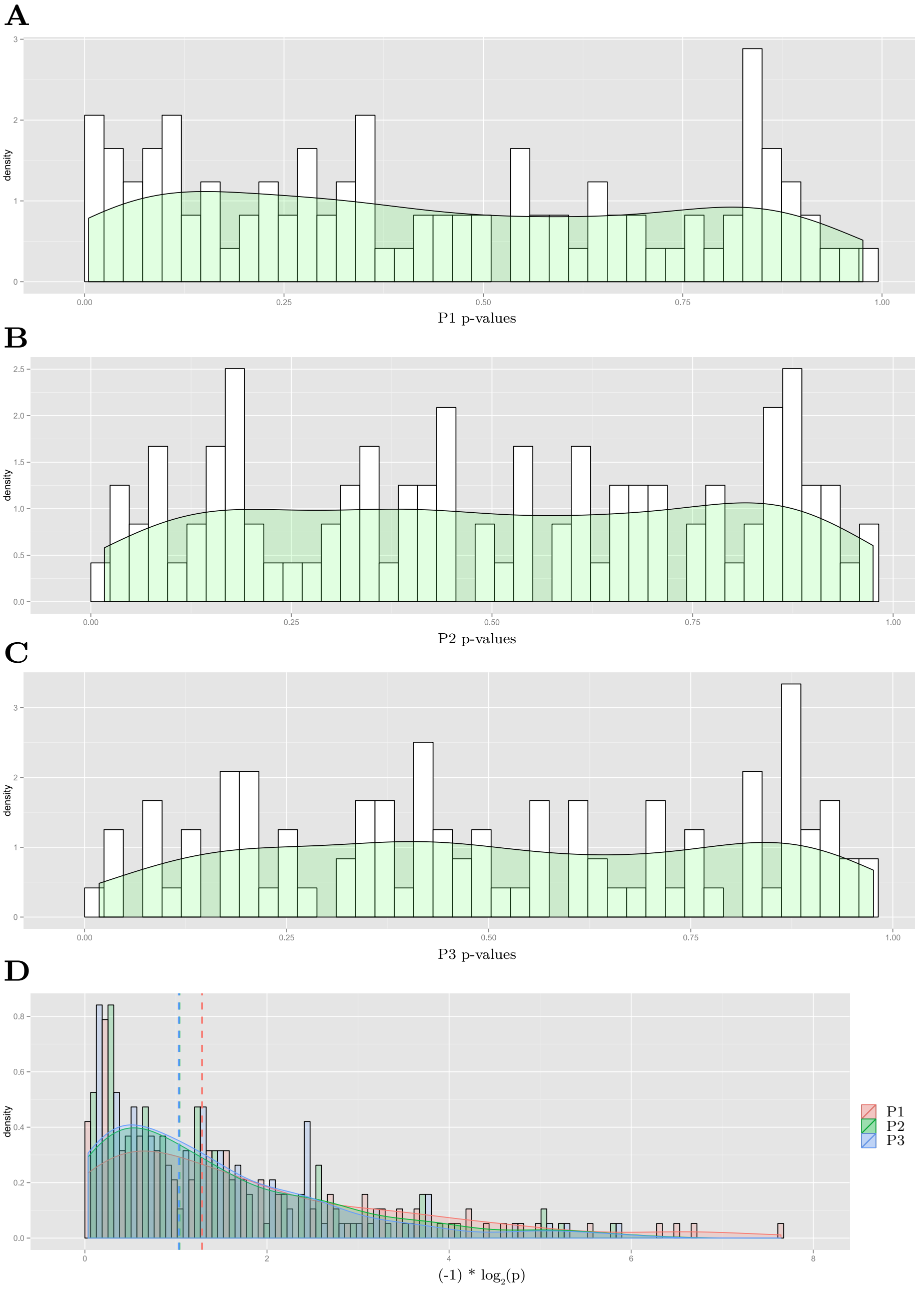


Figure 4.10: P-value distributions of motif `M152_0.6`: (**A**) Fisher's combined p-value: 0.01236, Kolmogorov-Smirnov test p-value: 0.133287. (**B**) Fisher's combined p-value: 0.567851, Kolmogorov-Smirnov test p-value: 0.992219. (**C**) Fisher's combined p-value: 0.786305, Kolmogorov-Smirnov test p-value: 0.70824.

4.6 RBPs and cisplatin

This section presents Transite analysis results of gene expression data from five different cell lines treated with the chemotherapeutic drug cis-diammine-dichloroplatinum(II), also known as cisplatin.

Cisplatin is a DNA-damaging agent that induces genotoxic stress. The alkylating agent predominantly causes the formation of intrastrand (but also interstrand) cross-links between two guanine residues [44] that in turn activate the DNA-damage response, which ultimately triggers apoptosis or cellular senescence [45].

4.6.1 Cell lines

A549 human non-small cell lung cancer cells

In GEO series `GSE6410`, A549 cells were treated with 50 µM of cisplatin for 1 hour. Control samples were treated with drug-free media for the same amount of time. After further 10 hours in drug-free media, gene expression changes were investigated using Affymetrix Human HG-Focus Target Array (GEO platform `GPL201`) [46].

SK-OV-3 human ovarian cancer cells

GEO series `GSE38545` investigates the transcriptional response of three human ovarian cancer cell lines to cisplatin. Dosage and duration of the cisplatin treatment were selected in a cell line specific manner to study the cisplatin-induced apoptotic death in each cell line. Illumina HumanRef-8 Expression BeadChips (GEO platform `GPL7192`) were used to investigate changes in expression.

NIH-OVCAR-3 human ovarian cancer cells

See previous section.

TOV-21G human ovarian cancer cells

See previous section.

U87 human primary glioblastoma cells

In GEO series `GSE66493`, U87 cells were treated with cisplatin for 24 hours at their IC_{50} concentration. The microarray platform Affymetrix Human Gene 1.0 ST Array (GEO platform `GPL6244`) was used.

4.6.2 Sample clustering

As a preliminary step, the samples of the three GEO series were clustered as described in section 2.1. The heatmaps of the sample distance clustering are shown in figure 4.11.

Based on Euclidean distance, the triplicates of cisplatin-treated samples of A549 cells cannot be separated from the triplicates of control samples of the same cells. This is a strong indication that the gene expression data from the `GSE6410` data series is too noisy to yield significant results.

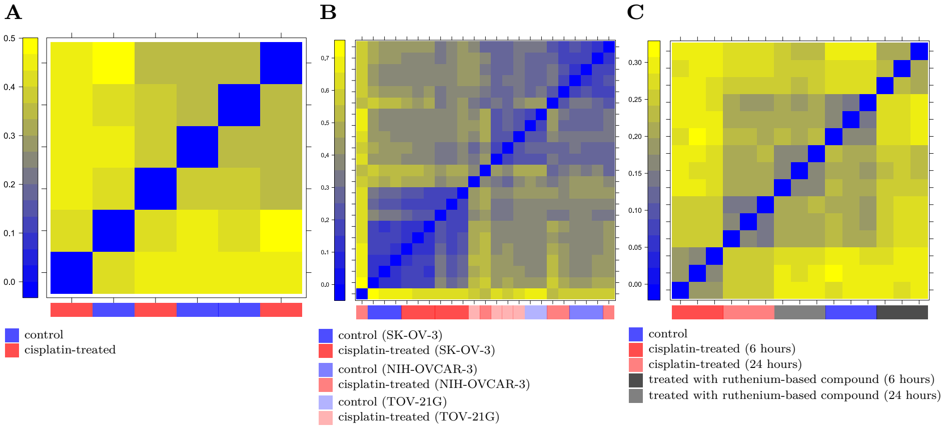


Figure 4.11: Sample distance heatmaps: (**A**) Treatment and control samples of GEO series `GSE6410` are scattered, which renders subsequent analysis useless. (**B**) Treatment and control samples of the various cell lines of GEO series `GSE38545` divide into *meaningful* (but noisy) groups. (**C**) In GEO series `GSE66493`, the sample labels are perfectly recapitulated in the sample distance clustering, where sample triplicates form definable 3 × 3 blocks in the distance heatmap.

4.6.3 Differential gene expression analysis

Differentially expressed genes were identified using the methodology described in section 2.3.

The overlap of differentially expressed genes between the cell lines is depicted in figure 4.12. A549 cells were discarded due to the weak discrimination between treatment and control samples, and—as a result—the low number of statistically significantly upregulated (six) and downregulated (two) genes.

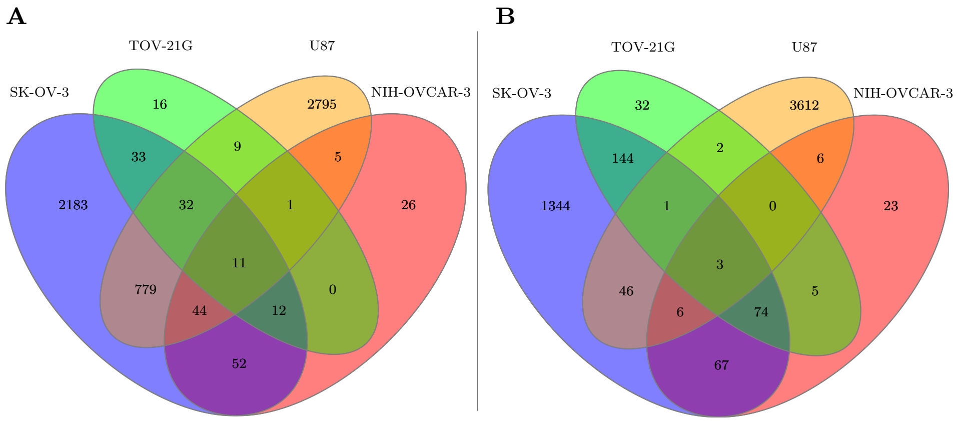


Figure 4.12: Overlap between sets of upregulated and downregulated genes of cisplatin-treated cell lines: (**A**) Statistically significantly upregulated genes (**B**) Statistically significantly downregulated genes

4.6.4 Transcript Set Motif Analysis

TSMA was performed on the 3' UTR sequences of upregulated and downregulated transcripts. The overlap between differentially expressed genes of the cell lines is shown in figure 4.13. Apart from U87 cells, there were no significantly over- or underrepresented RBP binding sites in upregulated transcripts (see panels A and B of figure 4.13). In downregulated transcripts, however, we found a consensus of 20 overrepresented RBP motifs across cell lines (panel C in the same figure, and tables 4.3 and 4.4), eight of which describe binding sites from RBPs of the ELAVL/Hu family. Because HuR/ELAVL1 is the only ubiquitously expressed member of that family— HuB/ELAVL2 and HuC/ELAVL3 are neuronal-specific— it can be assumed that ELAVL1 acts as the major regulator, which also recognizes and binds to ELAVL2 and ELAVL3 sites. [47, 48]. It has been shown previously that cisplatin treatment hinders ELAVL1 activity, causing its mRNA targets to decay [49]. This finding is recapitulated *in silico* by TSMA. Figure 4.14 shows overrepresented ELAVL binding sites in downregulated transcripts. Put differently, ELAVL mRNA targets are downregulated after cisplatin treatment, as expected. Similar to ELAVL1, HNRNPC is involved in 3'-UTR-mediated mRNA stabilization [50]. Moreover, it is found to be a regulator of homologous recombination based DNA repair [51], which is the only adequate mechanism to repair cisplatin-induced interstrand cross-links [52]. Binding sites for alternative splicing regulators SRSF1, SRSF4, and SRSF6 [53] are statistically significantly underrepresented in downregulated transcripts. Underrepresentation in downregulated transcripts is usually linked to overrepresentation in upregulated transcripts, albeit insignificant overrepresentation. The role of these splicing regulators in the context of cisplatin-induced DNA damage is unclear. The distribution of SRSF4/SRSF6-motif-associated hexamers in downregulated transcripts of different cell lines is shown in figure 4.15.

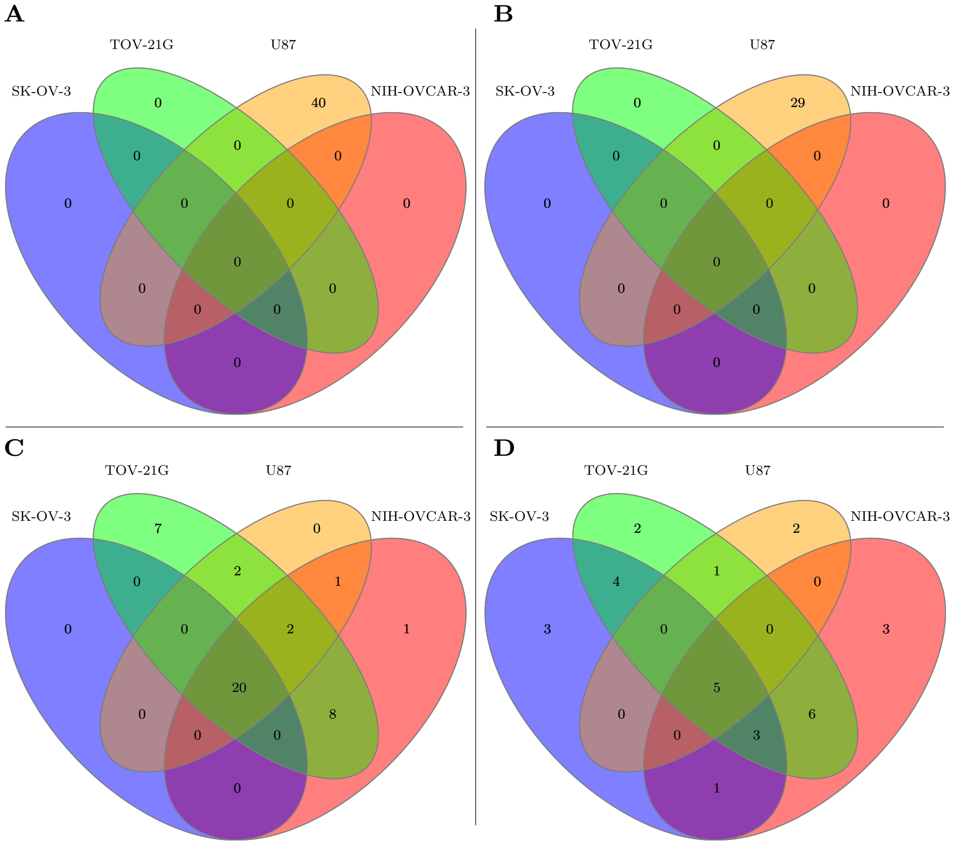


Figure 4.13: Overlap between sets of upregulated and downregulated genes of cisplatin-treated cell lines: (**A**) Statistically significantly overrepresented binding sites in upregulated transcripts (**B**) Statistically significantly underrepresented binding sites in upregulated transcripts (**C**) Statistically significantly overrepresented binding sites in downregulated transcripts (**D**) Statistically significantly underrepresented binding sites in downregulated transcripts

Table 4.3: Consensus RBPs with overrepresented binding sites in downregulated transcripts after cisplatin treatment

motif ID	RBPs
782_8497264	ELAVL2
784_7972035	ELAVL2
LC1	ELAVL1
M012_0.6	CPEB3, CPEB2
M025_0.6	HNRNPC
M075_0.6	TIA1
M077_0.6	U2AF2
M079_0.6	CELF3
M108_0.6	ELAVL1, ELAVL3
M112_0.6	ELAVL1, ELAVL3
M120_0.6	CPEB3
M124_0.6	ELAVL3
M127_0.6	ELAVL1, ELAVL3
M149_0.6	CPEB3, CPEB4
M150_0.6	RALY
M156_0.6	TIA1
M158_0.6	HNRNPCL1
M227_0.6	PTBP1, PTBP2, ROD1
M228_0.6	PTBP1, PTBP2, ROD1
M232_0.6	ELAVL1, ELAVL3

Table 4.4: Consensus RBPs with underrepresented binding sites in downregulated transcripts after cisplatin treatment

motif ID	RBPs
M061_0.6	SAMD4A, SAMD4B
M151_0.6	HNRNPH2, HNRNPH1, HNRNPF
M153_0.6	LIN28A, LIN28B
M154_0.6	SRSF1
M334_0.6	SRSF4, SRSF6

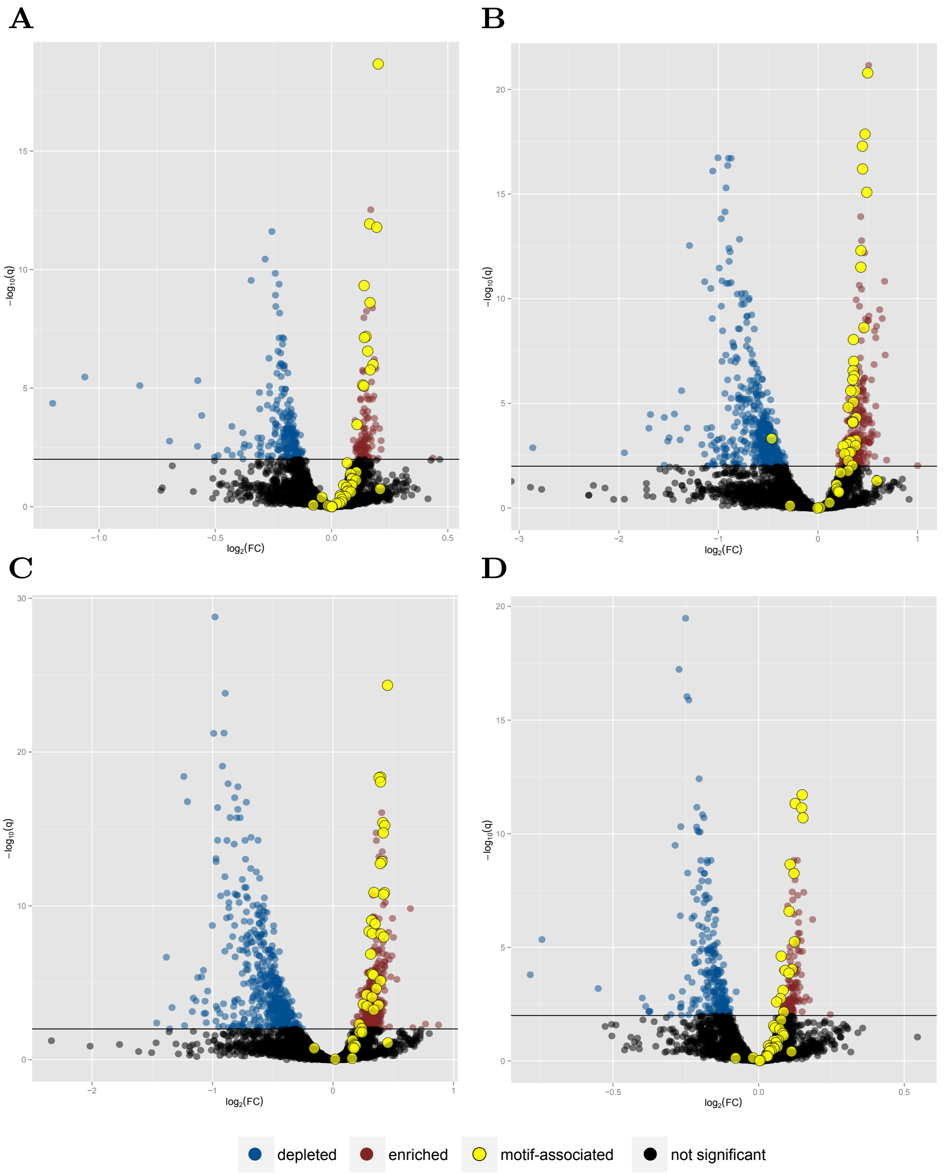


Figure 4.14: ELAVL2 motif (motif id: 782_8497264) in downregulated transcripts: Motif-associated hexamers are predominantly found among enriched hexamers. (**A**) SK-OV-3 cells (**B**) NIH-OVCAR-3 cells (**C**) TOV-21G cells (**D**) U87 cells

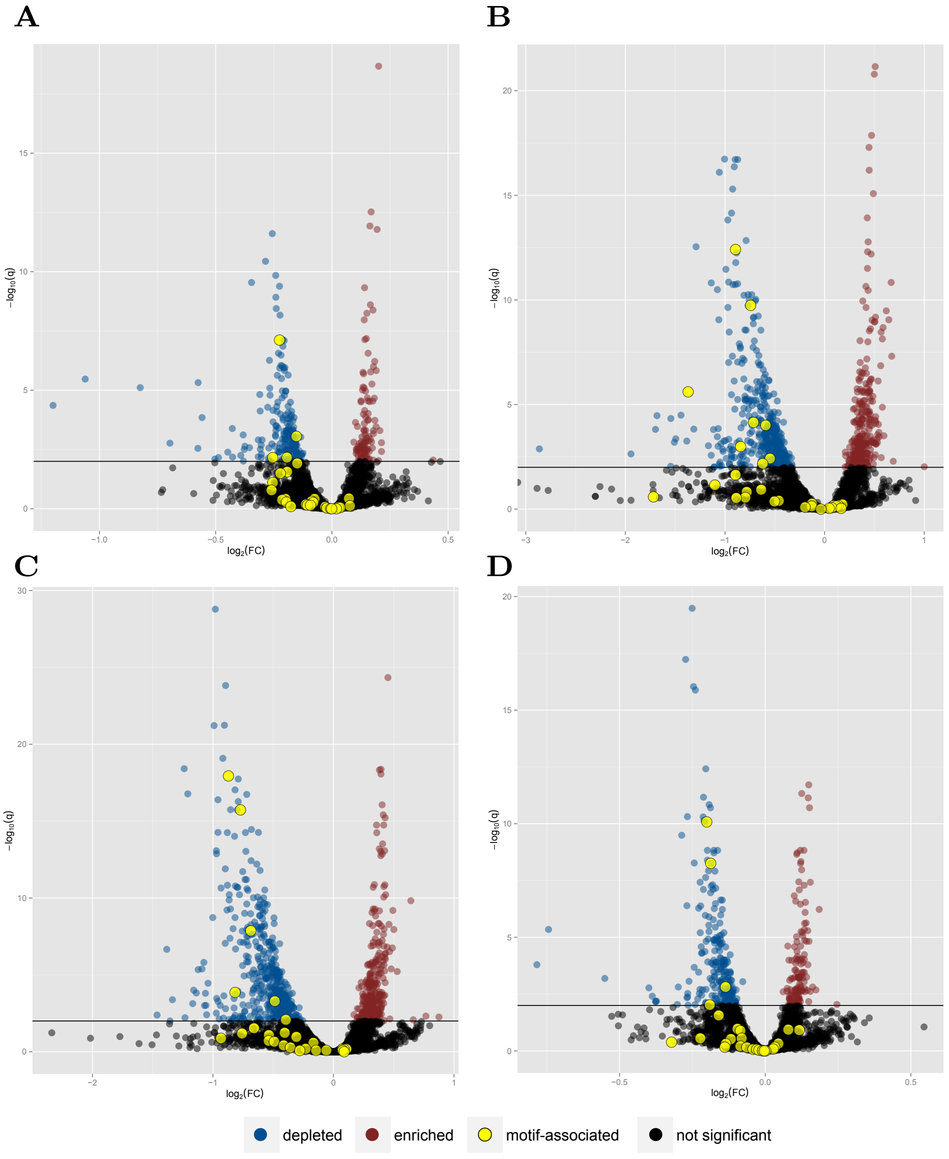


Figure 4.15: SRSF4/SRSF6 motif (motif id: M334_0.6) in downregulated transcripts: Motif-associated hexamers are predominantly found among depleted hexamers. (**A**) SK-OV-3 cells (**B**) NIH-OVCAR-3 cells (**C**) TOV-21G cells (**D**) U87 cells

4.6.5 Spectrum Motif Analysis

There was no overlap between the four cell lines, especially U87 cells showed a different behavior with only one RBP motif meeting the requirements to be labeled *non-random* (see section 4.2). However, there was a substantial overlap between NIH-OVCAR-3 and TOV-21G cells, 36 non-random spectrum plots that were common to both cell lines, 15 specific to NIH-OVCAR-3 cells, and 14 specific to TOV-21G cells. Representative examples of spectrum plots are shown for RBPs exhibiting decreasing linear relationship between differential regulation after cisplatin and abundance of binding sites across all transcripts (see figure 4.16), and an increasing linear relationship (see figure 4.17).

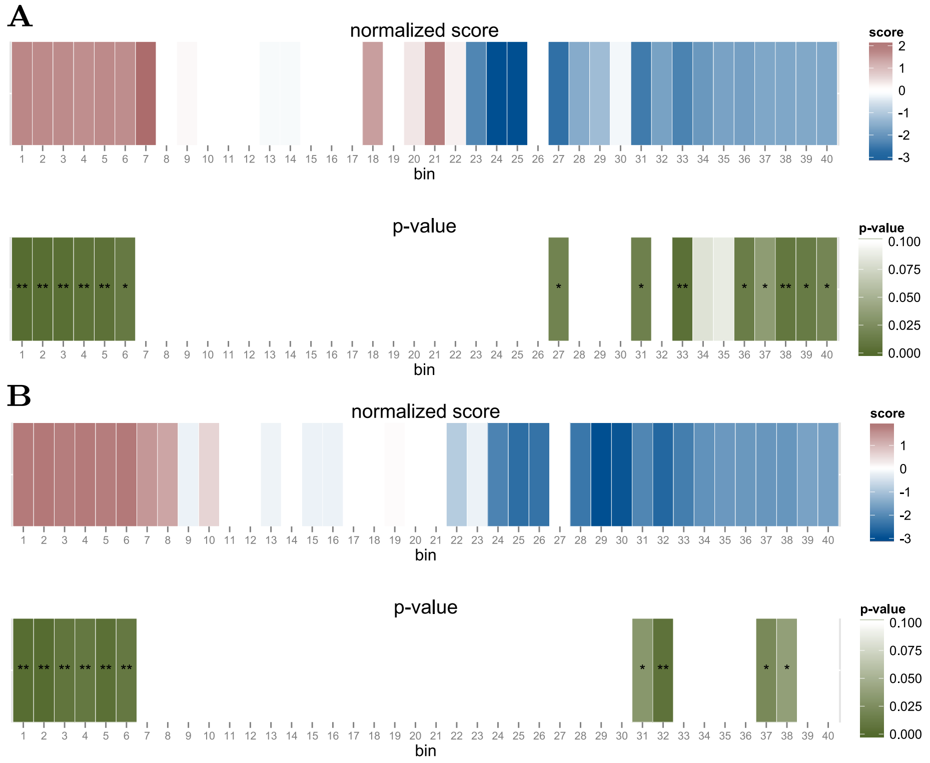


Figure 4.16: Spectrum plots of ELAVL2 motif (motif id: `782_8497264`): Transcripts were sorted according to their fold change (most downregulated after cisplatin treatment on the left, most upregulated on the right), they exhibit a gradient of ELAVL2 binding sites. (**A**) NIH-OVCAR-3 cells (**B**) TOV-21G cells

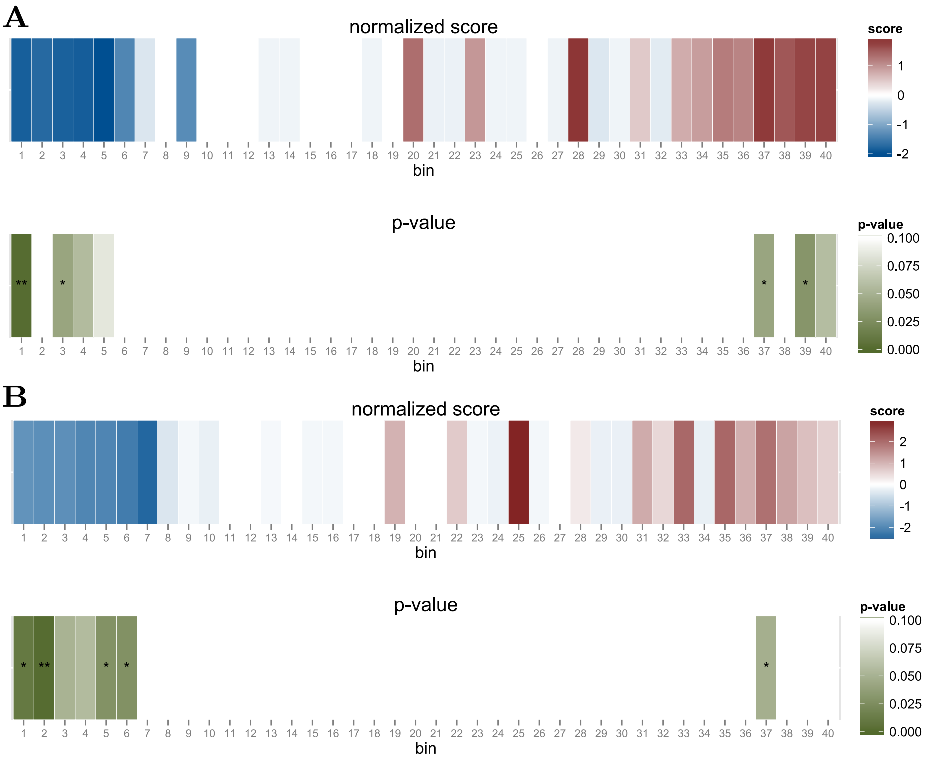


Figure 4.17: Spectrum plots of SRSF4/SRSF6 motif: (**A**) NIH-OVCAR-3 cells (**B**) TOV-21G cells

4.7 R package Transite

Transite was developed in R 3.2. The package development process was streamlined by `devtools` [54].

Function documentation files were generated with the documentation system `roxygen`. Vignettes were created with `rmarkdown` and `knitr`.

Functions from the `parallel` R standard library package were used to parallelize tasks like *k*-mer enrichment calculation and *k*-mer-based and transcript-based score calculation.

Computationally expensive algorithms or algorithms that cannot be vectorized were implemented in C++. These include the transcript-based scoring algorithm, the local consistency score Monte Carlo test, and the *k*-mer-based scoring algorithm. The C++ code integration was facilitated by the R package `Rcpp` [55].

In order to further decrease run-time, hash tables were used to cache (1) motif scores of all hexamers, and (2) transcript-based hit counts for sequence regions (5' UTRs, intronic regions, 3' UTRs) of all human and murine RefSeq identifiers.

4.7.1 Package dependencies

Three CRAN packages are listed in the *imports* section of the package description file:

dplyr 0.4.2	A Grammar of Data Manipulation
ggplot2 1.0.1	An Implementation of the Grammar of Graphics
Rcpp 0.12.0	Seamless R and C++ Integration

And one package from Bioconductor is imported:

Biostrings 2.36.4	String objects representing biological sequences, and matching algorithms

4.8 Transite website

The aim of the Transite website is to make the functionality of the Transite R/Bioconductor package available to a broader circle of scientists, including people outside the R community.

It will be available at `http://transite.mit.edu`. The software will be hosted on servers provided by the David H. Koch Institute for Integrative Cancer Research at Massachusetts Institute of Technology.

The website was developed in R with the reactive web application framework `shiny` [56] from RStudio. The components of the graphical user interface were provided by `shiny` and `shinyBS`, which serves as an R wrapper for the Twitter Bootstrap HTML/CSS/JavaScript components.

4.8.1 Requirements

Apart from Transite and the R standard library package `grid`, the following R packages from the CRAN package repository are required for the Transite website:

`shiny` 0.12.2	Web Application Framework for R
`shinyBS` 0.61	Twitter Bootstrap Components for Shiny
`shinyjs` 0.1.0	Perform Common JavaScript Operations in Shiny Apps using Plain R Code
`knitr` 1.11	A General-Purpose Package for Dynamic Report Generation in R
`DT` 0.1	A Wrapper of the JavaScript Library 'DataTables'
`rmarkdown` 0.8	Dynamic Documents for R (requires pandoc)
`gridExtra` 2.0.0	Miscellaneous Functions for "Grid" Graphics
`dplyr` 0.4.2	A Grammar of Data Manipulation
`ggplot2` 1.0.1	An Implementation of the Grammar of Graphics
`scales` 0.3.0	Scale Functions for Visualization
`stringr` 1.0.0	Simple, Consistent Wrappers for Common String Operations
`mailR` 0.4.1	A Utility to Send Emails from R

The package `rmarkdown` relies on the stand-alone document converter pandoc to convert markdown to HTML documents.

These R/Bioconductor packages are used to retrieve and handle sequence data and identifiers of platform transcripts:

`TxDb.Hsapiens.UCSC.hg38.knownGene` 3.1.2
Annotation package for TxDb object(s)

`TxDb.Mmusculus.UCSC.mm10.knownGene` 3.1.2
Annotation package for TxDb object(s)

`BSgenome.Hsapiens.NCBI.GRCh38` 1.3.1000
Full genome sequences for Homo sapiens (GRCh38)

`BSgenome.Mmusculus.UCSC.mm10` 1.4.0
Full genome sequences for Mus musculus (UCSC version mm10)

`org.Hs.eg.db` 3.1.2	Genome wide annotation for Human
`org.Mm.eg.db` 3.1.2	Genome wide annotation for Mouse
`BSgenome` 1.36.3	Infrastructure for Biostrings-based genome data packages
`AnnotationDbi` 1.30.1	Annotation Database Interface
`GenomicFeatures` 1.20.3	Tools for making and manipulating transcript centric annotations
`Biostrings` 2.36.4	String objects representing biological sequences, and matching algorithms

Table 4.5: Human microarray platform backgrounds for TSMA

Platform	**GEO accession**
Affymetrix Human Genome U133 Plus 2.0	GPL570
Affymetrix Human Genome U133A	GPL96
Affymetrix HT Human Genome U133A	GPL3921
Affymetrix HT Human Genome U133B	GPL9197
Affymetrix Human 35K SubC	GPL98
Affymetrix Human 35K SubB	GPL99
Affymetrix Human 35K SubC	GPL100
Affymetrix Human 35K SubD	GPL101
Affymetrix Human Genome U133A	2.0 GPL571
Affymetrix Human Genome U133B	GPL97
Affymetrix Human Genome U95A	GPL91
Affymetrix Human Genome U95B	GPL92
Affymetrix Human Genome U95C	GPL93
Affymetrix Human Genome U95D	GPL94
Affymetrix Human Genome U95E	GPL95
Affymetrix Human X3P	GPL1352
Illumina HumanHT-12 v3.0 expression beadchip	GPL18461
Illumina HumanRef-8 WG-DASL v3.0	GPL8432
Illumina HumanWG-6 v2.0 expression beadchip	GPL13376

4.8.2 Features

The three major features, Transcript Set Motif Analysis, Spectrum Motif Analysis and Single Transcript Motif Analysis are accessible via the horizontal main menu. The feature-specific pages provide forms to configure and customize Transite runs (see figures 4.18 and 4.19). After the user filled in the form and submitted the analysis run, the web application writes the Transite job description to an `rmarkdown` file, which includes all the necessary commands and settings for the Transite run. As soon as enough resources are available, a separate process executes the job. Upon completion, the Transite analysis report is sent to the e-mail address provided by the submitter. The report consists of an HTML document with figures, created with `pandoc`, `rmarkdown`, `knitr`, and `DT`, and supplemental text-based, tab-delimited data tables of intermediate and final results.

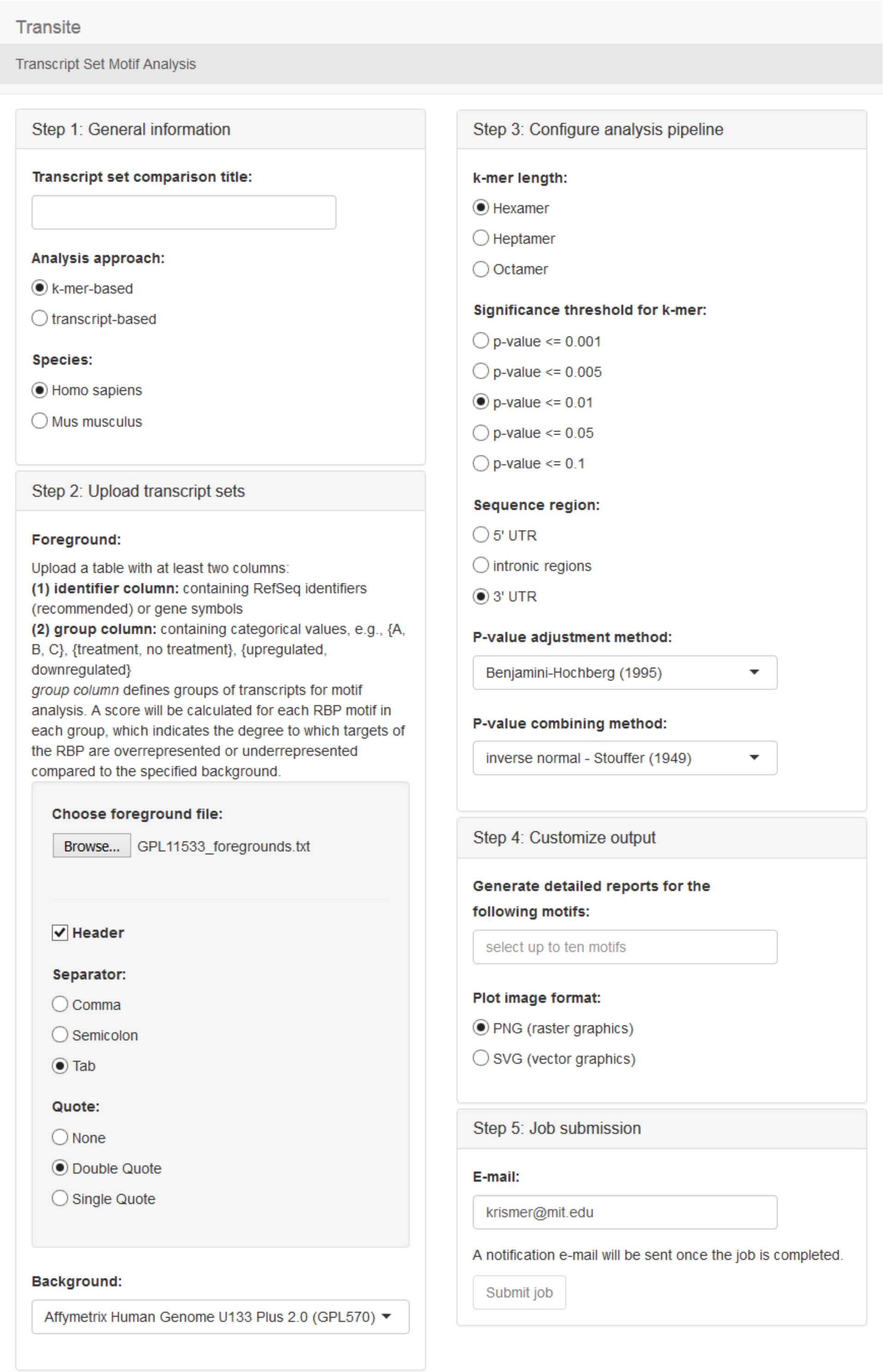
For TSMA, the user is asked to upload a text file with two columns: an identifier column with either RefSeq identifiers or gene symbols (HGNC symbols for human transcripts, MGI symbols for murine transcripts), and a group label column to identify the transcript groups, i.e., nominal labels like *upregulated*, *downregulated*. Furthermore, it is necessary to select a matching background gene set. Several human and murine microarray backgrounds are predefined (see tables 4.5 and 4.6). If the correct background is not part of the list, a custom background file can be uploaded.

SPMA requires a text file with an identifier column and a *value* column

Table 4.6: Murine microarray platform backgrounds for TSMA

Platform	**GEO accession**
Affymetrix Mouse Genome 430 2.0	GPL1261
Affymetrix Murine 19K SubA	GPL77
Affymetrix Murine 19K SubB	GPL78
Affymetrix Murine 19K SubC	GPL79
Affymetrix HT Mouse Genome MG-430A	GPL8759
Affymetrix HT Mouse Genome MG-430B	GPL8760
Affymetrix Mouse Genome 430A 2.0	GPL8321
Affymetrix Mouse Expression 430A	GPL339
Affymetrix Mouse Expression 430B	GPL340
Affymetrix Murine Genome U74A Version 2	GPL81
Affymetrix Murine Genome U74B Version 2	GPL82
Affymetrix Murine Genome U74C Version 2	GPL83
Affymetrix Murine 11K SubA	GPL75
Affymetrix Murine 11K SubB	GPL76
Affymetrix Murine Genome U74A	GPL32
Affymetrix Murine Genome U74B	GPL33
Affymetrix Murine Genome U74C	GPL34
Illumina MouseWG-6 v2.0 expression beadchip	GPL6887
Illumina MouseRef-8 v2.0 expression beadchip	GPL6885

with an ordinal attribute like fold change, which will be used for sorting the transcripts. The definition of the background gene set is not necessary, since they are inherently defined by the main text file.



Transite

Transcript Set Motif Analysis

Step 1: General information

Transcript set comparison title:

Analysis approach:

● k-mer-based

○ transcript-based

Species:

● Homo sapiens

○ Mus musculus

Step 2: Upload transcript sets

Foreground:

Upload a table with at least two columns:
(1) identifier column: containing RefSeq identifiers (recommended) or gene symbols
(2) group column: containing categorical values, e.g., {A, B, C}, {treatment, no treatment}, {upregulated, downregulated}
group column defines groups of transcripts for motif analysis. A score will be calculated for each RBP motif in each group, which indicates the degree to which targets of the RBP are overrepresented or underrepresented compared to the specified background.

Choose foreground file:

Browse... GPL11533_foregrounds.txt

☑ **Header**

Separator:

○ Comma

○ Semicolon

● Tab

Quote:

○ None

● Double Quote

○ Single Quote

Background:

Affymetrix Human Genome U133 Plus 2.0 (GPL570)

Step 3: Configure analysis pipeline

k-mer length:

● Hexamer

○ Heptamer

○ Octamer

Significance threshold for k-mer:

○ p-value <= 0.001

○ p-value <= 0.005

● p-value <= 0.01

○ p-value <= 0.05

○ p-value <= 0.1

Sequence region:

○ 5' UTR

○ intronic regions

● 3' UTR

P-value adjustment method:

Benjamini-Hochberg (1995)

P-value combining method:

inverse normal - Stouffer (1949)

Step 4: Customize output

Generate detailed reports for the following motifs:

select up to ten motifs

Plot image format:

● PNG (raster graphics)

○ SVG (vector graphics)

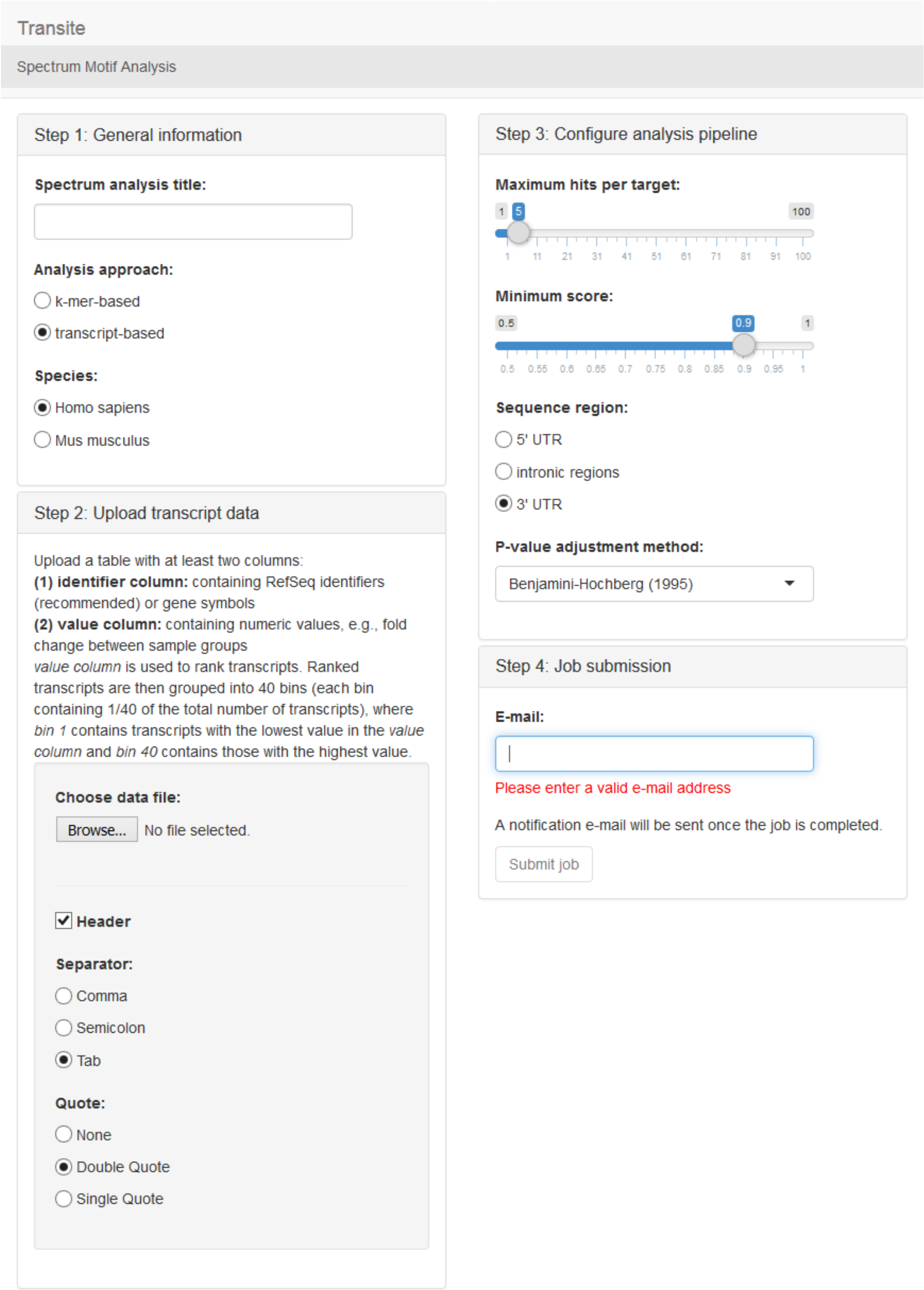
Step 5: Job submission

E-mail:

krismer@mit.edu

A notification e-mail will be sent once the job is completed.

Figure 4.18: *TSMA* form with *k*-mer-based analysis approach.



Transite

Spectrum Motif Analysis

Step 1: General information

Spectrum analysis title:

Analysis approach:

○ k-mer-based

● transcript-based

Species:

● Homo sapiens

○ Mus musculus

Step 2: Upload transcript data

Upload a table with at least two columns:
(1) identifier column: containing RefSeq identifiers (recommended) or gene symbols
(2) value column: containing numeric values, e.g., fold change between sample groups
value column is used to rank transcripts. Ranked transcripts are then grouped into 40 bins (each bin containing 1/40 of the total number of transcripts), where *bin 1* contains transcripts with the lowest value in the *value column* and *bin 40* contains those with the highest value.

Choose data file:

Browse... No file selected.

☑ **Header**

Separator:

○ Comma

○ Semicolon

● Tab

Quote:

○ None

● Double Quote

○ Single Quote

Step 3: Configure analysis pipeline

Maximum hits per target: 5

Minimum score: 0.9

Sequence region:

○ 5' UTR

○ intronic regions

● 3' UTR

P-value adjustment method:

Benjamini-Hochberg (1995)

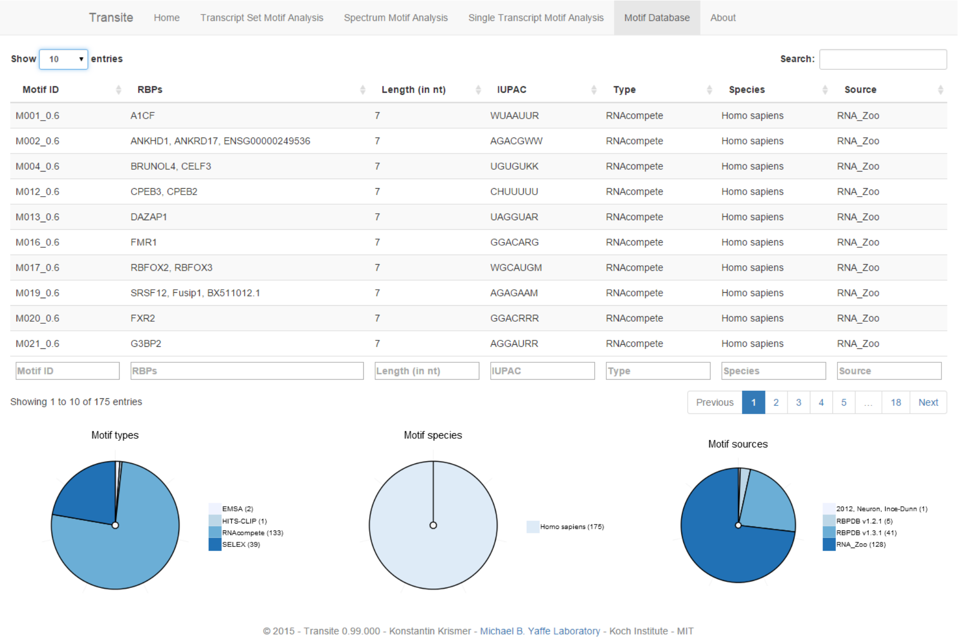
Step 4: Job submission

E-mail:

Please enter a valid e-mail address

A notification e-mail will be sent once the job is completed.

Figure 4.19: *SPMA* form with transcript-based analysis approach.



Motif ID	RBPs	Length (in nt)	IUPAC	Type	Species	Source
M001_0.6	A1CF	7	WUAAUUR	RNAcompete	Homo sapiens	RNA_Zoo
M002_0.6	ANKHD1, ANKRD17, ENSG00000249536	7	AGACGWW	RNAcompete	Homo sapiens	RNA_Zoo
M004_0.6	BRUNOL4, CELF3	7	UGUGUKK	RNAcompete	Homo sapiens	RNA_Zoo
M012_0.6	CPEB3, CPEB2	7	CHUUUUU	RNAcompete	Homo sapiens	RNA_Zoo
M013_0.6	DAZAP1	7	UAGGUAR	RNAcompete	Homo sapiens	RNA_Zoo
M016_0.6	FMR1	7	GGACARG	RNAcompete	Homo sapiens	RNA_Zoo
M017_0.6	RBFOX2, RBFOX3	7	WGCAUGM	RNAcompete	Homo sapiens	RNA_Zoo
M019_0.6	SRSF12, Fusip1, BX511012.1	7	AGAGAAM	RNAcompete	Homo sapiens	RNA_Zoo
M020_0.6	FXR2	7	GGACRRR	RNAcompete	Homo sapiens	RNA_Zoo
M021_0.6	G3BP2	7	AGGAURR	RNAcompete	Homo sapiens	RNA_Zoo
Motif ID	RBPs	Length (in nt)	IUPAC	Type	Species	Source

Showing 1 to 10 of 175 entries

© 2015 - Transite 0.99.000 - Konstantin Krismer - Michael B. Yaffe Laboratory - Koch Institute - MIT

Figure 4.20: ***Motif database*** **webpage:** Displays all motifs in the Transite database in searchable and sortable table.

Chapter 5

Discussion

Transite is a computational method for the analysis of RBP-mediated mRNA stability changes in various cellular processes. Hypotheses are generated regarding the role of RBPs in these changes by using existing knowledge of RBP binding preferences in combination with the vast body of publicly available gene expression data. Subsequent independent experimental validation of these hypotheses are required. Instead of relying on the analysis of a single data set from one cell line, it is advisable to include as many cell lines and independent data sets as possible. The consensus results are less prone to idiosyncratic behavior pertaining to a specific cell line - treatment combination, thus reducing the risk of false positives.

Transite results depend on the published sequence motifs that describe the binding preferences of RBPs. Some RBPs have not been described yet, others might be described incorrectly. Another limitation to the approach described in this thesis is the possibility that the RBP-induced change of mRNA abundance is overshadowed by a simultaneous change of its transcription rate.

Popular alternatives to the frequently used Fisher's conditional exact test are unconditional exact tests by Barnard [57] and Boschloo [58], which are uniformly more powerful for 2 × 2 contingency tables [59, 60]. Unlike conditional tests, which assume both row and column margins of the contingency table to be known in advance, unconditional tests assume either the row or the column margin, or—using a multinomial model—only the total sample size. Both Barnard's and Boschloo's tests are computationally demanding compared to Fisher's exact test [61], which is why the latter is used in Transite.

With a few minor adjustments the Transite pipeline can also be used to investigate the distribution of complementary sequences of microRNA seed regions in transcripts. This topic was intentionally left out, since it would go beyond the scope of this thesis.

Chapter 6

Conclusion and Outlook

The aim of this study is to develop a computational method for identifying RBPs as key post-transcriptional players in the concerted regulation and function of cellular processes. Based on previous work [17] in the Michael B. Yaffe laboratory, the analytical pipeline Transite has been devised to provide insights into the regulatory role of RBPs in various cellular processes by leveraging gene expression data and current knowledge of RBP binding preferences. A comprehensive analysis of RBPs in the context of cisplatin treatment has been carried out using gene expression data from five different cancer cell lines. Transite will be available as an R package to enable a seamless integration in preexisting workflows.

The Transite website will be deployed on http://transite.mit.edu. After the successful completion of unit and load tests, and a beta testing phase with members of the Michael B. Yaffe laboratory, the website will be available to the general public. Furthermore, the Transite R package will be submitted to the R bioinformatics package repository Bioconductor [62].

References

[1] M. B. Yaffe et al. “A motif-based profile scanning approach for genome-wide prediction of signaling pathways”. In: *Nat. Biotechnol.* 19.4 (Apr. 2001), pp. 348–353 (cit. on p. 1).

[2] B. D. Berkovits and C. Mayr. “Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization”. In: *Nature* 522.7556 (June 2015), pp. 363–367 (cit. on p. 1).

[3] E. T. Wang et al. “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221 (Nov. 2008), pp. 470–476 (cit. on p. 1).

[4] M. Dutertre et al. “DNA damage: RNA-binding proteins protect from near and far”. In: *Trends Biochem. Sci.* 39.3 (Mar. 2014), pp. 141–149 (cit. on pp. 1, 2).

[5] D. Ray et al. “A compendium of RNA-binding motifs for decoding gene regulation”. In: *Nature* 499.7457 (July 2013), pp. 172–177 (cit. on p. 1).

[6] H. Tran, F. Maurer, and Y. Nagamine. “Stabilization of urokinase and urokinase receptor mRNAs by HuR is linked to its cytoplasmic accumulation induced by activated mitogen-activated protein kinase-activated protein kinase 2”. In: *Mol. Cell. Biol.* 23.20 (Oct. 2003), pp. 7177–7188 (cit. on p. 2).

[7] J. Guhaniyogi and G. Brewer. “Regulation of mRNA stability in mammalian cells”. In: *Gene* 265.1-2 (Mar. 2001), pp. 11–23 (cit. on p. 2).

[8] A. Hasan et al. “Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability”. In: *PLoS Genet.* 10.11 (Nov. 2014), e1004684 (cit. on p. 2).

[9] Sean Davis and Paul Meltzer. “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”. In: *Bioinformatics* 14 (2007), pp. 1846–1847 (cit. on p. 7).

[10] T. Barrett et al. “NCBI GEO: archive for functional genomics data sets–update”. In: *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D991–995 (cit. on p. 7).

[11] John W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977 (cit. on p. 7).

[12] Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. “arrayQualityMetrics–a bioconductor package for quality assessment of microarray data”. In: *Bioinformatics* 25.3 (2009), pp. 415–6 (cit. on p. 7).

[13] M. E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Res.* (Jan. 2015) (cit. on p. 9).

[14] G. K. Smyth. “Linear models and empirical bayes methods for assessing differential expression in microarray experiments”. In: *Stat Appl Genet Mol Biol* 3 (2004), Article3 (cit. on p. 10).

[15] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300 (cit. on p. 10).

[16] J. A. Blake et al. “Gene Ontology Consortium: going forward”. In: *Nucleic Acids Res.* 43.Database issue (Jan. 2015), pp. D1049–1056 (cit. on p. 10).

[17] Anna Gattinger. “Motif-Based Analysis of Post-Transcriptional Control of the DNA Damage Response”. MA thesis. University of Applied Sciences Upper Austria, 2014 (cit. on pp. 14, 69).

[18] V. Gotea et al. “Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers”. In: *Genome Res.* 20.5 (May 2010), pp. 565–577 (cit. on p. 17).

[19] J Chambers, T Hastie, and D Pregibon. “Statistical Models in S”. In: Springer. 1990, pp. 317–321 (cit. on p. 20).

[20] K. S. Pollard et al. “Detection of nonneutral substitution rates on mammalian phylogenies”. In: *Genome Res.* 20.1 (Jan. 2010), pp. 110–121 (cit. on p. 26).

[21] A. Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome Res.* 15.8 (Aug. 2005), pp. 1034–1050 (cit. on p. 26).

[22] M. T. Weirauch et al. “Determination and inference of eukaryotic transcription factor sequence specificity”. In: *Cell* 158.6 (Sept. 2014), pp. 1431–1443 (cit. on p. 27).

[23] K. B. Cook et al. “RBPDB: a database of RNA-binding specificities”. In: *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D301–308 (cit. on p. 27).

[24] C. Tuerk and L. Gold. “Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase”. In: *Science* 249.4968 (Aug. 1990), pp. 505–510 (cit. on p. 27).

[25] D. Ray et al. “Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins”. In: *Nat. Biotechnol.* 27.7 (July 2009), pp. 667–670 (cit. on p. 27).

[26] M. M. Garner and A. Revzin. “A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system”. In: *Nucleic Acids Res.* 9.13 (July 1981), pp. 3047–3060 (cit. on p. 27).

[27] R.A. Fisher. *Statistical methods for research workers.* Edinburgh Oliver & Boyd, 1932 (cit. on p. 28).

[28] S. Stouffer, L. DeVinney, and E. Suchmen. *The American soldier: Adjustment during army life.* Princeton University Press, 1949 (cit. on p. 28).

[29] T. Lipták. “On the combination of independent tests”. In: *Magyar Tudományos Akadémia Matematikai Kutatóintézet Közleményei* (1958), pp. 171–196 (cit. on p. 28).

[30] G. S. Mudholkar and E. O. George. “The logit method for combining probabilities”. In: *Symposium on optimizing methods in statistics.* 1979, pp. 345–366 (cit. on p. 29).

[31] Eugene S. Edgington. “An Additive Method for Combining Probability Values from Independent Experiments”. In: *The Journal of Psychology* 80.2 (1972), pp. 351–363 (cit. on p. 29).

[32] L. Tippett. *The methods of statistics.* Williams and Norgate, 1931 (cit. on p. 29).

[33] S. Holm. “A simple sequentially rejective multiple test procedure”. In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70 (cit. on p. 30).

[34] Yosef Hochberg. “A sharper Bonferroni procedure for multiple tests of significance”. In: *Biometrika* 75.4 (1988), pp. 800–802 (cit. on p. 30).

[35] Jean Dunn and Olive Jean Dunn. “Multiple Comparisons Among Means”. In: *American Statistical Association* (1961), pp. 52–64 (cit. on p. 30).

[36] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300 (cit. on p. 31).

[37] Yoav Benjamini and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency”. In: *Ann. Statist.* 29.4 (Aug. 2001), pp. 1165–1188 (cit. on p. 31).

[38] B. W. Matthews. “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. In: *Biochim. Biophys. Acta* 405.2 (Oct. 1975), pp. 442–451 (cit. on p. 32).

[39] K. Nishida, M. C. Frith, and K. Nakai. “Pseudocounts for transcription factor binding sites”. In: *Nucleic Acids Res.* 37.3 (Feb. 2009), pp. 939–944 (cit. on p. 33).

[40] B. Phipson and G. K. Smyth. “Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn”. In: *Stat Appl Genet Mol Biol* 9 (2010), Article39 (cit. on p. 34).

[41] C. J. Clopper and E. S. Pearson. “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial”. In: *Biometrika* 26.4 (1934), pp. 404–413 (cit. on p. 35).

[42] Thomas M. Loughin. “A systematic comparison of methods for combining p-values from independent tests”. In: *Computational Statistics & Data Analysis* 47.3 (2004), pp. 467–485 (cit. on p. 37).

[43] N. Mukherjee et al. “Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability”. In: *Mol. Cell* 43.3 (Aug. 2011), pp. 327–339 (cit. on p. 44).

[44] N. Poklar et al. “Influence of cisplatin intrastrand crosslinking on the conformation, thermal stability, and energetics of a 20-mer DNA duplex”. In: *Proc. Natl. Acad. Sci. U.S.A.* 93.15 (July 1996), pp. 7606–7611 (cit. on p. 52).

[45] S. P. Jackson and J. Bartek. “The DNA-damage response in human biology and disease”. In: *Nature* 461.7267 (Oct. 2009), pp. 1071–1078 (cit. on p. 52).

[46] G. M. Almeida et al. “Multiple end-point analysis reveals cisplatin damage tolerance to be a chemoresistance mechanism in a NSCLC model: implications for predictive testing”. In: *Int. J. Cancer* 122.8 (Apr. 2008), pp. 1810–1819 (cit. on p. 52).

[47] C. M. Brennan and J. A. Steitz. “HuR and mRNA stability”. In: *Cell. Mol. Life Sci.* 58.2 (Feb. 2001), pp. 266–277 (cit. on p. 54).

[48] M. N. Hinman and H. Lou. “Diverse molecular functions of Hu proteins”. In: *Cell. Mol. Life Sci.* 65.20 (Oct. 2008), pp. 3168–3181 (cit. on p. 54).

[49] I. Lopez de Silanes et al. “The RNA-binding protein HuR regulates DNA methylation through stabilization of DNMT3b mRNA”. In: *Nucleic Acids Res.* 37.8 (May 2009), pp. 2658–2671 (cit. on p. 54).

[50] S. Shetty. “Regulation of urokinase receptor mRNA stability by hnRNP C in lung epithelial cells”. In: *Mol. Cell. Biochem.* 272.1-2 (Apr. 2005), pp. 107–118 (cit. on p. 54).

[51] R. W. Anantha et al. “Requirement of heterogeneous nuclear ribonucleoprotein C for BRCA gene expression and homologous recombination”. In: *PLoS ONE* 8.4 (2013), e61368 (cit. on p. 54).

[52] N. C. Turner and A. N. Tutt. “Platinum chemotherapy for BRCA1-related breast cancer: do we need more evidence?” In: *Breast Cancer Res.* 14.6 (2012), p. 115 (cit. on p. 54).

[53] J. Wang et al. “Tau exon 10, whose missplicing causes frontotemporal dementia, is regulated by an intricate interplay of cis elements and trans factors”. In: *J. Neurochem.* 88.5 (Mar. 2004), pp. 1078–1090 (cit. on p. 54).

[54] Hadley Wickham and Winston Chang. *devtools: Tools to Make Developing R Packages Easier.* R package version 1.9.0. URL: http://github.com/hadley/devtools (visited on 09/12/2015) (cit. on p. 60).

[55] Dirk Eddelbuettel and Romain Francois. “Rcpp: Seamless R and C++ Integration”. In: *Journal of Statistical Software* 40.8 (Apr. 13, 2011), pp. 1–18 (cit. on p. 60).

[56] Winston Chang et al. *shiny: Web Application Framework for R.* R package version 0.12.2. 2015. URL: http://CRAN.R-project.org/package=shiny (visited on 09/12/2015) (cit. on p. 61).

[57] G. A. Barnard. “A New Test for 2 × 2 Tables”. In: *Nature* 156 (1945), p. 177 (cit. on p. 68).

[58] R. D. Boschloo. “Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities”. In: *Statistica Neerlandica* 24 (1 1970), pp. 1–9 (cit. on p. 68).

[59] N. Rabbee et al. “Power and sample size for ordered categorical data”. In: *Stat Methods Med Res* 12.1 (Jan. 2003), pp. 73–84 (cit. on p. 68).

[60] D. V. Mehrotra, I. S. Chan, and R. L. Berger. “A cautionary note on exact unconditional inference for a difference between two independent binomial proportions”. In: *Biometrics* 59.2 (June 2003), pp. 441–450 (cit. on p. 68).

[61] S. Lydersen, M. W. Fagerland, and P. Laake. “Recommended tests for association in 2 x 2 tables”. In: *Stat Med* 28.7 (Mar. 2009), pp. 1159–1175 (cit. on p. 68).

[62] Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nature Methods* 12.2 (2015), pp. 115–121. URL: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html (cit. on p. 69).