A Bioinformatics Pipeline for Identifying Dysregulated Pathways in Cancer from Comparative RNA-Seq Transcriptome Analysis

Dissertation

for the award of the degree
Doctor rerum naturalium (Dr. rer. nat.)
of the Georg-August-Universität Göttingen

within the PhD Programme in Environmental Informatics (PEI)
of the Georg-August-University School of Science (GAUSS)

submitted by

Darius Wlochowitz
from Gleiwitz (Poland)

Göttingen
2022

Thesis Committee:

Prof. Dr. Edgar Wingender,
Department of Medical Bioinformatics, University Medical Center Göttingen

Prof. Dr. Stephan Waack,
Institute of Computer Science, Georg August University Göttingen

Prof. Dr. Mehmet Gültas,
Faculty of Agriculture, South Westphalia University of Applied Sciences

Prof. Dr. Tim Beißbarth,
Department of Medical Bioinformatics, University Medical Center Göttingen

Members of the Examination Board:

1st Referee: Prof. Dr. Edgar Wingender,
University Medical Center Göttingen.

2nd Referee: Prof. Dr. Stephan Waack,
Georg August University Göttingen.

Further members of the Examination Board:

Prof. Dr. Mehmet Gültas,
Faculty of Agriculture, South Westphalia University of Applied Sciences

Prof. Dr. Tim Beißbarth,
Department of Medical Bioinformatics, University Medical Center Göttingen

Prof. Dr. Ulrich Sax,
Department of Medical Informatics, University Medical Center Göttingen

Prof. Dr. Winfried Kurth,
Department Ecoinformatics, Biometrics and Forest Growth, Georg August University Göttingen

Date of oral examination: 29th of March 2022

Abstract

Cancer is characterized as a multifactorial disease which undergoes genetic and epigenetic changes during invasive tumor growth. Thus, numerous tumor samples have been profiled using high-throughput sequencing technologies such as microarray and *RNA sequencing* (RNA-Seq) to obtain their transcriptomes. However, disentangling such high-dimensional data to identify dysregulated signaling pathways remains a difficult task. To close this gap, bioinformatics pipelines are needed to uncover gene misregulation by establishing causal regulatory links between transcription factors (TFs) and their target genes. TFs are proteins that control gene expression by recognizing short motifs called transcription factor binding sites (TFBSs) in DNA regulatory regions like promoters, enhancers, and silencers.

To this end, the goal of this thesis was to establish and evaluate a bioinformatics pipeline for comparing phenotypes based on RNA-Seq. The individual workflows of the pipeline comprise methods in RNA-Seq data analysis, promoter analysis, comprehensive functional categorization, and *master regulator analysis* (MRA), thereby identifying differentially expressed genes (DEGs), TFs, biological processes, and master regulators (MRs). For promoter analysis, a discriminative motif discovery approach using the *Boruta* feature selection algorithm is proposed, which distinguishes two DEG promoter sequence datasets based on TFBS patterns. In addition, a gene clustering approach is proposed using the *Jensen-Shannon divergence* (JSD), *principal component analysis* (PCA), and the k-means algorithm, which groups DEG promoters based on TFBS patterns related to the discriminative motifs. The gene clusters obtained are subjected to *Gene Ontology* (GO) functional categorization and MRA.

The utility of the pipeline was demonstrated using three heterogenous gene expression studies that are characterized by distinct signaling pathway activity in cancer. In the course of promoter analysis, the results indicated that Boruta's ranking-based importance scores can be used to identify biologically relevant TFs. Furthermore, the results indicated clearly separated gene clusters characterized by uniquely significant GO terms and MRs.
In conclusion, the pipeline provides a useful bioinformatics framework for the comparative study of phenotypes based on RNA-Seq to reveal variations in transcriptional regulation and pathway repertoire.

Zusammenfassung

Krebs ist eine multifaktorielle Erkrankung, die während des invasiven Tumorwachstums genetische und epigenetische Veränderungen durchläuft. So wurden zahlreiche Tumorproben mit Hilfe von Hochdurchsatz-Sequenzierungstechnologien wie Microarray und RNA-Sequenzierung (RNA-Seq) profiliert, um ihre Transkriptome zu erhalten. Es bleibt jedoch eine schwierige Aufgabe, solche hochdimensionalen Daten zu entschlüsseln, um dysregulierte Signalwege zu identifizieren. Um diese Lücke zu schließen, werden bioinformatische Pipelines benötigt, die die Fehlregulierung von Genen aufdecken, indem sie kausale regulatorische Verbindungen zwischen Transkriptionsfaktoren (TFs) und ihren Zielgenen herstellen. TFs sind Proteine, die die Genexpression steuern, indem sie kurze Motive, so genannte Transkriptionsfaktor-Bindungsstellen (TFBS), in DNA-Regulationsregionen wie z.B. *Promotoren*, *Enhancern* und *Silencern* erkennen.

Ziel dieser Arbeit war es daher, eine Bioinformatik-Pipeline für den Vergleich von Phänotypen auf der Grundlage von RNA-Seq zu entwickeln und zu bewerten. Die einzelnen Arbeitsabläufe der Pipeline umfassen Methoden in der RNA-Seq-Datenanalyse, der Promotoranalyse, der umfassenden Funktionsanalyse und der *Master-Regulator*-Analyse (MRA), wodurch differenziell exprimierte Gene (DEG), TFs, biologische Prozesse und *Master-Regulatoren* (MR) identifiziert werden. Für die Promoter-Analyse wird ein diskriminierender Motiv-Entdeckungsansatz mit dem *Boruta feature selection*-Algorithmus vorgeschlagen, der zwei DEG-Promotersequenzdatensätze auf der Grundlage von TFBS-Mustern unterscheidet. Darüber hinaus wird ein Gengruppierungsansatz vorgeschlagen, bei dem *Jensen-Shannon-Divergenz* (JSD), Hauptkomponentenanalyse (PCA) und der *k-Means*-Algorithmus verwendet werden. Dieser Ansatz gruppiert die DEG-Promotoren auf der Grundlage von TFBS-Mustern, welche mit den diskriminierenden Motiven zusammenhängen. Die erhaltenen Gengruppen werden dann einer funktionalen Kategorisierung mit Hilfe der *Gene Ontology* (GO) und einer MRA unterzogen.

Die Nützlichkeit der Pipeline wurde anhand von drei heterogenen Genexpressionsstudien demonstriert, die sich durch eine unterschiedliche Aktivität der Signalwege bei Krebs auszeichnen. Im Verlauf der Promotoranalyse zeigten die Ergebnisse, dass die auf der Rangfolge basierenden Wichtigkeitsscores von Boruta verwendet werden können, um biologisch relevante TFs zu identifizieren. Darüber hinaus wiesen die Ergebnisse auf klar getrennte Gengruppen hin, die durch eindeutig signifikante GO-Begrifflichkeiten und MRs gekennzeichnet sind. Zusammenfassend lässt sich sagen, dass die Pipeline einen nützlichen bioinformatischen Rahmen für die vergleichende Untersuchung von Phänotypen auf der Grundlage von RNA-Seq bietet, um Variationen in der Transkriptionsregulation und im Repertoire der Signalwege aufzudecken.

Acknowledgements

I would like to take this opportunity to thank everyone who accompanied and supported me during my time as a PhD student. Looking back, in the course of this long time, more and more people important to me privately have joined me, whom I can only thank personally in an appropriate way.

First and foremost, I would like to thank Prof. Dr. Edgar Wingender, who welcomed me as a PhD student and made my scientific specialization in the field of Bioinformatics possible in the first place. It was a great pleasure and honor for me to be supervised by him in the research area of gene regulation and, in this context, especially to learn about these amazingly somehow “collectable” transcription factors.

I would also like to thank Prof. Dr. Mehmet Gültas for his support throughout the duration of my dissertation. I was able to benefit greatly from his profound methodological knowledge. Moreover, he helped me significantly to find my methodological focus in the dissertation. I consider myself lucky that he was willing to supervise me and was also available for the examination committee without hesitation.

Furthermore, I would like to thank Prof. Dr. Tim Beißbarth for his support after he took over the new department. Through this, I have once again gained many new valuable impulses and was able to enrich myself especially in the area of analysis of high-dimensional data and clinical annotations. I thank him for his patience and scientific advice.

I would like to thank Prof. Dr. Stephan Waack for his willingness to be the co-referee of my dissertation. Furthermore, I thank him for the time and effort he put into my TAC meetings.

I would also like to thank the further members of the examination committee:
Prof. Dr. Ulrich Sax and Prof. Dr. Winfried Kurth. Thank you very much for your time and quick willingness to participate in the commission.

Many thanks to all collaborators for their project contributions, their valuable clinical insights, and the datasets that I was able to include in this thesis.

To cut a long story short, I would like to thank Dr. Martin Haubrock for his many years of mentoring and friendship.

Finally, I would like to thank all former and new members of the department for their support and the almost always pleasant interpersonal environment. At the same time, I would like to thank all the PhD students of my era, and all those who followed, for the pleasant moments.

Contents

List of Figures

List of Tables

Acronyms

BH *Benjamini-Hochberg*

BP *biological process*

BRE *TFIIB recognition element*

CC *cellular component*

CRC *colorectal cancer*

CRLM *colorectal cancer liver metastases*

CRMs *cis-regulatory modules*

CSS *core similarity score*

CTD *carboxy-terminal domain*

DAG *directed cyclic graph*

DBDs *DNA-binding domains*

DE *Differential expression*

DEG *differentially expressed gene*

DGI *decrease of Gini impurity*

DNA *deoxyribonucleic acid*

DPE *downstream core promoter element*

GO *Gene Ontology*

Inr *initiator*

JSD *Jensen-Shannon divergence*

MDA *mean decrease in accuracy*

MDG *mean decrease in Gini*

MF *molecular function*

minFN *minimize false negatives*

minFP *minimize false positives*

minSUM *minimize the sum of both error rates*

MIRA *maximal importance of the random attributes*

miRNAs *microRNAs*

ML *machine learning*

MR *master regulator*

MRA *master regulator analysis*

mRNA *messenger RNA*

MSS *matrix similarity score*

OOB *out-of-bag*

ORA *overrepresentation analysis*

PCA *principal component analysis*

PFM *position frequency matrix*

PMI *pointwise mutual information*

Pol II *RNA polymerase II*

PWM *position weight matrix*

RF *random forest*

RNA *ribonucleic acid*

siRNAs *small interfering RNAs*

RNA-Seq *RNA sequencing*

SNPs *single nucleotide polymorphisms*

TAFs *TBP-associated factors*

TBP *TATA-box binding protein*

TF *transcription factor*

TFBS *transcription factor binding site*

TRN *transcriptional regulatory network*

TSS *transcription start site*

1. Introduction

Cancer is characterized as a multifactorial disease which undergoes genetic and epigenetic changes during tumorigenesis. These changes enable tumor cells to evade the immune system, create the tumor microenvironment, invade the tissues surrounding it, and spread to remote sites depending on the type of cancer. Many of these hallmarks of cancer are caused by a dysregulation of signal transduction pathways that are entangled by cross-talks in large networks which process signals from both outside and inside the cell, thereby coordinating biological processes within the cell [11].

Transcription factors (TFs) are regulatory proteins that act as central components in signaling pathways. Most TFs bind to DNA regulatory regions proximal to the TSS of genes, called promoters, but also to distant regions (enhancers or silencers), by recognizing short motifs called TFBSs. As a result, TF binding controls gene expression by activating or repressing target genes. Understanding gene regulation is critical in cancer management, with TFs being particularly desired targets for cancer gene therapy [12].

High-throughput technologies have revolutionized clinical practice by allowing for the customization of cancer patient diagnosis, risk assessment, and therapy [13]. Hence, clinicians rely on computational methods to analyze high-dimensional data and discover changes of clinical relevance. RNA-Seq data analysis is an example of a high-throughput application that demands sophisticated bioinformatics workflows. These workflows process primary and secondary inputs through a series of data processing steps utilizing various mathematical and statistical methods to analyze large sequencing data and related biological annotations, with some of the outputs serving as inputs to subsequent steps. A pipeline is defined as a collection of such workflows that work together to process data in a predetermined order.
Using high-throughput sequencing technologies like RNA-Seq, thousands of mRNA transcripts may be analyzed simultaneously [14]. Since mRNA translation yields all proteins in cells, mRNA expression levels are a reliable indicator of protein abundance. Most importantly, RNA-Seq can be used to derive DEGs by comparing the gene expression levels of transcripts between sample groups using statistical methods [15].

Typically, RNA-Seq data analysis yields DEGs that are anticipated to have a specific role in clinical application. However, viewing these DEGs as a simple list fails to give mechanistic insights into the biology under investigation. As a result, a feasible next step is to assess gene overrepresentation or enrichment in biological processes, pathways, or disease phenotypes. The key aspect of coupling genes is that it is usually easier and more relevant to investigate associations to known functions or functional categories rather than deducing roles of single genes or the entire gene list from scratch [16]. GO *overrepresentation analysis* (ORA) is a method for assessing the statistical significance of the overlap of a set of genes of interest with the set of genes known to be associated with, for example, a particular biological process [16].

Furthermore, computational analysis of DEG promoters derived from RNA-Seq is a widely used strategy to evaluate the relevance of TFs based on their binding site occurrence patterns across promoter sequences [17, 18, 19]. The underlying conception here is that a group of genes may be co-expressed because their regulatory regions share a common TF or set of TFs. To computationally detect potential TFBSs using known motifs, PWMs have been widely utilized as a mathematical representation of TF binding specificity [20, 21, 22, 23]. A PWM is basically a scoring matrix that contains the *log-likelihood* of each nucleotide (A, G, T, C) in a motif. However, selecting an adequate PWM score cutoff to determine whether a given motif match in a sequence may be considered as a potential TFBS is a major obstacle [24]. Employing a high PWM score cutoff means that the majority of weak binding sites are discarded, while using a lower cutoff means that predictions are excessively noisy and biological results are biased. Therefore, detection of *overrepresented sites* in a set of target promoters using a background set has been proposed to optimize PWM score cutoffs and identify TFs likely to regulate the target promoters [25]. The purpose of conventional methods for *promoter analysis* is usually to find motifs (potential TFBSs) in which the number of motif occurrences in the target promoters is significantly greater than in the background promoters. To obtain reliable significance measures, different statistical methods have been used such as Fisher exact test, Welch t-test, multihypergeometric test, Wilcoxon rank sum test and Z-score-based methods [26, 19, 27, 25, 28]. Since these methods also require an adequate background, their performance strongly depends on it.

Moreover, the so-called *upstream analysis* strategy has proven to be an effective tool to establish TF-target interactions, enabling the identification of MRs as well as their regulated pathways [29, 30, 17, 18]. MRs are defined as molecules that can be found at the top of the regulatory hierarchy in signaling pathways and may coordinate gene expression changes at several levels. The upstream analysis strategy consists of two fundamental steps: (i) analysis of DEG promoters to discover relevant TFs using the TRANSFAC® database [22, 23] and the MATCH™ algorithm [22, 24]; (ii) discovery of signal transduction pathways that activate these TFs, and discovery of MRs that regulate these pathways using the TRANSPATH® database [31]. We have previously evaluated this strategy for the analysis of different *colorectal cancer* (CRC) phenotypes [17]. The results revealed interesting MR candidates and signaling pathways whose activity may explain differences in tumor development and progression underlying the CRC phenotypes.

In bioinformatics, the application of *machine learning* (ML) is extensively investigated using supervised and unsupervised methods. Some common unsupervised methods in transcriptome profiling are PCA, *hierarchical clustering*, and *k-means clustering* [32, 33]. These methods focus on the similarity of gene expression patterns to group genes across biological samples, suggesting that genes with similar patterns are linked to similar regulatory processes and biological functions.

Supervised methods like *support vector machines* [34] and *artificial neural networks* [35] are frequently employed in classification problems where the goal is to determine a robust classifier on predetermined sample groups in order to assign every new sample to the correct group. In addition, *feature selection* by *random forest* (RF) [36] has proven to be effective in data preprocessing to reduce noise or redundant data. For example, RF has previously been utilized in gene expression studies to select subsets of genes based on gene relevance rankings for sample classification [37].

Furthermore, information theory-based measures like the JSD are widely applied in the field of bioinformatics to measure the similarity (or dissimilarity) between two probability distributions [38, 39, 40, 41, 42, 43].

1.1. Objectives and Structure of the Thesis

1.1.1. Objectives

Inspired by the upstream analysis strategy, the main objective of this thesis was to develop a bioinformatics pipeline that links transcriptome profiling by RNA-Seq to biological processes and MRs through promoter analysis, thereby uncovering dysregulated signaling pathways underyling the phenotypes under study.

In the following, I will introduce and motivate two new ideas that represent methodological modifications to the upstream analysis and which have been incorporated into the pipeline. A detailed description of the full pipeline can be found in Section 4.1.

Since RF and JSD have already shown promising results in gene expression studies with various applications for feature selection and clustering, their performance and functionality was evaluated for the pipeline:

- For promoter analysis, a discriminative motif discovery approach using the *Boruta* feature selection algorithm [44] is proposed, which distinguishes two DEG promoter sequence datasets based on TFBS patterns. An TFBS pattern is characterized by the frequency distribution formed by TFBS occurrences across a promoter set. Furthermore, Boruta's ranking-based importance scores are used to capture and prioritize discriminative motifs (characterized by PWMs). This approach was compared to conventional overrepresentation methods to evaluate whether the TFBS patterns associated with the most relevant motifs fall inside or outside of the concept of overrepresentation (or underrepresentation), or may not be discovered by conventional methods.

- For gene clustering, an approach is proposed to connect the results obtained from promoter analysis to ORA and MRA. To this end, genes (related to DEG promoters) are clustered based on similarities between TFBS patterns related to the most relevant motifs using JSD, PCA, and k-means. The objective here was to investigate whether specific GO terms and MR candidates are uniquely significant for the gene clusters obtained.

1.2. Structure of the Thesis

I laid out the remainder of the thesis as follows. In Chapter 2, I will provide an overview of the biological background necessary to motivate and understand fundamental concepts in biology, focusing on gene regulation, molecular networks, and cancer. In Chapter 3, I will provide an overview of the computational prerequisites which includes methods, workflows in Bioinformatics, and databases necessary to motivate and understand the bioinformatics pipeline developed in my thesis. Chapter 4 provides a detailed description of the pipeline, which covers utilized tools and their parameters. In addition, I will introduce the reference datasets used to evaluate the performance and functionality of the pipeline. Chapter 5 covers the results obtained after applying the pipeline to the reference datasets. This chapter also highlights some of the key findings and their biological relevance in light of previous studies. Afterwards, I will discuss the results in Chapter 6, addressing several methodological and biological aspects of the pipeline. In addition, I will also discuss the choice of methods, including parameters and databases, and their main limitations. Finally, I will summarize the conclusions and provide an outlook for further work in Chapter 7.

1.3. Journal Articles

First authorship and shared first authorship

- Wlochowitz D, Haubrock M, Arackal J, Bleckmann A, Wolff A, Beißbarth T, Wingender E, Gültas M. **Computational Identification of Key Regulators in Two Different Colorectal Cancer Cell Lines**. Front Genet. 2016 Apr 5;7:42. doi: 10.3389/fgene.2016.00042.
- Menck K, Wlochowitz D, Wachter A, Conradi LC, Wolff A, Peeck M, Scheel A, Schlüter K, Korf U, Wiemann S, Schildhaus HU, Wingender E, Pukrop T, Homayounfar K, Beißbarth T, Bleckmann A. **High-throughput profiling of colorectal cancer liver metastases reveals intra- and interpatient heterogeneity in the EGFR- and WNT-pathway associated with clinical outcome**. [In preparation (2022), Abstract B.1].

Co-authorship

- Blazquez R, Wlochowitz D, Wolff A, Seitz S, Wachter A, Perera-Bel J, Bleckmann A, Beißbarth T, Salinas G, Riemenschneider MJ, Pröscholdt M, Evert M, Utpatel K, Siam L, Schatlo B, Balkenhol M, Stadelmann C, Schildhaus HU, Korf U, Reinz E, Wiemann S, Vollmer E, Schulz M, Ritter U, Hanisch UK, Pukrop T. **PI3K: A master regulator of brain metastasis-promoting macrophages/microglia**. Glia. 2018 Nov;66(11):2438-2455. doi: 10.1002/glia.23485. Epub 2018 Oct 25.
- Blazquez R, Rietkötter E, Wenske B, Wlochowitz D, Sparrer D, Vollmer E, Müller G, Seegerer J, Sun X, Dettmer K, Barrantes-Freer A, Stange L, Utpatel K, Bleckmann A, Treiber H, Bohnenberger H, Lenz C, Schulz M, Reimelt C, Hackl C, Grade M, Büyüktas D, Siam L, Balkenhol M, Stadelmann C, Kube D, Krahn MP, Pröscholdt MA, Riemenschneider MJ, Evert M, Öfner PJ, Klein CA, Hanisch UK, Binder C, Pukrop T. **LEF1 supports metastatic brain colonization by regulating glutathione metabolism and increasing ROS resistance in breast cancer**. Int J Cancer. 2020 Jun 1;146(11):3170-3183. doi: 10.1002/ijc.32742. Epub 2019 Nov 11.
- Kalya M, Kel A, Wlochowitz D, Wingender E, Beißbarth T. **IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme**. Front Genet. 2021 Jun 15;12:670240. doi: 10.3389/fgene.2021.670240.
- Menck K, Heinrichs S, Wlochowitz D, Sitte M, Nöding H, Janshoff A, Treiber H, Ruhwedel T, Schatlo B, von der Brelie C, Wiemann S, Pukrop T, Beißbarth T, Binder C, Bleckmann A. **WNT11/ROR2 signaling is associated with tumor invasion and poor survival in breast cancer**. J Exp Clin Cancer Res. 2021 Dec 15;40(1):395. doi: 10.1186/s13046-021-02187-z.

- Büntzel J, Klemp HG, Krätzner R, Schulz M, Dihazi GH, Streit F, Bleckmann A, Menck K, Wlochowitz D, Binder C. **Metabolomic Profiling of Blood-Derived Microvesicles in Breast Cancer Patients**. Int J Mol Sci. 2021 Dec 17;22(24):13540. doi: 10.3390/ijms222413540.

2. Biological Background

2.1. Molecular Biology of Gene Regulation

At various scales, our environment is a collection of densely interconnected complex systems. While the exact disciplines of physics and chemistry explain our environment from the subatomic to molecular level, biology is responsible for dealing with an inexact and complex world. Nonetheless, biology may be described with a degree of clarity at various degrees of detail. The molecular level of *deoxyribonucleic acid* (DNA), *ribonucleic acid* (RNA), proteins, and metabolites is the lowest biological level of detail. All of these molecules make up a cell, which is part of a tissue. Organs in an organism are made up of different tissues, and the ecosystem is made up of many species.

2.1.1. DNA, RNA, and the Flow of Genetic Information

DNA stores information about how an organism is assembled. DNA forms a coiled ladder, called a *double helix*, that consists of two sugar phosphate backbones and pairs of nucleotide bases, i.e., adenine (A), cytosine (C), guanine (G), and thymine (T). The nucleotide A only couples or pairs with T, while C only pairs with G. Proteins catalyze processes and are responsible for many other functions in the cell, while DNA is the passive portion of the biochemistry of a cell. Gene expression is the process of transmitting information from DNA to proteins (Figure 2.1).

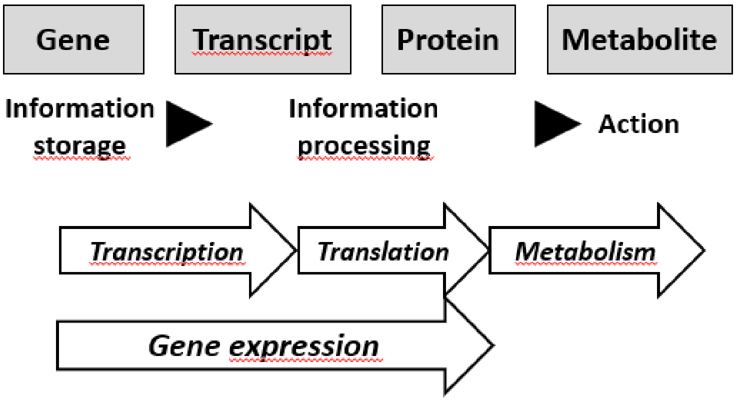


Figure 2.1.: Information flow from transcription to metabolism.

Transcription is a controlled process in which an RNA polymerase-containing protein complex opens the DNA helix, reads one strand, and synthesizes a matching RNA. It starts and ends at signal sequences known as promoters and terminators, respectively. In eukaryotes, a transcript is the RNA that corresponds to a certain gene (Figure 2.2). The introns are excised from the RNA in eukaryotes, leaving only the exons.

Ribosomes generate amino acid chains from spliced RNA during translation. The information in RNA is read in triplets (codons), which have 43 possible combinations. These are used to code for 20 amino acids, one start codon, and three stop codons (more than one codon may stand for one amino acid). The primary structure of a protein is its amino acid sequence, whereas the secondary structure is made up of regular three-dimensional patterns like loops, helices, and sheets. The tertiary structure defines how these patterns are organized in space to create a protein or a subunit of a protein. Finally, the quaternary structure shows how the subunits' amino acid chains are organized to form an active protein complex. Moreover, proteins can perform a variety of tasks in the cell. For example, structural proteins maintain the cell's structure, TFs control transcription, and enzymes catalyze the conversion of one metabolite (e.g., sugars and amino acids) to another.

2.1.2. Transcriptional Regulation

Gene expression differs between cell types or in response to various stimuli. To be translated into proteins, genetic information that an individual inherits as DNA must first be translated into an RNA product. The process of converting a specific region of a double-stranded DNA sequence into a single-stranded RNA sequence is known as RNA synthesis or transcription (Figure 2.2).

TFs are specific proteins that either enable or inhibit the recruitment of the *RNA polymerase II* (Pol II) complex, thereby coordinating the transcriptional regulation mechanism. In addition, certain RNA products act post-transcriptionally, which includes *microRNAs* (miRNAs) and *small interfering RNAs* (siRNAs) (Figure 2.2). As a result, transcriptional regulation is the most important control point in gene expression, and it varies depending on the cell type and conditions.

2.1.3. Epigenetic Regulation

When cells divide, epigenetic processes can ensure that differential expression patterns are transmitted in a stable manner. Chromatin remodeling and DNA methylation are the two epigenetic processes that preserve heritable transcription states. The size of the molecule is limited by the packaging of DNA into chromatin, which is roughly two meters of DNA per human cell. The chromatin is made up of nucleosomes, which are repeating units of histones and DNA. Heterochromatin is transcriptionally silent because it is highly compacted chromatin. Additional histone tail modifications attract or repel chromatin remodeling complex regulatory proteins [42]. Moreover, chromatin-remodeling factors and enzymes that cova-

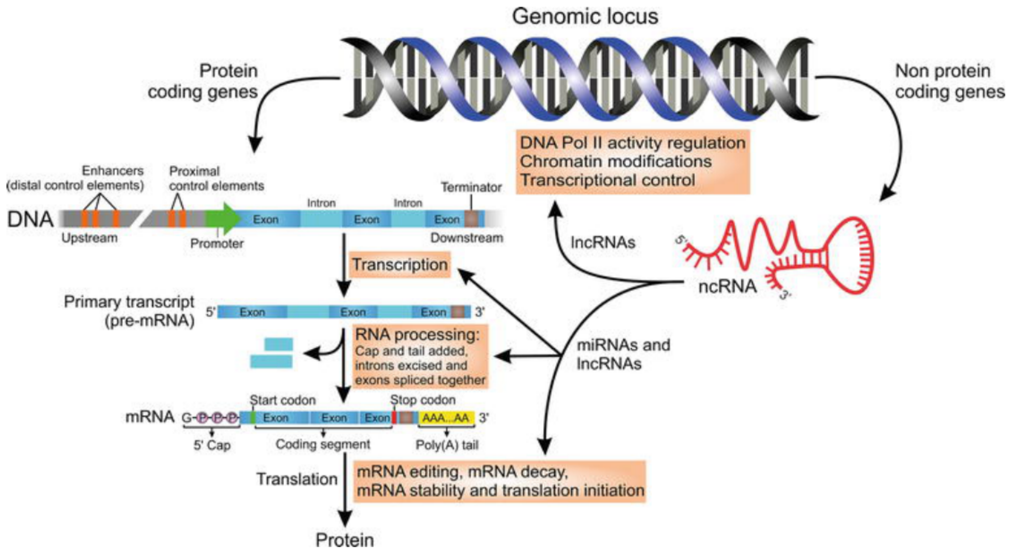


Figure 2.2.: Control of gene expression. The schematic representation depicts the main molecular mechanisms that control gene expression. Gene expression is controlled by elements located proximal to the TSS of a gene (promoter region), as well as distal elements (enhancer region) that may be located far away from a gene. Regulatory elements in the pre-mRNA sequence dictate which exons are retained or spliced out during alternative splicing. Several alternative exons in a single pre-mRNA reveal distinct forms of alternative-splicing patterns. The attachment of the 5' Cap and Poly(A) tail is a tightly controlled process that is critical for mRNA stability and the transport of the mRNA from the cytoplasm to the nucleus. *Non-coding RNAs* (ncRNAs) can regulate various events that result in the synthesis of distinct proteins. ncRNAs guide multiprotein complexes to specific genomic loci, inducing chromatin alterations (i.e., methylation and acetylation) and DNA polymerase II activity. Adapted from [1].

lently modify histones are recruited by specific TFs bound to defined regions in the genome. This event causes additional TFs to bind to chromatin, resulting in either a permissive or non-permissive environment for gene expression [2]. To create a permissive state, chromatin modifying factors are linked with the elongation complex prior to RNA polymerase II during RNA synthesis.

2.1.4. DNA Regulatory Elements: Promoters, Enhancers, Silencers, and Insulators

Upstream and downstream of the TSS, each gene has one or more regulatory regions: promoters, enhancers, silencers, and insulators (Figure 2.3). The promoter region of a gene is located upstream of the coding sequence. At the promoter region, transcription adapter proteins assist to recruit the *pre-initiation complex*'s (PIC) TFs. The PIC directs the RNA polymerase II complex to the TSS at the basal promoter, called the core promoter. The rest of the promoter consists of TFBSs. In the absence of certain TFs, there is no transcriptional activity in eukaryotes. As a result, the transcription is turned off by default [45]. Furthermore, the majority of core promoters has several initiation sites rather than a single TSS.

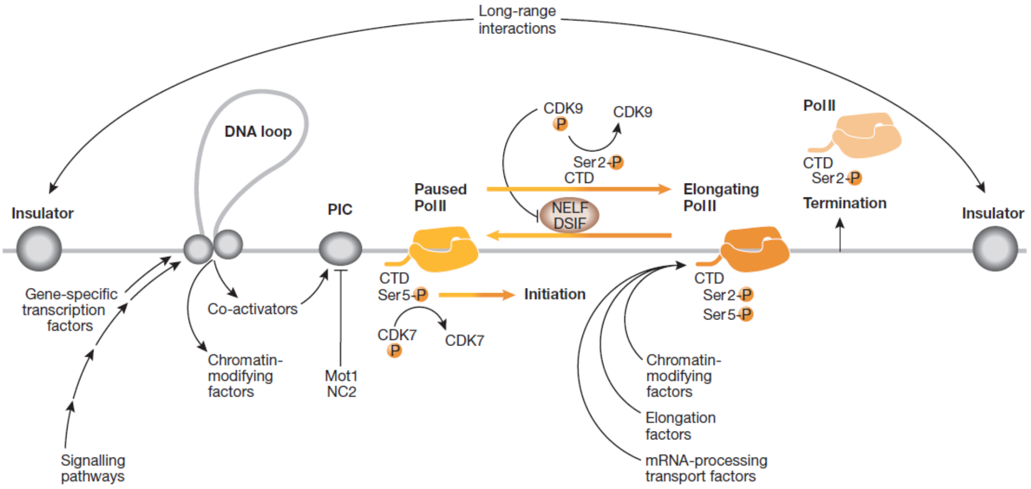


Figure 2.3.: Transcriptional regulation. Specific TFs that bind to regulatory region are activated via signaling pathways. Through DNA looping, distant TFs can interact with more proximal TFs. Chromatin-modifying factors and transcriptional coactivator complexes can be recruited by TFs. The co-activators assist to recruit and boost the transcription activity of the PIC. Specific repressors can control PIC assembly (Mot1 and NC2). A cyclin-dependent kinase phosphorylates the carboxyl-terminal domain (CTD) of Pol II at the start of transcription (CDK7). In higher eukaryotes, transcription elongation is frequently inhibited, resulting in a halted polymerase. Several components involved in chromatin modification, transcription elongation, mRNA processing, mRNA transport, and termination are recruited by the elongating Pol II with the CTD phosphorylated at particular serines. Long-range chromatin interactions, such as those between insulator elements, serve to segregate chromatin regions and prevent regulatory signals from spreading. Adapted from [2].

Enhancers can be located hundreds of thousands of base pairs (bp) upstream or downstream of the promoters they regulate [46]. Insulators, on the other hand, are regulatory elements that, when positioned between an enhancer and a promoter, can block enhancer–promoter interactions. Furthermore, silencers are defined as genomic regions where specific TFs bind and inhibit or slow down transcription.

All of these regulatory regions are organized in a modular manner, with each one consisting of one or more distinct regions known as CRMs [47]. CRMs generate expression patterns in a spatial and temporal manner by reading-out the concentrations of various TFs under distinct conditions [47].

Several TFBSs may be found in each CRM. TFs often interact in synergy, meaning that their total impact is greater than the sum of their separate effects. In contrast, antagonistic effects may take place when TFs bind to overlapping sites [48]. TFs compete for regulatory control since they may identify the same sites with varying affinities in many situations. Likewise, repressors can suppress gene expression by competing with activators for DNA binding, concealing the activation interface [49].

Eukaryotic Promoters

In gene promoters of eukaryotes, two functional components are always present, although they are often difficult to detect just based on the information in the DNA sequence. The RNA polymerase complex is recruited to the core promoter, while the other component corresponds to modules that confer transcription specificity. In various genes, the content and structure of these modules and TFBSs varies greatly [45]. Sequence conservation upstream of the TSS is linked to gene function in mammals [50]. Because they contain more TFBSs, promoters implicated in biological processes like development or cell-cell communication are more conserved. On the other hand, promoters implicated in basic biological processes, such as ribosome metabolism, show minimal conservation. Many of these genes are housekeeping genes, meaning they are expressed in many tissues and hence do not require specialized regulation [51].

The promoter region between -500 and +50 bp relative to the TSS is adequate to trigger transcription of most human genes, according to *in vitro* studies [52]. The *TATA box*, *initiator* (Inr), *TFIIB recognition element* (BRE), and *downstream core promoter element* (DPE) are all typical CRMs present in eukaryotic core promoters (Figure 2.4). Some, but not all, core promoters have each of these core promoter components. Each of these motifs has a role in the transcription initiation process. Proteins that interact with each other to control the transcription initiation complex need different distances between binding sites. TFIID, a basic TF, is a component of the Pol II initiation complex. *TATA-box binding protein* (TBP) and other *TBP-associated factors* (TAFs) are included. At a certain distance, a complex consisting of TFIID/TBP, TAF150, and TAF250 binds to the TATA-box and the Inr. TFII-I and YY1, two DNA-binding factors, interact with the Inr. TFII-I is a *basic-helix-loop-helix* protein that activates transcription *in vitro* by binding to the Inr and

E-box motifs.

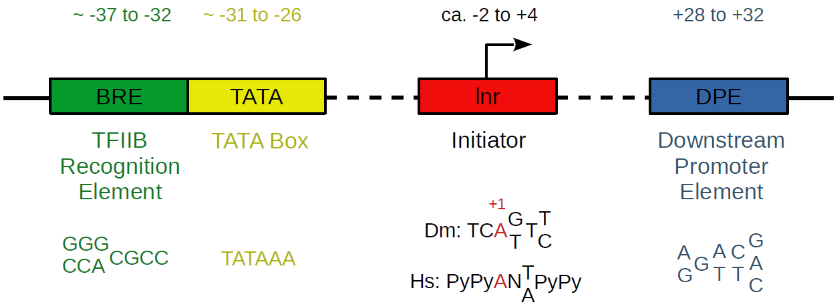


Figure 2.4.: The core promoter. In core promoters, common CRMs are frequently present at certain distances from the TSS. Some, all, or none of these motifs may be found in any given core promoter. The BRE is a subset of TATA boxes that has been extended upstream. In other genes, distinct TFs bind upstream near the TATA-box. Drosophila core promoters were found to have the DPE consensus. For both Drosophila (*Dm*) and humans (*Hs*), the Inr consensus sequence is displayed. Inspired by [3].

2.1.5. Transcription Factors

Transcription factors TFs are regulatory proteins that are essential for regulating the specific transcriptional program of a cell. Most of them bind to regulatory regions located near the TSS of genes, but also to distant regions (enhancers or silencers), by recognizing to short DNA regions (5-20 bp) of low information content, known as transcription factor binding sites TFBSs. Acting alone or as part of a complex, they can promote or inhibit the recruitment of Pol II to promoter regions. According to a comprehensive study, the human genome contains around 1,400 TFs [53]. Their binding sites are usually grouped into CRMs which are DNA segments that control gene expression. Functioning of CRMs is determined by both the accessibility of CRMs and the relative number of active TFs.

Transcription Factor Classification

A vast number of different genes encoding TFs can be found in the genomes of multi-cellular organisms. Each TF contains one or more *DNA-binding domains* (DBDs) that characterize sequence specificity. How the DBD interacts with the DNA recognition motif is determined by the structure of the domain. The major factors of the specificity of an amino acid sequence to a specific DNA motif are hydrogen bridges and *van der Waals force*. TFs can bind to the DNA as dimers or protein complexes, affecting their protein structure and DBD accessibility.

Moreover, TFs can be classified based on the structure of their DBDs, which can reveal information about their functions; for example, homeodomain-containing TFs are frequently linked to developmental processes, while those in the interferon regulatory factor families are linked to inducing immune responses to viral infections [54].

First proposed by E. Wingender in 1997 and further extended in 2013, TFs can be classified into different structural groups according to their DBDs, called *superclasses* [10, 55, 56, 57]. These can be further classified into the general levels *classes*, *families*, and *subfamilies* (Table 2.1). According to the database TFClass, the three largest superclasses are *Basic domain*, *Zinc-coordinating domain*, *Helix-turn-helix domain* with 11%, 52%, and 27% human TF genes, respectively [10].

Table 2.1.: TFClass classification for human. The table indicates the number of classes, families, and subfamilies in each superclass. Adopted from [10].

Superclasses	**Classes**	**Families**	**Subfamilies**
Basic domain	3	18	36
Zinc-coordinating domain	8	25	130
Helix-turn-helix domain	7	22	143
Other all-α-helical DBD	2	8	11
α-Helices exposed by β-structures	2	7	4
Immunoglobulin fold	7	16	6
β-Hairpin exposed by an α/β-scaffold	2	3	3
β-Sheet binding to DNA	2	2	0
β-Barrel DBD	1	1	3
Yet undefined DBD	5	10	0
All	39	112	336

2.2. Molecular Networks

Molecular processes and cellular functions depend on the coordinated effects as well as interactions of their chemical components. We are now able to identify the chemical composition of cells on a genome scale using a variety of high-throughput experimental techniques. These methods include the measurement of messenger RNA molecules synthesized under specific conditions (transcriptomics), whole genome annotation and sequencing (genomics), measurements of protein interactions, abundance and functional states (proteomics) as well as of the concentration and presence of metabolites (metabolomics). All of these techniques can be utilized to reconstruct networks of biochemical reactions that operate in cells. In this section I will give a short introduction about the reconstruction of signaling, metabolic, and transcriptional regulatory networks. It is noteworthy to mention that all of these networks interact with one another and are therefore neither independent nor separate.

2.2.1. Metabolic Networks

The functions of a cell comprise complex networks of interacting chemical reactions precisely organized in time and space. In other words, these networks of biochemical reactions create observable cellular functions. The process of identifying all of the chemical reactions that create a network is known as network reconstruction. The network reconstruction in metabolomics has been implemented and developed for many different organisms. *Intermediary metabolism* may be thought of as a "biochemical engine" that transforms raw products into building blocks and energy which are required for maintaining cells, performing different cellular functions, and creating biological structures. Intermediary metabolism is highly dynamic, follows the laws of chemistry and physics, and is therefore constrained by physicochemical limitations. It has an intricate regulatory system that enables it to respond to a wide range of external stimuli.

2.2.2. Transcriptional Regulatory Networks

Chemical reactions in metabolic networks involve the dismemberment and reassembling of small molecules through a sequence of chemical conversions. Transcriptional regulatory networks (TRNs), on the other hand, entail the interaction and association of large molecules, such as protein-protein and DNA-protein interactions. However, metabolites do engage in some of these chemical conversions directly. In response to different environmental and developmental factors, a complex TRN regulates which genes are expressed. Even though the underlying chemical reactions of metabolic and regulatory networks are the same, the reaction types that create building blocks of TRN differ from those of metabolism. The promoter region of genes, which includes the *cis*-regulatory binding sites for TFs regulating a gene's expression, is the fundamental functional block of TRNs. A directed graph (Section 3.4) with vertices representing genes and directed edges indicating regulation can be utilized to model TRNs.

2.2.3. Signaling Networks

In *signaling networks*, a signal is transduced from the outside of the cell to the inside. A wide range of signals are carried from the cellular environment to the nucleus or several other organelles and functions inside the cell. These environmental signals may be physicochemical (e.g., osmotic pressure or pH) or biological, such as chemo- and cytokines. A cascade of regulatory events occur when a cell encounters an external signal, for example, the binding of a growth factor or a hormone to an extracellular receptor. These regulatory events can be complex such as a network of phosphorylations or as elementary as the opening of an ion channel. The transmission of a signal often entails: (i) the binding of a ligand to an external receptor; (ii) the following phosphorylation of an intracellular enzyme; (iii) signal transmission and amplification; and (iv) resulting changes in cellular functions, i.e. increased gene expression.

2.3. Cancer Biology

Cancer is a broad term that refers to a range of diseases in which cells proliferate uncontrollably. Cancers are categorized based on the cells or organs that they come from. There are approximately 100 distinct forms of cancers due to the fact that malignant development may arise in different places in and on the body [58]. Although cancer is a vastly complicated and diverse disease, virtually all cancers have a set of characteristics which are known as cancer hallmarks. The classic hallmarks of cancer are the ability of a cancer cell to unlimitedly multiply, evasion of programmed cell death, persistent angiogenesis, insensitivity to growth-inhibitory signals, spread, tissue invasion, and metastasis [11].

Through advanced sequencing technologies, our understanding of the underlying genomic, epigenomic, and proteomic diversity of a tumor cell underlying these alterations has greatly improved in recent years. Only a minority of the mutations in a cancer cell's genome are drivers that are considered to be causative and important for tumorigenesis by contributing to growth advantage. The others are known as passengers, and they are described as those that do not influence the cell's fitness but are acquired along the way, for example, through genomic instability. Moreover, it has been demonstrated that the hallmarks are influenced not only by the cancer cell itself, but also by the surrounding tumor microenvironment. Another important factor in cancer biology is the inflammatory state which is triggered by immune system cells and allows tumor growth in a variety of ways [11].

Metastasis

Cancer cells go through a series of stages to develop *metastasis*, from local invasion to metastatic growth. Mutations in numerous genes and pathways are driving this process. More than 90% of cancer-related deaths are caused by metastasis [58]. The primary tumor cells breach the walls of blood vessels, survive, and invade and adapt to the microenvironment of a remote site to form a metastatic lesion. Clinical management of tumors is usually determined by the primary site. In the presence of metastatic lesions, however, no identifiable primary site may be detected in 3-5% of all diagnosed malignancies [59].

3. Computational Prerequisites

3.1. Computational Prediction of TFBSs

Pattern matching algorithms have served in most approaches for predicting potential TFBSs. In this regard, the alignment of experimentally verified sites can be used to obtain a consensus binding sequence that describes a pattern. Each position of the consensus sequence is allocated an IUPAC code letter [60] that represents the nucleotide (A, G, T, or C) composition of each column of the alignment. The relative frequencies of nucleotides at each position are lost in the consensus representation. Based on the normalized frequencies of nucleotides, the *position weight matrix* PWM model better characterizes the DNA binding preferences at each position. Several resources include libraries of PWMs from various species. The most important PWM collections for vertebrate TFs are TRANSFAC® [22] and JASPAR [61]. Furthermore, MATCH™ [22, 24], ConSite2 [62], and PROMO [63] are some of the most important methods that utilize PWMs to predict potential sites.

3.1.1. Position Weight Matrix (PWM)

The most often utilized model to characterize the DNA binding specificity of a TF is the PWM [64, 65]. A *position frequency matrix* PFM is created from a set of known TFBSs that represents the frequency of each nucleotide at each position in the motif (Figure 3.1). Using these frequencies, a PWM is constructed, which gives each nucleotide at each position a log-scale value. A score equivalent to the sum of all values at each position is determined based on the PWM for any given sequence. A sequence logo can be used to create a graphical representation of the motif from the PWM. In the logo, the information content of a position is represented by the y-axis, and the height of each nucleotide is proportionate to its frequency in the motif's respective position.

3.1.2. The TRANSFAC® Database

TRANSFAC® is database focused on the relationship between TFs and their DNA-binding sites and DNA-binding profiles [22, 23]. The structural and functional characteristics of each TF are described in detail where scientific literature supports them. The exact location of a site, its nucleotide sequence, and the methods that led to its identification are provided in the database. An annotation team generated a part of the PWMs in TRANSFAC®, while others were obtained from scientific literature.

A

Site 1	T	T	T	G	T	C	G	C	A	T
Site 2	T	T	T	G	T	C	G	A	A	T
Site 3	C	A	T	G	T	C	G	G	T	C
Site 4	C	A	T	G	T	C	G	G	T	C
…	…	…	…	…	…	…	…	…	…	…
Site N	G	T	T	G	T	C	T	G	G	A
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

B

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	68	102	2	3	1	4	2	53	131	239
C	182	49	4	2	3	596	4	25	173	65
G	65	26	44	597	2	2	467	494	145	67
T	292	430	557	5	601	5	134	35	158	236

C

$$W(b.i) = \ln \frac{f(b.i)}{f_{exp}(b)}$$

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	-0.8	-0.4	-4.3	-3.9	-5.0	-3.6	-4.3	-1.1	-0.1	0.5
C	0.2	-1.1	-3.6	-4.3	-3.9	1.4	-3.6	-1.8	0.1	-0.8
G	-0.8	-1.8	-1.2	1.4	-4.3	-4.3	1.1	1.2	0.0	-0.8
T	0.7	1.0	1.3	-3.4	1.4	-3.4	-0.1	-1.5	0.0	0.4

D

$$S = \sum_{i}^{n} W(b_i.i)$$

A	T	T	A	G	G	C	C	C	G
-0.8	1.0	1.3	-3.4	-4.3	-4.3	3.6	-1.8	0.1	-0.8

$\sum = -9.4$

E

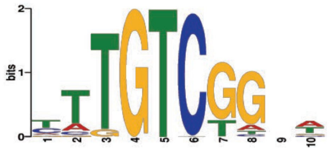


Figure 3.1.: Motif representation. The process of creating a motif for a TF based on known instances of the factor's genomic binding sites. DNA affinity purification sequencing is used to find sequences bound by ARF5 [4]. First, *N* distinct ARF5 binding sites are aligned (A). The PFM is then created by computing the nucleotide frequency at each position of the binding site alignment. (B). Next, the PFM is converted to a PWM. In the formula, the weight of nucleotide *b* at position *i* is *W(b, i)*, the frequency of this nucleotide is *f(b,i)*, and the expected background frequency of the given nucleotide is f_{exp} (C). The total of the corresponding PWM weights can be used to score a sequence (D). The sequence logo shows the TF's preference for each binding site position (E). This approach is among multiple alternatives for calculating a sequence score [5, 6, 7]. Adapted from [8].

TRANSFAC® features comprise more than 10,000 PWMs that correspond to over 49,000 factors in different species (Table 3.1). The TRANSFAC® database serves as an encyclopedia of TFs; however, when combined with a tool like MATCH™, it can be effectively used to predict sites in regulatory regions of the DNA. For each TF, the target sequences and regulated genes are specified, enabling for the creation of comprehensive datasets for each TF binding sequence. Moreover, TRNs have been built and studied using the TF-target gene relationships reported in TRANSFAC®. TRANSFAC® is currently maintained by geneXplain GmbH (`https://genexplain.com`).

Table 3.1.: Statistics of TRANSFAC® release 2021.1 (Source: `https://genexplain.com/wp-content/uploads/2021/01/TRANSFAC_statistics.pdf`).

Category	Entries
Factors	49,098
miRNAs	1,772
DNA Sites	50,904
mRNA Sites	67,820
Factor-DNA Site Links	68,917
miRNA-mRNA Site Links	74,672
Genes	102,887
ChIP TFBS	91,809,826
DNase Hypersensitivity Sites	15,376,241
Histone Modification Fragments	1,071,162
DNA Methylation Fragments	51,926
Matrices	10,290
References	40,657

3.1.3. MATCH™

MATCH™ provides a search algorithm for the detection of potential TFBSs in any given set of genomic sequences using PWMs [22, 24]. MATCH™ can be deployed using precompiled TRANSFAC® PWM libraries with different *matrix similarity score* (MSS) cutoff values to enable a range of search modes with varying degrees of stringency. It is also possible to create a custom user profile with subsets of PWMs with default or user-defined cutoff values. The MSS and the *core similarity score* (CSS) are the score types used by the MATCH™ algorithm. They range from 0.0 to 1.0, with 1.0 being an exact motif match. The top five most conserved consecutive positions of a matrix are designated as the core of that matrix. The MSS and CSS scores are computed using the same formula as shown below. The CSS is computed using only the core positions of the matrix, whereas the MSS uses all of the matrix positions. Briefly, the CSS is determined for each pentanucleotide, and the associated values are stored in the hash table. Each occurrence of this pentanucleotide is searched up in the sequence and lengthened at both ends to suit the matrix length for each entry of the hash table with the CSS greater than the cutoff [24]. Following that, the MSS is computed. In the MATCH™ output, only matches with a MSS greater than a specified cutoff are reported as potential TFBSs by the search algorithm. For the subsequence x of the length l, the *MSS* (and the *CSS*) can be calculated as follows (see [24] for more details):

$$MSS = \frac{Current - Min}{Max - Min}, \tag{3.1.1}$$

$$Current : \sum_{i=1}^{l} I(i) f_{i,b_i}. \tag{3.1.2}$$

The parameter $f_{i,N}$ indicates the frequency of a nucleotide N $(N \in \{A, T, C, G\})$ to occur in the matrix at the position i.

$$Min : \sum_{i=1}^{l} I(i) f_i^{\min} \tag{3.1.3}$$

$$Max : \sum_{i=1}^{l} I(i) f_i^{\max} \tag{3.1.4}$$

Parameters $f_i^{\min}$ and $f_i^{\max}$ describe the lowest and highest frequency of the nucleotide N at the position i of the matrix, respectively. The information vector $\mathbb{I}(i)$ defines the conservation of positions i in a matrix, and can be calculated as given in equation 3.1.5.

$$\mathbb{I}(i) = \sum_{N \in \{A,T,C,G\}} f_{i,N} \ln\left(4 f_{i,N}\right), \quad i = 1, 2, ..., l. \tag{3.1.5}$$

Mismatches in less conserved regions are accepted more readily when the frequencies are multiplied with the information vector, whereas mismatches in highly conserved regions are strongly discouraged [24]. As a result, the performance in recognizing the binding sites of TFs is improved, compared to methods which do not employ the information vector.

TRANSFAC® PWM Profiles

When using the MATCH™ algorithm, it is essential to assess whether a motif in any given sequence is a potential TFBS or not. For this purpose, the user must specify appropriate cutoff values for PWM scores. Noteworthy, the number of site predictions is inversely proportional to the cutoff values. The MATCH™ algorithm comes with pre-calculated cutoffs that allow to *minimize false negatives* (minFN), *minimize the sum of both error rates* (minSUM), and *minimize false positives* (minFP).

- **minimize false negatives (minFN)**
  The minFN profile can be used to minimize the number of false negative sites in sequences of interest (target set). This profile was created using known binding sites of TFs. SELEX sites or sets of oligonucleotides based on the nucleotide distribution in the PWM were utilized to compensate for insufficient genomic binding sites (< 10) and estimate minFN. The cutoff is defined as the score at which at least 90% of the target set is recognized, equivalent to a 10% false negative rate [24].

- **minimize false positives (minFP)**
  The minFP profile can be used to minimize the number of false positive sites in sequences of interest (target set). This profile was created using an estimation of the false positive rate in upstream sequences by applying the MATCH™ algorithm. The cutoff is defined as the score that gives 1% of hits in the target set relative to the number of hits received at the minFN cutoff [24].

- **minimize the sum of both error rates (minSUM)**
  The minSUM profile can be used to minimize the sum of both error rates mentioned above. For each cutoff ranging from minFN to minFP, the sum of corresponding percentages for false positives and false negatives is determined, with the false positive rate at minFN (10% false negative rate) set as 100%. The cutoff is determined by the score at which this sum is the smallest [24].

3.1.4. Definition and Construction of a Site Count Matrix

A site count matrix $\mathbb{M}$ characterizes the TFBS frequency distributions in genomic sequences acquired from computational methods such as MATCH™ [66]. The rows of $\mathbb{M}$ correspond to sequences labeled with their IDs and columns correspond to site models (the PWMs). According to the observed frequencies or counts of predicted sites in sequences, the entries of $\mathbb{M}$ are determined as follows.

Let m be the number of sequences under study and s_i $(i = 1,\dots,m)$ be a sequence. Further, let t_j $(j = 1,\dots,n)$ be a site predicted using PWM j, where n is the number of PWMs in the library. The entry of $\mathbb{M}$ at position (i, j), f_{ij}, is the frequency of t_j in s_i.

3.1.5. Promoter Analysis: Discovery of Site Overrepresention

Site overrepresentation methods make use of the underlying idea that a shared collection of TFs is expected to control a group of transcriptionally co-regulated genes under certain circumstances or conditions. Thus, overrepresented sites are likely to be functionally meaningful after background noise is taken into consideration. There exists a variety of methods that discover overrepresented sites within the promoters of co-expressed genes, with oPOSSUM [19, 67, 68] and F-Match [25] being two of the most commonly utilized ones. oPOSSUM combines phylogenetic footprinting with statistical methods to calculate site overrepresentation and the fraction of genes having sites in a target set as compared with a random background set using the Z-score and Fisher exact test, respectively [19]. F-Match, on the other hand, is built on top of the MATCH™ algorithm and first evaluates the sequence sets to find optimal PWM score cutoffs.

The F-Match algorithm maximizes and minimizes the ratio between the frequency of sites in the sequences of the target set as well as the background set to account for overrepresented and underrepresented sites, respectively. It is important to mention that overrepresentation methods are extremely dependent on the background. In addition, many binding sites may be detected by their overrepresentation or conservation, while others may be more difficult to spot due to low affinity binding sites, alternate recognition motifs, and non-conserved functional binding sites in mammalian regulatory regions [69]. The selection of an appropriate background sequence set with contextual and compositional genomic characteristics similar to those of the target sequence set is essential for accurate detection of overrepresented sites [70].

Given a variety of possible sequence biases in sequences, utilization of genomically matched background sequences improves the effectiveness to discover biologically relevant sites in the target [71, 72]. For example, the GC content might vary greatly among genomic sequences [73], and if not taken into account, can considerably bias the detection of GC-rich sequence motifs in target sequences. In this regard, shuffling of sequences with preservation of dinucleotide frequencies is a common approach to reducing GC content bias [74, 75, 76, 77]. However, since randomized sequences generated are not derived from genomic regions, they may not be appropriately adjusted for shape preferences [78], periodicity in AT/CG content within nucleosomal DNA [79] and other higher-level DNA sequence characteristics. Furthermore, the utilization of supervised methods that consider a semi-random selection of background sequences from the genome like HOMER [80] and BiasAway [81] usually perform better compared to dinucleotide shuffle methods.

3.1.6. F-Match

The F-Match method is usually applied to compare the number of sites identified in query (or target) sequences to those found in background sequences [25]. If a specific TF plays a biological role in the regulation of a certain set of target promoters, we can assume that the frequency of the corresponding binding sites identified in the target sequences will be considerably higher than predicted by random chance. The stringency of this TF's interaction with these sequences is often unknown, resulting in ambiguity when establishing score thresholds with the MATCH™ algorithm. F-Match examines sets of promoter sequences and attempts to identify two thresholds for each PWM: *th-min* and *th-max*. The minimum ratio between the frequency of binding sites in the target set (T) and background set (B) is provided by *th-min*, whereas the threshold *t-max* maximizes the same ratio. The

MATCH™ algorithm is first applied to both sets to determine the thresholds *th-min* and *th-max*. Hereby, the lowest PWM thresholds is used that correlates to the minimum of false negative rates. A set of predicted K and M sites in promoter sets T and B with the corresponding scores $\mathbb{MSS}$ is obtained for each PWM as a result (see [25] for more details):

$$\mathbb{MSS}_{T,1}, \mathbb{MSS}_{T,2}, ..., \mathbb{MSS}_{T,K}; \tag{3.1.6}$$

$$\mathbb{MSS}_{B,1}, \mathbb{MSS}_{B,2}, ..., \mathbb{MSS}_{B,M}. \tag{3.1.7}$$

F-Match performs an extensive search through all observed scores in sequence sets. The algorithm computes the number of sites k and m identified in T $(\mathbb{MSS}_{T,j} \geqslant th/j = 1, K)$ and B $(\mathbb{MSS}_{B,j} \geqslant th/j = 1, M)$, respectively, using each observed score as a threshold th. In the event of equal distribution of sites between both sets, the expected number of sites k_{exp} within set T can be calculated as

$$k_{\text{exp}} = n \cdot f, \tag{3.1.8}$$

where the parameters n and f are defined as

$$n = (k + m),$$

$$f = \frac{|T|}{|T| + |B|}.$$

The p-value of identifying the observed number of sites, which is higher for overrepresented matches $(k > k_{\text{exp}})$ and lower for underrepresented matches $(k < k_{\text{exp}})$ can be calculated under the assumption of a binomial distribution of sites between both sets. The two p-values are determined from Equations 3.1.9 and 3.1.10

$$\text{if } k < k_{\text{exp}}: \quad p(-) = \sum_{i=0}^{k} \binom{n}{i} f^i (1-f)^{n-i}, \tag{3.1.9}$$

$$\text{if } k > k_{\text{exp}}: \quad p(+) = \sum_{i=k}^{n} \binom{n}{i} f^i (1-f)^{n-i}, \tag{3.1.10}$$

where $p(-)$ and $p(+)$ define the p-values of under- and overrepresentation of matches in the promoter set T, respectively.

The F-Match algorithm determines thresholds th_{min} and th_{max} that minimize and maximize the ratio k/k_{exp} for a certain significance level ξ, respectively, given that the p-value is smaller than ξ $(p < \xi)$. If the necessary significance level for a particular PWM is not reached, the optimal threshold cannot be determined with F-Match.

3.2. Gene Expression Analysis

Using high-throughput methods such as RNA-Seq, thousands of mRNA transcripts may be analyzed simultaneously. Since mRNA translation yields all proteins in cells, mRNA expression levels are a reliable indicator of protein abundance. Nagalakshmi *et al.* [82] (2008) were the first to propose RNA-Seq for this task. In comparison to microarray, which has been used for differential gene expression analysis since the mid-1990s [83, 84, 85], RNA-Seq has several advantages, including reduced cross-hybridization artifacts, the ability to detect low-abundance transcripts without having to know the sequence in advance, and the ability to reduce excessive background noise in the experiment [86].

In RNA-seq, purified RNA from a cell or tissue is sheared and converted to cDNA before being sequenced on a platform such as Roche 454 or Illumina. Although the platforms' biochemistry and processing procedures differ, they all produce millions of short reads from one or both ends of each cDNA fragment. The RNA-seq experimental protocol consists of three parts [87]: (i) extraction of target RNAs; (ii) conversion of RNAs into cDNAs and binding of adapters to one or both ends of the cDNAs; and (iii) amplification using PCR and sequencing of cDNAs. Alternative splice junctions, mutations, *single nucleotide polymorphisms* (SNPs), gene fusions, and post-transcriptional changes are all detected using RNA-Seq [15]. Moreover, RNA-Seq allows researchers to assemble the whole transcriptome and find genes without having access to a genome.

The number of sequencing reads mapped to each gene (contig) is counted to measure gene expression (or transcript abundance). Raw read counts usually contain artifacts and errors. As a result, to get a credible gene expression level, the raw counts must be normalized. A few examples of common normalization methods are *reads per kilobase million* (RPKM) [88], *transcripts per million* (TPM) [15], DESeq [89], and *trimmed Mean of M values* (TMM) [90]. To reduce feature-length and library-size effects within samples, RPKM is utilized. TPM can be used to convert to RPKM, but they are better for reducing sample-to-sample errors [15]. *Differentially expressed genes* (DEGs) can be discovered by comparing the normalized gene expression levels between sample groups using statistical methods. In more detail, the key to determining if a difference in read count is meaningful in any given gene is to test for differential expression. If the difference is higher than what would be expected if it were solely due to natural random variation, it is considered significant [89].

3.3. Functional Categorization

Typically, gene expression data analysis leads to the identification of a reduced set of genes or proteins that have altered their status or abundance in response to a pathological or experimental condition, and therefore have gained a specific role in clinical application. However, viewing these genes as a simple list fails to give mechanistic insights into the biology under investigation. Selecting only familiar genes, on the other hand, tends to exclude less well-known genes from researchers' consideration.

As a result, a feasible next analytical step is to assess gene enrichment in biological processes, pathways, or disease phenotypes. The key aspect of coupling these genes is that it is usually easier and more relevant to investigate associations to known functions or functional categories given as a complete property rather than deducing roles of individual gene or the entire gene list from scratch [16].

3.3.1. The Gene Ontology (GO)

The GO is a freely available resource that offers a structured and controlled vocabulary for defining the functions of genes and gene products. The notion that homologous genes (paralogs and orthologs) usually have conserved functions across species was the driving force behind the first efforts of the GO consortium [91]. As a result, the integration of annotations from all organisms into a single central repository allowed for knowledge exchange and the deduction of functions for genes of interest.

An ontology is composed of defined terms connected by specific relationships between the terms. GO terms are categorized into the three main non-overlapping ontologies *cellular component* (CC), *biological process* (BP), and *molecular function* (MF) [91].
BP describes a set of molecular events with a defined start and end that are important for the proper functioning of living units (e.g., cells, tissues, and organs). Examples for BP are processes such as signal transduction and stem cell maintenance. MF describes a set of biochemical functions of a gene product. Examples for MF are functions such as catalysis and transporter activity. CC refers to the parts of a cell (intracellular and extracellular). Examples for CC are cell components such as the nucleus and cytoplasm.

The GO is structured as a *directed cyclic graph* (DAG), with nodes representing GO terms and edges representing relationships between terms. In addition, evidence from experiments, literature, databases, or computational methods are included in GO. The GO's generic and species-specific versions are updated on a regular basis. The GO is maintained on the web page (`www.geneontology.org`).

3.3.2. Functional Overrepresentation Analysis (ORA)

When investigating a subset of genes of interest to investigate if they are present in a set of relevant annotations, it might be challenging to comprehend the results based on raw numbers because a certain number of genes will be included in the annotations by random. GO ORA is an approach for assessing the statistical significance of the overlap of a set of genes of interest with the set of genes known to be associated with, e.g., a particular BP [16].

ORA's main assumption is that genes are independent and equally relevant in biological processes. Briefly, ORA finds relevant BPs by comparing the number of genes of interest found in a biological process against a background list, which is often the total number of genes present in the genome of interest. The Fisher exact test (or hypergeometric test) is the most widely used statistical test to assess whether a number of genes of interest are overrepresented in a gene set from an annotation database. Hypergeometric tests with multiple correction adjustments are usually used to obtain *p*-values assessing the significance of overrepresentation of biological processes [16].

3.4. Graph Search Algorithm

In this section, I will introduce some key definitions and concepts from the graph theory. A graph is the most fundamental mathematical method used to model a molecular pathway which consists of a set of *vertices* (also called points or nodes) as well as a set of *edges* (also called links or arcs). When an edge e is located between vertices a and b then a,b are adjacent (neighbors) to one another and incident with e. The vertices incident to e are defined as its end-vertices, whereby the number of edges with v as an end-vertex indicate the degree of a vertex v. A loop describes the edge with the same two end-vertices on both ends. In a bipartite graph the vertices are separated into two disjoint sets A and B so that every edge connects a node in both sets. The networks or graphs can be split into two categories: *directed* and *undirected graphs*, which are depicted in Figure 3.2.

Figure 3.2.: A directed (left) and an undirected graph (right), containing two vertices a and b, as well as one edge.

A directed graph consists of a set of vertices a and b connected by directed arcs which are generally illustrated with an arrow on the edge. In other words, we can think of an edge (ab) as connecting the starting link a with the terminal link b. Similarly, an undirected graph also comprises a vertex set as well as an edge set. However, in an undirected graph the edges are not associated with direction. Hence, rather than ordered pairs of vertices a and b, the elements of an edge set are essentially two-element subsets of a vertex. As in the case with directed graphs, the notation *ab* (or *ba* if the direction is unimportant) is used to define the edge $\{a,b\}$ in an undirected graph.

The shortest path problem in graph theory is the task of identifying a path between two nodes in a graph that minimizes the sum of the weights of its constituent edges [92]. Here, the vertices correspond to junctions, while the edges represent road segments, each of which is weighted by its length. Commonly, the problem of finding the shortest paths between vertices is solved by applying Dijkstra's algorithm [92]. It may be used to identify the shortest path from a single vertex to a single destination vertex, with the Dijkstra's algorithm stopping once the shortest path to the target vertex has been determined.

3.4.1. Master Regulator Analysis (MRA)

The MRA is integrated in the geneXplain platform (https://genexplain.com) and it enables a fast search for key molecules within a signal transduction network [25, 29, 30, 93]. A weighted graph defines the signal transduction network as $\mathbb{G} = (V,E,C)$ where the parameter V indicates the set of vertices (molecules in the TRANSPATH® database [31]), E is the set of arcs (reactions between molecules in TRANSPATH®), and C defines the cost function of a non-negative value for every arc $(C : E \to \boldsymbol{R}^+ \cup \{0\})$ [93]. The initial values of C for each direct reaction are set to 1.0 in the simplest setup of the algorithm. As a result, the cost of any path through subsequent reactions is equal to the number of reactions. A weighted graph is utilized to perform the method in the geneXplain platform, which encodes the direct or indirect reactions in TRANSPATH® into distinct edge weights (costs) [25].

MRA is based on the Dijkstra's shortest path algorithm (see [93] for more details). The key node search method begins with each molecule from the set of selected molecules, i.e. subset V_x of V, and builds the shortest path to all vertices i of V that are within a defined radius. After all vertices of V_x are evaluated, the algorithm computes the number of visits N_i for each vertice i of V. This corresponds to the number of selected molecules in the input set V_x that can transfer the signal to the vertice i. The values N_i can already be utilized as a ranking score for potential key nodes, i.e. to compute the MR score. A greater value of N_i increases the likelihood that the molecule i transmits its signal to the molecules in the input set V_x. However, using N_i directly as the MR score should be avoided since it will rank those molecules higher in score that are generally highly connected within the network, thus resulting in false positives (e.g., hub molecules). Consequently, the total number of molecule nodes is included in the whole network V that can be accessed from the molecule i in the score. The score $S(r)$ is calculated for each potential MR, and represents a certain balance between the distinction and sensitivity of signal transmission from the master node to the molecule. $S(r)$ is defined as

$$S(r) = \sum_{r=1}^{r_{\max}} \frac{V_r}{\left(1+\kappa \frac{N_r}{N_{\max,r}}\right) V_{\max,r}}, \tag{3.4.1}$$

where the parameter r indicates the radius of pathway steps from the master vertice to the molecule, V_r is the number of input molecules that receive a signal from the master vertice within r steps, and N_r defines the total number of all potential molecules in TRANSPATH® that have received a signal from the master node within r pathway steps. The maximum values among all potential MR vertices are indicated by $V_{\max,r}$ and $N_{\max,r}$ which enable to normalize the score to the $(0,1)$-interval. The higher the score in this interval, the more distinct and sensitive the MR is for the vertice set of input molecules. The parameter κ defines a penalty, with a value of 0.1 set by default.

3.4.2. TRANSPATH®

The TRANSPATH® database is one of the best scientifically conceptualized pathway libraries, ideal for a variety of applications due to its sophisticated structure. TRANSPATH® was created by E. Wingender and F. Schacherer as a part of the POET system's object-oriented database [31]. It is currently maintained by geneXplain GmbH (https://genexplain.com). TRANSPATH® stores information about more than 80,000 genes and 298,000 molecules involved in metabolic or signal transduction pathways of human, mouse, and rat. The information on molecules, genes and biochemical reactions is organized according to various hierarchies (Figure 3.3).

Individual reactions are meticulously documented with all details on the experiments, including the taxonomic origin of molecules as well as all reaction partners as stated in the published experiment ("molecular evidence level"). To establish a more thorough and complete overview, all evidences for a pathway step are assembled ("pathway step level"). The "semantic projection" focuses just on essential components whilst ignoring smaller abundant molecules and mechanistic details.

3.4.3. Upstream Analysis

DE analysis (Section 3.2), promoter analysis (Section 3.1.5), and MR search (Section 3.4.1) can be combined in a single workflow called *upstream analysis* to discover MR candidates that may control the activity of their regulated TFs, which in turn might regulate DEGs. Thus, MRs are potentially responsible for the gene expression changes observed in a comparative transcriptome study (Figure 3.4). The upstream analysis has been described earlier in [17].

This strategy involves three steps: (i) identification of DEGs; (ii) promoter analysis using F-Match (or other site overrepresentation methods); and (iii) TRANSPATH® MR search in pathways using the geneXplain platform.

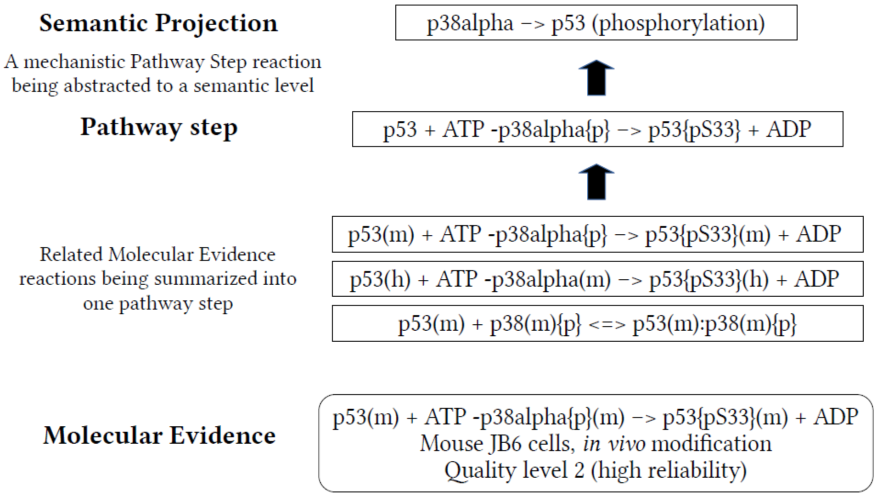

Semantic Projection
A mechanistic Pathway Step reaction being abstracted to a semantic level
p38alpha –> p53 (phosphorylation)
Pathway step
p53 + ATP -p38alpha{p} –> p53{pS33} + ADP
Related Molecular Evidence reactions being summarized into one pathway step
p53(m) + ATP -p38alpha{p} –> p53{pS33}(m) + ADP
p53(h) + ATP -p38alpha(m) –> p53{pS33}(h) + ADP
p53(m) + p38(m){p} <=> p53(m):p38(m){p}
Molecular Evidence
p53(m) + ATP -p38alpha{p}(m) –> p53{pS33}(m) + ADP
Mouse JB6 cells, *in vivo* modification
Quality level 2 (high reliability)

Figure 3.3.: The reaction hierarchy on molecular pathways in the TRANSPATH® database (Source: https://genexplain.com/transpath/).

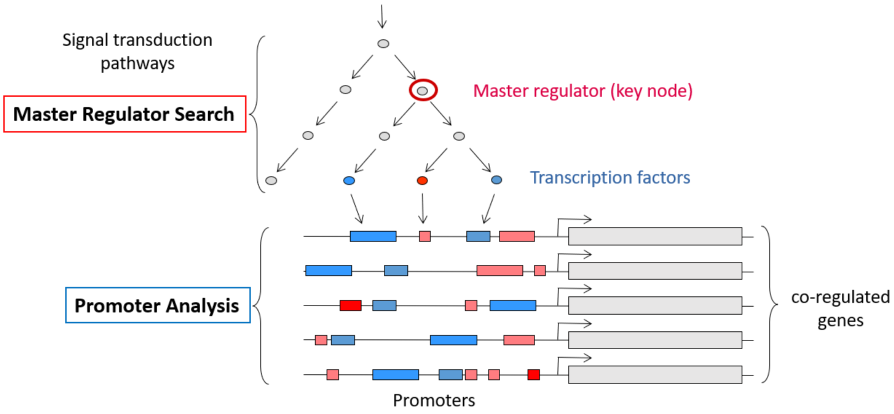


Figure 3.4.: Upstream analysis. Firstly, DE analysis is performed to identify DEGs. Secondly, promoter analysis is performed to find relevant TFs that are potential regulators of co-regulated genes. Thirdly, MR search is performed to find MR candidates in signal transduction pathways.

3.5. Supervised Learning

In this section, I will introduce different methods for supervised learning, which belong to a subcategory of ML. Supervised learning is defined by its application of labeled datasets to train algorithms for data classification or the accurate prediction of outcomes. It adjusts the weights of the input data until the model is well fitted, which occurs during the cross validation process. The training dataset comprises both inputs and correct outputs, allowing the model to learn over time. The algorithm of supervised learning quantifies the accuracy of the training dataset through a loss function, and it is adjusted until the error is substantially minimized. When it comes to data mining, supervised learning can be divided into two groups of problems – classification and regression. Classification employs an algorithm to correctly assign test data into specific categories. It distinguishes specific entities within the dataset, and makes conclusions on how these entities should be defined or labeled. Regression, on the other hand, is utilized to further understand the relationship between dependent and independent variables (also called features).

3.5.1. Random Forest

Random forest (RF) is a nonparametric technique developed by statistician Leo Breiman (2001) [36]. This predictive modeling approach has become a standard tool for data processing in the field of bioinformatics. RF is built on decision trees which represent a specific instance of the classification of data. It analyzes each instance separately, selecting the one with the majority of votes as the confirmed prediction. RF creates an ensemble of decision trees from bagged samples of the training data and random selections of variables. It has demonstrated excellent performance in instances where the number of variables exceeds the number of observations, and can handle complicated interaction structures as well as highly correlated values, and returns quantifications of variable or feature importance.

The principles of the RF algorithm are depicted in Figure 3.5. Each tree in the original RF approach [36] represents a standard classification or regression tree (CART) that selects the splitting predictor from a randomly selected subset of predictors, and employs the *decrease of Gini impurity* (DGI) as a splitting criterion [9]. The hyperparameter *mtry* indicates the number of randomly sampled variables as candidate predictors at each split. Each tree is created from a bootstrap sample chosen with replacement from the original dataset, and finally, the predictions of all trees are aggregated utilizing majority voting [9]. The *out-of-bag* (OOB) error is a key feature of RF. For some of the trees, each observation is an OOB observation, i.e. it was not utilized to build them. Thus OOB may be regarded as an internal validation dataset for these trees. In other words, the OOB error is essentially the average error rate acquired when the observations from the dataset are predicted using the trees for which they are out-of-bag.

Variable importance quantifications in RF are commonly used to rank variables according to their relevance to a classification problem, reducing the number of model inputs in

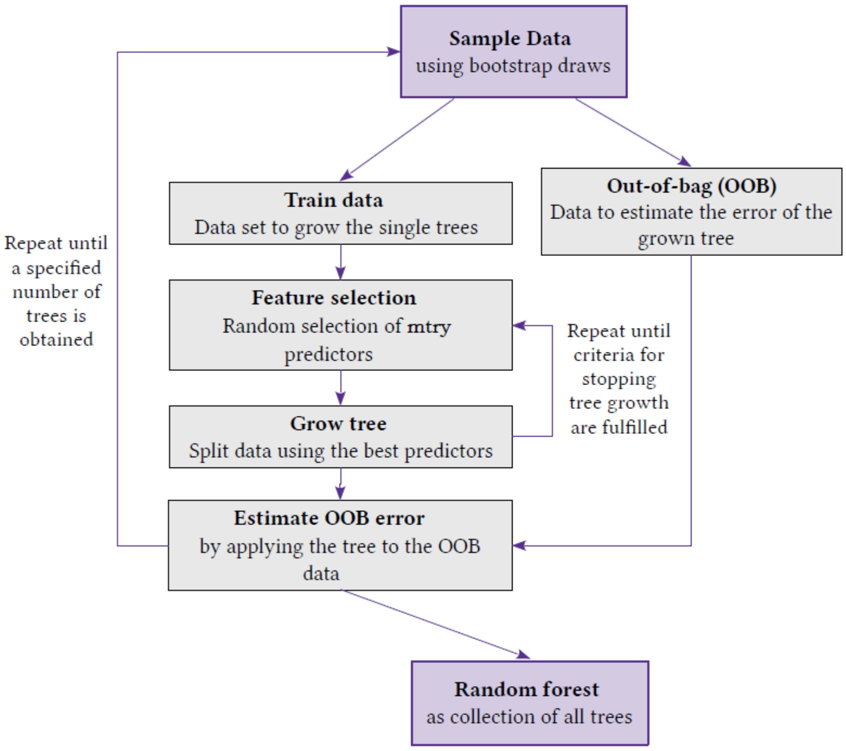

Sample Data
using bootstrap draws
Train data
Data set to grow the single trees
Out-of-bag (OOB)
Data to estimate the error of the grown tree
Repeat until a specified number of trees is obtained
Feature selection
Random selection of mtry predictors
Repeat until criteria for stopping tree growth are fulfilled
Grow tree
Split data using the best predictors
Estimate OOB error
by applying the tree to the OOB data
Random forest
as collection of all trees

Figure 3.5.: A schematic of the random forest algorithm. Inspired by [9].

datasets of high dimensionality, thus improving computational efficiency. The *mean decrease in Gini* (MDG) and the *mean decrease in accuracy* (MDA) are two widely used measures of variable importance. I will briefly explain only the method of MDA since it is of higher relevance to the thesis. MDA is known as the *permutation importance*. The calculation of MDA requires that the values of the variable are randomly permutated in the OOB samples to obtain quantification of the change in prediction accuracy. Variables of higher importance have a significant influence on the outcome values. Low-importance values, on the other hand, can be discarded from the model, making it easier as well as faster to fit and predict. Mean importance values of variables $\overline{VI}_p(i)$ can be calculated from the following equation

$$\overline{VI}_p(i) = \left(\sum_{j=1}^{i} VI_p(j) \right) / i, \tag{3.5.1}$$

where the parameter p is the predictor variable of interest, VI_p indicates the corresponding variable importance value for an individual run j, and $\overline{VI}_p(i)$ defines the mean importance variable over i runs [94].

3.5.2. Boruta Feature Selection Algorithm

The importance score in RF alone is insufficient to find significant correlations between the decision attribute and variables/features. This problem can be solved by applying the *Boruta* algorithm [44, 95] which provides selection criteria for important attributes. This method originates from the notion of RF, in which problems are solved by adding more randomness to the system [95]. The basic concept of Boruta consists of creating a randomized copy of the system, and merging it with the original to develop the classifier for this extended system. We compare the importance of the variable in the original system to that of the randomized variables to assess its importance. Only those variables with a greater importance than that of the randomized variables are considered important. The procedure of the Boruta feature selection algorithm is explained below (see [95] for more details).

- An extended system is created in which each descriptive variable is duplicated. The values of copied variables are then permuted randomly across objects, resulting in random correlations between the decision attribute and the duplicated variables.
- Several RF runs are performed. Before each run, the duplicated variables are randomized, and as a result, the random component of the system is different for each run of the RF.
- The importance of all attributes is calculated for each run.
- If the importance of the attribute is higher than the maximal importance of all randomized attributes for a single run then this attribute is considered important.
- A statistical test is performed for all attributes. The null hypothesis states that the variable's importance is equal to the *maximal importance of the random attributes* (MIRA). The test is a two-sided equality test, meaning the hypothesis can be rejected when the attribute's importance is significantly higher or lower than MIRA. We then count for each attribute the amount of times the attribute's importance was higher than MIRA (a hit is recorded for the variable). When the number of hits is substantially greater than the expected value, the variable is considered important (confirmed). However, when the number of hits is considerably lower than the expected value, the variable is deemed unimportant (rejected). The calculation of limits is straightforward for the confirming and rejecting variable for an user-defined number of iterations and confidence level.
- Unimportant variables are erased from the information system, often along with their randomized mirror pair. In some instances, randomized variables can be maintained in the system, which can reduce the number of important variables while preserving the accuracy of the classifier.
- The Boruta algorithm is repeated for an user-defined number of iterations, or until all attributes are either conclusively considered important or rejected, whichever occurs first.

3.6. Unsupervised Learning

Unsupervised learning algorithms are used in machine learning to cluster and analyze unlabeled and non-classified datasets. These algorithms are able to identify data groupings or previously unknown patterns with minimal human intervention. Unsupervised learning algorithms are commonly utilized for clustering, dimensionality reduction, and association problems. In this section, I will introduce the algorithms in unsupervised learning which I used in the thesis.

3.6.1. Principal Component Analysis

Principal component analysis (PCA) is an unsupervised learning algorithm used in ML to reduce dimensionality [96]. This non-parametric statistical technique makes use of orthogonal transformation to convert the observations of correlated features into a set of linearly uncorrelated data. The transformed features are called *principal components* (PCs).

The dimensionality reduction using PCA is performed as follows:

(i) **Standardizing the data**: Scaling the data so that all the variables and their values lie within a similar range.

(ii) **Generating the covariance matrix**: The correlation between the different variables in the data set is expressed by a covariance matrix. It is necessary to identify highly dependent variables since they include biased and redundant information, thus lowering the model's overall performance.

(iii) **Computing the eigenvectors and eigenvalues**: The mathematical components eigenvectors and eigenvalues must be calculated from the covariance matrix to determine the principal component of the data set.

(iv) **Calculating the PCs**: The PCs reduce and contain the majority of the valuable information distributed across the initial variables. The principal components are calculated by maintaining newly obtained variables highly significant and unrelated to one another. After computing the eigenvectors and eigenvalues, they are ordered descendingly in which the eigenvector with the highest eigenvalue is the most significant and thus forms the first PC. The dimensions of the data can be reduced by eliminating the less significant PCs. Finally, the feature matrix is constructed which comprises all the significant data variables with maximal information.

(v) **Reducing the dimensionality of the data set**: The original data is rearranged with the obtained PCs. The transpose of the original data set is multiplied by the transpose of the acquired feature vector to replace the original data axis with the previously determined PCs.

3.6.2. K-Means Clustering

K-means clustering is an unsupervised learning algorithm used in ML to solve clustering problems [97, 98]. This iterative algorithm partitions the n unlabeled dataset into k different clusters in which each data point belongs to the cluster with the nearest mean. The k-means clustering algorithm determines the correct value of k-center points or centroids, and assigns each data point to its nearest k-center. The data points closest to the specific k-center form a cluster. The k-means clustering technique comprises the following steps:

(i) Choosing the value for k clusters based off the dataset.
(ii) Assigning each data point to the k cluster with the closest centroid.
(iii) Recalculating the positions of the k centroids.
(iv) Repeating steps (ii) and (iii) until the centroids no longer shift. This generates a division of the data points into groups from which the metric to be minimized can be determined.

The aim of k-means is to minimize the squared error function, i.e., the total intra-cluster variance. The objective function, J, is defined as the sum of the *Euclidean distance* of each case i which is summed over k clusters. J can be computed from equation 3.6.1 where k indicates the number of clusters, n the number of cases, x_i defines the data point of the i'th case, and c_j is the centroid for cluster j. If x_i is assigned to cluster j $(x_i \in j)$ then the parameter ω_{ij} is equal to 1, otherwise $(x_i \notin j)$, $\omega_{ij} = 0$.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \omega_{ij} \left\| x_i - c_j \right\|^2 \tag{3.6.1}$$

The minimization of the objective function is performed in two parts of iterations. The goal is to keep minimizing the objective function after each complete iteration. Firstly, J is minimized with respect to ω_{ij}, and c_j is treated fixed (eq. 3.6.2). The total Euclidean distance can only be minimized when $\omega_{ij} = 1$ which occurs when k is equal to the distance function yielding a minimum. In this step the cluster assignments are updated.

$$\frac{\partial J}{\partial \omega_{ij}} = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i - c_j \right\|^2 \tag{3.6.2}$$

$$\Rightarrow \omega_{ij} = \begin{cases} 1, & \text{if } k = \text{argmin} \left\| x_i - c_j \right\|^2 \\ 0, & \text{otherwise.} \end{cases}$$

In the second step, J is minimized with respect to c_j, and the assigned data points (ω_{ij}) are treated fixed (eq. 3.6.3). To determine the c_j values for the minima, the gradient is set to be 0. In this step the centroids are recomputed after the cluster assignments from the first step.

$$\frac{\partial J}{\partial c_j} = 0 \tag{3.6.3}$$

3.6.3. Silhouette Method

The *silhouette* method is used to determine the optimal number of clusters, as well as to analyze and verify the consistency within data clusters [99, 100, 101]. This technique calculates the silhouette coefficients for each data point, which indicate how similarly a point corresponds to its own cluster (cohesion) relative to other clusters (separation). These results are represented concisely in a plot where the values of the silhouette coefficients range between [-1, 1]. The silhouette coefficients close to the value of +1 indicate that the object is well fitted to its own cluster and is far distant from neighboring clusters. The negative values reflect that the objects were assigned to the incorrect cluster, whereas a value of 0 shows that the object is on or near the decision border between the to neighboring clusters. If most samples have a high positive value, then the clustering configuration is suitable. The silhouette coefficient of an i'th point, $s(i)$, can be calculated from equation 3.6.4 where the parameters $a(i)$ and $b(i)$ indicate the average distance of the i'th point with all the other points in the same cluster and in the cluster nearest to the i'th cluster, respectively.

$$s(i) = \frac{b(i) - a(i)}{\max\left(b(i), a(i)\right)} \tag{3.6.4}$$

The silhouette analysis can be applied to calculate the hyperparameter, k, for the k-means clustering algorithm. After plotting the silhouette coefficients for each k-clustering model, the fluctuations and outliers of each cluster are observed.

3.7. Information Theory

In this section, I will introduce the most basic definitions in the information theory which is a study of communication and quantification of digital information. The notion of information is comprehensive and therefore difficult to define by a single entity. However, we define a measure called *entropy* for any probability distribution, which has many features that match the intuitive concept of what a quantity of of information should be.

3.7.1. Entropy

The notion of information entropy was first introduced by mathematician C. Shannon (1948) [102, 103]. The entropy is defined as a measure of the uncertainty of a random variable. Assume that X is a discrete random variable. After learning the value of X, the entropy of X quantifies the amount of the information acquired. In other words, the entropy of X gives an estimate for the amount of uncertainty about X before its value is known.

Entropy: *Let X be a discrete random variable with alphabet χ and probability function $p(x) = \mathrm{P}\{X = x\}, x \in \chi$, where the probability fulfills $0 \leqslant p(x) \leqslant 1$ and $\sum_{x\in\chi} p(x) = 1$. The Entropy $\mathbb{H}(X)$ of* variable *X is defined as*

$$\mathbb{H}(X) = -\sum_{x\in\chi} p(x) \log p(x). \tag{3.7.1}$$

In the equation 3.7.1, log is defined in the base of two ($\log_2$). If the realizations have a zero probability, $p(x) = 0$, then we obtain $p(x)\log p(x) = 0$. The equation 3.7.1 indicates that the calculation of the entropy $\mathbb{H}$ is dependent on the probability distribution of X, and has a maximum value when all probabilities $p(x_i)$ are equal.

Fundamental properties of entropy:

(i) *Non-negativity*: *For any $p(x)$, the Shannon entropy is always positive or equal to zero*:

$$\mathbb{H}(X) \geqslant 0 \tag{3.7.2}$$

(ii) *Upper bound*: *The entropy $\mathbb{H}(X)$ has a maximum value of* $\log(n)$ *for a random variable X with alphabet size n*:

$$\mathbb{H}(X) \leqslant \log(n). \tag{3.7.3}$$

(iii) *Conditioning decreases the entropy of a random variable X*:

$$\mathbb{H}(X) \geqslant \mathbb{H}(X|Y). \tag{3.7.4}$$

(iv) *Subadditivity*:

$$\mathbb{H}(X,Y) \leqslant \mathbb{H}(X) + \mathbb{H}(Y). \tag{3.7.5}$$

The equality can only be achieved when the random variables X and Y are independent.

(v) *The entropy is smaller in value or equal to the joint entropy*:

$$\mathbb{H}(X) \leqslant \mathbb{H}(X,Y). \tag{3.7.6}$$

The equality can only be obtained when the random variables X and Y are independent.

(vi) *The Shannon entropy $\mathbb{H}(X)$ is concave in the mass probability function $p(x)$.*

3.7.2. Relative Entropy

The relative entropy $\mathbb{KL}(p||q)$ measures the distance between two probability distributions $p(x)$ and $q(x)$ over the same alphabet χ. In other words, $\mathbb{KL}(p||q)$ measures how inefficient it is to assume that the distribution is q when p is the true distribution.

Kullback-Leibler divergence: *The relative entropy or Kullback-Leibner (KL) divergence $\mathbb{KL}(p||q)$ between two mass probability functions $p(x)$ and $q(x)$ with alphabet χ is defined as*

$$\begin{aligned} \mathbb{KL}(p||q) &= \sum_{x\in\chi} p(x)\log\frac{p(x)}{q(x)} \\ &= \mathbb{H}(X) - \sum_{x\in\chi} p(x)\log q(x). \end{aligned} \tag{3.7.7}$$

In the above equation 3.7.7, we employ the conventions of $0\log\frac{0}{0} = 0$ and (based on continuity arguments) of $0\log\frac{0}{q} = 0$ and $p\log\frac{p}{0} = \infty$ if $p(x) > 0$.

Fundamental properties of Kullback-Leibler divergence:

(i) *Non-negativity*: *The KL divergence can only have a positive value or be equal to zero*:

$$\mathbb{KL}(p||q) \geqslant 0. \tag{3.7.8}$$

The equality to zero, $\mathbb{KL}(p||q) = 0$, is only obtained if $p(x) = q(x)$.

(ii) *Asymmetry*: *If $p(x) \neq q(x)$ then the KL divergence is not symmetric*:

$$\mathbb{KL}(p||q) \neq \mathbb{KL}(q||p) \tag{3.7.9}$$

(iii) *The KL divergence does not satisfy the triangle inequality.*

Jensen-Shannon divergence: *The Jensen-Shannon (JS) divergence between two (or more) mass probability functions is a symmetrized, bounded, and smoothed variant of the*

Kullback-Leibler divergence. The JS divergence $\mathbb{JS}(p||q)$ between two probability functions $p(x)$ and $q(x)$ with weights π_1 and π_2 is defined as

$$\mathbb{JS}(p||q) = \pi_1 \mathbb{KL}\left(p(x)||\frac{p(x)+q(x)}{2}\right) + \pi_2 \mathbb{KL}\left(q(x)||\frac{p(x)+q(x)}{2}\right), \tag{3.7.10}$$

where π_1 and π_2 fulfill the limitations $\pi_1 + \pi_2 = 1$, and $0 \leqslant \pi_i \leqslant 1$.

Another form of the *JS* divergence is given in equation 3.7.11 where $\mathbb{H}(p) = \sum_i p_i \log p_i$ defines the Shannon entropy and $\pi_1 = \pi_2 = \frac{1}{2}$.

$$\mathbb{JS}(p||q) = \mathbb{H}\left(\frac{p+q}{2}\right) - \frac{1}{2}\mathbb{H}(p) - \frac{1}{2}\mathbb{H}(q) \tag{3.7.11}$$

Fundamental properties of Jensen-Shannon divergence:

(i) *Non-negativity*: *The JS (Jensen-Shannon) divergence of two probability distributions $p(x)$ and $q(x)$ is always either positive or equal to zero*:

$$\mathbb{JS}(p||q) \geqslant 0. \tag{3.7.12}$$

The equality to zero, $\mathbb{JS}(p||q) = 0$, is acquired only if $p = q$.

(ii) *Symmetry*: *The JS divergence is symmetric*:

$$\mathbb{JS}(p||q) = \mathbb{JS}(q||p). \tag{3.7.13}$$

(iii) *The values of the JS divergence of $p(x)$ and $q(x)$ are bounded between* 0 *and* 1:

$$0 \leqslant \mathbb{JS}(p||q) \leqslant 1. \tag{3.7.14}$$

(iv) *The JS divergence does not satisfy the triangle inequality whereas the square root of the JS divergence, $\sqrt{\mathbb{JS}(p||q)}$, does.*

(v) *To measure the difference between more than two mass probability functions $p_1, p_2, ..., p_n$ with weights $\pi_1, \pi_2, ..., \pi_n$, the JS divergence is brought to a more generalized form*:

$$\mathbb{JS}(p_1, p_2, ..., p_n) = \mathbb{H}\left[\sum_{i=1}^{n} \pi_i p_i\right] - \left[\sum_{i=1}^{n} \pi_i \mathbb{H}[p_i]\right], \tag{3.7.15}$$

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

4. Materials and Methods

4.1. Bioinformatics Pipeline

The bioinformatics pipeline presented in the thesis consists of five different theme-oriented workflows. In the following, I will go through each step of the pipeline's workflows (Sections 4.1.1 - 4.1.5) and present their functionalities. Finally, I will discuss the reference datasets that were used to evaluate the performance of the pipeline (Section 4.1.7).

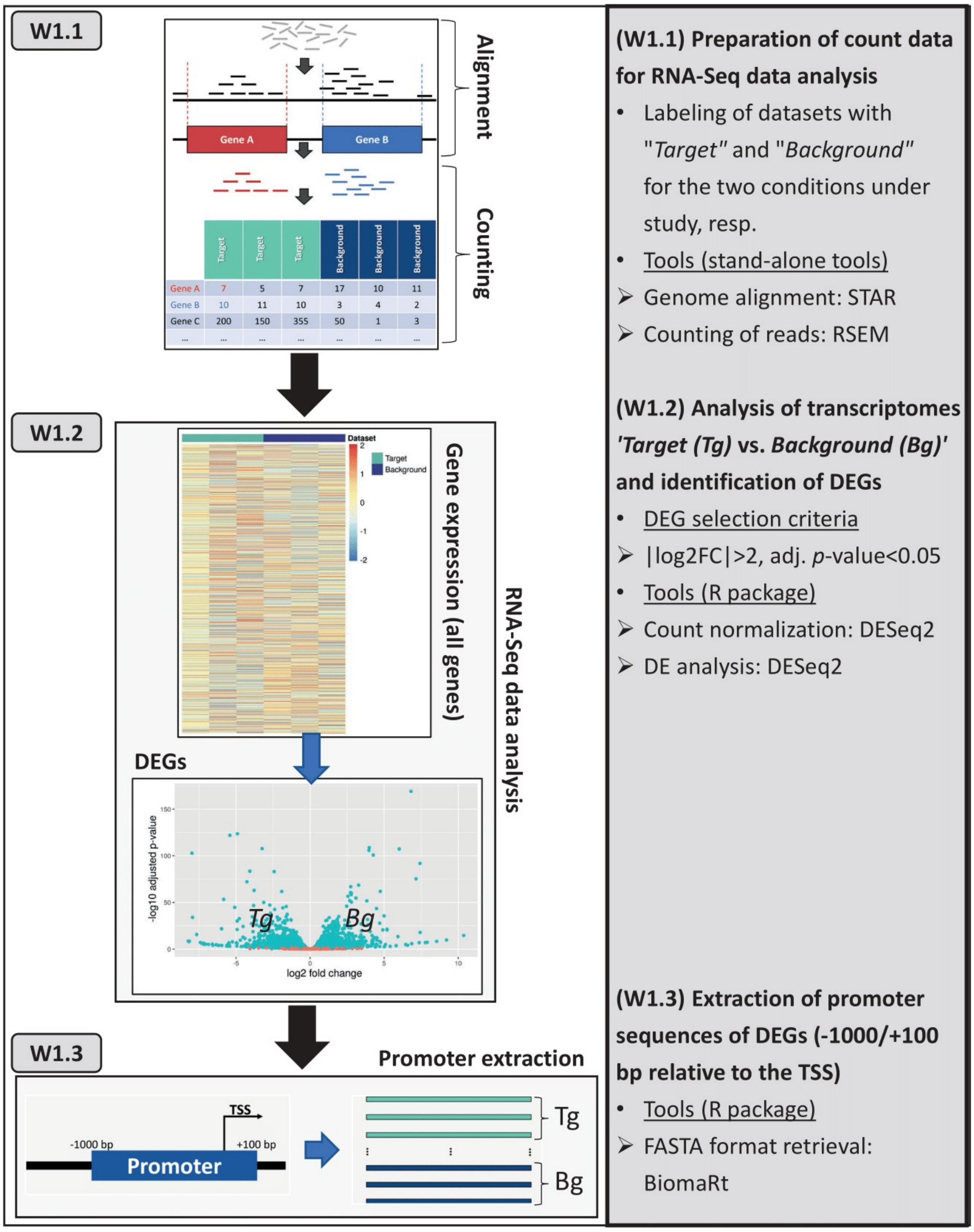
4.1.1. Workflow 1 – RNA-Seq Data Analysis and Promoter Sequence Extraction

Workflow 1 is dedicated to the analysis of RNA-Seq data and the extraction of promoter sequences of DEGs (Figure 4.1). RNA-Seq data analysis enables both the comparison of transcriptomes across different conditions (e.g., phenotypes, disease states, treatment groups etc.) and the identification of DEGs whose observed expression patterns differ between respective conditions.

The first step (*W1.1*) summarizes the procedure that is used to obtain RNA-Seq count data from FASTQ files [104], which involves two initial pre-processing steps: (i) mapping of sequencing reads against a reference genome using STAR (*v2.7.9a*) [105]; and (ii) counting of mapped reads using RSEM (*v1.3.3*) [106]. For simplicity, the two conditions of interest (here referred to as datasets) are labeled "*Target*" and "*Background*". *Target* corresponds to the dataset of greater interest (e.g., samples from drug-treated cell lines, highly invasive cell lines, poor-outcome patients etc.), whereas *Background* corresponds to the control dataset (e.g., samples from non-treated cell lines, non-invasive cell lines, good-outcome patients etc.).

Once obtained after mapping and counting, gene-level RSEM count data is provided to the second step (*W.1.2*) where the count data is variance stabilized and corrected for size factors and normalization factors using the Bioconductor package DESeq2 [107] (*v1.32.0*, `https://bioconductor.org/packages/DESeq2/`). The normalized data is then used to visualize the expression patterns of all genes across *Target* and *Background*. Thereafter, significantly upregulated DEGs ($|\log 2 \text{ fold change}| > 2, \text{adj. } p\text{-value} < 0.05$) in the *Target* and *Background* are identified using DESeq2. Normalization and DE analysis are performed as outlined in [108]). Briefly, DEGs between *Target* and *Background* are identified using the DESeq2 *results* function with the explicit parameters: $alpha = 0.05$ and $lfcThreshold = \log 2\ (1.5)$ where *alpha* and *lfcThreshold* specify the significance cutoff for adjusted p-values and the $\log 2$ fold change cutoff, respectively. Fold changes are subsequently shrunk using the *lfcShrink* function in the Bioconductor package apeglm [109] (*v 1.14.0*, https://bioconductor.org/packages/apeglm/). All other parameters of the two functions *results* and *lfcShrink* are left at their defaults. The principles underlying RNA-Seq data analysis including DE analysis can be found in Section 3.2.

The last step (*W1.3*) describes the retrieval of promoter sequences that correspond to DEGs. Promoters are here defined as the regions of 1000 (-) base pairs (bp) upstream and 100 (+) bp downstream relative to annotated TSSs of genes. The extraction of the corresponding -1000/+100 bp promoter sequences in FASTA format is performed based on the genome builds hg38 (human) or mm10 (mouse) in the Ensembl genome database [110] using the Bioconductor package BiomaRt [111] (*v2.48.3*, https://bioconductor.org/packages/biomaRt/). In more detail, promoters of genes are retrieved using the *getSequence* function in BiomaRt with the explicit parameters: *type* = "ensembl_gene_-id", *upstream* = 1000, *downstream* = 100, and *seqType* = "coding_gene_flank". All other parameters of the *getSequence* function are left at their defaults.



	Target	Target	Target	Background	Background	Background
Gene A	7	5	7	17	10	11
Gene B	10	11	10	3	4	2
Gene C	200	150	355	50	1	3
...	...	...	...	...	...	...

(W1.1) Preparation of count data for RNA-Seq data analysis

- Labeling of datasets with "*Target*" and "*Background*" for the two conditions under study, resp.
- Tools (stand-alone tools)

➢ Genome alignment: STAR

➢ Counting of reads: RSEM

(W1.2) Analysis of transcriptomes '*Target (Tg)* vs. *Background (Bg)*' and identification of DEGs

- DEG selection criteria

➢ |log2FC|>2, adj. *p*-value<0.05

- Tools (R package)

➢ Count normalization: DESeq2

➢ DE analysis: DESeq2

(W1.3) Extraction of promoter sequences of DEGs (-1000/+100 bp relative to the TSS)

- Tools (R package)

➢ FASTA format retrieval: BiomaRt

Figure 4.1.: Workflow 1. RNA-Seq data analysis and extraction of promoter regions of DEGs. Details on data processing steps including tools are provided in the grey box.

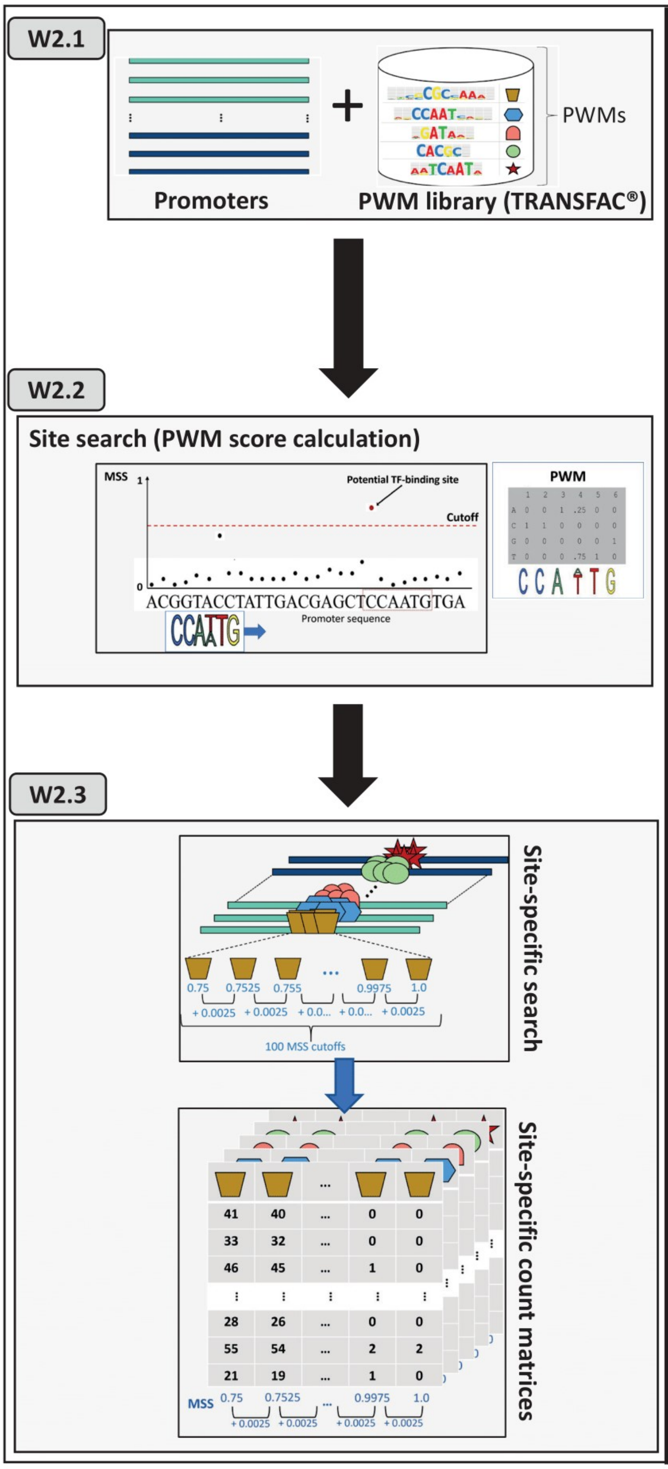
4.1.2. Workflow 2 – TFBS search using the MATCH™ Algorithm

Workflow 2 is dedicated to the search for potential TFBSs and the construction of so-called *site count matrices* which summarize the results from a site search (Figure 4.2).

The first step (*W2.1*) describes the preparation of input data required to perform the site search, which involves the selection of promoter sequences in *Target* and *Background*, as well as the selection of a PWM library. The selection of promoter sequences is performed in a way that if the number of sequences in *Target* and *Background* differ, the larger of the two sets (sorted by decreasing log 2 fold changes of DEGs) is truncated at the bottom to match the number of sequences in the smaller set, ultimately obtaining a 50:50 ratio of sequences in both sets. A library of 163 non-redundant vertebrate PWMs from TRANSFAC® (release *2019.3*) is selected as the default library to grant an optimal tradeoff between performance and run time for any sequence sets under study. The PWM library and sequence logos can be found in Table B1.

In the second step (*W2.2*), the MATCH™ algorithm (command line *v1.0*, `https://genexplain.com/`) is applied to the promoter sequences using the specified TRANSFAC® PWM library with all MSS cutoffs set to an initial value of 0.75, thereby disregarding all matches which score a similarity of less than 0.75 (75% similarity) during site search. The CSS cutoffs defined in the TRANSFAC® minFN profile are left at their default values. The output of MATCH™ ultimately stores the combined results from the MATCH™ run involving predictions for all 163 PWMs. In Sections 3.1.1, 3.1.2, and 3.1.3, further information on PWMs, the TRANSFAC® database, and the MATCH™ algorithm is provided, respectively.

The third step (*W2.3*) describes the processing of the combined MATCH™ results to construct 163 site-specific count matrices, each of which stores site predictions for different PWMs. More information on the general definition and construction of a site count matrix can be found in Section 3.1.4. Each site-specific count matrix contains the site predictions for a specific PWM, with each column in a matrix characterizing the site frequency distribution acquired above a distinct MSS cutoff, ranging from 0.75 (the first column) to 1 (the last column) by an increment of 0.0025.



(W2.1) Preparation of input data for TF-binding site search

- Sequence set size adjustment
- ➢ Adjust set sizes to 50:50 (*Target*:*Background*)
- PWM library selection:
- ➢ Select 163 non-redundant PWMs from TRANSFAC® (2019.3) as the default library

(W2.2) TF-binding site search

- Set all MSS cutoffs to an initial value of 0.75
- Run site search and store combined results from site search involving predictions using all 163 PWMs
- Tools (stand-alone tool)
- ➢ MATCH™

(W2.3) Construction of site-specific count matrices

- Create separate outputs for site-specific predictions
- Count number of site occurrences within each promoter sequence that are above different MSS cutoffs (*N*=100) ranging from 0.75-1 by an increment of 0.0025
- Construct site-specific count matrices where the rows and columns correspond to promoters and number of site occurrences observed above a specified MSS cutoff, resp.

Figure 4.2.: Workflow 2. TFBS search and construction of site-specific count matrices. Details on data processing steps including tools are provided in the grey box.

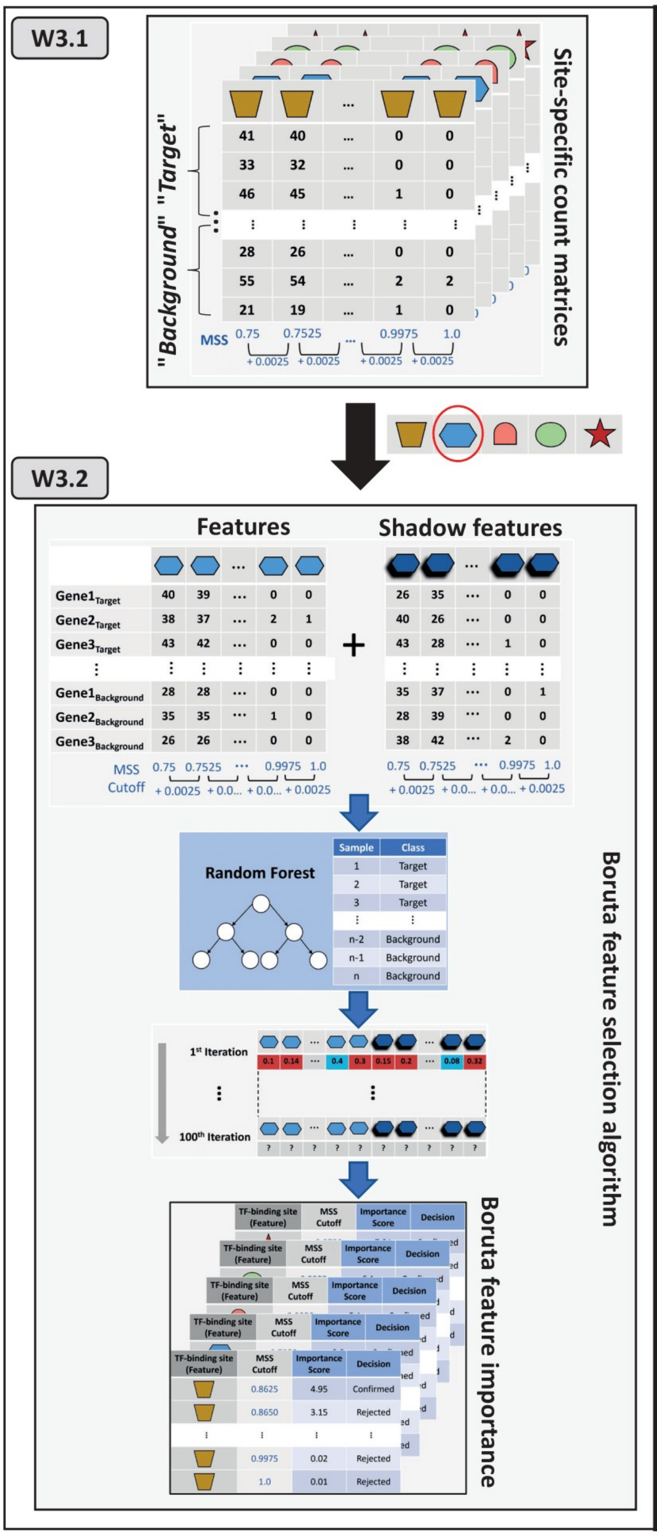
4.1.3. Workflow 3 — MSS Cutoff Optimization using Feature Selection

Workflow 3 is dedicated to the optimization of MSS cutoffs using the Boruta feature selection algorithm. The objective of this workflow is to find the MSS cutoff for each PWM at which the site frequency distributions observed discriminate between *Target* and *Background* (Figure 4.3). Consequently, if none of the 100 different MSS cutoffs in a site-specific count matrix falls under this criteria, the respective PWM will be eliminated.

In the first step (*W3.1*), site-specific matrices are prepared by introducing the two class labels "*Target*" and "*Background*" for promoters that belong to *Target* and *Background*, respectively. Subsequently, count matrices are provided one by one (per PWM) to the Boruta algorithm for feature selection in the second step (*W3.2*). The Boruta algorithm is deployed using the R package Boruta [44] (*v7.0.0*, https://CRAN.R-project.org/package=Boruta) and with *maxRuns* set to 100 (the default value) where *maxRuns* correspond to the number of iterations. All other parameters are left at their defaults.

Boruta considers each column label (i.e., PWM at specified MSS cutoff) of a site-specific matrix as unique feature identifiers which can be selected (confirmed) or not (rejected) by the feature selection algorithm. Noteworthy, it is possible that Boruta may select multiple features at the same time, if any at all. The Boruta algorithm starts by duplicating an original site-specific matrix and randomly shuffles the count values in each column of the duplicate to create a second matrix with shadow features. Subsequently, both matrices are attached to each other and Boruta trains a random forest classifier on the combined dataset. The Boruta algorithm is repeated up to 100 times (the predefined maximal number of iterations performed) and, at each iteration, it calculates an importance score for each feature. Thereafter, it examines each of the original features to check if they are more important than their shadow features. This procedure may result in a *hit* stored in a vector that reflects whether a feature is doing better than the best shadow feature. The algorithm can stop after any number of iterations, or when all the features have been confirmed or rejected. Boruta's decision is made by using a binomial distribution to compare the number of times a feature outperformed the shadow features. Finally, the algorithm generates a results table per PWM which contains the median importance score for each feature as well as the decision for each MSS cutoff.

Random forest and Boruta are described in more detail in Sections 3.5.1 and 3.5.2, respectively.



(W3.1) Preparation of site-specific matrices as input data for feature selection

- Introduce class labels *"Target"* and *"Background"* for the two promoter sets, resp.

(W3.2) Boruta feature selection algorithm (site-specific setup)

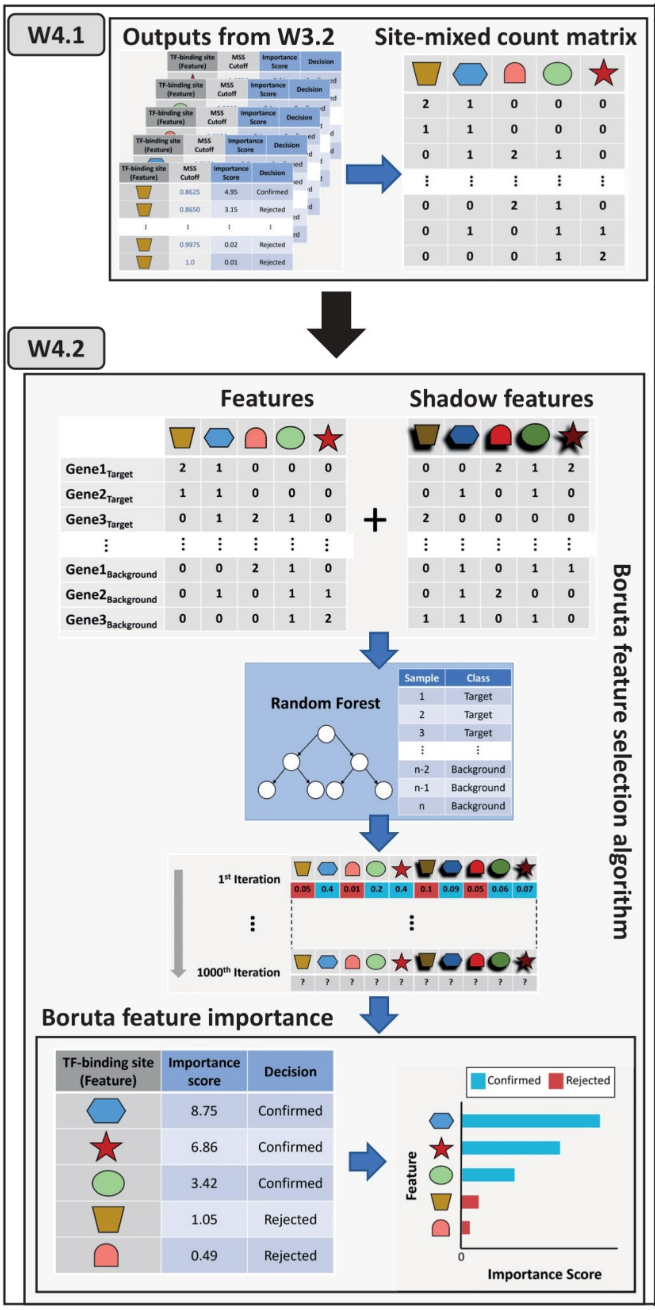
- Setup for 163 Boruta runs
  - ➢ Boruta creates a second count matrix with shadow features by randomly shuffling each feature
  - ➢ Count matrices are attached to each other and the Random Forest algorithm is deployed on the combined dataset
  - ➢ Boruta's algorithm is repeated up to 100 times (predefined max. number of iterations) and, at each iteration, it calculates importance scores to check if the original features are more important than their shadow features, resulting in a 'hit' stored in a vector
  - ➢ Boruta’s decision is made by using a binomial distribution to compare the number of times a feature outperformed the shadow features
  - ➢ Boruta creates a results table per PWM (*N*=163) that contains the median importance score per feature and the decision, i.e., 'confirmed' or 'rejected'
- Tools
  - ➢ Boruta (R package)

Figure 4.3.: Workflow 3. Matrix similarity score (MSS) cutoff optimization using Boruta feature selection algorithm. Details on data processing steps including tools are provided in the grey box.

4.1.4. Workflow 4 – Identifying the most relevant Site Models (PWMs) using Feature Selection

Workflow 4 is dedicated to the identification of the most relevant site models according to the Boruta feature selection algorithm (Figure 4.4). In this regard, Boruta is now being applied to a single site-mixed count matrix using optimized MSS cutoff values.

In the first step of *Workflow 4*, the site-mixed count matrix is built from the confirmed features with the highest median importance score (per PWM at optimized MSS cutoff value) obtained from *Workflow 3* (*W4.1*). The most relevant features from the site-mixed matrix are then selected using the Boruta feature selection algorithm as outlined in *W3.2*, with the exception that the maximum number of iterations (*maxRuns*) is increased to 1000 (*W4.2*). Finally, Boruta generates a results table with the median importance score for each feature and the decision.



(W4.1) Preparation of a site-mixed count matrix as input data for feature selection

- Keep only one confirmed feature (per PWM) with the highest median importance score observed in the final outputs from W3.2
- Build a site-mixed count matrix from all kept features across different site models

(W4.2) Boruta feature selection algorithm (site-mixed setup)

- Setup for a final Boruta run
  - ➢ Run Boruta feature selection algorithm as outlined in *Workflow 3* but, this time, using the site-mixed count matrix as the input data
  - ➢ Set number of max. iterations to 1000 and run Boruta
  - ➢ Boruta creates a final results table with the median importance score for each feature and the decision, i.e., 'confirmed' or 'rejected'
- Tools
  - ➢ Boruta (R package)

Figure 4.4.: Workflow 4. Identifying the most relevant site models (the PWMs) using Boruta feature selection algorithm. Details on data processing steps including tools are provided in the grey box.

4.1.5. Workflow 5 – Identification of Gene Clusters and Dysregulated Molecular Pathways

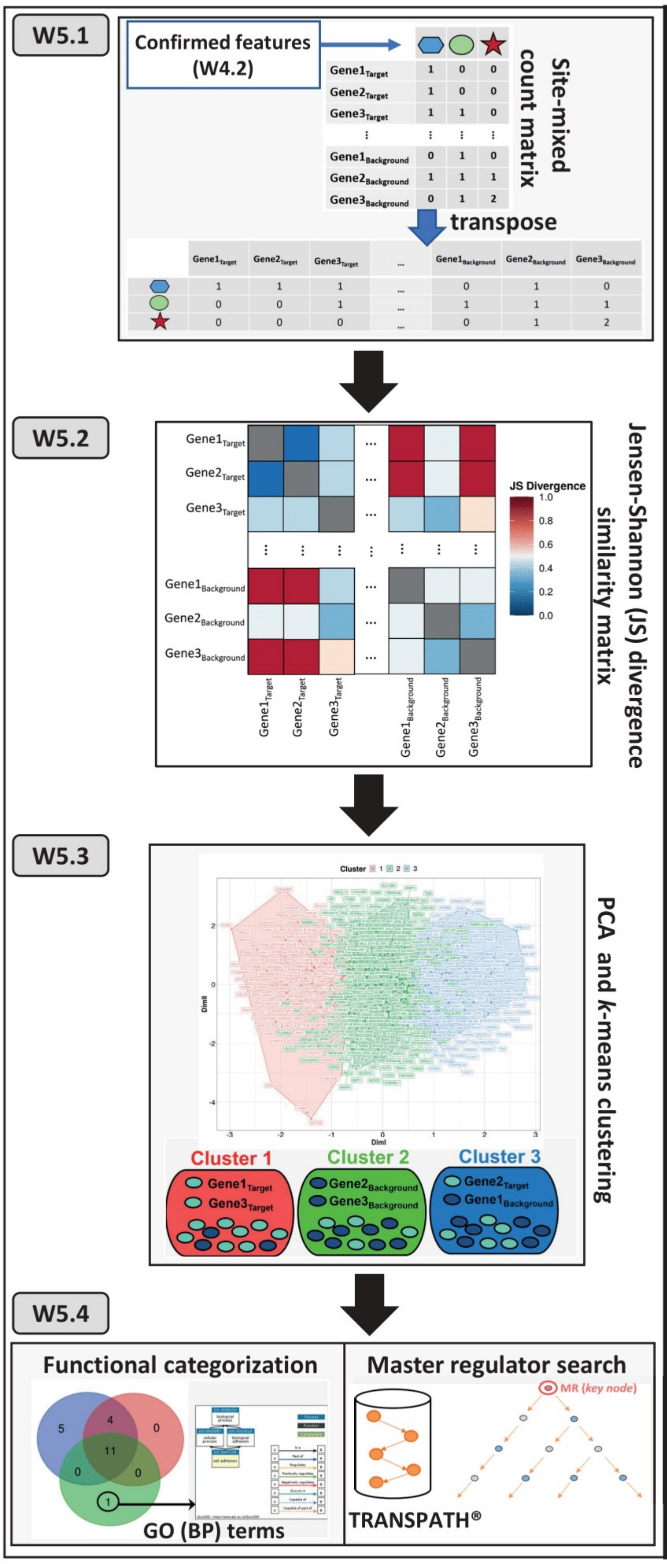
Workflow 5 is dedicated to the identification of gene clusters consisting of any possible combination of DEGs (defined in *Workflow 1*) across *Target* and *Background* (Figure 4.5). Following clustering, this workflow is also used to identify relevant biological processes and MRs that may regulate the activity of these processes.

In the first step of *Workflow 5*, input data is prepared by retaining all confirmed features in the site-mixed matrix acquired from the final results table in step *W4.2* (Figure 4.4) and transposing the matrix (*W5.1*). The JSD is then used to calculate similarities between probability distributions (i.e., site frequency distributions as probability distributions), leading to the creation of a JSD similarity matrix (*W5.2*). The theory underlying the JSD is provided in Section 3.7.

In the next step, PCA is applied to the similarity matrix to perform a dimensionality reduction step, thereby improving the clustering (*W5.3*). The theory underlying PCA is provided in Section 3.6.1. Thereafter, k-means clustering is applied on top of PCA to investigate the relationship in the PCA projection, yielding a visualization of the gene clusters on a two-dimensional plane by selecting the principal components that account for the most variability as the axes. In addition, the silhouette method is utilized to determine the optimal number of k clusters which is supplied to the k-means algorithm. The theory underlying k-means clustering and the silhouette method are provided in the Sections 3.6.2 and 3.6.3, respectively. The construction of the JSD similarity matrix, PCA, and k-means clustering are performed using the R package immunarch (*v0.6.6*, `https://CRAN.R-project.org/package=immunarch`). In more detail, gene clusters are obtained using the *geneUsageAnalysis* function in the package immunarch with the explicit parameters: *method* = "js+pca+kmeans", *do.norm* = TRUE, and *laplace* = 1e-12. The parameter *method* specifies the stream of analyses to be performed and *do.norm* specifies if the data should be normalized using Laplace smoothing [112] to acquire probability distributions. The parameter *laplace* specifies the pseudocount which is used for Laplace smoothing, and which will be added to every count value in the matrix. All other parameters of the function are left at their defaults except for the number of k that was set to the optimal number of clusters provided by the silhouette method.

After clustering, each gene cluster is first subjected to functional ORA to categorize genes using the annotations in the category BP of the GO database. In this regard, ORA is used to determine whether gene sets associated with a specific BP term are statistically overrepresented in the identified gene clusters. Overrepresented BP terms are retrieved for each cluster and compared between clusters using the Bioconductor package clusterProfiler [113] (*4.0.5*, `https://bioconductor.org/packages/clusterProfiler/`). To this end, the *compareCluster* function in the package clusterProfiler is used with the explicit parameters: *fun* = "enrichGO", *ont* = "BP”, and *OrgDb* = org.Mm.eg.db (or org.Hs.eg.db) where *fun* is set to perform ORA using the BP annotations, which is specified in the parameter *ont*. The parameter *OrgDb* is either set to org.Mm.eg.db or org.Hs.eg.db to retrieve genome wide annotations for mouse or human, respectively. In addition, the *simplify* function in clusterProfiler is used with default parameter settings to remove redundant overrepresented BP terms from the results. All other parameters are left at their defaults and only the top 10 most significant BP terms are reported for each gene cluster after calculating adjusted *p*-values following the *Benjamini-Hochberg* (BH) method [114]. More information on the definition of the Gene Ontology, its annotations, and their utilization in the context of ORA are provided in the Sections 3.3.1 and 3.3.2, respectively.

Thereafter, MRs are searched upstream of the genes in each cluster, utilizing an analysis workflow of the geneXplain platform (web edition 6.3, `https://genexplain.gwdg.de/bioumlweb/`). MRs are molecules that reside at the top of the regulatory hierarchy in signal transduction pathways, with the ability to potentially influence genes at several levels in a pathway. MRs are found for a set of genes in each cluster using a modified shortest path algorithm that examines the graph for common regulators or key nodes utilizing the TRANSPATH® database's pathway knowledge. To this end, the MR search workflow ("*Regulator search*") of the geneXplain is applied to the clusters with the FDR cutoff set to 0.05 and with the maximum search radius set to 10 steps upstream of genes. All other parameters of the master regulator search algorithm are left at their defaults. Further information on the TRANSPATH® database and the master regulator search algorithm are provided in the Sections 3.4.1 and 3.4.2, respectively.



	⬢	●	★
$\text{Gene1}_{\text{Target}}$	1	0	0
$\text{Gene2}_{\text{Target}}$	1	0	0
$\text{Gene3}_{\text{Target}}$	1	1	0
⋮	⋮	⋮	⋮
$\text{Gene1}_{\text{Background}}$	0	1	0
$\text{Gene2}_{\text{Background}}$	1	1	1
$\text{Gene3}_{\text{Background}}$	0	1	2

	$\text{Gene1}_{\text{Target}}$	$\text{Gene2}_{\text{Target}}$	$\text{Gene3}_{\text{Target}}$	...	$\text{Gene1}_{\text{Background}}$	$\text{Gene2}_{\text{Background}}$	$\text{Gene3}_{\text{Background}}$
⬢	1	1	1	...	0	1	0
●	0	0	1	...	1	1	1
★	0	0	0	...	0	1	2

(W5.1) Preparation of a transposed site-mixed count matrix as input data for gene clustering

- Keep only confirmed features obtained from W4.2 in site-mixed count matrix and transpose the matrix

(W5.2) Construction of a Jensen-Shannon divergence (JSD) square similarity matrix

- Calculate JSD values (column-wise) between distributions
- Tools
- ➢ immunarch (R package)

(W5.3) Clustering of genes

- Perform dimensionality reduction using PCA based on JSD matrix
- Perform *k*-means on top of PCA
- Tools
- ➢ immunarch (R package)

(W5.4) Functional categorization and master regulator analysis of gene clusters

- Perform over-representation analysis (ORA) using the Gene Ontology (GO)
- Retrieve significant GO BP (Biological Process) terms
- Search for master regulators (10 steps upstream) using the TRANSPATH® database
- Tools
- ➢ clusterProfiler (R package)
- ➢ geneXplain 6.3 web edition

Figure 4.5.: Workflow 5. Identification of gene clusters, their functional categorization into biological processes, as well as their upstream master regulators. Details on data processing steps including tools are provided in the grey box.

4.1.6. Statistical Analysis and Visualizations

Unless otherwise stated, analyses were performed using the free statistical software R (*v3.6.2*, http://www.R-project.org/). Heatmaps were visualized with the R package pheatmap (*v1.0.12*, https://CRAN.R-project.org/package=pheatmap). All plots relating to Jensen-Shannon divergence, PCA, and k-means were visualized with the R package immunarch (*v0.6.6*, https://CRAN.R-project.org/package=immunarch). Venn diagrams were visualized using a webtool available at http://bioinformatics.psb.ugent.be/webtools/Venn/. Logo plots and master regulatory networks were visualized using the geneXplain platform (web edition 6.3, https://genexplain.gwdg.de/bioumlweb/) workflows "*Create matrix logo*" and "*Visualize results*", respectively. All other plots were visualized with the R package ggplot2 (*v3.3.5*, https://CRAN.R-project.org/package=ggplot2).

4.1.7. Reference Datasets

I evaluated the performance of the pipeline by applying it to three reference data sets which are presented below. All datasets are associated with activity of specific signaling pathways in cancer.

1. NF-κB pathway-related dataset
The first dataset corresponds to a NF-κB pathway-related gene set that was derived from microarray data analysis described in [19]. This gene set comprises significantly downregulated genes in HUVEC cells stimulated with Interleukin-1β (IL-1β) after treatment with an NF-κB inhibitor in comparison to IL-1β stimulated cells. The rationale behind this experiment is that treatment of HUVEC cells with Interleukin-1 (IL-1) produces an inflammatory response, which is manifested as an increase in mRNA expression. Most importantly, the underlying biological mechanism behind this response can be modulated by the inhibition of the NF-κB signaling pathway. Significantly downregulated genes were directly obtained from the study and defined as the target set.

2. WNT pathway-related CRC dataset
The second dataset was derived from an RNA-Seq study which involves the comparison of two murine CRC cell lines, namely CMT-93 and 1638N-T1 [17]. We have previously analyzed this dataset by applying the upstream analysis strategy to identify important regulators implicated in the canonical Wnt/β-catenin signaling pathway in CRC. Significantly upregulated genes in 1638N-T1 were defined as the target set due to specific predisposing mutations in key genes of the WNT signaling pathway. Likewise, significantly upregulated genes in CMT-93 were defined as the background set.

3. EGFR/WNT pathway-related CRLM dataset

The third dataset was derived from an unpublished RNA-Seq study which compared sample groups related to two patients with *colorectal cancer liver metastases* (CRLM) [115]. We have previously analyzed this dataset by applying the master regulator analysis to identify important regulators implicated in the EGFR/WNT signaling pathways. Significantly upregulated genes for poor-outcome patient I were defined as the target set. Likewise, significantly upregulated genes for good-outcome patient II were defined as the background set.

5. Results

In this chapter, I will evaluate the performance and functionality of the bioinformatics pipeline. A detailed description of its components and implementation is provided in Section 4.1. The PWM library and sequence logos used for the analyses can be found in Table B1. Furthermore, the reference datasets used for the evalutation are provided in Section 4.1.7.

5.1. Analysis of the NF-κB Pathway-related Dataset

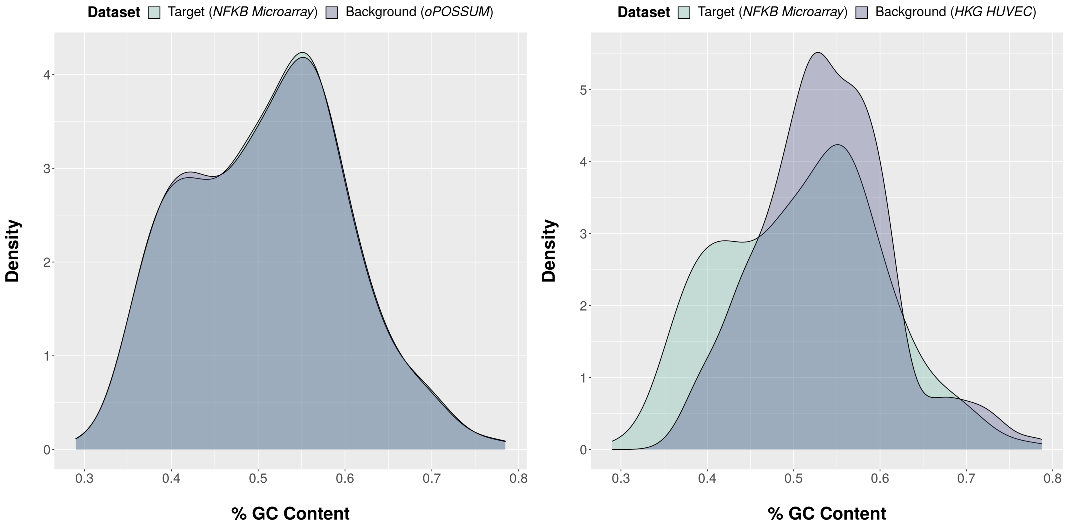
The main objective for the analysis of the NF-κB pathway-related dataset was to evaluate the utility of the discriminative motif discovery approach as a tool for promoter analysis. The secondary objective was to compare the results of the pipeline with those of the two conventional promoter analysis tools F-Match and oPOSSUM (Section 3.1.5). It should be noted at this point that the setup for this analysis does not reflect the standard application of the pipeline since different random promoter sets were used as backgrounds to evaluate the effects of the background selections on the results. To this end, only the step *W1.3* of *Workflow 1* (Section 4.1.1), *Workflow 2* (Section 4.1.2), *Workflow 3* (Section 4.1.3), and *Workflow 4* (Section 4.1.4) were deployed.

For the analysis, 288 promoter sequences were acquired that correspond to the down-regulated genes presented in [19]. This promoter set was used as the target set, labeled "*Target (NF-κB Microarray)*". Two different random background sets were created: (i) a background composed of promoters matching the GC composition of the target promoters using the tool BiasAway [116], labeled "*Background (oPOSSUM)*"; and (ii) a background composed of promoters related to non-overlapping housekeeping genes in HUVEC cells from the HRT Atlas [117], labeled "*Background (HKG HUVEC)*".

5.1.1. Analyzing the GC Content

To examine whether there was a potential bias in GC content between promoter sequences in *Target (NF-κB Microarray)* and both background random sets, I analyzed the density distributions of the GC content which were computed and visualized by kernel density estimate, i.e. a smoothed version of the histogram (Figures 5.1a and b).

As indicated by the density distributions, the GC content between *Target (NF-κB Microarray)* and *Background (oPOSSUM)* was well matched, with no significant shift in the



(a) **(b)**

Figure 5.1.: Visualization of the density distributions of the GC content for promoter sequences (1100 bp in length) in comparisons of *Target (NF-κB Microarray)* with *Background (oPOSSUM)* (a) or *Background (HKG HUVEC)* (b).

curve towards the left or right, which would have otherwise indicated an overall lower or higher GC content, respectively (Figure 5.1). In contrast, the GC content between *Target (NF-κB Microarray)* and *Background (HKG HUVEC)* was not that well matched and showed a more prominent shift of the curve for *Background (HKG HUVEC)* towards the right, indicating an overall higher GC content.

5.1.2. Optimizing MSS Cutoffs

After site search using the MATCH™ algorithm in *Workflow 2* (Section 4.1.2), the resulting site-specific matrices were subjected to Boruta in *Workflow 3* (Section 4.1.3) to find the most relevant MSS cutoffs per site model (or PWM) that best discriminated between target and background sets.

Applying Boruta using the GC content-matched background *Background (oPOSSUM)*, 19 out of 163 PWMs were retained. For four of these 19 PWMs (V$DELTAEF1_01, V$HELIOSA_02, V$LUN1_01, and V$PIT1_Q6_01), multiple MSS cutoffs were found to be relevant (Figure 5.2). These were ranked by their median importance scores, which reflected their importance over 100 Boruta iterations. I decided to select only those who had the highest median importance score per PWM as the most relevant MSS cutoffs (defined as the optimized MSS cutoffs). In contrast, applying Boruta using the housekeeping gene-related background *Background (HKG HUVEC)*, 95 out of 163 PWMs were retained. Figure 5.3 shows only the most relevant MSS cutoffs according to highest median importance score for these 95 PWMs.

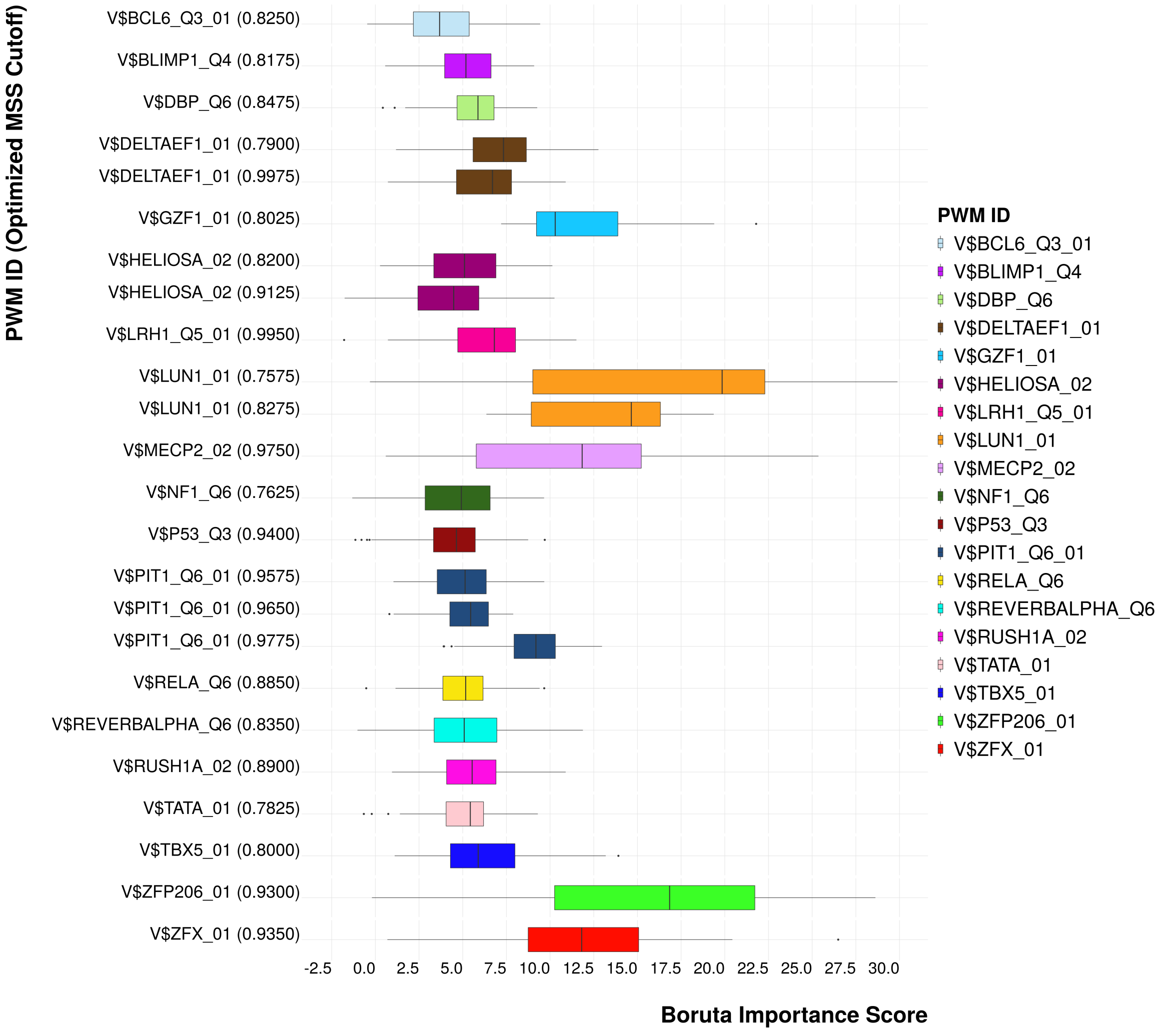


Figure 5.2.: Boruta results for the MSS cutoff optimization step using *Target (NF-κB Microarray)* and *Background (oPOSSUM)*. The forest plot summarizes all the PWMs retained at their relevant MSS cutoffs after 163 site-specific Boruta runs. The boxplots summarize the importance scores obtained over 100 Boruta iterations in each run.

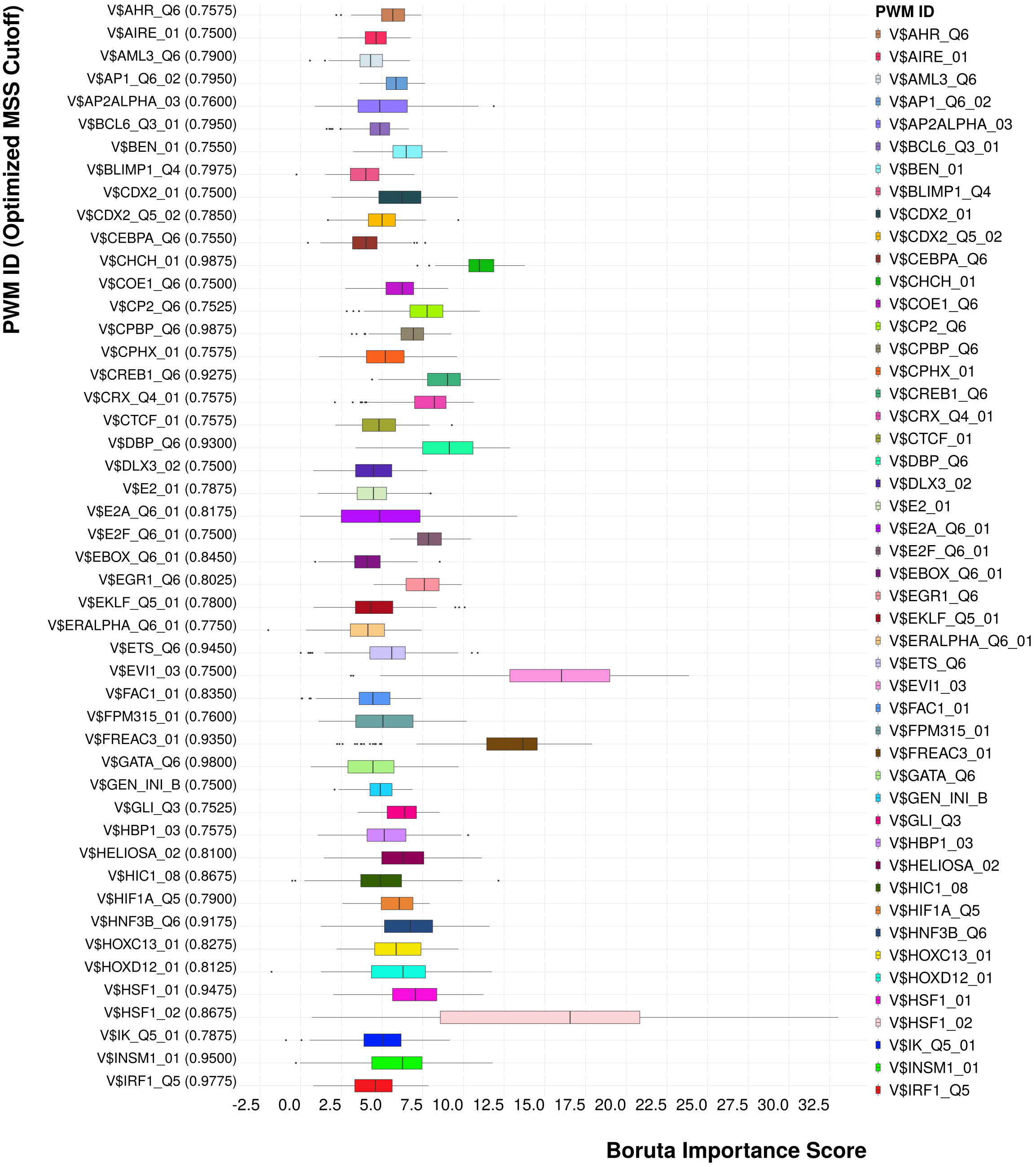

PWM ID (Optimized MSS Cutoff)
V$AHR_Q6 (0.7575)
V$AIRE_01 (0.7500)
V$AML3_Q6 (0.7900)
V$AP1_Q6_02 (0.7950)
V$AP2ALPHA_03 (0.7600)
V$BCL6_Q3_01 (0.7950)
V$BEN_01 (0.7550)
V$BLIMP1_Q4 (0.7975)
V$CDX2_01 (0.7500)
V$CDX2_Q5_02 (0.7850)
V$CEBPA_Q6 (0.7550)
V$CHCH_01 (0.9875)
V$COE1_Q6 (0.7500)
V$CP2_Q6 (0.7525)
V$CPBP_Q6 (0.9875)
V$CPHX_01 (0.7575)
V$CREB1_Q6 (0.9275)
V$CRX_Q4_01 (0.7575)
V$CTCF_01 (0.7575)
V$DBP_Q6 (0.9300)
V$DLX3_02 (0.7500)
V$E2_01 (0.7875)
V$E2A_Q6_01 (0.8175)
V$E2F_Q6_01 (0.7500)
V$EBOX_Q6_01 (0.8450)
V$EGR1_Q6 (0.8025)
V$EKLF_Q5_01 (0.7800)
V$ERALPHA_Q6_01 (0.7750)
V$ETS_Q6 (0.9450)
V$EVI1_03 (0.7500)
V$FAC1_01 (0.8350)
V$FPM315_01 (0.7600)
V$FREAC3_01 (0.9350)
V$GATA_Q6 (0.9800)
V$GEN_INI_B (0.7500)
V$GLI_Q3 (0.7525)
V$HBP1_03 (0.7575)
V$HELIOSA_02 (0.8100)
V$HIC1_08 (0.8675)
V$HIF1A_Q5 (0.7900)
V$HNF3B_Q6 (0.9175)
V$HOXC13_01 (0.8275)
V$HOXD12_01 (0.8125)
V$HSF1_01 (0.9475)
V$HSF1_02 (0.8675)
V$IK_Q5_01 (0.7875)
V$INSM1_01 (0.9500)
V$IRF1_Q5 (0.9775)
-2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
Boruta Importance Score
PWM ID
V$AHR_Q6
V$AIRE_01
V$AML3_Q6
V$AP1_Q6_02
V$AP2ALPHA_03
V$BCL6_Q3_01
V$BEN_01
V$BLIMP1_Q4
V$CDX2_01
V$CDX2_Q5_02
V$CEBPA_Q6
V$CHCH_01
V$COE1_Q6
V$CP2_Q6
V$CPBP_Q6
V$CPHX_01
V$CREB1_Q6
V$CRX_Q4_01
V$CTCF_01
V$DBP_Q6
V$DLX3_02
V$E2_01
V$E2A_Q6_01
V$E2F_Q6_01
V$EBOX_Q6_01
V$EGR1_Q6
V$EKLF_Q5_01
V$ERALPHA_Q6_01
V$ETS_Q6
V$EVI1_03
V$FAC1_01
V$FPM315_01
V$FREAC3_01
V$GATA_Q6
V$GEN_INI_B
V$GLI_Q3
V$HBP1_03
V$HELIOSA_02
V$HIC1_08
V$HIF1A_Q5
V$HNF3B_Q6
V$HOXC13_01
V$HOXD12_01
V$HSF1_01
V$HSF1_02
V$IK_Q5_01
V$INSM1_01
V$IRF1_Q5

Fig. 5.3 (*cont'd*)

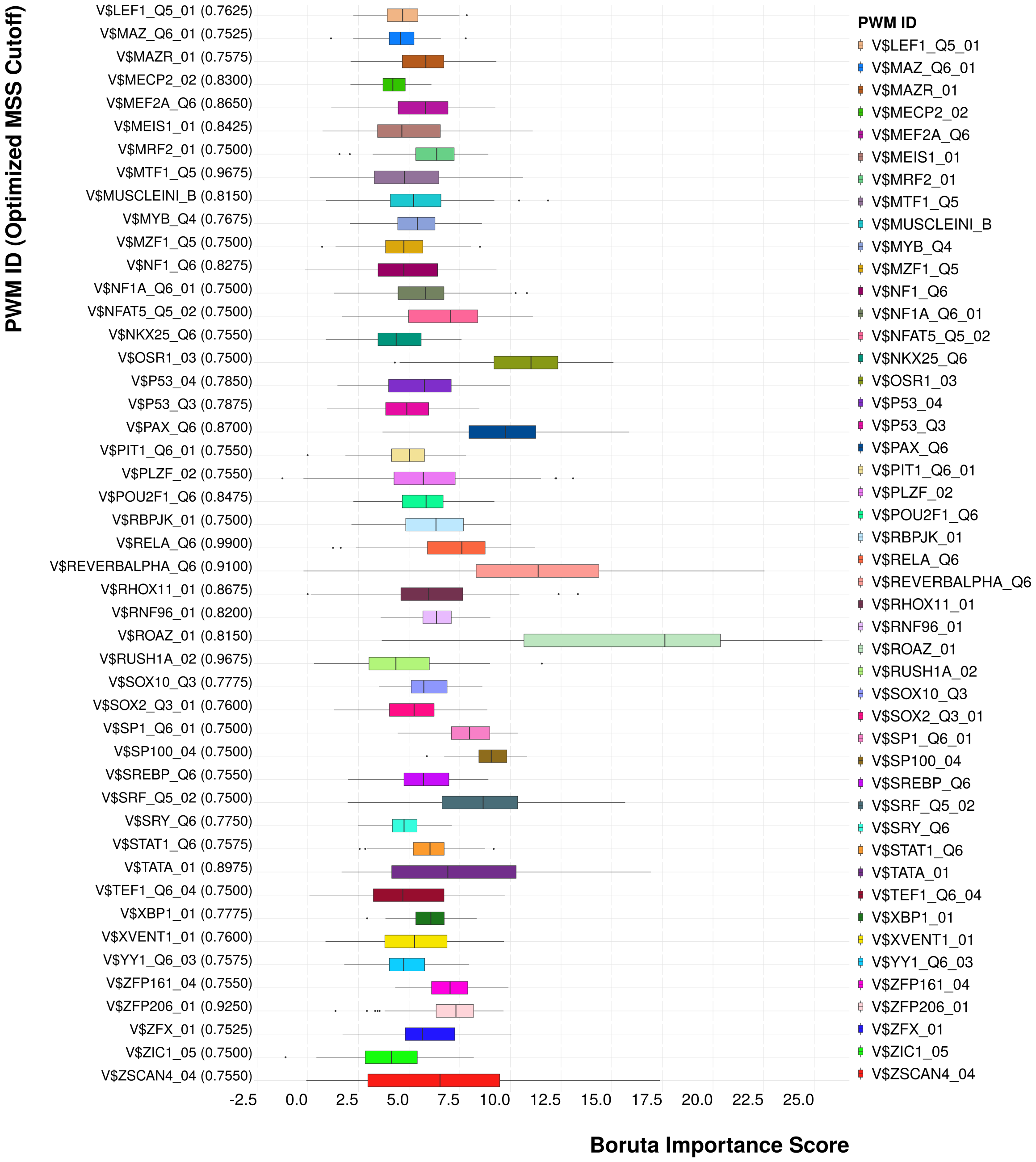


Figure 5.3.: Boruta results for the MSS cutoff optimization step using *Target (NF-κB Microarray)* and *Background (HKG HUVEC)*. Results are sorted alphabetically by PWM. The forest plot summarizes all the PWMs retained at their most relevant MSS cutoffs after 163 site-specific Boruta runs. The boxplots summarize the importance scores obtained over 100 Boruta iterations in each run.

5.1.3. Identifying the most relevant Site Models

Proceeding with *Workflow 4* (Figure 4.4), the most relevant PWMs were identified in both analyses. Applying Boruta using the GC content-matched background *Background (oPOSSUM)*, 14 out of 19 PWMs were retained, with V$RELA_Q6 (MSS cutoff of 0.885) being the most relevant one according to the median importance score (Figure 5.4). V$RELA_Q6 corresponds to the transcription factor p65, a subunit of the NF-κB transcription complex, which is involved in NF-κB activation and nuclear translocation [118]. At this point, it should be noted that V$RELA_Q6 is the only PWM in the library that refers to the NF-κB family of TFs and, thus, serves as the reference PWM for this dataset.

Applying Boruta using the housekeeping gene-related background *Background (HKG HUVEC)*, 36 out of 95 PWMs were retained, with V$SP100_04 (MSS cutoff of 0.75) being the most relevant one according to the median importance score (Figure 5.5). V$SP100_04 corresponds to the SP100 nuclear antigen which is associated with innate immune response [119]. For this analysis, V$RELA_Q6 (MSS cutoff of 0.99) was found on the 9th rank.

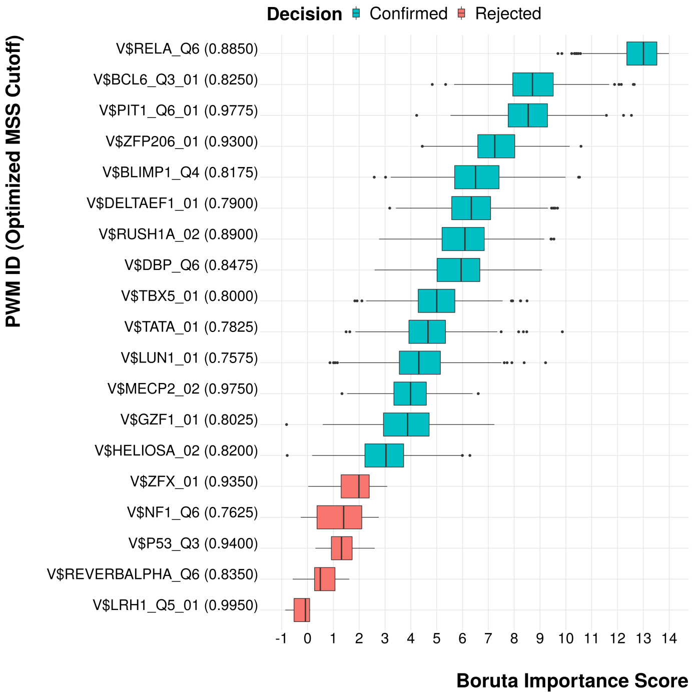
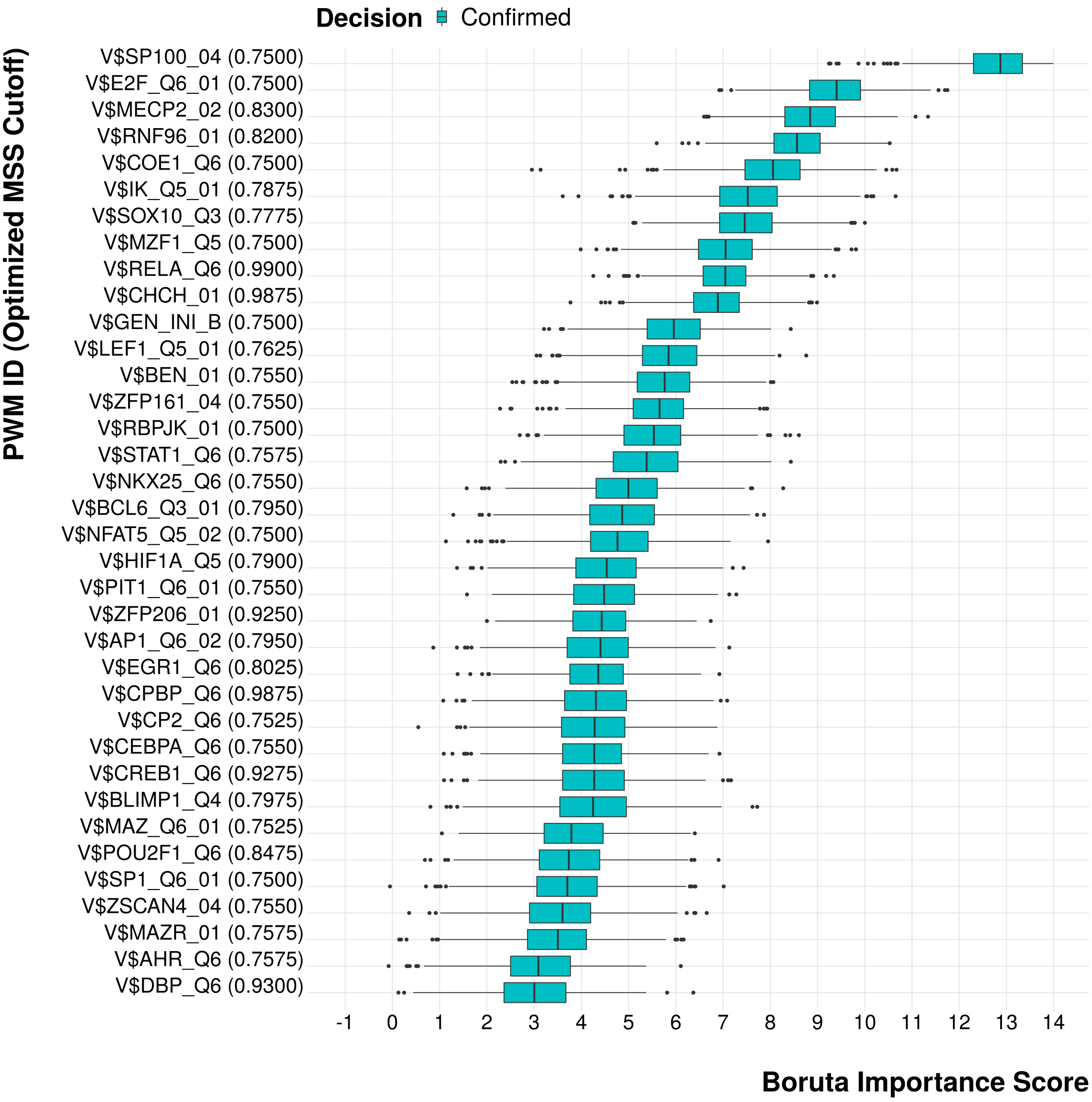
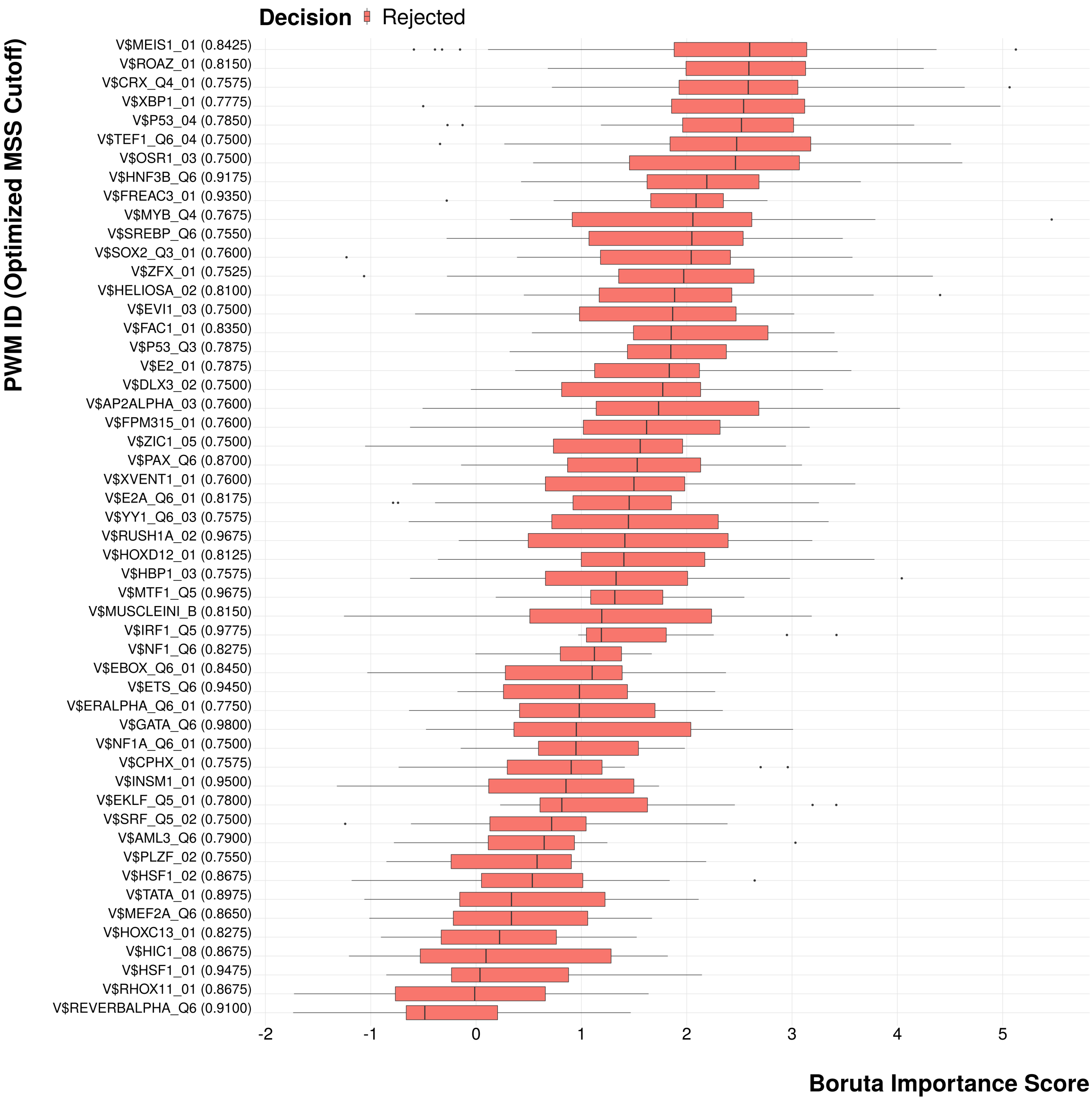


Figure 5.4.: Boruta results showing the most relevant PWMs in the analysis using the GC content-matched *Background (oPOSSUM)*. The forest plot shows PWMs at their optimized MSS cutoffs (the features) that were either confirmed or rejected after the site-mixed Boruta run. The boxplots summarize their importance scores using 1000 iterations in the Boruta run. PWMs are sorted by their median importance score.


Decision
Confirmed
PWM ID (Optimized MSS Cutoff)
V$SP100_04 (0.7500)
V$E2F_Q6_01 (0.7500)
V$MECP2_02 (0.8300)
V$RNF96_01 (0.8200)
V$COE1_Q6 (0.7500)
V$IK_Q5_01 (0.7875)
V$SOX10_Q3 (0.7775)
V$MZF1_Q5 (0.7500)
V$RELA_Q6 (0.9900)
V$CHCH_01 (0.9875)
V$GEN_INI_B (0.7500)
V$LEF1_Q5_01 (0.7625)
V$BEN_01 (0.7550)
V$ZFP161_04 (0.7550)
V$RBPJK_01 (0.7500)
V$STAT1_Q6 (0.7575)
V$NKX25_Q6 (0.7550)
V$BCL6_Q3_01 (0.7950)
V$NFAT5_Q5_02 (0.7500)
V$HIF1A_Q5 (0.7900)
V$PIT1_Q6_01 (0.7550)
V$ZFP206_01 (0.9250)
V$AP1_Q6_02 (0.7950)
V$EGR1_Q6 (0.8025)
V$CPBP_Q6 (0.9875)
V$CP2_Q6 (0.7525)
V$CEBPA_Q6 (0.7550)
V$CREB1_Q6 (0.9275)
V$BLIMP1_Q4 (0.7975)
V$MAZ_Q6_01 (0.7525)
V$POU2F1_Q6 (0.8475)
V$SP1_Q6_01 (0.7500)
V$ZSCAN4_04 (0.7550)
V$MAZR_01 (0.7575)
V$AHR_Q6 (0.7575)
V$DBP_Q6 (0.9300)
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Boruta Importance Score

(a)

Fig. 5.5 (*cont'd*)



(b)

Figure 5.5.: Boruta results showing the most relevant PWMs in the analysis using the housekeeping gene-related *Background (HKG HUVEC)*. The forest plot shows PWMs at their optimized MSS cutoffs (the features) that were either confirmed (a) or rejected (b) after the site-mixed Boruta run. The boxplots summarize their importance scores using 1000 iterations in the Boruta run. PWMs are sorted by their median importance score.

5.1.4. Analyzing Site Over- and Underrepresentation

To reflect *site over- and underrepresentation* (also called *site enrichment*), target-to-background ratios (*T/B*) were calculated to compare the average site frequencies that corresponded to the most relevant PWMs in *Target* to those in *Background*. The average site frequency was calculated by adding up all site occurrences observed in a respective column of a site count matrix divided by the number of promoter sequences in a set (*N*=288). The target-to-background ratio was then calculated by dividing the average site frequency in *Target* by the average site frequency in the *Background*.

The results indicated that 7 out of 14 PWMs were associated with prominent site enrichments according to the *T/B* in *Target (NF-κB Microarray)*. V$RELA_Q6 (MSS cutoff of 0.885) was associated with the third highest *T/B* ratio of approximately 1.43 for the first analysis using *Background (oPOSSUM)* (Figure 5.6 and Table A1).

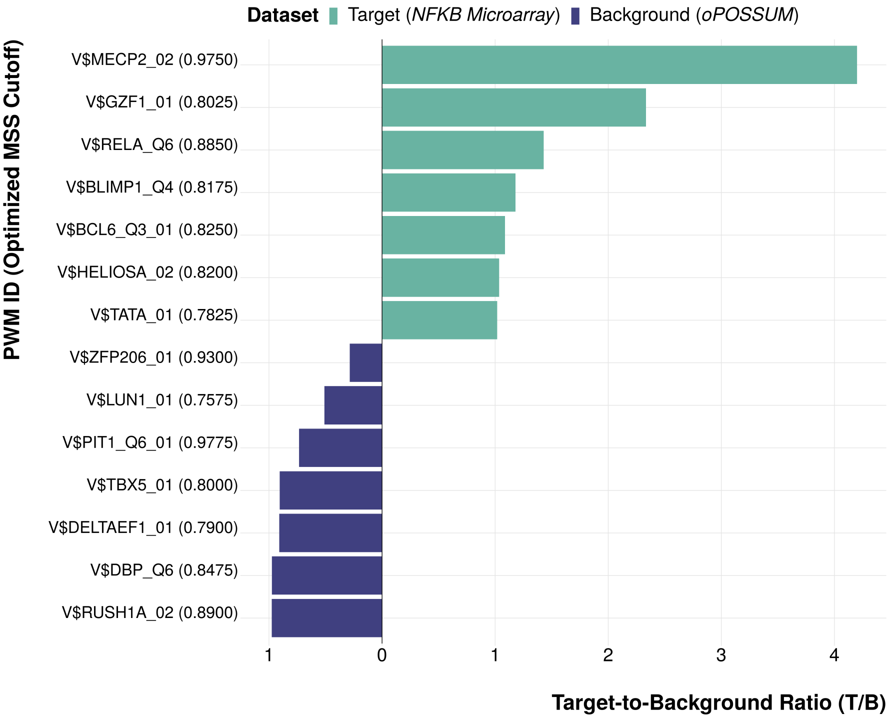


Figure 5.6.: Double-sided bar plot showing target-to-background ratios (*T/B*) for the most relevant PWMs in the analysis using *Target (NF-κB Microarray)* and the GC content-matched *Background (oPOSSUM)* .

For the second analysis using *Background (HKG HUVEC)*, 17 out of 36 PWMs were associated with prominent enrichments in *Target (NF-κB Microarray)*. Interestingly, V$RELA_Q6 (MSS cutoff of 0.99) was associated with the highest (*T/B*) ratio of 10 (Figure 5.7 and Table A2).

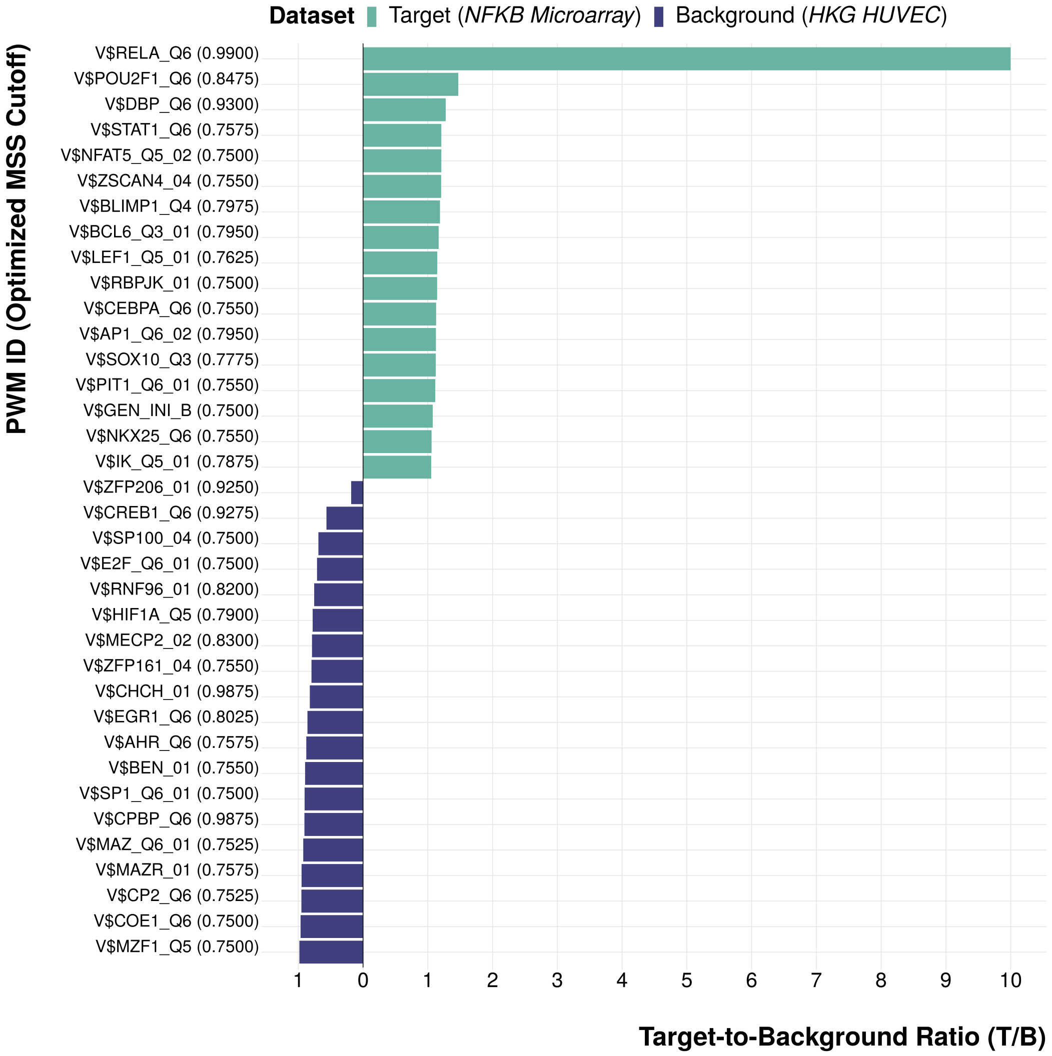


Figure 5.7.: Double-sided bar plot showing target-to-background ratios (*T/B*) for the most relevant site models (PWMs) in the analysis using *Target (NF-κB Microarray)* and the housekeeping gene-related *Background (HKG HUVEC)* .

5.1.5. Comparing Methods in Promoter Analysis

I repeated both analyses for the two conventional site overrepresentation methods F-Match and oPOSSUM. Thereafter, I performed a comparison between the three methods (i.e., the pipeline, F-Match, and oPOSSUM). More details on F-Match and oPOSSUM are provided in Sections 3.1.5 and 3.1.6.

The pipeline and F-Match use the MATCH™ algorithm and incorporate a MSS cutoff optimization step. Therefore, F-Match was also applied with initial MSS cutoffs of 0.75 (and default CSS cutoffs from the minFN profile), and overrepresented sites were reported using an FDR cutoff of less than 0.05. oPOSSUM, on the other hand, uses its own implementation for site search and reports sites above a fixed cutoff score with a minimum required value of 0.75 for motif matches. To obtain comparable results, oPOSSUM was applied using a motif match cutoff of 0.75, a Fisher p-value less than 0.01, and a Z-score greater than 10. F-Match and oPOSSUM were deployed using the web-based systems of *Sequence-based Single Site Analysis* (`http://opossum.cisreg.ca/cgibin/oPOSSUM3/opossum_seq_ssa` and the geneXplain platform (`https://genexplain.com`, respectively.

For the analysis using the GC content-matched *Background (oPOSSUM)*, F-Match and oPOSSUM identified 34 and 85 relevant PWMs that are linked to site overrepresentation, respectively (Figure 5.8). Considering only the top 10, both the pipeline and F-Match identified V$RELA_Q6 at the 1st rank, while the results of oPOSSUM did not include this PWM amongst the top 10 (Table 5.1).

For the analysis using the housekeeping gene-related *Background (HKG HUVEC)*, F-Match and oPOSSUM identified 103 and 100 relevant PWMs, respectively (Figure 5.9). Considering only the top 10, the pipeline and F-Match identified V$RELA_Q6 at the 9th rank and 2nd rank, respectively, while the results of oPOSSUM did not include this PWM amongst the top 10 (Table 5.2).

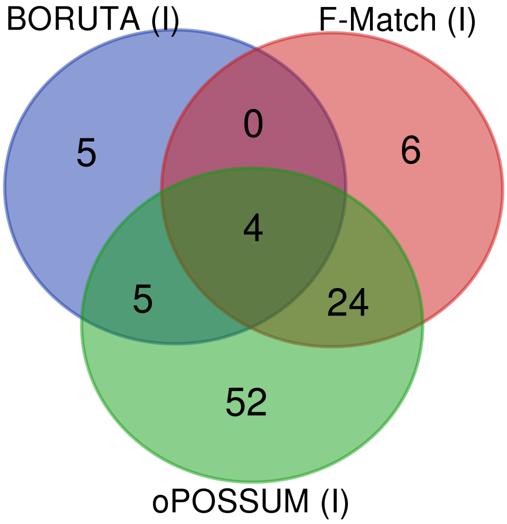


Figure 5.8.: Total number of overlapping relevant PWMs acquired from analyses using the pipeline (denoted *Boruta*), F-Match, and oPOSSUM in the analysis using *Target (NF-κB Microarray)* and the GC content-matched *Background (oPOSSUM)*.

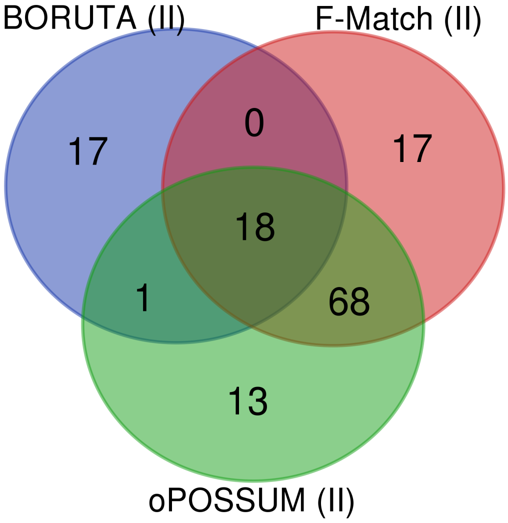


Figure 5.9.: Total number of overlapping relevant PWMs acquired from analyses using the pipeline (denoted *Boruta*), F-Match, and oPOSSUM in the analysis using *Target (NF-κB Microarray)* and the housekeeping gene-related *Background (HKG HUVEC)*.

Table 5.1.: Top 10 relevant PWMs identified by the pipeline (denoted *Boruta*), F-Match, and oPOSSUM in the analysis using *Target (NF-κB Microarray)* and the GC content-matched *Background (oPOSSUM)*.

Rank	Pipeline (Boruta)	F-Match	oPOSSUM
1	V$RELA_Q6	V$RELA_Q6	V$IRF1_Q5
2	V$BCL6_Q3_01	V$HSF1_01	V$SRF_Q5_02
3	V$PIT1_Q6_01	V$ETS_Q6	V$NFAT5_Q5_02
4	V$ZFP206_01	V$HMGIY_Q3	V$BLIMP1_Q4
5	V$BLIMP1_Q4	V$IRF1_Q5	V$ZFP161_04
6	V$DELTAEF1_01	V$INSM1_01	V$CEBPA_Q6
7	V$RUSH1A_02	V$AP2ALPHA_03	V$NFAT1_Q4
8	V$DBP_Q6	V$MAZ_Q6_01	V$BCL6_Q3_01
9	V$TBX5_01	V$CEBPA_Q6	V$XVENT1_01
10	V$TATA_01	V$BLIMP1_Q4	V$STAT1_Q6

Table 5.2.: Top 10 relevant PWMs identified by the pipeline (denoted *Boruta*), F-Match, and oPOSSUM in the analysis using *Target (NF-κB Microarray)* and the housekeeping gene-related *Background (HKG HUVEC)*.

Rank	Pipeline (Boruta)	F-Match	oPOSSUM
1	V$SP100_04	V$HSF1_01	V$HMGIY_Q3
2	V$E2F_Q6_01	V$RELA_Q6	V$POU2F1_Q6
3	V$MECP2_02	V$HSF1_02	V$PIT1_Q6_01
4	V$RNF96_01	V$NR3C1_03	V$HNF3B_Q6
5	V$COE1_Q6	V$DMRT4_01	V$TATA_01
6	V$IK_Q5_01	V$POU2F1_Q6	V$CDX2_Q5_02
7	V$SOX10_Q3	V$BBX_03	V$SRY_Q6
8	V$MZF1_Q5	V$COE1_Q6	V$IPF1_Q5
9	V$RELA_Q6	V$HDX_01	V$STAT1_Q6
10	V$CHCH_01	V$ROAZ_01	V$CEBPA_Q6

5.1.6. Interim Conclusion

In this section, the pipeline's utility for promoter analysis was evaluated using a set of downregulated genes (the target set) from an experiment comparing Interleukin-1β (IL1B)-stimulated HUVEC cells treated with a NF-κB inhibitor versus IL1B-stimulated HUVEC cells not treated with this inhibitor. When evaluated using a conventional site overrepresentation method for promoter analysis, binding sites for the NF-κB family of TFs were previously characterized as the most biologically relevant ones in the target set.

Given the NF-κB pathway-related target set, I assessed whether the Boruta feature ranking and selection algorithm as a part of the pipeline can be applied to identify discriminatory sites that relate to the NF-κB family of TFs. For this purposes, I retained a naïve background set of random promoter sequences matched to the GC content of the target set in the assessment. A second background was generated by selecting promoter sequences that correspond to non-modulated housekeeping genes in HUVEC.

When the pipeline was applied to the target set in conjunction with either one of the two background sets, the PWM V$RELA_Q6 which corresponds to members of the NF-κB family of TFs (NF-κB, c-REL, p65, and p50) was successfully identified in both analyses. Applying Boruta using the GC content-matched background, V$RELA_Q6 was identified as the most relevant PWM within the ranked list (i.e., ranked by Boruta's median importance score). In contrast, applying Boruta using the housekeeping gene-related background, another PWM, namely V$SP100_04 (SP100 nuclear antigen), was identified as the most relevant PWM within the ranked list. Since SP100 nuclear antigen is involved in the immune response, it might interact with or act independently of the NF-κB pathway, thereby regulating the host's intrinsic immune response.

Focusing on site overrepresentation, the methods comparison between the pipeline and the two conventional overrepresentation methods, F-Match and oPOSSUM, showed that all three methods can identify *a priori* known biologically relevant sites. As a result, it can be concluded that several of the pipeline's relevant sites may potentially fall within the basic concept of site overrepresentation.

After demonstrating the pipeline's utility for promoter analysis to identify the TFs potentially driving gene expression changes in a heterogeneous dataset, it was then applied to two additional reference datasets with a different setup, where the target and background sets correspond to significantly upregulated and significantly downregulated genes (the DEGs) in an experiment, respectively. This setup reflects the standard application of the pipeline.

5.2. Analysis of the WNT Pathway-related CRC Dataset

In this section, I evaluated the utility of the pipeline to identify relevant BPs as well as MRs of signaling pathways in CRC based on RNA-Seq data. For this purpose, I performed a comparative analysis of two distinct murine CRC cell lines, namely *1638N-T1* and *CMT-93*, which represent suitable models to study determinants of cancer cell invasiveness. The cell line *1638N-T1* carries specific predisposing mutations in key genes of the WNT pathway that lead to hyperactivation of this pathway and, consequently, to a greater invasive capacity of *1638N-T1* cells in comparison to *CMT-93* cells.

The differences at both the transcriptional regulation and pathway levels between the two cell lines were examined for the comparison "*1638N-T1* versus *CMT-93*". Consequently, the upregulated genes in *1638N-T1* were used as the target set, whereas the downregulated genes were used as the background set. When the direction of the comparison is changed, these downregulated genes can likewise represent upregulated genes in *CMT-93*.

5.2.1. RNA-Seq data analysis and Promoter Sequence Extraction

For the analysis of the WNT pathway-related CRC dataset, RSEM-gene level count data was acquired from [17] and processes as outlined in *Workflow 1* (Figure 4.1). Thereafter, RNA-Seq data analysis was performed, which involved the comparative transcriptome analysis of two CRC cell lines, denoted "*1638N-T1*" and "*CMT-93*". The *1638N-T1*-related samples were considered as the target set, denoted "*Target (1638N-T1)*". The '*CMT-93*-related samples were considered as the background set, denoted "*Background (CMT-93)*".
After initial explorative analysis of the gene expression levels across the samples, 2650 DEGs were identified between the two sample groups, of which 1326 and 1324 were significantly upregulated for *Target (1638N-T1)* and *Background (CMT-93)*, respectively (Figure 5.10). After retrieving the corresponding promoter sequences for these genes, the *Target (1638N-T1)*-related set with 1326 sequences (sorted by decreasing log 2 fold changes of the corresponding DEGs) was truncated at the bottom to match the number of 1322 sequences in the *Background (CMT-93)*-related set. For further analysis, the same labels were used for the two sequence sets, i.e., *Target (1638N-T1)* and *Background (CMT-93)*.

5.2.2. Analyzing the GC Content

To examine whether there was a potential bias in GC content between promoter sequences in *Target (1638N-T1)* and *Background (CMT-93)*, I analyzed the density distributions of the GC content which were computed and visualized by kernel density estimate (Figure 5.11). As indicated by the density distributions, the GC content in *Background (CMT-93)* shows a slight shift in the curve towards the right, indicating an overall higher GC content when compared to *Target (1638N-T1)*.

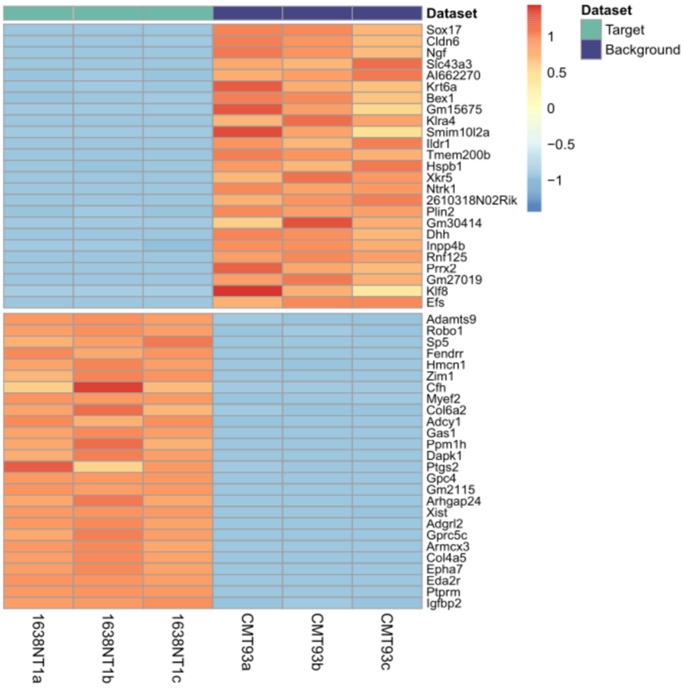


Figure 5.10.: Comparative transcriptomic analysis of samples related to *Target (1638N-T1)* and *Background (CMT-93)*. The heatmap visualizes the top 25 upregulated DEGs and the top 25 downregulated DEGs based on log 2 fold changes in the comparison. Count values related to gene expression were centered and scaled in row direction.

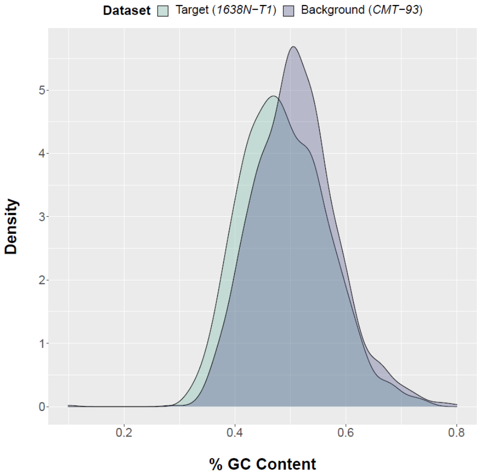


Figure 5.11.: Visualization of the density distributions of the GC content for promoter sequences (1100 bp in length) in *Target (1638N-T1)* and *Background (CMT-93)*.

5.2.3. Optimizing MSS Cutoffs

After site search using the MATCH™ algorithm in *Workflow 2* (Section 4.1.2), the resulting site-specific matrices were subjected to Boruta in *Workflow 3* (Section 4.1.3) to find the most relevant MSS cutoffs per site model (or PWM) that best discriminated between target and background sets. When Boruta was used for the MSS cutoff optimization step in *Workflow 3* (Figure 4.3), 85 relevant site models (PWMs) were retained (Figure 5.12).

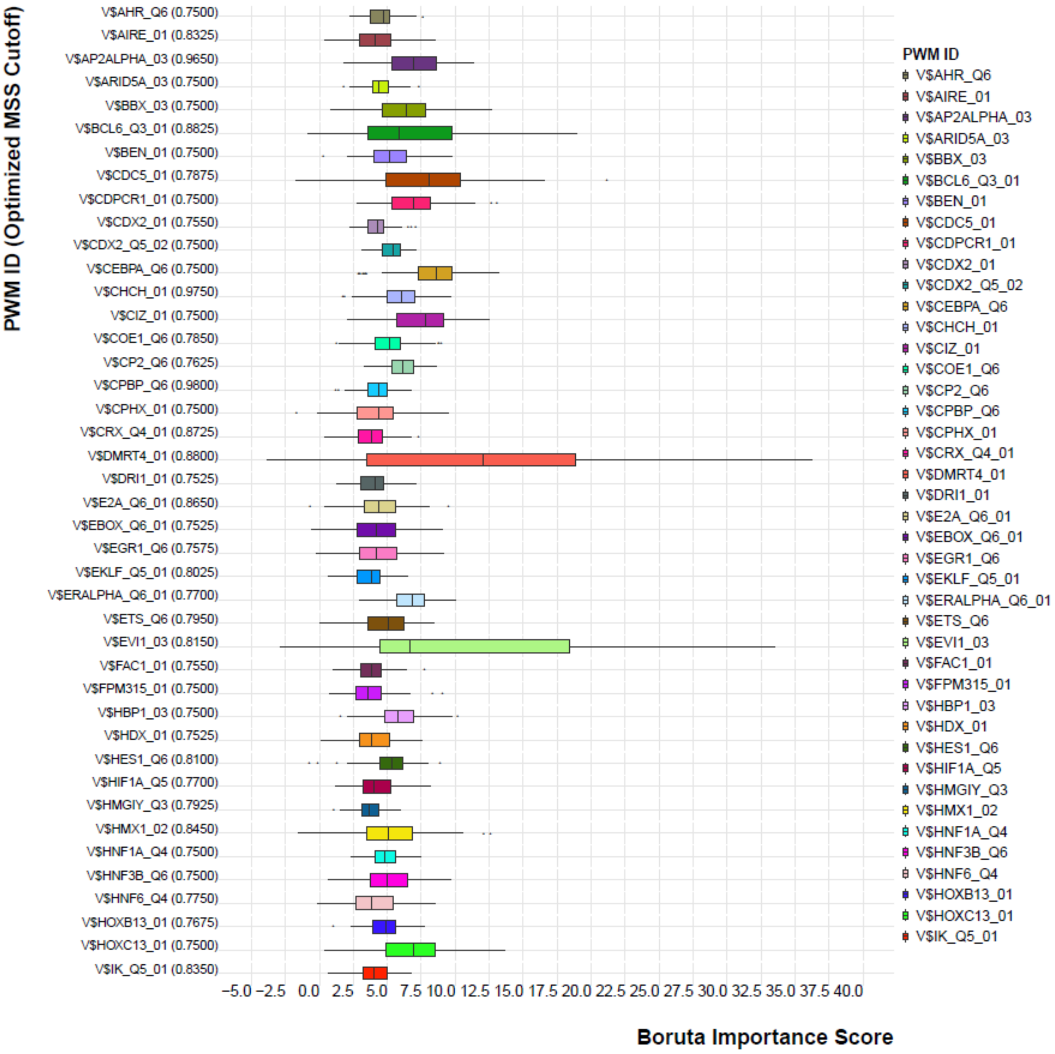


Fig. 5.12 (*cont'd*)

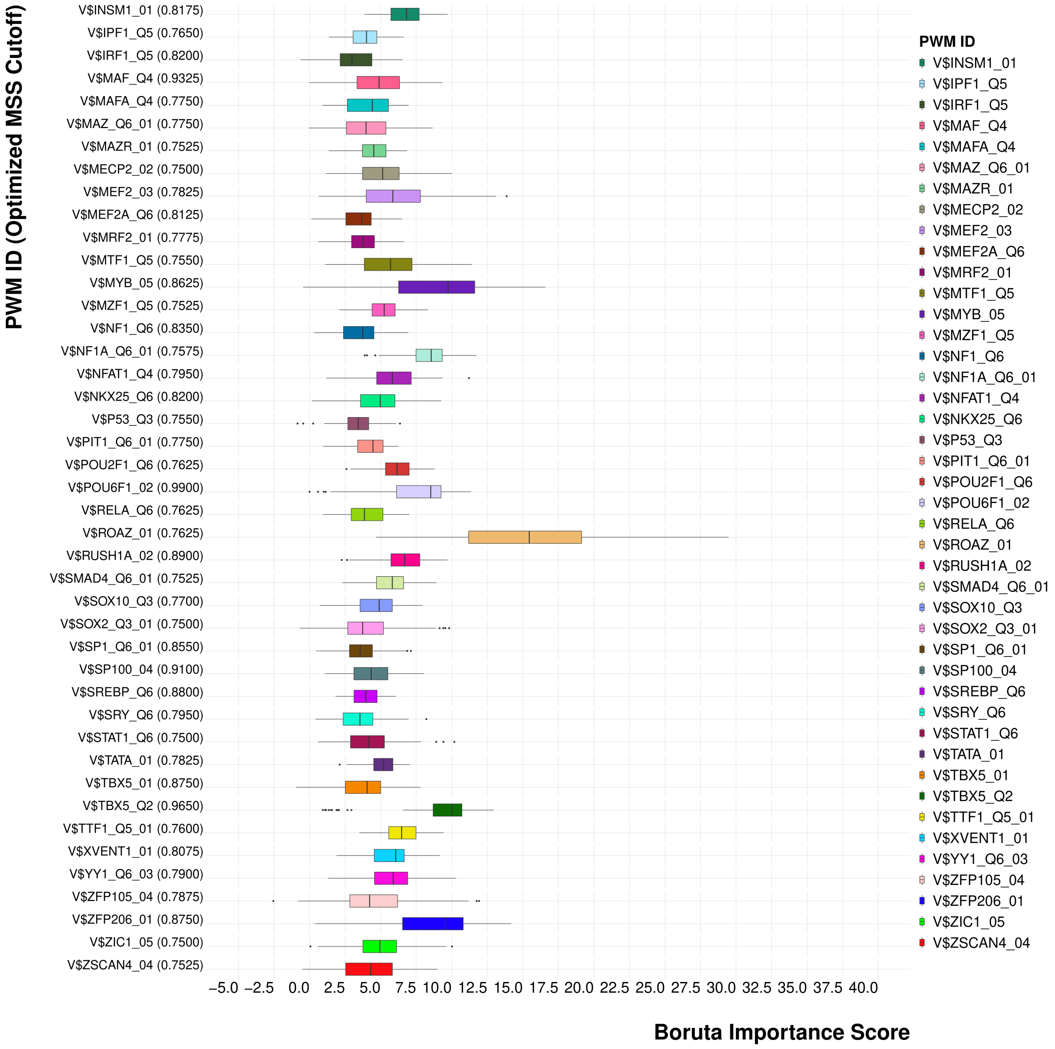
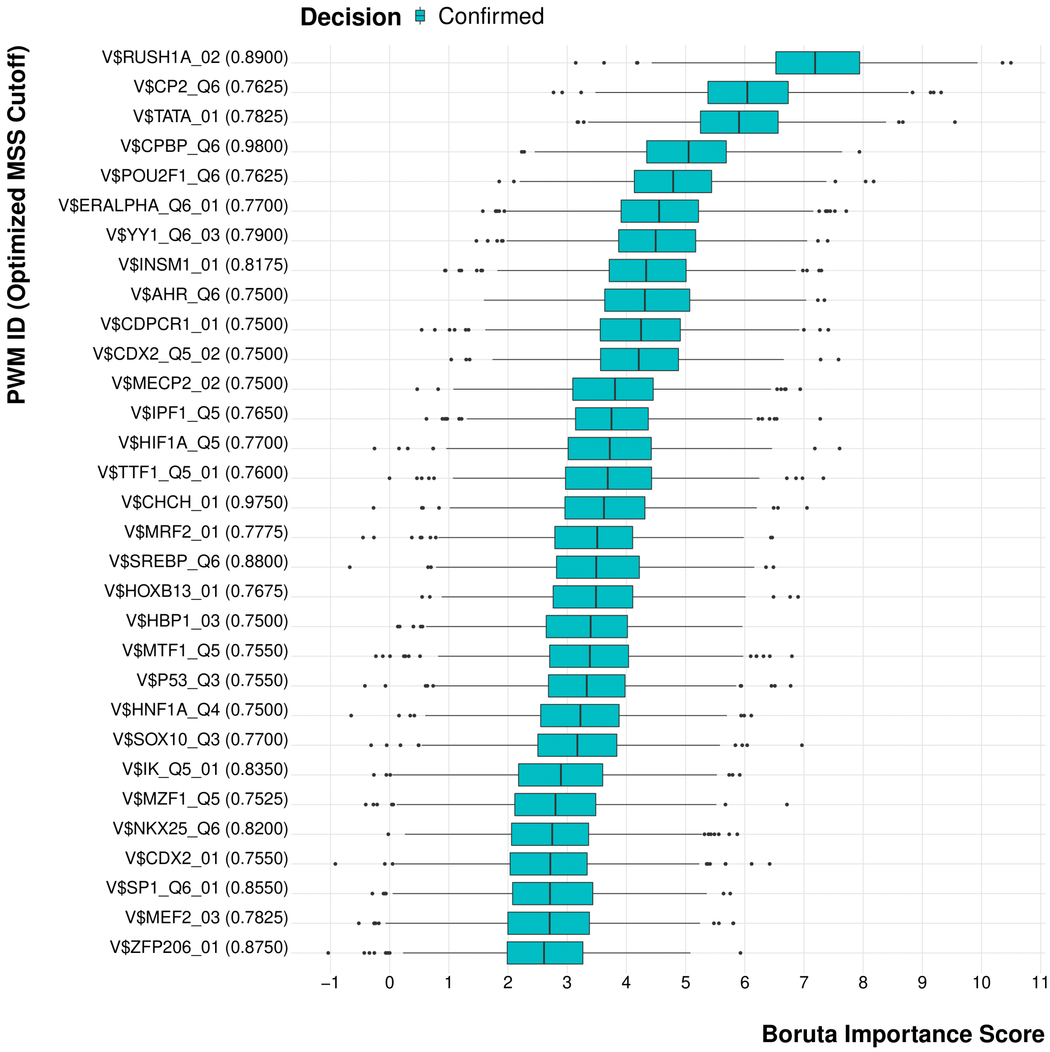


Figure 5.12.: Boruta results for the MSS cutoff optimization step using *Target (1638N-T1)* and *Background (CMT-93)*. Results are sorted alphabetically by PWM. The forest plot summarizes all the PWMs retained at their most relevant MSS cutoffs after 163 site-specific Boruta runs. The boxplots summarize the importance scores obtained over 100 Boruta iterations in each run.

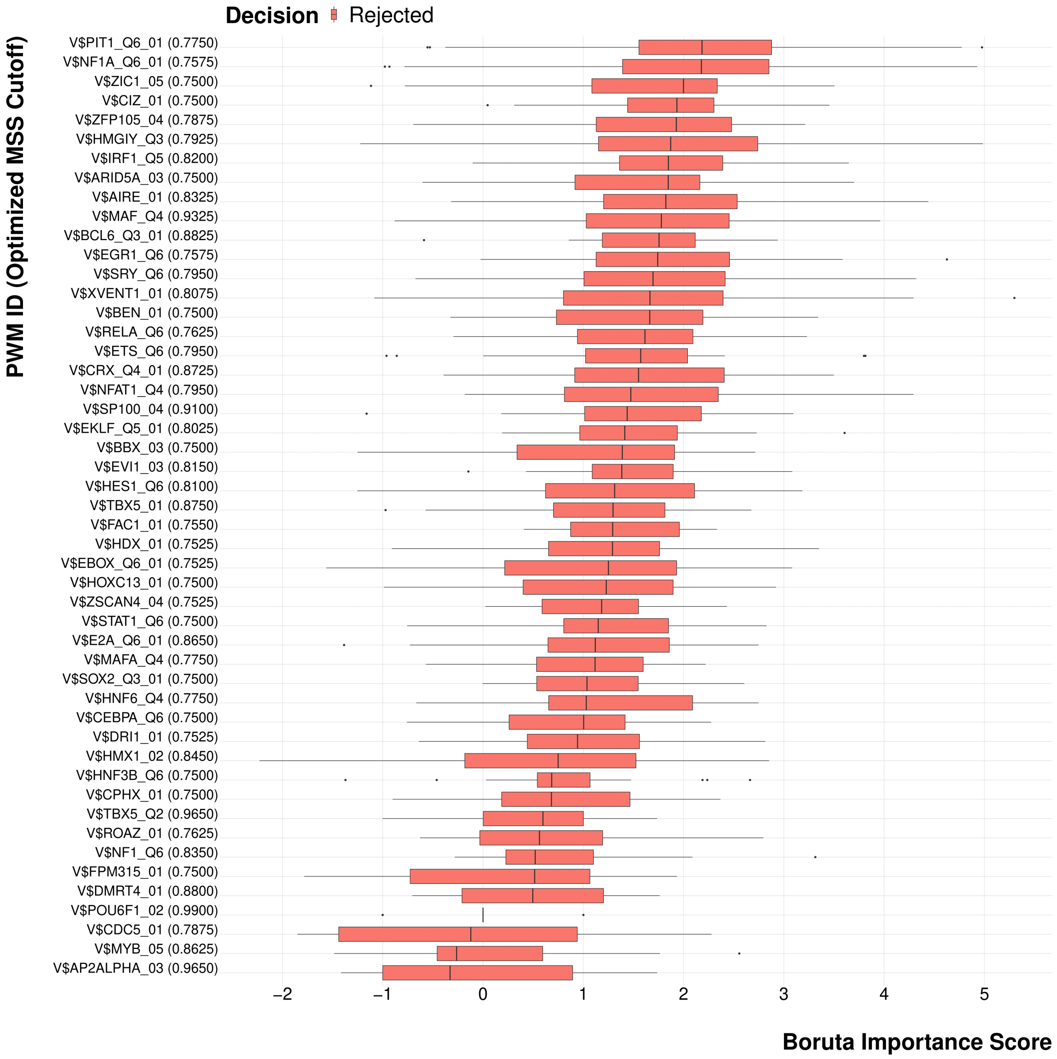
5.2.4. Identifying the most relevant Site Models

Proceeding with *Workflow 4* (Figure 4.4), the most relevant PWMs were identified in the analysis. Applying Boruta to the site-mixed count matrix, 31 out of 85 PWMs were retained, with V$RUSH1A_02 (MSS cutoff: 0.89) being the most relevant one according to the median importance score (Figure 5.13). V$RUSH1A_02 corresponds to the helicase-like TF (HLTF) whose enzymatic functions have been implicated in both translocase and ubiquitin ligase activity [120, 121]. Furthermore, HLTF has been previously associated with migration and invasion in CRC via TGF-β/SMAD signaling [122].



(a)

Fig. 5.13 (*cont'd*)



(b)

Figure 5.13.: Boruta results showing the most relevant PWMs. The forest plot shows PWMs at their optimized MSS cutoffs (the features) that were either confirmed (a) or rejected (b) after the site-mixed Boruta run. The boxplots summarize their importance scores using 1000 iterations in the Boruta run. PWMs are sorted by their median importance score.

5.2.5. Analyzing Site Over- and Underrepresentation

To further investigate how the average site frequencies that corresponded to the 31 relevant PWMs in *Target (1638N-T1)* compare to those in *Background (CMT-93)*, target-to-background ratios (*T/B*) were calculated.

The results indicated that 14 out of 31 were associated with prominent site enrichments according to the *T/B* in *Target (1638N-T1)*, with V$MEF2_03 (MSS cutoff of 0.7825) having the highest *T/B* ratio of approximately 1.31 (Figure 5.14 and Table A3).

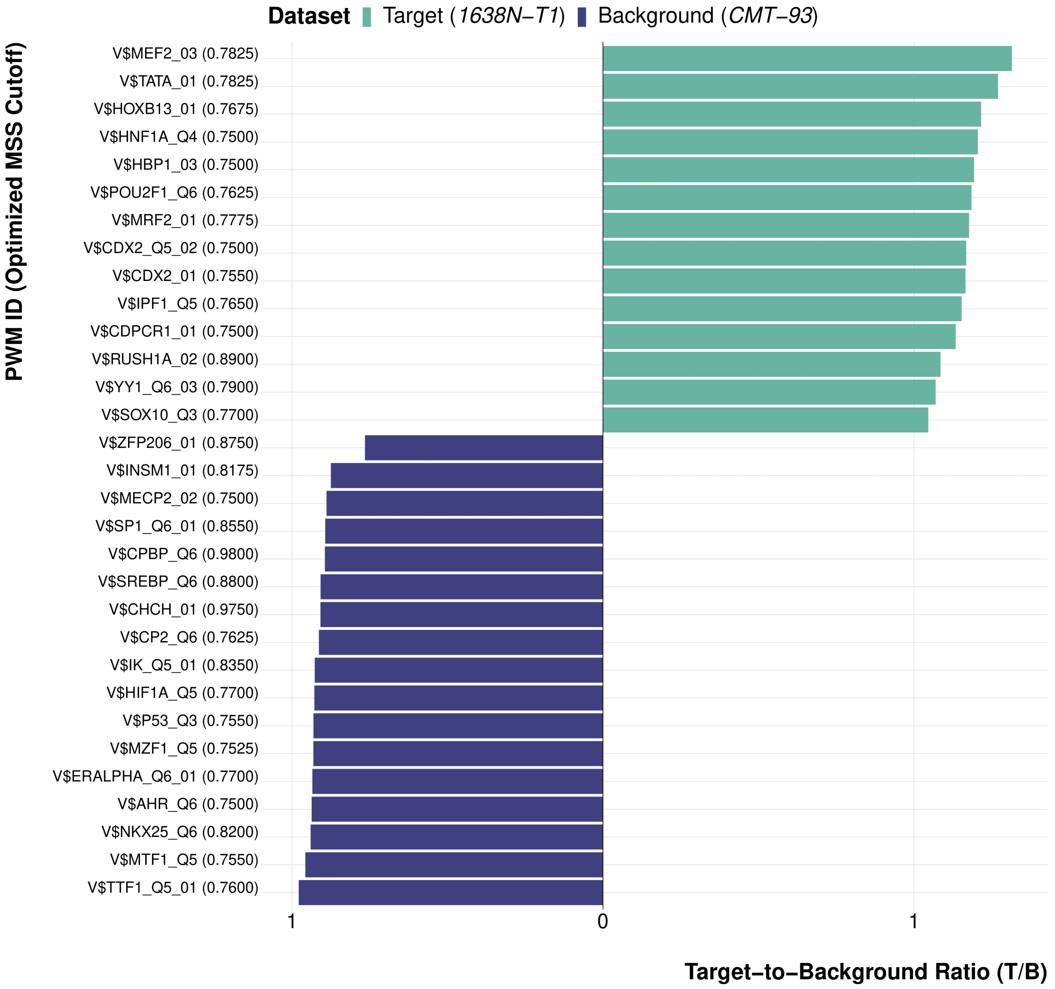


Figure 5.14.: Double-sided bar plot showing target-to-background ratios (*T/B*) for the most relevant PWMs.

5.2.6. Retrieving Gene Clusters

Proceeding with *Workflow 5*, gene clusters consisting of any possible combination of DEGs (defined in *Workflow 1*) across *Target* and *Background* (Figure 4.5) were identified. To this end, the relevant PWMs in the *Target*- and *Background*-related site-mixed count matrices were retained. The matrices were then attached side by side and the combined matrix was transposed. Thereafter, the JSD was used to calculate the similarities between all probability distributions that correspond to the columns (gene promoters) in the site-mixed matrix, leading to the creation of a JSD similarity matrix. To better distinguish between gene clusters, k-means clustering was performed on top of PCA to investigate the relationship in the PCA projection. Consequently, a visualization of the gene clusters on a two-dimensional plane was obtained by selecting the principal components that account for the most variability as the axes. In addition, the silhouette method was utilized to determine the optimal number of k clusters which was supplied to the k-means algorithm. As a results, the 2644 DEGs were clustered into three gene clusters that contained 592 (*Cluster 1*), 507 (*Cluster 2*), and 1545 (*Cluster 3*) genes (Figure 5.15).

Next, the overlap in genes between the 1322 DEGs related to *Target (1638N-T1)* and *Background (CMT-93)*, respectively, and the three clusters was analyzed. Only *Cluster 1* showed a greater overlap with the DEGs related to *Target (1638N-T1)* while having a smaller overlap with the DEGs related to *Background (CMT-93)* among the three clusters (Table 5.3).

Table 5.3.: Gene cluster affiliations of DEGs related to *Target (1638N-T1)* and *Background (CMT-93)*.

Dataset	*Cluster 1*	*Cluster 2*	*Cluster 3*
Target (1638N-T1)	381	209	732
Background (CMT-93)	211	298	813

To test whether there was a significant association between the observed overlaps in Table 5.3, a Pearson's *Chi-squared* test of independence was performed using the cluster affiliations as three variables in a three-way contingency table. The *Chi-squared* test resulted in a p-value of 1.215e-15, which provides sufficient evidence to conclude that the observed overlaps are not likely due to chance.

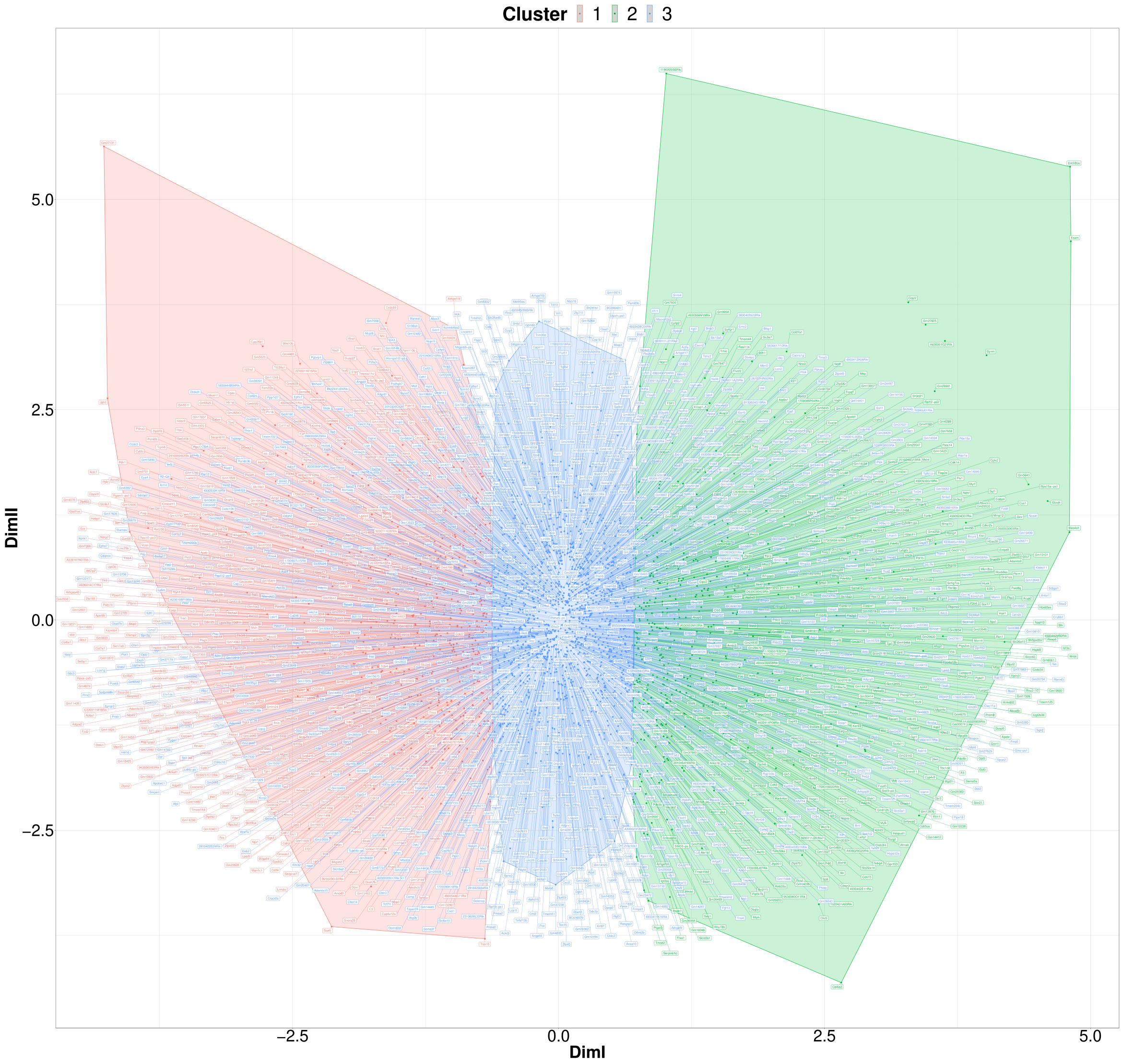


Figure 5.15.: PCA-based k-means clustering of 2644 DEGs. Calculations are based on JSD-derived similarity measures between the probability distributions that relate to the 31 relevant PWMs. Different colors indicate the three different gene clusters. The optimal number of *k* clusters was determined by the silhouette method.

5.2.7. Analyzing GO Overrepresentation

To further assess whether the three clusters were associated with different biological themes, a comparative functional categorization was perfomred using ORA in combination with annotations defined in GO BP. Noteworthy, the R package clusterProfiler only considered 409/592 (*Cluster 1*), 374/507 (*Cluster 2*), and 1130/1545 (*Cluster 3*) genes in each cluster for ORA after gene ID conversion. As shown in Figure 5.16, the most discriminative BP terms are listed in the top 10 significant BP terms (BH-adjusted p-value < 0.05) for each cluster. Interestingly, the results revealed that the top three BP terms ("renal system development", "endothelial cell migration", and "tissue migration") for *Cluster 1* were not included in the BP terms for the other two clusters. The top 20 significant BP terms are shown for *Cluster 1* in Table A4.

5.2.8. Analyzing Dysregulated Pathways

Finally, MRs were searched for each gene cluster using the geneXplain platform to construct dysregulated pathways in cancer. As a result, 78, 66, and 51 MRs were identified for *Cluster 1*, *Cluster 2*, and *Cluster 3*, respectively. Thereafter, I analyzed the overlap between MRs related to the three clusters (Figure 5.17), which showed that most MRs were unique for each cluster. Since the previous results suggested a higher relevance of *Cluster 1* for *Target (1638N-T1)*, I sought to combine the results from ORA and the MR search to find the most promising candidates that might be implicated in cell or tissue migration.

In total, seven MRs (FLT4, GATA3, MMP9, PRKD1, PTGS2, PTPRM, and WNT5A) were associated with the gene "hits" in the ORA results that corresponded to the two BP terms "endothelial cell migration" and 'tissue migration". For example, FLT4 encodes the tyrosine kinase receptor VEGFR-3 which has been suggested as a metastatic marker as well as a response marker for small molecule therapeutics to suppress metastasis in CRC [123]. In addition, it has been reported that the expression of FLT4 (VEGFR-3) is controlled by upstream Wnt/β-catenin signaling in endothelial cell migration during angiogenesis [124]. Most importantly, our previous study confirms the biological relevance of VEGF for the underyling dataset [17]. The significant MRs for *Cluster 1* in are included in Table A5. With FLT4 (VEGFR-3) being one promising MR, I also visualized its corresponding master regulatory pathway using the geneXplain platform (Figure 5.18).

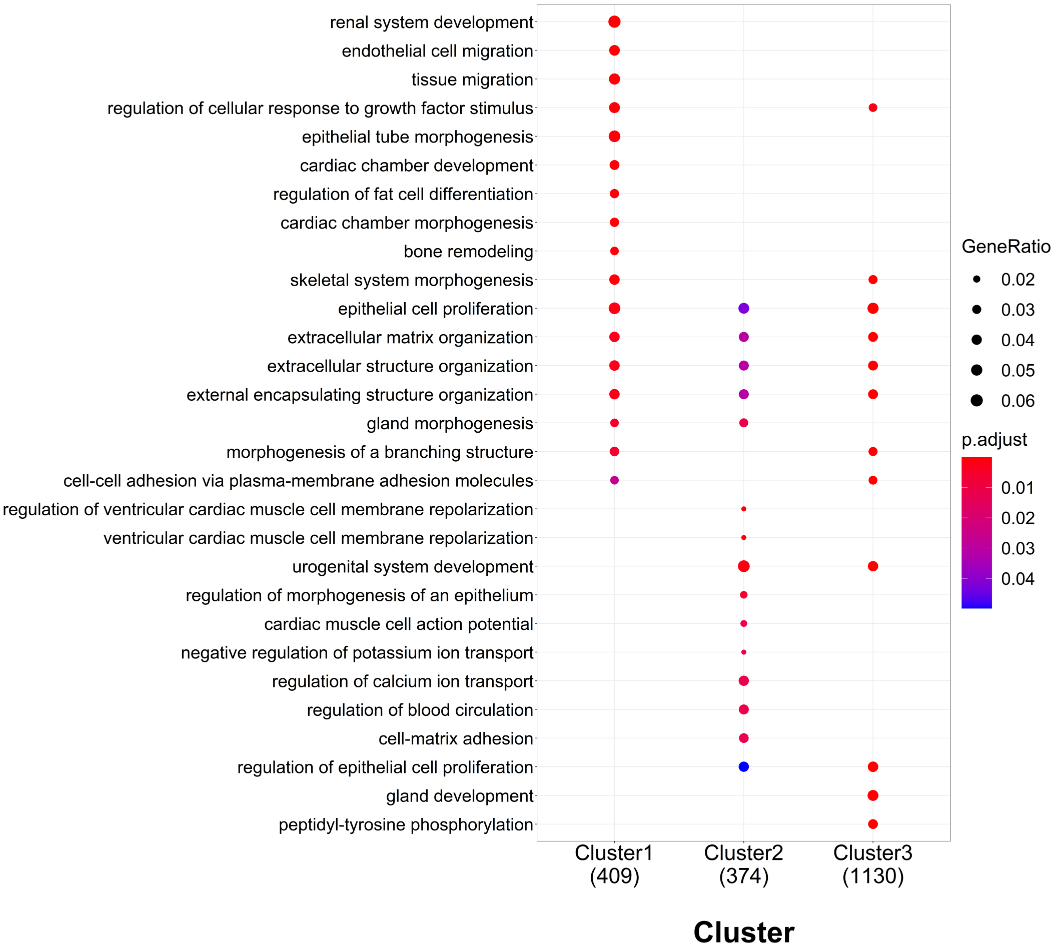


Figure 5.16.: Comparative functional categorization of gene clusters using GO BP terms. The most discriminative BP terms are listed in the top 10 significant BP terms (BH-adjusted p-value < 0.05) for each cluster. The *geneRatio* reflects the ratio of genes in each cluster (i.e., 409, 374, and 1130) that are annotated in a BP term.

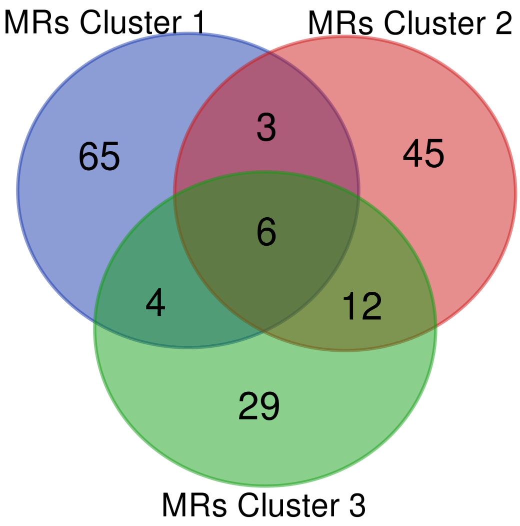


Figure 5.17.: Venn diagram of MRs that were obtained for the three gene clusters after MR search.

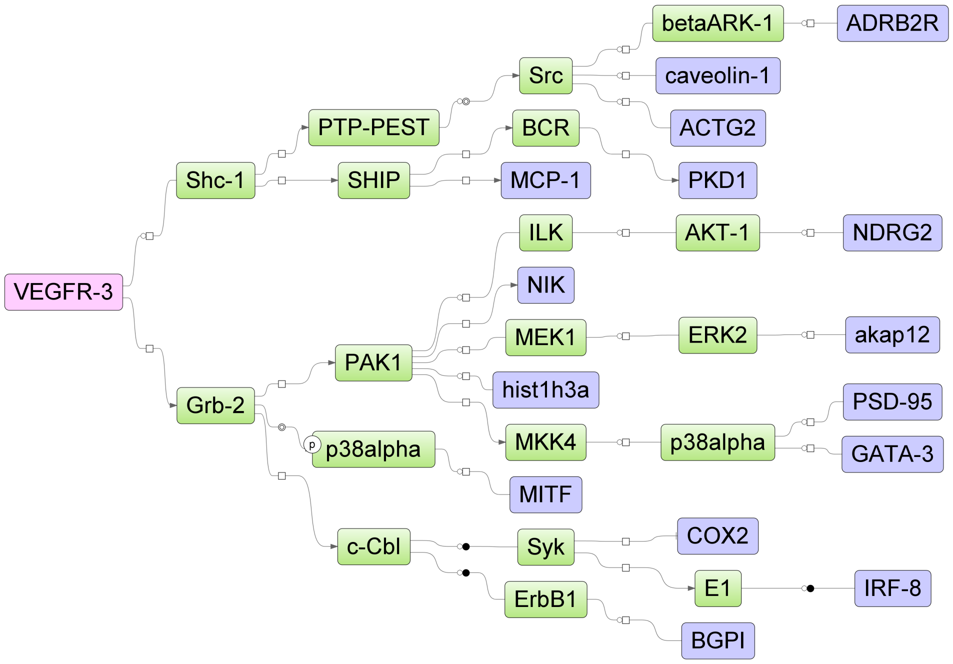


Figure 5.18.: VEGFR-3 (FLT4)-related master regulatory pathway based on genes in *Cluster 1*. Red, blue and green node colors represent the MR, regulated genes in *Cluster 1*, and connecting molecules from TRANSPATH®, respectively.

5.3. EGFR/WNT Pathway-related CRLM Dataset

In this section, I evaluated the utility of the pipeline to identify relevant BPs as well as MRs of signaling pathways in CRLM based on RNA-Seq data. For this purpose, I performed a comparative analysis of two patients with CRLM, namely *Patient I* and *Patient II*, for whom a poor outcome after treatment (i.e., cancer-related death) and a good outcome after treatment were described, respectively.

Alterations in the regulation of the EGFR pathway and the WNT pathway often contribute to the hallmarks of cancer and, in particular, to cancer progression. Therefore, the differences at both the transcriptional regulation and pathway levels between the two patients were examined for the comparison "*Patient I* versus *Patient II*". Consequently, the upregulated genes in *Patient I* were used as the target set, whereas the downregulated genes were used as the background set. When the direction of the comparison is changed, these downregulated genes can likewise represent upregulated genes in *Patient II*.

5.3.1. RNA-Seq data analysis and Promoter Sequence Extraction

For the analysis of the EGFR/WNT pathway-related CRLM dataset, RSEM-gene level count data was acquired from [115] and processes as outlined in *Workflow 1* (Figure 4.1). Thereafter, RNA-Seq data analysis was performed which involved the comparative transcriptome analysis of two patients with CRLM, denoted "*Patient I*" and "*Patient II*". The *Patient I*-related samples were considered as the target set, denoted "*Patient I)*". The '*Patient II*-related samples were considered as the background set, denoted "*Patient II*".
After initial explorative analysis of the gene expression levels across the samples, 1082 DEGs were identified between the two sample groups, of which 430 and 652 were significantly upregulated for *Target (Patient I)* and *Background (Patient II)*, respectively (Figure 5.19). After retrieving the corresponding promoter sequences for these genes, the *Target (Patient II)*-related set with 652 sequences (sorted by decreasing log 2 fold changes of the corresponding DEGs) was truncated at the bottom to match the number of 430 sequences in the *Background (Patient I)*-related set. For further analysis, the same labels were used for the two sequence sets, i.e., *Target (Patient I)* and *Background (Patient II)*.

5.3.2. Analyzing the GC Content

To examine whether there was a potential bias in GC content between promoter sequences in *Target (Patient I)* and *Background (Patient II)*, I analyzed the density distributions of the GC content which were computed and visualized by kernel density estimate (Figure 5.20). As indicated by the density distributions, the GC content in *Background (Patient II)* shows a shift in the curve towards the left, indicating an overall lower GC content when compared to *Target (Patient I)*.

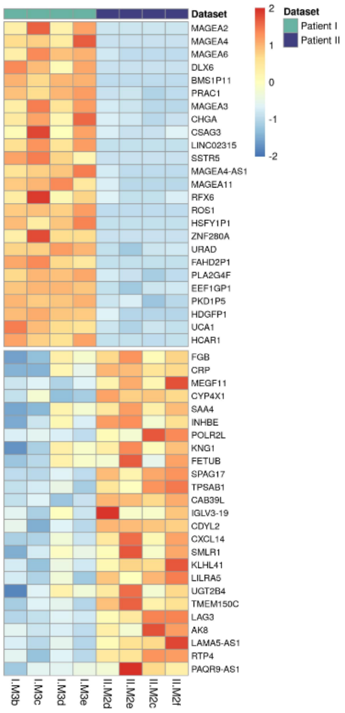


Figure 5.19.: Comparative transcriptomic analysis of samples from *Target (Patient I)* and *Background (Patient II)*. The heatmap visualizes the top 25 upregulated DEGs and the top 25 downregulated DEGs based on log 2 fold changes in the comparison. Count values related to gene expression were centered and scaled in row direction.

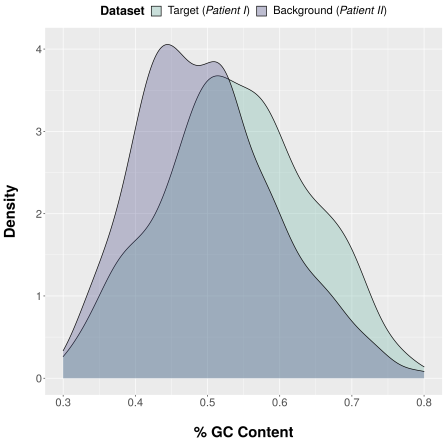


Figure 5.20.: Visualization of the density distributions of the GC content for promoter sequences (1100 bp in length) in *Target (Patient I)* and *Background (Patient II)*.

5.3.3. Optimizing MSS Cutoffs

After site search using the MATCH™ algorithm in *Workflow 2* (Section 4.1.2), the resulting site-specific matrices were subjected to Boruta in *Workflow 3* (Section 4.1.3) to find the most relevant MSS cutoffs per site model (or PWM) that best discriminated between target and background sets. When Boruta was used for the MSS cutoff optimization step in *Workflow 3* (Figure 4.3), 106 relevant PWMs were retained (Figure 5.21).

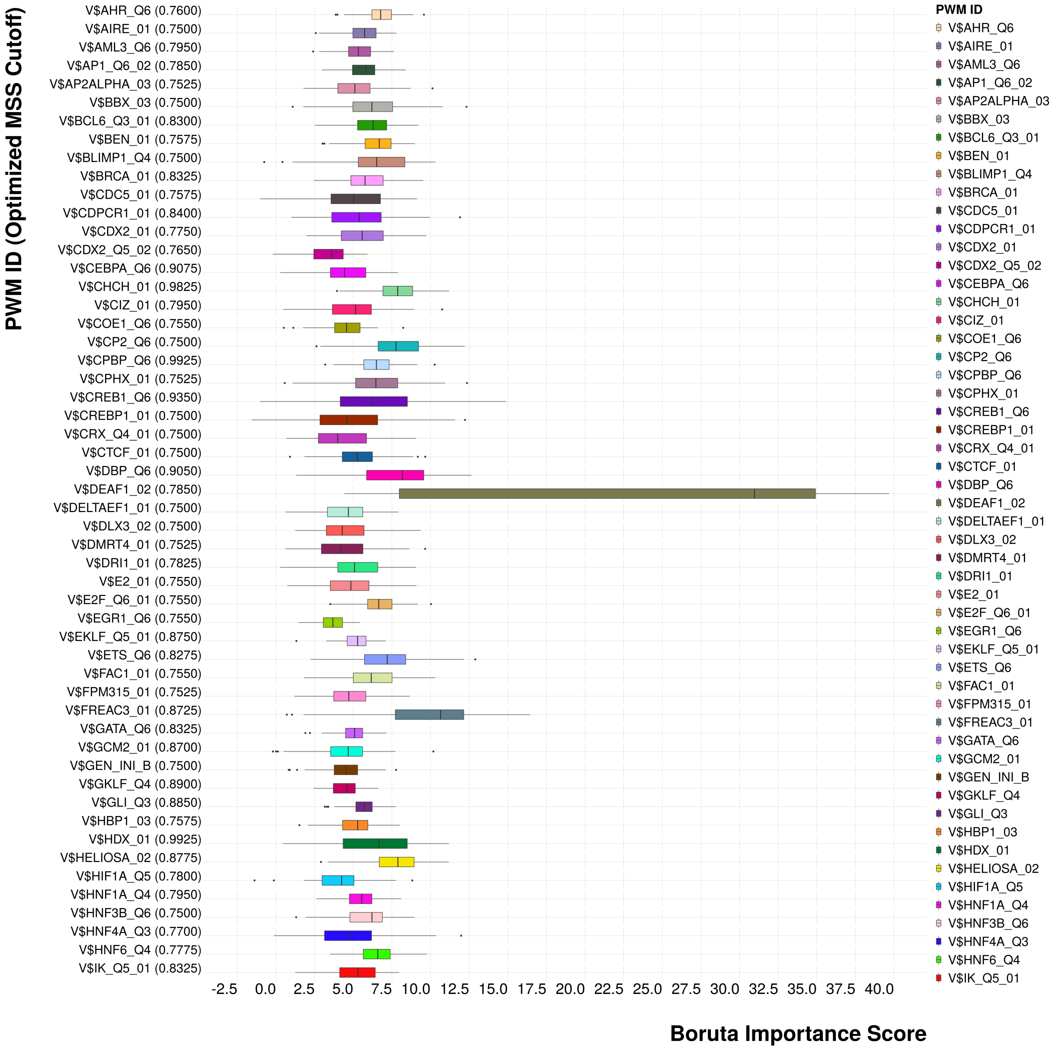


Fig. 5.21 (*cont'd*)

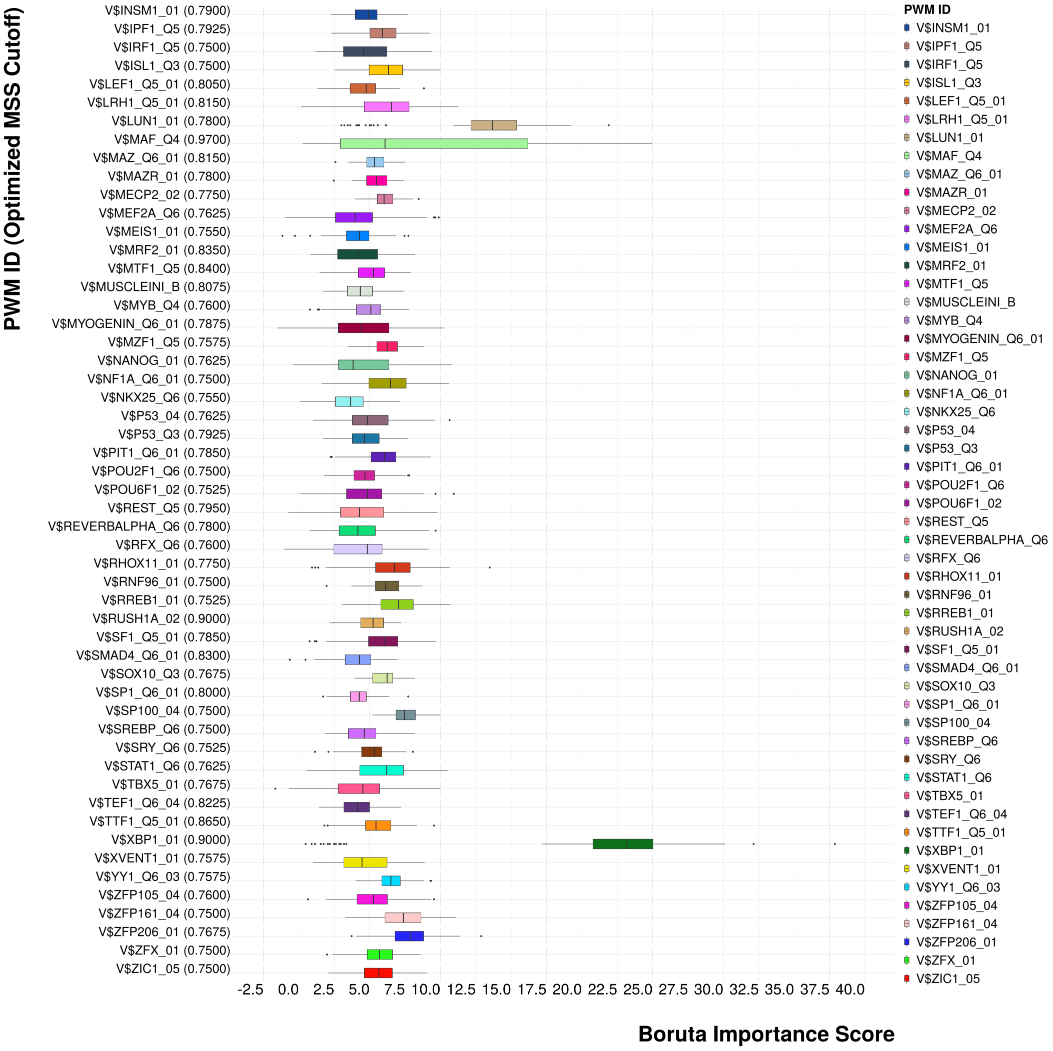


Figure 5.21.: Boruta results for the MSS cutoff optimization step using *Target (Patient I)* and *Background (Patient II)*. Results are sorted alphabetically by PWM. The forest plot summarizes all the PWMs retained at their most relevant MSS cutoffs after 163 site-specific Boruta runs. The boxplots summarize the importance scores obtained over 100 Boruta iterations in each run.

5.3.4. Identifying the most relevant Site Models

Proceeding with *Workflow 4* (Figure 4.4), the most relevant PWMs were identified in the analysis. Applying Boruta to the site-mixed count matrix, 37 out of 106 PWMs were retained, with V$MECP2_02 (MSS cutoff of 0.775) being the most relevant one according to the median importance score (Figure 5.22). V$MECP2_02 represents binding sites of the methyl-CpG binding protein 2 (MeCP2) that binds to methylated DNA or promoters to control chromatin organization and transcription [125]. Moreover, MeCP2 has been reported to promote CRC metastasis and its upregulation has been associated with poor prognosis [125].

5.3.5. Analyzing Site Over- and Underrepresentation

To further investigate how the average site frequencies that corresponded to the 37 relevant PWMs in *Target (Patient I)* compare to those in *Background (Patient II)*, target-to-background ratios (*T/B*) were calculated.

The results indicated that 24 out of 37 PWMs were associated with prominent site enrichments according to the *T/B* in *Target (Patient I)*, with V$DEAF1_02 (MSS cutoff of 0.785) having the highest *T/B* ratio of approximately 2.81 (Figure 5.23 and Table A6).

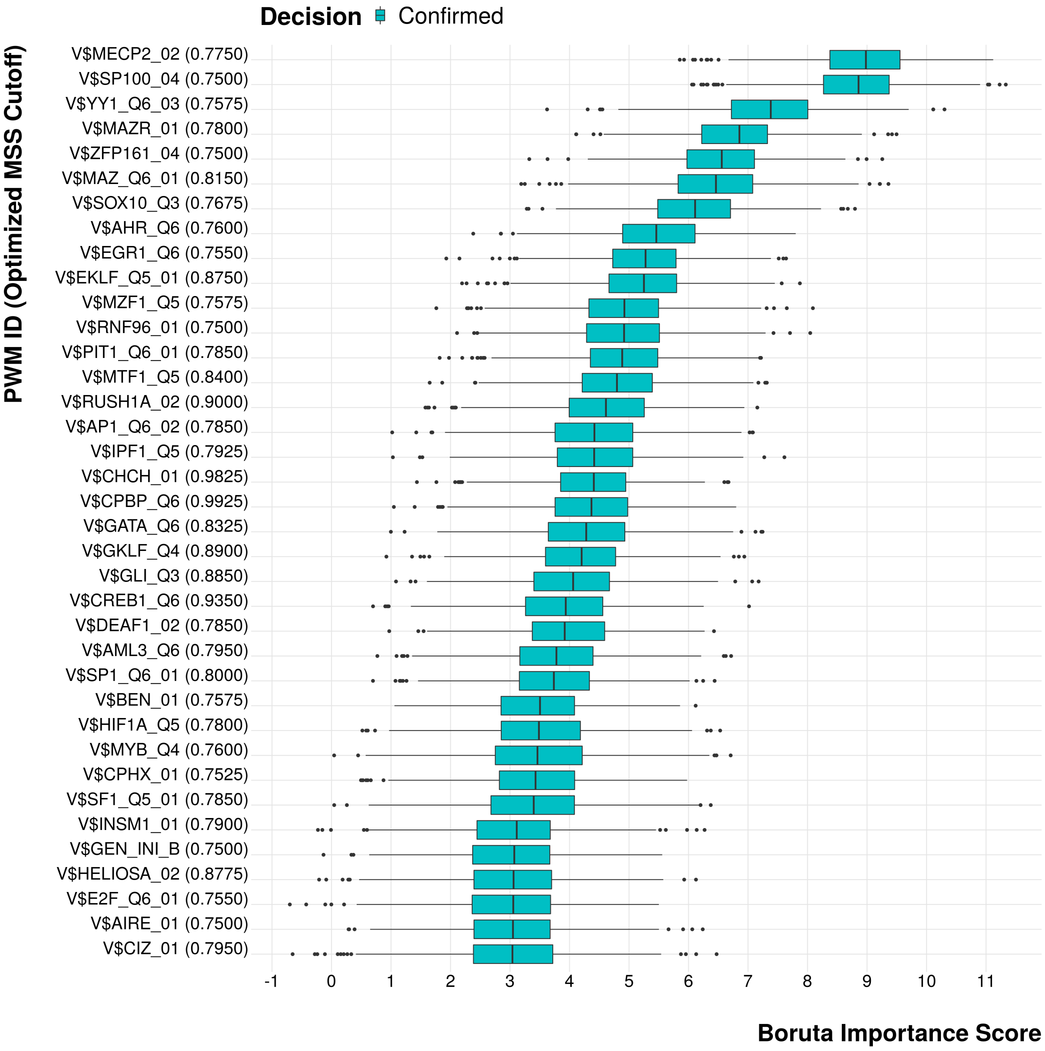
5.3.6. Retrieving Gene Clusters

Proceeding with *Workflow 5*, gene clusters consisting of any possible combination of DEGs (defined in *Workflow 1*) across *Target (Patient I)* and/or *Background (Patient II)* (Figure 4.5) were identified. As a results, the 860 DEGs were clustered into three gene clusters that contained 146 (*Cluster 1*), 222 (*Cluster 2*), and 492 (*Cluster 3*) genes (Figure 5.24).

Next, the overlap in genes between the 860 DEGs related to *Target (Patient I)* and *Background (Patient II)*, respectively, and the three clusters. *Cluster 1* showed the largest discrepancy between the overlaps. The overlap of *Cluster 1* with the DEGs in *Target (Patient I)* was more than twice as large as the overlap with the DEGs in *Background (Patient II)* (Table 5.4).

Table 5.4.: Gene cluster affiliations of DEGs related to *Target (Patient I)* and *Background (Patient II)*.

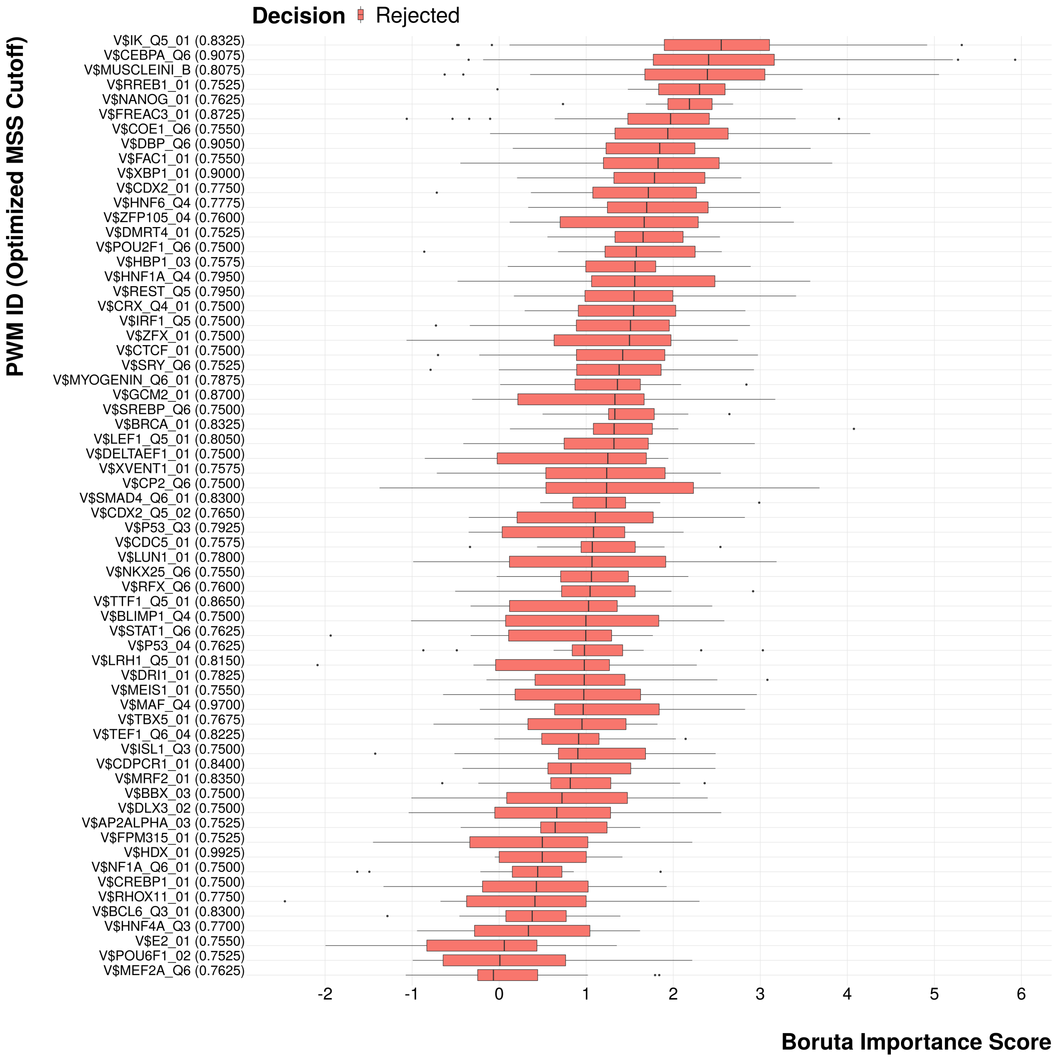
Dataset	***Cluster 1***	***Cluster 2***	***Cluster 3***
Target (Patient I)	98	76	256
Background (Patient II)	48	146	236



(a)

To test whether there was a significant association between the observed overlaps in Table 5.4, I performed a Pearson's *Chi-squared* of independence using the cluster affiliations as three variables in a three-way contingency table. The *Chi-squared* resulted in a *p*-value of 2.053e-09, which provides sufficient evidence to conclude that the observed overlaps are not likely due to chance.

Fig. 5.22 (*cont'd*)



(b)

Figure 5.22.: Boruta results showing the most relevant PWMs. The forest plot shows PWMs at their optimized MSS cutoffs (the features) that were either confirmed (a) or rejected (b) after the site-mixed Boruta run. The boxplots summarize their importance scores features using 1000 iterations in the Boruta run. PWMs are sorted by their median importance score.

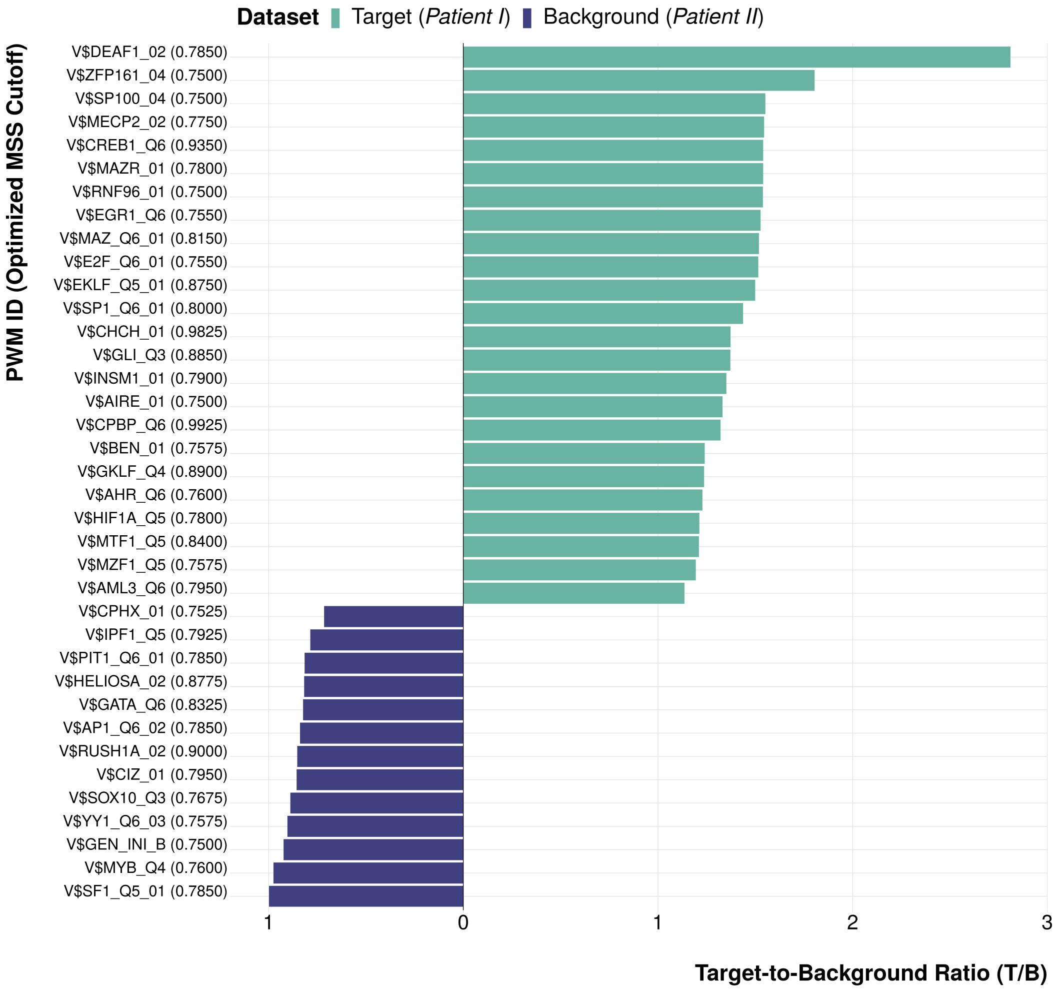


Figure 5.23.: Double-sided bar plot showing target-to-background ratios (*T/B*) for the most relevant PWMs.

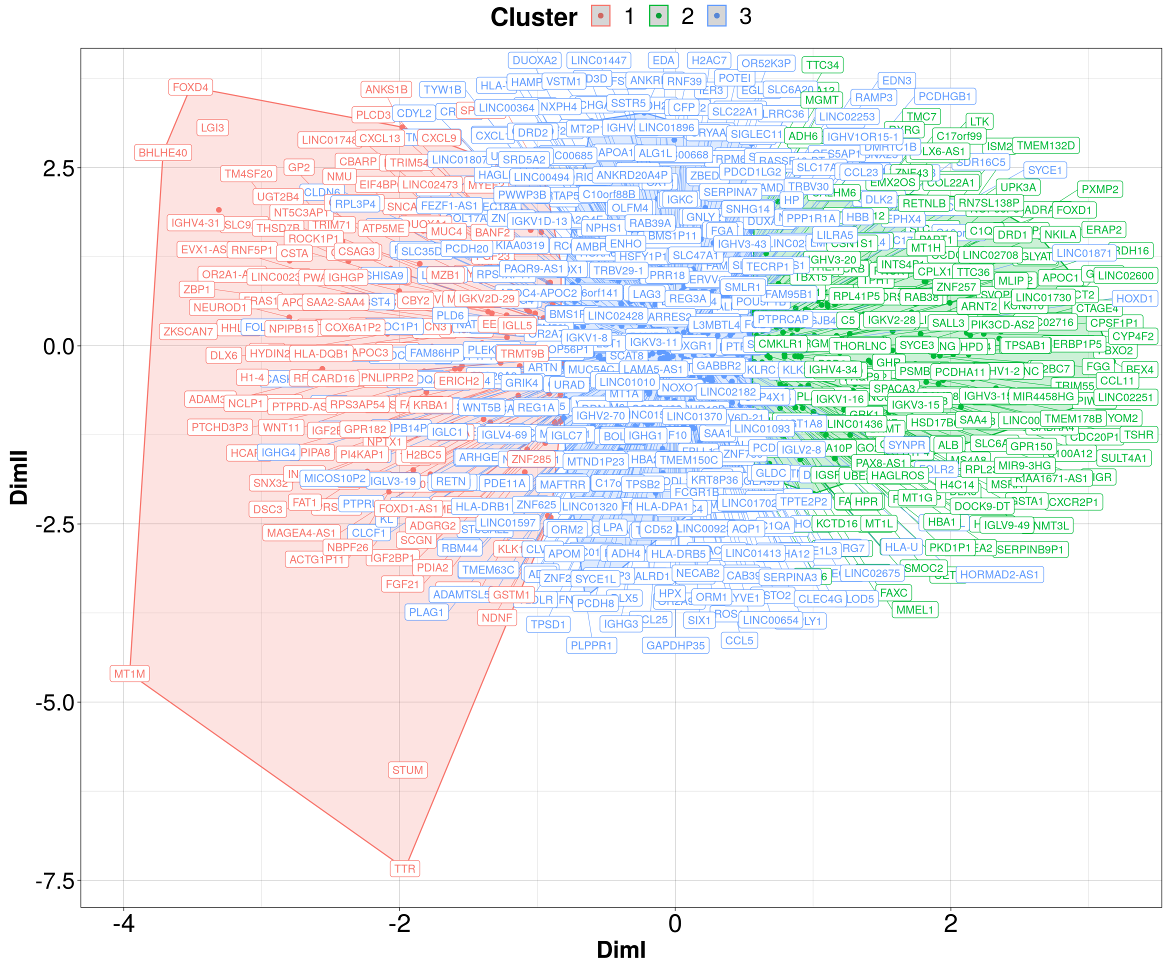


Figure 5.24.: PCA-based k-means clustering of 860 DEGs. Calculations are based on JSD-derived similarity measures between the probability distributions that relate to the 37 relevant PWMs. Different colors indicate the three different gene clusters. The optimal number of *k* clusters was determined by the silhouette method.

5.3.7. Analyzing GO Overrepresentation

To further assess whether the three clusters were associated with different biological themes, a comparative functional categorization was performed using ORA in combination with annotations defined in GO BP. Noteworthy, the R package clusterProfiler only considered 105/146 (*Cluster 1*), 164/222 (*Cluster 2*), and 367/492 (*Cluster 3*) in each cluster for ORA after gene ID conversion. As shown in Figure 5.25, the most discriminative BP terms are listed in the top 10 significant BP terms (BH-adjusted p-value < 0.05) for each cluster. Interestingly, the results revealed that the top 10 BP terms, e.g., “enteroendocrine cell differentiation”, “heart induction”, “secondary heart field specification”, “regulation of heart morphogenesis”, and “mesoderm formation”, for *Cluster 1* were not included in the significant BP terms for the other two clusters. The top 20 significant BP terms are shown for *Cluster 1* in Table A7.

5.3.8. Analyzing Dysregulated Pathways

Finally, MRs were searched for each gene cluster using the geneXplain platform to construct dysregulated pathways in cancer. As a result, 85, 133, and 46 MRs were identified for *Cluster 1*, *Cluster 2*, and *Cluster 3*, respectively. Thereafter, I analyzed the overlap between MRs related to the three clusters (Figure 5.26), which showed that most MRs were unique for each cluster. Since the previous results indicated a high relevance of *Cluster 1* for *Target (Patient I)*, I combined the results from ORA and the MR search to find the most promising MRs that might be associated with the most significant GO BP term, namely “enteroendocrine cell differentiation”, that was obtained for *Cluster 1*.

In total, the four MRs (BMP4, INSM1, NEUROD1, and RFX6) were associated with the gene “hits” in the ORA results that corresponded to this BP term. For example, the bone morphogenetic protein 4 (BMP4) is a ligand of the TGF-β signaling pathway that regulates the recruitment and activation of TFs that belong to the SMAD family. It has been suggested that BMP4 is involved in CRC metastasis via the activation of Rho and ROCK [126]. In addition, BMP4 overexpression has also been associated with Wnt/β-catenin signaling in CRLM, rendering it as an attractive therapeutic target for cancer intervention [127]. Most importantly, our previous study confirms the biological relevance of TGF-β/SMAD signaling and Rho/ROCK signaling for the underyling dataset [115]. I included the significant MR candidates for *Cluster 1* in Table A8. With BMP4 being one promising MR candidate at hand, I also visualized its corresponding master regulatory network using the geneXplain platform (Figure 5.27).

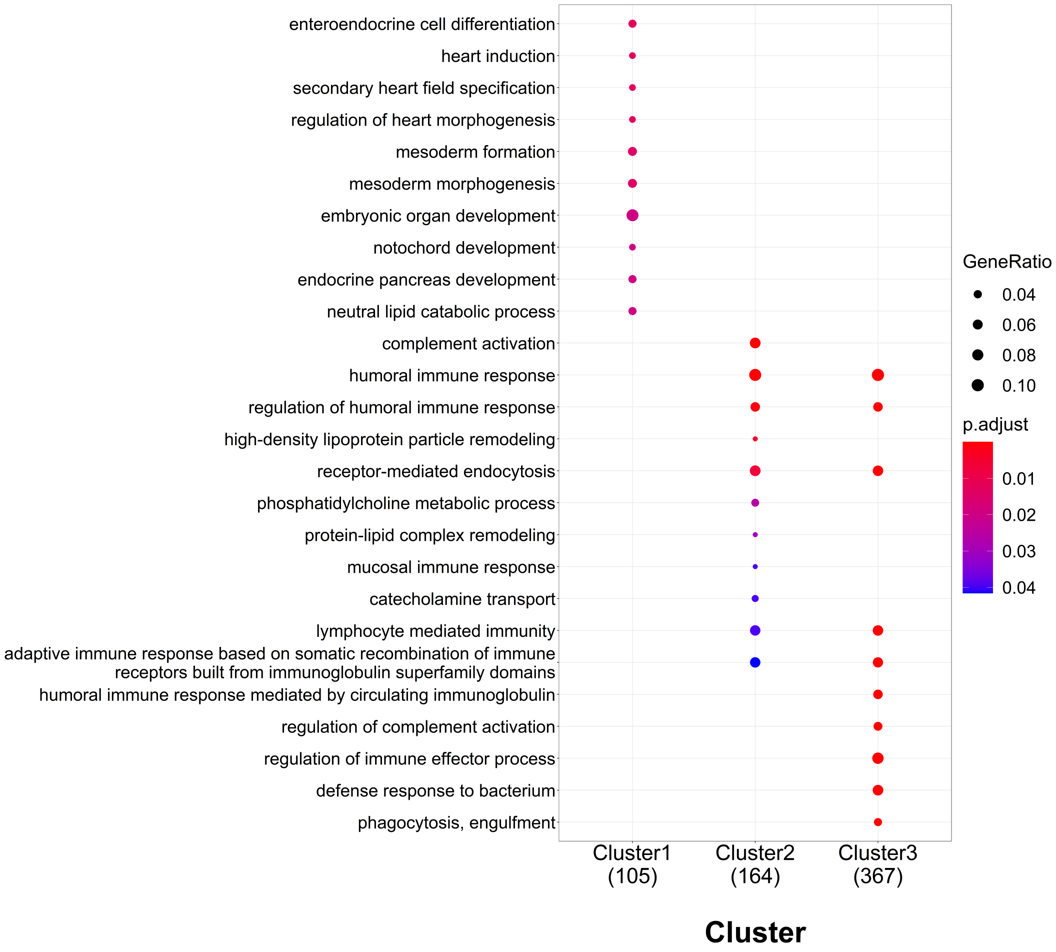


Figure 5.25.: Comparative functional categorization of gene clusters using GO BP terms. The most discriminative BP terms are listed in the top 10 significant BP terms (BH-adjusted *p*-value < 0.05) for each cluster. The *geneRatio* reflects the ratio of genes in each cluster (i.e., 105, 164, and 367) that are annotated in a BP term.

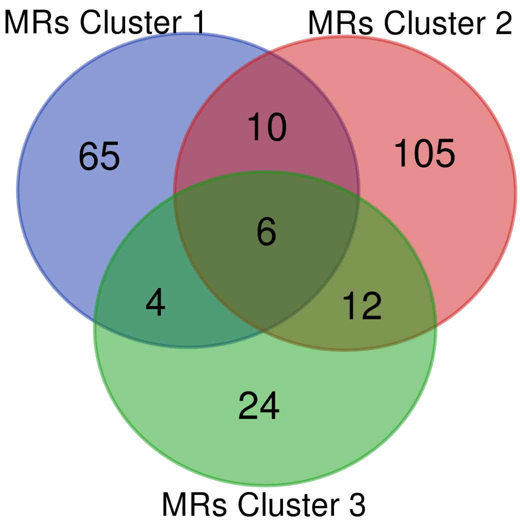


Figure 5.26.: Venn diagram of MRs that were obtained for the three gene clusters after MR search.

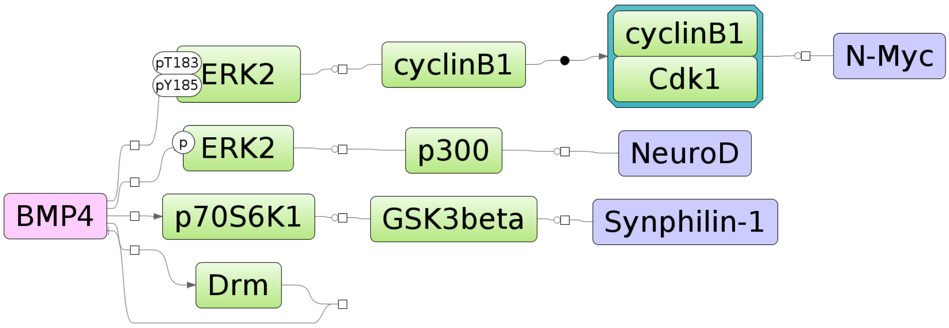


Figure 5.27.: BMP4-related master regulatory network based on genes in Cluster. Red, blue, and green node colors represent the MR, regulated genes in *Cluster 1*, and connecting molecules from TRANSPATH®, respectively.

6. Discussion

High-throughput technologies have revolutionized clinical practice by allowing for the customization of cancer patient diagnosis, risk assessment, and therapy [13]. Hence, clinicians rely on computational methods to analyze high-dimensional data and discover changes of clinical relevance. RNA-Seq is an example of a high-throughput application that demands sophisticated workflows. In the bioinformatics community, connecting multiple workflows to provide novel tools for studying phenotypes has become common practice. For this purpose, promoter analysis is often employed downstream of RNA-Seq data analysis to construct molecular networks from the results [17, 18, 29, 25, 30]. Promoter analysis usually yields a list of ranked TFs (or their associated site models) based on some importance measures, as well as a variety of intermediate outputs. Although a list of relevant TFs and their target genes is sufficient to construct molecular pathways, processing of intermediate outputs such as the number of site occurrences in promoters can be used to generate new types of input data for the linkage of downstream workflows. In addition, most methods in promoter analysis focus on the identification of overrepresented sites that discriminate between a set of target sequences and a set of background sequences. While there is a basic concept of site overrepresentation (or underrepresentation), it is unclear if there are other relevant hidden site patterns inside or outside of this concept that may not be found using conventional methods.

The main objective of this thesis was to develop a bioinformatics pipeline that links transcriptome profiling by RNA-Seq to crucial biological processes and dysregulated pathways through promoter analysis, thereby uncovering molecular key players and their interactions in pathways. At the same time, another objective was to evaluate an algorithm for promoter analysis based on feature selection which might extend the concept of site overrepresentation or relevance.

To evaluate the utility of the pipeline, it was applied to three reference data sets, with lists of relevant DEGs, TFBSs, biological processes (BPs), and master regulators (MRs) generated at the end of each workflow. All three reference datasets have already been characterized and analyzed in previous studies [19, 17, 115] with similar objectives. Therefore, some of the key findings of the pipeline could already be compared and validated with these studies (see Sections 5.1, 5.2, and 5.3).

In this Chapter, I will mainly address several methodological and biological aspects of each workflow of the pipeline. In addition, I will also discuss the choice of methods, including parameters and databases, and their main limitations.

6.1. Workflow 1 – RNA-Seq Data Analysis and Promoter Sequence Extraction

6.1.1. RNA-Seq Data Analysis

The pipeline was constructed with sequence data (FastQ format) from RNA-Seq as the entry point in mind; however, a list of DEGs can also be supplied as an entry point. In its standard application, DEGs are split into target and background sets that relate to upregulated genes and downregulated genes, respectively, which are supplied to the promoter extraction step. There are many existing standard workflows available to perform RNA-Seq data analysis and retrieve DEGs. However, there is no single workflow that works equally well for all use cases [15]. Therefore, the choice fell back on popular methods with a broad applicability, such as STAR [105], RSEM [106], and DESeq2 [107]. A description of the main challenges associated with RNA-Seq data analysis may be found in [15] for further reference. While any other combination of existing methods could have been used for these steps, the selection of cutoff parameters for the adjusted p-values and the $\log 2$ fold change should reflect reasonable choices to capture significant biological changes between the sets, especially when followed by promoter analysis. I adopted the analysis steps and the parameters from previous studies to identify DEGs [17, 108, 115].

6.1.2. Promoter Regions

The definition of a promoter length can vary and it will have an impact on the results. Larger promoter regions are more likely to contain an amount of erroneous binding site predictions while also capturing more true binding sites [128]. In this regard, the usage of vast areas upstream of the transcription start site (TSS) is especially inappropriate for promoter analysis [128]. In agreement with a previous study, promoters were defined as 1000 bp upstream and 100 bp downstream of the TSS [17]. As currently implemented, all promoters that correspond to the genes in both sets were considered for promoter analysis. However, it might be advantageous to filter out overlapping promoters that correspond to a single gene, thereby handling a possible overestimation of site occurrences. Another improvement for the pipeline that could additionally be implemented is an approach in comparative genomics called *phylogenetic footprinting* [129, 130, 131, 132]. It is based on the basic idea that functional regions in the DNA are subject to greater selective pressure and thus are conserved across moderately diverged species [132]. These conserved DNA regions can be uncovered by comparing sequences from orthologous genes. It has been reported for the oPOSSUM method that its performance is improved by restricting the search space to only conserved regions, thereby eliminating most non-functional sites [19].

6.2. Workflow 2 – TFBS search using the MATCH™ Algorithm

6.2.1. Single-Site Promoter Analysis

The bioinformatics pipeline identifies potential binding sites for several TFs independently, which is commonly known as *single-site promoter analysis*. This type of promoter analysis ignores the possibility of combinations of sites that may be required for TF binding. Several methods offer an alternative paradigm, with the objective of detecting cis-regulatory modules (CRMs) formed by clusters or co-occurrences of sites from multiple TFs [133, 134, 135, 66]. However, these methods have to deal with additional obstacles that include similarities between motifs, biases induced by site overrepresentation, and the difficulty of applying statistical tests to evaluate the relevance of observed site co-occurrences [133]. In addition, these methods must also deal with preferred distance arrangements within CRMs, which remains a difficult task, given that variable configurations of different module sizes exist in promoters.

Another general drawback associated with promoter analysis is that TF binding to distant regulatory regions, like enhancers, is usually not taken into consideration. Hence, enhancers are not included in the sets of promoter sequences. However, it should be emphasized that the discovery of enhancers is not as intuitive as the discovery of promoters [136]. Firstly, the general sequence code of enhancers is poorly understood. Secondly, enhancers are dispersed across the 98% of the human genome that does not encode proteins, creating a vast search space. Thirdly, they can be located upstream or downstream of genes as well as within introns. Forthly, they do not always operate on the nearest promoter of a gene and they can regulate multiple target genes. Most importantly, their activity might be limited to a certain tissue or cell type, a specific period in life, or specific physiological, pathological, or environmental circumstances [136]. While nowadays resources exist that could be used to derive enhancer-promoter interactions [137, 138, 139], available data on tissue or cell-type-specific enhancers is still too limited to ensure a broad applicability.

However, these limitations should be acknowledged, and overcoming them could drastically improve the utility of the pipeline by reducing the number of erroneous site predictions. Likewise, the FANTOM5 project [138, 139] could be used for single cases to retrieve tissue-specific promoters. Nevertheless, single-site promoter analysis is still commonly used as presented in the pipeline and has been shown to be quite effective in predicting biologically relevant sites in most use cases [26, 19, 27, 25, 28].

6.2.2. Site Models

PWMs are frequently utilized as a mathematical representation of TF binding specificity; however, there are several recognized limitations when searching for short or highly degenerate motifs in genomic sequences. For example, conventional PWM models consider that positions inside the motif are independent where each position contributes to the PWM score, which measures binding affinity, individually [140]. Therefore, a variety of proposed modifications of the PWM model have been proposed, such as Bayesian Markov models [141] and *protein-binding microarray* (PBM) models [142]. These models use combinations of the conventional PWM models to address the difficulties associated with the simplified assumptions underlying PWM models [21]. In addition, methods that combine conventional PWM models with additional restrictions on the physical characteristics of the DNA strand bound by a TF have been presented [143]. While these modified models can be beneficial in a few circumstances, it has been demonstrated that the conventional PWM models provide a better description of the binding site in the vast majority of biological use cases [20, 21]. Above all, conventional PWM models have been utilized in nearly unmodified form for more than three decades. Furthermore, compilations containing hundreds of PWMs are currently available in a number of databases [144, 22, 145, 71, 146, 147, 148, 149]. Since eukaryotes have a large number of TFs, site search often requires a large number of PWMs to cover the huge spectrum of TF binding motifs which leads to a high level of redundancy for promoter analysis [150]. For this reason, the TRANSFAC® vertebrate non-redundant PWM library was selected for this task. This library not only provides a pre-defined compilation of quality-checked models but also a maximal coverage, while reducing redundancy.

6.2.3. Site Search Algorithms

There are several pattern matching algorithms that utilize PWMs to predict potential sites, such as MATCH™ [22, 24], ConSite2 [62], or PROMO [63]. The MATCH™ algorithm was selected for site search because it can be easily deployed using pre-compiled TRANSFAC® PWM libraries with different MSS cutoff values to enable a range of search modes with varying degrees of stringency. The MATCH™ algorithm considers that mismatches in less conserved regions of sequences are accepted more readily when the frequencies defined by a PWM are multiplied with a information vector, whereas mismatches in highly conserved regions are strongly discouraged [24]. As a result, the performance in recognizing the sites is improved, compared to methods which do not employ the information vector [24]. Above all, the methods F-Match [24] and CiiiDER [151] have previously demonstrated the utility of the MATCH™ algorithm in combination with TRANSFAC® PWMs for promoter analysis.

When using the MATCH™ algorithm, a major challenge in site search is the selection of adequate MSS cutoffs. Employing a high PWM score cutoff means that the majority of weak binding sites are discarded, while using a lower cutoff means that predictions are excessively noisy and biological results are biased. To deal with this difficulty, for example, the cutoff can be set so that one site occurrence is predicted every 1 or 10 kb of the genome under study [148]. The site overrepresentation method F-Match sets the MSS cutoffs to reasonable initial values and then optimizes the cutoffs based on the notion that a certain site is overrepresented in the target promoters compared to the background promoters. The biological assumption here is that several target promoters are co-regulated by the same TF, which provides more confidence in the predictions because it is more likely in this case that the TF plays a relevant biological role for the promoters under study. Inspired by this idea, a similar approach was implemented for *Workflow 3* which starts with initial MSS cutoffs of 0.75 for site search. Other initial cutoffs values were not tested in the scope of the thesis but minimum values of 0.7 or 0.75 have been previously suggested for site search [19, 152, 153, 154].

6.3. Workflows 3 and 4 – MSS Cutoff Optimization and Identifying the most relevant Site Models (PWMs) using Feature Selection

At this point, I will cover both workflows together, since they are identical in terms of implementation and differ only in terms of the input site count matrices. Both workflows are basically deployed in the same way to select adequate MSS cutoffs and relevant sites. Since there are significant differences in length, GC content, information content, and other PWM characteristics such as the presence of half-sites, I opted to run each workflow independently rather than combining them to avoid a bias in *Workflow 3*.

6.3.1. Feature Selection for Promoter Analysis

The bioinformatics pipeline uses an approach that is based on feature selection within a random forest (RF) framework for promoter analysis. To this end, feature selection was applied with the proposed objectives of selecting adequate MSS cutoffs and relevant sites which are defined as the most relevant predictor features based on their importance scores. In this context, importance scores reflect how well the features partitioned the data into the two set classes (i.e., the target and background sets).

RF has previously been utilized in gene expression studies to select subsets of genes based on gene relevance rankings for sample classification [37]. As a result, RF has proven to be a well-suited algorithm for gene expression data analysis by providing measures of feature importance. Most importantly, it has performed well, even when the majority of the predictive features are highly correlated or present noise, and the number of features is significantly greater than the number of observations [37]. Applying RF to gene expression data has also suggested that the default parameters do require little or no fine-tuning at all [37]. In analogy to gene expression data, the site count matrix that is subjected to the feature selection algorithm in the pipeline presents itself with similar characteristics where features are redundant, highly correlated and noisy. The Boruta feature selection algorithm was used for promoter analysis because it implements a wrapper around RF which allows to select features based on their importance of classification. It discovers discriminative patterns from the sets and makes a classification decision according to these discriminative patterns. Most importantly, Boruta can recursively eliminate features over a series of iterations which did not perform well, thereby reducing the error of the RF model. Therefore, the Boruta algorithm is ideal for analyzing the site count matrix with irregular and hidden patterns.

The results presented in Sections 5.1, 5.2, and 5.3 indicated that Boruta's ranking-based importance scores can be used to identify relevant sites that exhibit discriminative patterns between the sets, while their associated *target-to-background* ratios can additionally indicate a potential over- and underrepresentation (Figures 5.6, 5.7, 5.14, and 5.23). The biological relevance of these sites was particularly evident in the analysis of the NF-κB pathway-related dataset, where the pipeline reported sites corresponding to the NF-κB family of TFs.

So far, the exact discriminative patterns on which Boruta's decision is based remain unknown in this thesis. Boruta might select patterns stemming from sites that are relatively abundant or overrepresented in target sequences, while being missing or underrepresented in the background sequences, or vice versa. Overall, the pipeline can select discriminative patterns that may relate to both site overrepresentation and underrepresentation in a single run, both of which are equally applicable for the gene clustering approach presented in *Worklow 5*.

Focusing on site overrepresentation, the methods comparison between the pipeline and the two conventional overrepresentation methods, F-Match and oPOSSUM, showed that all three methods can identify *a priori* known biologically relevant sites (Tables 5.1 and 5.2). However, they also identified a considerable number of unique sites (Figures 5.8 and 5.9). The latter observation suggests that each method might consider different aspects of overrepresentation or relevance, which further underpins that the notion of overrepresentation (or underrepresentation) is mainly reliant on how the different algorithms are implemented. As a result, it remains difficult to compare and interpret the results of different algorithms and to determine whether one tool performs better than the others when applied to heterogeneous datasets. Nevertheless, it would be beneficial to interpret Boruta's decision as future work, i.e., which patterns between target and background sets have lead to the selection of the relevant sites. Understanding the feature selection process may provide at least the possibility to better characterize the concept of site over- and underrepresentation. In addition, the insights may also help to fine-tune the feature selection procedure of the pipeline. Given that random forest is built on decision trees, it leaves space for the interpretation of its predictions [155, 156, 157]. Deciphering the path (or paths) from the root to the leaf in the decision tree is possible and it may may disclose intriguing patterns in promoter analysis.

However, categorizing the predictions as biologically relevant or not remains problematic without further manual validation using literature or experimental validation, especially for datasets with such a multitude of variables. Therefore, as demonstrated in *Workflow 5*, functional categorization and pathway analysis are viable steps in downstream analysis to biologically interpret the performance of the pipeline at a more comprehensive level.

6.3.2. The Role of the Background Set for Promoter Analysis

Choosing a suitable background set for promoter analysis is not an easy task since there is no gold standard choice and most methods follow various guidelines in this regard. Furthermore, there is a strong implication that promoter analysis might be affected by differences in GC content between the target and background sets [19, 81, 80].

The results revealed that using a non-GC content-matched background set instead of a matched one resulted in the prediction of more relevant sites, as observed for the pipeline, F-Match, and oPOSSUM (Figures 5.8 and 5.9). Hence, it cannot be ruled out that this observation is at least partly attributable to the use of a non-GC content-matched background set without analyzing additional datasets. However, while the results of the pipeline showed that the reference PWM V$RELA_Q6 (NF-$\kappa$B, c-REL, p65, and p50) ranked higher in the top ten for the GC content-matched background set based on the importance score, the corresponding *target-to-background* ratio was approximately seven times higher in the presence of a higher optimized MSS cutoff when the non-GC content-matched background set was utilized (Figures 5.6 and A2). This observation highlights the importance of the cutoff optimization step during promoter analysis and the usefulness of the *target-to-background* ratio as an independent and simple measure for site overrepresentation and underrepresentation.

The question of whether or not to employ a GC-matched background set with regard to the pipeline cannot be completely resolved at this point since additional non-GC content-matched background sets have to be analyzed for the same dataset. In this context, it should be highlighted the top 10 results did not indicate a particular bias towards the identification of GC-rich sites in the analyses (Tables 5.1, 5.2, and B1). Considering the main findings, it is still more advisable to use a background set that is biologically related to the target set (e.g., DEG promoters) than a random promoter set adjusted for GC content to reduce the risk of missing biologically relevant GC-rich sites in promoter analysis.

6.4. Workflow 5 – Identification of Gene Clusters and Dysregulated Molecular Pathways

6.4.1. Gene Clusters

To connect the results obtained from promoter analysis to the next steps which involve comprehensive functional categorization and pathway analysis, genes are clustered based on similarities between binding site patterns across target and background sets. In this context, high similarity between any two genes reflects that most TFs likely to bind to the corresponding relevant sites regulate the genes in a similar manner, with gene co-regulation defined here as two genes sharing at least one TF. Importantly, co-regulation might involve TFs that act as transcriptional activators or repressors, thereby influencing the expression of individual genes as a result of their combined activity.

In gene expression studies, clustering of genes based on their gene expression patterns across different groups of samples is a common practice [158, 159, 160, 161, 162, 163]. The rationale behind this approach is that genes with highly similar gene expression patterns are likely controlled by the same biological processes, form protein complexes, or have similar topological characteristics in protein-protein interaction networks, and are also more likely to be co-regulated by the same TF [159, 160]. However, it has also been reported that genes that can be functionally categorized together do not always show correlated gene expression [164].

Hence, the objective of the types of analysis presented in *Workflow 5* was to investigate whether genes that are similar in terms of their binding site patterns are also more similar to each other in functional terms, analogous to gene expression studies.

For gene clustering, a combinatorial usage of JSD, PCA, and k-means clustering is proposed. The application of information theory-based measures such as *pointwise mutual information* (PMI) to site count matrices has already been demonstrated in the context of promoter analysis [165]. Likewise, the JSD is widely applied in the field of bioinformatics [38, 39, 40, 41, 42, 43] to measure the similarity between two probability distributions. A combinatorial usage of Jensen-Shannon distance (i.e., the square root of the JSD) and PCA has been previously proposed for clustering of count values that reflect abundance levels of bacteria [43]. Noteworthy, the JSD was applied with *Laplace smoothing* [112] to the count matrices in the thesis, which adds a small pseudocount to each count to overcome the problem introduced by the presence of zeros counts (Section 4.1.5).

Data pre-processing with PCA is a widely utilized exploratory tool in multivariate analysis to project data to a lower dimensional subspace, and clustering via the k-means algorithm has already been shown to operate effectively on data with low effective dimensionality [166, 167, 168, 169, 170]. The construction of the JSD similarity matrix, PCA, and k-means clustering was performed using the framework presented in the R package immunarch (`https://CRAN.R-project.org/package=immunarch`).

6.4.2. Biological Processes and Master Regulators

The understanding of large lists of genes results can be immensely assisted by prior biological knowledge from gene set annotations and pathway knowledge. Functional ORA is a method for assessing the statistical significance of the overlap of a set of genes of interest with the set of genes known to be associated with, e.g., a particular BP [16].

Comparing biological themes among gene clusters after clustering analysis is commonly used to reveal differences between two conditions in cancer. However, a large number of highly similar GO BP terms are often dominating the top overrepresented terms [113, 171, 172, 173, 174]. To overcome this issue, the R package clusterProfiler (`https://bioconductor.org/packages/clusterProfiler/`) was utilized because it can eliminate redundant terms, thereby retaining only representative terms.

The ORA results uncovered unique BP terms for each gene cluster that have been previously associated with the phenotypes under study in the analysis of the WNT pathway-related CRC dataset (5.2 and the EGFR/WNT pathway-related CRLM dataset (5.3) [17, 115].

To demonstrate that the gene clusters could be governed by unique MRs in signaling pathways, the MRA of the geneXplain platform (`https://genexplain.com`) was applied to the gene clusters. This type of analysis has often been used in combination with promoter analysis to establish causal regulatory links between TFs and their gene targets, enabling the identification of MRs and their dysregulated pathways. A popular strategy used in comparative cancer studies consists of two fundamental steps [17, 18]: (i) analysis of DEG promoters to discover relevant TFs using the TRANSFAC® database [22, 23] and the MATCH™ algorithm [22, 24]; (ii) discovery of signal transduction pathways that activate these TFs, and discovery of MRs that regulate these pathways using the TRANSPATH® database [31]

A modified strategy is proposed in the thesis as part of the pipeline that covers all the same steps from previous studies but performs the MRA based on the gene clusters instead of lists of TFs. The results indicated that most MRs were unique for each gene cluster in both analyses (Figures 5.17 and 5.26). This is an intriguing finding since it implies that the gene clusters obtained are subject to different regulatory pathways. To select the most promising MRs for a gene cluster, the results of ORA and MR search can be coupled by manually searching for MRs included in the gene hits of significant GO terms. This additional step revealed potentially dysregulated pathways, providing insight into the regulatory interactions of key players (Figures 5.18 and 5.27).

7. Conclusion

7.1. Summary

In this thesis, a bioinformatics pipeline was presented for identifying dysregulated pathways in cancer from comparative RNA-Seq transcriptome analysis. The individual workflows of the pipeline comprise methods in RNA-Seq data analysis, promoter analysis, comprehensive functional categorization, and pathway analysis.

RNA-Seq data analysis is performed using popular methods (STAR, RSEM, and DESeq2) with a broad applicability for the pre- and post-processing steps to identify DEGs. Promoter analysis represents the core of the pipeline and serves as a basis for linking transcriptome profiles generated from RNA-Seq to important biological processes and dysregulated pathways in cancer. To computationally identify TFBSs in target and background promoter sets that relate to upregulated and downregulated genes, respectively, the MATCH™ algorithm in conjunction with *position weight matrices* (PWMs) from the TRANSFAC® database is applied to promoter sequences. This setup is used to create a site count matrix which characterizes frequencies of site occurrences spanning both sets.
Next, the Boruta feature selection algorithm is applied to the data matrix with the stated goals of optimizing PWM score cutoffs and identifying relevant sites based on their discriminative site patterns, i.e., how effectively they partition the data into the two set classes. Thereafter, gene clustering is performed based on the assumption that genes associated with similar site patterns are presumably regulated by the same biological processes and pathways. To this end, a clustering approach based on the JSD, PCA, and k-means algorithm is applied to the patterns of relevant sites to identify gene clusters. To discover biological processes and master regulators in pathways, GO ORA using the R package clusterProfiler and TRANSPATH® MRA using the geneXplain platform are applied to each gene cluster.

The utility of the pipeline was demonstrated using three heterogenous gene expression studies that are characterized by distinct pathway activity in cancer. In the course of promoter analysis, the results indicated that Boruta's ranking-based importance scores can be used to identify biologically relevant sites that exhibit discriminative patterns between the sets, while their corresponding *target-to-background* ratios can additionally indicate a potential over- and underrepresentation. In the course of gene clustering, the results indicated clearly separated clusters that were characterized by uniquely significant BP terms and master regulatory pathways.

In conclusion, the pipeline provides a useful tool for the comparative study of phenotypes in cancer and uncovers variations in transcriptional regulation and pathway repertoire.

7.2. Outlook

Crucially, the bioinformatics pipeline is capable of analyzing heterogeneous datasets based on gene expression data *on-the-fly*. In addition, it does not rely on external data resources, which in many cases are difficult to obtain. I discussed several potential improvements for the steps in promoter analysis that could be realized to reduce the number of erroneous predictions without major compromises in terms of broad applicability. These potential improvements include: (i) obtaining tissue-specific promoters from the FANTOM5 database; (ii) filtering out overlapping promoters; and (iii) filtering out non-conserved promoter regions via phylogenetic footprinting.

Although feature selection was effective in achieving the critical steps in promoter analysis, i.e. the optimization of PWM score cutoffs and the identification of important sites, there is still a need to interpret Boruta's decision-making. Interpreting the RF process may provide further insight into the relevant site patterns, which in turn can be useful for fine-tuning of the feature selection procedure.

Bibliography

[1] DeOcesano-Pereira C, Velloso FJ, Carreira ACO, Ribeiro CSP, Winnischofer SMB, Sogayar MC, Trombetta-Lima M: **Post-Transcriptional Control of RNA Expression in Cancer**. In *Gene Expression and Regulation in Mammalian Cells - Transcription From General Aspects*, InTech 2018.

[2] Hahn S: **Transcriptional regulation. Meeting on regulatory mechanisms in eukaryotic transcription.** *EMBO reports* 2008, **9**:612–616.

[3] Butler JEF, Kadonaga JT: **The RNA polymerase II core promoter: a key component in the regulation of gene expression.** *Genes & development* 2002, **16**:2583–2592.

[4] O'Malley RC, Huang SSC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR: **Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape.** *Cell* 2016, **165**:1280–1292.

[5] Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics (Oxford, England)* 2000, **16**:16–23.

[6] Stormo GD: **Modeling the specificity of protein-DNA interactions.** *Quantitative biology (Beijing, China)* 2013, **1**:115–130.

[7] Ruan S, Stormo GD: **Inherent limitations of probabilistic models for protein-DNA binding specificity.** *PLoS computational biology* 2017, **13**:e1005638.

[8] Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, Zubieta C, Kaufmann K, Parcy F: **Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants.** *Molecular plant* 2019, **12**:743–763.

[9] Boulesteix AL, Janitza S, Kruppa J, König IR: **Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2012, **2**(6):493–507.

[10] Wingender E, Schoeps T, Haubrock M, Dönitz J: **TFClass: a classification of human transcription factors and their rodent orthologs.** *Nucleic acids research* 2015, **43**:D97–102.

[11] Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144**:646–674.

[12] Libermann TA, Zerbini LF: **Targeting transcription factors for cancer gene therapy.** *Current gene therapy* 2006, **6**:17–33.

[13] Gagan J, Van Allen EM: **Next-generation sequencing to guide cancer therapy.** *Genome medicine* 2015, **7**:80.

[14] Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature reviews. Genetics* 2009, **10**:57–63.

[15] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcze?niak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: **A survey of best practices for RNA-seq data analysis.** *Genome biology* 2016, **17**:13.

[16] Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS computational biology* 2012, **8**:e1002375.

[17] Wlochowitz D, Haubrock M, Arackal J, Bleckmann A, Wolff A, Beißbarth T, Wingender E, Gültas M: **Computational Identification of Key Regulators in Two Different Colorectal Cancer Cell Lines.** *Frontiers in genetics* 2016, **7**:42.

[18] Kalya M, Kel A, Wlochowitz D, Wingender E, Beißbarth T: **IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme.** *Frontiers in genetics* 2021, **12**:670240.

[19] Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.** *Nucleic acids research* 2005, **33**:3154–3164.

[20] Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity.** *Nature biotechnology* 2011, **29**:480–483.

[21] Dabrowski M, Dojer N, Krystkowiak I, Kaminska B, Wilczynski B: **Optimally choosing PWM motif databases and sequence scanning approaches based on ChIP-seq data.** *BMC bioinformatics* 2015, **16**:140.

[22] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic acids research* 2006, **34**:D108–D110.

[23] Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic acids research* 1996, **24**:238–241.

[24] Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic acids research* 2003, **31**:3576–3579.

[25] Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E: **Beyond microarrays: find key transcription factors controlling signal transduction pathways.** *BMC bioinformatics* 2006, **7 Suppl 2**:S13.

[26] Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells.** *Genome research* 2003, **13**:773–780.

[27] Joshi H, Nord SH, Frigessi A, Børresen-Dale AL, Kristensen VN: **Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes.** *BMC genomics* 2012, **13**:199.

[28] Zambelli F, Pesole G, Pavesi G: **Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes.** *Nucleic acids research* 2009, **37**:W247–W252.

[29] Kel AE, Stegmaier P, Valeev T, Koschmann J, Poroikov V, Kel-Margoulis OV, Wingender E: **Multi-omics "upstream analysis" of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer.** *EuPA open proteomics* 2016, **13**:1–13.

[30] Koschmann J, Bhar A, Stegmaier P, Kel AE, Wingender E: **"Upstream Analysis": An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data.** *Microarrays (Basel, Switzerland)* 2015, **4**:270–286.

[31] Schacherer F, Choi C, Götze U, Krull M, Pistor S, Wingender E: **The TRANSPATH signal transduction database: a knowledge base on signal transduction networks.** *Bioinformatics (Oxford, England)* 2001, **17**:1053–1057.

[32] Leung YF, Cavalieri D: **Fundamentals of cDNA microarray data analysis.** *Trends in genetics : TIG* 2003, **19**:649–659.

[33] Quackenbush J: **Computational analysis of microarray data.** *Nature reviews. Genetics* 2001, **2**:418–427.

[34] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D: **Knowledge-based analysis of microarray gene expression data by using**

support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:262–267.

[35] Vohradský J: **Neural network model of gene expression.** *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2001, **15**:846–854.

[36] Breiman L: **Random forests**. *Machine Learning* 2001, **45**:5–32.

[37] Díaz-Uriarte R, Alvarez de Andrés S: **Gene selection and classification of microarray data using random forest.** *BMC bioinformatics* 2006, **7**:3.

[38] Cumbo F, Vergni D, Santoni D: **Investigating transcription factor synergism in humans.** *DNA research : an international journal for rapid publication of reports on genes and genomes* 2018, **25**:103–112.

[39] D'Alessio AC, Fan ZP, Wert KJ, Baranov P, Cohen MA, Saini JS, Cohick E, Charniga C, Dadon D, Hannett NM, Young MJ, Temple S, Jaenisch R, Lee TI, Young RA: **A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity.** *Stem cell reports* 2015, **5**:763–775.

[40] Gültas M, Düzgün G, Herzog S, Jäger SJ, Meckbach C, Wingender E, Waack S: **Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming**. *BMC Bioinformatics* 2014, **15**.

[41] Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics (Oxford, England)* 2007, **23**:1875–1882.

[42] Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, Sperling S: **Combinatorial effects of four histone modifications in transcription and differentiation.** *Genomics* 2008, **91**:41–51.

[43] Arumugam M, , Raes J, Pelletier E, Paslier DL, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J, Weissenbach J, Ehrlich SD, Bork P: **Enterotypes of the human gut microbiome**. *Nature* 2011, **473**(7346):174–180.

[44] Kursa MB, Rudnicki WR: **Feature Selection with theBorutaPackage**. *Journal of Statistical Software* 2010, **36**(11).

[45] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Molecular biology and evolution* 2003, **20**:1377–1419.

[46] Atchison ML: **Enhancers: mechanisms of action and cell specificity.** *Annual review of cell biology* 1988, **4**:127–153.

[47] Yuh CH, Bolouri H, Davidson EH: **Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.** *Science (New York, N.Y.)* 1998, **279**:1896–1902.

[48] Masquilier D, Sassone-Corsi P: **Transcriptional cross-talk: nuclear factors CREM and CREB bind to AP-1 sites and inhibit activation by Jun.** *The Journal of biological chemistry* 1992, **267**:22460–22466.

[49] Wang D, Hu Y, Zheng F, Zhou K, Kohlhaw GB: **Evidence that intramolecular interactions are involved in masking the activation domain of transcriptional activator Leu3p.** *The Journal of biological chemistry* 1997, **272**:19383–19392.

[50] Lee S, Kohane I, Kasif S: **Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes.** *BMC genomics* 2005, **6**:168.

[51] Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM: **Housekeeping genes tend to show reduced upstream sequence conservation.** *Genome biology* 2007, **8**:R140.

[52] Trinklein ND, Aldred SJF, Saldanha AJ, Myers RM: **Identification and functional analysis of human transcriptional promoters.** *Genome research* 2003, **13**:308–312.

[53] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nature reviews. Genetics* 2009, **10**:252–263.

[54] Luscombe NM, Austin SE, Berman HM, Thornton JM: **An overview of the structures of protein-DNA complexes.** *Genome biology* 2000, **1**:REVIEWS001.

[55] Wingender E: **Criteria for an updated classification of human transcription factor DNA-binding domains.** *Journal of bioinformatics and computational biology* 2013, **11**:1340007.

[56] Wingender E, Schoeps T, Haubrock M, Krull M, Dönitz J: **TFClass: expanding the classification of human transcription factors to their mammalian orthologs.** *Nucleic acids research* 2018, **46**:D343–D347.

[57] Wingender E, Schoeps T, Dönitz J: **TFClass: an expandable hierarchical classification of human transcription factors.** *Nucleic acids research* 2013, **41**:D165–D170.

[58] Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57–70.

[59] Pavlidis N, Pentheroudakis G: **Cancer of unknown primary site: 20 questions to be answered.** *Annals of oncology : official journal of the European Society for Medical Oncology* 2010, **21 Suppl 7**:vii303–vii307.

[60] Cornish-Bowden A: **Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.** *Nucleic acids research* 1985, **13**:3021–3030.

[61] Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic acids research* 2006, **34**:D95–D97.

[62] Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison.** *Nucleic acids research* 2004, **32**:W249–W252.

[63] Farré D, Roset R, Huerta M, Adsuara JE, Roselló L, Albà MM, Messeguer X: **Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN.** *Nucleic acids research* 2003, **31**:3651–3653.

[64] Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic acids research* 1990, **18**:6097–6100.

[65] Stormo GD, Zhao Y: **Determining the specificity of protein-DNA interactions.** *Nature reviews. Genetics* 2010, **11**:751–760.

[66] Meckbach C, Tacke R, Hua X, Waack S, Wingender E, Gültas M: **PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information.** *BMC bioinformatics* 2015, **16**:400.

[67] Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW: **oPOSSUM: integrated tools for analysis of regulatory motif over-representation.** *Nucleic acids research* 2007, **35**:W245–W252.

[68] Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW: **oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets.** *G3 (Bethesda, Md.)* 2012, **2**:987–1002.

[69] Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse.** *Nature genetics* 2007, **39**:730–732.

[70] Hung JH, Weng Z: **Motif Finding.** *Cold Spring Harbor protocols* 2017, **2017**.

[71] Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML: **Diversity and complexity in DNA recognition by transcription factors.** *Science (New York, N.Y.)* 2009, **324**:1720–1723.

[72] Plotkin JB, Kudla G: **Synonymous but not the same: the causes and consequences of codon bias.** *Nature reviews. Genetics* 2011, **12**:32–42.

[73] Nekrutenko A, Li WH: **Assessment of compositional heterogeneity within and between eukaryotic genomes.** *Genome research* 2000, **10**:1986–1995.

[74] Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data.** *Bioinformatics (Oxford, England)* 2011, **27**:1653–1659.

[75] Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, Woodard J, Mariani L, Kock KH, Inukai S, Siggers T, Shokri L, Gordân R, Sahni N, Cotsapas C, Hao T, Yi S, Kellis M, Daly MJ, Vidal M, Hill DE, Bulyk ML: **Survey of variation in human transcription factors reveals prevalent DNA binding changes.** *Science (New York, N.Y.)* 2016, **351**:1450–1454.

[76] Jiang M, Anderson J, Gillespie J, Mayne M: **uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts.** *BMC bioinformatics* 2008, **9**:192.

[77] Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S, Consortium D, Bussemaker HJ, Morris QD, Bulyk ML, Stolovitzky G, Hughes TR: **Evaluation of methods for modeling transcription factor sequence specificity.** *Nature biotechnology* 2013, **31**:126–134.

[78] Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS: **Deconvolving the recognition of DNA shape from sequence.** *Cell* 2015, **161**:307–318.

[79] Struhl K, Segal E: **Determinants of nucleosome positioning.** *Nature structural & molecular biology* 2013, **20**:267–273.

[80] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription**

factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* 2010, **38**:576–589.

[81] Worsley Hunt R, Mathelier A, Del Peso L, Wasserman WW: **Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment.** *BMC genomics* 2014, **15**:472.

[82] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science (New York, N.Y.)* 2008, **320**:1344–1349.

[83] Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP: **Using oligonucleotide probe arrays to access genetic diversity.** *BioTechniques* 1995, **19**:442–447.

[84] Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science (New York, N.Y.)* 1995, **270**:467–470.

[85] Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP: **Accessing genetic information with high-density DNA arrays.** *Science (New York, N.Y.)* 1996, **274**:610–614.

[86] Kogenaru S, Qing Y, Guo Y, Wang N: **RNA-seq and microarray complement each other in transcriptome profiling.** *BMC genomics* 2012, **13**:629.

[87] Hou Z, Jiang P, Swanson SA, Elwell AL, Nguyen BKS, Bolin JM, Stewart R, Thomson JA: **A cost-effective RNA sequencing protocol for large-scale gene expression studies.** *Scientific reports* 2015, **5**:9570.

[88] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature methods* 2008, **5**:621–628.

[89] Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome biology* 2010, **11**:R106.

[90] Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome biology* 2010, **11**:R25.

[91] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**:25–29.

[92] Dijkstra EW: **A note on two problems in connexion with graphs**. *Numerische Mathematik* 1959, **1**:269–271.

[93] Tatarinova T: *Biological networks and pathway analysis*. New York, NY: Humana Press 2017.

[94] Behnamian A, Millard K, Banks SN, White L, Richardson M, Pasher J: **A Systematic Approach for Variable Selection With Random Forests: Achieving Stable Variable Importance Values**. *IEEE Geoscience and Remote Sensing Letters* 2017, **14**(11):1988–1992.

[95] Kursa MB, Jankowski A, Rudnicki WR: **Boruta – A System for Feature Selection**. *Fundamenta Informaticae* 2010, **101**(4):271–285.

[96] Lever J, Krzywinski M, Altman N: **Principal component analysis**. *Nature Methods* 2017, **14**(7):641–642.

[97] Lloyd S: **Least squares quantization in PCM**. *IEEE Transactions on Information Theory* 1982, **28**(2):129–137.

[98] Alkhalifah S, Aloraini A: **Graphical Model and Clustering-Regression based Methods for Causal Interactions: Breast Cancer Case Study**. *International Journal of Artificial Intelligence & Applications* 2020, **11**(3):11–27.

[99] Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. *Journal of Computational and Applied Mathematics* 1987, **20**:53–65.

[100] Kaufman L: *Finding groups in data : an introduction to cluster analysis*. New York: Wiley 1990.

[101] Shahapure KR, Nicholas C: **Cluster Quality Analysis Using Silhouette Score**. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE 2020.

[102] Shannon CE: **A Mathematical Theory of Communication**. *Bell System Technical Journal* 1948, **27**(3):379–423.

[103] Pathria RK: *Statistical mechanics*. Boston: Academic Press 2011.

[104] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic acids research* 2010, **38**:1767–1771.

[105] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics (Oxford, England)* 2013, **29**:15–21.

[106] Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC bioinformatics* 2011, **12**:323.

[107] Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15**:550.

[108] Menck K, Heinrichs S, Wlochowitz D, Sitte M, Noeding H, Janshoff A, Treiber H, Ruhwedel T, Schatlo B, von der Brelie C, Wiemann S, Pukrop T, Beißbarth T, Binder C, Bleckmann A: **WNT11/ROR2 signaling is associated with tumor invasion and poor survival in breast cancer.** *Journal of experimental & clinical cancer research : CR* 2021, **40**:395.

[109] Zhu A, Ibrahim JG, Love MI: **Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences.** *Bioinformatics (Oxford, England)* 2019, **35**:2084–2092.

[110] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project.** *Nucleic acids research* 2002, **30**:38–41.

[111] Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nature protocols* 2009, **4**:1184–1191.

[112] Manning CD, Raghavan P, Schütze H: *Introduction to Information Retrieval.* Cambridge University Pr. 2008, [`https://www.ebook.de/de/product/7455223/christopher_d_manning_prabhakar_raghavan_hinrich_schuetze_introduction_to_information_retrieval.html`].

[113] Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *Omics : a journal of integrative biology* 2012, **16**:284–287.

[114] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995, **57**:289–300.

[115] Menck K, Wlochowitz D, Wachter A, Conradi LC, Wolff A, Peeck M, Scheel AH, Schlüter K, Korf U, Wiemann S, Schildhaus HU, Wingender E, Pukrop T, Homayounfar K, Beißbarth T, Bleckmann A: **High-throughput profiling of colorectal cancer liver metastases reveals intra- and interpatient heterogeneity in the EGFR- and WNT-pathway associated with clinical outcome**. [In preparation].

[116] Khan A, Riudavets Puig R, Boddie P, Mathelier A: **BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences.** *Bioinformatics (Oxford, England)* 2021, **37**:1607–1609.

[117] Hounkpe BW, Chenou F, de Lima F, De Paula EV: **HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets.** *Nucleic acids research* 2021, **49**:D947–D955.

[118] Chen LF, Greene WC: **Shaping the nuclear action of NF-kappaB.** *Nature reviews. Molecular cell biology* 2004, **5**:392–401.

[119] Kozlyuk N, Monteith AJ, Garcia V, Damo SM, Skaar EP, Chazin WJ: **S100 Proteins in the Innate Immune Response to Pathogens.** *Methods in molecular biology (Clifton, N.J.)* 2019, **1929**:275–290.

[120] Unk I, Hajdú I, Fátyol K, Hurwitz J, Yoon JH, Prakash L, Prakash S, Haracska L: **Human HLTF functions as a ubiquitin ligase for proliferating cell nuclear antigen polyubiquitination.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:3768–3773.

[121] Blastyák A, Hajdú I, Unk I, Haracska L: **Role of double-stranded DNA translocase activity of human HLTF in replication of damaged DNA.** *Molecular and cellular biology* 2010, **30**:684–693.

[122] Liu L, Liu H, Zhou Y, He J, Liu Q, Wang J, Zeng M, Yuan D, Tan F, Zhou Y, Pei H, Zhu H: **HLTF suppresses the migration and invasion of colorectal cancer cells via TGF-β/SMAD signaling in vitro.** *International journal of oncology* 2018, **53**:2780–2788.

[123] Xiao X, Liu Z, Wang R, Wang J, Zhang S, Cai X, Wu K, Bergan RC, Xu L, Fan D: **Genistein suppresses FLT4 and inhibits human colorectal cancer metastasis.** *Oncotarget* 2015, **6**:3225–3239.

[124] Reis M, Czupalla CJ, Ziegler N, Devraj K, Zinke J, Seidel S, Heck R, Thom S, Macas J, Bockamp E, Fruttiger M, Taketo MM, Dimmeler S, Plate KH, Liebner S: **Endothelial Wnt/β-catenin signaling inhibits glioma angiogenesis and normalizes tumor blood vessels by inducing PDGF-B expression.** *The Journal of experimental medicine* 2012, **209**:1611–1627.

[125] Luo D, Ge W: **MeCP2 Promotes Colorectal Cancer Metastasis by Modulating ZEB1 Transcription.** *Cancers* 2020, **12**.

[126] Voorneveld PW, Kodach LL, Jacobs RJ, Liv N, Zonnevylle AC, Hoogenboom JP, Biemond I, Verspaget HW, Hommes DW, de Rooij K, van Noesel CJM, Morreau H, van Wezel T, Offerhaus GJA, van den Brink GR, Peppelenbosch MP, Ten Dijke P, Hardwick JCH: **Loss of SMAD4 alters BMP signaling to promote colorectal cancer cell metastasis via activation of Rho and ROCK.** *Gastroenterology* 2014, **147**:196–208.e13.

[127] Lorente-Trigos A, Varnat F, Melotti A, Ruiz i Altaba A: **BMP signaling promotes the growth of primary human colon carcinomas in vivo.** *Journal of molecular cell biology* 2010, **2**:318–332.

[128] Zhang MQ: **Identification of human gene core promoters in silico.** *Genome research* 1998, **8**:319–326.

[129] Koop BF: **Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution.** *Trends in genetics : TIG* 1995, **11**:367–371.

[130] Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nature genetics* 2000, **26**:225–228.

[131] Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome.** *Genome research* 1997, **7**:959–966.

[132] Ho Sui SJ: **In silico approaches to investigating mechanisms of gene regulation** 2008.

[133] Vandenbon A, Kumagai Y, Akira S, Standley DM: **A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription.** *BMC genomics* 2012, **13 Suppl 7**:S11.

[134] Murakami K, Imanishi T, Gojobori T, Nakai K: **Two different classes of co-occurring motif pairs found by a novel visualization method in human promoter regions.** *BMC genomics* 2008, **9**:112.

[135] Ha N, Polychronidou M, Lohmann I: **COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets.** *PloS one* 2012, **7**:e52055.

[136] Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G: **Enhancers: five essential questions.** *Nature reviews. Genetics* 2013, **14**:288–295.

[137] Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

[138] Andersson R, , Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Forrest ARR, Carninci P, Rehli M, Sandelin A: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**(7493):455–461.

[139] Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, Ishikawa-Kato S, Kaida K, Kaiho A, Kanamori-Katayama M, Kawashima T, Kojima M, Kubosaki A, Manabe RI, Murata M, Nagao-Sato S, Nakazato K, Ninomiya N, Nishiyori-Sueki H, Noma S, Saijyo E, Saka A, Sakai M, Simon C, Suzuki N, Tagami M, Watanabe S, Yoshida S, Arner P, Axton RA, Babina M, Baillie JK, Barnett TC, Beckhouse AG, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Carlisle AJ, Clevers HC, Davis CA, Detmar M, Dohi T, Edge ASB, Edinger M, Ehrlund A, Ekwall K, Endoh M, Enomoto H, Eslami A, Fagiolini M, Fairbairn L, Farach-Carson MC, Faulkner GJ, Ferrai C, Fisher ME, Forrester LM, Fujita R, Furusawa JI, Geijtenbeek TB, Gingeras T, Goldowitz D, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Hasegawa Y, Herlyn M, Heutink P, Hitchens KJ, Hume DA, Ikawa T, Ishizu Y, Kai C, Kawamoto H, Kawamura YI, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klein S, Klinken SP, Knox AJ, Kojima S, Koseki H, Koyasu S, Lee W, Lennartsson A, Mackay-Sim A, Mejhert N, Mizuno Y, Morikawa H, Morimoto M, Moro K, Morris KJ, Motohashi H, Mummery CL, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nozaki T, Ogishima S, Ohkura N, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Passier R, Patrikakis M, Pombo A, Pradhan-Bhatt S, Qin XY, Rehli M, Rizzu P, Roy S, Sajantila A, Sakaguchi S, Sato H, Satoh H, Savvi S, Saxena A, Schmidl C, Schneider C, Schulze-Tanzil GG, Schwegmann A, Sheng G, Shin JW, Sugiyama D, Sugiyama T, Summers KM, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tomoiu A, Toyoda H, van de Wetering M, van den Berg LM, Verardo R, Vijayan D, Wells CA, Winteringham LN, Wolvetang E, Yamaguchi Y, Yamamoto M, Yanagi-Mizuochi C, Yoneda M, Yonekura Y, Zhang PG, Zucchelli S, Abugessaisa I, Arner E, Harshbarger J, Kondo A, Lassmann T, Lizio M, Sahin S, Sengstag T, Severin J, Shimoji H, Suzuki M, Suzuki H, Kawai J, Kondo N, Itoh M, Daub CO, Kasukawa T, Kawaji H, Carninci P, Forrest ARR, Hayashizaki Y: **FANTOM5 CAGE profiles of human and mouse samples.** *Scientific data* 2017, **4**:170112.

[140] Hooghe B, Broos S, van Roy F, De Bleser P: **A flexible integrative approach based on random forest improves prediction of transcription factor binding sites.** *Nucleic acids research* 2012, **40**:e106.

[141] Siebert M, Söding J: **Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences.** *Nucleic acids research* 2016, **44**:6055–6069.

[142] Zhao Y, Ruan S, Pandey M, Stormo GD: **Improved models for transcription factor binding site identification using nonindependent interactions.** *Genetics* 2012, **191**:781–790.

[143] Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordân R, Rohs R: **TFBSshape: a motif database for DNA shape features of transcription factor binding sites.** *Nucleic acids research* 2014, **42**:D148–D155.

[144] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals.** *Nature* 2005, **434**:338–345.

[145] Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR: **Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences.** *Cell* 2008, **133**:1266–1276.

[146] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic acids research* 2010, **38**:D105–D110.

[147] Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152**:327–339.

[148] Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ: **HOCOMOCO: a comprehensive collection of human transcription factor binding sites models.** *Nucleic acids research* 2013, **41**:D195–D202.

[149] Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E: **SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates.** *Nucleic acids research* 2013, **41**:D214–D220.

[150] Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, Essack M, Gao X, Bajic VB: **A novel method for improved accuracy of transcription factor binding site prediction.** *Nucleic acids research* 2018, **46**:e72.

[151] Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, Gould JA, Forster SC, Hertzog PJ: **CiiiDER: A tool for predicting and analysing transcription factor binding sites.** *PloS one* 2019, **14**:e0215495.

[152] Oh YM, Kim JK, Choi S, Yoo JY: **Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices**. *Nucleic Acids Research* 2011, **40**(5):e38–e38.

[153] Zabet NR, Adryan B: **Estimating binding properties of transcription factors from genome-wide binding profiles.** *Nucleic acids research* 2015, **43**:84–94.

[154] Garcia-Alcalde F, Blanco A, Shepherd AJ: **An intuitionistic approach to scoring DNA sequences against transcription factor binding site motifs.** *BMC bioinformatics* 2010, **11**:551.

[155] Schonlau M, Zou RY: **The random forest algorithm for statistical learning**. *The Stata Journal: Promoting communications on statistics and Stata* 2020, **20**:3–29.

[156] Orlenko A, Moore JH: **A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions**. *BioData Mining* 2021, **14**.

[157] Palczewska A, Palczewski J, Robinson RM, Neagu D: **Interpreting Random Forest Classification Models Using a Feature Contribution Method**. In *Integration of Reusable Systems*, Springer International Publishing 2014:193–218.

[158] Yadav R, Srivastava P: **Clustering, Pathway Enrichment, and Protein-Protein Interaction Analysis of Gene Expression in Neurodevelopmental Disorders.** *Advances in pharmacological sciences* 2018, **2018**:3632159.

[159] Allocco DJ, Kohane IS, Butte AJ: **Quantifying the relationship between co-expression, co-regulation and gene function.** *BMC bioinformatics* 2004, **5**:18.

[160] Zhang C, Lee S, Mardinoglu A, Hua Q: **Investigating the Combinatory Effects of Biological Networks on Gene Co-expression.** *Frontiers in physiology* 2016, **7**:160.

[161] Kumar SU, Kumar DT, Bithia R, Sankar S, Magesh R, Sidenna M, Doss CGP, Zayed H: **Analysis of Differentially Expressed Genes and Molecular Pathways in Familial Hypercholesterolemia Involved in Atherosclerosis: A Systematic and Bioinformatics Approach**. *Frontiers in Genetics* 2020, **11**.

[162] Saelens W, Cannoodt R, Saeys Y: **A comprehensive evaluation of module detection methods for gene expression data.** *Nature communications* 2018, **9**:1090.

[163] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP: **Gene co-expression analysis for functional classification and gene-disease predictions.** *Briefings in bioinformatics* 2018, **19**:575–592.

[164] Yin W, Mendoza L, Monzon-Sandoval J, Urrutia AO, Gutierrez H: **Emergence of co-expression in gene regulatory networks.** *PloS one* 2021, **16**:e0247671.

[165] Meckbach C, Wingender E, Gültas M: **Removing Background Co-occurrences of Transcription Factor Binding Sites Greatly Improves the Prediction of Specific Transcription Factor Cooperations**. *Frontiers in Genetics* 2018, **9**.

[166] Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics (Oxford, England)* 2001, **17**:763–774.

[167] Mitsuhiro M, Yadohisa H: **Reduced $$k$$ k -means clustering with MCA in a low-dimensional space**. *Computational Statistics* 2014, **30**(2):463–475.

[168] Ding C, He X: **K-means clustering via principal component analysis**. In *Twenty-first international conference on Machine learning - ICML '04*, ACM Press 2004.

[169] Zhang N, Leatham K, Xiong J, Zhong J: **PCA-K-Means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals**. In *2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*, IEEE 2018.

[170] Voss-Fels K, Frisch M, Qian L, Kontowski S, Friedt W, Gottwald S, Snowdon RJ: **Subgenomic Diversity Patterns Caused by Directional Selection in Bread Wheat Gene Pools**. *The Plant Genome* 2015, **8**(2).

[171] Jantzen SG, Sutherland BJ, Minkley DR, Koop BF: **GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets**. *BMC Research Notes* 2011, **4**.

[172] Supek F, Bošnjak M, Škunca N, Šmuc T: **REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms**. *PLoS ONE* 2011, **6**(7):e21800.

[173] Reijnders MJMF, Waterhouse RM: **Summary Visualizations of Gene Ontology Terms With GO-Figure!** *Frontiers in Bioinformatics* 2021, **1**.

[174] Sun D, Liu Y, Zhang XS, Wu LY: **CEA: Combination-based gene set functional enrichment analysis**. *Scientific Reports* 2018, **8**.

A. Appendix A: Computational Pipeline

A.1. Analysis of the NF-κB Pathway-related Dataset

Table A1.: ***Target-to-background*** **ratios of corresponding relevant PWMs in the analysis of the NF-κB pathway-related dataset using** ***Target (NF-κB Microarray)*** **and the GC content-matched** ***Background (oPOSSUM).*** ***optBOR.conf***: optimized MSS cutoffs for a list of 14 confirmed PWMs by Boruta; ***TSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the target set; ***BSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the background set; ***TaF***: average site frequency that corresponds to the target set; ***BaF***: average site frequency that corresponds to the background set; ***target-to-backround ratio*** **(T/B ratio)**: target-to-background ratio: average site frequency in the target set divided by average site frequency in the background set.

PWM ID	***optBOR.conf***	***TSP%***	***BSP%***	***TaF***	***BaF***	***T/B ratio***
V$MECP2_02	0.9750	5.9028	1.7361	0.0729	0.0174	4.2000
V$GZF1_01	0.8025	9.0278	3.4722	0.0972	0.0417	2.3333
V$RELA_Q6	0.8850	71.8750	68.0556	1.7674	1.2361	1.4298
V$BLIMP1_Q4	0.8175	99.6528	98.2639	7.0590	5.9826	1.1799
V$BCL6_Q3_01	0.8250	100.0000	100.0000	15.8299	14.5625	1.0870
V$HELIOSA_02	0.8200	100.0000	100.0000	9.3021	8.9826	1.0356
V$TATA_01	0.7825	95.1389	99.3056	11.1979	11.0000	1.0180
V$RUSH1A_02	0.8900	100.0000	100.0000	24.3264	24.9549	0.9748
V$DBP_Q6	0.8475	100.0000	100.0000	15.6840	16.1076	0.9737
V$DELTAEF1_01	0.7900	100.0000	100.0000	10.2431	11.2743	0.9085
V$TBX5_01	0.8000	99.6528	100.0000	6.1250	6.7708	0.9046
V$PIT1_Q6_01	0.9775	24.6528	21.5278	0.2778	0.3785	0.7339
V$LUN1_01	0.7575	7.9861	16.6667	0.0938	0.1840	0.5094
V$ZFP206_01	0.9300	2.7778	9.3750	0.0347	0.1215	0.2857

Table A2.: ***Target-to-background*** **ratios of corresponding relevant PWMs in the analysis of the NF-κB pathway-related dataset using** ***Target (NF-κB Microarray)*** **and the housekeeping gene-related** ***Background (HKG HUVEC)***. ***optBOR.conf***: optimized MSS cutoffs for a list of 36 confirmed PWMs by Boruta; ***TSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the target set; ***BSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the background set; ***TaF***: average site frequency that corresponds to the target set; ***BaF***: average site frequency that corresponds to the background set; ***target-to-backround ratio*** **(T/B ratio)**: target-to-background ratio: average site frequency in the target set divided by average site frequency in the background set.

PWM ID	***optBOR.conf***	***TSP%***	***BSP%***	***TaF***	***BaF***	***T/B ratio***
V$RELA_Q6	0.99	6.9444	0.6944	0.0694	0.0069	10.0000
V$POU2F1_Q6	0.8475	82.9861	72.9167	2.4097	1.6389	1.4703
V$DBP_Q6	0.93	96.8750	95.4861	4.2396	3.3229	1.2759
V$STAT1_Q6	0.7575	100.0000	99.6528	17.8160	14.7326	1.2093
V$NFAT5_Q5_02	0.75	99.6528	99.6528	7.7847	6.4479	1.2073
V$ZSCAN4_04	0.755	99.6528	99.3056	7.5174	6.2292	1.2068
V$BLIMP1_Q4	0.7975	100.0000	97.9167	8.9896	7.5729	1.1871
V$BCL6_Q3_01	0.795	100.0000	100.0000	25.0521	21.4861	1.1660
V$LEF1_Q5_01	0.7625	100.0000	100.0000	36.7188	32.0208	1.1467
V$RBPJK_01	0.75	100.0000	100.0000	7.7569	6.7778	1.1445
V$CEBPA_Q6	0.755	100.0000	100.0000	73.9479	65.5486	1.1281
V$AP1_Q6_02	0.795	100.0000	100.0000	17.9514	15.9583	1.1249
V$SOX10_Q3	0.7775	100.0000	100.0000	64.9688	57.9271	1.1216
V$PIT1_Q6_01	0.755	100.0000	100.0000	38.0382	34.1424	1.1141
V$GEN_INI_B	0.75	100.0000	100.0000	79.4271	73.7708	1.0767
V$NKX25_Q6	0.755	100.0000	100.0000	26.8194	25.3403	1.0584
V$IK_Q5_01	0.7875	100.0000	100.0000	26.2361	24.8993	1.0537
V$MZF1_Q5	0.75	100.0000	100.0000	51.0903	51.9931	0.9826
V$COE1_Q6	0.75	99.6528	100.0000	24.9965	25.9062	0.9649
V$CP2_Q6	0.7525	100.0000	100.0000	49.6875	52.2153	0.9516
V$MAZR_01	0.7575	97.5694	99.6528	20.3646	21.4583	0.9490
V$MAZ_Q6_01	0.7525	99.3056	100.0000	23.1632	25.0590	0.9243
V$CPBP_Q6	0.9875	100.0000	100.0000	30.1944	33.3299	0.9059
V$SP1_Q6_01	0.75	99.6528	100.0000	29.1146	32.2812	0.9019
V$BEN_01	0.755	100.0000	100.0000	53.0069	59.2882	0.8941
V$AHR_Q6	0.7575	100.0000	100.0000	33.0312	37.6910	0.8764
V$EGR1_Q6	0.8025	90.9722	98.9583	8.6562	10.0729	0.8594
V$CHCH_01	0.9875	85.7639	98.2639	8.7917	10.6875	0.8226
V$ZFP161_04	0.755	71.8750	90.2778	4.9444	6.2153	0.7955
V$MECP2_02	0.83	78.1250	97.5694	6.0069	7.6215	0.7882
V$HIF1A_Q5	0.79	95.1389	97.9167	5.0069	6.4410	0.7774
V$RNF96_01	0.82	85.7639	99.3056	6.6042	8.7535	0.7545
V$E2F_Q6_01	0.75	90.6250	98.2639	7.2049	10.1181	0.7121
V$SP100_04	0.75	96.8750	100.0000	16.2882	23.6007	0.6902
V$CREB1_Q6	0.9275	46.8750	62.8472	0.6250	1.1076	0.5643
V$ZFP206_01	0.925	3.1250	15.2778	0.0556	0.3021	0.1839

A.2. Analysis of the WNT Pathway-related CRC Dataset

Table A3.: ***Target-to-background*** **ratios of corresponding relevant PWMs in the analysis of the WNT pathway-related CRC dataset using** ***Target (1638N-T1)*** **and** ***Background (CMT-93)***. ***optBOR.conf***: optimized MSS cutoffs for a list of 31 confirmed PWMs by Boruta; ***TSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the target set; ***BSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the background set; ***TaF***: average site frequency that corresponds to the target set; ***BaF***: average site frequency that corresponds to the background set; ***target-to-backround ratio*** **(T/B ratio)**: target-to-background ratio: average site frequency in the target set divided by average site frequency in the background set.

PWM ID	***optBOR.conf***	***TSP%***	***BSP%***	***TaF***	***BaF***	***T/B ratio***
V$MEF2_03	0.7825	63.6157	53.8578	1.1467	0.8722	1.3148
V$TATA_01	0.7825	98.6384	96.5961	12.3873	9.7526	1.2701
V$HOXB13_01	0.7675	95.6127	93.2678	5.8805	4.8366	1.2158
V$HNF1A_Q4	0.75	98.1089	97.8064	10.7995	8.9607	1.2052
V$HBP1_03	0.75	97.4281	95.9909	6.441	5.3986	1.1931
V$POU2F1_Q6	0.7625	99.5461	99.6218	13.1558	11.1044	1.1847
V$MRF2_01	0.7775	99.1679	98.8654	9.969	8.469	1.1771
V$CDX2_Q5_02	0.75	100	100	45.0076	38.5386	1.1679
V$CDX2_01	0.755	98.4115	98.1089	8.1982	7.0333	1.1656
V$IPF1_Q5	0.765	99.9244	99.9244	22.2156	19.2655	1.1531
V$CDPCR1_01	0.75	99.2436	99.3192	9.677	8.5325	1.1341
V$RUSH1A_02	0.89	100	99.9244	28.0113	25.8101	1.0853
V$YY1_Q6_03	0.79	100	100	21.91	20.4849	1.0696
V$SOX10_Q3	0.77	100	100	81.6301	78.0393	1.046
V$TTF1_Q5_01	0.76	100	100	146.4955	149.7655	0.9782
V$MTF1_Q5	0.755	100	99.9244	20.7973	21.733	0.9569
V$NKX25_Q6	0.82	100	99.8487	7.9909	8.5023	0.9399
V$AHR_Q6	0.75	100	100	30.3154	32.3669	0.9366
V$ERALPHA_Q6_01	0.77	100	100	25.3366	27.1172	0.9343
V$MZF1_Q5	0.7525	100	100	45.7557	49.1422	0.9311
V$P53_Q3	0.755	100	100	45.1884	48.5454	0.9308
V$HIF1A_Q5	0.77	100	100	11.5681	12.4667	0.9279
V$IK_Q5_01	0.835	100	100	13.9592	15.0643	0.9266
V$CP2_Q6	0.7625	100	100	41.0719	44.969	0.9133
V$CHCH_01	0.975	100	100	36.9092	40.6316	0.9084
V$SREBP_Q6	0.88	100	99.9244	8.4251	9.2791	0.908
V$CPBP_Q6	0.98	99.9244	100	29.2678	32.736	0.8941
V$SP1_Q6_01	0.855	86.4599	92.8896	5.826	6.5242	0.893
V$MECP2_02	0.75	99.3949	99.8487	10.5968	11.9221	0.8888
V$INSM1_01	0.8175	84.0393	88.6536	2.8926	3.3064	0.8749
V$ZFP206_01	0.875	10.1362	10.7413	0.1997	0.261	0.7652

Table A4.: Top 20 significant Gene Ontology (GO) Biological Process (BP) terms obtained for the gene cluster *Cluster 1* after overrepresentation analysis (ORA) in the analysis of the WNT pathway-related CRC dataset.

GO Term ID	Description	p.adjust	Gene_Hits
GO:0072001	renal system development	5.8927e-06	Adamts16,Bmp7,Cyp4a12a,Fbn1,Foxf1,Gata3,Gdf11,Gpc3,Gsta3,Hoxa11,Irx3,Lgr5,Mmp17, Mmp9,Nog,Notch3,Plxnd1,Robo1,Slc12a1,Slit2,Spp1,Stra6,Vcan,Wnt1,Wnt5a
GO:0043542	endothelial cell migration	6.247e-05	Adgra2,Ccn3,Cdh13,Ceacam1,Flt4,Gata3,Hdac9,Lgals12,Plxnd1,Prcp,Prkd1,Ptgs2,Ptp4a3,Ptprm,Robo1,Slit2,Stc1,Wnt5a
GO:0090130	tissue migration	0.0004	Adgra2,Ccn3,Cdh13,Ceacam1,Flt4,Foxf1,Gata3,Hdac9,Lgals12,Mmp9,Plxnd1,Prcp,Prkd1,Ptgs2,Ptp4a3,Ptprm,Robo1,Slit2,Stc1,Wnt5a
GO:0090287	regulation of cellular response to growth factor stimulus	0.0005	Adgra2,Bambi,Cav1,Cdkn2b,Fbn1,Fbn2,Gata3,Gpc3,Msx2,Nog,Onecut1,Ptp4a3,Rgma,Robo1,Slit2,Vegfd,Vsir,Wnt1,Wnt5a
GO:0060562	epithelial tube morphogenesis	0.0005	Adamts16,Bmp7,Flt1,Foxf1,Fzd2,Gata3,Gpc3,Hoxa11,Irx3,Lgr5,Msx2,Nog,Plxnd1,Rgma,Robo1,Shank3,Slit2,Spint1,St14,Tbx2,Wnt1,Wnt5a
GO:0003205	cardiac chamber development	0.0005	Bmp7,Ccm2l,Dsp,Foxf1,Fzd2,Gata3,Msx2,Nog,Plxnd1,Robo1,Slit2,Stra6,Tbx2,Wnt5a,Zfpm2
GO:0045598	regulation of fat cell differentiation	0.0005	Bmp7,Enpp1,Fndc5,Gata3,Lgals12,Msx2,Ptgs2,Tph1,Trib3,Wnt1,Wnt10b,Wnt5a,Zfpm2
GO:0003206	cardiac chamber morphogenesis	0.0005	Bmp7,Ccm2l,Dsp,Foxf1,Fzd2,Gata3,Msx2,Nog,Robo1,Slit2,Tbx2,Wnt5a,Zfpm2
GO:0046849	bone remodeling	0.0005	Adrb2,Car2,Ceacam1,Enpp1,Htr1b,Inpp4b,Mitf,Ptger4,Spp1,Syt7,Tph1
GO:0048705	skeletal system morphogenesis	0.0005	Anxa6,Barx2,Bmp7,Col11a2,Cyp26b1,Fbn2,Gas1,Hoxa11,Hoxb3,Hoxb6,Hoxc9,Msx2,Nog,Rflna,Scx,Stc1,Wnt10b
GO:0060840	artery development	0.0005	Apob,Efemp2,Fkbp10,Foxf1,Nog,Notch3,Plxnd1,Ptger4,Robo1,Stra6,Tbx2
GO:0045444	fat cell differentiation	0.0006	Adrb2,Bbs1,Bmp7,Dysf,Enpp1,Fndc5,Gata3,Lgals12,Msx2,Ptgs2,Scd1,Tph1,Trib3,Wnt1,Wnt10b,Wnt5a,Zfpm2
GO:0051216	cartilage development	0.0006	Anxa6,Barx2,Bbs1,Bmp7,Ccn3,Col11a2,Hoxa11,Hoxb3,Mboat2,Msx2,Nog,Rflna,Scx,Stc1,Wnt5a
GO:0040013	negative regulation of locomotion	0.0012	Ccn3,Dpysl3,Gata3,Il1rn,Il33,Mitf,Nexmif,Nog,Ptger4,Ptprm,Robo1,Sema6c,Sema6d,Slit2,Slurp1,Stc1,Tmeff2,Trp53inp1,Wnt5a
GO:0007411	axon guidance	0.0013	Alcam,Bmp7,Gas1,Gata3,Lama3,Nog,Notch3,Plxna4,Plxnd1,Ptprm,Robo1,Rtn4rl1,Sema6c,Sema6d,Slit2,Wnt5a
GO:0090092	regulation of transmembrane receptor protein serine-threonine kinase signaling pathway	0.0013	Bambi,Bmp7,Cav1,Ccn3,Cdkn2b,Fbn1,Fbn2,Gdf11,Gpc3, Msx2,Nog,Onecut1,Rgma,Vsir,Wnt1,Wnt5a
GO:0097485	neuron projection guidance	0.0013	Alcam,Bmp7,Gas1,Gata3,Lama3,Nog,Notch3,Plxna4,Plxnd1,Ptprm,Robo1,Rtn4rl1,Sema6c,Sema6d,Slit2,Wnt5a
GO:0050919	negative chemotaxis	0.0013	Plxna4,Robo1,Rtn4rl1,Sema6c,Sema6d,Slit2,Wnt5a
GO:0050680	negative regulation of epithelial cell proliferation	0.0014	Cav1,Cdkn2b,Ceacam1,Flt1,Gas1,Gata3,Gpc3,Ptprm,Robo1,Slurp1,Wnt10b,Wnt5a
GO:0050673	epithelial cell proliferation	0.0014	Cav1,Ccl2,Ccn3,Cdh13,Cdkn2b,Ceacam1,Dysf,Flt1,Flt4,Gas1,Gata3,Gpc3,Lgr5,Nog,Prkd1,Ptprm,Robo1,Sgpp2,Slurp1,Vegfd,Wnt10b,Wnt5a

Table A5.: Master regulator (MR) candidates obtained for the gene cluster *Cluster 1* after MR search in the analysis of the WNT pathway-related CRC dataset. MRs are alphabhtically sorted by HGNC gene name.

Gene Name	Master molecule name	Reached from set	FDR
Adrb2	ADRB2R(m)	19	0
Birc2	cIAP-1(m)	24	0.014
Bmp7	BMP7(m)	13	0
Cacna1c	CACNL1A1(m)	2	0
Carm1	carm1(m)	22	0.015
Cav1	caveolin-1(m)	29	0
Ccl2	MCP-1(m)	10	0
Cdh2	N-cadherin(m)	11	0.001
Cdkn2b	p15INK4b(m)	13	0.001
Ceacam1	BGPI(m)	24	0.002
Cybb	gp91phox(m)	9	0
Dlg4	PSD-95(m)	10	0.002
Egfr	ErbB1(m)	34	0.025
Erbb3	ErbB3(m)	26	0.029
Fcer1g	FCRG(m)	25	0.037
Fgf2	FGF-2(m)	24	0.007
Flt4	VEGFR-3(m)	16	0
Furin	PACE(m)	24	0.006
Gata3	GATA-3(m)	3	0
Gda	guanine deaminase(m)	2	0.001
Grk2	betaARK-1(m)	29	0.029
Grk5	GPRK5(m)	19	0.008
Hipk2	hipk2(m)p	22	0.011
Ifng	IFNgamma(m)	24	0.018
Igfbp2	IGFBP-2(m)	12	0
Il10	IL-10(m)	28	0
Il1rn	IL-1RA(m)	10	0
Il33	IL-33(m)	10	0
Irf8	IRF-8(m)	9	0
Lamb1	Laminin-beta1(m)	5	0
Lgals3bp	Lgals3bp(m)	2	0
Map2k3	MKK3(m)p	27	0.018
Map2k6	MKK6(m)p	27	0.018
Map3k14	NIK(m)	25	0
Map3k7	TAK1(m)	29	0.042
Mapk12	p38gamma(m)	29	0.007
Mapk13	p38delta(m)	30	0.005

Gene name	Master molecule name	Reached from set	FDR
Mapk14	p38alpha(m)	32	0.03
Mitf	MITF(m)	4	0.001
Mmp9	gelatinase B(m)	8	0
Ncstn	gamma-secretase(m)	20	0.005
Npr1	atrial natriuretic peptide receptor A(m)	6	0
Padi4	Padi4(m)	2	0.001
Pik3cg	p110gamma(m)	18	0.004
Pik3r5	p101(m)	2	0
Pik3r6	Pik3r6(m)	2	0
Ppm1b	PP2Cbeta1(m)	27	0.038
Prkaca	PKACA(m)	33	0.041
Prkcd	PKCdelta(m)pY311	26	0.007
Prkcq	PKCtheta(m)	26	0.024
Prkd1	PKD1(m)	25	0
Psen1	gamma-secretase(m)	20	0.005
Psenen	gamma-secretase(m)	20	0.005
Ptger4	EP4(m)	17	0
Ptgs2	COX2(m)	7	0
Ptprb	RPTP-beta(m)	20	0.003
Ptprm	RPTPmu(m)	3	0.001
Rac2	Rac2(m)	25	0.012
Rhoa	RhoA(m)	25	0.033
Rhoc	RhoC(m)	25	0.01
Rhog	Rhog(m)	25	0.01
Rps6ka5	MSK1(m)	34	0.027
Scd1	acyl-CoA desaturase 1(m)	2	0
Sgk1	SGK-1(m)	28	0.004
Sirt6	SIR2L6(m)	20	0.001
Spp1	Osteopontin(m)	14	0
Stk11	LKB1(m)	27	0.048
Stk4	MST1(m)	27	0.01
Tab1	TAB1(m)	25	0.026
Tgfb1	TGFbeta1(m)	25	0.005
Tnf	TNF-alpha(m)	25	0.007
Traf1	TRAF1(m)	4	0.001
Traf2	TRAF2(m)	23	0.035

Gene name	Master molecule name	Reached from set	FDR
Vsir	GI24(m)	2	0
Wnt1	Wnt-1(m)	8	0
Wnt5a	wnt5a(m)	8	0
Zdhhc3	DHHC3(m)	30	0.047
Zdhhc7	DHHC7(m)	30	0.049

A.3. EGFR/WNT Pathway-related CRLM Dataset

Table A6.: ***Target-to-background* ratios of corresponding relevant PWMs in the analysis of the EGFR/WNT pathway-related CRLM dataset using *Target (Patient I)* and *Background (Patient II)*. *optBOR.conf***: optimized MSS cutoffs for a list of 37 confirmed PWMs by Boruta; ***TSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the target set; ***BSP%***: total percentage of promoter sequences with at least a site prediction that corresponds to the background set; ***TaF***: average site frequency that corresponds to the target set; ***BaF***: average site frequency that corresponds to the background set; ***target-to-backround ratio* (T/B ratio)**: target-to-background ratio: average site frequency in the target set divided by average site frequency in the background set.

PWM ID	*optBOR.conf*	*TSP%*	*BSP%*	*TaF*	*BaF*	*T/B ratio*
V$DEAF1_02	0.7850	20.6977	7.4419	0.2419	0.0860	2.8108
V$ZFP161_04	0.7500	78.1395	60.6977	6.3535	3.5209	1.8045
V$SP100_04	0.7500	99.3023	95.1163	17.6581	11.3791	1.5518
V$MECP2_02	0.7750	96.2791	91.1628	11.7581	7.6093	1.5452
V$CREB1_Q6	0.9350	43.0233	31.6279	0.6163	0.4000	1.5407
V$MAZR_01	0.7800	96.2791	93.7209	17.6256	11.4419	1.5404
V$RNF96_01	0.7500	97.6744	95.3488	18.2302	11.8442	1.5392
V$EGR1_Q6	0.7550	97.6744	95.5814	22.0721	14.4558	1.5269
V$MAZ_Q6_01	0.8150	94.6512	90.0000	11.8767	7.8209	1.5186
V$E2F_Q6_01	0.7550	92.3256	83.7209	6.8465	4.5186	1.5152
V$EKLF_Q5_01	0.8750	89.3023	78.3721	4.3837	2.9233	1.4996
V$SP1_Q6_01	0.8000	98.3721	95.8140	17.2070	11.9721	1.4373
V$CHCH_01	0.9825	100.0000	99.3023	28.1605	20.5163	1.3726
V$GLI_Q3	0.8850	96.2791	90.2326	7.8070	5.6907	1.3719
V$INSM1_01	0.7900	93.7209	90.9302	7.4605	5.5209	1.3513
V$AIRE_01	0.7500	100.0000	99.7674	26.9000	20.2000	1.3317
V$CPBP_Q6	0.9925	100.0000	99.3023	24.4721	18.5256	1.3210
V$BEN_01	0.7575	100.0000	100.0000	57.3000	46.1953	1.2404
V$GKLF_Q4	0.8900	100.0000	100.0000	38.6721	31.2721	1.2366
V$AHR_Q6	0.7600	100.0000	100.0000	35.1395	28.6047	1.2285
V$HIF1A_Q5	0.7800	99.3023	97.4419	7.5651	6.2372	1.2129
V$MTF1_Q5	0.8400	99.0698	97.6744	6.8256	5.6395	1.2103
V$MZF1_Q5	0.7575	100.0000	100.0000	51.3186	42.9605	1.1946
V$AML3_Q6	0.7950	100.0000	100.0000	23.3279	20.5326	1.1361
V$SF1_Q5_01	0.7850	100.0000	100.0000	13.8023	13.8349	0.9976
V$MYB_Q4	0.7600	100.0000	100.0000	37.1581	38.1023	0.9752
V$GEN_INI_B	0.7500	100.0000	100.0000	78.0744	84.5628	0.9233
V$YY1_Q6_03	0.7575	100.0000	100.0000	42.8512	47.4186	0.9037
V$SOX10_Q3	0.7675	100.0000	100.0000	72.6860	81.8070	0.8885
V$CIZ_01	0.7950	99.5349	100.0000	25.7023	30.0302	0.8559
V$RUSH1A_02	0.9000	100.0000	100.0000	16.1721	18.9558	0.8531
V$AP1_Q6_02	0.7850	100.0000	100.0000	17.8349	21.2628	0.8388
V$GATA_Q6	0.8325	99.5349	99.7674	9.7767	11.8814	0.8229
V$HELIOSA_02	0.8775	95.5814	96.5116	3.3326	4.0767	0.8175
V$PIT1_Q6_01	0.7850	100.0000	99.7674	20.3884	25.0163	0.8150
V$IPF1_Q5	0.7925	99.3023	98.8372	13.2256	16.8349	0.7856
V$CPHX_01	0.7525	73.4884	85.3488	2.5070	3.5047	0.7153

Table A7.: Top 20 significant Gene Ontology (GO) Biological Process (BP) terms obtained for the gene cluster *Cluster 1* after overrepresentation analysis (ORA) in the analysis of the EGFR/WNT pathway-related CRLM dataset.

GO Term ID	Description	p.adjust	Gene_Hits
GO:0035883	enteroendocrine cell differentiation	0.01264	BMP4,INSM1,NEUROD1,RFX6
GO:0003129	heart induction	0.01264	BMP4,MESP1,WNT11
GO:0003139	secondary heart field specification	0.01264	BMP4,MESP1,WNT11
GO:2000826	regulation of heart morphogenesis	0.01264	BMP4,MESP1,WNT11
GO:0001707	mesoderm formation	0.01388	BMP4,EPHA2,MESP1,TLX2,WNT11
GO:0048332	mesoderm morphogenesis	0.01388	BMP4,EPHA2,MESP1,TLX2,WNT11
GO:0048568	embryonic organ development	0.01971	BMP4,DLX6,EPHA2,HOXB6,ID3,MESP1,NEUROD1,PLCD3,TBX4,WNT11
GO:0030903	notochord development	0.01971	EPHA2,ID3,WNT11
GO:0031018	endocrine pancreas development	0.01971	BMP4,INSM1,NEUROD1,RFX6
GO:0046461	neutral lipid catabolic process	0.01971	APOA2,APOC3,FGF21,PNLIPRP2
GO:0046464	acylglycerol catabolic process	0.01971	APOA2,APOC3,FGF21,PNLIPRP2
GO:1904036	negative regulation of epithelial cell apoptotic process	0.02158	FGF21,NDNF,NEUROD1,TERT
GO:0035196	production of miRNAs involved in gene silencing by miRNA	0.02302	BMP4,MYCN,TERT,TRIM71
GO:0061311	cell surface receptor signaling pathway involved in heart development	0.038	BMP4,MESP1,WNT11
GO:2000637	positive regulation of gene silencing by miRNA	0.038	BMP4,MYCN,TRIM71
GO:0060148	positive regulation of posttranscriptional gene silencing	0.038	BMP4,MYCN,TRIM71
GO:0007498	mesoderm development	0.038	BMP4,EPHA2,MESP1,TLX2,WNT11
GO:0006721	terpenoid metabolic process	0.038	APOA2,APOC3,CYP2C19,LRP8,TTR
GO:0010453	regulation of cell fate commitment	0.04	BMP4,MESP1,WNT11
GO:0031128	developmental induction	0.04	BMP4,MESP1,WNT11

Table A8.: Master regulator (MR) candidates obtained for the gene cluster *Cluster 1* after MR search in the analysis of the EGFR/WNT pathway-related CRLM dataset. MRs are alphabhtically sorted by HGNC gene name.

Gene Name	Master molecule name	Reached from set	FDR
ACP1	LMW-PTP-isoform1(h)	12	0.0
ADAM10	adam10(h)	10	0.0
AKT1	AKT-1(h)p	13	0.0
APOC3	APOC3(h)	5	0.0
AURKB	Aurora-B(h)	14	0.022
BMP4	BMP4(h)	5	0.0
CAPN1	calpain-1(h)	13	0.011
CBL	c-Cbl(h)	18	0.03
CCNA1	cyclinA1(h)	13	0.014
CCNA2	cyclinA2(h)	13	0.037
CCNB1	cyclinB1(h):Cdk1(h)	12	0.014
CDK1	cyclinA(h):Cdk1(h)	12	0.0
COPS2	CSN(h)	11	0.004
COPS5	CSN(h)	11	0.004
CSNK2A1	CKII-alpha(h):CKII-alpha2(h):CKII-beta(h)	14	0.001
CSNK2A2	CKII-alpha(h):CKII-alpha2(h):CKII-beta(h)	14	0.001
CSNK2B	CKII-alpha(h):CKII-alpha2(h):CKII-beta(h)	14	0.001
CXCL13	BCA-1(h)	3	0.0
CXCL9	MIG(h)	5	0.0
EHMT2	G9A(h)	9	0.0
EP300	p300(h)	12	0.014
EPHA2	Eck(h)	7	0.0
FAT1	fat1(h)	3	0.0
FURIN	PACE(h)	13	0.001
GRIN1	NR1(h)	11	0.009
GRK2	betaARK-1(h)	15	0.027
GSK3B	GSK3beta-isoform2(h)	11	0.003

Gene Name	Master molecule name	Reached from set	FDR
HUWE1	MULE(h)	9	0.001
IGF2BP1	IMP-1(h)	10	0.001
IL10	IL-10(h)	15	0.008
KAT2A	GCN5(h)	9	0.0
KAT2B	p/CAF(h)	13	0.006
KLK6	kallikrein-6-isoform1(h)	11	0.014
MAP3K10	MLK2(h)	8	0.006
MAPK6	ERK3(h)	15	0.006
MAPK7	ERK5(h)	15	0.007
MUC4	MUC4(h)	11	0.0
MYCN	N-Myc(h)	4	0.0
NEUROD1	NeuroD(h)ace	1	0.0
NFKBIA	IkappaB-alpha(h)	14	0.001
NTRK2	trkB(h)	12	0.017
PDGFRA	PDGFRalpha(h)	14	0.008
PDGFRB	PDGFRbeta(h)	14	0.006
PHEX	Pex(h)	1	0.0
PINK1	Pink1(h)	10	0.001
PRKACA	PKACA-isoform2(h)	11	0.007
PRKCZ	PKCzeta-isoform1(h)	13	0.029
PRKDC	DNA-PKcs-isoform1(h)	10	0.017
PRKN	parkin(h)	13	0.001
PTPRF	LAR(h)	14	0.0
PTPRO	PTPRO(h)	11	0.005
PTPRZ1	RPTPzeta-L(h)	14	0.043
RCHY1	PIRH2(h)	16	0.018
RNF19A	dorfin(h)	10	0.0
RNF41	Nrdp1(h)	16	0.04

Gene Name	Master molecule name	Reached from set	FDR
SIAH1	Siah1(h)	16	0.0
SIAH2	Siah-2(h)	14	0.0
SIRT6	SIR2L6-isoform1(h)	12	0.0
SNCA	alpha-synuclein(h)	15	0.007
SNCAIP	Synphilin-1(h)	12	0.0
SPRY2	Sprouty2(h)	14	0.044
SRC	Src(h)pY419	15	0.022
STUB1	CHIP(h)	14	0.009
SYK	Syk-isoform1(h)	18	0.005
TERT	tert-isoform1(h)	2	0.0
TGM2	TGC(h)	11	0.027
TNF	TNF-alpha(h)	12	0.003
UBA1	E1-isoform1(h)ub	14	0.016
UBE2A	rad6A-isoform1(h)	14	0.028
UBE2C	E2-C-isoform1(h)	16	0.05
UBE2D2	Ubc5B(h)ub(1)	12	0.0
UBE2D3	Ubc5C(h)ub(1)	14	0.01
UBE2E3	UBE2E3(h)	14	0.025
UBE2G1	Ubc7(h)	11	0.001
UBE2G2	UBE2G2(h)	14	0.001
UBE2J1	Ubc6(h)	14	0.02
UBE2K	HIP2-isoform1(h)	14	0.029
UBE2L3	UbcH7(h)	16	0.019
UBE2N	Uev1-isoform1(h):Ubc13(h)	11	0.011
UBE2Q2	UBE2Q2-isoform1(h)	14	0.016
UBE2R2	Ubc3B(h)	14	0.018
UBE2S	E2-EPF(h)	14	0.026
UBE2V1	Uev1-isoform1(h):Ubc13(h)	11	0.011
UBE3A	E6AP(h)	17	0.048
WNT11	WNT11(h)	4	0.0

B. Appendix B: Miscellaneous

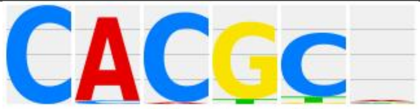
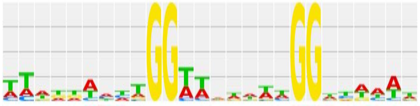
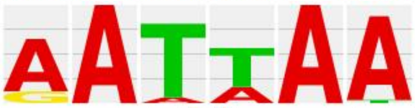
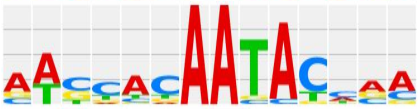
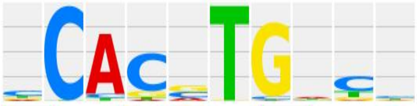
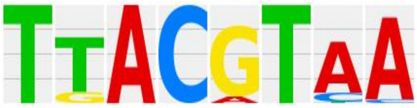
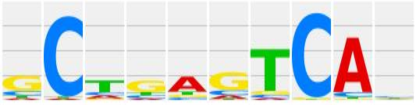
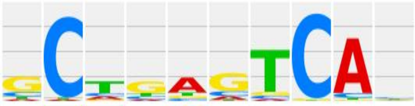
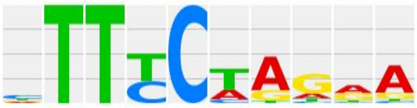
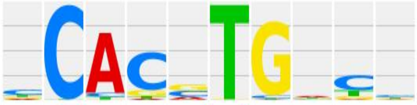
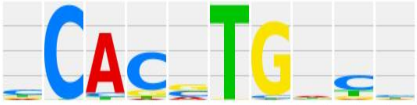
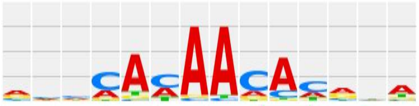
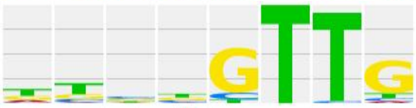
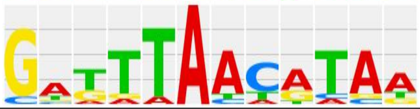
B.1. Abstract

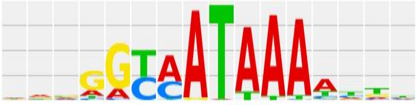
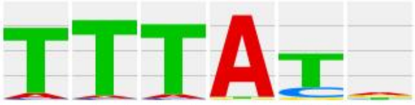
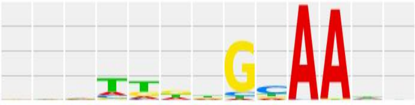
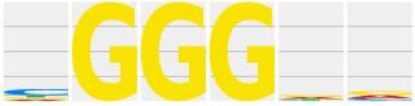
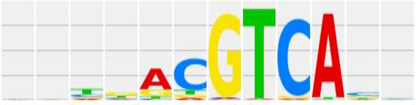
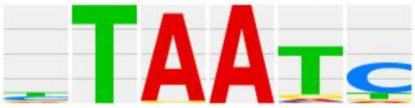
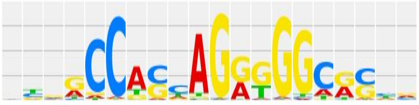
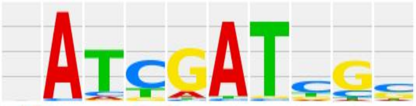
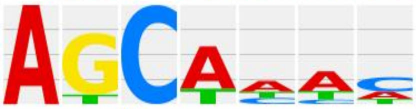
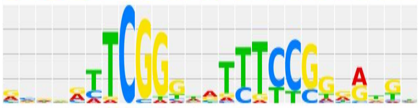
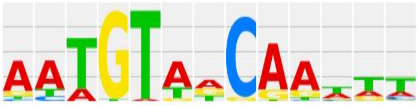
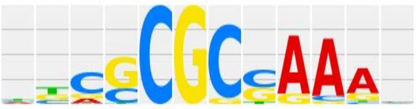
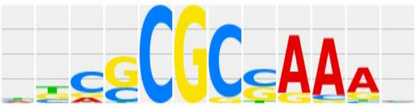
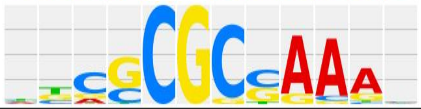
Distant metastases are a decisive event in cancer progression, and 70% of patients with colorectal cancer develop liver metastases (CRLM). Therapy outcome is largely influenced by tumor heterogeneity, however, intra- and interpatient heterogeneity of CRLM have been poorly studied. In particular, the contribution of the WNT and EGFR pathway, which both are frequently deregulated in colorectal cancer, has not been addressed in this context yet. To this end, we comprehensively characterized normal liver tissue and eight CRLM from two patients by standardized histopathological, molecular and proteomic subtyping. Suitable fresh frozen tissue samples were profiled by transcriptome sequencing (RNA-Seq) and proteomic profiling with reverse phase protein arrays (RPPA) combined with bioinformatic analyses to assess tumor heterogeneity and identify WNT- and EGFR-related master regulators and metastatic effectors. A standardized data analysis pipeline for the integration of RNA-Seq with clinical and genetic data was established. Dimensionality reduction of the transcriptome data revealed a distinct signature for CRLM compared to normal liver tissue and pointed at a high degree of tumor heterogeneity. WNT and EGFR signaling were highly active in CRLM and genes of both pathways were heterogeneously expressed between both patients and between synchronous metastases of a single patient. An analysis of the master regulators and metastatic effectors implicated in the regulation of these genes revealed a set of genes (SFN, IGF2BP1, PRKCB, STAT1, PIK3CG) that were differentially expressed in CRLM and associated with clinical outcome in a large cohort of colorectal cancer patients. In conclusion, high-throughput profiling allows the definition of a CRLM-specific signature and reveals WNT and EGFR genes associated with inter- and intra-patient heterogeneity that are expressed in a large cohort of colorectal cancer patients and are of prognostic relevance.

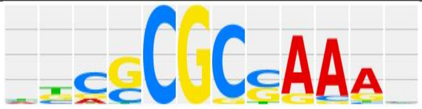
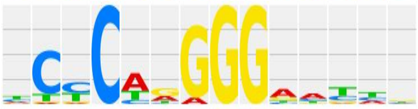
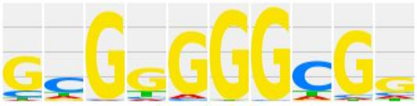
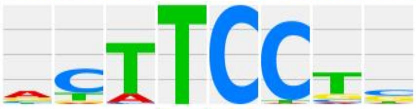
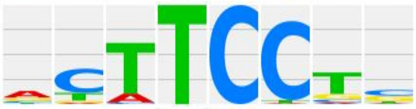
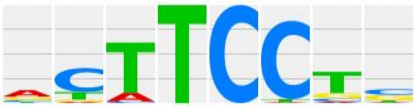
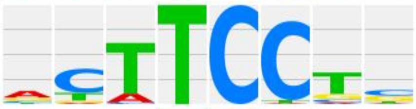
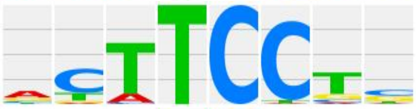
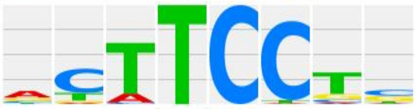
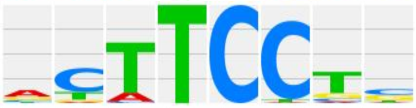
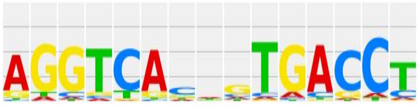
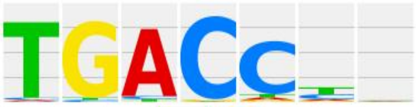
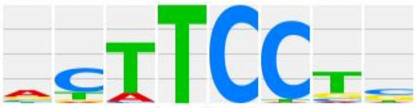
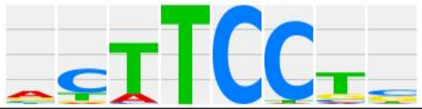
- Menck K, Wlochowitz D, Wachter A, Conradi LC, Wolff A, Peeck M, Scheel A, Schlüter K, Korf U, Wiemann S, Schildhaus HU, Wingender E, Pukrop T, Homayounfar K, Beißbarth T, Bleckmann A. **High-throughput profiling of colorectal cancer liver metastases reveals intra- and interpatient heterogeneity in the EGFR- and WNT-pathway associated with clinical outcome**. [In preparation (2022), [115]].

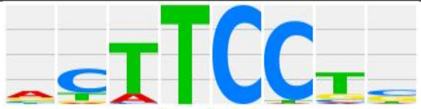
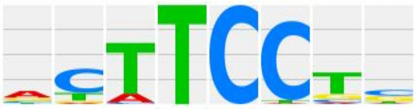
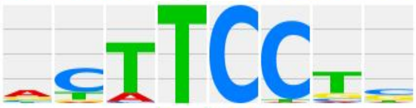
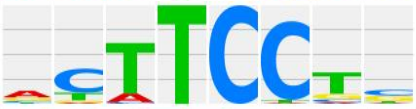
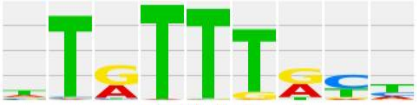
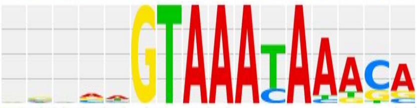
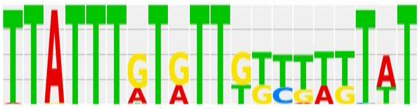
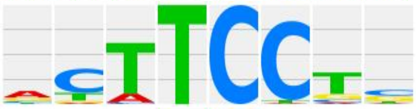
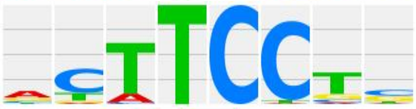
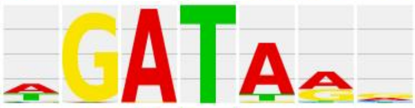
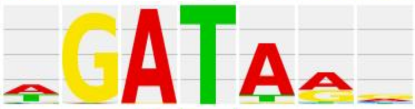
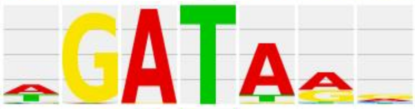
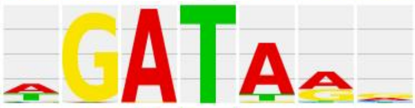
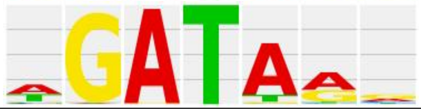
B.2. Annotations for TRANSFAC$^{®}$ PWM Library

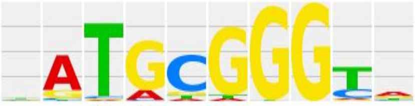
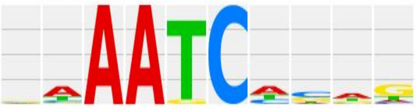
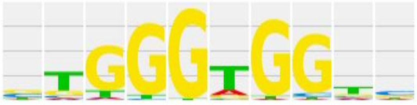
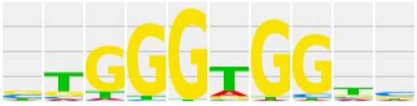
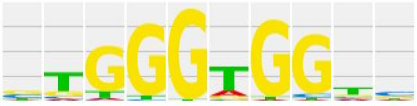
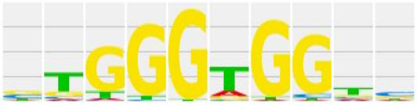
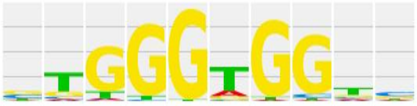
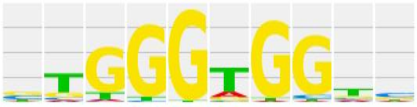
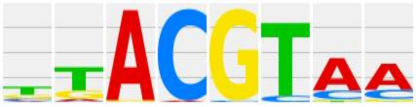
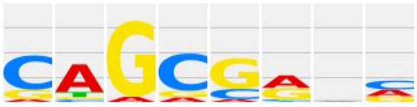
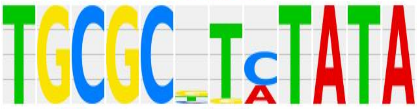
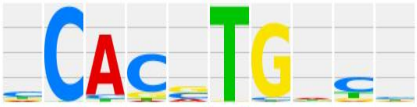
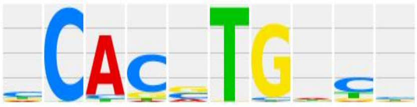
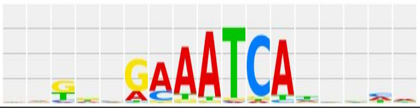
Table B1.: Annotations for TRANSFAC® vertebrate non-redundant library (release 2019.3) alphabetically sorted by HGNC gene name.

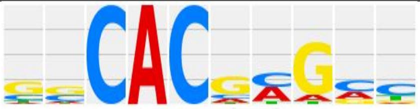
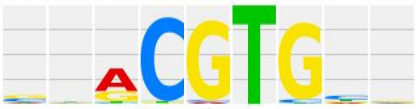
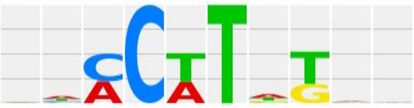
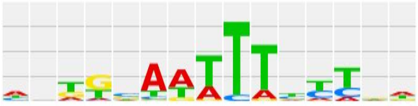
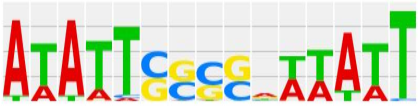
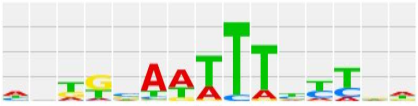
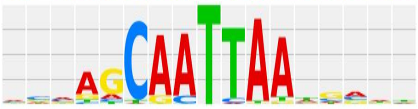
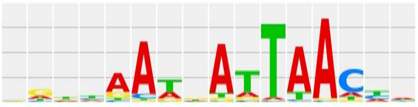
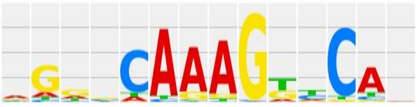
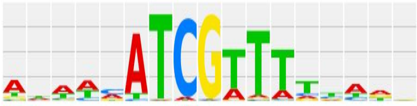
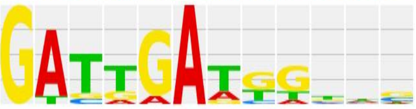
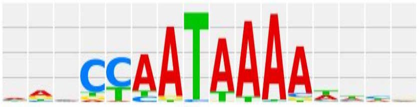
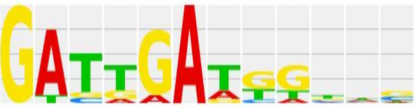
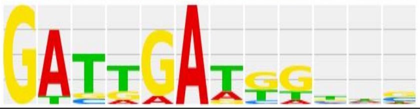
Gene name	PWM	Sequence logo
AHR	V$AHR_Q6	
AIRE	V$AIRE_01	
ARID3A	V$DRI1_01	
ARID5B	V$MRF2_01	
ASCL1	V$EBOX_Q6_01	
ATF2	V$CREBP1_01	
BACH1	V$MAF_Q4	
BACH2	V$MAF_Q4	
BCL6	V$BCL6_Q3_01	
BHLHE40	V$EBOX_Q6_01	
BHLHE41	V$EBOX_Q6_01	
BPTF	V$FAC1_01	
BRCA1	V$BRCA_01	
CDC5L	V$CDC5_01	

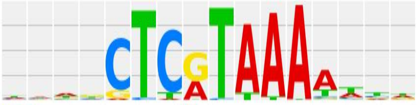
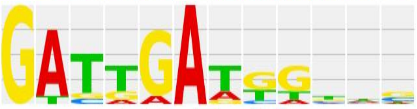
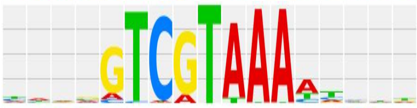
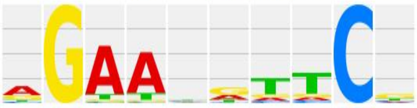
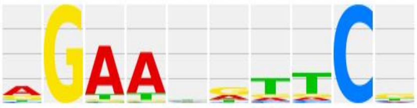
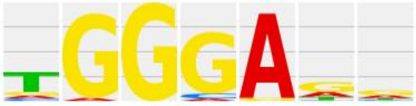
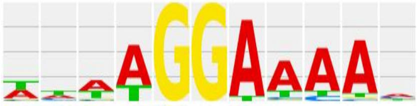
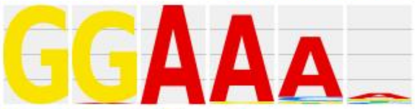
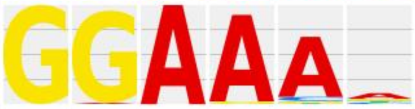
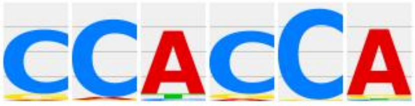
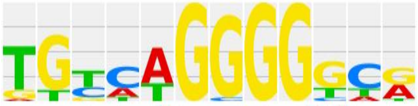
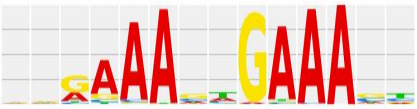
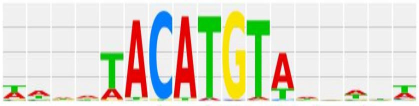
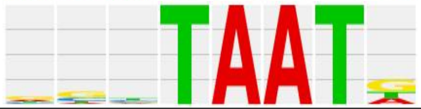
Gene name	PWM	Sequence logo
CDX2	V$CDX2_01	
CDX2	V$CDX2_Q5_02	
CEBPA	V$CEBPA_Q6	
CHURC1	V$CHCH_01	
CREB1	V$CREB1_Q6	
CRX	V$CRX_Q4_01	
CTCF	V$CTCF_01	
CUX1	V$CDPCR1_01	
DBP	V$DBP_Q6	
DEAF1	V$DEAF1_02	
DMRTA1	V$DMRT4_01	
E2F1	V$E2F_Q6_01	
E2F3	V$E2F_Q6_01	
E2F4	V$E2F_Q6_01	

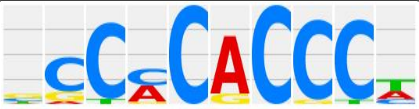
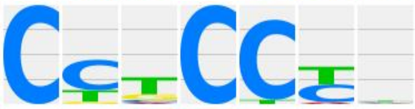
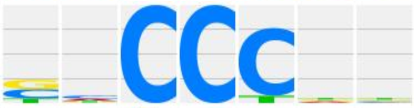
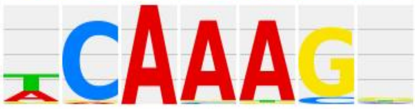
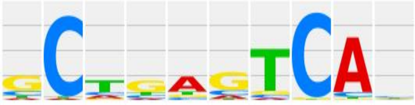
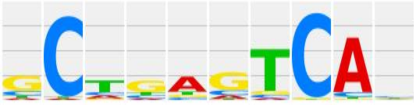
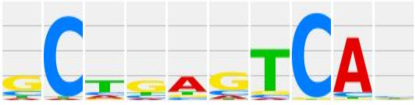
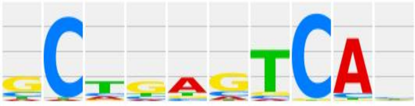
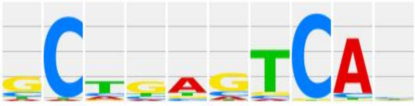
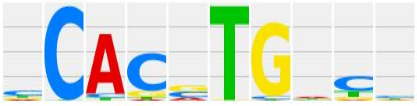
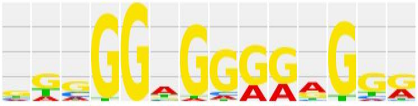
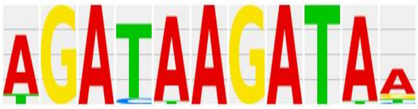
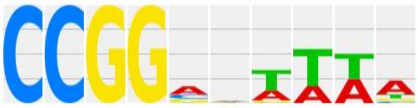
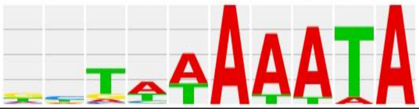
Gene name	PWM	Sequence logo
E2F7	V$E2F_Q6_01	
EBF1	V$COE1_Q6	
EGR1	V$EGR1_Q6	
ELF1	V$ETS_Q6	
ELF2	V$ETS_Q6	
ELF4	V$ETS_Q6	
ELK1	V$ETS_Q6	
ELK3	V$ETS_Q6	
ELK4	V$ETS_Q6	
ERG	V$ETS_Q6	
ESR1	V$ERALPHA_01	
ESR1	V$ERALPHA_Q6_01	
ETS1	V$ETS_Q6	
ETS2	V$ETS_Q6	

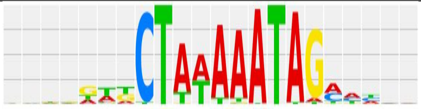
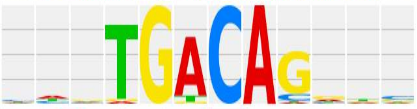
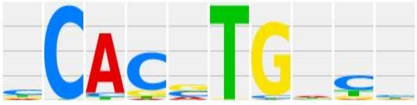
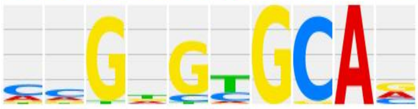
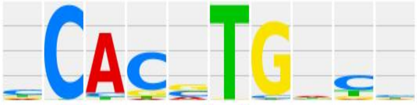
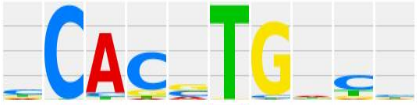
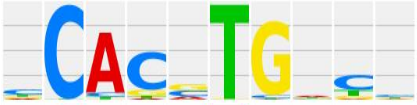
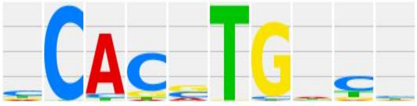
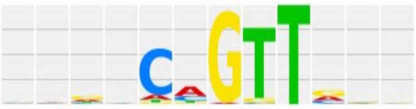
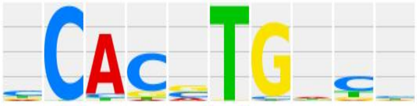
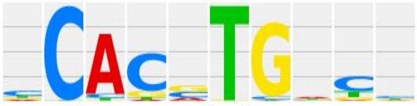
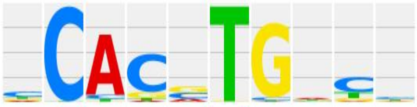
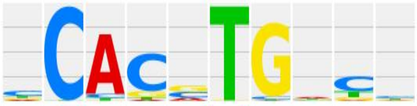
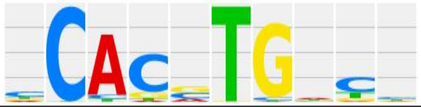
Gene name	PWM	Sequence logo
ETV4	V$ETS_Q6	
ETV6	V$ETS_Q6	
ETV7	V$ETS_Q6	
FLI1	V$ETS_Q6	
FOXA2	V$HNF3B_Q6	
FOXC1	V$FREAC3_01	
FOXP1	V$FOXP1_01	
GABPA	V$ETS_Q6	
GABPB1	V$ETS_Q6	
GATA1	V$GATA_Q6	
GATA2	V$GATA_Q6	
GATA3	V$GATA_Q6	
GATA4	V$GATA_Q6	
GATA6	V$GATA_Q6	

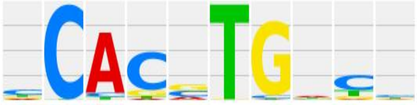
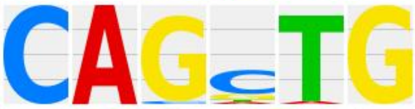
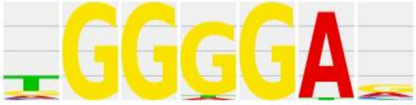
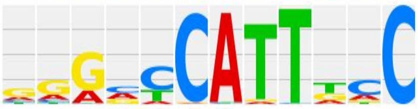
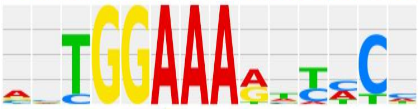
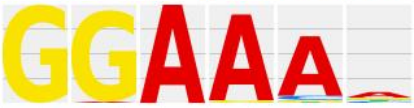
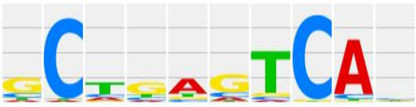
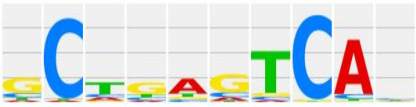
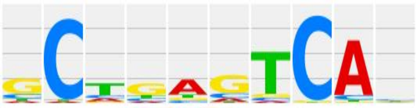
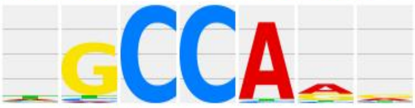
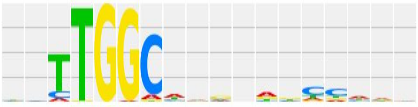
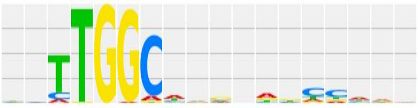
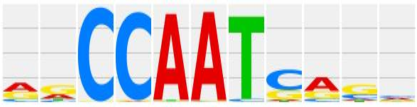
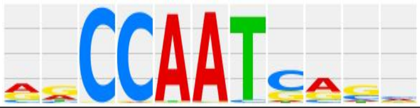
Gene name	PWM	Sequence logo
GCM2	V$GCM2_01	
GFI1	V$GFI1_Q6_01	
GLI1	V$GLI_Q3	
GLI2	V$GLI_Q3	
GLI3	V$GLI_Q3	
GLIS1	V$GLI_Q3	
GLIS2	V$GLI_Q3	
GLIS3	V$GLI_Q3	
GMEB2	V$GMEB2_04	
GTF2IRD1	V$BEN_01	
GZF1	V$GZF1_01	
HAND1	V$EBOX_Q6_01	
HAND2	V$EBOX_Q6_01	
HDX	V$HDX_01	

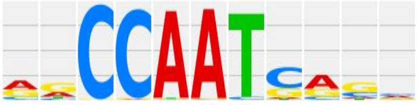
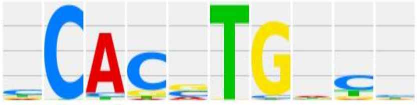
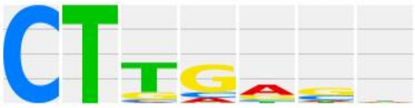
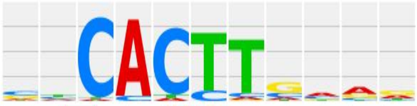
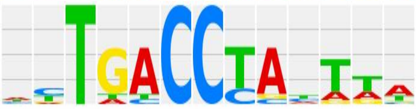
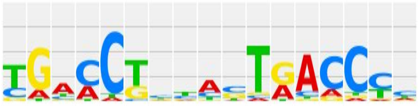
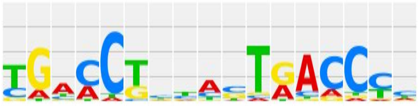
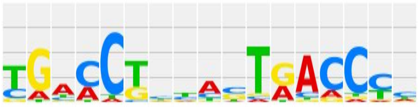
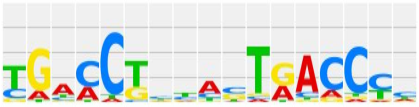
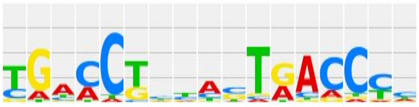
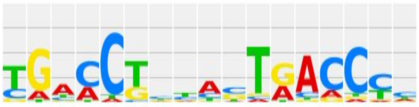
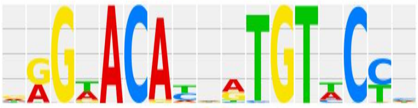
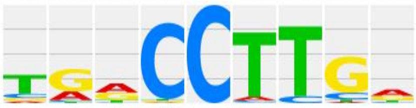
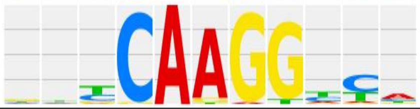
Gene name	PWM	Sequence logo
HES1	V$HES1_Q6	
HIF1A	V$HIF1A_Q5	
HLTF	V$RUSH1A_02	
HMGA1	V$HMGIY_Q3	
HMGA2	V$HMGA2_01	
HMGA2	V$HMGIY_Q3	
HMX1	V$HMX1_02	
HNF1A	V$HNF1A_Q4	
HNF4A	V$HNF4A_Q3	
HOMEZ	V$HOMEZ_01	
HOXB1	V$PBX_Q3	
HOXB13	V$HOXB13_01	
HOXB7	V$PBX_Q3	
HOXB8	V$PBX_Q3	

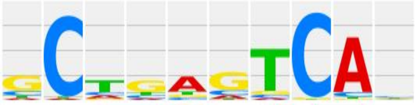
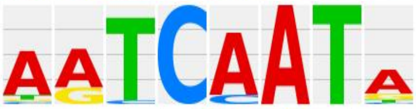
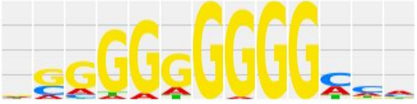
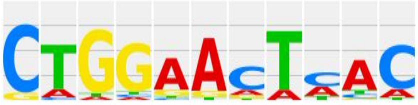
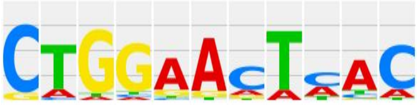
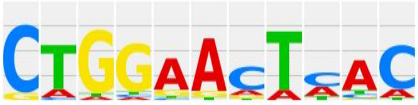
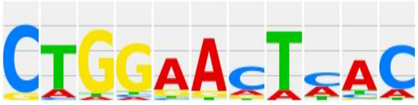
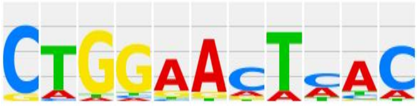
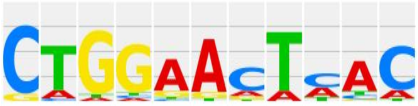
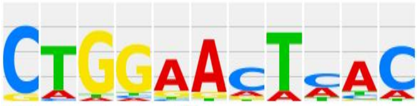
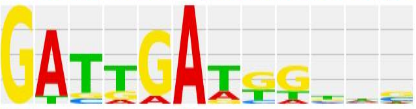
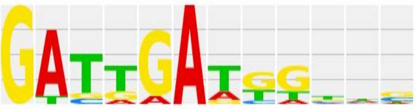
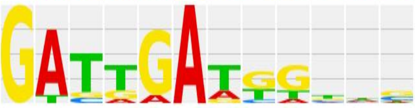
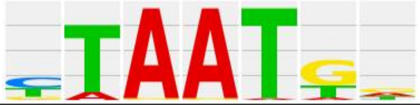
Gene name	PWM	Sequence logo
HOXC13	V$HOXC13_01	
HOXC6	V$PBX_Q3	
HOXD12	V$HOXD12_01	
HSF1	V$HSF1_01	
HSF1	V$HSF1_02	
IKZF1	V$IK_Q5_01	
IKZF2	V$HELIOSA_02	
ILF2	V$NFAT1_Q4	
ILF3	V$NFAT1_Q4	
ING4	V$ING4_01	
INSM1	V$INSM1_01	
IRF1	V$IRF1_Q5	
IRX2	V$IRX2_01	
ISL1	V$ISL1_Q3	

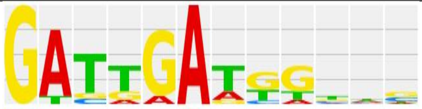
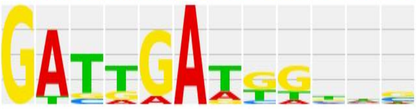
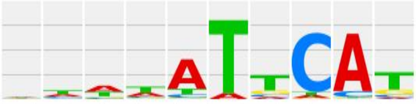
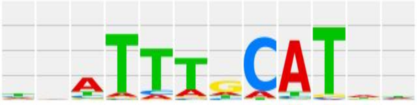
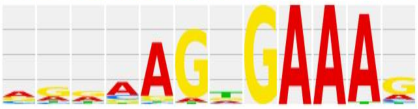
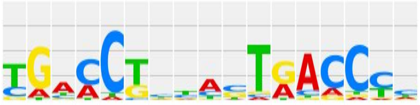
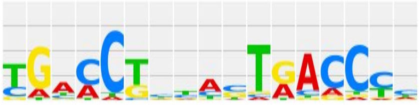
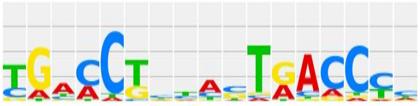
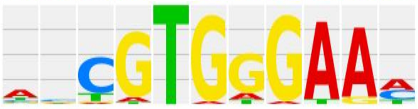
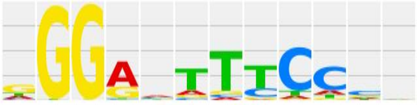
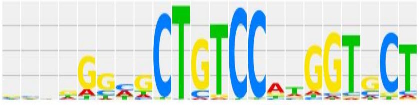
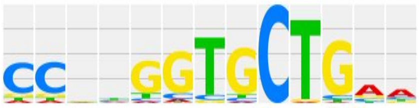
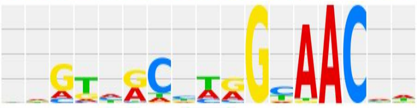
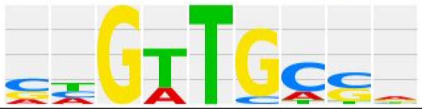
Gene name	PWM	Sequence logo
KLF1	V$EKLF_Q5_01	
KLF4	V$GKLF_Q4	
KLF6	V$CPBP_Q6	
LEF1	V$LEF1_Q5_01	
MAF	V$MAF_Q4	
MAFA	V$MAFA_Q4	
MAFB	V$MAF_Q4	
MAFG	V$MAF_Q4	
MAFK	V$MAF_Q4	
MAX	V$EBOX_Q6_01	
MAZ	V$MAZ_Q6_01	
MECOM	V$EVI1_03	
MECP2	V$MECP2_02	
MEF2A	V$MEF2A_Q6	

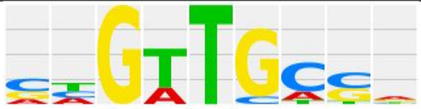
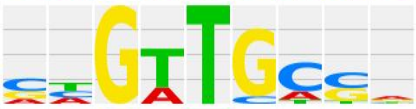
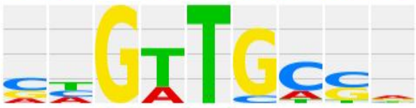
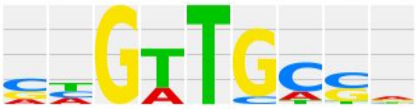
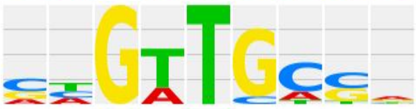
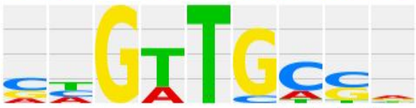
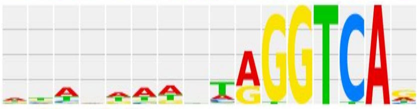
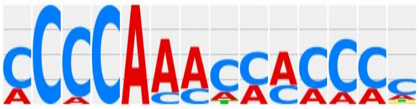
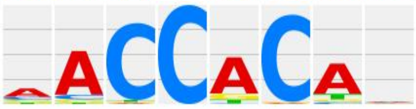
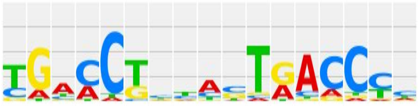
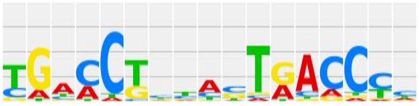
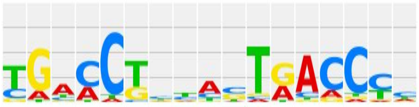
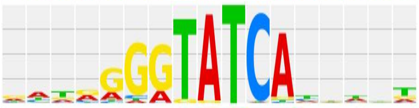
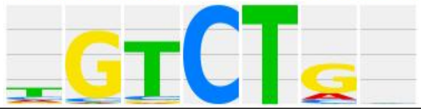
Gene name	PWM	Sequence logo
MEF2A	V$MEF2_03	
MEIS1	V$MEIS1_01	
MITF	V$EBOX_Q6_01	
MTF1	V$MTF1_Q5	
MXD1	V$EBOX_Q6_01	
MXD3	V$EBOX_Q6_01	
MXD4	V$EBOX_Q6_01	
MXI1	V$EBOX_Q6_01	
MYB	V$MYB_Q4	
MYC	V$EBOX_Q6_01	
MYCN	V$EBOX_Q6_01	
MYF5	V$EBOX_Q6_01	
MYF6	V$EBOX_Q6_01	
MYOD1	V$EBOX_Q6_01	

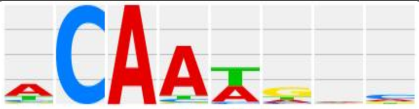
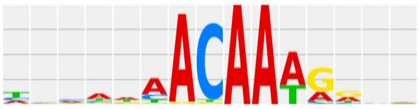
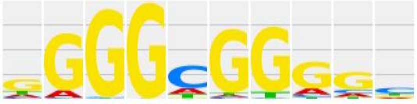
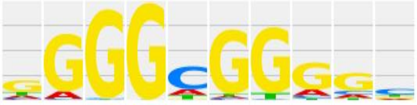
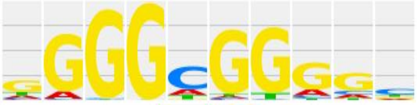
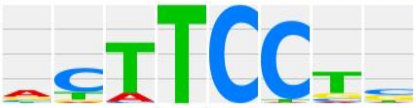
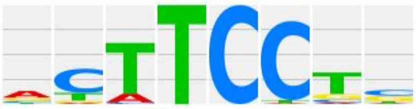
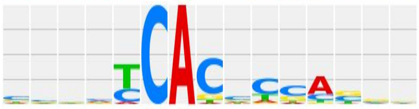
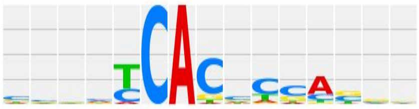
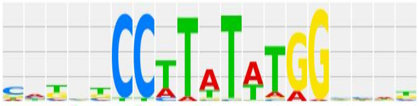
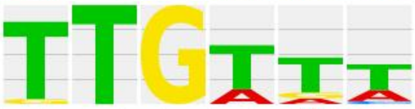
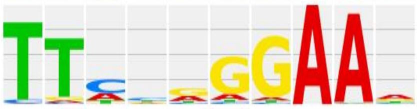
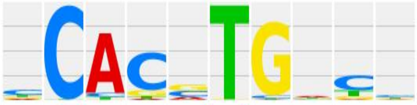
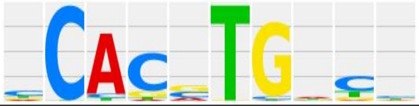
Gene name	PWM	Sequence logo
MYOG	V$EBOX_Q6_01	
MYOG	V$MYOGENIN_Q6_01	
MZF1	V$MZF1_Q5	
NANOG	V$NANOG_01	
NFAT5	V$NFAT5_Q5_02	
NFATC2	V$NFAT1_Q4	
NFE2	V$MAF_Q4	
NFE2L1	V$MAF_Q4	
NFE2L2	V$MAF_Q4	
NFIA	V$NF1A_Q6_01	
NFIA	V$NF1_Q6	
NFIC	V$NF1_Q6	
NFYA	V$NFY_Q3	
NFYB	V$NFY_Q3	

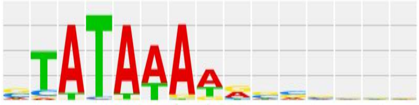
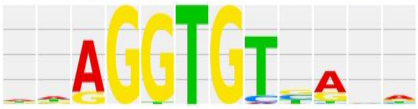
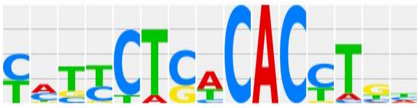
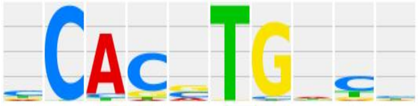
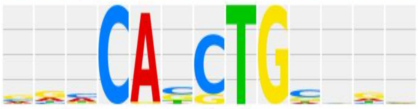
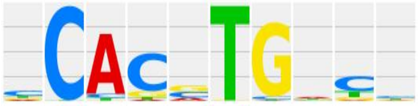
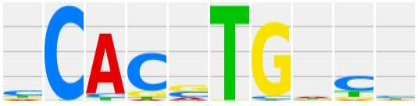
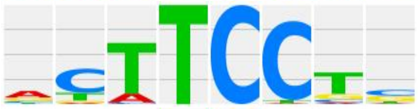
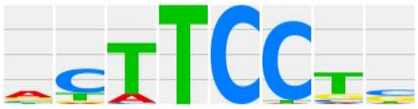
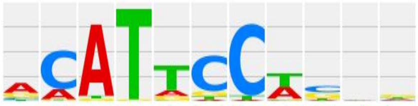
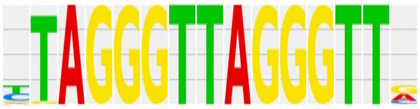
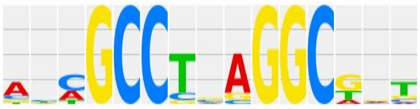
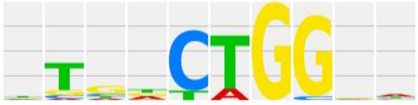
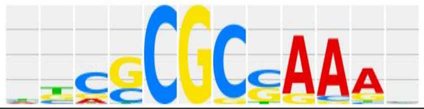
Gene name	PWM	Sequence logo
NFYC	V$NFY_Q3	
NHLH1	V$EBOX_Q6_01	
NKX2-1	V$TTF1_Q5_01	
NKX2-5	V$NKX25_Q6	
NR1D1	V$REVERBALPHA_Q6	
NR1H2	V$DR4_Q2	
NR1H3	V$DR4_Q2	
NR1I2	V$DR4_Q2	
NR1I3	V$DR4_Q2	
NR2F1	V$DR4_Q2	
NR2F2	V$DR4_Q2	
NR3C1	V$NR3C1_03	
NR5A1	V$SF1_Q5_01	
NR5A2	V$LRH1_Q5_01	

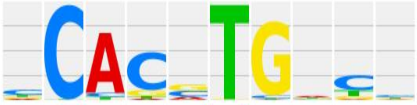
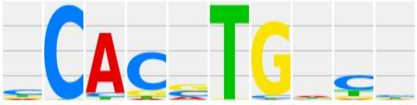
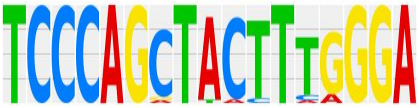
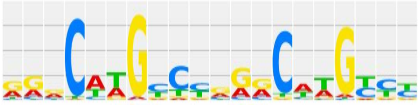
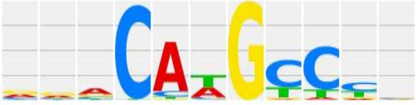
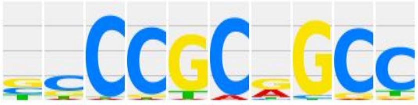
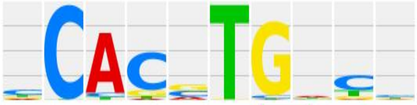
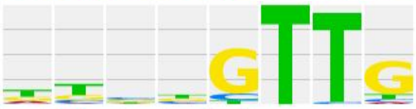
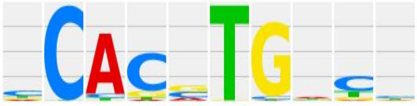
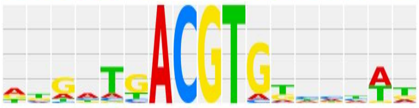
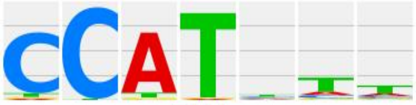
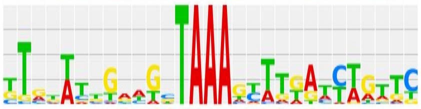
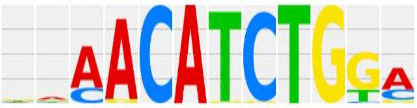
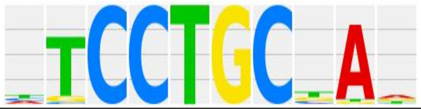
Gene name	PWM	Sequence logo
NRF1	V$MAF_Q4	
ONECUT1	V$HNF6_Q4	
PATZ1	V$MAZR_01	
PAX1	V$PAX_Q6	
PAX2	V$PAX_Q6	
PAX3	V$PAX_Q6	
PAX4	V$PAX_Q6	
PAX5	V$PAX_Q6	
PAX6	V$PAX_Q6	
PAX8	V$PAX_Q6	
PBX1	V$PBX_Q3	
PBX2	V$PBX_Q3	
PBX3	V$PBX_Q3	
PDX1	V$IPF1_Q5	

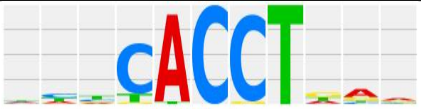
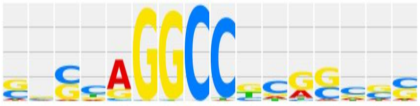
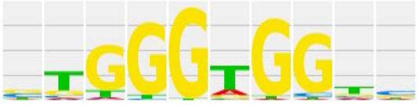
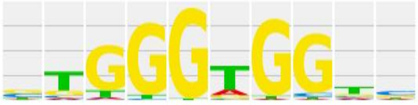
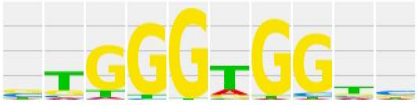
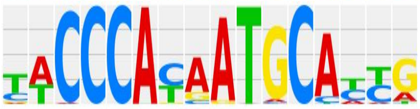
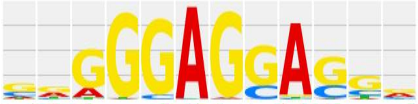
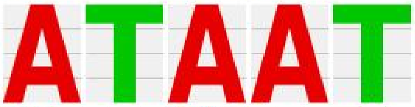
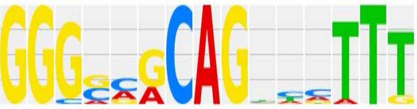
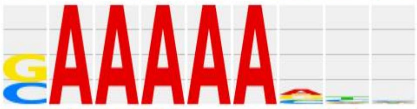
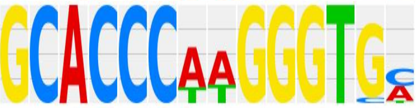
Gene name	PWM	Sequence logo
PDX1	V$PBX_Q3	
PKNOX1	V$PBX_Q3	
POU1F1	V$PIT1_Q6_01	
POU2F1	V$POU2F1_Q6	
PRDM1	V$BLIMP1_Q4	
RARA	V$DR4_Q2	
RARB	V$DR4_Q2	
RARG	V$DR4_Q2	
RBPJ	V$RBPJK_01	
RELA	V$RELA_Q6	
REST	V$REST_01	
REST	V$REST_Q5	
RFX1	V$RFX1_01	
RFX1	V$RFX_Q6	

Gene name	PWM	Sequence logo
RFX2	V$RFX_Q6	
RFX3	V$RFX_Q6	
RFX4	V$RFX_Q6	
RFX5	V$RFX_Q6	
RFXANK	V$RFX_Q6	
RFXAP	V$RFX_Q6	
RORA	V$RORALPHA_Q4	
RREB1	V$RREB1_01	
RUNX2	V$AML3_Q6	
RXRA	V$DR4_Q2	
RXRB	V$DR4_Q2	
RXRG	V$DR4_Q2	
SIX1	V$SIX1_01	
SMAD4	V$SMAD4_Q6_01	

Gene name	PWM	Sequence logo
SOX10	V$SOX10_Q3	
SOX2	V$SOX2_Q3_01	
SP1	V$SP1_Q6_01	
SP3	V$SP1_Q6_01	
SP4	V$SP1_Q6_01	
SPI1	V$ETS_Q6	
SPIB	V$ETS_Q6	
SREBF1	V$SREBP_Q6	
SREBF2	V$SREBP_Q6	
SRF	V$SRF_Q5_02	
SRY	V$SRY_Q6	
STAT1	V$STAT1_Q6	
TAL1	V$EBOX_Q6_01	
TAL2	V$EBOX_Q6_01	

Gene name	PWM	Sequence logo
TBP	V$TATA_01	
TBX5	V$TBX5_01	
TBX5	V$TBX5_Q2	
TCF12	V$EBOX_Q6_01	
TCF3	V$E2A_Q6_01	
TCF3	V$EBOX_Q6_01	
TCF4	V$EBOX_Q6_01	
TCF7	V$ETS_Q6	
TCF7L2	V$ETS_Q6	
TEAD1	V$TEF1_Q6_04	
TERF1	V$TRF1_01	
TFAP2A	V$AP2ALPHA_03	
TFCP2	V$CP2_Q6	
TFDP1	V$E2F_Q6_01	

Gene name	PWM	Sequence logo
TFE3	V$EBOX_Q6_01	
TFEB	V$EBOX_Q6_01	
TOPORS	V$LUN1_01	
TP53	V$P53_04	
TP53	V$P53_Q3	
TRIM28	V$RNF96_01	
USF1	V$EBOX_Q6_01	
USF2	V$BRCA_01	
USF2	V$EBOX_Q6_01	
XBP1	V$XBP1_01	
YY1	V$YY1_Q6_03	
ZBTB16	V$PLZF_02	
ZBTB18	V$RP58_01	
ZBTB33	V$KAISO_01	

Gene name	PWM	Sequence logo
ZEB1	V$DELTAEF1_01	
ZFX	V$ZFX_01	
ZIC1	V$GLI_Q3	
ZIC2	V$GLI_Q3	
ZIC3	V$GLI_Q3	
ZNF143	V$ZNF143_03	
ZNF263	V$FPM315_01	
ZNF333	V$ZNF333_01	
ZNF350	V$ZBRK1_01	
ZNF384	V$CIZ_01	
ZNF423	V$ROAZ_01	
ZSCAN10	V$ZFP206_01