Freie Universität Berlin

# Prediction of Transcription Factor Binding Site to enrich wTO package

**Thesis submitted as fulfillment for the degree**
**Bachelor of Science in Bioinformatics**
**by**
**Jin Soo Park**

**26.04.2022**

1st Reviewer:
Prof. Dr. Katja Nowick
Freie Universität Berlin

2nd Reviewer:
PhD Vladimir Bajić
Freie Universität Berlin

This thesis was prepared at the Freie Universität Berlin, Institut for Biology, Department of Human Biology, under the supervision of Prof. Dr. Katja Nowick and PhD Vladimir Bajić.

1st Reviewer:
Prof. Dr. Katja Nowick
Freie Universität Berlin
Königin-Luise-Str. 1-3
14195 Berlin

2nd Reviewer:
PhD. Vladimir Bajić
Freie Universität Berlin
Königin-Luise-Str. 1-3
14195 Berlin

# Eidesstattliche Erklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Berlin, den 26.04.2022

Jin Soo Park

# List of Figures

# List of Tables

# Abstract

The major central question of biology is to systematically identify how genotype influences phenotype. However, since gene regulatory procedure can involve complex interaction of a larger number of genes, spotting metabolic networks and ecological networks can be challenging. This regulation of expressed genes is mostly based on the DNA - protein interaction. To understand the mediation of biological processes requires comprehending how transcription factors (TFs) function sequence-specifically. This specific sequence recognized by TFs is called motif, whereas transcription factor binding sites (TFBS) are characterized. Co-expressed genes are more likely to contain shared TFBS and TFBS can be predicted and associated computationally. Gene pairs with a very high expression correlation show a significant excess of shared binding sites. It is functional to identify TFBS via the presence of common sequence motifs to correlate the expression levels of genes. Discovering the TFs can be the initial step to acknowledge the basic language of the co-express network. The wTO package supports beneficial features for generation and analysis of such networks. Different bioinformatic tools were applied to enhance the wTO package.

# Zusammenfassung

Die zentrale Frage der Biologie besteht darin, systematisch zu ermitteln, wie der Genotyp den Phänotyp beeinflusst. Da die Genregulation jedoch eine komplexe Interaktion zwischen einer größeren Anzahl von Genen umfassen kann, ist die Erkennung von metabolischen Netzwerken und ökologischen Netzwerken eine Herausforderung. Diese Regulierung exprimierter Gene basiert meist auf der DNA-Protein-Interaktion. Um die Vermittlung biologischer Prozesse zu verstehen, muss man begreifen, wie Transkriptionsfaktoren (TFs) sequenzspezifisch funktionieren. Diese besondere sequenzspezifische Erkennung von TFs wird als Motiv bezeichnet, während Transkriptionsfaktor-Bindungsstellen (TFBS) charakterisiert werden. Bei gemeinsam exprimierten Genen ist wahrscheinlicher, dass sie gemeinsame TFBS enthalten, und TFBS können rechnerisch vorhergesagt und zugeordnet werden. Genpaare mit einer sehr hohen Expressionskorrelation weisen einen erheblichen Überschuss an gemeinsamen Bindungsstellen. Es ist funktionell, TFBS über das Vorhandensein gemeinsamer Sequenzmotive zu identifizieren, um die Expressionsniveaus von Genen zu korrelieren. Die Entdeckung der TFs kann der erste Schritt sein, um die grundlegende Sprache des Co-Express-Netzwerks anzuerkennen. Das wTO-Paket unterstützt nützliche Funktionen für die Generierung und Analyse solcher Netzwerke. Zur Verbesserung des wTO-Pakets wurden verschiedene bioinformatische Werkzeuge eingesetzt.

# Table of Contents

# Abbreviation

**BD : Bipolar disease**
**BDNF : Brain-derived neurotrophic factor**
**BP : Base pair**
**Chip-seq: Chromatin Immunoprecipitation sequencing**
**CN : Consensus network**
**CSV : Comma separated values**
**CT : Control**
**DNA : Deoxyribonucleic acid**
**ER : Endoplasmic reticulum**
**FDR : False discovery rates**
**FIMO : Find individual motif occurrence**
**FUB: Freie universität berlin**
**GRF: gene regulatory factor**
**GTRD : Gene transcription regulation database**
**HMM : Hidden markov model**
**ID : Identifier**
**KBP: Kilo base pair**
**MEME : Multiple expectation maximization for motif elicitation**
**NT : Nucleotide**
**PCM : Position count matrix**
**PFM : Position frequency matrix**
**PSSM : Position specific scoring matrix**
**PWM : Position weight matrix**
**RF : Random forest**
**RNA : Ribonucleic acid**
**SNV : Single nucleotide variation**
**SZ : Schizophrenia**
**TF : Transcription factor**
**TFFM :Transcript factor flexible model**
**TSS: Transcription starting site**
**TSV : Tab separated values**
**TFBS : Transcription factor binding sites**
**Unix-OS : Unix operating system**
**wTO : Weight topology overlaps**

# 1. Introduction

## 1.1. Biological background of transcription factors

Our genetic information is stored in the form of double helix, so-called deoxyribonucleic acid (DNA), and this information is crucial for inheritance, protein coding, and providing instruction for all living organisms. The process that DNAs carrying genetic instructions are read into ribonucleic acid (RNA) is transcription, and RNAs to proteins is translation. This information transfer is directed, meaning it's usually not reversible (Crick, 1970). In multicellular organisms, nearly every cell contains the same genome and the same genes (Collins et al, 2003; Ricroch et al, 2007). However, not every gene is transcriptionally active in every cell. Some cells show variation of gene expression patterns by controlling which genes are transcribed and which transcripts are translated. Underneath this variation, the spacious range of physical, biochemical, and developmental differences among numerous cells and tissues may play an important role, and they can also be key features of disease development and species-specific changes (Hanson et al, 2014). Unlike the genome, the transcriptome is more complex because four nucleotide (nt) codes of DNAs and mRNAs are translated into a much more complex code of 20 amino acids. This complex code system results in primary sequence polypeptides of varying lengths folded into one of a startlingly large number of possible conformations and chemical modifications to produce a final functional protein (Manzoni et al, 2018). This complexity may undoubtedly provide a key for understanding genetic information and pleiotropy (Carninci et al, 2005). These genetic information readouts are called transcripts, and a transcriptome is a collection of all the gene readouts present in a cell. The observation of this collection directs the quantification of how genetic variability may impact the transcriptome at the single cell level (Dey et al, 2015). With observation, it can measure the counts of transcript to determine the amount of the gene activity with control of genetic activity of molecular signals (Williams et al, 2007).

The major central question of biology is to understand how genotype influences phenotype, and the insight into the transcriptome is important to answer this challenging question because it offers major insight on gene structure, expression, and regulation in many organisms (Wang et al, 2019). Determining the pattern and timing of gene expression can be mostly accomplished by tracking the enzymatic activity of its protein product. Therefore, tracking the process of this protein, in which gene information is

expressed, is an important scientific topic (Mitsis et al, 2020). As for encoded information converted to its functional product by gene, the transcription factors (TFs) are primarily well known as major gene regulators. TFs are proteins that can have specific binding affinity to certain DNA sequences and regulate its gene expression (Mitsis, Thanasis, et al. 2020). A TF usually targets specific transcription factor binding sites (TFBSs) where the binding event of the TF with its specific DNA occurs. Consequently, recognizing the TF-TFBS pairs is a determining step in understanding the potential function of TFs and the regulatory network in an organism (Yu & Li, 2017). Due to different binding affinities of TFs, TF binding events can occur cell type specifically for regulators of biological processes and to its transcriptional regulatory regions, e.g. promoters, enhancers, to control the expression of target genes (Inukai et al, 2017). Regulation of transcription is the most common form of gene activity control, and the activity of TFs allows genes to be specifically regulated. Especially in eukaryotes, the binding of TF to the promoter sequence results in the formation of protein complexes involving several protein-DNA and protein-protein interactions (Veerla & Höglund, 2006). The TF-DNA indicate interaction variation are important mediators of phenotypic differences. Promoter sequences are typically located directly upstream or downstream of transcription starting site (TSS) of DNAs, where RNA polymerase and the necessary TF binds to initiate the transcription. Regarding transcription initiation to understand the transcriptional process, it is necessary to define transcript regions or at least TSS. The core promoter elements, which include sites where transcription initiation occurs, can also be bound by certain TFs (Lee & Young, 2013). Precise knowledge of TSSs is indispensable for understanding how transcription is regulated (Kapranov, 2009).

Large degree of gene expression is regulated by the specific binding of protein TF to cis -regulatory TFBS in gene promotor regions, and sequence motif can indicate sequence-specific binding sites for TF (D'haeseleer, 2006; Lis & Walther, 2016). Binding motif of TF has specific binding preferences to a specific set of DNA. Particular binding affinity is distinct by the binding motif. Motif can cause multiple TSSs to converge at one site and enhance transcription (Yukawa et al, 2011). TSS as the likely source of the identified motif orientation preferences (Lis & Walther, 2016; Stormo, 2000; Aerts, 2012). Indeed, TFs can also attach to enhancer sequences positioned upstream or downstream from an associated gene, resulting in stimulation or enhancement of transcription of related genes. Moreover, profiling TF’s direct binding preferences provides an additional layer to understand one of the fundamental mechanisms of gene regulation and a comprehensive course of gene-regulation evolution including disease development (Lee & Young, 2013). Misregulation of gene expression programs can cause a broad range of diseases, for example bipolar (BD) is determined by gene environment interactions

(D'Addario et al, 2018) and changes in the regulation of gene expression at multiple levels is involved in schizophrenia (SZ) (Wood, 2019). Furthermore, as the TFs are conserved over larger evolutionary distances, identifying binding site accessibility will be an initial and indispensable step to picture the direction of evolution (Hobert, 2008; Chen & Rajewsky, 2007). Around 1,600 TFs have been documented for humans (Lambert et al, 2018) and it is estimated that there are 2,000 - 3,000 (Rodriguez-Caso et al, 2005). Increase in the number of TFs is associated with higher control of gene regulation (Swift & Coruzzi, 2017). Both theoretical arguments and practical observation indicate that TFs work together to achieve needed specificity in both DNA binding and effector function. For identifying true TF- TF interaction, the weighted topological overlap (wTO) method can be used to discover the overlaps among classes of transcript (Gysi et al, 2018).

## 1.2. wTO R package

Complex network analysis methods, especially co-expression networks, provide important and insightful new knowledge of the functioning and interactions of genes. A co-expression network identifies which genes tend to show a similar expression pattern across samples, representing a gene–gene similarity matrix. The term “co-expression” refers to a similarity of gene expression patterns across a variety of experimental conditions, and the “network” to an illustration of regulatory relationship between genes (Aoki et al, 2007). The network is an undirected graph composed of nodes and links representing genes and mutual co-expression relationship. Topology means the patterns of node-to-node connectivity, or configuration of links.

In the first step of the co-expression network construction, individual connections between genes are defined based on correlation measures or mutual information between each pair of genes. These relationships describe the similarity and distance between expression patterns of the gene pair across all the samples (Steuer et al, 2002). In the second step, a network graph is created based on the co-expression interactions calculated in the previous step, and this is composed of two major components - nodes and edges. Nodes represent co-expressed genes, and edges the presence and the strength of the co-expression relationship (Albert & Barabási, 2002). In the third step, modules, meaning groups of co-expressed genes, are identified using one of several available clustering techniques. Clustering in co-expression analyses is used to group genes with similar expression patterns across multiple samples. The clustering method needs to be chosen with consideration of evaluation based on internal and external criteria because it can greatly influence the outcome and meaning

of the analysis (D'haeseleer, 2005). In the end, modules can subsequently be interpreted by the functional procedure of wTO, a method to identify and rank overrepresented functional interaction patterns among a list of genes. The analysis using this R package is especially beneficial for establishing network architecture for system-level cellular organization (Ravasz et al, 2002). In an inspection of wTO-analysis, the new linkage- weight for a pair of linked nodes is decided through a normalized process that accounts for all common network substracts. Moreover, the wTO method can be drawn on to discover the overlap among neighbors of transcripts, e.g. TFs. Consequently, wTO network presents an additional solid representation of the connection and association between the node-sets of interest than frank correlation network analysis (Nowick et al, 2009).

The wTO package maintained by Gysi (Gysi et al, 2018; wTO package CRAN) scrutinizes and discusses the necessity of distinction, which allows for both positive or negative correlations. Using only absolute correlation would falsify the biological insight in symbiotic or predator-prey relationships (Gysi et al, 2018). Particularly, calculations for networks in wTO package are using the wTO calculation, which is established by Nowick and collaborators (Nowick et al, 2009). This measurement informs the user whether the interested network is activating or inhibiting/repressing. For example, in gene regulatory networks or in metabolic networks, the increase of a substance can have consequences to an increase or decrease of other physiological and biochemical properties (Gysi et al, 2018).

By constructing co-expression networks, following packages are habitually used : WGCNA (Langfelder & Horvath, 2008) and ARACNE (Margolin et al, 2006). The prior package provides functions to find modules of highly correlated genes, resulting in an unsigned network, and also network modules will be calculated to measure module membership (Tab. 1, second row: signed topological overlap). The latter offers to reconstruct transcriptional regulatory networks using mammalian cellular networks. Distinctly to these common packages, the wTO package offers several major advantages. Foremost, it avoids undesirable network results, which can be caused by technical preprocess within data or biological differences from confounding factors. Most common biological dissimilarity are resulted from sex, age, and geographical origin of the individual. The package supplies signed and unsigned wTO network as well as the consensus network (CN) to elude unwanted results (Tab. 1, third row: consensus topological overlap ). CN is a fundamental structural property for differential network analysis, which captures systematic differences between two or more networks in an integrated network. It is also used to identify changes in connectivity patterns or consistency of module structure between different conditions (Langfelder & Horvath,

2008). By WGCNA, consensus wTO (in their nomenclature, the “topological overlap matrix”) is assigned to define each edge of the consensus network, which will discard some gene pairs in networks because of its strict unanimity with overlaps. This loss of such information is not only unwanted, but possible to lead to biased results. Further, the package has included the option to determine p-values for each pairwise wTO value with the choice of using Pearson correlation. The p-value is established with the probability of the bootstrapping model (Tab. 1, fourth row: pairwise p-values ), the null hypothesis is set true, when the statistical summary would be equal to or more extreme than the actual observed result (Wasserstein & Lazar, 2016). The computation of p-values for each link is based on its empirical distribution, allowing to reduce false positive links in wTO network. Especially, by identifying TF-TF interactions with the transcriptional regulation database of *Escherichia coli*, the wTO package shows significantly better demonstration in qualitative and quantitative analysis with identifying true positive and false negative than WGCNA, ARACNE (Gysi et al, 2018). Here true positive interaction is defined, when the result matches from the transcriptional regulation database and true negative is described when these interactions are not validated in the dataset. Normally, the interest of running wTO lies on a subset of nodes. The package provides the ability to visualize the topology network and calculation of wTO value with reasonable computational time. This node-and-edge type view of co-expression networks enables distinct comparison (Tab. 1, fifth row: network view), showing important structural differences between interactive networks to users (Xulvi-Brunet & Li, 2009).

**Table 1. Available features of co-expression network constructing tools**

Each column represents packages for constructing co-expression networks. The first row shows the presence of topological overlap function in each package, and except ARACNE the other two packages provide this function, which is important for quantifying co-expression at gene-pair level. The second row shows the capability of computing signed and unsigned wTO networks. Only the wTO package offers signed output which is helpful for the construction of a CN and visualization. The third row shows the difference of the calculation of a consensus network, which WGCNA uses a strict version of consensus, in that it will harshly discard any weak gene pair. The fourth row shows the existence of bootstrap calculation, which is also only provided in the wTO package, and helps to avoid multiple comparisons problems. The fifth row shows the possibility of a node-and-edge type view of a co-expression network. The wTO package has internal functions to visualize, while other WGCNA and ARACNE must be exported to Cytoscape for network view. The sixth row shows the presence of a thresholding procedure, which assists choosing parameters in an unweighted network. The wTO package is not the case as it retains the score by definition. The seventh row shows calculation options on link weights. The eighth row shows the distinct allowance of calculation of networks from time series data without replicates.

| Method | wTO | WGCNA | ARACNE |
|---|---|---|---|
| Topological overlap | Yes | Yes | No |
| Signed topological overlap | Optional | No | No |
| Consensus topological overlap | Weighted sum | Minimum weight | No |
| Pairwise p-values | Yes | No | Used to filter MI |
| Network view | Native | Exported to Cytoscape | Exported to Cytoscape |
| Soft thresholding | No | Optional | No |
| Correlation choices | Spearman, Pearson | Bicor, Pearson | Spearman, Pearson, Kendall |
| Capable with time series | Yes | No | No |

## 1.3. Motivation

Given that most biological functions cannot be attributed to a single gene, a module is likely to represent a set of genes having discrete functions that arises from interactions among them (Hartwell et al, 1999). Mutually, the analysis of local modules may be more informative with respect to the regulatory mechanisms of the specific biological processes. Gene co-expression networks have various purposes and usages, including candidate disease gene prioritization, prediction for association, functional gene annotation and the identification of regulatory genes. Nevertheless, co-expression networks are effectively able to identify correlations with network construction. If indicated genes are active simultaneously, they can often be suggested to be active in the same biological processes (Amar et al, 2013). Although, presence of a functional linkage can not be conferred about causality or distinguish between regulatory and regulated genes (van Dam et al, 2017). In co-expression analysis, similarity of gene expression profiles is measured using correlation coefficients. A co-expression network thus illustrates correlation patterns between genes, and so represents the complexity of

a cellular transcriptional network. One important finding is that a gene co-expression network has the universal topological features of complex network systems characterized by modularity (Jeong et al, 2000). Topological properties such as degree, network density (Barnes, 1969), betweenness and clustering coefficient (Watts & Strogatz, 1998) should be specifically identified to reach the primary goal of co-expression network analysis : identification of the modular structure (Aoki et al, 2007).

Yet, results of wTO is based on the similarity of co-expression patterns and the TFs with its network (Fig. 2). The patterns are indicated through correlation scores, which are generated from the co-existence of its shared gene ID (or gene names or gene symbols). The wTO package displays an intuitive way to represent complex co-expression patterns between inserted genes; it can meet challenges to explain the hierarchy of complex biological networks (Barabási & Oltvai, 2004). The outcome of function *wTO.Complete* and *wTO.Consensus* gives users a fairly promising description about the list of interactions between Node 1 and Node 2 with its link's weight and the p-value in *data.table* format. Clearly, the output can be filtered out with any value within raw p-value, adjusted p-value or weight by needs. Even the result can be strained through wished significance or weight, the correlation may prevail unwanted outcomes. Moreover, the drawback of two possible correlations are still underlying, resulting in false discovery. Pearson correlation is sensitive to outliers, which can overestimate or understate the robustness of the relationship. The data for spearman rank correlation is advised to be monotonically correlated, skewed or ordinal. To achieve and fulfill the primary goal of co-expression network analysis with wTO packages, hereby recommend two additional aspects : 1) comparison of motif similarity and 2) TFBS prediction within obtained results. These additional aspects can be applied directly to results of *wTO.complete* function to enrich the robustness of returned value and secure false positive rate.

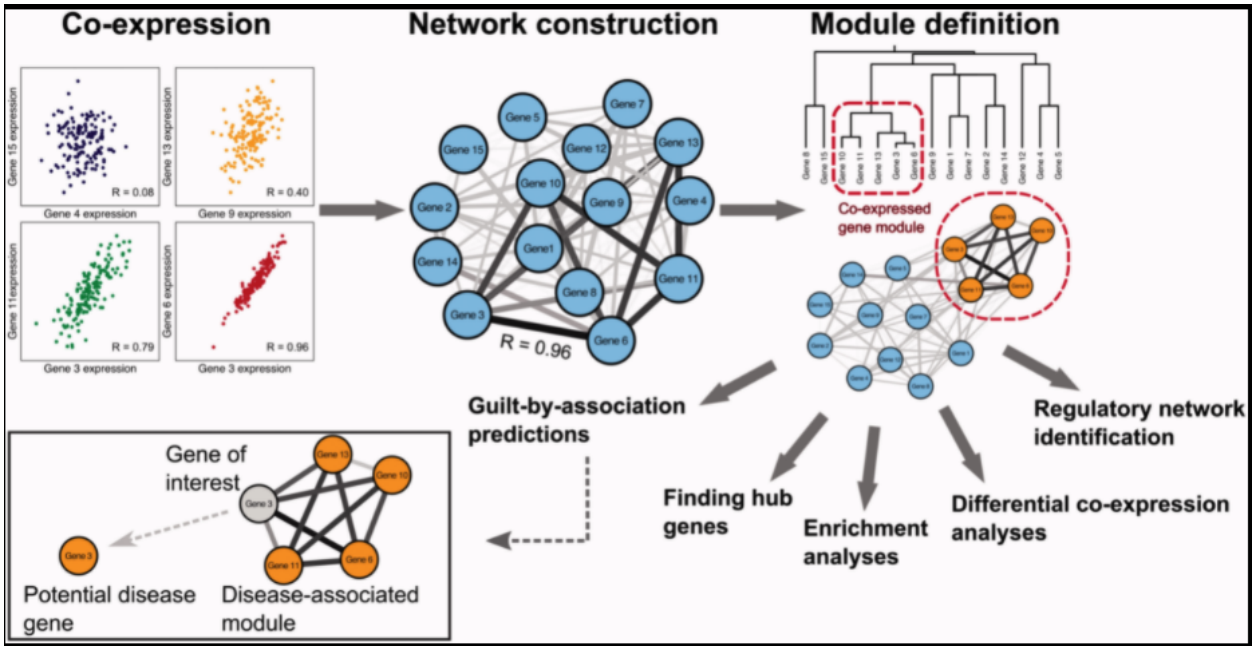


**Figure 1. Schematic example of a co-expression network analysis**
**Co-expression)** First, pairwise correlation is determined for each possible gene pair in the expression data. The R parameter represents the strength and direction of correlation between gene pairs. The coefficient R = 0.96, for instance, indicates a strong and positive linear relationship displayed under the blue network beneath the "network construction". These relationships describe the similarity between expression patterns of the gene pair across all the samples. **Network construction)** Then co-expression associations are used to construct a network, where each node represents a gene and each edge represents the relationship. **Module definition)** Modules within these networks are defined using clustering analysis to group genes with similar expression patterns. The network and modules can be interrogated to identify regulators, functional enrichment and hub genes. Differential co-expression analysis can be used to identify modules that behave differently under different conditions. Potential disease genes can be commonly identified using a guilt-by-association approach that highlights particular genes that are co-expressed with multiple disease genes. Modified from (Barabási & Oltvai, 2004).

## 1.4. Aim of Study

Many eukaryotic genes are co-regulated by multiple TFs in a cell type-specific manner (Maston et al, 2006). Among TFs, multiple molecules of the same TF can also occupy neighboring sites, which two scenarios may result in motifs: two TFs bind to neighboring sites (co-binding), or one TF protein binds to another that in turn binds to DNA (tethered binding) (Nature, 2019). Combinatorial regulation by multiple TFs can explain how a relatively small number of TFs can govern gene expression under diverse conditions (Odom et al, 2006; Zinzen et al, 2009). Therefore detecting motifs of TFBS will be one of the most widely studied flavors of the problem, both for its biological significance and for its bioinformatic hardness (Zambelli et al, 2012; Boeva et al, 2010; Mitra et al, 2018).

The aim of this study is :

1. To overview, whether the TFBS of TFs that are connected in wTO are often co-located within the promoters of the genes correlated with both TFs.
2. To discover the most cooperative and collaborative TFs from different conditions and locate their TFBS coordinate.
3. To explore the possibility of the detection of TFBS to enrich co-regulation calculation of wTO networks.
4. To test, whether predicted TFBSs and its motif sequence differ from each group.
5. To quantify common co-expressed genes by regulation of TFs

# 2. Materials

## 2.1. File formats

Bioinformatic methods utilize various file formats for different purposes. The distinct information that can be collected from each of the formats employed while preprocessing the data is outlined in the following section.

### 2.1.1. FASTA

FASTA is a very common text based file format to describe nucleotides or amino acid sequences of proteins. Each FASTA file starts with the “>” symbol indicating the head line, followed by an identifier of the sequence and a brief description of the sequence. The nt sequence is encoded in characters (A, C, G, T, and N ) and printed in the second line (Stormo et al,1982). Multiple sequences can be stored in one FASTA file.

### 2.1.2. CSV

Comma separated values (CSV) are essentially plaintext files that can easily be arranged into a spreadsheet-like array, Every texline in the CSV file represents a row, and columns are delimited in a line with the use of the comma “,” character. Comma form is prominent and irregular whitespaces are ignored.

### 2.1.3. TSV

Tab separated value (TSV) is a common method to exchange data among spreadsheets, databases, and word processors. Each and every record is illustrated as a single line. Every single field value is represented as a text.

### 2.1.4. PWM

Position Weight Matrix (PWM) provides a simple and consistent interface to the various motif types. It is originally induced for the Position Count Matrix (PCM), in which a simple basic motif is shown for each position the total counts of letters. The PCM then converted to Position Probability Matrix (PPM), which Equation (1.1) would be used to calculate the probability ($P$) of each letter $N$ at an individual position from counts ($C$).

$$P(N) = \frac{C_N}{\sum C} \text{ ( Equation 1.1)}$$

As the next step followed by PWM, every letter has a 'score' which can be used to evaluate how well a sequence matches a motif (Fig. 2). The scores (*S*) calculate to the log of each probability, correcting for background frequencies (*B*). This results in the following calculation:

$$P(N) = log_2 \frac{P(C_N)}{B_N} \text{ (Equation 1.2)}$$

To avoid -$Inf$ at PWM, pseudocount 1 can be used in both numerator and denominator (Tremblay, 2019).

A)

| # | Sequence |
|---|---|
| 1 | AAGAAT |
| 2 | ATCATA |
| 3 | AAGTAA |
| 4 | AACAAA |
| 5 | ATTAAA |
| 6 | AAGAAT |

B)

| Position | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 6 | 4 | 0 | 5 | 5 | 4 |
| C | 0 | 0 | 2 | 0 | 0 | 0 |
| G | 0 | 0 | 3 | 0 | 0 | 0 |
| T | 0 | 2 | 1 | 1 | 1 | 2 |

C)

| Position | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1.00 | 0.67 | 0.00 | 0.83 | 0.83 | 0.66 |
| C | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | 0.00 |
| G | 0.00 | 0.00 | 0.50 | 0.00 | 0.00 | 0.00 |
| T | 0.00 | 0.33 | 0.17 | 0.17 | 0.17 | 0.33 |

D)

| Position | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 2 | 1.425 | -Inf | 1.737 | 1.737 | 1.415 |
| C | -Inf | -Inf | 0.415 | -Inf | -Inf | -Inf |
| G | -Inf | -Inf | 1.000 | -Inf | -Inf | -Inf |
| T | -Inf | 0.415 | -0.585 | -0.585 | -0.595 | 0.415 |

**Figure 2. Representation of TFBS motifs**
**A)** List of six sequences as an example and can be basically represented in B). **B)** PCM. Simple counting of total occurrence for each position **C)** PPM. The probability for each letter at each position of B) and this can be represented like C) with equation 1.1. **D)** PWM. This shows every letter with a score (equation 1.2), which can be used to evaluate how well a sequence matches a motif.The log of fractions where the probability of a certain letter in a sequence is higher than that of the background probability of that letter result in positive scores, and vice versa for negative scores.

## 2.2. Tools and functions

To test hypotheses and to conduct analysis, diverse tools are employed. Most of the analysis was coded based on the statistical programming language R and implemented code was tested in the Unix operating system (Unix-OS) using the server, which is provided and run by Freie Universität Berlin (FUB). A brief explanation of each tool is shown in table 2 and 3. The extensive descriptions and their computation are explained in the method part more precisely. In case of installation of new software besides of pre-installed tools, the anaconda tool was used for some of the packages for R in the server that aims to set the package management and deployment. After creating a new environment for the new tool, the installation source of the software was searched in the anaconda cloud individually. Such a tool installed in an anaconda environment can only be accessible by activating a customized environment, and this property facilitates the package management.

**Table 2. List of software tools used in the Unix - OS environment**

| Tool | Version | Description | Reference |
|---|---|---|---|
| Anaconda Navigator - Individual edition (Unix - OS) | 2.1.1 | Package and environment management system for server | Anaconda Software Distribution 2021 |

The R version (Team, R. Core, 2013) used in this research is version 4.1.3. The R package *memes* should be exclusively downloaded locally and installed additionally with terminal commands. The package *magrittr* is optional to use in code, as it is a pipe operator which only aims to decrease development time and to improve readability and maintainability of the code. Excluding this package would not affect the result.

**Table 3. List of R packages used for analysis.**

| Tool | Version | Description< | Reference |
|---|---|---|---|
| GenomicFeatures | 1.46 | A package provide set of tools and methods for making and manipulating transcript centrix annotation | Lawrence et al, 2013 |
| TxDb.Hsapiens.UCSC.hg38.knownGene | 3.14 | A package containing a TxDb objects for annotation database generated from UCSC | Bioconductor Core Team, 2019 |
| BSgenome.Hsapiens.UCSC.hg38 | 1.4.4 | A package containing full genome sequences for Homo sapiens as provided by UCSC (hg38, based on GRCh38.p13) | Team TBD, 2021 |
| plyranges | 1.14.0 | A package providing a grammatical and consistent way of manipulating for Bioconductor classed and GenomicRanges | Lee et al, 2019 |
| biomaRt | 2.50.3 | A package containing generic and scalable systems and enables advanced querying of biological data sources - Uniprot, Ensemble, etc. | Durinck et al, 2009 |
| Repitools | 1.40.0 | A package providing tools for the analysis of enrichment based epigenomic data. | Statham et al, 2010 |
| MotifDb | 1.36.0 | A package containing more than 9900 annotated position frequency matrices from 14 public sources for multiple organisms. | Shannon & Richards, 2021 |
| stringr | 1.4.0 | A package providing simple and consistent wrappers for a common string operation | Wickham, 2021 |
| dplyr | 1.0.8 | A package providing grammars for data manipulation | Wickham et al, 2019 |
| universalmotif | 1.12.4 | A package allowing for importing common motif types from various classes | Tremblay, 2022 |
| memes | 1.2.5 | A package providing interfaces to the MEM suite family of tools for motif analysis on DNA, RNA, and protein sequence | Nystrom S, 2022 |

| | | | |
|---|---|---|---|
| magrittr (optional) | 2.0.2 | A package providing a mechanism for chaining commands with a new for- ward pipe operator | Bache et al, 2014 |

# 3. Methods

## 3.1. Input data

The main data was provided by Rebecca Saager, current member of the AG Nowick. The primary data consist of two dataset: 1) *wTO::wTO.fast* result which was generated from a signed and weighted co-expression network in three different groups Control (CT), bipolar (BD), and schizophrenia (SZ) with wTO method. (Tab. 4) 2) correlation degrees of common TFs in each group. The analysis originally was aimed to study to determine the role of long non-coding RNAs and gene regulatory factor (GRF) proteins in the cingulate cortex and the hippocampus for BD and SZ. GRFs are obtained through the expression patterns from respective cells or tissues of interest, which regulate the transcription of other genes either in an activating or repressing way (Jovanovic et al, 2021). Degree of common TFs was counted by how many genes sharing the same TF name.

**Table 4. Example of five rows representing actual result of *wTO::wTO.fast* from group control**
The first row shows the column names of the data. Depending on the combination of the first and second column, the value of wTO, pval and pval.adj correspond.

| Node.1 | Node.2 | wTO | pval | pval.adj |
|---|---|---|---|---|
| AL627309.5 | LINC01409 | 0.292 | 0.076 | 0.34665784 |
| AL627309.5 | FAM87B | 0.025 | 0.243 | 0.35701545 |
| LINC01409 | FAM87B | 0.210 | 0.304 | 0.38227119 |
| AL627309.5 | LINC01128 | 0.466 | 0.216 | 0.35040828 |
| LINC01409 | LINC01128 | 0.490 | 0.137 | 0.34665784 |

The calculation of *wTO::wTO.fast* is formed on the correlation coefficients of the total gene set from the input data (Gysi et al, 2018). The Pearson correlation coefficient (“wTO” column in Tab. 4) (*r)* is calculated to give correlation type and to determine the linear relationship between two genes from insert (Node 1, Node 2) (Equation 3.1) (Heumann et al, 2016).

$$\gamma = \frac{\sum_{i=1..N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1..n}(x_i-\bar{x})^2 \sum_{i=1..n}(y_i-\bar{y})^2}}$$ (Equation 3.1)

The calculated correlation coefficient formulates the adjacency matrix A = [$a_{i,j}$] with :

$$a_{i,j} = (if\ i \neq j,\ then\ \gamma_{i,j} \vee if\ i = j,\ then\ 0)$$ (Equation 3.2)

For every pair of Node.1 and Node.2, *wTO::wto.fast* also computes the wTO score $w_{i,j}$ based on the correlation values in the adjacency matrix.

$$w_{i,j} = \frac{\sum_{u=1..N} a_{i,u}a_{u,j} + a_{i,j}}{min(k_i,k_j) + 1 - |a_{i,j}|}$$ (Equation 3.3)

and

$$k_i = \sum_{j=1..N} |a_{i,j}|$$ (Equation 3.4)

Given equation explicitly includes the correlations of both genes in the pair (*i , j*) with all other genes in the gene set (*u*), regardless of whether or not they belong to the genes of interest (Nowick et al, 2009). The p-value for the actual wTO score is enumerated by dividing the number of deviant scores by the number of bootstraps. All of the p-values were adjusted to enhance reliability and reproducibility of analysis findings from multiple testing using the Benjamini-Hochberg procedure (Jafari, Mohieddin, & Ansari-Pour, 2019). Also these adjusted p-values were ordered and each p-value (*p*) was multiplied with the number of total p-values (*m*) and divided by the respective order of the list (*k*) (Benjamini & Hochberg, 1995) (Equation 3.5). In analysis, only the links with an adjusted p-value of < 1 x $10^{-2}$ were considered as significant.

$$pval.adj = p * \frac{m}{k}$$ (Equation 3.5)

## 3.2. Motif databases and HOCOMOCO

Homo sapiens comprehensive model collection (HOCOMOCO) is the collection that consists of patterns describing DNA binding specificities for humans. The database is gathered through systematic motif discovery from more than five thousand ChIP-Seq experiments using uniformly processed within the cross-validation pipeline. Cross validation is a re-sampling method that uses different portions of the data to test and train a model on different iterations such as JASPAR, TRANSFAC, SwissRegulon and others (Anthony, 2014; Wingender, 2008; Rapicavoli et al, 2013; Stormo et al, 2014). Mainly, it is utilized with several Chromatin Immunoprecipitation followed by next generation sequencing (ChIP-seq) peak calling tools and aggregated in the Gene Transcription Regulation Database (GTRD) (Kulakovskiy et al, 2017). ChIP-seq is a technique widely used to identify genomic binding sites for epigenetic regulators, including TFs and DNA/RNA binding proteins (Robinson, 2012). ChIP-seq enables the discovery of the interaction between protein complexes and DNA regulatory elements, and their gene regulatory networks with multiple TFBSs (Valouev et al, 2008). HOCOMOCO also provides PWM models for binding sites of 680 human TFs and includes 1,302 mononucleotide and 576 dinucleotide position weight matrices, which describe primary binding preferences of each transcription factor and reliable alternative binding specificities. Compared to other known motif collections, HOCOMOCO offers several advantages to other known motif datasets (Kulakovskiy et al, 2017). First, it covers a comparably large systematically derived collection of TFBS models for humans in overview of the collections (Tab. 5). Furthermore, in terms of evaluation performance between other collections, JASPAR and HOCOMOCO were best scored in numerical estimation of binding site recognition performance for 145 human mononucleotide PWMs and for all dinucleotide PWMs. Also, the database supports the observation of TF-specific performance of scoring functions, algorithms and DBs (Kibet & Machanick, 2015). Finally, an interactive interface and bulk downloads are accessible in the browser, which allow users to download on the following url : *https://hocomoco11.autosome.org/*. For this thesis, HOCOMOCO v11 is used.

**Table 5. The number of human Transcription factors covered by TFBS models of known motif collections.**

| Collection | Number of human TFs covered by TFBS models |
|---|---|
| HOMER | 123 |
| JASPAR | 130 |
| SWISSREGULON | 337 |
| HT-SELEX | 404 |
| HOCOMOCO V10 | 601 |
| HOCOMOCO V11 | 680 |

## 3.3. Tidy sequence

In the three different groups, control, bipolar and schizophrenia, the result of *wTO::wTO.fast* was overviewed and filtered by adjusted p-value of $< 1 \times 10^{-2}$ and absolute wTO score of $\geqq 5 \times 10^{-1}$ using *dplyr::filter* and merged through *base::unique*. This was considered as the significant result (Gysi et al, 2018). wTO score closer to |1| indicate a strong relationship, while closer to 0 indicate weak to no relationship. Then, the next step is extracting corresponding promoter sequences to each gene with upstream and downstream of interest. To retrieve the sequence, a local function is written taking input with three variables; character list of gene names, numeric value of upstream, and of downstream. The optimal promoter search space for potential TF binding sites should be ±5 kilo base pair (Kbp) from the TSS of human genes, and this promoter size is used to pre-compute the mechanistic TF regulatory network. The upstream and downstream for promoters beyond +5 Kbp can cause significant decrease of the sensitivity and specificity with p-value $1.5 \times 10^{-2}$ (Plaisier et al, 2016). Therefore, the default value for upstream is set to 5 Kbp and for 3 Kbp downstream from TFBS to map TFBS experimentally and to detect mostly, which also includes the localization in noncoding exons (Levy & Hannenhalli, 2002). The function filters the given character list of gene names as symbols from a committee of the Human Genome Organization (HUGO) (Tweedie et al, 2020). Then, the Ensembl human gene annotation data (Rigden & Fernández, 2020) was selected and called by *biomaRt::useDataset* from

the Ensembl BioMart server. The Ensemble data and HUGO symbol were then used to build the query for *biomaRt::getBM*, which will return the vector of attributes in the official Ensembl transcript identifier (ID). Each transcript ID includes a unique, stable 11 digit number. Transcripts beginning with ENST are human transcripts, for example ENST00000303660 for ZEB2. Transcript ID is genomic variant or splice variant, so-called isoform, of a corresponding gene with ENSG ID. Next, the function uses *GenomicFeatures::transcriptsBy* to extract the genomic feature by gene type from *TxDb.Hsapiens.UCSC.hg38.knownGene*, which is an annotation database generated from UCSC by a SQlite database and this manages genomic locations and the relationship between pre-processed mRNA transcript, exons, protein coding sequences, and their related gene identifiers (Bioconductor Core Team, 2019; Kent, 2002). The function *GenomicFeatures::transcriptsBy* returns a *GRangesList* object, which is the form of multiple *GRange* objects with hierarchical structure. *GRange* provides core and fundamental infrastructure for the integrative statistical analysis of range-based genomic data, which offers a diverse ecosystem with various packages such as quality assessment, sequence analysis, variant calling and other tasks. After *GRangesList* was produced, *plyrange::as_grange* constructed a *Granges* object to represent genomic intervals, which have sequence name, strand and additional attribute of human transcript ID. Yet, human transcript ID contains an internal version value in the end of each official Ensembl name with a bullet point symbol, which does not guarantee to be unique or even defined. The *plyrange::mutate* helps to manipulate all of the transcript ID in the *Grange* object. The *base::gsub* is used to modify and to replace version information after a match of a string. Afterwards, the verb *plyrange::filter* restricts the rows from *GRange* object to keep only unique strings of Ensembl transcript ID from official Ensembl transcript ID of interest. Subsequently, the dependency package *GenomeInfoDb* of *GenomeFeatures* subsets of the standard chromosome. The function *GenomeInfoDb::keepStandardChormosomes* internally uses the predefined tables inside of the package to find correct *seqlevels* according to the sequence style of the object. It retains exclusively on *Homo sapiens* chromosomes and drops any other species with coase pruning mode to preserve exact ranges and the order inside the individual list elements of *GRange* in each group.

## 3.4. Promoter annotation

Now, the genomic interval needed to be modified by altering the width of the interval of interest while leaving the start, end or midpoint of the coordinates unchanged. This was achieved with the *plyrange::stretch* verb along with *anchor_** adverbs. Utmost *plyrange::anchor_start* fixed the start coordinate of the *GRange* object and *plyrange::stretch* extended the interval in direction by an integer amount of upstream, here 5 Kbp as default is input. For downstream, *plyrange::anchor_end* preserved the end coordinate and widened the anchor for 3 Kbp. Finally, the *meme::get_seqeunce* extracted a set of sequences from *BSgenome.Hsapiens.UCSC.hg38*, a sequence container*,* which stores full genome sequences for humans as provided by USCS (Arora et al, 2022; Lee et al, 2022). This reference assembly is based on Genome Reference Consortium Human Build 38 patch release 13, which is a composite genome, derived from multiple genomic clone libraries and is intended to provide representation for the pan-human genome (Schneider et al, 2017). Overall, *Biostrings::DNAStringSet* object was returned from a manual function with names corresponding to genomic coordinates. Notably, the *Biostrings* package was not used in universal function, however only as a container for storing a set of sequences for its representation and efficiency (Arora et al, 2022).

## 3.5. Motif discovery by using FIMO

Find individual motif occurrences (FIMO) (Grant et al, 2011; Bailey et al, 2015) is an efficient, statistically rigorous motif discovery software for any motif-based sequence analysis, which is supported and maintained by MEME Suite and National Institutes of Health (Jayaram et al, 2016). To run FIMO *meme::runFimo* in R, the user is required to provide two inputs here: a fasta-format sequence with optional genomic coordinate headers, and a set of motifs to detect within the input sequences. Basically, a collection of the DNA or protein sequence input for *meme::runFimo* function can be as a path to a *.fasta* formatted file, or as a *Biostrings::XStringSet* object*,* which was provided and controlled from the promoter annotation part. The function will parse genomic coordinates from sequence entries from data's header. The function *meme::runFimo* takes as input one or more fixed-length motifs, which are represented as position-specific scoring matrices (PSSM). The motif can be as a path to a *.meme* formatted file, generated from the MEME motif discovery algorithms, extracted from an

existing motif database *MotifDb::MotifDb* or using a simple own created format with *universalmotif::create_motif*. *MotifDb* collects approximately 2328 PFM from diverse public sources of humans, with ample accompanying metadata (Wingender, Edgar et al. 2015). To discover PWM of interest *MotifDb::query*, the names of the matrices should be the same as row names of the metadata *DataFrame*, and are chosen to balance the needs of concision and full description, including the organism in which the motif was discovered, the data source, and the name of the motif in the data source from which it was obtained. Each input motif was then converted into a log-odds PSSM *universalmotif::converts_motif* and used each PSSM to independently scan each input sequence. It reports all positions in each sequence that match a motif with a statistically significant log-odds score (Tab. 6) (Boeva, 2016). The program stores a collection of the DNA or protein sequences for occurrences of one or more similar motifs in a database of known motifs and computes a log-likelihood ratio score for each motif with respect to each sequence position and converts these scores to p-value using dynamic programming method (Staden, 1994), which assumes a zero-order null model in which sequences are generated at random with user-specified per-letter background frequencies. Furthermore, FIMO employs a bootstrap method (Storey, 2003) to estimate false discovery rates (FDRs), q-value. The q-value is not monotonic relative to p-value. It is generally recommended to use a q-value threshold rather than a *p*-value threshold. However, to compute *q*-values, FIMO has to hold all the matches in memory. When scanning a full genome, this can easily require tens of gigabytes of memory, which was not available by FUB server. By default, FIMO uses a p-value threshold of 0.0001, and scans both DNA strands, meaning it will expect about one "match" every five promoters simply by chance, which might be a tolerable false-positive rate (20 %), depending on what user intend to do next. The formula is:

$$\frac{\#\,false\,positive}{Promoter} = (2 * Promoter\ length) * p\ value\ threshold \quad \text{( Equation 3.6)}$$

If the given promoters are longer than 1,000 bp, it would need to decrease the p-value threshold to compensate. For example, for 8,000 bp promoters would need to divide the p-value threshold by eight, giving a threshold of $1.25 \times 10^{-5}$. The problem can be, for some transcription factor motifs, the best possible match to the motif is not significant at this level (Noble, 2009). The returns of FIMO were then transformed to dataframe *Repitools::annoGR2DF* for easier access and modification. In the end, all of the *dataframes* from each motif and promoter sequences were merged to one *dataframe* by using *dplyr::bind_rows* and allocated with a unique ID, combination of matched sequence and motif name *stringr::str_c*.

**Table 6. The present nomenclature for PWM/motif**

This table summarizes the present nomenclature. This is formulated to deal with incomplete specification of bases in nucleic acid sequences. In cases where two or more bases are permitted at a particular position the nomenclature permits the allocation of a single-letter symbol. The nomenclature may also be applied where uncertainty exists as to extent and/or identity.

| Symbol | Meaning | Origin of designation |
|---|---|---|
| G | G | Guanine |
| A | A | Adenine |
| T | T | Thymine |
| C | C | Cytosine |
| R | G or A | puRine |
| Y | T or C | pYrimidine |
| M | A or C | aMino |
| K | G or T | Keto |
| S | G or C | Strong interaction(3 H bonds) |
| W | A or T | Weak interaction (2 H bonds) |
| H | A or C or T | not G,H follows G in the alphabet |
| B | G or T or C | not A, B follows A |
| V | G or C or A | not-T (not-U), V follows U |
| D | G or T or A | not-C, D follows C |
| N | G or A or T or C | aNy |

# 4. Results

## 4.1. Promoter annotation

The promoter annotation by using multiple packages was based on the result of *wTO::wTO.fast* from each group (Tab. 7, columns: CT/BD/SZ). This was the very first step for input data preparation for the downstream analysis using FIMO. First, only the correlated genes after a filtering the adjusted p-value and γ were used to collect the most significant information based on wTO results. (Equation 3.1, Equation 3.5) (in code $adj.p\,value < 0.01$ and $|wTO| \geq 0.5$). Even though the initial data from wTO had identical counts in Node. 1, Node. 2, and row counts, after filtration mentioned above, the count of row and the combination have been altered throughout. For instance, original raw data consisted of 53,659,453 rows from a combination of 10,263 genes in each group, however around at most < 0.01 % of a gene pair (< 787,056 / 52,659,453 by CT) and utmost 35 % of genes (< 3,566 / 10,263 by BD) were cut out referring to the significance and the strength of correlation score (Tab. 7, first & second & third & fourth row: number of rows of CT & unique number of each nodes of CT & number of rows - after filtration of BD & unique number of each nodes - after filtration of BD). Then, the genes with HGNC symbol were enquired to check if they're already known and available on the database. Since the promoter sequence of each gene can be retrieved exclusively with corresponding transcript ID from UCSC annotation database, the number of obtainable genes were different in each group (Tab. 7, fifth row: number of transcript names). Notably, a single unique Ensembl gene ID contains multiple spliced transcript names - for example 3,566 of BD genes showed 14,121 transcripts ID, meaning particular transcripts of a gene can differ by its location of transcription start and end sites. Furthermore, some of the chromosome annotations were from fixed patch, which were accessioned scaffold sequences that represent assembly updates and represent changes to assembly sequence. As this update might affect the interpretation of the result and is not a preliminary interest of the thesis, all of the scaffold sequences were excluded. Accordingly, the count of retrieved promoter sequences of genes on each group eventually varied.

**Table 7. Result of promoter annotation**
The first row shows the total number of rows on raw wTO results. The count of rows is identical in each group. The second row shows the count of the set of nodes. Each group contained 52,659,453 combinations from 10,263 genes. The third row shows the remaining count of row, after filtering with $adj.p\,value$ and $wTO$. The fourth row shows the remaining gene count after filtration. Each set of genes produced the combination of the third row. The fifth row shows the total count of genomic variants, which was induced by the unique number of each node after filtering. The sixth row shows the total counts of promoter sequence after pruning unordinary chromosomes.

| | CT | BD | SZ |
|---|---|---|---|
| Number of rows | 52,659,453 | 52,659,453 | 52,659,453 |
| Unique number of each nodes (count of unique gene list from whole nodes) | Node.1 - 10,262<br>Node.2 - 10,262<br><br>(10,263) | Node.1 - 10,262<br>Node.2 - 10,262<br><br>(10,263) | Node.1 - 10,262<br>Node.2 - 10,262<br><br>(10,263) |
| Number of rows - after filtration | 787,056 | 586,466 | 452,508 |
| Unique number of each nodes - after filtration (count of unique gene list from whole nodes) | Node.1 - 2,864<br>Node.2 - 2,886<br>(2,999) | Node.1 - 3,379<br>Node.2 - 3,362<br>(3,566) | Node.1 - 2,347<br>Node.2 - 2,302<br>(2,486) |
| Number of transcript names | 13,208 | 14,121 | 12,191 |
| Number of transcript names with standard chromosome | 12,920 | 13,759 | 11,964 |

## 4.2. Transcription factors from hub

The TFs of interest from the wTO data result were in total 102 (Tab. 8, first row: counts of TFs of interest. CT, n= 32/ BD, n= 37/ SZ, n= 33). With a given TFs of interest, TFs could be represented in various names, and correspondingly the amount of search terms for FIMO varied too (Tab. 8, second row & third row: counts of TF of interest in

HGNC symbol & count of TFs of interest in uniprot swissprot). For *MotifDb::query*, the official external ID was used and for promoter annotation, transcript ID was used due to the function *GenomicFeatures::transcriptsBy*. Yet, only a few TFs were identified by the motif annotated database from UCSC for each group. One explanation for the loss of information from input to outcome of FIMO is that the motif for TF of interest has not so far been defined in the metadata of *MotifDb*.

**Table 8. Count of TFs changes throughout databases and its unique ID**
The TFs from each of the hubs were only partially discovered from FIMO. Only < 10 % of TFs of interest were able to be used for further analysis. Rows of FIMO result shows that the number of predicted TFBS on each group were divergent. The seventh row shows spliced variants from each TFs, which were uncovered from FIMO. The eight rows represent the counts of genomic regions, which searched TFs of interest targets.

| | CT | BD | SZ |
|---|---|---|---|
| Count of TFs of interest | 32 | 37 | 33 |
| Count of TF of interest in HGNC symbol | 32 | 34 | 33 |
| Count of TFs of interest in uniprot swissprot | 33 | 34 | 34 |
| Count of Transcripts ID | 183 | 177 | 259 |
| TF of interest searched in FIMO (loss of information) | 3 (9 %) | 2 (5 %) | 3 (9 %) |
| Rows of FIMO result | 315,496 | 170,584 | 477,651 |
| Transcript ID (FIMO) | 12,496 | 12,555 | 11,339 |
| Ensembl gene ID corresponding to Transcript ID (FIMO) | 1,881 | 1,925 | 1,689 |

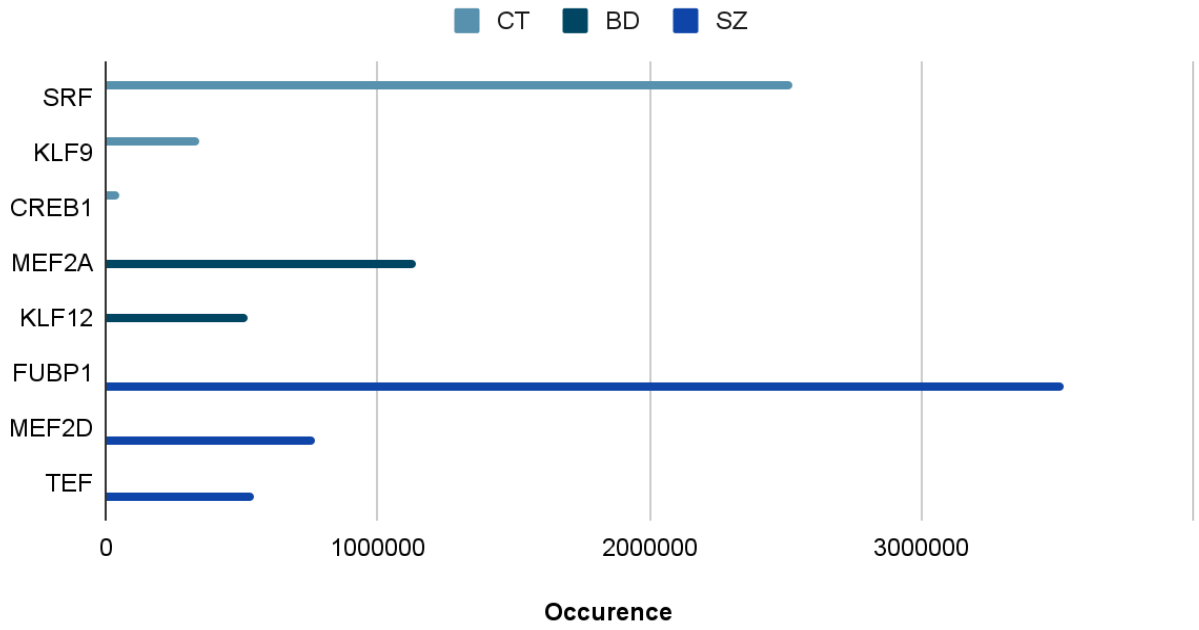


**Figure 3. Occurrence of motif from each groups**
The occurrences (x axis) of the TF binding motif displays dissimilarity of each group. Each group has its own abundance of TFs (y axis).

As a result of *meme::runFIMO*, the set of ranked sequences for occurrence of known motifs were returned. Each motif was utilized independently and consisted of the location of the chromosome, its start site and end site, width, strand, corresponding motif ID, log-likelihood ratio score, p-value, and matched sequence. Before comparison on groups and for simplification, the peak information from flank sequence was integrated to the FIMO result and it was transformed to *Grange* objects into *tibble* form. In search of TFs of interest in each group, most of the TFs were not found due to loss of information (Fig. 3). Only a handful of TFs were discovered from the database with corresponding motif ID (Fig. 3). Nevertheless, each TF showed a distinct motif sequence mapping, designated transcript ID and p-value. For instance, three TFs from CT group exhibit association with 12,496 transcript ID and 1,881 genes (Fig. 3, CT: SRF, KLF9, CREB1). Moreover, the frequency of TFs in the BD and SZ group can be compared comprehensively.

## 4.3. TFBS prediction

FIMO is by no means a general-purpose tool for identifying individual binding sites or protein motifs. Among several TFBS cluster prediction tools, FIMO is evaluated as one of the best performing (Jayaram, Narayan et al. 2016). However, compared to motif discovery of TFs of interest, the analysis with the whole known TFs required another level of computation power. It took approximately 131 hours with virtual memory of 4 gigabytes of memory at the FUB server. After feeding FIMO with promoter sequence and all known TF from humans, the data is collected thoroughly with distinct ID combinations. Figure 4 represents the distribution of TFBS and its overlap in the three groups in the venn diagram. Approximately 52 % (Fig. 4, left figure: 414/769) of all known TFs motifs were able to be searched on every group with individual matching sequences.

The average length of binding sites are typically 10 nt long in both eukaryotes and prokaryotes and empirically distributed from as few as 5 and maximum to 30 nt (D'haeseleer, 2006). Considerably, the binding sites were characterized by their information content, which is determined by balance between specificity and mutation. The number of different bases that can occur at each nt and still functional binding is important (Lis & Walther, 2016). In order to secure the magnitude of specificity from mutation, which can be increased by increasing binding site length, and to consist of a systematic data set, the range of observed motifs are filtered to a maximum 20 nt (Stewart et al, 2012). Figure. 5 shows that the distribution of binding length has been less kurtosis after excluding motifs longer than 20 bp. To profile the distinct patterns, all intersections between groups were excluded and the data have been saved separately (Tab. 10). Notably, the number of rows in CT and SZ is reduced individually on a scale to $1 \times 10^{-1}$ th and $1 \times 10^{-2}$ th of original fimo data. Next, then extracted the top ranked TFs with its occurrences and its matched sequence (Fig. 6). On average, TFs in BP showed more binding occurrences than any other groups. However, if the data was sorted with its motif id and corresponding sequence, the rank and average occurrences change. In CT, the TF TBX15 occurs multiple times with simple modification of nt. In BD, some of TFs e.g. SOX21, RARG, ZNF85 are presented multiple times with individual sequences. In SZ, the occurrences of TFs are more diverse and show less occurrences in general. The Sequence-level information can be used to visualize all sequences and their contribution to the final PWM, which can demonstrate the primary difference and help to understand the variability of motif matches (Fig. 7)

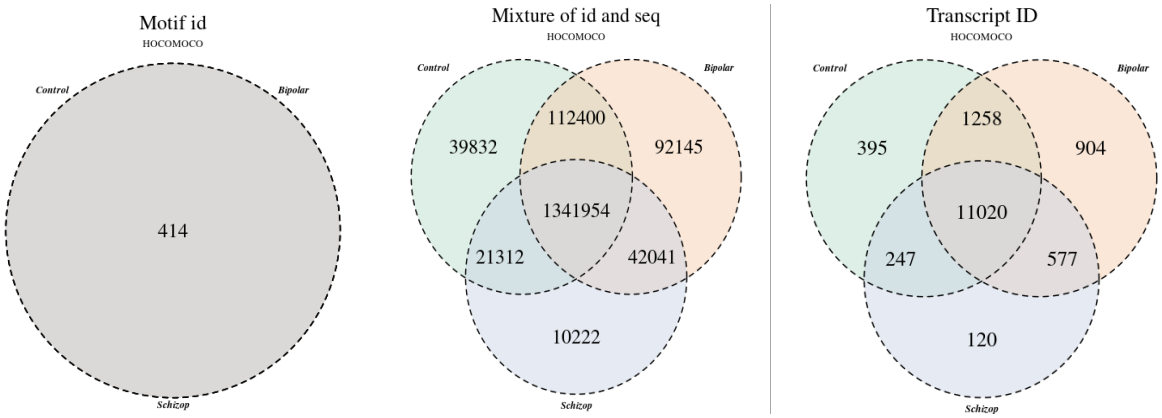


**Figure 4. The distribution of TFBS and its matched sequence in three groups**
The left figure shows that 414 TFs motifs were predicted to bind in all three groups commonly. To distinguish the distinctness of each motif and its sequence, the modif ID and its match sequences are prepared as the ID column of each row. Even though some of the motif co-occurs between groups, they are clear markers on each group (Middle venn diagram). On the right, it compares the transcript ID between groups. Most genomic variants overlap between groups.

**Table 10. Count of rows and motif changes throughout filter**
The first row shows the total number of rows, which implies binding occurrence of single TFBS. Group BD showed the most binding frequency than other groups. The second row shows the total count of motifs on each group. Remarkably all three groups had shown the same types of TFs, however different on motif sequences. The third row shows the count of rows, which primarily filtered the motif length longer than 20 bp. The fourth row shows the number of rows, after excluding all the intersections. The combination of motif id and its matched sequence is used to check the intersection. The promotion of loss of information is distinctive. The fifth row shows the total set of motifs, after excluding the intersected motif and its sequence. The sixth row shows the number of TFs, which are predicted commonly in every group.

| | CT | BD | SZ |
|---|---|---|---|
| Number of rows - raw | 42,697,524 | 44,974,329 | 39,445,102 |
| Number of Motif | 446 | 446 | 446 |
| Number of rows - filter motif length | 33,307,777 | 35,059,405 | 30,780,237 |
| Number of rows- after excluding all intersect | 23,922,062 | 3,359,096 | 354,425 |

| | | | |
|---|---|---|---|
| Number of Motif - after excluding all intersect | 218 | 239 | 192 |
| Number of motif in common | 192 | 192 | 192 |

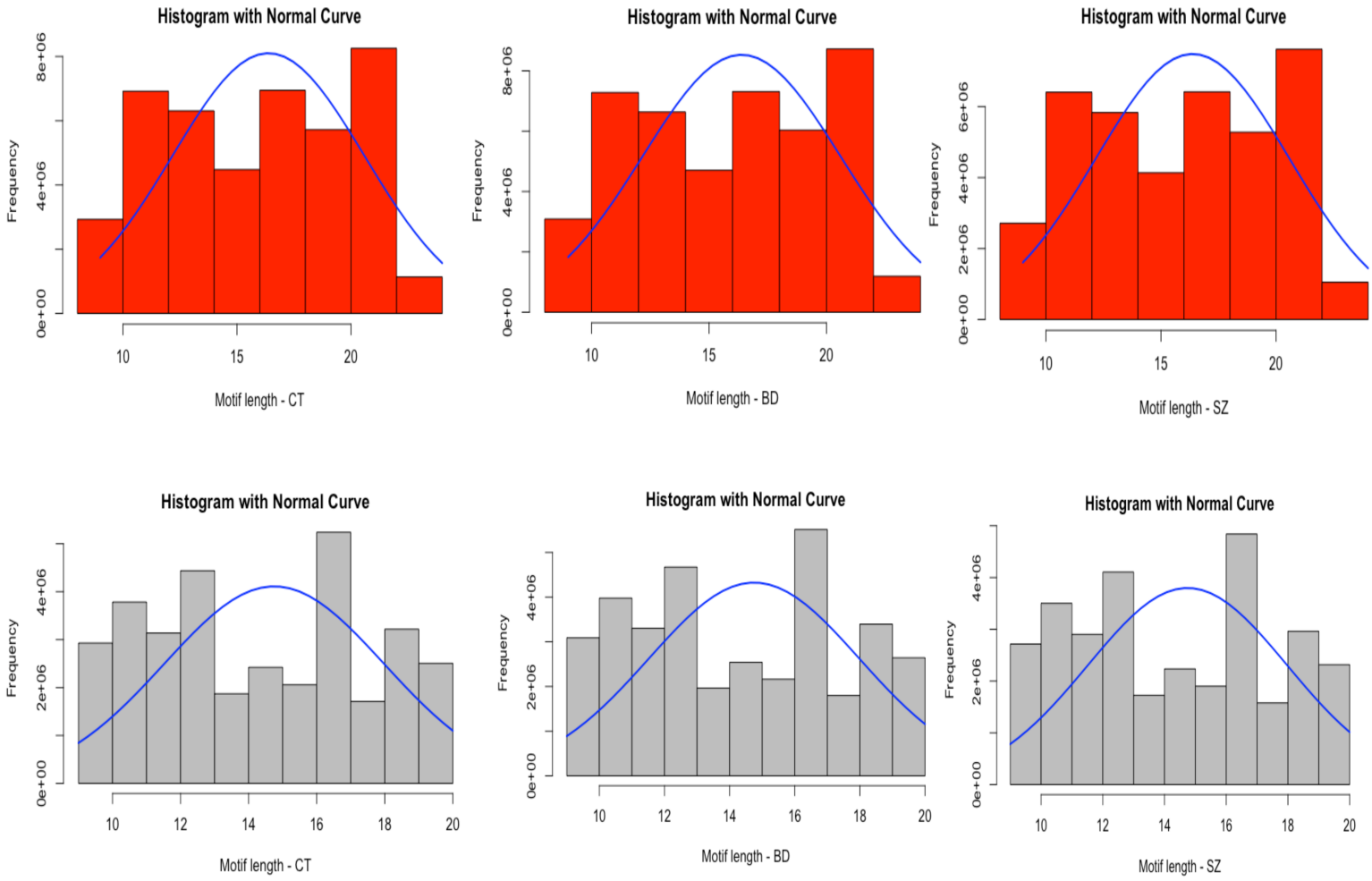


**Figure 5. Histogram and density plot of motif length occurrence on each group**
The original distribution of motif length on each group was scaled similarly from minimum 9 bp to 24 bp and average in 16 bp (red upper histograms). After excluding motif length above 20 bp, the empirical distribution of motif length shows flatter and more spread out (gray lower histograms). However the analysis failed to capture the motifs, which are shorter than 9 bp.

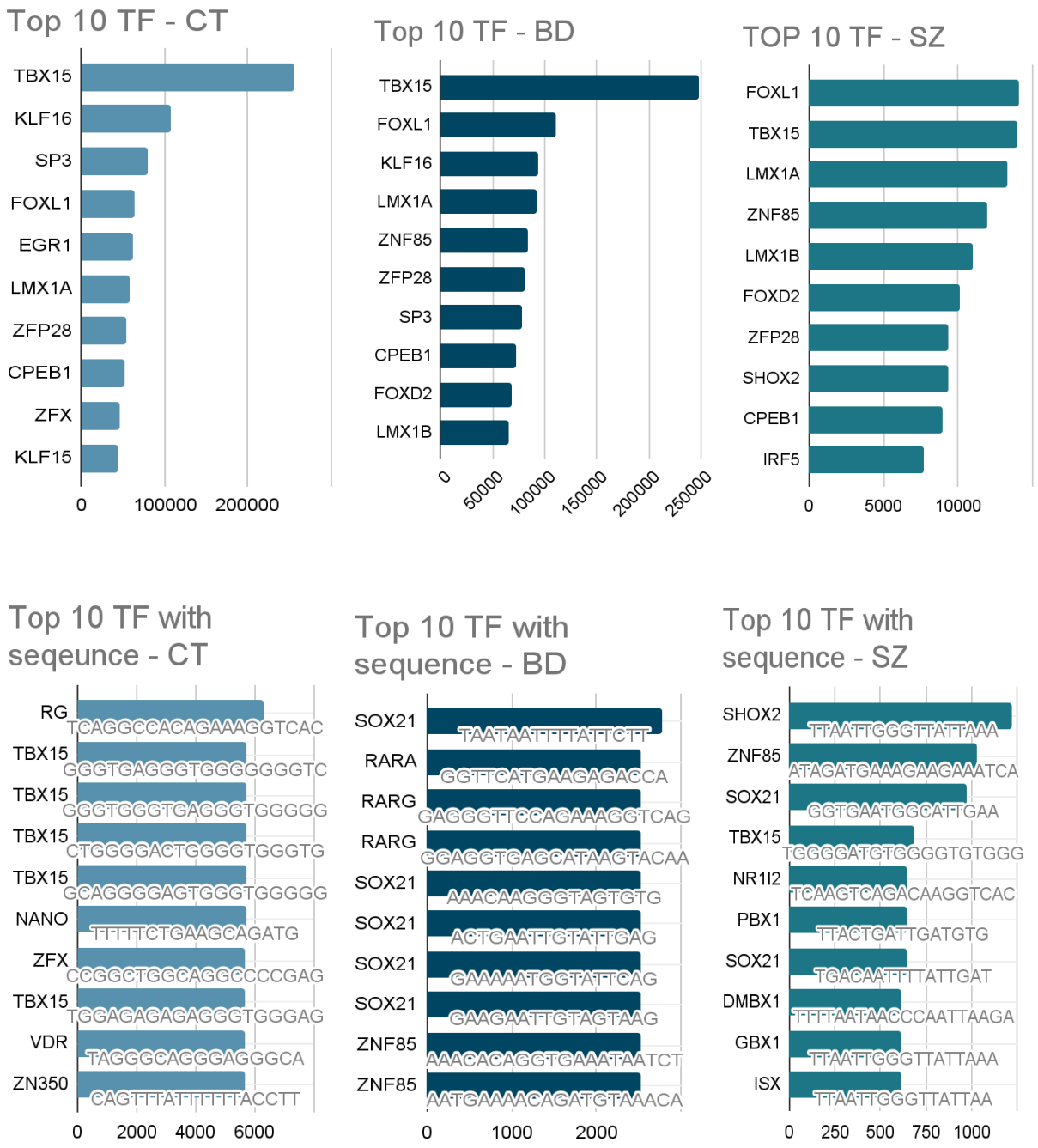


**Figure 6. Histogram of motif occurrences on each group**

Every group shows a different representative TFBS, which is simply sorted by frequency of binding sites. Each x axis represents binding occurrences and y axis names of TFs. The plots from the second row are ranked TF with its matched sequence.

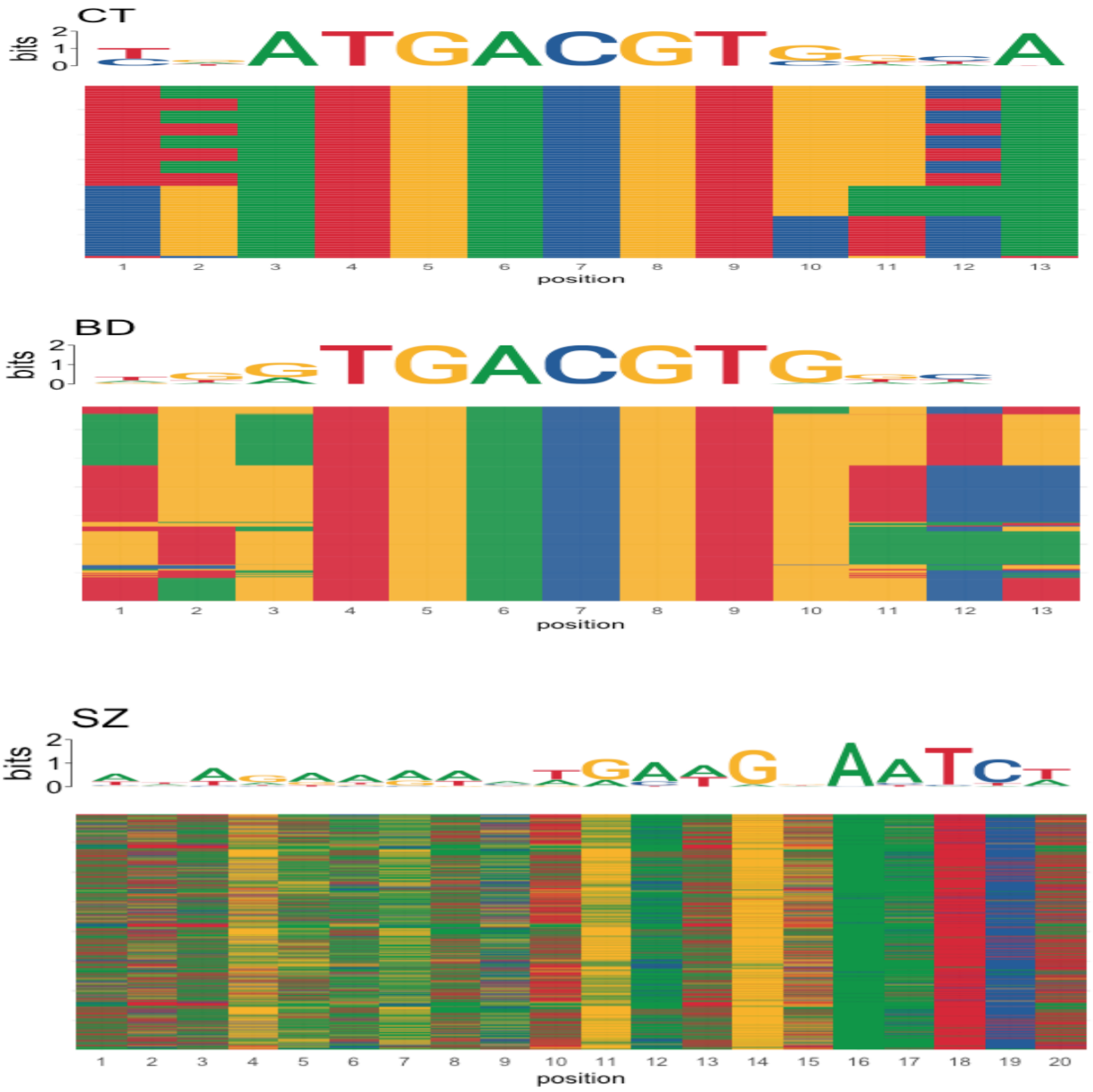


**Figure 7. Visualization of a heatmap of the aligned sequence of CREB3 in each group.** Motif match of TF CREB3 is visualized to observe how sequence variability differs from each group. Heatmap was created and generated from the PWM of each group. The Cells are colored by the sequence at the positions (x axis). The size of the logo represents the size of the information content in bits (y axis). In comparison to CT and BD, the sequence logo from SZ does not show distinct conservation on position.

## 4.4. Integration of FIMO and wTO

For understanding biological processes, global gene expression profiling is used with its similar expression patterns (Veerla & Höglund, 2006). The wTO method allowed users to visualize the interaction between TFs in a large network by not just displaying possible direct TF correlations, but by taking the overlap between TF-associated gene sets into account. These TF-associated genes used to construct the network are co-expressed and commonality in the associated gene sets for two TFs is represented to suppress false positives (Nowick et al, 2009).

With the result of FIMO, the identification of common and significant TFBS by co-expressed genes can be improved with experimental investigation. The existence of common target genes between TFs can capture the co-regulation possibility and can be quantified through frequency of occurrences. The measure of TFBS co-occurrence can be also calculated as frequency ratio from each TFs (Vandenbon et al, 2012). As FIMO presents the match motif patterns, the patterns can be also used to create new PWM and visualize it as a sequence logo. Therefore, an additional function for scanning common motif sequences of co-expressed genes between two TFs would be advantageous for users for securing the significant results of wTO outcome.

The written function returns a formatted table with a co-expressed and overlapped gene from two TFs, first TF, second TF, and matched sequences from each TF. The function checks all possible TF combinations, which is discovered from FIMO within the promoter of each gene. Assuming two TFs (TF1, TF2) were predicted to bind in the promoter of gene “gene1” with motifs (TF1 - motif A, TF1 - motif B, TF2 - motif C), then the table will simply return with following format:

| Gene | Node1 | Node 2 | Occurrence | sequence |
|---|---|---|---|---|
| gene1 | TF1 | TF2 | 3 | motif A, motif B, motif C |

To feature the gene expression profile specifically, the table only contains distinct combinations of TFs and counts sequence occurrences by each TF. Aiming to reinforce the probability of each one of link scores, which is computed by wTO, and to avoid any spurious correlation by random bootstrap or resampling (Efron & Tibshirani, 1994). Moreover, the coordinates of each gene expression have been saved for further investigation of quantifying the actual neighbor site and distance of TFs. The original result of wTO network was TF-focused based on basic correlation (Equation 3.1), now

with written function, it will be able to capture the TFs binding patterns of correlated genes and furthermore can gauge a wider collection of genes with coherence matrices. The coherence matrix reserves all counts of occurrences between TFs and its target gene and constructs basic structure for network visualization. The absolute count of co-occurrences are based on solid p-values (Equation 3.6) and the co-occurrences can be relatively compared to each other throughout using the matrix. Consequently, the feature of the table can be visualized in a simple network, showing the general trend of TFs interaction and its profound binding occurrence(Fig. 8).

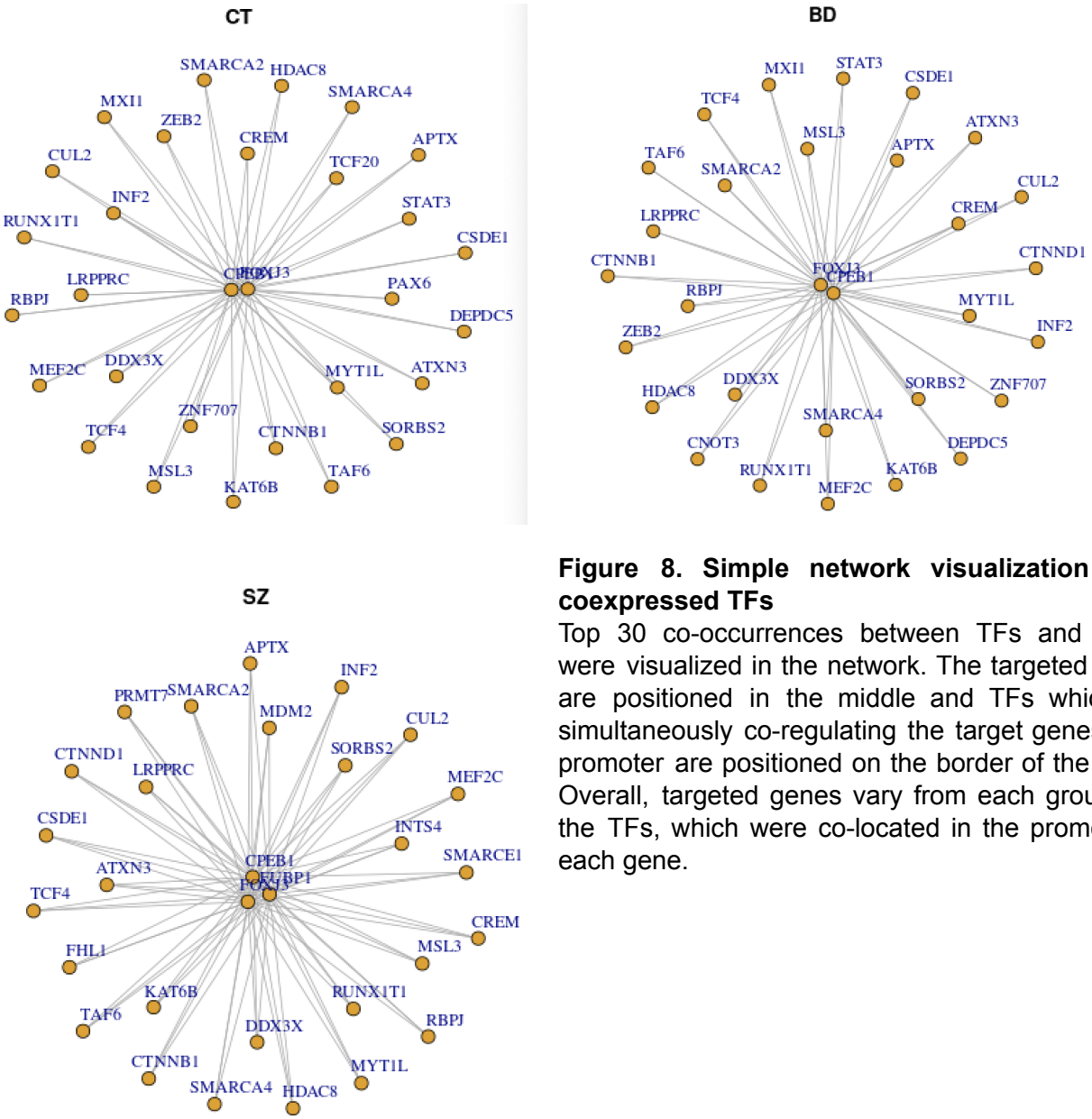


**Figure 8. Simple network visualization of a coexpressed TFs**

Top 30 co-occurrences between TFs and genes were visualized in the network. The targeted genes are positioned in the middle and TFs which are simultaneously co-regulating the target genes in its promoter are positioned on the border of the circle. Overall, targeted genes vary from each group and the TFs, which were co-located in the promoter of each gene.

# 5. Discussion

## 5.1. Accuracy of TFBS prediction

Accurate prediction of TFBS for gene finding is required to understand the regulation and the applications such as evaluating the effects of single nucleotide variations (SNVs), which is causing differential expression (Worsley-Hunt et al, 2011) and leading to various diseases (Jarinova & Ekker, 2012). The result of this thesis is based on the protocol of FIMO from the MEME suite. The performance is assessed by calculating the ratio of predicted binding sites that has a minimum overlap of 70 % of bases pairs with known binding sites from PAZAR, a diagnostic test with true and negative predicted binding sites, gaps between actual location of a motif and proportion of predicted and actual negative sites. It shows that FIMO carries out the performance metric value of average sensitivity (0.816), positive predictive volume (0.734), geometric accuracy (0.774) and false positive rate (0.015) using scrambled sequences (Jayaram et al, 2016; Sand et al, 2008). Also, general promoter annotation depends on official gene names of different databases and still, in terms of the human genome, it is only a trivial proportion of TFBS to be a global reference (Vierstra et al, 2020; Vierstra & Stamatoyannopoulos, 2016). Despite its authority from each collection, the accuracy of peak detection can vary by names from wTO results. Determining the location of the TSS is a critical step in identifying the promoter region, the precise information or study, of which is inevitably necessary to spot the gene expression patterns, regulatory networks, cell differentiation, and development (Xiaoyue et al, 2007). Further motif length may have little relevance on the quality of motif, independent of other factors, such as number of binding sites for the gene regulation, type of targeted gene, and potential of mutation( Wunderlich & Zeba, 2009). However, specific motifs may affect the performance of log-odds scores with the PWM model, which various ability of PWM models for some TF motifs influences the discriminative power of analysis (Kibet & Machanick, 2015). Classically, computational prediction of TFBSs is based on basic PWMs, which quantitatively score binding motifs based on the observed nt patterns in a set of TFBSs for the corresponding TF. PWM models make the robust presumption that each nt engages independently in the corresponding DNA-protein interaction and do not consider flexible binding characteristics of motifs. Even though it is not a general rule, some TFs exemplify a different scenario. Therefore, it can be considered to use the transcription

factor flexible model (TFFM) or DRAF model as an option to target performance of DNA-protein interaction. TFFM is originated to address the confounding properties of nt conformation, inter-positional sequence dependence and variable lengths observed in the more comprehensive experimental data, which is driven from thousands of experimentally validated DNA-TF interaction sequences from ENCODE ChIP-seq data set with Hidden Markov model (HMM). HMM offers the probabilistic relationship between states to model sequential data and has been widely used in computational biology for the prediction of protein motifs. Advantage of such a model is that the model both positions interdependence within TFBSs and variable length motifs within a single dedicated framework (Mathelier & Wasserman, 2013; Rabiner & Juang, 1986). The DRAF model is also a TFBS prediction tool, which uses physicochemical properties of TFs and binary representation of TFBSs with random forest (RFs) machine learning models (Mathelier & Wasserman, 2013). False positive rate can be decreased compared to FIMO by building multiple decision trees based on RF (Breiman, 2001). In addition, the length of the promoter can potentially impact the accuracy of TFBS prediction. Along with FIMO's highest precision compared to other mapping tools, such as Cluster-Buster, Motif Occurrence Detection, Matrix- Scan, Ensembl, reportedly the median precision for FIMO is 5 % - compared with 2.2 % to 2.4 % for the other tools. Although recall is low with the FIMO results as a consequence to the tool prediction, approximately 10 fold fewer true motif matches relative to the other tools (22 % median recall versus 36 %–48 % for the other tools). Recommendable Cluster-Buster can be conducted additionally with FIMO to overcome the recall problem of FIMO. Cluster-Buster is beneficial to discriminate motif clusters from background DNA by searching for regions of the sequence that resembles a statistical model of a motif cluster more than they favor a model of background DNA, which is DNA left behind from an action of interest (Frith, 2003). In the analysis, 5 Kbp upstream and 3 Kbp downstream sequence was used to scan and to predict the overall motif matches. The 5 Kbp upstream promoter region of the TSS of the gene is crucial for transcription initiation and defined as the longest transcript for the gene (Patel & Bush,2021). For length of downstream, 3 Kbp is set to default based on consideration of Brain-derived neurotrophic factor (BDNF). BDNF is a secreted protein of the neurophin family which controls the survival, growth, and function of neurons both during the development and in the adult nervous system. Defects in BDNF expression and signaling have been implicated in various neuropsychiatric and neurodegenerative diseases, including major depression, schizophrenia, Alzheimer's disease, and Huntington's disease (Tuvikene et al, 2021; Autry & Monteggia, 2012). The 3 Kbp region is hypersensitive to DNaseI, indicative of an open chromatin structure. For further points of its possible importance, based on mouse brain tissue data, 3 Kbp downstream can show enhancer marks from histons, higher activation of transcription due to interaction with RNA polymerase II, epigenetic

modification following regulation of gene expression, and the region is conserved between mammals. However, FIMO preferably takes 1,000 bp as default sequence with threshold p-value of 0.0001 (Grant, Charles E., Timothy L. Bailey 2011). FIMO internally cut off the restricted spurious matches with short motifs and low complexity motifs due to increased threshold by length, which should have appeared as real TFBS based on ChIP-seq data (Kulkarni et al, 2019). Also the length of the promoter had a significant effect on running time and computational power. Therefore, for later prediction with FIMO, users should adjust the length of the scanned window to be less than 8 Kbp or examine conditionally in extra. Particularly, the distance of the promoter region for brain derived neurotrophic factor spans divergently by purpose of studies (Abdolmaleky et al, 2006; Stopkova et al, 2005). As well the precision of identifying TFBS can be increased by cross validation with tool TSSFinder. The TSSFinder is a tool, which concentrates on location TSS to model the core promoter region and to predict TSS signals of eukaryotic genes with a probabilistic model on characterization of gene structures (de Medeiros Oliveira et al, 2021).

## 5.2. Structural motif

In analysis with FIMO, the TFBSs shorter than 9 bp were not detected because possible best matches for short motifs were not identified as statistically significant by edited p-value input (Equation 3.6). However, experimentally to argue, most TFs bind to short sequences around 5 - 10 bp, and degenerate sequence motifs that occur very frequently in the human genome (Loots et al, 2002). Nonetheless, the basic characteristics of TFBSs by higher eukaryotes are remarkably conserved across life, despite vast differences in population size, genome size, regulatory complexity, and selective regimes (Blanchette et al, 2006). Natural selection and function of individual TFBS plays a role by determining the length and information content of a TFBS (Shultzaberger et al, 2012). Hence, missing out part of empirical distribution of binding site lengths can cause selection bias of underlying mechanistic parameters in the result (Sella & Hirsh, 2005). Consequently, supplemental analysis on shorter motifs should be included to assist the FIMO result. Moreover, scrutiny of motif length can be pivoted besides prediction of binding site. The evolution of binding sites length and information content have altered the fate of TFs. Different TFs can partially share their binding sequences due to their common evolutionary origins. This “redundancy” of binding defines a way of organizing TFs in “motif families” by grouping TFs with similar binding preferences (Rosanova et al, 2017). For example, TFs of *Escherichia coli* regulate a greater number of target genes that exhibit less total information content in their binding

site, meaning factors that serve as important master regulators of many genes achieve this by increased promiscuity of binding. Therefore, such as mutation rate per length, degeneracy per nt, and explanation of specificity and promiscuity can be inspected along binding sites. Often mutation can lead to a collection of new motifs, which can expand the curiosity, whether previously identified TFBS motif resembles any functional or biological measurement. Subsequently, studying the similarity of TFs motif give several fundamental opportunity, such as elucidating the relationship between sequence and function of TFs (Weirauch et al, 2014), assigning further known TF to *de novo* discovered motif (Gupta et al, 2007) and estimating performance of in silico motif prediction approaches (Najafabadi et al, 2015). In particular, the *de novo* motif identification can be conducted to compare whether newly discovered and putative motifs resemble any previously discovered motif in an existing database (Gupta et al, 2007). Additionally, *In silico* motif prediction can supplement the analysis, aiming to predict the secondary and tertiary structure of primary amino acid sequence (Mooney et al, 2011). Moreover, motifs can have sufficient conservation, in consequences affecting the protein function and its prediction (Golovin & Henrick, 2008; Fox-Erlich & Schiller, 2009). Another pinpoint is about co-occurrences that are also strongly influenced by motif similarity (Vandenbon et al, 2012). Namely the result of FIMO can be further quantified with the TOMTOM tool from the MEME suite, for searching through a database of similar motifs with a given query and searching further co-occurrences (Bailey et al, 2015).

## 5.3 Binding Affinity

The outcome of FIMO contains the coordinate of the TFs binding site with the location of the chromosome and its range on each matched sequence. Approaching to model the DNA-binding specificity of TFs by grouping and tallying the TFs and its chromosome location showed compulsive counts. Noticeably, an integrative analysis of TFBS cluster region chromatin landscape can give birds eye view to transcriptional regulation. In the character of some TFs bound to adjacent sites interact cooperatively at the protein level to improve the affinity of the TF complex to specific sites in the genome, for instance co-binding of TFs to a common cofactor or complex (Spitz & Furlong, 2012). For example, TFs were particularly interesting as effectors of broad phenotypic changes, due to the large number of genes they regulate. It is thus possible that by themselves, or in oligogenic combinations, they can account for complex disorders such as bipolar disorder or schizophrenia (Le-Niculescu et al, 2009). Foremost, the MEF2 family of TF are shown in BD and SZ (Fig. 3 name of TF in BD and name of TF in SZ). MEF2

regulates large programs of gene expression pivotal for the development and maintenance of the brain (Assali et al, 2019). MEF2 proteins are known to regulate by neuronal synaptic activity, and they recruit several epigenetic enzymes to influence chromatin structure and is a candidate risk gene for several common mental disorder, inclusind bipolar disorder, schizophrenia, attention deficit and hyperactivity disorder, major depressive disorder, and alzheimer disease (Ahlem et al, 2020; Nurnberger et al, 2014; Xie et al, 2017; Mitchell et al, 2017; Purcell et al, 2014; Nature, 2014; Deczkowska et al, 2017). KLF12 (Fig. 3 BD) is a transcription factor, more specifically a zinc finger transcriptional repressor, which is one of top candidate genes for bipolar disorder identified by Genome wide Association studies. In evidence of mouse genetics, KLF12 maps to a mouse for abnormal emotion and affect of behavior. In human studies, increased expression in low mood, i.e. depression (Le-Niculescu et al, 2009). TEF (Fig. 3 SZ) is reported in the polymorphisms of a clock-related gene, which contributes to the risk of sleep disturbance and depression in Parkinson disease (Hua et al, 2014). A deficiency of TEF, TF family group appears to have decreased brain levels of serotonin, and dopamine and such changes previously been reported to cause epilepsies in other systems by mice (Gachon et al, 2004). Epilepsy is also associated with risk of developing SZ (Andersen et al, 2019). Expression of FUBP1 (Fig. 3 SZ) is dynamically regulated during adult neurogenesis. An overexpression of TF contributes to promoting neoplastic cell proliferation and is observed in many human cancers (Hwang et al, 2018; Ma et al, 2021). Also in Figure 7, the CREB3 shows individually binding characters. The role of CREB3 is known to induce transcription of the endoplasmic reticulum(ER) to stress, which is also involved in the unfolded protein response (Reiling et al. 2013). Disruption of ER in mice exhibited a blunted stress response characterized by low levels of both anxiety and depressive- like behavior in addition to lower corticosterone levels (Penney et al. 2017). In humans, corticosterone is an important intermediate in the synthesis of cortisol secretion and could be important to understand neurodevelopmental stress and BDNF deficiency in cognitive symptoms of schizophrenia (Klug et al. 2012). Identifying such different measures of TFs can be beneficial for understanding causal links of pheno- and genotype. Moreover genetic variants in the binding motif (motif dependent) disrupt the TF recognition of a motif, but the variants beyond the binding motif (motif independent) induce either proximal (< 200 bp) or distal variants of TF-DNA binding, which induce cooperative and collaborative binding and co-existence of common motifs in TFs and genes will give higher probability of binding affinity (Ibarra et al, 2020). The strength of the binding interaction can be presented with binding affinity, which can be a candidate for evaluating the coexpression. These binary interaction metrics can also be used to construct reliable quality measurement of clustering the protein–protein interaction networks (Şenbabaoğlu et al, 2016). As a consequence, the result raises the enthusiasm to

investigate a larger number of ChIP-seq of TFs of interest to orient the TF binding motif/sites.

## 5.3. Future work

This thesis identified possible TFBS between BD and SZ in a heuristic way. Regardless, it also revealed a way of improving the workflow and adjustments that was used to enrich the wTO package and TFBS prediction. Consequently, a repetition of the analysis with an adapted pipeline is beneficial in order to enhance the results. This includes mixing TFBS prediction tools, fine-tune of promoter length, higher sample size of TFs, different filtering strategies, consideration of TF evolution and a more thorough analysis of the individual TF - TF interaction. Furthermore, a validation of the target genes and actual TFBS were not thoroughly examined, which could supplement the network visualization in wTO with TFBS co-occurrence. Consequently, since TFBS prediction generated in this study identified different co-expression patterns for the gene from the TFs of interest and all known TFs, comparing existing wTO networks and the patterns could uncover the underlying biological or characteristics of TFs. As a next step to consolidate the findings, comparing bootstrapped gene samples and results of co-expression patterns is recommendable as this would minimize the insecurities that were caused in the symbol transformation. Finally, the resulting TFBS prediction reveals new candidate co-expression for genes of interest. Since most of them have not considered themselves to be connected or cooperate, further analysis on these would provide a new approach to strengthen the wTO package.

# 6. References


1. Crick, Francis. "Central dogma of molecular biology." *Nature* 227.5258 (1970): 561-563.
2. Collins, Francis S., Michael Morgan, and Aristides Patrinos. "The Human Genome Project: lessons from large-scale biology." *Science* 300.5617 (2003): 286-290.
3. Ricroch, A., et al. "Evolution of genome size across some cultivated Allium species." *Genome* 48.3 (2005): 511-520.
4. Hanson, MA and, and P. D. Gluckman. "Early developmental conditioning of later health and disease: physiology or pathophysiology?." *Physiological reviews* (2014).
5. Manzoni, Claudia, et al. "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences." *Briefings in bioinformatics* 19.2 (2018): 286-302.
6. Dey, Siddharth S., et al. "Integrated genome and transcriptome sequencing of the same cell." *Nature biotechnology* 33.3 (2015): 285-289.
7. Williams, Rohan BH, et al. "The influence of genetic variation on gene expression." *Genome research* 17.12 (2007): 1707-1716.
8. Wang, Bo, et al. "Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing." *Frontiers in genetics* 10 (2019): 384.
9. Mitsis, Thanasis, et al. "Transcription factors and evolution: an integral part of gene expression." *World Academy of Sciences Journal* 2.1 (2020): 3-8.
10. Yu, Chun-Ping, and Wen-Hsiung Li. "Predicting transcription factor binding sites and their cognate transcription factors using gene expression data." *Plant Gene Regulatory Networks*. Humana Press, New York, NY, 2017. 271-282.
11. Inukai, Sachi, Kian Hong Kock, and Martha L. Bulyk. "Transcription factor–DNA binding: beyond binding site motifs." *Current opinion in genetics & development* 43 (2017): 110-119.
12. Veerla, Srinivas, and Mattias Höglund. "Analysis of promoter regions of co-expressed genes identified by microarray analysis." *BMC bioinformatics* 7.1 (2006): 1-15.
13. Lee, Tong Ihn, and Richard A. Young. "Transcriptional regulation and its misregulation in disease." *Cell* 152.6 (2013): 1237-1251.
14. Kapranov, Philipp. "From transcription start site to cell biology." *Genome Biology* 10.4 (2009): 1-4.
15. D'addario, Claudio, et al. "Regulation of gene transcription in bipolar disorders: role of DNA methylation in the relationship between prodynorphin and brain derived neurotrophic factor." *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 82 (2018): 314-321.

16. D'haeseleer, Patrik. "What are DNA sequence motifs?." *Nature biotechnology* 24.4 (2006): 423-425.
17. Lis, Monika, and Dirk Walther. "The orientation of transcription factor binding site motifs in gene promoter regions: does it matter?." *BMC genomics* 17.1 (2016): 1-21.
18. Yukawa, Yasushi, et al. "A common sequence motif involved in selection of transcription start sites of Arabidopsis and budding yeast tRNA genes." *Genomics* 97.3 (2011): 166-172.
19. Stormo, Gary D. "DNA binding sites: representation and discovery." *Bioinformatics* 16.1 (2000): 16-23.
20. Aerts, Stein. "Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets." *Current topics in developmental biology* 98 (2012): 121-145.
21. Wood, Heather. "New insights into altered gene regulation in schizophrenia." *Nature Reviews Neurology* 15.11 (2019): 619-619.
22. Hobert, Oliver. "Gene regulation by transcription factors and microRNAs." *Science* 319.5871 (2008): 1785-1786.
23. Chen, Kevin, and Nikolaus Rajewsky. "The evolution of gene regulation by transcription factors and microRNAs." *Nature Reviews Genetics* 8.2 (2007): 93-103.
24. Lambert, Samuel A., et al. "The human transcription factors." *Cell* 172.4 (2018): 650-665.
25. Rodriguez-Caso, Carlos, Miguel A. Medina, and Ricard V. Sole. "Topology, tinkering and evolution of the human transcription factor network." *The FEBS journal* 272.24 (2005): 6423-6434.
26. Swift, Joseph, and Gloria M. Coruzzi. "A matter of time—how transient transcription factor interactions create dynamic gene regulatory networks." *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1860.1 (2017): 75-83.
27. Gysi, Deisy Morselli, et al. "wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool." *BMC bioinformatics* 19.1 (2018): 1-16.
28. Aoki, Koh, Yoshiyuki Ogata, and Daisuke Shibata. "Approaches for extracting practical information from gene co-expression networks in plant biology." *Plant and Cell Physiology* 48.3 (2007): 381-390.
29. Steuer, Ralf, et al. "The mutual information: detecting and evaluating dependencies between variables." *Bioinformatics* 18.suppl_2 (2002): S231-S240.
30. D'haeseleer, Patrik, Shoudan Liang, and Roland Somogyi. "Genetic network inference: from co-expression clustering to reverse engineering." *Bioinformatics* 16.8 (2000): 707-726.
31. Şenbabaoğlu, Yasin, et al. "A multi-method approach for proteomic network inference in 11 human cancers." *PLoS computational biology* 12.2 (2016): e1004765.
32. Albert, Réka, and Albert-László Barabási. "Statistical mechanics of complex networks." *Reviews of modern physics* 74.1 (2002): 47.

33. D'haeseleer, Patrik. "How does gene expression clustering work?." *Nature biotechnology* 23.12 (2005): 1499-1501.
34. Ravasz, Erzsébet, et al. "Hierarchical organization of modularity in metabolic networks." *science* 297.5586 (2002): 1551-1555.
35. Nowick, Katja, et al. "Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain." *Proceedings of the National Academy of Sciences* 106.52 (2009): 22358-22363.
36. https://cran.r-project.org/web/packages/wTO/wTO.pdf
37. Langfelder, Peter, and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis." *BMC bioinformatics* 9.1 (2008): 1-13.
38. Xulvi-Brunet, Ramon, and Hongzhe Li. "Co-expression networks: graph properties and topological comparisons." *Bioinformatics* 26.2 (2010): 205-214.
39. Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA statement on p-values: context, process, and purpose." *The American Statistician* 70.2 (2016): 129-133.
40. Langfelder, Peter, and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis." *BMC bioinformatics* 9.1 (2008): 1-13.
41. Margolin, Adam A., et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." *BMC bioinformatics*. Vol. 7. No. 1. BioMed Central, 2006.
42. Amar, David, Hershel Safer, and Ron Shamir. "Dissection of regulatory networks that are altered in disease via differential co-expression." *PLoS computational biology* 9.3 (2013): e1002955.
43. Hartwell, Leland H., et al. "From molecular to modular cell biology." *Nature* 402.6761 (1999): C47-C52.
44. Van Dam, Sipko, et al. "Gene co-expression analysis for functional classification and gene–disease predictions." *Briefings in bioinformatics* 19.4 (2018): 575-592.
45. Barnes, John A. "Graph theory and social networks: A technical comment on connectedness and connectivity." *Sociology* 3.2 (1969): 215-232.
46. Jeong, Hawoong, et al. "The large-scale organization of metabolic networks." *Nature* 407.6804 (2000): 651-654.
47. Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." *nature* 393.6684 (1998): 440-442.
48. Aoki, Koh, Yoshiyuki Ogata, and Daisuke Shibata. "Approaches for extracting practical information from gene co-expression networks in plant biology." *Plant and Cell Physiology* 48.3 (2007): 381-390.
49. Barabasi, Albert-Laszlo, and Zoltan N. Oltvai. "Network biology: understanding the cell's functional organization." *Nature reviews genetics* 5.2 (2004): 101-113.
50. Van Dam, Sipko, et al. "Gene co-expression analysis for functional classification and gene–disease predictions." *Briefings in bioinformatics* 19.4 (2018): 575-592.
51. 1 Transcription factor motifs. *Nature* (2019). https://doi.org/10.1038/nature28170
52. Maston, Glenn A., Sara K. Evans, and Michael R. Green. "Transcriptional regulatory elements in the human genome." *Annu. Rev. Genomics Hum. Genet.* 7 (2006): 29-59.
53. Odom, Duncan T., et al. "Core transcriptional regulatory circuitry in human hepatocytes." *Molecular systems biology* 2.1 (2006): 2006-0017.

54. Zinzen, Robert P., et al. "Combinatorial binding predicts spatio-temporal cis-regulatory activity." *Nature* 462.7269 (2009): 65-70.
55. Zambelli, Federico, Graziano Pesole, and Giulio Pavesi. "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era." *Briefings in bioinformatics* 14.2 (2013): 225-237.
56. Boeva, Valentina, et al. "De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis." *Nucleic acids research* 38.11 (2010): e126-e126.
57. Mitra, Sneha, Anushua Biswas, and Leelavati Narlikar. "DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP." *PLoS computational biology* 14.4 (2018): e1006090.
58. Spitz, François, and Eileen EM Furlong. "Transcription factors: from enhancer binding to developmental control." *Nature reviews genetics* 13.9 (2012): 613-626.
59. Ibarra, Ignacio L., et al. "Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions." *Nature communications* 11.1 (2020): 1-16
60. Weirauch, Matthew T., et al. "Determination and inference of eukaryotic transcription factor sequence specificity." *Cell* 158.6 (2014): 1431-1443.
61. Gupta, Shobhit, et al. "Quantifying similarity between motifs." *Genome biology* 8.2 (2007): 1-9.
62. Najafabadi, Hamed S., et al. "C2H2 zinc finger proteins greatly expand the human regulatory lexicon." *Nature biotechnology* 33.5 (2015): 555-562.
63. Mooney, Catherine, et al. "In silico protein motif discovery and structural analysis." *In Silico Tools for Gene Discovery*. Humana Press, 2011. 341-353.
64. Zhang, Hongen. "Overview of sequence data formats." *Statistical Genomics*. Humana Press, New York, NY, 2016. 3-17.
65. Stormo, Gary D., et al. "Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli." *Nucleic acids research* 10.9 (1982): 2997-3011.
66. Tremblay, B. J. M. (2019). Introduction to sequence motifs.
67. Team, R. Core. "R: A language and environment for statistical computing." (2013): 201.
68. Lawrence, Michael, et al. "Software for computing and annotating genomic ranges." *PLoS computational biology* 9.8 (2013): e1003118.
69. Team, Bioconductor Core, and Bioconductor Package Maintainer. "TxDb. Hsapiens. UCSC. hg38. knownGene: Annotation package for TxDb object (s)." *R package* (2019).
70. Team TBD (2021). *BSgenome.Hsapiens.UCSC.hg38: Full genome sequences for Homo sapiens (UCSC version hg38, based on GRCh38.p13)*. R package version 1.4.4.
71. Lee, Stuart, Dianne Cook, and Michael Lawrence. "Plyranges: A grammar of genomic data transformation." *Genome biology* 20.1 (2019): 1-10.
72. Durinck, Steffen, et al. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." *Nature protocols* 4.8 (2009): 1184-1191.

73. Statham, Aaron L., et al. "Repitools: an R package for the analysis of enrichment-based epigenomic data." *Bioinformatics* 26.13 (2010): 1662-1663.
74. Shannon P, Richards M (2021). *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs*. R package version 1.34.0.
75. Wickham H (2022). *stringr: Simple, Consistent Wrappers for Common String Operations*. http://stringr.tidyverse.org, https://github.com/tidyverse/stringr.
76. Wickham, Hadley, et al. "dplyr: A Grammar of Data Manipulation. R package version 0.8. 0.1." *Retrieved January* 13 (2019): 2020.
77. Tremblay BJ (2022). *universalmotif: Import, Modify, and Export Motifs with R*. R package version 1.12.4, https://bioconductor.org/packages/universalmotif/.
78. Nystrom S (2022). *memes: motif matching, comparison, and de novo discovery using the MEME Suite*. https://snystrom.github.io/memes/, https://github.com/snystrom/memes.
79. Bache, Stefan Milton, and Hadley Wickham. "magrittr: a forward-pipe operator for R." *R package version* 1.1 (2014).
80. Jovanovic, Vladimir M., et al. "Positive selection in gene regulatory factors suggests adaptive pleiotropic changes during human evolution." *Frontiers in genetics* 12 (2021): 753.
81. Heumann, Christian, and Michael Schomaker. *Introduction to statistics and data analysis*. Springer International Publishing Switzerland, 2016.
82. Jafari, Mohieddin, and Naser Ansari-Pour. "Why, when and how to adjust your P values?." *Cell Journal (Yakhteh)* 20.4 (2019): 604.
83. Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995): 289-300.
84. Anthony, Peter D. *The ideology of work*. Routledge, 2014.
85. Wingender, Edgar. "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation." *Briefings in bioinformatics* 9.4 (2008): 326-332.
86. Rapicavoli, Nicole A., et al. "A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics." *elife* 2 (2013): e00762.
87. Stormo, Gary D., Zheng Zuo, and Yiming Kenny Chang. "Spec-seq: determining protein–DNA-binding specificity by sequencing." *Briefings in functional genomics* 14.1 (2015): 30-38.
88. Kulakovskiy, Ivan V., et al. "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis." *Nucleic acids research* 46.D1 (2018): D252-D259.
89. Landt, Stephen G., et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." *Genome research* 22.9 (2012): 1813-1831.
90. Valouev, Anton, et al. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." *Nature methods* 5.9 (2008): 829-834.
91. Kibet, Caleb Kipkurui, and Philip Machanick. "Transcription factor motif quality assessment requires systematic comparative analysis." *F1000Research* 4 (2015).

92. Plaisier, Christopher L., et al. "Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis." *Cell systems* 3.2 (2016): 172-186.
93. Levy, Samuel, and Sridhar Hannenhalli. "Identification of transcription factor binding sites in the human genome sequence." *Mammalian Genome* 13.9 (2002): 510-514.
94. Patel, Neel, and William S. Bush. "Modeling transcriptional regulation using gene regulatory networks based on multi-omics data sources." *BMC bioinformatics* 22.1 (2021): 1-19.
95. Tuvikene, Jürgen, et al. "Intronic enhancer region governs transcript-specific Bdnf expression in rodent neurons." *Elife* 10 (2021): e65161.
96. Autry, Anita E., and Lisa M. Monteggia. "Brain-derived neurotrophic factor and neuropsychiatric disorders." *Pharmacological reviews* 64.2 (2012): 238-258.
97. Tweedie, Susan, et al. "Genenames. org: the HGNC and VGNC resources in 2021." *Nucleic acids research* 49.D1 (2021): D939-D946.
98. Howe, Kevin L., et al. "Ensembl 2021." *Nucleic acids research* 49.D1 (2021): D884-D891.
99. Kent, W. James, et al. "The human genome browser at UCSC." *Genome research* 12.6 (2002): 996-1006.
100. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., ... & Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome research, 27(5), 849-864.
101. Arora S, Morgan M, Carlson M, Pagès H (2022). *GenomeInfoDb: Utilities for manipulating chromosome names, including modifying them to follow a particular naming style*. R package version 1.30.1, https://bioconductor.org/packages/GenomeInfoDb.
102. Lee, Brian T., et al. "The UCSC Genome Browser database: 2022 update." *Nucleic acids research* 50.D1 (2022): D1115-D1122.
103. Schneider, Valerie A., et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly." *Genome research* 27.5 (2017): 849-864.
104. Padmalatha, Kethireddy Venkata, et al. "Genome-wide transcriptomic analysis of cotton under drought stress reveal significant down-regulation of genes and pathways involved in fibre elongation and up-regulation of defense responsive genes." *Plant molecular biology* 78.3 (2012): 223-246.
105. Jayaram, Narayan, Daniel Usvyat, and Andrew CR Martin. "Evaluating tools for transcription factor binding site prediction." *BMC bioinformatics* 17.1 (2016): 1-12.
106. Chiu, Tsu-Pei, et al. "TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites." *Nucleic acids research* 48.D1 (2020): D246-D255.
107. Portales-Casamar, Elodie, et al. "The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences." *Nucleic acids research* 37.suppl_1 (2009): D54-D60.

108. Griffith, Obi L., et al. "ORegAnno: an open-access community-driven resource for regulatory annotation." *Nucleic acids research* 36.suppl_1 (2007): D107-D113.
109. Vierstra, Jeff, et al. "Global reference mapping of human transcription factor footprints." *Nature* 583.7818 (2020): 729-736.
110. Vierstra, Jeff, and John A. Stamatoyannopoulos. "Genomic footprinting." *Nature methods* 13.3 (2016): 213-221.
111. Yevshin, Ivan, et al. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments." *Nucleic acids research* (2016): gkw951.
112. Worsley-Hunt, Rebecca, Virginie Bernard, and Wyeth W. Wasserman. "Identification of cis-regulatory sequence variations in individual genome sequences." *Genome Medicine* 3.10 (2011): 1-14.
113. Jarinova, Olga, and Marc Ekker. "Regulatory variations in the era of next generation sequencing: Implications for clinical molecular diagnostics." *Human mutation* 33.7 (2012): 1021-1030.
114. Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif." *Bioinformatics* 27.7 (2011): 1017-1018.
115. Timothy L. Bailey, James Johnson, Charles E. Grant, William S. Noble, "The MEME Suite", Nucleic Acids Research, 43(W1):W39-W49, 2015
116. Sand, Olivier, J-V. Turatsinze, and J. van Helden. "Evaluating the prediction of cis-acting regulatory elements in genome sequences." *Modern genome annotation*. Springer, Vienna, 2008. 55-89.
117. Wingender, Edgar, et al. "TFClass: a classification of human transcription factors and their rodent orthologs." *Nucleic acids research* 43.D1 (2015): D97-D102.
118. https://www.wipo.int/export/sites/www/standards/en/pdf/03-25-01.pdf
119. Boeva, Valentina. "Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells." *Frontiers in genetics* 7 (2016): 24.
120. Staden, Rodger. "Staden: searching for motifs in nucleic acid sequences." *Computer Analysis of Sequence Data*. Springer, Totowa, NJ, 1994. 93-102.
121. Storey, John D. "The positive false discovery rate: a Bayesian interpretation and the q-value." *The Annals of Statistics* 31.6 (2003): 2013-2035.
122. Noble, William S. "How does multiple testing correction work?." *Nature biotechnology* 27.12 (2009): 1135-1137.
123. Le Niculescu, H., Patel, S. D., Bhat, M., Kuczenski, R., Faraone, S. V., Tsuang, M. T., ... & Niculescu Iii, A. B. (2009). Convergent functional genomics of genomewide association data for bipolar disorder: Comprehensive identification of candidate genes, pathways and mechanisms. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 150(2), 155-181.
124. Assali, Ahlem, Adam J. Harrington, and Christopher W. Cowan. "Emerging roles for MEF2 in brain development and mental disorders." *Current opinion in neurobiology* 59 (2019): 49-58.

125. Assali, A., Harrington, A. J., & Cowan, C. W. (2019). Emerging roles for MEF2 in brain development and mental disorders. Current opinion in neurobiology, 59, 49-58.
126. Nurnberger, J. I., Koller, D. L., Jung, J., Edenberg, H. J., Foroud, T., Guella, I., ... & Kelsoe, J. R. (2014). Identification of pathways for bipolar disorder: a meta-analysis. JAMA psychiatry, 71(6), 657-664.
127. Xie, Z., Yang, X., Deng, X., Ma, M., & Shu, K. (2017). A genome-wide association study and complex network identify four core hub genes in bipolar disorder. International Journal of Molecular Sciences, 18(12), 2763.
128. Mitchell, A. C., Javidfar, B., Pothula, V., Ibi, D., Shen, E. Y., Peter, C. J., ... & Akbarian, S. (2018). MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice. Molecular psychiatry, 23(1), 123-132.
129. Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., ... & Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature, 506(7487), 185-190.
130. Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K. H., Holmans, P. A., ... & Milanova, V. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421.
131. Deczkowska, A., Matcovitch-Natan, O., Tsitsou-Kampeli, A., Ben-Hamo, S., Dvir-Szternfeld, R., Spinrad, A., ... & Schwartz, M. (2017). Mef2C restrains microglial inflammatory response and is lost in brain ageing in an IFN-I-dependent manner. Nature communications, 8(1), 1-13.
132. Hua, P., Liu, W., Chen, D., Zhao, Y., Chen, L., Zhang, N., ... & Kuo, S. H. (2014). Cry1 and Tef gene polymorphisms are associated with major depressive disorder in the Chinese population. Journal of affective disorders, 157, 100-103.
133. Gachon, F., Fonjallaz, P., Damiola, F., Gos, P., Kodama, T., Zakany, J., ... & Schibler, U. (2004). The loss of circadian PAR bZip transcription factors results in epilepsy. Genes & development, 18(12), 1397-1412.
134. Andersen, K. M., Petersen, L. V., Vestergaard, M., Pedersen, C. B., & Christensen, J. (2019). Premature mortality in persons with epilepsy and schizophrenia: A population‐based nationwide cohort study. Epilepsia, 60(6), 1200-1208.
135. Hwang, I., Cao, D., Na, Y., Kim, D. Y., Zhang, T., Yao, J., ... & Paik, J. (2018). Far upstream element-binding protein 1 regulates LSD1 alternative splicing to promote terminal differentiation of neural progenitors. Stem cell reports, 10(4), 1208-1221.
136. Ma, C., Huang, Z., Wu, Z., Di, C., Lin, X., Huang, M., ... & Yin, H. (2021). Overexpression of FUBP1 is associated with human cervical carcinoma development and prognosis. Life Sciences, 269, 119098.
137. Stewart, A. J., Hannenhalli, S., & Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. Genetics, 192(3), 973-985.
138. Veerla, S., & Höglund, M. (2006). Analysis of promoter regions of co-expressed genes identified by microarray analysis. BMC bioinformatics, 7(1), 1-15.

139. Vandenbon, A., Kumagai, Y., Akira, S., & Standley, D. M. (2012, December). A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription. In BMC genomics (Vol. 13, No. 7, pp. 1-15). BioMed Central.
140. Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.
141. Worsley-Hunt, R., Bernard, V., & Wasserman, W. W. (2011). Identification of cis-regulatory sequence variations in individual genome sequences. Genome Medicine, 3(10), 1-14.
142. Jarinova, O., & Ekker, M. (2012). Regulatory variations in the era of next generation sequencing: Implications for clinical molecular diagnostics. Human mutation, 33(7), 1021-1030.
143. Kibet, C. K., & Machanick, P. (2015). Transcription factor motif quality assessment requires systematic comparative analysis. F1000Research, 4.
144. Mathelier, A., & Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. PLoS computational biology, 9(9), e1003214.
145. Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. ieee assp magazine, 3(1), 4-16.
146. Mathelier, A., & Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. PLoS computational biology, 9(9), e1003214.
147. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
148. Frith, M. C., Li, M. C., & Weng, Z. (2003). Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic acids research, 31(13), 3666-3668.
149. Kulkarni, S. R., Jones, D. M., & Vandepoele, K. (2019). Enhanced maps of transcription factor binding sites improve regulatory networks learned from accessible chromatin data. Plant physiology, 181(2), 412-425.
150. Abdolmaleky, H. M., Cheng, K. H., Faraone, S. V., Wilcox, M., Glatt, S. J., Gao, F., ... & Thiagalingam, S. (2006). Hypomethylation of MB-COMT promoter is a major risk factor for schizophrenia and bipolar disorder. Human molecular genetics, 15(21), 3132-3145.
151. Stopkova, P., Vevera, J., Paclt, I., Zukov, I., Papolos, D. F., Saito, T., & Lachman, H. M. (2005). Screening of PIP5K2A promoter region for mutations in bipolar disorder and schizophrenia. Psychiatric genetics, 15(3), 223-227.
152. de Medeiros Oliveira, M., Bonadio, I., Lie de Melo, A., Mendes Souza, G., & Durham, A. M. (2021). TSSFinder—fast and accurate ab initio prediction of the core promoter in eukaryotic genomes. Briefings in bioinformatics, 22(6), bbab198.
153. Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I., & Rubin, E. M. (2002). rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome research, 12(5), 832-839.
154. Shultzaberger, R. K., Maerkl, S. J., Kirsch, J. F., & Eisen, M. B. (2012). Probing the Informational and regulatory plasticity of a transcription factor DNA–binding domain. PLoS genetics, 8(3), e1002614.

155. Sella, G., & Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences, 102(27), 9541-9546.
156. Rosanova, A., Colliva, A., Osella, M., & Caselle, M. (2017). Modelling the evolution of transcription factor binding preferences in complex eukaryotes. Scientific reports, 7(1), 1-12.
157. Golovin, A., & Henrick, K. (2008). MSDmotif: exploring protein sites and motifs. BMC bioinformatics, 9(1), 1-11.
158. Vandenbon, A., Kumagai, Y., Akira, S., & Standley, D. M. (2012, December). A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription. In BMC genomics (Vol. 13, No. 7, pp. 1-15). BioMed Central.
159. Reiling, Jan H., et al. "A CREB3–ARF4 signalling pathway mediates the response to Golgi stress and susceptibility to pathogens." *Nature cell biology* 15.12 (2013): 1473-1485.
160. Penney, Jenna, et al. "LUMAN/CREB3 is a key regulator of glucocorticoid-mediated stress responses." *Molecular and cellular endocrinology* 439 (2017): 95-104.
161. Klug, Maren, et al. "Long-term behavioral and NMDA receptor effects of young-adult corticosterone treatment in BDNF heterozygous mice." *Neurobiology of disease* 46.3 (2012): 722-731.
162. Fox-Erlich, Susan, Martin R. Schiller, and Michael R. Gryk. "Structural conservation of a short, functional, peptide-sequence motif." *Frontiers in bioscience: a journal and virtual library* 14 (2009): 1143.
163. Wunderlich, Zeba, and Leonid A. Mirny. "Different gene regulation strategies revealed by analysis of binding motifs." *Trends in genetics* 25.10 (2009): 434-440.
164. Blanchette, Mathieu, et al. "Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression." *Genome research* 16.5 (2006): 656-668.

Freie Universität Berlin

**Fachbereich Mathematik, Informatik und Physik**

**SELBSTSTÄNDIGKEITSERKLÄRUNG**

| | |
|---|---|
| Name: Park | (BITTE nur Block- oder Maschinenschrift verwenden.) |
| Vorname(n): Jin Soo | |
| Studiengang: Bioinformatik | |
| Matr. Nr.: 5009608 | |

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Bachelorarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe.

Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keiner anderen Universität als Prüfungsleistung eingereicht.

Datum: 26.04.2022 Unterschrift: Jin Soo Park