NOVEL METHODS FOR CONSTRUCTING, COMBINING AND COMPARING CO-EXPRESSION NETWORKS

Towards uncovering the molecular basis of human cognition
Der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

INFORMATIK

Vorgelegt

von B.Sc Deisy Morselli Gysi

geboren am 13. Juni 1990 in Curitiba, Brasilien

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter Stadler, Universität Leipzig
2. Prof. Dr. Martin Middendorf, Universität Leipzig
3. Prof. Dr. Katja Nowick, Freie Universität Berlin
4. Prof. Dr. Ulf Leser, Humboldt-Universität zu Berlin

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 4. April 2019 mit dem Gesamtprädikat *magna cum laude.*

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

(Ort, Datum)

(Unterschrift)

Zahid has a theory that autistic people are the normal ones because they see things as they really are while neurotypicals add an extra layer of meaning on to things.

Atypical

Acknowledgements

Firstly, I would like to thank my advisor Prof. Dr. Katja Nowick for the continuous support, patience, motivation and guiding me throughout this journey. I am also thankful for having Prof. Dr. Martin Middendorf all his great advises, discussions and adopting me as if one of his kin. I thank Prof. Dr. Peter Stadler for all his insightful comments, great discussions and encouragement. I am profoundly grateful to Prof. Dr. Eivind Almaas for hosting, supervising and being so extremely helpful in my internship in the land where the sun is not to be seen. I could not have imagined having a better set of advisor and mentors for my Ph.D. Thank you all, from the bottom of my heart. Without all your precious support it would not be possible to come to an end of this research.

I'm grateful for having nice collaborations during this time! Thanks, Prof. Dr. Volker Busskamp, for all your patience in explaining the biology of neurons, so I could convert it into network science, and for listening to the results patiently and translating back to biology! Prof. Dr. Marlis Reich, it was a major pleasure working with the enchanted world of magic mushrooms! I am also very grateful for all your insights, patience and kindness and for converting network metrics and statistical measures into biological communities. It was a pleasure finding the bad guys of the ocean. Tim and the bees, it was a great time being part of such a sweet project :P

Thank you Jens, Petra and Ute for all the help with all the bureaucratic and technical support. Thank you CNPq, for providing me with financial support for this research.

Of course, I cannot forget the discussions with my labmates from NowickLab, Swarm Intelligence, Bioinf, AlmaasLab and Regenerative Therapies. I'm grateful for my office mates (Tobias, Thanh, Nic and Tom) filling this journey with riddles. This period could not be so fun and filled of great moments without EVOP. Thanks, Katja, Rui and Vladi for organising this amazing platform for deep learning. I sincerely have to thank Jose, *ora pois*, and Kerry-Berry for always keeping in touch and sharing experiences. Thank you Vladi for proofreading the text.

A big thank you for the Research Academy of Leipzig, that provide us with amazing life-changing workshops. A special thanks, Susanne Skoruppa, for helping me to find and connect with my real self. I will never ever forget you.

My very special thanks to my two new non-blood brothers, that we adopted each other as a great family: Rituparno and Ali. Love you guys. This would never have the *cha-cha* fun without the salsa nights with Eugenio and Wilmer. Thank you, Tiago Fragoso for listening to my problems carefully and being a great friend and collaborator. Thank you also for proof-

reading the text! A special thanks for my special soulmate and favourite Drag Queen Elivane Victor, who has always been there. All the time. When things were great or bad. For supporting me and listening to my problems over a coffee on the weekends. *Soulmates never die*. Bel and Flavião I am also very thankful that we were all simultaneously in the same (but different) boats, sharing the freaking out moments, decisions and indecision.

Of course, I cannot forget my lovely and sweet fiancee, Wesley Bertoli, who decided to pursue his crazy Ph.D journey at the same time. And proposed twice during this period just to make sure that i) I did not change my mind; ii) I was not yet completely crazy. Yes, I do. I love you, sweetheart. Thank you for being the best boyfriend in the world, for supporting me and being my *safe harbour*.

Por último, mas não menos importante, sou eternamente grata a minha família linda e amada pelo seu apoio, dedicação e por compreender a minha ausência nesses últimos anos. Meus pais, Carlos e Denise, meus irmãos Matheus e Rodrigo e meu avô, Raul, que mesmo sem terem muita certeza do que eu faço, estao sempre orgulhosos e torcendo por mim. E um agradecimento especial a minha avó, AdeLinda, que decidiu não me acompanhar mais de longe e sim com sua presença espiritual. *You will always gonna be missed.* Amo muito vocês. Muito obrigada.

Abstract

Network analyses, such as of **gene co-expression networks** are an important approach for the systems-level study of biological data. For example, understanding patterns of co-regulation in mental disorders can contribute to the development of new therapies and treatments.

In a **gene regulatory** process a particular Transcription Factor (TF) or non-coding RNA (ncRNA) can up- or down-regulate other genes, therefore it is important to explicitly consider both positive and negative interactions. Although exists a variety of software and libraries for constructing and investigating such networks, none considers the sign of interaction. It is also required that the represented networks have high accuracy, where the interactions found have to be relevant and not found by chance or background noise. Another issue derived from building co-expression networks is the reproducibility of those. When constructing independent networks for the same phenotype, though, using different expression datasets, the output network can be remarkably distinct due to biological or technical noise in the data. However, most of the times the interest is not only to characterise a network but to compare its features to others. A series of questions arise from understanding phenotypes using co-expression networks: i) **how to construct highly accurate networks**; ii) **how to combine multiple networks derived from different platforms**; iii) **how to compare multiple networks**.

For answering those questions, i) I improved the weighted Topological Overlap (wTO) method to construct highly accurate networks, where now each interaction in a network receives a probability. This method showed to be much more efficient in finding correct interactions than other well-known methods; ii) I developed a method that is able to combine multiple networks into one building a Consensus Network (CN). This method enables the correction for background noise; iii) I developed a completely novel method for the comparison of multiple co-expression networks, Co-expression Differential Network Analysis (CoDiNA). This method identifies genes specific to at least one network. It is natural that after associating genes to phenotypes, an inference whether those genes are enriched for a particular disorder is needed. I also present here a tool, RichR, that enables enrichment analysis and background correction.

I applied the methods proposed here in two important studies. In the first one, the aim was to understand the neurogenesis process and how certain genes would affect it. The combination of the methods shown here pointed one particular TF, ZN787, as playing an important role in this process.

Moreover, the application of this toolset to networks derived from brain samples of individuals with cognitive disorders identified genes and network connections that are specific to certain disorders, but also found an overlap between neurodegenerative disorders and brain development and between evolutionary changes and psychological disorders. CoDiNA also pointed out that there are genes involved in those disorders that are not only human-specific.

Keywords: Co-expression networks; Networks Construction; Networks Comparison; Consensus Network; Human Cognition.

Contents

List of Figures

List of Tables

Acronyms

A	Adenine
AD	Alzheimer’s Disease
AGO2	Argonaute-2
ARACNe	Algorithm for the Reconstruction of Accurate Cellular Networks
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve

BD	Bipolar Disorder
BH	Benjamini-Hochberg adjusted $p-$value
bp	base pair

C	Cytosine
cDNA	complementary DNA
CER	Cerebellum
ChIP-seq	Chromatin Immunoprecipitation Followed by Sequencing
CMP	Chimpanzee
CN	Consensus Network
CoDiNA	Co-expression Differential Network Analysis
CRAN	Compreensive R Archive Network
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeat
CTR	Control

DBD	DNA-Binding Domain
DBP	DNA-Binding Protein
DE	Differential Expression
DisGeNET	Disease to Genes Network
DNA	Deoxyribonucleic Acid
DPI	Data Processing Inequality
dpi	days post induction

DrL	Distributed Recursive (Graph) Layout
ENCODE	Encyclopedia of DNA Elements
FPKM	Fragments Per Kilobase Million
G	Guanine
g2d	Gene-to-Disorder
GAN	Ganglia
GDA	Gene-Disease Associations
GEO	Gene Expression Omnibus
GLMM	Generalised Linear Mixed Model
GO	Gene Ontology
GRF	Gene Regulatory Factor
GS2D	Gene Set to Disease
GWAS	Genome Wide Association Study
HIP	Hippocampus
hiPSC	human induced Pluripotent Stem Cell
iNGN	inducible-Neurogenin Cell Line
iPSC	induced Pluripotent Stem Cell
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	Knock Out
LOESS	Locally Estimated Scatterplot Smoothing
LOWESS	Locally Weighted Scatterplot Smoothing
MAS5	MicroArray Suite 5
MDD	Major Depression Disorder
MI	Mutual Information
miRNA	micro RNA
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
nFC	normalised Fold Change

NGS	Next Generation Sequencing
NLM	National Library of Medicine
NLME	Nonlinear Mixed Effect Model
OE	Over Expressing
OMIM	Online Mendelian Inheritance in Man
OTU	Operational Taxonomic Unit
PCA	Principal Component Analysis
PD	Parkinson Disease
PFC	Prefrontal Cortex
PPI	Protein-Protein Interactions
PsyGeNET	Psychiatric disorders Gene association NETwork
ReViGO	Reduce + Visualise Gene Ontology
RH	Rhesus macaque
RIP-seq	RNA-Interaction Protein Immunoprecipitation and subsequent sequencing
RMA	Robust Multi-array Average
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
ROC	Receiver Operator Curve
RPKM	Reads Per Kilobase Million
rRNA	ribossomal RNA
SCZ	Schizophrenia
SimRel	Semantic similarity scores
SOM	Self Organising Maps
SPACE	Sparse Partial Correlation Estimation
T	Thymine
TC	Temporal Cortex
TF	Transcription Factor
TO	Topological Overlap
TOM	Topological Overlap Matrix
TPM	Transcripts Per Kilobase Million
tRNA	transfer RNA

U	Uracil
UTR	Untranslated Region
WGCNA	Weighted Gene Co-Expression Network Analysis
WT	Wild Type
wTO	weighted Topological Overlap

Symbols

B^*	Each bootstrap realisation
H_0	Null Hypothesis
H_a	Alternative Hypothesis
Δ	Euclidean distance
Δ^*	Penalised Euclidean distance
Δ^{**}	Normalised penalised Euclidean distance
$\Delta_{\widetilde{\rho}}$	Euclidean distance from the link weight to the $\widetilde{\rho}$
$\Omega_{i,j}$	The CN coefficient of node i and node j
Φ	Category in CoDiNA
α	One of the three Φ categories. It means that a particular link or node is common to all conditions under comparison
β	One of the three Φ categories. It means that a particular link or node has a different sign in at least one condition under comparison
δ	Maximal distance from the original wTO accepted
γ	One of the three Φ categories. It means that a particular link or node does not exist in all conditions under comparison
$\mathbb{A}$	Adjacency matrix
$\mathbb{B}$	The bipartide network for gene and diseases
$\mathbb{D}$	The adjacency matrix for diseases
$\mathbb{D}^*$	The normalised adjacency matrix for diseases
$\mathbb{L}$	A set of links in a network. $\mathbb{L} = L_1, \ldots, L_l$
$\mathbb{N}$	A set of nodes. $\mathbb{N} = N_1, \ldots, N_n$
$\mathbb{W}$	A set of networks. $\mathbb{W} = (W_1, \ldots, W_w)$
$\omega_{i,j}$	The wTO coefficient of node i and node j
$\omega^*_{i,j}$	Each $\omega_{i,j}$ generated in a bootstrap
$\rho_{i,j}$	Correlation of nodes i and j
τ	Minimum absolute value for slicing the $\rho_{ij,k}$ into $\widetilde{\rho}_{ij,k}$
$\widetilde{\Phi}$	Subcategory in CoDiNA
$\widetilde{\rho}_{ij,k}$	Categorical correlation of nodes i and j in network k

i	Index on nodes
j	Index on nodes
k	Index on networks
l	Number of links in a network
m	The size of a sample
n	Number of nodes in a network
o	Index on diseases
p	Index on diseases
u	Index on nodes
w	Number of networks in a set

1

Introduction

COMPLEX-SYSTEMS AND PHENOTYPES can be better understood by using co-expression network analysis. In a co-expression network, genes are represented as nodes and two genes are connected if they have a significant correlation on its expression. In this approach the interest lies in pinpointing how two genes share co-expression patterns, by doing so, we are allowed to characterise the network of a particular phenotype. Although network approaches became central for understanding those phenotypes, methods for constructing highly accurate networks are still lacking.

Unfortunately, the construction of such networks depends on the data used as input. This reliance on data results in networks that are not reproducible, and therefore, might not represent the real gene-gene interactions in a particular system. Therefore, methods that enable the construction of a network that combines multiple networks, and hence multiple datasets can result in a more informative and reproducible network.

Finally, in most of the studies, the interest does not lie in characterising one complex-system but in comparing it under different systems. However, no method is available for comparing multiple networks and classifying interactions and genes according to its presence in the conditions under investigation. Finding genes that are different or specific to certain conditions can be used as signature genes that can allow for better treatments and diagnosis in complex disorders, such as, mental illness.

I applied those methods in two studies. The first study concerns about understanding the neurogenesis process. The interest lied in characterising new genes that highly involved in that process. In a second study, the aim was to find signature genes for mental disorders and understand if there is a co-regulation pattern of those disorders that are similar to infants or other primates.

1.1 Organisation of this PhD dissertation

This PhD dissertation is organised in three parts. In the first part, two chapters comprise a short introduction to the underlying biology of the transcriptomics world (Chapter 2) and an introduction to network theory that are used to model gene interactions derived from co-expression analysis (Chapter 3). Few definitions that did not completely fit in the text can be found in the Glossary.

The second part of this dissertation contains the three methods I developed. All three methods are available as R packages and can be easily downloaded from the Compreensive R Archive Network (CRAN). The vignettes are available inside the package and a detailed manual can be found on my web page https://deisygysi.github.io/rpackages/ and in Appendix A, Appendix B. Chapter 4 refers to the wTO approach for constructing networks and the CN, implemented in the wTO R package. This package is able to construct highly accurate co-expression networks from expression and co-occurrence networks from abundance data in metagenomics, for independent samples, repeated measures and time series. Just by characterising one co-expression network one is allowed to better comprehend the phenotype under study. However, often the interest lies in the comparison of at least two networks and understanding its similarities, differences and specificity. Therefore, a method that allows for these comparisons had to be developed and it is described in Chapter 5, my second methodology and R package CoDiNA. Chapter 6 concerns about a gene enrichment method, implemented in a package, RichR, that given a list of genes and a list of disorders allows for understanding if a particular disorder is enriched in the dataset. It is particularly useful in a combination with CoDiNA for understanding the patterns of differential or specific co-expression.

The third part of this dissertation focus on the application of the methods. In the first application (Chapter 7), the objective is to describe and understand how the micro RNA 124 (miRNA-124) can shape the neurogenesis process. For that, induced Pluripotent Stem Cell (iPSC) cells wild-type and knocked out for this micro RNA (miRNA) were followed for 4 days, I build the networks using the wTO package. Later, I applied CoDiNA for the identification of specific genes and one gene pointed out by CoDiNA was over-expressed in wild-type cells and followed for 14 days. The cells had higher levels of apoptosis and were not perfectly functional neurons. In a second application (Chapter 8), I used a combination of the three methods I developed to understand how cognitive disorders might be related. In the first moment, the interest lies in neurodevelopmental disorders such as Parkinson's and Alzheimer's diseases. Later, I am interested in pinpointing similarities and specificities of other mental illness such as Autism Spectrum Disorder, Bipolar Disorder, Major Depression Disorder and Schizophrenia. In the last part of this chapter, I want to understand the patterns of co-regulation of the Prefrontal Cortex (PFC) that kept constant in the human lineage when compared to other primates.

The conclusion and further perspectives are presented on Chapter 9 and focus on the power of the toolset developed in this PhD project; it also shows other applications of the methods in the following studies: HIV and Tuberculosis co-infection; Glioma and its subkinds; a metagenomics time series study of the fungi environment in a marine ecosystem.

1.2 List of publications

Additional to the three R packages that were developed in this PhD project, this work also led to two published research articles, four research articles are under preparation one review article. Moreover, I also had two talks presented in conferences, five posters presented at scientific conferences, all listed below.

Part of the literature review presented in Chapter 3 is part of the invited headline review under preparation.

- **D. M. Gysi**, T. M. Fragoso and K. Nowick. "Construction, comparison and evolution of networks in biology, social sciences, economy, and humanities - or: what can we learn from other disciplines". *Journal of Royal Society Interface.* Invited headline review. 2019.

The `wTO` R package, described in Chapter 4, is public available in CRAN, `https://cran.r-project.org/package=wTO`. The well-described manual can be found in Appendix A and at `https://deisygysi.github.io/rpackages/Pack-1`. A publication describing the methodology is available.

- **D. M. Gysi**, A. Voigt, T. M. Fragoso, E. Almaas and K. Nowick. "wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool". *BMC Bioinformatics*, 19(1), 392, 2018. `https://doi.org/10.1186/s12859-018-2351-7`.

The `CoDiNA` R package, described in Chapter 5, is public available at CRAN, `https://cran.r-project.org/web/packages/CoDiNA`, a well described manual can be found Appendix B and `https://deisygysi.github.io/rpackages/Pack-2`. A publication is submitted.

- **D. M. Gysi**, T. M. Fragoso, V. Busskamp, E. Almaas and K. Nowick. "Co-expression Differential Network Analysis: How to compare multiple networks simultaneously?". *Submitted.* 2019.

The `RichR` R package is public available at `https://cran.r-project.org/web/packages/RichR` and a publication is under preparation.

- **D. M. Gysi** and K. Nowick. "Make me RichR: an R package for gene-disease enrichment". *In preparation.* 2019.

The application of the previous methodologies generated three manuscripts. The neurogenesis study, Chapter 7, is published.

- L. K. Kutsche*, **D. M. Gysi***, J. Fallmann, K. Lenk, R. Petri, A. Swiersy, S. D. Klapper, K. Pircs, S. Khattak, P. F. Stadler, J. Jakobsson, K. Nowick, and V. Busskamp. "Combined experimental and system-level analyses reveal the complex regulatory network of miR-124 during human neurogenesis". *Cell Systems*, 7(4), 438–452, 2018.
  *Both authors contributed equally.

The manuscripts that concern applications of the methods in the cognition network and the marine fungi ecosystem are under preparation.

- **D. M. Gysi** and K. Nowick. "Evolution of gene-co-expression networks implicated in cognitive functions in primates. *In preparation*. 2019.

- S. Banos, **D. M. Gysi**, T. Richter-Heitmann, K. Nowick, M. Friedrich, F. O. Glöckner, M. Boersma, K. H. Wiltshire, A. Wichels, G. Gerdts and M. Reich. "Dynamics observed in a pelagic marine fungal community: an interplay of oscillation types, stability, resilient, and biotic interactions". *In preparation*, 2019.

Another collaboration is also under preparation.

- A. Geffre, T. Gernat, A. Toth, G. Robinson, B. Bonning, A. Hamilton, B. Jones, **D. M. Gysi** and A. Dolezal. "Pathogen manipulation in the Anthropocene: Viruses of managed honey bees alter host social behaviour". *In preparation*, 2019.

A list of the posters and oral presentations given is shown below.

1. **D. M. Gysi**, T. M. Fragoso, E. Almaas and K. Nowick. "Co-expression Differential Network Analysis". *XXIX International Biometric Conference*, 2018.

2. **D. M. Gysi**, T. M. Fragoso, E. Almaas and K. Nowick. "Comparing multiple co-expression networks". *4th Summer School in Complex Networks*, 2018.

3. **D. M. Gysi**, T. M. Fragoso, E. Almaas and K. Nowick. "How to build and compare co-expression networks". *BenGenDiv*, Berlin, 2018.

4. **D. M. Gysi**, A. Voigt, T. M. Fragoso, E. Almaas and K. Nowick. "An R package for calculating the Weighted Topological Overlap Network with a visualization tool", *CompleNet'*18, Boston, 2018.

5. **Gysi, D. M.**, A. Voigt, T. M. Fragoso, E. Almaas and K. Nowick. "wTO, an R package for computing the weighted Topological Overlap and Consensus Networks". *NORBIS annual conference*, Trømso, 2017.

6. L. K. Kutsche, **D. M. Gysi**, K. Lenk, R. Petri, J. Jakobsson, K. Nowick and V. Busskamp. "A systems level view on miR-124 function during neuronal differentiation from human iPS cells". *Intelligent Systems for Molecular Biology*, Prague, 2017.

7. L. K. Kutsche, **D. M. Gysi**, K. Lenk, R. Petri, J. Jakobsson, K. Nowick and V. Busskamp. “A systems level view on miR-124 function during neuronal differentiation from human iPS cells”. *Gene regulatory mechanisms in neural fate decisions*, San Juan de Alicante, 2017.

I

Elements of transcriptomics and network theory

2

An introduction to transcriptomics

"Begin at the beginning, " the King said, gravely, "and go on till you come to an end; then stop."

– Lewis Carroll, *Alice in Wonderland*

2.1 From cells to genes

All living organisms consist of cells. Like us, they can grow, reproduce, respond to stimuli and at some point, die. Organisms can have only one cell that is able to cope with their daily life or can be multicellular, like us, humans, where their cells re-organise in particular ways that become specialists in a few tasks. For example, blood is composed of many cell types: erythrocytes (red blood cells) are responsible for transporting the oxygen; leukocytes (also known as white cells) are part of the immune system and are important for fighting infections, and platelets play an important role in clotting the blood at a wound.

What makes those cells different and able to perform well in different tasks are the distinct set of proteins they express. Proteins serve as structural components in a cell and can change its shape according to temperature, ion concentrations and other stimuli and as a consequence, act as a sensor. Proteins can be responsible for what comes in and leave a cell; can act as catalysts allowing reactions to occur or enzymes that make those reactions faster. Another important function of a protein is that it acts as a switch by turning a gene on or off. I will discuss more this function in Section 2.2.

The information associated with each protein function, when or how it should be produced is stored in a long polymer known as Deoxyribonucleic Acid (DNA). This polymer's three-dimensional structure is a double helix consisting of two anti-parallel and reverse complementary strands, each having $5'$ and $3'$ ends that coil around a common axis. The

DNA is constituted of nucleotides, often referred to as bases, because they contain cyclic organic bases. These bases are commonly abbreviated as Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) and are attached to the DNA structure by a covalent bond and projected out of the helical backbone. Each one of those nucleotides binds to another nucleotide of the other strand by a hydrogen bond in a base pairing that follows a general rule, A pairs with T and C with G. This complementary matching is extremely strong, and when this bound is broken, they spontaneously bind back, under certain biochemical conditions. This strong hybridisation process is useful for detecting one strand on the absence of the other. The DNA is able to copy itself in a process called **replication**. This process is important to keep the exact same information in all cells.

The DNA structures itself in the way that its information is linearly stored into discrete *functional* units called genes. A gene consists of DNA sequence and can be a recipe for constructing proteins, however, not all genes code for proteins. Genes vary a lot in size, in humans, they can be few hundreds DNA base pairs (bps) to more than 2 million bps. The Human Genome Project estimated that humans have around $20,000$ and $25,000$ genes. The DNA can be found inside the nucleus of a cell, it complexes with a protein set called **histone** to form a **nucleosome**. Another histone, the H1 histone, binds on the outside of the nucleosome and forms a **chromatosome**, that folds into itself to make an extra tight folding that results in a **chromosome**.

Two coupled processes are used to transform the stored information inside the DNA into proteins, together they are called gene expression. The first part of this process is called **transcription** and the second **translation**. This two process coupled with the replication of the DNA is so important that is commonly referred to as the **central dogma of biology** and it is represented in Figure 2.1.

In the transcription process, the coding part of a gene is copied into a Ribonucleic Acid (RNA). A large enzyme, RNA polymerase, binds to the DNA and uses it as a template to catalyse the nucleotides' linkages into an RNA molecule.

An RNA strand is similar to the complementary $3' \rightarrow 5'$ DNA, however, it is a single strand version and does not contain Timine (T) but Uracil (U) instead. The RNA has different functions than the DNA, one of these functions, is to store short copies of parts of DNA for a small time frame. In eukaryotes, this initial RNA is processed into a smaller version, called messenger RNA (mRNA), that carries the information out of the nucleus to the cytoplasm.

Translation, the second process starts in the cytoplasm. There is a molecular machinery, part protein and part ribossomal RNA (rRNA), called ribosome decodes the mRNA into an amino acid chain, also called polypeptide. Every three nucleotides in the RNA form a codon. The ribosome decodes the information by binding the complementary transfer RNA (tRNA) anticodon sequences to mRNA codons. The other side of the tRNA has a specific amino acid according to a general rule called genetic code (Figure 2.2). Every amino acid can be generated by multiple codes, but one code can generate one amino acid, it means that it is redundant, but not ambiguous, hence the genetic code is said to be degenerate. The process of binding an amino acid into each other is executed until a stop codon is reached; at this moment the ribosome releases the polypeptide into the cytoplasm, that can be later processed into a functional protein.

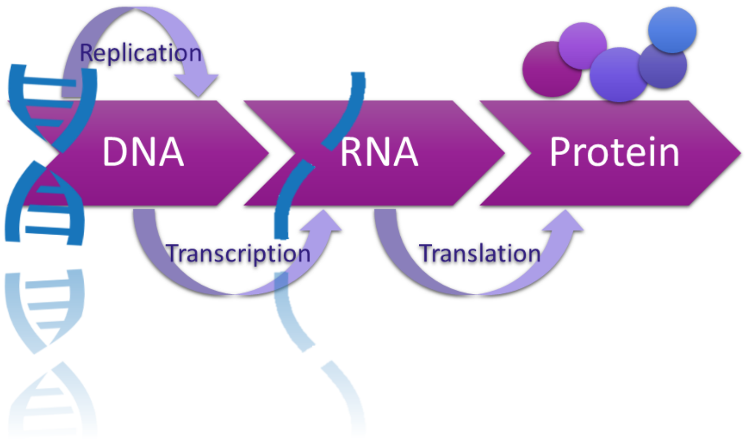


FIGURE 2.1: Central dogma of biology, from DNA to proteins: DNA is copied into DNA by the replication process. The information contained in the DNA is copied into an RNA by the transcription process, and the RNA is decrypted into proteins by the translation process.

2.2 Gene Regulation

Even though our cells are specialised in different functions, we have exactly the same collection of genes, the genome, in almost all our cells. However, not all genes are being expressed throughout the whole organism, in all its cells or at the same time. Hence, the organisms know **when** and **where** particular genes have to be transcribed. The set of genes that are active or inactive in each cell is different according to their functions, and most of the cells can change its response according to external signals or different conditions. It means that the genes that are active or inactive depend not only on the cell type but also on its environment or time-point. For example, the brain is composed by different types of cells such as neurons, glia, astrocytes, oligodendrocytes, etc and they arrange in diverse ways inside the brain, creating areas that are responsible for specific functions. Moreover, the gene expression of a particular brain area of an infant is different from the same area in an adult.

The control of gene activity depends on Transcription Factors (TFs). They are DNA-Binding Protein (DBP) that, as the name says, are proteins that bind to the DNA and can activate or repress the transcription of particular genes. These proteins contain a domain, DNA-Binding Domain (DBD), that is specifically designed for a set of genes, instead of recognising all the genes. This domains recognise a short specific DNA sequence containing around $6-12$ bps. Those sequences are specific and do not occur often in the DNA, assuring the TF specificity to activation or repression. Proteins that are essential to the gene regulatory process, but do not have a DBD (such as coactivators, chromatin remodellers, histone acetyltransferases, histone deacetylases, kinases and methylases) are not TF. TFs can work alone

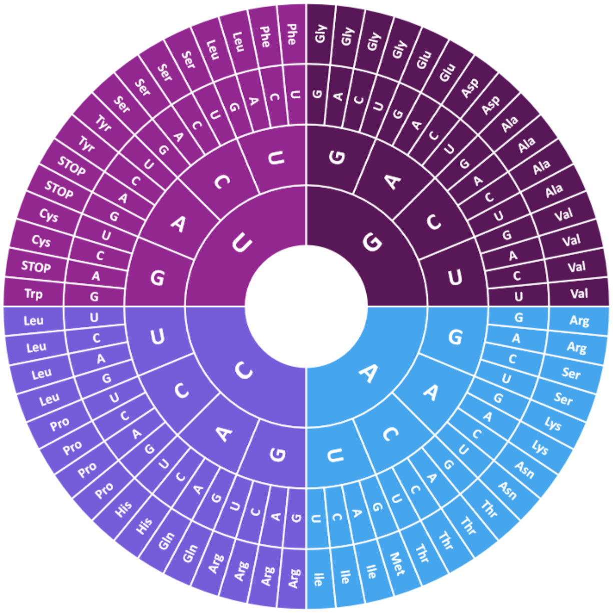


FIGURE 2.2: Genetic code, from nucleotides to amino acids:. Starting from the centre of the circle and working to the outermost circle, this figure represents which amino acid's are encoded by which codon.

or be combined into multiprotein complexes to mediate the gene regulatory process.

A **promoter** is a DNA sequence where the RNA polymerase binds and starts the transcription of a gene. Oftentimes, the promoter is located far away from where the TF binds, this regulatory sites can be tens of thousands of base pairs either upstream (opposite to the direction of transcription) or downstream (in the same direction as transcription) from the promoter. The regulation of a single promoter can be a result of a series of TFs allowing, as a consequence, complex control of gene expression.

This complex control can be formed by the promoter and a series of control elements, located near transcription start sites and by sequences located far from the regulated genes. Such elements can be: **TATA box** (a conserved sequence found around 25 to 35 bp upstream of the start site) or **initiators** (unconserved sequence containing a C at the -1 position and a A at $+1$). The genes that are regulated using the TATA box or initiator have a well-defined start and an end. However, this is not the case for all protein-coding genes. Many of these genes can start its transcription at any possible site over an extended region, often in a range of $20-200$ bp. Genes with this set-up result in mRNAs with multiple $5'$ alternative ends and do not have an initiator a TATA box, instead, they contain a region $(20-50$ bp around 100 bp upstream the start-site region) rich in the bases G and C. The dinucleotide GC content is underrepresented in vertebrates, therefore regions rich in GC, also called CpG island, upstream a start-site can suggest a transcription-initiation region. The *housekeeping genes*, genes required for maintenance of basic cellular function, are normally controlled by CpG islands.

Two general mechanisms are associated with the activation or repression of the regulatory process of the gene expression. In the first process, the regulatory proteins act along with other proteins and modulate the chromatin structure, the second influences the ability of general TF to bind to promoters.

Nevertheless, one TF alone binding to the DNA is, often, not enough to infer direct functional effects on the levels of the gene expression of genes that are close to each other, that are under control of multiple TFs simultaneously[1]. One way to measure the interaction of TFs with binding sites is by using Chromatin Immunoprecipitation Followed by Sequencing (ChIP-seq)[2–4]. This is a technique that uses an antibody that recognises either a TFs or the histone modification for identifying binding locations by pulling down attached DNA.

2.3 Measuring gene expression

The genome is static and is exactly the same across almost all the cells. Opposite to it, the transcriptome is extremely dynamic. A **transcriptome** or **expression profile** is the collection of genes expressed or transcribed from the DNA and it is the major determinant of the cellular function and its phenotype. Hence, differences in the gene expression are responsible for morphological and phenotypic responses to different environmental stimuli, perturbations. The transcriptome changes vastly according to the cell function and environment, it is also highly dynamic and can change extremely fast to quickly respond to the environmental conditions[5–8].

How much of a gene is being transcribed can be measured by the amount of its mRNA in the cell[9]. Two main procedures can be used for it. The former is done using microarrays and the latter by sequencing the mRNA called RNA sequencing (RNA-seq).

By understanding the expression levels of genes in particular conditions or when and where that gene is (over or under) expressed, we can comprehend more about its function[8]. Moreover, the joint changes of the co-expression of a set of genes can shed light on the regulatory mechanisms, broader functions inside the cell and consequently biochemical pathways. I will explore more about it on Section 2.5 and Section 3.3. And in Subsection 2.3.1 and Subsection 2.3.2 I will cover how the expression is measured and how to pre-process and access the quality control of those data.

2.3.1 Microarrays

Microarrays or gene chips are, as the name says, chips with thousands of tiny spots in predefined positions, with each spot containing a known DNA sequence or gene. The probes, DNA molecules attached to each spot, can hybridise to a particular transcript (or a set of mRNAs)[10]. There are three main kinds of microarrays: the two-channels, one-channel and the oligonucleotide. In general, when performing a two-channel microarray analysis, the mRNA molecules are collected from a reference and treatment samples. Both mRNA samples are converted into complementary DNA (cDNA), and different fluorophores are used for labelling each probe. It means, for example, the treatment samples cDNA can be labelled with a red fluorescent dye (Cy5), while the reference, green (Cy3). Both samples are then mixed together and the cDNA molecules hybridise to the microarray. Later, the chip is scanned to measure the expression of each gene contained in the microarray. The strength of the signal from a spot (gene) depends on the amount of target sample binding to the probes present on that spot. In our example, when the expression of a particular gene is higher in the experimental sample than in the reference sample the corresponding spot on the microarray appears in red. However, if the expression in the reference sample is higher than in the experimental sample, the spot appears green. While in a case where the two samples have the same expression the spot turns yellow[8;10;11]. To identify up-regulated and down-regulated genes, the relative intensity of each fluorophore can be used in a ratio-based analysis[12].

The one-colour microarrays provide the intensity value for each probe (or probe set) that indicates the hybridisation level of a labelled target. However, the true abundance level is not indicated by this value, but a relative abundance when contrasted to other samples or conditions when processed in the same experiment. Therefore, the comparison of two conditions requires two separated single-dye hybridisation. On the other hand, the usage of one-channel arrays also have some strength, for example, i) a sample that has a bad quality does not affect the raw data of the other samples, opposed to the two channels, where one low-quality sample can reduce the quality a whole array; ii) data from different experiments can be directly compared (when it is accounted for batch effects)[13].

The last array type is the oligonucleotide. This array oftentimes carries probes that hybridise with RNA spike-ins and the target probes are normalised by the hybridisation of control probes. It also measures the absolute level of gene expression across different spots and can be corrected by a normalisation within samples and between samples[14].

None of the microarrays directly measures the amount of mRNA, but the fluorescence from the hybridisation of the cDNA to the probe DNA in the array. Oftentimes the probes are not specific to one gene, but a set of genes. Therefore, this kind of data carries background noise and often the expression values are representative of a set of genes instead of one gene. Because of that, adequate quality control and filtering of the data is extremely important.

The quality control for microarrays follows different steps according to the array type, platform, company etc. where the expression was measured. Independent of the platform, the first step is to remove the background noise, followed by some normalisation on the data. In the background correction, the background intensity of each spot is removed. This can be done using a local subtraction, that estimates the background for each colour and removes it from the foreground or a model based correction, where different (linear) models can compute the signal component and a noise component. The normalisation removes systematic variations of the microarray that affects the expression levels of the measured gene. The normalisation can be simple as a i) quantile normalisation, that aims to correct the variation between the arrays and make the samples comparable; ii) probe normalisation that aims to correct for the variation within probe sets and it equalises the behaviour of the probes between the arrays. More sophisticated models exist such as z-ratio, Locally Estimated Scatterplot Smoothing (LOESS), Locally Weighted Scatterplot Smoothing (LOWESS)[15–17], Robust Multi-array Average (RMA) and MicroArray Suite 5 (MAS5). The two last ones are more robust and therefore, the focus here.

The RMA uses a model to normalise the whole chip, removes all probes that do not completely hybridise and returns a $\log_2$ intensity value. The MAS5 normalises each array separated and sequentially using all the data (even mismatched probes) and returns a robust average. Also, computes a matrix where allows the users to filter for probes that are present, marginal or absent.

In the statistical environment `R`[18] plenty of packages are available for dealing with microarrays. The packages `affy`[19] or `oligo`[20] can be used for an Affymetrix array and `lumi`[21–23] for an Illumina array.

2.3.2 RNA-seq

Because the microarrays contain noise and also the DNA sequence has to be known prior to the hybridisation, better methods had to be developed. The RNA-seq is a Next Generation Sequencing (NGS) technique, where the whole mRNA is converted into cDNA and then sequenced using different platforms, that can be chosen accordingly to the researcher goals[24–26], such as Roche 454, Illumina, Helicos and PacBio (Pacific Biosciences) that uses a DNA polymerase to drive their sequencing reaction or SOLiD (Life Technologies) and Complete Genomics use a DNA ligase. Those platforms can be categorised as single molecule-based (as the name says, that sequence a single molecule, such as Helicos and PacBio) or ensemble-based (that sequence multiple identical copies of a DNA molecule, such as Illumina and SOLiD)[26].

Independent of the platform, the resulting data are often in FASTQ format. This format contains a number that identifies each read, the read sequence and the quality score for each base. In order to get informative data that can be used for further analysis, two main preprocessing steps are often performed. The first step includes removing artefacts (such as sequence adaptor), low complexity reads and check for contamination on the samples. Public tools such as cutadapt[27], Quake[28], SeqTrim[29], TagDust[30] etc. can be used for solving these issues. The sequencing errors can be removed in the intermediate step, based on the quality score of each base, also, small reads should be removed. The tool FASTQC[31] can be used in order to solve this problem.

The next step is the alignment of the processed reads to a reference genome using an appropriate aligner tool prior to performing the downstream analysis. The choice of an appropriate tool depends on the RNA-seq platform, the application and the organism under study. Most of the commercial platforms have their own series of tools and analysis (pipeline). The most used alignment tools are Bowtie[32], TopHat[33], Segemehl[34–36] and STAR[37]. After that, for a gene expression analysis, it is required to count the amount of mRNA that was mapped to each gene. This can be done using rnacounter[38], for example. The amount of each mRNA counted per gene can be directly used for an analysis where the interest lies on understanding (and pinpointing) which gene has a differential expression in two different conditions (or tissues). Most of the times a normalisation on the data is needed. The most common normalisation used in RNA-seq data are: Reads Per Kilobase Million (RPKM), Fragments Per Kilobase Million (FPKM) and Transcripts Per Kilobase Million (TPM). Another method, Kallisto, uses a pseudo-alignment, and is able to quantify and normalise RNA-seq much faster than alignment methods[39].

2.4 Gene Differential Expression

In many studies, the interest lies in understanding the difference between the gene expression levels of two or more conditions. In a Differential Expression (DE) analysis, the interest lies on associating genes to a phenotype using statistical modelling. The models to be used depends on the kind of data and, of course, the nature of the trait under study. In a nutshell, if the probability function of the expression data is normally distributed (after normalising microarray data, for example), one can use linear models, when the probability function of the expression data follows a Negative Binomial distribution (in RNA-seq experiments) specific models can be used. Other set-up models can be used to solve it, for example, Generalised Linear Mixed Model (GLMM) or Nonlinear Mixed Effect Model (NLME). Although, it is recognised that one gene is not responsible for a phenotype, rather a collection of genes acting together might shape a phenotype, DE is still widely used to pinpoint associations.

2.5 Gene co-expression networks

Gene co-expression networks can be used for understanding *how* and *which* gene interactions are involved in a particular system. Those networks have received much attention[40;41] because they can shed light on the molecular mechanisms that underlie biological processes or on disease associations of gene sets. Those networks enabled the identification of gene functions and refine its annotation in plants such as cotton[42], maize[43] and papaya[44]. They also added knowledge on complex disorders, such as Diabetes[45], Major Depressive Disorder[46;47], Bipolar Disorder[48], Schizophrenia[49–51], and revealed many important hub-genes in cancers[40;52;53]. They were used to comprehend how expression patterns in primate brains have changed over the course of evolution[54–56] and how soil treatments can affect the rice production[57]. Co-expression networks were also applied for providing insights into the neurogenesis process in humans[58] and tissue-specific regulatory processes[59–61].

In general, a network that is not from molecular data are often able to be observed, and therefore, the weights and interactions are deterministic, meanwhile, in molecular biology, most of the interactions are stochastic and indirectly measured, hence, associated to a higher level of uncertainty. This might be a noise intrinsic to the network and should be filtered out. For example, in a co-expression network, the weight of interaction is based on estimating a correlation of the gene expression. Therefore, carries more noise than if the interaction could be directly measured. The reasons for that noise are manifold and include biological differences of the individuals such as gender, age or ethnic group, demographic or technical differences such as array-platform, data quality and facilities that performed the analysis. Hence, co-expression networks need correction approaches to reduce noise, such as a consensus network, that combines multiple independent networks[54;62].

A series of questions arise from understanding systems using co-expression networks: i) **how to build highly accurate networks**; ii) **how to combine multiple networks derived from different platforms**; iii) **how to compare multiple networks**. But, before we are able to answer them, I will take a journey into network science on Chapter 3.

3
Gene interactions using a network approach

"Although this detail has no connection whatever with the real substance of what we are about to relate, it will not be superfluous, if merely for the sake of exactness in all points."

– Les Misérables, *Victor Hugo*

NETWORKS ARE BROADLY employed in many fields: from understanding *how friends connect in a party* to *how animals relate to each other*, *how super-heroes appear in the same comic books* to *how genes can be associated to the same biological process.* Network analyses are especially beneficial for understanding complex-systems in any research field. Examples of complex biological or medical systems include gene regulatory, ecological and psychometry networks. Social networks can include scientific collaborations, relationships between actors, sexual partnerships and also can explore the relationship of social insects. Gastronomy is also employing network analysis tools, for instance, in recipe building towards the *perfect recipe*, a recipe that tastes extremely good. In finance, the interest in using networks often lies in predicting economic crises. Most networks are not static, and they can differ from each other, for example, how genes connect in a particular disorder is different than in controls, or how ingredients pair in a specific culture is different than in another. Moreover, networks can change over time, be it within minutes, for instance, when the cell reacts to an environmental change or the stock market to the introduction of new company assets, or within years, such as over the course of a lifetime or evolution.

In this chapter, I will give an introduction to the network theory and its applications in biology. This chapter is based on my review, where a much broad application of networks in other fields is available.

- **D. M. Gysi**, T. M. Fragoso and K. Nowick. "Construction, comparison and evolution of networks in biology, social sciences, economy, and humanities - or: what can we learn from other disciplines". *Journal of Royal Society Interface*. Invited headline review. 2019.

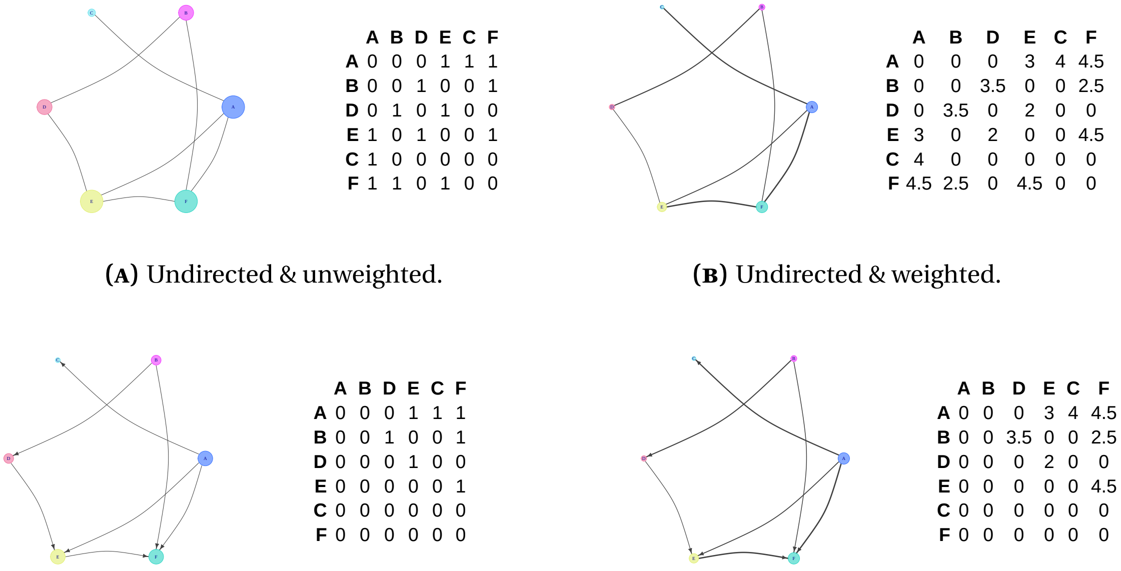
3.1 Basic definitions on networks

The set of interactions among a set of entities is, in general, called a **graph** or a **network**[63;64]. In graph theory, each entity is called a **vertex**, while in network notation, it is called a **node**[63]. Accordingly, the connections between two entities are called **edge** or **link**, respectively[63]. In this dissertation I will always refer to them using the **network notation**, unless required otherwise (data formats, for example). The total number of nodes in a network will be denoted as n and the number of links in a network will be denoted as l. While nodes can receive a label, links in general, are not labelled[63] (although, in many cases, weights can also be perceived as a label and De Bruijn graphs are labelled).

A network can be represented i) mathematically as an adjacency matrix (usually denoted as $\mathbb{A}$) or an edge list; or ii) visually as a graph (Figure 3.1). Links of a network can possess a direction (normally depicted by an arrow), which indicates that the interaction is asymmetric, e.g. one gene is regulating another gene or a person follows somebody else in a social network. Networks with directed links are called **directed networks**, while networks without directed interactions or in which the direction is not known are referred to as **undirected networks**, e.g. collaboration in the same paper or interactions between proteins. The links can also have a weight to express the strength of the interaction, which results in a weighted network[63;64]. Usually, the weight is graphically displayed as the thickness or the length of the links.

A network is a pair $G = \{\mathbb{N}, \mathbb{L}\}$ of a set $\mathbb{N}$ of nodes connected by a set $\mathbb{L}$ of links. A link can have a **weight**, the weight is a measure of how strong a particular interaction is[65], a link can also have a *direction*, that specifies the source (starting point) and a target (endpoint) where the interaction occurs[66]. Two nodes are considered to be **neighbours** if they are connected. The **degree** of a node a is the number of nodes it interacts with[65] and the **strength** of a node is the sum of the weights attached to links belonging to a node[67]. **Hubs** are nodes with a much larger degree compared with the average degree value[67], those are nodes, that in general are important to keep the topology of a network. A set of highly interconnected nodes is a **module** or **cluster**[68], the **clustering coefficient** describes the degree with which a node is connected to all its neighbours[69] and the **global clustering coefficient** measures the total number of triangles in a network[63]. The average clustering coefficient, as the name says, is the average of the clustering coefficient of all nodes in a network[69]. Two nodes are connected, in a network, if a sequence of adjacent nodes, a path, connects them[69], the **shortest path length** is the number of edges along the shortest path connecting them[69], the **average path length** is the average of the shortest paths between all pairs of nodes[69] and the **diameter** is the maximum distance between two nodes[65].

A **bipartite network** is a network where the nodes can be divided into two disjoint sets of nodes such that links connect nodes from the two sets to each other, but never inside the same set[63]; The Topological Overlap (TO) is a measure of how interconnected two nodes are based on common neighbours[68;70], details are given in Chapter 4. In general, **Global measures** are measures that describe the whole network, for example, degree distribution; average clustering coefficient; path length; modularity index. The **Local measures** are characteristics of individual nodes of a network, such as their degree and centrality.



	A	B	D	E	C	F
A	0	0	0	1	1	1
B	0	0	1	0	0	1
D	0	1	0	1	0	0
E	1	0	1	0	0	1
C	1	0	0	0	0	0
F	1	1	0	1	0	0

(A) Undirected & unweighted.

	A	B	D	E	C	F
A	0	0	0	3	4	4.5
B	0	0	3.5	0	0	2.5
D	0	3.5	0	2	0	0
E	3	0	2	0	0	4.5
C	4	0	0	0	0	0
F	4.5	2.5	0	4.5	0	0

(B) Undirected & weighted.

	A	B	D	E	C	F
A	0	0	0	1	1	1
B	0	0	1	0	0	1
D	0	0	0	1	0	0
E	0	0	0	0	0	1
C	0	0	0	0	0	0
F	0	0	0	0	0	0

(C) Directed & unweighted.

	A	B	D	E	C	F
A	0	0	0	3	4	4.5
B	0	0	3.5	0	0	2.5
D	0	0	0	2	0	0
E	0	0	0	0	0	4.5
C	0	0	0	0	0	0
F	0	0	0	0	0	0

(D) Directed & weighted.

FIGURE 3.1: Graph example: Four networks are represented here beside its adjacency matrix. All networks contain 6 nodes and 7 links. Networks on the left side are unweighted and on the right side are weighted. The width of the links is proportional to the weight of the links. Networks on the top are undirected and on the bottom directed. The arrows represent the direction of the interaction.

3.2 Notation

The notation used throughout the whole thesis is presented here. I will refer to it whenever I use it. Let $\mathbb{N}$ be a set of nodes $\mathbb{N} = \{N_1, \ldots, N_n\}$. The indices in the nodes in this thesis are denoted by i, j or u. The total number of nodes in a network is n. Also, let $\mathbb{L}$ be the set of links $\mathbb{L} = \{L_1, \ldots, L_l\}$ and l the number of links in $\mathbb{L}$. Moreover, let $\mathbb{W}$ be a set of independent networks $\mathbb{W} = \{W_1, \ldots, W_w\}$. The index for the networks is k and the total number of networks is w. $\mathbb{A}$ is an adjacency matrix. Constructed using a correlation measure $\rho_{i,j}$ from a set of measures in a sample of size m.

A bipartide network adjacency matrix is represented here by $\mathbb{B}$. A Disease weighted network that is constructed from the $\mathbb{B}$ network is defined as $\mathbb{D}$ and its index are represented by o and p.

3.3 The use of networks in biological sciences

Recent applications of complex network analysis methods have provided important new knowledge of the function and interactions of genes at the systems level[69;71–73]. Favourites amongst biologists include Protein-Protein Interactions (PPI) networks[74;75], metabolic networks[76;77], and co-expression networks[44;50].

In **PPI** networks, the nodes represent proteins and they are connected by a link if they physically interact with each other[78]. Typically, these interactions are measured experimentally, for instance with Yeast-Two-Hybrid systems, but interactions can also be inferred computationally based on sequence similarity[79]. PPI can be used to infer gene functions and the association of sub-networks to diseases[74]. Gene duplication and sequence divergence can shape such networks over time. For example, the PPI network of *Saccharomyces cerevisae* showed that the evolutionary older a particular protein is, the more connections it has with time[80]. Interaction networks of some eukaryotic Transcription Factors (TFs) became more and more complex due to the duplication of genes encoding for TFs of these networks[81]. Results of both studies can be related to the phenomena known as preferential attachment, which states that nodes that already have many links will attract more new links over time than other nodes[82]. Duplication of single genes or of whole genomes has also been proposed to be the driving force of gene regulatory networks[83;84]. The preferential attachment has also been observed in metabolic networks[85]. This process can directly affect the formation of more complex protein structures and pathways and lead to the evolution of more complex organisms. In a metabolic network the nodes describe the metabolites (biomolecules) and the links represents enzymes (proteins) that are able to catalyse a biochemical reaction[86]. This network type contains the stoichiometry of the reactions necessary for the synthesis and degradation of basic metabolites or complex compounds such as proteins[87;88].

To describe metabolic processes, **metabolic networks** have proved to be valuable. In a metabolic network, the nodes describe the metabolites (biomolecules) and the links represent enzymes (proteins) that are able to catalyse a biochemical reaction[86]. These networks contain the stoichiometry of reactions necessary for the synthesis and degradation of basic metabolites or complex compounds such as proteins[87;88]. With the availability of annotated genomes, it became possible to construct genome-scale metabolic networks. They combine inferred or measured gene-protein-reaction relationships, transport reactions, and an estimated biomass composition[89;90]. These reconstructions have been successfully used in biotechnological applications, mainly targeting the over-production of metabolites[91;92]. The computational approach for analysing this kind of networks is, in general, a constraint-based analysis[89;93].

A different way of investigating evolutionary changes of gene networks was proposed by Andreas Wagner with his, so-called, genotype-phenotype maps[94;95]. In these networks, the nodes represent sequences, e.g. genes or binding sites, and they are linked if they only differ in one position of their sequence. It can then be analysed whether neighbouring sequences encode for the same phenotype, e.g. the same gene function or binding of the same factor. These analyses revealed that these gene networks show a certain level of evolvability and robustness: while changed sequences can still lead to the same phenotype (robustness), it is also possible that evolutionary changes of single sequence positions can create new phenotypes. For these networks, both genotype and phenotype information needs to be available.

In co-expression networks, a pair of nodes is typically connected by a link if the genes they represent a significantly correlated expression pattern across a set of biological samples of interest. They can be built from RNA sequencing or microarray data[52;62;96]. Often the links have a weight, which can be estimated from the correlation and represents the strength of a gene-pair association. The sign of the link can be indicative of whether a gene pair is regulated in the same direction or oppositely controlled[52;96]. Most of the methods for building co-expression networks are based on a similarity measure, such as mutual information or correlation (Pearson, Spearman, Bicor, etc)[97–99]. Co-expression networks are an example of undirected and weighted networks. To reduce noise, one can choose to represent the TO of nodes instead of each interaction. The TO expresses how similar two nodes are in their set of neighbours, such that a link is drawn between two nodes if they share many interactions[68;100]. A comparison of those methods for building networks is presented in Chapter 4.

3.4 Measuring the changes in a network

Co-expression networks are constantly changing. Biologists are often interested in comparing them between a healthy and diseased status or among tissues, or in comprehending changes in the gene co-expression during development or evolution. In many studies, global network features are compared, such as the degree of a node, degree distribution, centrality, modules and pinpointing hubs. However, these measures do not have a real biological meaning. Instead, rewiring in the topology of the co-expression networks, including identifying which nodes have altered links and who changed the neighbours give much more biological insights. That information cannot be obtained using other approaches such as gene Differential Expression (DE). Consequently, a method that classifies nodes and links according to the concepts of being present, different or absent in some networks is essential for understanding how different phenotypes are affected by the gene regulatory processes[52;101].

Evolutionary analyses of co-expression networks are relatively new because comparable transcriptome datasets from different species are still rare. My method, Co-expression Differential Network Analysis (CoDiNA)[52] (Chapter 5) has been recently applied to unravel changes in co-expression network topology during development and differentiation of neurons from induced pluripotent stem cells. In addition, CoDiNA could pinpoint at which time point of the differentiation process the knock-out of the micro RNA (miRNA) 124 has the biggest effect on network rewiring[58] (Chapter 7). Another study compared the weighted Topological Overlap (wTO) networks of the prefrontal cortex of humans, chimpanzees and rhesus macaque to infer their ancestral networks and species-specific links[102].

A comparison of methods for comparing networks is shown in Chapter 5.

II

Development of methods for constructing, combining and comparing co-expression networks

4

Building and combining highly accurate networks

"Let me interact
How can I connect?
Let me interact
How can I connect?"

– Human Connect To Human, *Tokio Hotel*

Network analyses, such as of **gene co-expression networks**, **metabolic networks** and **co-occurrence networks** became an important approach for the systems-level study of biological data. Several software and libraries exist for constructing and investigating such networks. In a **gene regulatory** process a particular gene can up- or down-regulate other genes, or in **ecology**, in a co-occurrence network, two species can compete for the same energy source or can live in symbiosis. In both examples, it is important to explicitly consider both positive and negative interactions. It is also required that the represented networks have high accuracy, it means that the interactions found have to be relevant and not found by chance or background noise. Another issue derived from building co-expression networks is the reproducibility of those. When constructing independent networks for the same phenotype using different expression datasets the output network can be remarkably distinct due to biological or technical noise in the data. In this chapter, I will present how I improved the weighted Topological Overlap (wTO) methodology by calculating probabilities associated with the randomness of a particular link, I also will compare this wTO approach with other state-of-art methods. Afterwards, I will show a novel method that allows the calculation of a network that reduces the noise by combining multiple independent networks into one Consensus Network (CN). The application of the wTO method will follow on Chapter 7 and Chapter 8. The application of the CN will follow on Chapter 8.

This chapter is based on my following publications.

- **D. M. Gysi**, A. Voigt, T. M. Fragoso, E. Almaas and K. Nowick. "wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool". *BMC Bioinformatics*, 19(1), 392, 2018. https://doi.org/10.1186/s12859-018-2351-7.

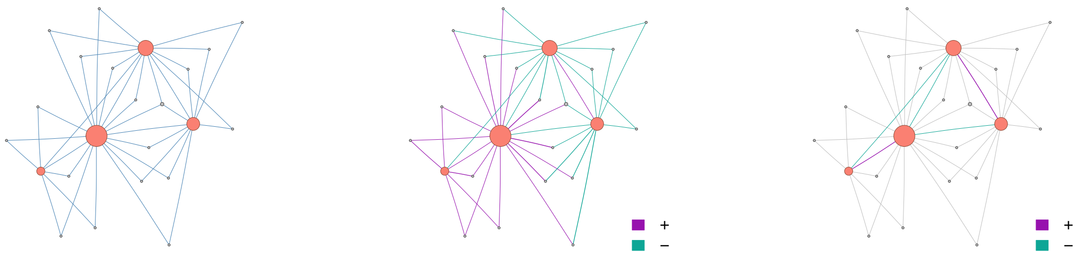
- **D. M. Gysi**, A. Voigt, T. M. Fragoso, E. Almaas and K. Nowick. "wTO: Computing Weighted Topological Overlaps (wTO) & Consensus wTO Network". 2017. Retrieved from https://cran.r-project.org/package=wTO.

This methodology is publicly available on the Compreensive R Archive Network (CRAN) as an R package, called wTO. A well-described manual can be found on Appendix A and https://deisygysi.github.io/rpackages/Pack-1.

4.1 Motivation

The analysis of interactions has been enabled in a system level of genes or species by recent applications of complex network analyses methodologies[69;71–73]. Within the biological network field, the analysis of co-expression[40;41] and co-occurrence networks[103–107] have received much attention in the complex-systems levels. In these particular areas, genes or species are represented by nodes and a pair of nodes is typically connected by a link if its nodes show a significant association expression pattern. Links can be represented as a binary relationship, where 1 denotes the **presence** and 0 the **absence** of a particular link, or alternatively, the link may have a numeric value (often called weight). The weight is a measure of association derived from a similarity measure analysis such as correlation or Mutual Information (MI) (Figure 4.1a). In a co-expression network, the strength of a gene-pair association can be represented by the link weight, and the sign as indicative of the type of associated gene interaction (Figure 4.1b): i) positive if the genes are co-regulated; ii) negative if they are oppositely controlled[96]. While in a co-occurrence network, the strength of association represent interactions among species: positive interactions might indicate commensalism or mutualism while negative interactions represent that the species compete for the same energy source or predator-prey interaction.

In many implementations of co-expression network analyses, one might primarily be interested in *a priori* defined subset of genes with a specific set of properties (Figure 4.1c). Examples include Transcription Factor (TF), non-coding RNA (ncRNA), genes with known orthologs in a set of organisms of interest or disease-associated genes[108;109]. For these situations, oftentimes the choice is made to only take into account direct interactions between the gene-subset of interest, instead of including the full set of interactions. A major drawback of such an approach is that relevant information contained in interaction patterns among excluded genes that would affect network topology and link strength values, is not incorporated in the network. The loss of such information is not only undesirable but may also lead to biased results. To reduce the noise and bias, a solution is to represent the Topological Overlap (TO) of nodes instead of each interaction. The TO expresses how similar two nodes are in their set of neighbours, such that a link is drawn between two nodes if they



(a) Unsigned network.

(b) Signed network.

(c) Topological Overlap.

Figure 4.1: Graphical representation of the topological overlap method: **(a)** represents a network, where the weight represents the association of a a node-pair. In this network all kinds of nodes can be connected; **(b)** shows a network where the association does not have only a weight but also a sign. The sign represents if a particular connection occurs in the same or opposite direction; **(c)** displays a TO network. In this network the interest lies in a subset of nodes (the red nodes). The interactions are weighted by the shared commonalities of those nodes and take into account the information that the grey links carry, however, because the interest lies only in the red nodes, only links between them are drawn (green and purple links).

share many interactions[68;100]. Figure 4.1 shows the graphical representation of the method. Just the calculating of the TO does not remove false associations of genes, therefore, a probability measure has to be associated with each link. The improvement of the method is shown on Subsection 4.2.1.

Networks differ when investigating independent networks derived from similar datasets, e.g. from a repeated experiment or independent studies on a similar subject[54]. These differences may arise from several sources: i) technical differences, such as the platform on which the expression data was measured, the facility where data was collected and prepared, or how data was processed; ii) biological differences from confounding factors, such as sex, age, and geographic origin of the individuals measured. In both cases, the network is different due to non-controllable or observable variables. Therefore, it is desirable to obtain an integrated network that considers all independently derived networks as biological replicates and systematically identifies their commonalities and is able to filter out the background noise. I developed a novel method to compute the network that captures all this information; I denote this as Consensus Network (CN). Details on that are presented on Section 4.3.

4.2 Methods for constructing co-expression networks

A variety of methods currently exist to analyse gene co-expression networks. Most are based on similarity measures. In a nutshell, we can split the similarity measures into two categories, the ones based on correlation (Sparse Partial Correlation Estimation (SPACE)[110] and TO methods such as Weighted Gene Co-Expression Network Analysis (WGCNA)[98;111] and weighted Topological Overlap (wTO)[55]) and the ones based on MI (Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe)[99;112]). These methods rest on a multitude of different mathematical principles, particularly with respect to how co-expression is quantified. However, co-expression networks based on the TO have been shown to compare favourably against other methods[113].

SPACE selects all partial correlations different from zero and fits a sparse regression model. This method allows for the identification of gene-hubs that can be further annotated or associated to the phenotype under study. ARACNe method builds the network using the MI and removing links that are indirect interactions using Data Processing Inequality (DPI) and only for not independent pairwise correlation. The wTO builds the adjacency matrix using the raw correlation, ρ, of all pairwise interactions and later computes the TO as

$$\omega_{i,j} = \frac{\sum_u a_{i,u}a_{u,j} + a_{i,j}}{\min\{k_i, k_j\} + 1 - |a_{i,j}|}, \tag{4.1}$$

where $k_i = \sum_j |a_{i,j}|$ and

$$\mathbb{A} = [a_{i,j}] = \begin{cases} \rho_{i,j} & i \neq j \\ 0 & i = j. \end{cases} \tag{4.2}$$

The WGCNA general framework consists of first calculating the pairwise correlations of the expression values of genes and applying a soft threshold on the correlation and has three options to calculate the adjacency matrix i) a signed version; ii) a hybrid version and iii) unsigned. In all three cases the weight values resulting from the TO are given by

$$\omega_{i,j} = \frac{\sum_u a_{i,u}a_{u,j} + a_{i,j}}{\min\{k_i, k_j\} + 1 - a_{i,j}}, \tag{4.3}$$

which lies on unit interval, since the adjacency matrix in WGCNA is defined only for positive values.

There are three possible adjacency matrix for this approach. The first one is the signed WGCNA, that considers only a soft threshold

$$a_{i,j}^{signed} = \left[\frac{\rho_{i,j} + 1}{2}\right]^{\beta},$$

where β is an integer that forces a network to fit into a power-law. The second one is a hybrid version that considers both a hard and a soft threshold

$$a_{i,j}^{signed\ hybrid} = \begin{cases} [\rho_{i,j}]^{\beta} & \text{for} \rho_{i,j} > 0 \\ 0 & \text{for} \rho_{i,j} \leqslant 0, \end{cases}$$

and the last one is an unsigned version, that considers only a soft threshold

$$a_{i,j}^{unsigned} = [\rho_{i,j}]^{\beta}.$$

In contrast to WGCNA, wTO uses the raw correlation as the adjacency matrix for building the Topological Overlap Matrix (TOM)[114;115]. Where the weight of a gene-gene interaction is then an average across all its neighbours. For both, wTO and WGCNA, the $\omega_{i,j} = 1$ if both the conditions are satisfied for the node with fewer links: i) all its neighbours are the same of the other node and ii) they are connected. However, $\omega_{i,j} = 0$ if they have no neighbours in common or are not connected.

Different from the methods above, Bayesian networks assume that the gene relationships are causal and a direction of the regulatory process can be captured in an acyclic graph. Each interaction takes into account all other possible gene interactions[116;117]. The main problem in using this method is the extremely long running time.

4.2.1 Weighted Topological Overlap calculation

Zhang et al.[118] first described the wTO method in 2005. The representation of interactions between a set of nodes by this method takes into account the overall commonality of all the links a node has, instead of basing the analysis only on calculating raw correlations among the nodes[115;118;119]. It thus provides a more comprehensive understanding of how two nodes are related. Therefore, it is expected that a network build from this method contains more robust information about the connections among nodes than what would result from simply taking direct correlations into account[55;118].

The wTO can be computed based on a similarity matrix, where the link weights are calculated using Pearson's product moment correlation coefficient or the Spearman Rank correlation. The first one measures the linear relationship between two genes. Note that, the Pearson's correlation coefficient is sensitive to extreme values, and therefore it can over or underestimate the strength of an association. The Spearman Rank correlation is recommended when data are monotonically correlated, skewed or ordinal, and it is less sensitive to extreme outliers than the Pearson coefficient[120–123].

Nowick et al.[55] improved the method by allowing the method to explicitly accommodate both positive and negative correlations. Later, I improved the method for allowing it to estimate if the $\omega_{i,j}$ value found is different from zero by calculating a $p-$value for each link. This calculation allows the final network to be filtered using a probability measure in each link instead of an overall correlation measure for all links.

I use the same notation defined in Section 3.2. Let $\mathbb{N}$ be a set of n nodes, $\mathbb{N} = \{N_1, \ldots, N_n\}$ and $\rho_{i,j}$ a correlation between a pair of nodes i and j. The adjacency matrix $\mathbb{A} = [a_{i,j}]$ is defined as in Equation 4.2.

Assuming that nodes i and j represent a sub-set of factors (e.g genes) of particular interest selected from the n nodes, the wTO ($\omega_{i,j}$) is calculated[55] between nodes i and j as presented in Equation 4.1.

Note that, this expression (Equation 4.1) explicitly includes both positive and negative correlations, and thus allows for $\omega_{i,j}$ to take both positive and negative values. Other software packages calculating the $\omega_{i,j}$ have implemented definitions of the TO method that does not allow for negative values[98], making my version more valuable for gene regulatory network analysis. The `wTO` package also calculates the unsigned network, and for that, it takes as an input the absolute correlation values ($|\rho_{i,j}|$).

Since (Equation 4.1) explicitly allows $a_{i,j} \leqslant 0$, therefore, is important to be aware of the limits of this expression. Consider three nodes i, j and u, and assume that $a_{ij} \leqslant 0$. All the terms in the numerator of (Equation 4.1) will be negative if $a_{iu}a_{uj} \leqslant 0$ for all nodes u. However, if $a_{iu}a_{uj} > 0$, then at least some contributions to the sum will cancel out. The same rationale applies for the case of $a_{ij} \geqslant 0$.

To systematically assess the potential effect of term cancellation in (Equation 4.1), I calculated the absolute wTO, $|\omega|$, which uses the absolute value of the correlations ($a_{i,j} = |\rho_{i,j}|$) as input for (Equation 4.1). In this case, the sign of the correlation is excluded from the analysis and only the magnitude of the link-strength is taken into account. Consequently, by generating a scatter plot of the signed and unsigned weights, it is possible to assess at which $\omega_{i,j}$-values term cancellations start affecting the results. Thus, for values of interest, the closer the plot of ω vs. $|\omega|$ is to the classic $y = |x|$, the better.

However, by just computing the wTO network all spurious correlations are not avoided. To overcome this issue, my package estimates the probability of each ω being zero by testing the hypothesis

$$\begin{cases} H_0 : \omega_{i,j} = 0 \\ H_a : \omega_{i,j} \neq 0, \end{cases}$$

of the null hypothesis (H_0) of no association against the two-sided alternative (H_a) of non-zero association. This can be computed by using a bootstrap approach[124].

Bootstrap is a method to measure the accuracy of statistical measures, such as mean, standard deviation, correlation etc. But it can also be used for more complex measures, such as the wTO. The idea behind the bootstrap is very simple. Assume that we have the gene expression of Gene A ($\mathbb{G}$) measured in m different individuals, it means, $\mathbb{G} = \{G_1, \ldots, G_m\}$. The process starts by generating multiple independent samples of size m **with** replacement, each resampling realisation is in the B^* bootstrap sample, $\{\mathbb{G}^{1^*}, \ldots, \mathbb{G}^{B^*}\}$. For each realisation, you compute the statistic $f(x^{B^*})$ of interest, such as the mean, median, standard deviation. With this collection of values, it is straightforward to derive an empirical distribution for the statistic under study. A visual representation of the method is on Figure 4.2. When the dataset is composed of correlated observations, such as time series and repeated measures, this naive independent sampling with replacement must be modified to sampling that allows for such particularities. The blocked bootstrap, as the name suggests, builds **blocks** that controls for the high dependency on data and each block is considered to be one individual.

The approximation of the weights' empirical distribution can be done by resampling the individuals with replicates and the probability that an observed weight is sufficiently distant from zero can be easily calculated. Each bootstrap realisation B^* estimates a $\omega^*_{i,j}$.

The absolute difference of $\omega_{i,j}$ and each $\omega^*_{i,j}$is bounded by a fixed confidence interval of $\delta - \omega_{i,j} \leqslant \omega^*_{i,j} \leqslant \delta + \omega_{i,j}$. This means that, the smaller δ is, the stronger the confidence is in a particular $\omega_{i,j}$. By default, the `wTO` package sets the δ to 0.2.

In short: for the ω case, I have drawn B^* new samples of size m with replacement from the original dataset. For each link, calculate the $\omega^*_{i,j}$, derive the empirical distribution for each one of the links. Define an interval of interest, δ, and estimate the proportion of $\omega^{B^*}_{i,j}$ outside the predefined interval.

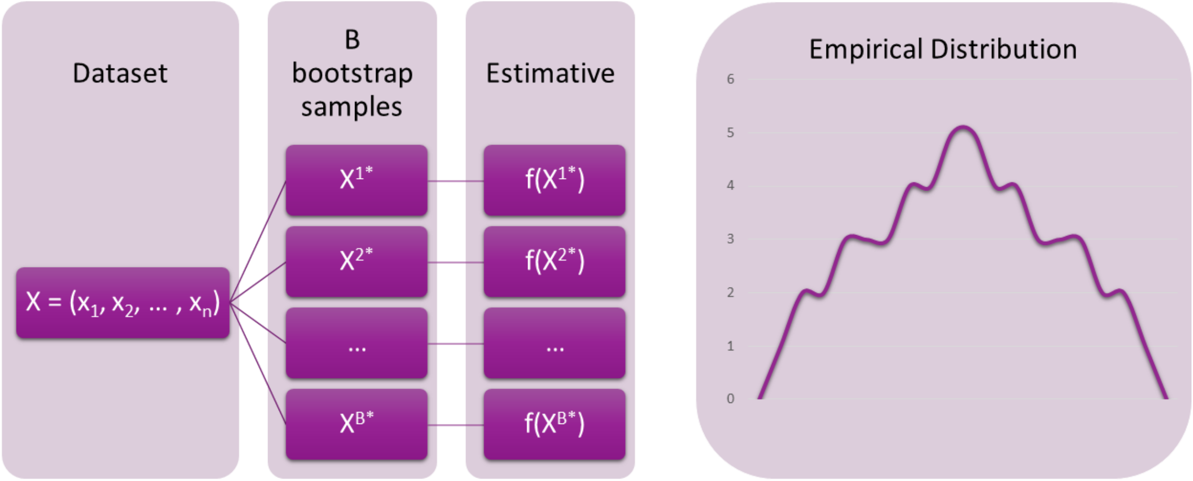


FIGURE 4.2: Schematic representation of the bootstrap method: Each B^* bootstrap sample is generated from the dataset. Each bootstrap sample contains m elements from the original data, with replicates. These samples are used for estimating the $f(x^*)$ statistics of interest. The result from each B^* statistics computed are used to build the empirical distribution.

Calculating networks for repeated measure and time series data

One advantage of the `wTO` package is its application in the analysis and construction of networks for time series and repeated measures data. For that, the implementation of blocked bootstrap resampling[124] is needed. This type of resampling is necessary mainly because there are two correlation components structures in these samples: the node's correlation and the autocorrelation (individuals or time). For the individual, the autocorrelation is given by the same individual being measured multiple times and its measures can be very similar, and thus, inflate (or deflate) the nodes correlation. While for a time series there is a tendency of consecutive values to be correlated. An important benefit of the presence of autocorrelations is that we may be able to identify patterns inside a time series, such as seasonality (patterns that repeat themselves at a periodic frequency). A way to measure the time dependency is by using a lag. Those are particularly helpful in time series analyses and can be chosen using a partial correlation of the time per sample. In both cases, the resampling is followed by calculating the wTO for a time series where the observations are **not** independent of each other.

4.2.2 Comparing wTO to other state-of-art approaches

In order to quantitatively compare the performance of wTO, WGCNA and ARACNe, I downloaded a gene expression dataset from *E. coli* from http://systemsbiology.ucsd.edu/InSilicoOrganisms/Ecoli/EcoliExpression2[125–128]. The data consist of 213 Affymetrix microarray gene expression profiles, corresponding to multiple different strains under different growth conditions, and contains gene expression data for $7,312$ distinct probes. Quality control of the probes was performed and $\log_2$ normalised previous to making the dataset available. Unspecific probe-set were removed and genes that were matched from more than one probe were combined using the mean of all its probes. The resulting dataset contained $4,356$ expressed genes.

To assess the capability of the three tools in identifying true TF-TF interactions, I used the RegulonDB[129] database, which contains experimental data from *E. coli*, as a reference. I define as True-Positive interactions those that are described in RegulonDB, and as True-Negatives all interactions that could not be experimentally validated in that dataset. For comparison, I also calculated networks using only Pearson's correlation without any modification. We generated the network for WGCNA following the steps described by the authors in the Tutorial[118;130] and used the functions pickSoftThreshold() and pickHardThreshold() for defining the power of the soft-threshold and for choosing the hard-threshold, respectively. The power was defined as 4 and the hard-threshold was set to 0.3.

The ARACNe network was built using the Pearson correlation with build.mim and ARACNe functions in the minet R package[131]. The wTO networks were built using $1,000$ simulations, Pearson correlation and filtered for Benjamini-Hochberg adjusted $p-$value (BH)[132] $\leqslant 0.01$ and the 90% quantile. One wTO network was constructed using a δ of 0.2, the default of the wTO package, and another network was built using a δ of 0.1. All networks were filtered to only contain TFs with information in the RegulonDB. To measure accuracy of the methods, the Receiver Operator Curve (ROC) curve was calculated using the pROC R package[133] (see Figure 4.3). The ROC curve is plotted, using different thresholds, the True Positive Rate, sensitivity, against the False Positive Rate, specificity. It is a plot of the power in function of the type I error.

ARACNe was able to better identify the number of true positives compared to WGCNA and wTO, but performs worse when finding true negatives, thus resulting in a larger number of false positives. (Figure 4.3, Table 4.1). WGCNA is better at finding true negatives, but does not identify many true links. The wTO method performs better than WGCNA in finding true positives and better than ARACNe in finding true negatives. It also finds fewer false positives than ARACNe. In general, even when using a large δ, that can be interpreted as a wide confidence interval, wTO performs better than the two other methods, as seen in the Area Under the Curve (AUC), the closer it is to unity, the better. This demonstrates that the use of the wTO method further reduces false effects coming from incorrectly assigned linked genes (false positives) when compared to ARACNe and raw correlations.

TABLE 4.1: Accuracy of the three methods and correlation.

Case	ReactomeDB (Total)	Pearson Correlation	ARACNe	WGCNA	wTO ($\delta = 0.1$)	wTO ($\delta = 0.2$)
True Negative	7234	2259	2633	7092	6520	5235
False Negative	0	216	245	321	318	288
False Positive	0	4975	4601	142	714	1999
True Positive	328	112	83	7	10	40
Total	7562	7562	7562	7562	7562	7562

4.3 A method for combining networks

Constructing independent co-expression networks for the same trait using different data leads to different networks. Therefore, it is well known that co-expression networks can carry background noise due to: i) biological differences, such as age, sex, gender etc; ii) technical differences such as the platform used for measuring the gene expression, facility. Those differences cannot be measured or be controlled when the network is constructed. This background noise that is intrinsic to each network leads to distinct networks for the same phenotype. Because of that, methods that are able to combine multiple networks into one network with reduced false positive rates and noise levels are required.

Berto et al.[54] described a consensus network based on gene-expression data from primates' frontal lobes by applying a Wilcoxon test on the links. It assumes that there are a sufficient amount of independent networks in two groups to be compared and therefore, combine the networks. However, there are not always a substantial amount of networks to be combined. Thus, a method that is able to build a CN from at least two networks, re-weight the links and compute a probability for them is needed.

The CN methodology I developed allows combing two or more networks, each generated from different and independent datasets, into a single CN. This method, penalises the links with opposite signs and also links that does not exist in all networks. According to the same rationale, links with the same sign among the multiple wTO-networks, will have their link weight values (ω) closer to the largest $|\omega_{i,j}|$ of the link in all k networks. In order to obtain the CN, the first step is to remove nodes that were not measured in all networks. Consequently, if a node is absent in at least one network, it is not possible to compute a consensus of the links that belong to that node (Figure 4.4). It is particularly important for avoiding the false associations of factors that were not measured in all networks.

In order to obtain a single integrated network derived from multiple independent wTO networks, we calculate a CN using the following approach.

Assume a set w of k independent networks $\mathbb{W} = \{W_1, \ldots, W_w\}$ replicated networks, the CN, $\Omega_{i,j}$ is defined as

$$\Omega_{i,j} = \sum_{k=1}^{w} \nu_{ij,k}\, \omega_{ij,k}, \tag{4.4}$$

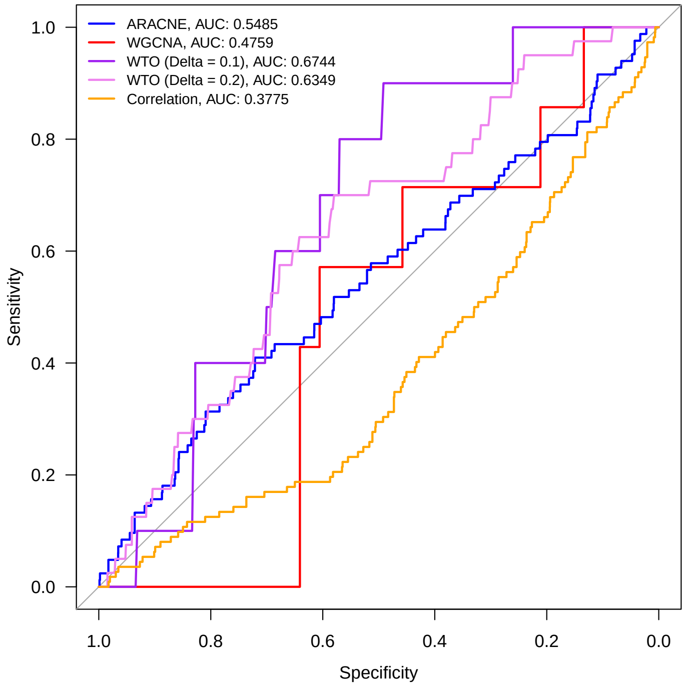


FIGURE 4.3: ROC curves for the comparison of methods. Overall, the `wTO` method performs better than `ARACNe`, `WGCNA` and raw Pearson correlations. `ARACNe` is better in finding true positives, while `WGCNA` is more conservative, and therefore better in finding true negatives but identifies fewer true positives.

where

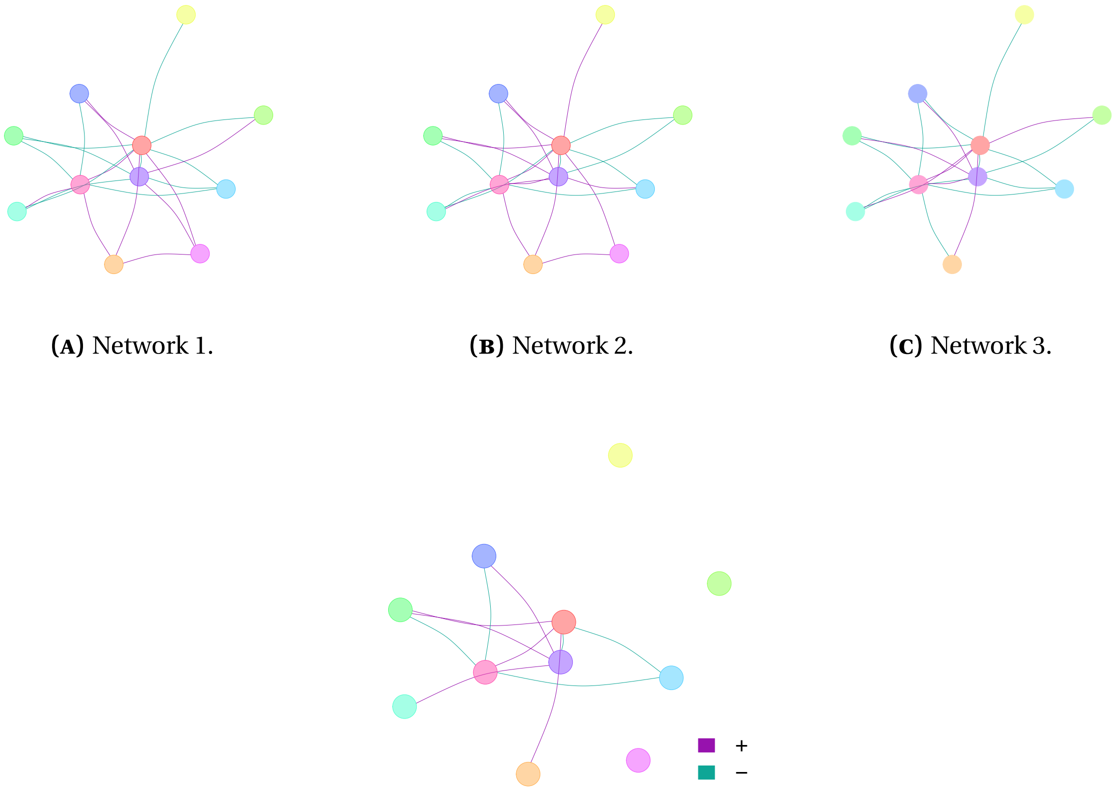
$$\nu_{ij,k} = \frac{|\omega_{ij,k}|}{\sum_{k=1}^{w} |\omega_{ij,k}|}. \tag{4.5}$$

A threshold can be applied to remove links with $\omega_{i,j}$ values close to zero, thus should not be included in the consensus network. To join networks that were generated with the wTO method into the CN, the $p-$values are combined using the Fisher's method[134]

$$\chi^2_{2w} \sim -2 \sum_{k=1}^{w} \log(p_k), \tag{4.6}$$

which can be used because i) the networks are independent; ii) Uniform distribution is assumed for the ω $p-$values.

A visual representation of the CN methodology is shown in Figure 4.4. The thicker the link between two nodes is, the stronger the correlation between them. The signs are represented by the colours green and violet, respectively. If a link has different signs in the networks, the strength of the link in the CN is close to zero. When all links agree to the same value or show little deviation, the strength of the resulting CN value is closer to the determined absolute maximum value. If a node is absent in at least one network, it is removed.



(a) Network 1.

(b) Network 2.

(c) Network 3.

(d) CN.

Figure 4.4: A schematic example of the CN method: Three independent networks are shown in **(a)**, **(b)** and **(c)** to be combined into one CN. Note that the rightmost network does not have the right bottom pink node; **(d)** is the resulting CN. Note that, the pink node node is present in the CN but does not contain any link. Also, only links that do not change sign between networks are present in the CN. For example, the link between the red and the yellow nodes is removed, because it has different signs in networks.

4.4 A package for constructing and combining networks

In order to make the wTO and CN methods available, an R package called wTO was developed. wTO is open source and freely available from CRAN https://cran.r-project.org/web/packages/wTO/ under the GPL-2 Open Source License, and it is platform independent.

4.4.1 Input data

The wTO R package can handle a wide range of input data. Data can be discrete or continuous values. However, it is recommended that the input data should be previously cleaned using the common steps for quality control and normalisation prior to the network construction. This step avoids background noise. For RNA sequencing (RNA-seq) data, wTO can handle normalised quantification, for example Reads Per Kilobase Million (RPKM), Fragments Per Kilobase Million (FPKM) and Transcripts Per Kilobase Million (TPM). For microarray data, $\log_2$, Robust Multi-array Average (RMA) or MicroArray Suite 5 (MAS5) values can be used. For metagenomics data, for instance for analysing co-occurrence networks, the recommended normalisation for the abundance data are per day/sample or Hellinger distance.

4.4.2 Functions

The function wTO() calculates the weights for all links according to (Equation 4.1) between a set of nodes for a given input dataset. A user that is **not** interested in having a highly accurate network can run this function.

To test whether the calculated $\omega_{i,j}$ is different than a random expectation and to decide on a suitable hard-threshold value for including link weights, the functions wTO.Complete() and wTO.fast() are implemented. Both this functions calculates the $\omega^*_{i,j}$ a number of times, specified by the user, by using either the method_resampling as (‘‘Bootstrap’’), or for time series or repeated measures data case (‘‘BlockBootstrap’’). In the last case a lag or an ID is required. The user may specify the correlation method (Pearson or Spearman) that this function should use, the Pearson correlation is the default choice.

Because resampling methods, such as bootstrap and permutations, are computationally expensive, the wTO.Complete() can also run in parallel over multiple cores to reduce the wall clock time. For running in parallel, the user may specify a given number of k computer threads to be used in the calculations. To implement the parallel function, the R package parallel[18] was used.

The execution of the wTO.Complete() function returns two outputs: i) a diagnosis set of plots (Figure 4.5) and ii) a list consisting of the following three objects:

- $Correlation is a data.table containing the Pearson or Spearman correlations between **all** the nodes, not only the set of interest. The $\omega_{i,j}$ values for the set of nodes of interest are based on these correlations. The default of this output is set to FALSE.

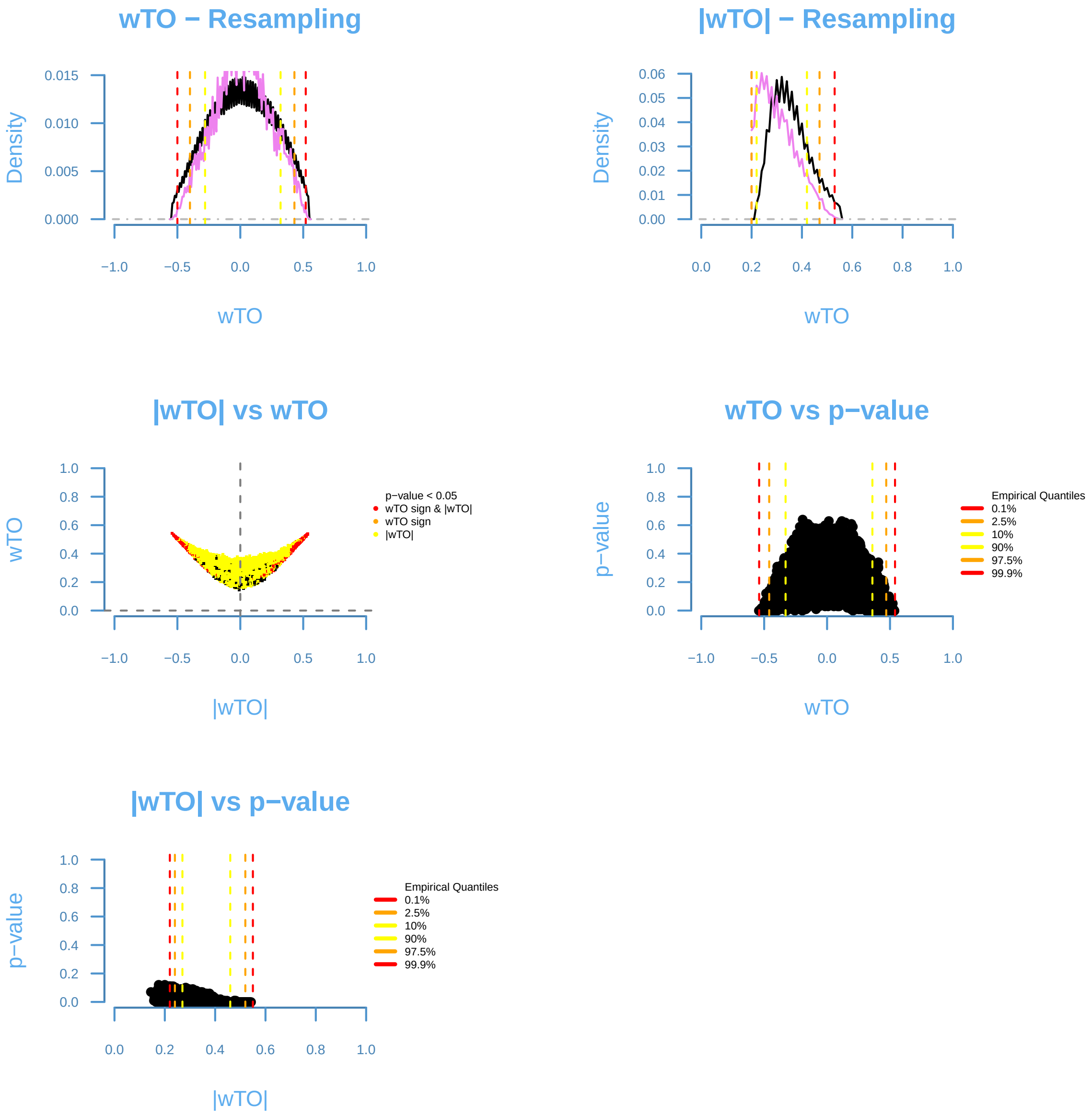


FIGURE 4.5: `wTO R` **package diagnosis plot**: The diagnosis plot shows the quality of the resampling (first two plots). The closer the purple line to the black line, the better. The $\omega_{i,j}$ vs $|\omega_{i,j}|$ shows the amount of $\omega_{i,j}$ being affected by cancellations on the heuristics of the method, the most similar to a *smile* plot, the better. Note that for both ω and $|\omega|$ when the $p-$value is lower than 0.05, the ω is closer to the diagonal line where the values are not affected by the signal cancellation of the method. The last two plots show the relationship of the $p-$values and the $\omega_{i,j}$. It is expected that higher ω will have lower $p-$values.

- `$wTO` is a `data.table` containing the nodes, the $\omega_{i,j}$ values (signed and unsigned), the $p-$values and the adjusted $p-$values computed using both signed and unsigned correlations.
- `$Quantile` is a table containing the quantiles for the empirical distribution, computed using the bootstrap and the quantiles for the real data: 0.1%, 2.5%, 10%, 90%, 97.5% and 99.9%. Those empirical values can be used as a threshold for the ω values when it is not desired to visualise small $\omega_{i,j}$.

The set of plots on Figure 4.5 indicates the quality of the resample: the closer the density of the resampled data is to the real data, the better. Another generated plot is the scatterplot of the $\omega_{i,j}$ vs $|\omega_{i,j}|$, as discussed in Subsection 4.2.1. The scatter plot of $p-$values against the $\omega_{i,j}$ and $|\omega_{i,j}|$ is also plotted along with suggested threshold values that are the quantiles based on the empirical distribution. Note that the values on the diagonal of this plot, have in most of the cases, low $p-$values, it shows that the implementation of the bootstrap *per se* can alleviate the sum problem presented before.

The `wTO.Consensus()` computes the CN. This function allows the user to give a list of networks in a `data.frame (edge.list)` format with: Node 1, Node 2, the link weight and the $p-$value. The output is a `data.table` containing the two nodes' names and the consensus weight, and the combined $p-$value. This allows the user to filter out the links that were not significant in part of the network.

The `wTO R` package also includes options to visualise the resulting networks. The function `NetVis()` generates an interactive graph using as input a list of links and their corresponding weights. The analysis functions `wTO.Complete()` and `wTO.Consensus()` both generate network data-structures (`edge.list`) that can be visualised with this function. The user needs to choose a relevant ω-threshold (the quantiles resulting from the bootstrap), $p-$value cut-off and/or $p_{adj}-$value to select the set of links to be plotted. Additionally, the user may choose a layout for the network visualisation from those available in the `igraph`[135] package. By default, the threshold value is set to 0.5, and the network layout-style is set to `layout_nicely`. To avoid false positives, we recommend to filter the data according to the desired significance $p-$value and to choose the wTO-threshold according to the computed empirical quantiles. The size of the nodes is relative to their degree. My package further includes an option for making clusters from the nodes; if allowed, nodes are coloured according to the cluster they belong to. The user can choose the method to create the clusters. The width of a link is relative to the $\omega_{i,j}$, and its colour is respective to its sign (if a signed network was calculated). Nodes can have different shapes, allowing for labelling nodes of different classes, for example, target genes or protein-coding and non-protein coding genes. Furthermore, the user may also zoom in and out of the network visualisation, drag nodes and links, edit nodes and links, and export the image as `.html` or `.png`. The package provides example datasets and an example of nodes of interest as well. The workflow of the data-analysis using the `wTO` package can be seen in Figure 4.6.

One important difference between the `wTO` package and the `WGCNA` package, is that `wTO` only use significant links for cluster (modules) network representation oposed to the full set of co-expressions, as in the `WGCNA` package.

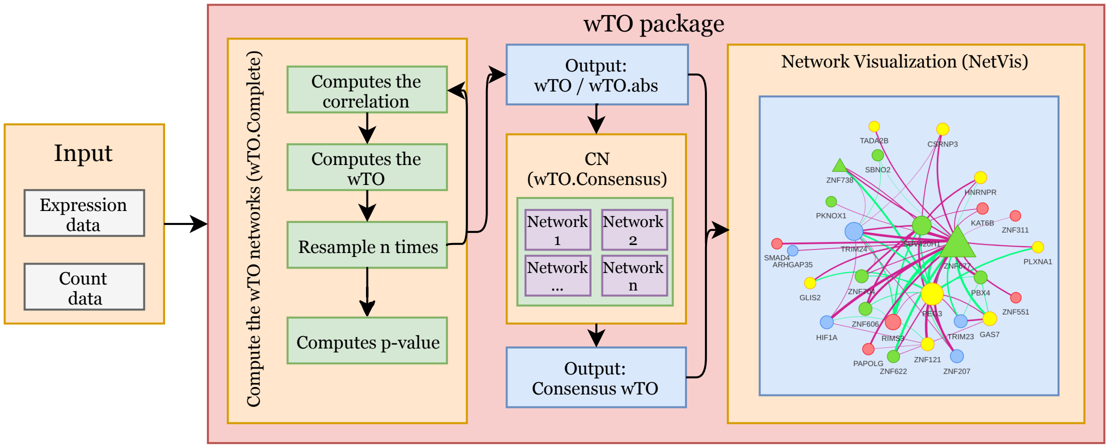


Figure 4.6: The wTO package workflow: Gray boxes refer to inputs, red boxes refer to content of the wTO package, yellow boxes are functions included in the package, blue boxes are outputs of those functions, and green boxes refer to methods internal to the package. `wTO` package can deal with multiple kinds of data, for example, RNA-seq counts or normalised values, microarray expression data, abundance data coming from metagenomic studies, and many more. All input data should be pre-processed with the quality control and normalisation methods recommended for each respective type of data. The function `wTO.Complete()` calculates the wTO values, as many times as desired. As output, the user will obtain an object containing the signed and absolute wTO values for each pair of nodes, $p-$values and $p_{adj}-$values for multiple testing. This output can be used for the construction of a CN from independent networks using the function `wTO.Consensus()`. Outputs from the wTO and CN networks can be used as an input for `NetVis()`, which is an integrated tool for plotting networks. As an interactive tool, it also allows the user to modify the network.

Relative to `WGCNA`, wTO provides three major additions: the determination of $p-$values (determined by bootstrapping) for each pairwise wTO value; the calculation of a consensus network, and the ability to visualise the topological overlap network (along with node grouping according to a choice of nine algorithms). While `WGCNA` provides a variety of tools for visualising the hierarchical tree forming the network, as well as for rendering the correlation matrix in heatmap form, it does not provide a node-and-edge type view of the co-expression network (but does allow for exporting networks into Cytoscape, in which network views are possible). Additionally, the consensus network as defined in Equation 4.5 differs from the consensus TOM defined in `WGCNA`, which simply assigns to each edge of the consensus network the minimal value of the topological overlap across the input conditions. This is a strict version of consensus (unanimity), in that it will discard any gene pair if the overlap is weak in even a single network. In contrast, while Equation 4.5 will remove contributions from networks where the TO is weak (or where the sign of the ω is in conflict with the other networks), an edge may still be included if it is sufficiently present across the other networks.

Further additions in wTO include the possibility of choosing the Spearman correlation as the basis of $\mathbb{A}$ (while `WGCNA` provides biweight midcorrelation, or bicor for short; both provide Pearson), as well as reducing computation time by the option of restricting the calculation of wTO scores to a set of genes of interest (while still including the adjacency to genes outside this set in each inter-set wTO score).

For an unweighted network, where $a_{i,j} = 0$ or $a_{i,j} = 1$ for all (i, j), this approximates to

$\omega_{i_i} \approx 1$ for large k_i. However, this is not the case for weighted networks. WGCNA differs from the wTO package in that $w_{i,i} = 1$ is explicitly set for all i, while the wTO package retains the score as defined by Equation 4.3.

A brief comparison of the main R packages for constructing networks can be found on Table 4.2.

TABLE 4.2: Comparison of key differences between wTO, WGCNA and ARACNe.

Method	**wTO**	**WGCNA**	**ARACNe**
TO	Yes	Yes	No
Signed TO	Optional	No	No
Consensus TO	Weighted sum	Minimum weight (strict)	No
Pairwise $p-$values	Yes	No	Used to filter MI
Network view	Native	Exported to Cytoscape	Exported to Cytoscape
Soft threshold	No	Optional (on by default)	No
Correlation choices	Spearman Pearson	Bicor Pearson	Spearman, Pearson Kendall
Deal with time series	Yes	No	No
Deal with repeated measures	Yes	No	No

4.4.3 Algorithm compute time with varying system size

Normally, when running the wTO, the interest lies on a subset of nodes of interest. In Figure 4.7, we show the run-time for different network sizes, and different proportions of nodes of interest. When running the wTO for all expressed genes coding for TF being the genes of interest, we have around 14% of nodes of interest. Using a standard laptop computer, it's possible to compute the wTO for a full network with $20,000$ nodes in 20 milliseconds per link. This shows that it is quite feasible to compute the full wTO for a realistic gene expression network.

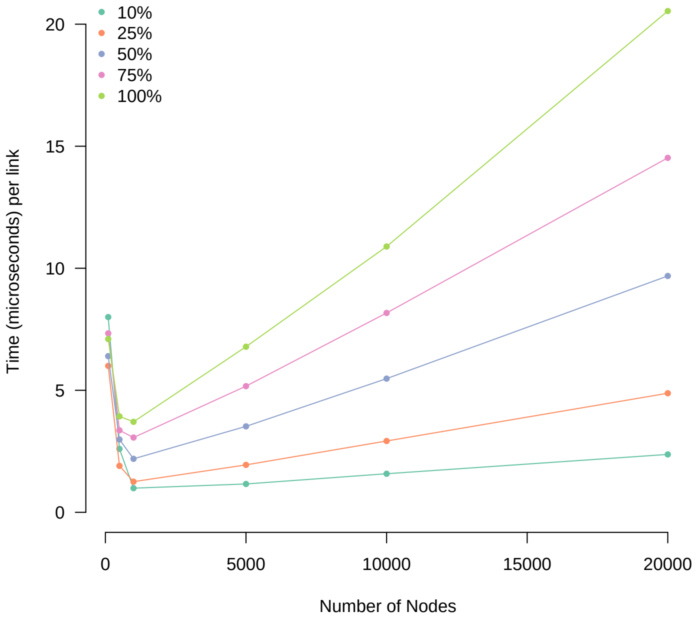


FIGURE 4.7: Computational time for the calculation of wTO for each link for different sizes of networks and proportions of sets of nodes of interest: The run time of the $\omega_{i,j}$ calculation increases with increasing proportion of nodes of interest. The graph presented here shows the time for computing each link for different sizes of nodes and proportions of subsets of nodes of interest.

5

Comparing highly accurate networks

> *"Six degrees of separation doesn't mean that everyone is linked to everyone else in just six steps. It means that a very small number of people are linked to everyone else in a few steps, and the rest of us are linked to the world through those special few."*
>
> – The Tipping Point: How Little Things Can Make a Big Difference, *Malcolm Gladwell*

Biological and medical sciences are increasingly recognising the relevance of gene co-expression-networks for the analysis of complex systems, phenotypes or diseases. Typically, complex phenotypes are investigated under varying conditions. While approaches for comparing **two** networks exist, this is not the case for multiple networks, although many studies consist of more than two datasets, for example, multiple tissues, treatments, time points, or species. In this Chapter, I present a method for the systematic comparison of an unlimited number of networks: Co-expression Differential Network Analysis (CoDiNA). In particular, CoDiNA detects links **and** nodes that are common, specific or different among the networks. CoDiNA includes a statistical framework to normalise between these different categories of common or changed network links and nodes, resulting in a comprehensive network analysis method, more sophisticated than simply comparing the presence or absence of network nodes.

In this Chapter, I will present how I developed the CoDiNA, a classifier method for links and nodes in a set of co-expression networks. I also will compare qualitatively my CoDiNA approach with other state-of-art methods.

Applying CoDiNA to a neurogenesis study we identified candidate genes involved in neuronal differentiation. Experimentally overexpressing one candidate resulted in a significant disturbance in the underlying gene regulatory network of neurogenesis (Chapter 7), moreover, applying CoDiNA to different mental disorders showed gene candidates that seem to be involved in modules involved with those disorders (Chapter 8).

This chapter is based on my following publications.

- **D. M. Gysi**, T. M. Fragoso, V. Busskamp, E. Almaas and K. Nowick. "Comparing multiple networks using the Co-expression Differential Network Analysis (CoDiNA)". *Submitted*. 2018.
- **D. M. Gysi**, T. M. Fragoso, E. Almaas and K. Nowick. "CoDiNA: Co-Expression Differential Network Analysis". *CRAN*. 2018. `https://cran.r-project.org/web/packages/CoDiNA/`.

This methodology is publicly available on the Compreensive R Archive Network (CRAN) as an `R` package, called CoDiNA. A well-described manual can be found on Appendix B and `https://deisygysi.github.io/rpackages/Pack-2`.

5.1 Motivation

Complex systems, exemplified by biological pathways, social interactions, and financial markets, can be expressed and analysed as systems of multi-component interactions[136]. In systems biology, it is necessary to develop a thorough understanding of the interactions between factors, such as genes or proteins. Gene co-expression networks have been especially effective in identifying those interactions[69;71–73], and as mentioned the sign of the interaction may suggest an up- or down-regulation of one factor by the other[96]. It has been shown that different conditions have distinct underlying regulatory patterns and therefore will lead to dissimilar networks even for a single system[54;62;136].

Differential network analyses are able to capture changes in gene relationships and are thus exceptionally suitable for understanding complex phenotypes and diseases[73]. And this cannot be done using a Differential Expression (DE) approach. Several methods for pairwise networks comparisons exists[101;137–145]. However, it is often of great interest to compare more than just two networks simultaneously, such as gene co-expression networks arising from different species, tissues or diseases, or co-existence networks from different environments. Unfortunately, the comparative method for more than two networks that account for both links and nodes were missing.

To overcome this issue, researchers have been contouring this using different approaches. For instance, an evolutionary study conducted pairwise comparisons between humans, chimpanzees (*Pan troglodytes*) and rhesus macaques (*Macaca mulatta*) to uncover similarities and differences in the Prefrontal Cortex (PFC)[102]. In a recent medical study, the authors compared enriched gene functions using the Gene Ontology[40] instead of comparing the networks of the multiple cancers. Another study generated a network that involved only differentially expressed genes[146] extracted from their multiple networks. Few problems arise from this: i) the lack of statistical power; ii) the accuracy is reduced by multiple comparisons. Therefore, these studies could have profited extensively from applying a method capable of systematically comparing multiple networks simultaneously.

Kuntal et al.[136] proposed a method, CompNet, that may address the comparison of multiple networks. However, the focus of CompNet is on the visualisation of the union, intersections and exclusive links of the analysed networks. ConMOd[147] has recently been developed to find conserved functional modules across multiple biological networks. However, a method that is capable of comparing both links **and** nodes of **any** number of networks is still lacking.

Here, I present a novel method for that purpose, Co-expression Differential Network Analysis (CoDiNA), implemented as an R package, CoDiNA, that also includes an interactive tool for network visualisation similar to the one presented in the wTO R package Chapter 4.

5.2 Comparing CoDiNA to other state-of-art approaches

Evaluating multiple co-expression network methodologies for comparing networks is considerably challenging due to the lack of a gold standard network for multiple conditions[148], of which all links are experimentally validated. Therefore, we are able to identify theoretical similarities and differences among the methods, and that is how I compare the methods here.

Few other tools, CompNet[136] and ConMod[147], allows for the comparison of more than two networks. The focus of CompNet is on the visualisation of pairwise Jaccard-similarities from the union, intersections and exclusive links of those networks. It includes features such as pie-nodes and links to allow the user to identify key elements of the network. Elements are identified by providing a distribution of global graph properties, such as the network number of nodes, number of links, density, clustering coefficient, average path length and diameter. Even though building a visualisation tool is not the focus of CoDiNA, we also incorporated an interactive tool for visualisation of the final network, and CoDiNA provides summary statistics of the network, such as the total number of links and nodes, degree of the nodes, as well as how many links and nodes have been classified as common, different or specific to each category. Other network statistics can be easily obtained using the igraph[135]. ConMod focuses on finding only common modules of different networks by building a consensus network from layering different networks and from that, selects nodes for building the conserved modules. Unlike CoDiNA and CompNet, ConMod does not perform an actual comparison.

Lichtblau et al.[148] compared ten differential network analysis methods that are able to perform a pairwise comparison. The authors split the methods into two main categories: Local search and Global search. Global methods focus on changes in the network topology while local methods search for changes in the nodes. CoDiNA combines both: it first searches for changes in the topology of the networks and then for the specificity of the nodes. This allows investigating both features with one powerful tool. Changes in network topology indicate alterations in affected pathways or regulatory relationships, while changes concerning specific nodes can evaluate the importance of genes for the network and suggest genes that might be responsible for the topology differences. Together, the local and global changes are crucial for understanding the functional effects of network changes.

As previously mentioned, most methods for distinguishing networks allow only pairwise comparisons, but few approaches distinguish multiple conditions. For a systematic comparison of links or nodes several methods can be considered: CoDiNA, CoXpress[139], Comp-

TABLE 5.1: Methods for comparing co-expression networks: Amount of networks to be compared; Statistical methodology used in the method; Focus on nodes or links; Output network of the method: Integrated visualisation tool and availability of the method.

Method	# Networks	Methods	Nodes / Links	Output	Visual	Available
CoDiNA	≥ 2	Geometrical transformation, Normalised scores for links and classification of nodes	Links and nodes	Full network	Yes	R package
CompNet	≥ 2	Jaccard-similarities from the union, intersections and exclusive links	Links	Full network	Yes	GUI
CoXpress	2	Hierarchical cluster analysis on the expression values	Nodes	Cluster of genes for hierarchical each group	Yes	R package
CSD	2	Score the links to construct a unified differential co-expression network	Links	Full network	No	In-house software
DiffCorr	2	Fisher's z-test	Links	Full network	Yes	R package
Gain	2	Calculates the Jaccard, Simpson, Geometric, Hypergeometric and Cosine indexes and Pearson correlation for links	Links	Full network	Yes	Web-based
MIMO	2	Sub-graph matching	Nodes	Sub-graph	No	In-house software
NetAlign	2	Identifies conserved structures from topology and sequence similarity	Nodes	Conserved Network Structures	No	Web-based
QNet	2	Computes graph similarities from trees for the nodes based on colouring graph theory	Nodes	Full network	No	In-house software
SAGA	2	Computes graph similarities for the nodes	Nodes	node gaps, node mismatches and graph structural differences	No	Web-based

Net[136], CSD[101], DiffCorr[145;149], Gain[144], MIMO[143], NetAlign[138], SAGA[140] and QNet[142]. From those, only the CoDiNA and CompNet are methods that were developed for the comparison of two or more networks simultaneously. Table 5.1 describes briefly the main methods for differential network analysis. However, as mentioned before, it is difficult to quantitatively evaluate the accuracy of any of these approaches, because a set of gold standard experimentally validated networks does not exist[148].

5.3 CoDiNA in a nutshell

I will first describe briefly the idea behind the method, and in Subsection 5.3.1 I describe the algorithm and the method in details.

To perform a comparison of co-expression networks, CoDiNA requires as input a set of networks to be assessed (Figures 5.1a, 5.1b, 5.1c). The networks can be constructed using a correlation method, but should only contain links that are statistically significant given a predefined $p-$value threshold, links that are not significant should have its weight set to zero.

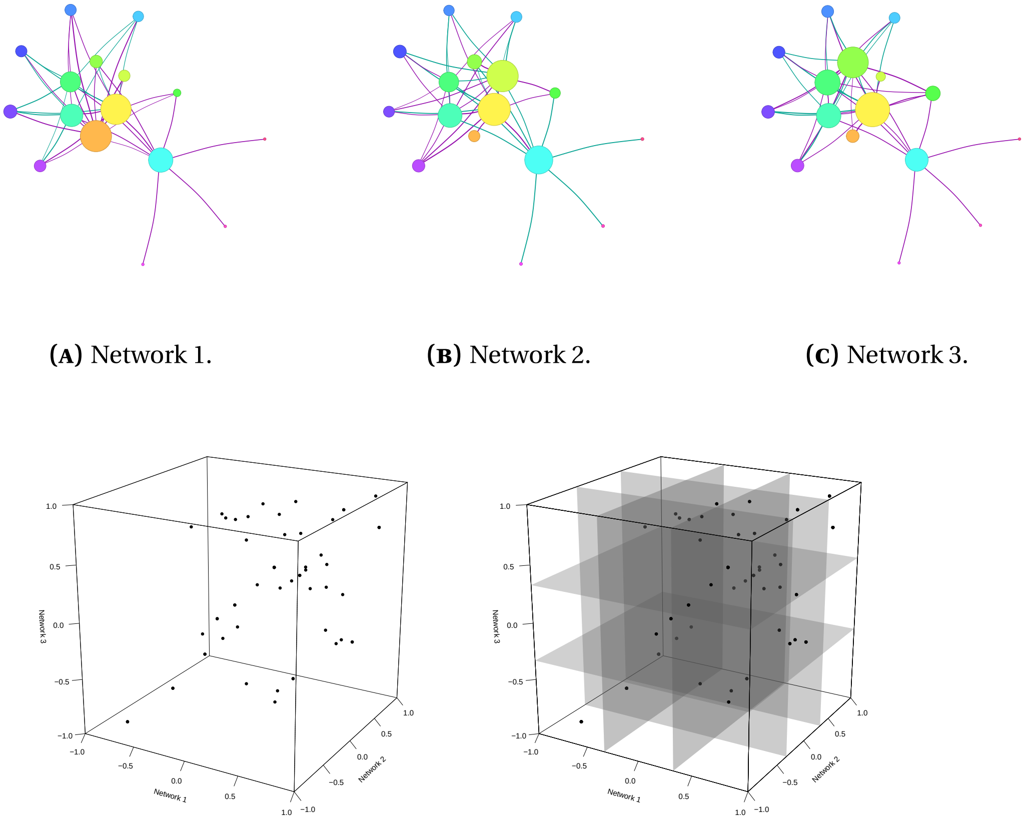
In order to avoid false associations, i.e. the incorrect inference that a particular gene is associated with a specific condition, the method requires that all investigated **nodes** are present in **all** networks: If a node is absent in at **least one** network, all of its links are removed from the networks in which it is present. This does not apply to nodes that are present but have no significant links. In this case, it is assigned a (weight) value of zero, thus allowing all measured node to be included in the analysis, even when they only have non-significant links.

The weight value of the link between the genes i and j, denoted by $\rho_{i,j}$, is defined within the interval $[-1, 1]$ (Figure 5.1d). To denote links as positive, negative, or neutral this interval is divided into three (equal) parts (Figure 5.1e). To compare networks whose intervals might vary, one option is to normalise the data inside the interval to $[-1, 1]$; I refer to this approach as *stretch*. This step is particularly important to compare networks that were not measured under identical and therefore comparable experimental conditions.

Each link is classified into one of three Φ categories based on its weight:

1. A link is classified as α if it is present in all networks with the same sign, i.e., it is an interaction that is common to all networks. In gene co-expression networks, it means that if the same interaction exists across all networks under comparison, probably, would be a high cost for changing this regulatory process (Figure 5.2a);

2. A link is called β if it is present in all networks but with different signs of the link's weight, i.e., it represents a different kind of interaction in at least one network. The biological interpretation of this category is that a particular gene changed its function so that a gene that up-regulates another gene in one condition down-regulates the same gene in another condition (or vice-versa). This is one of the most important interactions one should look for. Changes in the regulatory process might indicate a change of the gene function (Figure 5.2b);



(a) Network 1.

(b) Network 2.

(c) Network 3.

(d) Weights scatterplot of the networks.

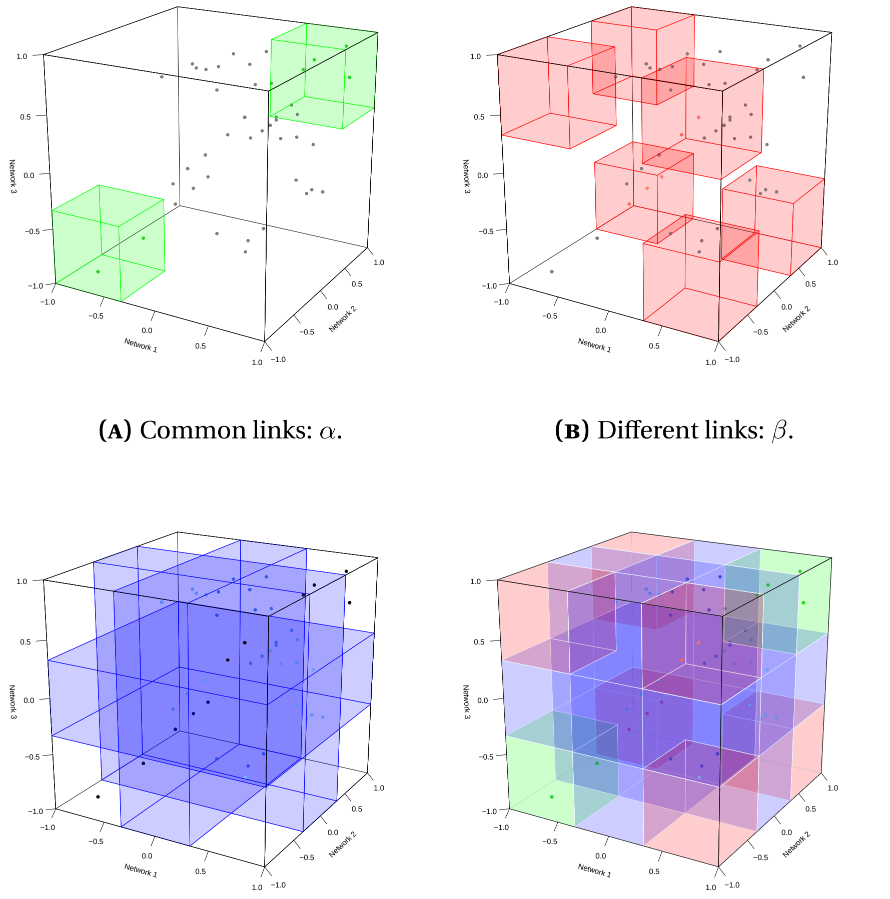
(e) Divide the intervals.

Figure 5.1: Visual representation of the CoDiNA method for a 3-network comparison: Areas definition. **5.1a**, **5.1b** and **5.1c** display three independent networks to be compared; violet links represent positively correlated gene-pairs, and green links negatively correlated ones. Node-size is relative to node strength. **5.1d** shows the geometrical representation of CoDiNA: a 3D scatter-plot that is derived from plotting the weights of each link in the three networks. **5.1e** displays the *slices* on the cube based on τ.

3. A link is considered a γ link, if it is present in some networks but not all, regardless of the sign of the link's weight, i.e., this link is specific to at least one network. This category pinpoints rewiring on the network topology, meaning that genes can start or stop regulating another gene and might indicate changes in metabolic pathways (Figure 5.2c).

To further characterise *how* a particular link is different or specific, a subcategory have to be assigned, $\widetilde{\Phi}$ (Figure 5.2d). This subcategory clarifies to which condition a link is specific or in which condition it has changed.

After all links are classified, they receive a score that is used to filter the networks for background noise (Figure 5.3a and Figure 5.3b). After applying the filter to the network (Figure 5.3c), the nodes have to be classified. This is done for each gene by testing the hypothesis



(A) Common links: α.

(B) Different links: β.

(C) Specific links: γ.

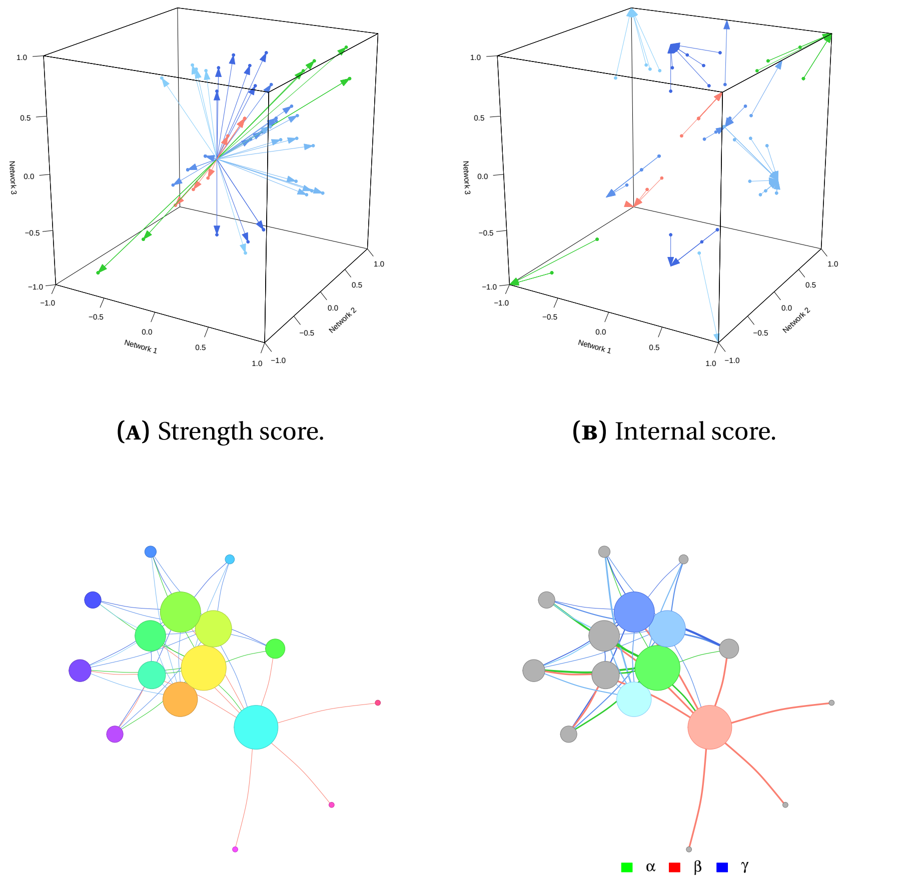
(D) All categories and subcategories.

FIGURE 5.2: Visual representation of the CoDiNA method for a 3-network comparison: Categories definition. **5.1e** displays the *slices* on the cube based on τ. **5.2a** represents where the α links lie on the 3D space; similarly, **5.2b** represents the β and **5.2c**, γ. The full Φ and $\widetilde{\Phi}$ positions can be seem on **5.2d**.

of the frequency of links in each Φ and $\widetilde{\Phi}$ categories are different than expected by chance, using a χ^2 goodness-of-fit (Figure 5.3d).

5.3.1 How the magic works

The notation used here is the same as previously presented in Section 3.2. For a set of nodes $\mathbb{N}$, $\mathbb{N} = \{N_1, \ldots, N_n\}$, where the indices in the nodes are denoted by: i or j. The total number of nodes in a network is n. $\mathbb{W}$ is a set of **independent** networks $\mathbb{W} = \{W_1, \ldots, W_w\}$. The index for the networks is k and the total number of networks is w. Each links has a specific link weight, $\rho_{ij,k} \in [-1, 1]$, connecting nodes i and j in a network k. If the compared networks have different link-weight ranges, these may be normalised by using a multiplicative



(A) Strength score.

(B) Internal score.

(C) CoDiNA network.

(D) CoDiNA network: Nodes classified.

FIGURE 5.3: Visual representation of the CoDiNA method for a 3-network comparison: Scores definition. The scores are shown in Δ^* 5.3a and $\Delta_{\widetilde{\rho}}$ 5.3b. The filtered network only for strong and well-classified links are displayed in 5.3c. Finally, the network where the nodes and links are classified is represented in 5.3d.

(*stretch*) parameter. This parameter forces the $\rho_{ij,k}$ values to be inside this interval. This is particularly important for comparing networks constructed from different and not directly comparable measures.

In order to avoid false associations, an important step to be aware of is that a *node* should be present in all networks; if a node is absent in at least one network, we remove all of its links in the networks where this node is present (Algorithm 1). This step is implemented to prevent the erroneous inference that a particular node is associated with a specific condition, when in fact that specific node possibly was not measured in the other conditions. If a link weight is found not to be significant, we assign it a (weight) value of zero, thus allowing all measured nodes to be included in the analysis even when only some of its links are significant.

Algorithm 1 Description of the RemoveNodes procedure

```
Input: Set of $\mathbb{N}$ nodes that belongs to the each network from the set $\mathbb{W}$
Output: Set of common nodes to all $w$ networks
procedure REMOVENODES($N_1, \cdots, N_n$)
    Set_Nodes = $\bigcap_{k=1}^{w}$ Nodes$_k$
    return Set_Nodes.
end procedure
```

Next, the link weight in each network is categorised. By default, the interval is partitioned into three equal parts ($\tau = 1/3$), which will be denoted as corresponding to a positive link, negative link or neutral link (Algorithm 2). Each link is categorised as

$$\widetilde{\rho}_{ij,k} = \begin{cases} -1 & \text{if } \rho_{ij,k} < -\tau \\ 1 & \text{if } \rho_{ij,k} > \tau \\ 0 & \text{otherwise,} \end{cases}$$

where $\widetilde{\rho}_{ij,k}$ is an integer transformation of the link weight based on the threshold, and τ. If a particular link categorical weight $\widetilde{\rho}_{ij,k}$ is zero in all the $w$networks, this link is removed from posterior analyses.

Algorithm 2 Description of the links categorisation algorithm

```
Input: Set of $\mathbb{W}$ networks with $\mathbb{N}$ nodes ($w \geqslant 2; n \geqslant 2$)
Output: Links weight categorised into $-1$, $0$ or $1$
Set $\tau > 0$;
By default $\tau \leftarrow 1/3$;
procedure ASSIGNCLASSES
    for $\rho_{ij} \leftarrow 1$ to $e$ do
        for $\rho_{ij}$ $k \leftarrow 1$ to $w$ do
            if $\rho_{ij,k} < \tau$ then
                $\widetilde{\rho}_{ij,k} \leftarrow -1$;
            else if $j > \tau$ then
                $\widetilde{\rho}_{ij,k} \leftarrow 1$;
            else
                $\widetilde{\rho}_{ij,k} \leftarrow 0$;
            end if
        end for
    end for
end procedure
```

After the correlation values are coded into the categorical variables $\widetilde{\rho}_{ij,k}$, each link is assigned to an additional subcategory, $\widetilde{\Phi}$, that shows in which condition the link is present and what is its sign, if present.

The classification approach assigns an α, β or γ to each of the links by defining Φ as

$$\Phi_{ij} = \begin{cases} \alpha & \text{if } \sum_{k=1}^{w} |\widetilde{\rho}_{ij,k}| = w \wedge |\sum_{k=1}^{w} \widetilde{\rho}_{ij,k}| = w \\ \beta & \text{if } \sum_{k=1}^{w} |\widetilde{\rho}_{ij,k}| = w \wedge |\sum_{k=1}^{w} \widetilde{\rho}_{ij,k}| < w \\ \gamma & \text{otherwise.} \end{cases}$$

The Algorithm 3 describes this process. Each link receives a subcategory, $\widetilde{\Phi}$, based on the pattern of networks in which that link exists. This makes it more straightforward to interpret the links in each of the Φ categories, and as a result, this improves our ability to identify links that are specific or it has a different behaviour to a subset of networks. This classification step is particularly important for links that are classified as β or γ type, because it is a clear identification in which network(s) the link is specific or different. Moreover, a group is given. The group gives the exact sign for a link in all networks. The maximum number of group is $(3^w - 1)$. Note that, the group where all categorical values are equal to zero is removed from analyses.

Algorithm 3 Description of the Φ algorithm

Input: Set of $\mathbb{W}$ networks with $\mathbb{N}$ nodes ($W \geqslant 2; N \geqslant 2$)
Output: Network with links categorised into α, β or γ

```
Set $\tau > 0$;
procedure PhiLinks
    for $\widetilde{\rho}\; i \leftarrow 1$ to $e$ do
        for $\widetilde{\rho_{i,j}}\; k \leftarrow 1$ to $w$ do
            if $\sum_{j}^{w} |\widetilde{\rho}_{ij,k}| = 0$ then
                remove link;
            else if $\sum_{j}^{w} |\widetilde{\rho}_{ij,k}| \;\&\; \sum_{j}^{w} \widetilde{\rho}_{ij,k} = w$ then
                $\Phi_{ij} \leftarrow \alpha$;
            else if $(\sum_{j}^{w} |\widetilde{\rho}_{ij,k}| \;\&\; \sum_{j}^{w} \widetilde{\rho}_{ij,k}) \neq w$ then
                $\Phi_{ij} \leftarrow \beta$;
            else
                $\Phi_{ij} \leftarrow \gamma$;
            end if
            Calculate the penalised Euclidean Distance $\Delta^{*}_{ij}$ (Equation 5.1).
            Calculate the normalised penalised Euclidean Distance $\Delta^{**}_{ij}$ (Equation 5.2).
        end for
    end for
end procedure
```

To illustrate the concept of subcategory, assume the following $\widetilde{\rho}$ of a particular link in 3 networks: Network$_A = 1$; Network$_B = 1$ and Network$_C = 1$. Because the value 1 is common in the three networks, this Φ category is clearly α, and no further explanation is needed. Now, take as a second example, Network$_A = 1$; Network$_B = -1$ and Network$_C = 1$. Its Φ class is β, but this class cannot help us understand where the change occurs, therefore, the $\widetilde{\Phi}$ is needed. Its $\widetilde{\Phi}$ class is β_{ν_B}. Important to note is that CoDiNA assumes the first network to be the reference network. And ν is a vector with all networks that the value is different from the reference network, in this case, Network$_A$. As a final example, assume that the $\widetilde{\rho}$ weight of the three networks are 0, 1 and 1 for Networks A, B and C, respectively. This link does not occur in network A, so it is a γ links, that is specific to networks B and C. But this is not possible to understand only by reading that its category is γ, therefore, its $\widetilde{\Phi}$ category is $\gamma_{\nu_{B.C}}$.

Let, $\mathbb{L}$, be the set of links $\mathbb{L} = \{L_1, \ldots, L_l\}$ and l the amount of links in $\mathbb{L}$. When all $\mathbb{L}$ links are assigned a Φ category and further subcategorised as $\widetilde{\Phi}$, it is necessary to score the links to identify those that are stronger. For every link $i = 1, \ldots, l$, we interpret the array of link weights $(\rho_{1i}, \ldots, \rho_{wi})$ as a point in a w-dimensional Euclidean space. In particular, as each link weight is bounded, all points are contained in the cube determined by the Cartesian product $[-1, 1]^w$.

As such, a link that is closer to the centre of the w-dimensional cube is weaker than a link closer to the links. Based on that, the Euclidean distance, Δ, to the origin of the space is calculated for all links E_{ij} as

$$\begin{aligned}\Delta_{ij} &= \sqrt{(\rho_{ij,1})^2 + \cdots + (\rho_{ij,w})^2} \\ &= \sqrt{\sum_{k=1}^{w} \rho_{ij,k}^2}.\end{aligned}$$

However, since links closer to corners will trivially have a larger Δ compared to the others, all distances are penalised by the maximum theoretical distance a link can assume in its category. Consequently, we define a penalised distance, Δ^*, as

$$\Delta_{ij}^* = \sqrt{\frac{\sum_{k=1}^{w} \rho_{ij,k}^2}{\sum_{k=1}^{w} |\widetilde{\rho}_{ij,k}|}}, \tag{5.1}$$

which lies in the unit interval.

A second step it to normalise the resulting values in each Φ and $\widetilde{\Phi}$ categories. I refer to it as Δ^{**}. Normalising the distance can be a way to overcome the challenge of some categories having more links than others. This measure is defined as

$$\Delta_{ij}^{**} = \frac{\Delta_{ij}^* - \min\{\Delta_{ij}^*\}}{\max\{\Delta_{ij}^*\} - \min\{\Delta_{ij}^*\}}. \tag{5.2}$$

Three different approaches may be applied to the normalisation:

- Normalise all the links together: Here, it is not considered if a complete cluster is situated near the surface or closer to the centre of the cube;

- Normalise links according to their Φ and $\widetilde{\Phi}$ class: In this alternative, all the categories are a part of the final output. This means that if one of the Φ groups lies inside the cube closer to its centre compared to the other Φ categories, it will be possible to see links that belong to this category in the final network.

Another important score calculated by CoDiNA, called internal Score, denoted by $\Delta_{\widetilde{\rho}}$, measures the distance from the link ij to the theoretical best well-clustered link in that particular $\widetilde{\Phi}$ category. In other words, if a link is considered an α with all positive links, we calculate its distance to the point $(1, 1, 1)$. This score allows us to identify links that are most well defined for each $\widetilde{\Phi}$ category.

Because the two scores Δ^{**} and $\Delta_{\widetilde{\rho}}$ are highly negatively correlated, the ratio between them also gives us a measure of the very best well-defined links. For a well defined not stretched CoDiNA network, this ratio should be greater or equal than 1.

Knowing only the links classification is not sufficient to describe a network; we are also interested in the nodes' classification. To define the Φ category of a particular node, we make a frequency table of how many times each node had a link in each Φ category and $\widetilde{\Phi}$ subcategory. Using a χ^2 goodness-of-fit test the hypothesis that the links of a node are distributed equally in all categories are tested. If the null hypothesis is rejected, the Φ -category with the maximum number of links is assigned to that particular node. Similarly, the same is done for the $\widetilde{\Phi}$ (Algorithm 4).

Algorithm 4 Description of the node-categorisation algorithm

```
Input: Set of ℕ nodes with 𝕃 links ($l \geqslant 2$; $n \geqslant 2$)
Output: Node classified as $\alpha$, $\beta$ or $\gamma$ type
procedure PhiNodes
    for $i \leftarrow 1$ to $n$ do
        $\Phi_\alpha$ = Count $\alpha$;
        $\Phi_\beta$ = Count $\beta$;
        $\Phi_\gamma$ = Count $\gamma$;
        Test if $\Phi_\alpha \neq \Phi_\beta \neq \Phi_\gamma$
    end for
end procedure
```

Algorithm 5 shows the complete pseudocode for the CoDiNA method.

Algorithm 5 Description of the `CoDiNA` algorithm

```
Call: RemoveNodes
Call: AssignClasses
Call: PhiLinks
Call: PhiNodes
```

5.4 A package to compare multiple networks

To make the proposed methodology publicly available, an R package, called CoDiNA, where all the presented steps are implemented. The R package also includes an interactive visualisation tool, similar to the one presented in Chapter 4, its workflow analysis is presented in Figure 5.4. The functions included in the package are:

- normalize(): Normalises a variable according to Equation 5.2;
- OrderNames(): Reorder the names of the nodes for each link in alphabetical order;
- MakeDiffNet(): Categorise all the links into Φ, $\widetilde{\Phi}$ and also the group. It also computes the normalised scores;
- plot: Classifies the nodes into Φ and $\widetilde{\Phi}$ following a user-defined cutoff for the chosen distance and plots the network in an interactive graph, where nodes and links can be dragged, clicked and chosen according to its group or classification. The size of a node is relative to its degree. Nodes and links that belong to the α (*common*) group are coloured in shades of green; Nodes belonging to the β (*different*) group are coloured in shades of red; Nodes of the γ (*specific*) group are coloured in shades of blue; Nodes have a category for group and Φ or $\widetilde{\Phi}$, according to a χ^2-goodness of fit test as defined above. If a node is group-undetermined and it is grey coloured. The user can also choose a layout for the network visualisation from those available in the igraph package[135]. It is further possible to cluster nodes, using the parameter MakeGroups, and the user may select among the following clustering algorithms: "walktrap"[150], "optimal"[151], "spinglass"[152–154], "edge.betweenness"[155;156], "fast_greedy"[157], "infomap"[158;159], "louvain"[160], "label_prop"[161] and "leading_eigen"[162]. These algorithms are implemented in the igraph package[135];

The CoDiNA package also contains three datasets for illustrative purposes.

- The AST data.table contains the nodes and the weighted Topological Overlap (wTO) of Transcription Factors (TFs), from GSE4290[163] for astrocytomas;
- The GLI data.table contains the nodes and the wTO of TFs, from GSE4290[163] for glioblastomas;
- The OLI data.table contains the nodes and the wTO of TFs, from GSE4290[163] for oligodendrogliomas;
- And the CTR data.table contains the nodes and the wTO of TFs, from GSE4290[163] for controls.

CoDiNA is open source and freely available from CRAN https://cran.r-project.org/web/packages/CoDiNA/ under the GPL-2 Open Source License, and it is platform independent.

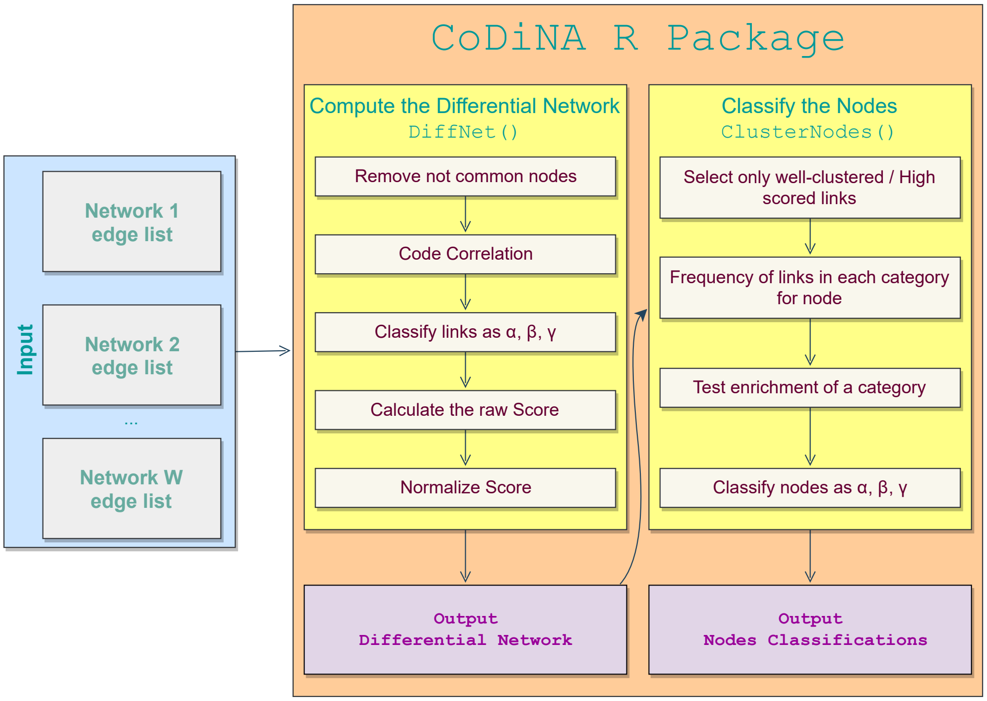

CoDiNA R Package
Input
Network 1 edge list
Network 2 edge list
...
Network W edge list
Compute the Differential Network
DiffNet()
Remove not common nodes
Code Correlation
Classify links as α, β, γ
Calculate the raw Score
Normalize Score
Classify the Nodes
ClusterNodes()
Select only well-clustered / High scored links
Frequency of links in each category for node
Test enrichment of a category
Classify nodes as α, β, γ
Output Differential Network
Output Nodes Classifications

FIGURE 5.4: Workflow process of the CoDiNA `R` package. Input data for the CoDiNA `R` package can be any network, filtered for containing only significant links. Edge list is a list containing all the links and its weights. To links for which the $p-$value is not significant, the user can assign a weight of zero. The function `MakeDiffNet()` clusters the links into the Φ and $\widetilde{\Phi}$ categories, calculates and normalises the scores. Its output is used as input for clustering the nodes into categories by the function `ClusterNodes()`. The `plot()` function can be used on the output from `MakeDiffNett()` and automatically calls the function `ClusterNodes()`.

6
Make me Rich

"I don't need no money
As long as I can feel the beat
I don't need no money
As long as I keep dancing"

– Cheap Thrills, *Sia*

AFTER CONSTRUCTING and comparing a set of networks it becomes natural to understand and make inference whether the transcripts associated with a particular phenotype are found by chance. Therefore an enrichment analysis is necessary. In this Chapter, I will present a robust and efficient approach that aims to find enrichment of gene-sets described as associated with disorders, Gene-Disease Associations (GDA). This methodology is publicly available as an `R` package, `RichR` that contains a dataset, Gene-to-Disorder (g2d), that was manually curated and combines information from the five most up-to-date studies on GDA into one.

This methodology is publicly available on the `CRAN` as an `R` package, called `RichR`. A publication regarding this methodology is under preparation.

- **D. M. Gysi** and K. Nowick. "Make me RichR: an R enrichment package". *In preparation.* 2019.

6.1 Motivation

A functional enrichment analysis can help to characterise large lists of candidate genes associated with functions, diseases, biological processes among others. In general, this analysis consists of comparing the gene-set against a background list and testing if there is a significant difference in gene functions[164].

The functional annotation can be retrieved from several databases such as Gene Ontology (GO), that returns information on molecular and biological functions (GO), metabolic pathway (Kyoto Encyclopedia of Genes and Genomes (KEGG)[165], WikiPathways[166]), experimental databases (Encyclopedia of DNA Elements (ENCODE)[167]). When a set of genes involved in the same function is also significant in the candidate list, it is more likely to be a relevant function in the candidate list.

The same rationale can be applied for a **disease enrichment**. In one hand many tools are able to make the enrichment analysis for GO (e. g. EasyGO[168], GOrilla[169], topGO[170], GSEA[171]) on the other very few are able to deal with disorders (DisGeNET[172], GS2D[164] and PsyGenNET[173]). Even though these tools are able to correct for the background, all have the background fixed and defined by the genes each database contains and it cannot be changed by the user. It becomes a problem when dealing with a subset of transcripts, such as expressed genes, Transcription Factors (TFs), non-coding RNA (ncRNA) or a set of orthologs. Therefore, a tool that allows the user to define its own background set is necessary.

6.2 Construction of a curated dataset of Gene-Disease association

I constructed and manually curated a database that associates genes to disorders using multiple tools and publication that incorporates information on genes and associated disorders. The datasets combined here collects information about genes and disorders from PubMed, Mesh and Genome Wide Association Study (GWAS) studies.

6.2.1 Data collection

Data of association of genes to disorders were retrieved from multiple tools. Those are the four most up-to-date and complete tools that combines information from Gene-Disease Associations (GDA): Gene Set to Disease (GS2D)[164], Disease to Genes Network (DisGeNET)[172], Berto2016[54] and Psychiatric disorders Gene association NETwork (PsyGeNET)[173]. Another important dataset is eDGAR[174] that could be included, however, it does not allow multiple searches and the download of a formated table.

The GS2D[164] is available online from http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=605 and contains data from National Center for Biotechnology Information (NCBI), PubMed and MeSH. All genes are identified by official Entrez Gene IDs, and the diseases are identified by MeSH C terms. The Disease-related citations are retrieved from annotations in PubMed. I filtered this dataset in order to retain only genes that had at least two publications that associated it with the same disease. Because this tool by its own may not reflect the most current and accurate biomedical/scientific data available from National Library of Medicine (NLM), I also included data from other datasets.

From DisGeNET, I used the curated data, that integrates data from UniProt, PsyGeNET, ClinVar, Orphanet, the GWAS Catalog, CTD (human data) and Human Phenotype Ontology. This database is homogeneously annotated with controlled vocabularies.

Berto et al.[54] made available a list containing GDA for mental disorders and brain development. For Autism Spectrum Disorder (ASD) his list was constructed using SFARI gene database[175–177] and a genome-wide differential expression study[178]. For other Gene Regulatory Factors (GRFs) associated with Parkinson's disease, Alzheimer's disease, and Schizophrenia they used results from GWAS studies[179–182] and independent publications[183–188], the brain development genes were manually selected from independent publications[181;189] and Online Mendelian Inheritance in Man (OMIM).

PsyGeNET is a tool that collects data for GDA of psychiatric diseases. This database has been developed by automatic extraction of information from the literature using the text mining tool BeFree and contains updated information on depression, bipolar disorder, alcohol use disorders and cocaine use disorders, and has been expanded to cover other psychiatric diseases of interest: bipolar disorder, schizophrenia, substance-induced depressive disorder and psychoses and cannabis use disorder.

The overlap of those databases is very small (0.004%), the biggest share of genes and diseases associations that is reported in only one database (95.98%) (Figure 6.1). This shows that it is important to consider all datasets in order to obtain a more robust Gene-to-Disorder (g2d) dataset. This lack of overlap is due to the different platforms and methods used by the different tools to retrieve data. Moreover, each tool collected the gene-disorder association from different sources.

6.2.2 Curating the data

The final database was filtered for diseases that had more than 4 genes associated with it; this step is important to assure that the final database contains only polygenic diseases.

The next step was to collect only the gene name, its associated diseases and the source of the data. The source of the data refers to the original database that contains this information. In all five databases, it is possible to retrieve the publications that identified the association of the gene to the disease.

Genes had all their names matched to Entrez ID using the `BioMart R` package[190;191]. Diseases kept the name it appeared in the original dataset, however, the same disease might have different orthography in the five databases. Thus, similar terms were merged into one. For example: *Major Depression* and *Major Depression Disorder* were considered to be the same disorder. After assuring that the same disorder would have the same term in all diseases, each gene-disorder pair was marked with the datasets that had this information.

In total, the new (cleaned) g2d database contains $13,028$ genes associated to $5,295$ diseases.

6.3 Enrichment function

In order to calculate the disease enrichment of a set of genes, I wrote the function `Enrichment()`. This function tests the hypothesis, for each disease, that the number of genes found is different than random. Genes are first corrected for the background gene list, given

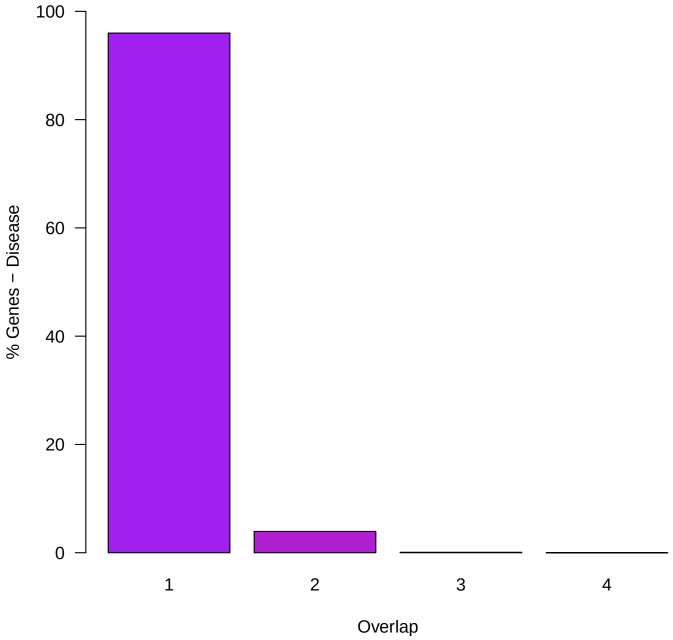


FIGURE 6.1: Overlap of Gene-Diseases in the four databases: The barplot shows the frequency of gene and diseases that repeats on the datasets, after curating diseases names. It becomes clear that the overlap of the information of gene and diseases is quite small, therefore, combining all the information into one dataset is desired.

by the user. Correcting for the background allow us, to identify disease enrichment even in a smaller subset of genes, for example, using only TFs.

The hypothesis test is done using a proportion test and Fisher's Exact test. Both $p-$values are returned and its $p-$values are corrected using Benjamini-Hochberg adjusted $p-$value (BH). The raw $p-$values are combined into one, using Fisher's method for combining $p-$value. Previously presented in Equation 4.6.

This tool is used in Chapter 8 to define gene enrichment for mental disorders.

III

Gene co-expression networks reveals important regulators in the brain

7

Neuronal Development

"Once you begin to appreciate the structure of the mind, there's no reason anything about us can't be changed. Pain can be destroyed. The mind can be solved."

– Maniac, *James Mantleray*

NON CODING RNAs can regulate many biological processes, including the one of an unspecialized cell differentiating itself in a neuron, known as neurogenesis. The brain-enriched miR−124, a particular micro RNA (miRNA), is assigned as a key player of neuronal differentiation via its complex, but little understood, regulation of thousands of annotated targets. To systematically understand its regulatory functions, I used Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas9 human stem cells where all the six miR−124 alleles were edited and disrupted to construct, analyse and compare its regulatory network. Under neuronal induction, miR−124-deleted cells experienced neurogenesis and became functional neurons, even though it had altered morphology and neurotransmitter specification. I performed a Transcription Factor (TF)-network analysis and revealed indirect miR−124 effects on apoptosis and neuronal sub-type differentiation. The results emphasise the need for combined experimental- and systems-level analyses to comprehensively untangle and reveal miRNA functions, including their involvement in the neurogenesis of diverse neuronal cell types found in the human brain.

The results presented in this Chapter are based on my shared first authorship paper published in *Cell Systems*.

- L. K. Kutsche*, **D. M. Gysi***, J. Fallmann, K. Lenk, R. Petri, A. Swiersy, S. D. Klapper, K. Pircs, S. Khattak, P. F. Stadler, J. Jakobsson, K. Nowick, and V. Busskamp. "Combined experimental and system-level analyses reveal the complex regulatory network of miR-124 during human neurogenesis". *Cell Systems*, 7(4), 438–452, 2018.
  *Both authors contributed equally.

7.1 Motivation

The human brain is constituted of more than 300 cell types of neurons, with an undetermined number of subtypes. Their underlying developmental features are essentially unknown. Up-to-date, micro RNAs (miRNAs) have been identified as playing a relevant role in the neurogenesis[192]. These miRNAs bind in a sequence-specific way to messenger RNA (mRNA) transcripts and negatively interfere with the concomitant translation of multiple target transcripts by annealing predominantly at the $3'$ Untranslated Region (UTR).

In the brain, the miR-124 is one of the most abundant miRNAs, and it is associated with processes such as neurogenesis[193;194], cancer[195;196], the control of synaptic functions in mature neurons in health[197–199] and disease[200;201] including cognitive impairment[40;202]. It is still not clear how, and at which developmental stages, miR-124 affects neurons. For instance, miR-124 has been shown to be involved in the initiation of neuronal differentiation[203–210], as well as in the maturation and survival of the differentiated neurons[211–216]. miR-124 different functions have been extensively studied in diverse model systems, such as mouse, chicken, frog, and human immortalised cell lines, with partially contradictory results about miR-124's importance in neurogenesis[199;204;205;207;208;212;214;215;217–225].

Despite previous knockout studies in mouse models were incomplete because not all three miR-124 paralogs, six alleles in total, coding for identical mature miRNAs were simultaneously removed[214]. There is also a high heterogeneity of neuronal ancestor cells *in vivo*, impeding studies of miR-124 in defined cell types[226]. Pooling heterogeneous cell types that differ in their coding and non-coding transcriptome likely results in incomplete views on miR-124's complex regulatory role: there are $4,024$ computationally predicted human transcripts with miR-124 binding sites, as well as further potential non-canonical binding events[227;228]. So far, most studies have experimentally validated only single or very few miR-124 targets at once. It remains unclear how many miR-124 targets are simultaneously regulated within a cell and what their composed impact is – direct and also indirect – via gene regulatory cascades. Therefore, it is essential to investigate miR-124's functions in a well-defined, homogeneous, and complete Knock Out (KO) model system. The data used was obtained from human induced Pluripotent Stem Cell (hiPSC)-based model system to mimic the neurogenesis of bipolar neurons under controlled and reproducible conditions[229]. A complete miR-124 KO (ΔmiR-124) was generated using CRISPR/CAS9 genome editing[230–232], where all the six alleles were deleted. By forced Transcription Factor (TF) induction, the neuronal differentiation in Wild Type (WT) and ΔmiR-124 cells was rapidly and robustly induced. Performing an in-depth molecular, cellular, and physiological characterisation of the ΔmiR-124 and isogenic WT lines revealed altered morphological and functional features, neurotransmitter specification, and decreased long-term viability. 98 miR-124-regulated targets were identified by the RNA-Interaction Protein Immunoprecipitation and subsequent sequencing (RIP-seq)[223;233] by capturing active miRNAs and their mRNA targets bounding to Argonaute-2 (AGO2).

Since my interest lied in functional differences between TF networks in WT and the ΔmiR-124 neurons, these networks were constructed using the weighted Topological Overlap (wTO) method, presented in Chapter 4, that distinguishes positive and negative correlations[54;55] and includes a $p-$value[62] followed by the comparative method I presented in Chapter 5, Co-expression Differential Network Analysis (CoDiNA)[52]. Using this computational approach, I was able to detect similarities and specific differences in WT and ΔmiR-124 neurogenesis at the level of genes and their regulatory connections including the impacts on neurogenesis of the uncharacterised TF ZNF787. This results highlights the complexity of the downstream effects of experimental miRNAs manipulations.

7.2 Methods

HiPSCs, generated from fibroblasts, were reprogrammed to become neurons according to protocols described in Kutsche et al.[58]. Before the neurogenesis was induced, the six alleles that codes for the miR-124 were removed. With that, two cell lines were originated and followed-up for transcriptomic analysis for the next four days and fourteen for morphological and physiological analysis. The cell line where it was KO for the miRNA under study is called ΔmiR-124 and the cells that did not undergo this process are called WT. Both cell colonies had its morphology and physiology analysed to ensure that the final cells were indeed neurons. For details of these experiments, please refer to the original paper.

Seven replicates were collected for each one of the 4 days of the neurogenesis induction for each group: WT and ΔmiR-124. The complementary DNA (cDNA) was sequenced; I processed the Fasta and controlled for quality as using with FASTQC[31] ($v0.11.4$, accessed $2016-09-10$), reads were mapped to the human genome assembly hg38/GRCh38 using Segemehl ($v0.2.0-418$,[34–36]). I extracted only uniquely mapped reads for further analysis. Counts were computed using rnacounter[38] using Reads Per Kilobase Million (RPKM) and raw counts, from the v25 gencode annotation[234]. Differential Expression (DE) was calculated using DESeq2 with raw counts[235]. The contrast I used was ΔmiR-124 versus WT. Normalised Fold Change (nFC) of the significant genes[132] (Benjamini & Hochberg $p_{adj}-$value $<$ 0.05) were used to construct the time series analysis. Z-scores were used to visualise expression analysis per gene. For the ZNF787-associated genes, counts were log-transformed with DESeq2 prior to standardisation.

TFs that were differentially expressed at all timepoints from 1 to 4 days post induction (dpi) were clustered according to their nFC pattern over time using the Self Organising Maps (SOM) algorithm[236], implemented in `R` using the package `SOM`[237]. I increased the number of clusters until the $q-$error of each group was reduced, with the average distortion measure under 10. The membership of the genes to each SOM cluster was used to colour the genes in the (target-) TF-TF network analysis.

I used my `R` package `wTO`[62;238], presented in Chapter 4, to calculate the wTO of the (target-) TF-TF networks. The correlation between a set of genes was corrected using all the other genes present, thus reducing the noise and the false positives, and taking into account the commonalities of those genes. For each one of the ten networks built (for both WT and ΔmiR-124 from day 0 to 4 dpi), the parameters used in this calculation were Pearson correlation coefficient and $1,000$ bootstrap resampling. The final results were filtered to a probability of 0.10 for random wTO. The wTO was calculated based on RPKM values. Genes with RPKM < 5 for each day were removed. From the total of $56,269$ mapped transcripts from the RNA sequencing (RNA-seq) dataset, $39,275$ are considered to be expressed. Only TFs from the target list (assembled from the Gene Regulatory Factor (GRF) Catalogue[54] were considered for the network analysis. The information for GRF Catalogue proteins was sourced from the most seminal studies in the area of human GRF inventories[239–245] and manually curated: these are associated with gene ontology terms for *regulation of transcription, DNA-dependent transcription, RNA polymerase II transcription co-factor and co-repressor activity, chromatin binding, modification, remodelling* or *silencing,* among others. The `wTO R` package was used to visualise the interactive plots and igraph[135] and network[246] `R` packages were used to visualise the steady developmental network. Human or primate specificity was judged according to the Uniprot database[247].

Later, the networks for each day were compared using CoDiNA, presented in Chapter 5. Where links and nodes are classified according to Φ categories, to its commonalities, differences and specificities. links are considered to be common (α) if they belong to a set of networks (WT and ΔmiR-124 for each dpi) with the same sign and similar strength. If the sign changes from one network to another, the link is considered different (β). If a particular link belongs to one network only, it is considered to be specific to this network (γ). α and β categories are intersections of correlated genes between WT and ΔmiR-124; γ links are exclusive for one condition. The classification of the interactions according to these concepts is central to understand how the ΔmiR-124 networks are affected in their TF interactions during the time course. Links were scored as previously described. Only links with a normalised Φ distance greater than 0.5 were kept for further analysis. In order to define the category a TF belongs to, a χ^2 goodness-of-fit test was used to test if the distribution of the links is different than $1/3$ for each category (p_{adj}−value < 0.05). Each TF is classified using the link category (Φ) and subcategory $(\widetilde{\Phi})$ that appears most frequently for that particular TF. Later, the correlation between TFs and genes was measured using a Pearson correlation coefficient. Only absolute correlations above 0.9 were considered for the following analyses. CoDiNA was computed separately for each day. In order to compare WT and ΔmiR-124 networks, the TFs were distinguished according to a category; the names of the genes correlated with each TF were retrieved.

In order to verify if any gene function was enriched in any of the DE genes, Φ and $\widetilde{\Phi}$ CoDiNA groups, Gene Ontology (GO) enrichment analysis was conducted using the `topGO`[170] `R` package using all expressed genes (average RPKM > 10) as background for each day. Semantic clustering was performed with Reduce + Visualise Gene Ontology (ReViGO)[248] using the Semantic similarity scores (SimRel) measure and allowed similarity of 0.9.

7.3 Results

The overall changes in gene expression in the absence of miR-124 was studied by analysing differential gene expression patterns between WT and ΔmiR-124 over neurogenesis course Figure 7.1. I found $2,884$ genes to be DE at 0 dpi, at 1 dpi $10,820$, at 2 dpi $10,897$, at 3 dpi

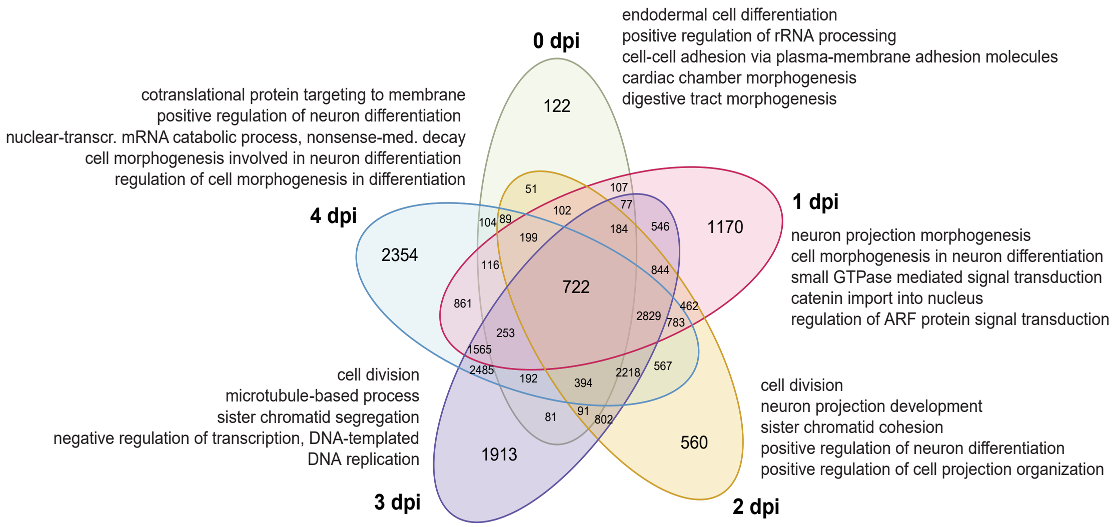

0 dpi
endodermal cell differentiation
positive regulation of rRNA processing
cell-cell adhesion via plasma-membrane adhesion molecules
cardiac chamber morphogenesis
digestive tract morphogenesis
cotranslational protein targeting to membrane
positive regulation of neuron differentiation
nuclear-transcr. mRNA catabolic process, nonsense-med. decay
cell morphogenesis involved in neuron differentiation
regulation of cell morphogenesis in differentiation
4 dpi
1 dpi
neuron projection morphogenesis
cell morphogenesis in neuron differentiation
small GTPase mediated signal transduction
catenin import into nucleus
regulation of ARF protein signal transduction
cell division
microtubule-based process
sister chromatid segregation
negative regulation of transcription, DNA-templated
DNA replication
cell division
neuron projection development
sister chromatid cohesion
positive regulation of neuron differentiation
positive regulation of cell projection organization
3 dpi
2 dpi
122
51
107
77
102
104
89
184
546
1170
199
2354
116
844
722
861
462
2829
783
253
1565
2485
567
2218
192
394
91
81
802
560
1913

FIGURE 7.1: Extended transcriptome analysis Venn diagram for differential gene expression over the time course of differentiation $(0-4$ dpi) comparing WT and ΔmiR-124 with associated GO terms.

$15,196$, and at 4 dpi $15,731$ ($p_{adj}-$value < 0.05; Figure 7.1). These genes showed GO enrichment for *differentiation*, *cell adhesion*, *morphogenesis*, and *cell division*.

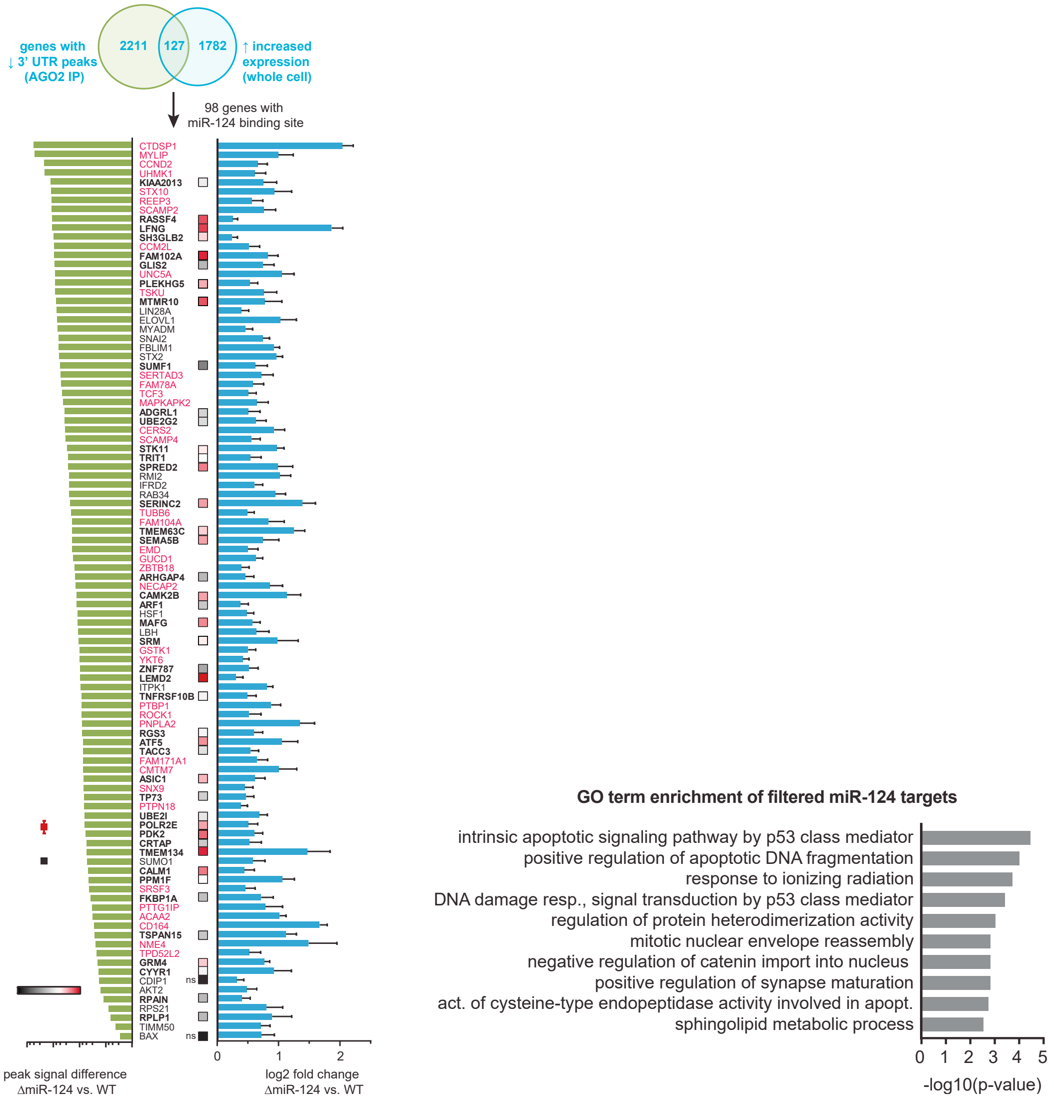
Because miRNAs repress gene expression, it is expected that direct miR-124 targets would show increased mRNA levels after miR-124 KO. It remains inconclusive to use computational predictions of miRNA target genes, mainly due to the unspecificity of miRNAs. In general, they can bind to thousands of targets and most of the mRNAs have multiple miRNA binding sites. Just for the miR-124 there are $4,024$ genes with possible binding sites annotated[249–252]. Using AGO2-RIP-seq for the identification of active miRNAs targets, it was identified 98 high-confidence active miR$-124-$targets (from those, 81 were validated) out of the $4,024$ annotated targets in inducible-Neurogenin Cell Line (iNGN) neurons.

The 98 high-probability targets were analysed to assess their involvement in biological processes using a GO term analysis Figure 7.2b. The GO enriched terms were associated with *synaptic maturation* and *apoptosis* at 4 dpi (Figure 7.2b).

7.3.1 A network of transcription factors is influenced by miR-124

A significant fraction of the identified miR-124 targets (24%, $m = 98$, $p-$value< 0.05; χ^2-test) coded for TFs, suggesting that miR-124 exerts much of its impact via influencing gene regulatory cascades involving many TFs. Measuring indirect miRNA effects is not trivial[253], but essential to understand the full spectrum of miRNA regulation. Therefore, I conducted a time series network analysis focusing on the 24 miR-124 targets coding for TFs to understand the miR-124 regulatory network that underlies the neurogenesis in humans. For each timepoint and, separately, for WT and ΔmiR-124 cells ($m = 7$ for each timepoint), the genes that correlated with the 24 TFs-targets were identified. Correlated genes represent potential target genes or interaction partners of each TFs.

To reveal how similar TFs were to their correlated target genes, the wTO networks were



(A) Differential Expression and miR−124 targets.

(B) GO enrichment analysis.

FIGURE 7.2: Differential Expression and GO enrichment analysis 7.2a Differentially expressed genes were filtered as shown in the Venn diagram. Transcripts with less 30 UTR signal in AGO2-IP data in miR−124 samples were intersected with significantly upregulated transcripts in the whole-cell samples. 127 transcripts overlapped, of which 98 were annotated miR−124 targets. And 98 high-probability miR−124 targets filtered for 30 UTR peak signal decrease (left) and increase in expression (right). Data are presented as mean ± $\log_2$-fold change standard error. Experimentally validated targets according to miRTarBase are indicated in red. 43 additional targets were validated by luciferase reporter assays; color code indicates relative luciferase signal reduction upon miR−124 overexpression. 7.2b GO term enrichment analysis of filtered miR-124 targets indicating their involvement in apoptosis and synaptic maturation.

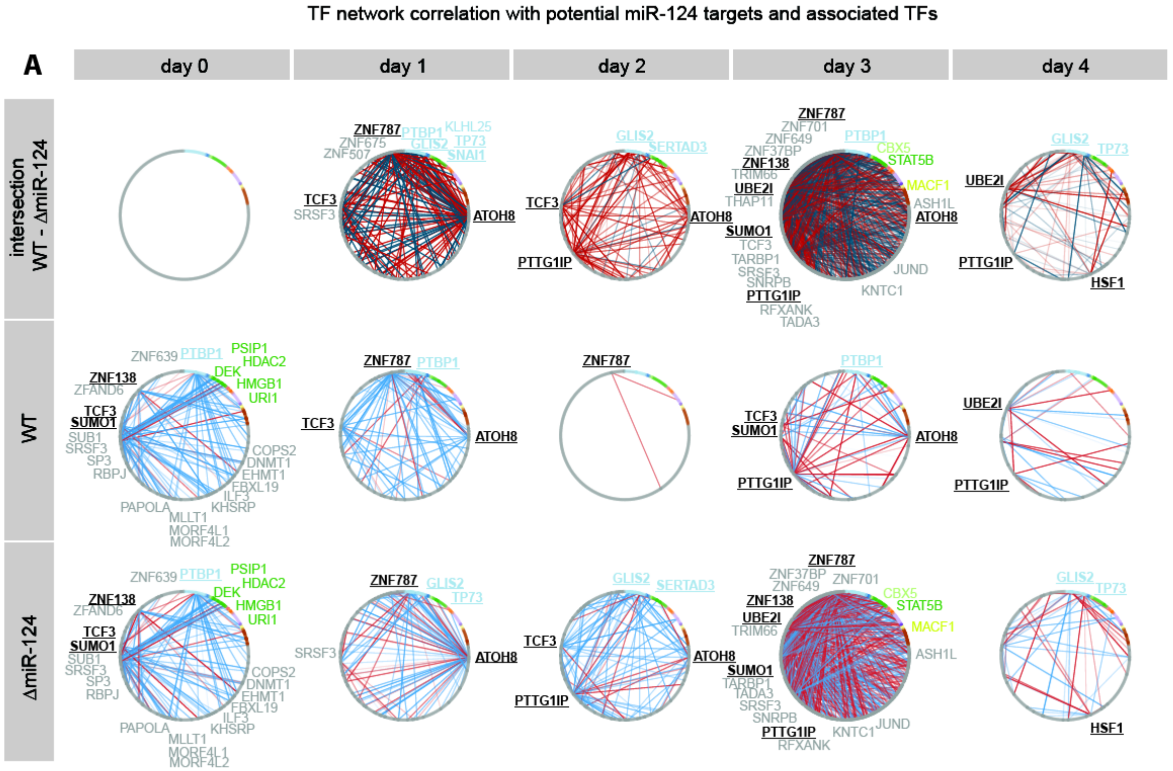
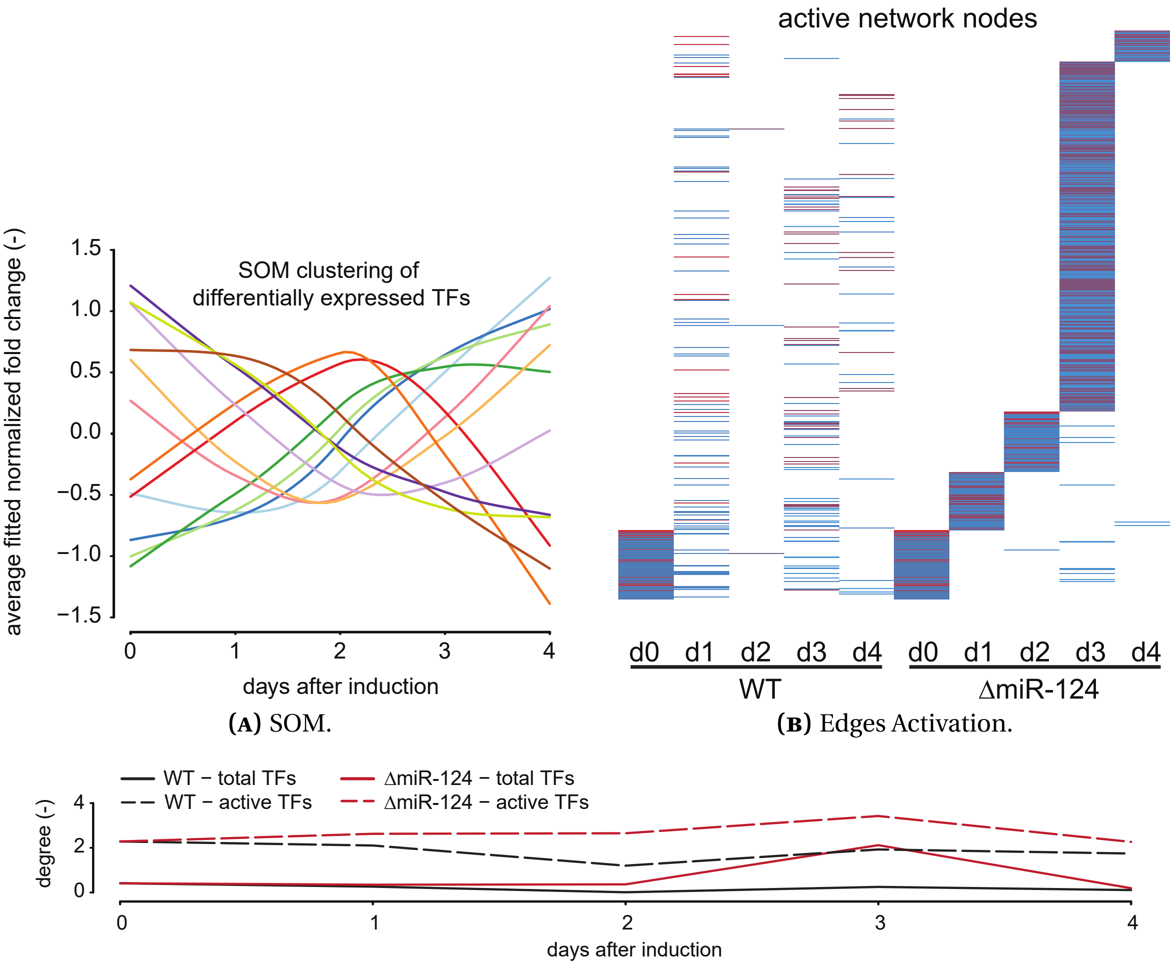


Figure 7.3: Time course network: Expression correlation (as weighted topological overlap, wTO) between TFs that were differentially expressed on at least one day between 1 dpi and 4 dpi, but not on 0 dpi. Differences in interaction ($|wTO^{WT} - wTO^{\Delta miR-124}| > 0.2$) are shown in the top panel. Every panel shows the development of the network during differentiation for the difference (top), WT (middle), and ΔmiR-124 (bottom). The opacity of the line indicates the wTO value. Coloured gene names represent a specific SOM cluster as shown in Figure 7.4a. Underlined TFs are miR-124 targets (Figure 7.2a).

constructed for all expressed TFs for each timepoint for WT and ΔmiR-124 cells separately (Figure 7.3). In these wTO networks, nodes represent TFs, and they are connected by a link if they share a significant number of correlated genes, i.e., are likely working concomitantly in regulating their target genes. As mentioned in Chapter 4, the benefits of Topological Overlap (TO) networks is that they result in more robust definitions of connections and interactions among genes than simple correlation networks, by the reduction of false positive inferences[55;62;115;118;119]. In contrast to the widely-used Weighted Gene Co-Expression Network Analysis (WGCNA)[98;254], the wTO network that I presented here accommodates both positive and negative correlations, which is essential for analysing TFs as they have both enhancing and repressing functions[55;62]. Furthermore, this method also specifies $p-$value to each link, resulting in a high-accuracy network based on seven replicates, which is essential for comparing networks with high confidence[62]. At 0 dpi, WT and ΔmiR-124 networks were identical. Differences on the networks started to appear from 1 dpi, peaked at 3 dpi, and decreased at 4 dpi, Figure 7.3.

To acknowledge the contrast between the WT and ΔmiR-124 networks, the wTO of ΔmiR-124 from the wTO of WT were subtracted, resulting in differential networks (Figure 7.3). Interesting to note is that, in the WT wTO networks, a parcel of links were reactivated at different timepoints, while in the ΔmiR-124 cells most links were unique, suggesting the behaviour of distinct regulatory modes and global changes after miR-124 loss Figure 7.4b.



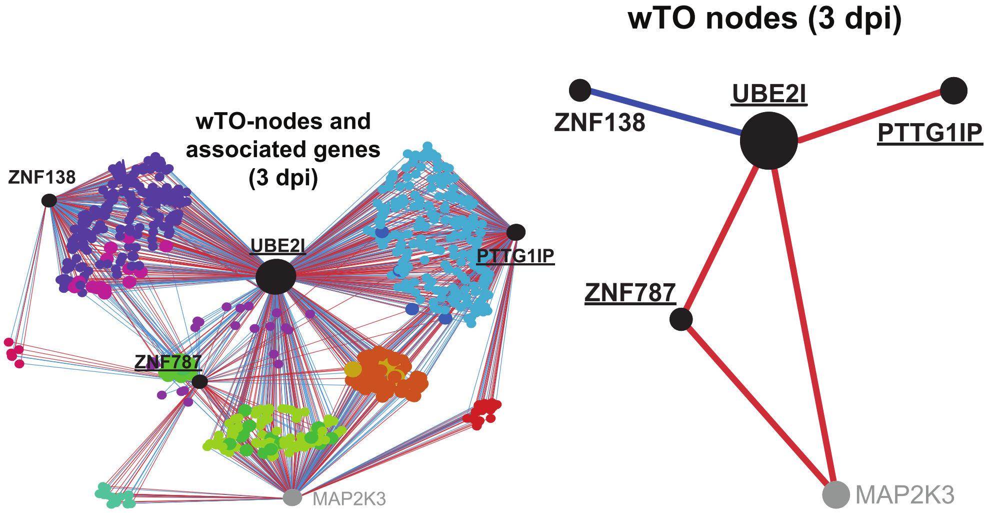
(A) SOM.

(B) Edges Activation.

(C) Average degree.

Figure 7.4: SOM classification, nodes activation 7.4a depics the Loess regression from Self Organising Maps calculated on the basis of normalised fold changes of permanently $(1 - 4$ dpi) differentially expressed TFs. Colour code represents the SOM categories. 7.4b shows the nodes activation of the network depicted gene relationships reappearing on different days. 7.4c presents the degree of nodes for each network for active genes and full set of genes over time, as a measure for network strength.

In the following step, I classified the differentially expressed TFs $(14.95\%, m = 3,145)$ using a SOM algorithm (Figure 7.4a). The SOM categories with a steady increase in DE over time were considered to be the best candidates for also being influenced by miR-124, as they were expected to be upregulated in the absence of the negative regulator (Figure 7.4a).



(A) Correlation network.

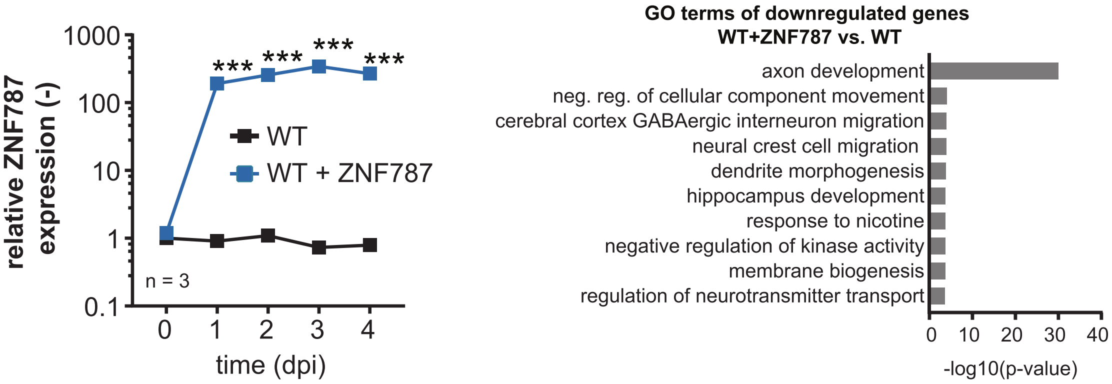
(B) wTO network.

Figure 7.5: Targets of ZNF787: 7.5a is an illustration of a miR−124 target-specific correlation subnetwork showing TF nodes at 3 dpi. Coloured lines indicate negative or positive correlations of underlying associated genes. 7.5b is an illustration of the subnetwork shown in 7.5a, including underlying associated genes.

Most miR−124 targets obtained suggested in by the analysis appeared in the monotonically increasing SOM categories, which is in line with the lack of repressing miR−124 (Figure 7.2a and Figure 7.4a): the targets previously described included PTBP1, which was identified as an important factor within the co-expression network (Figure 7.3). Nodes including GLIS2, SERTAD3, and TP73 appeared to be very important, as these genes fulfilled all criteria: i) they were filtered and validated targets (Figure 7.2a); ii) were top hits in the network analysis; iii) followed a rising trend in the SOM clustering.

Because most of the network difference was detected to happen at 3 dpi, it became the centre for the subsequent analysis (Figure 7.3, Figure 7.5a and Figure 7.5b). I also detected some miR−124 targeted TFs with unknown functions (ZNF787) and a indirectly targeted human-specific (ZNF138) within the dense 3 dpi network of the ΔmiR−124 samples[240;247]. For example, the miR−124 target ZNF787 connects to ZNF138 via UBE2 (Figure 7.5a). Other correlated genes (Figure 7.5b) were extracted from the wTO analysis. This visualisation emphasises how integrated the ZNF787 was within the gene regulatory network upon miR−124 depletion at 3 dpi.

The importance of ZNF787 was evaluated experimentally by perturbating the ZNF787 node by Over Expressing (OE) ZNF787 in WT iNGN cells (Figure 7.6a). The WT-ZNF787-OE cells underwent neurogenesis and were positive for the neuronal marker MAP2. I performed GO term analyses on DE genes between WT and WT-ZNF787-OE ($m = 3$ biological replicates, 4 dpi). Particularly, concentrating on downregulated genes, many neuronal biological processes were significantly repressed (Figure 7.6b). Hence, the data indicated that ZNF787 acts as a neuronal feature repressor. This was in line with ZNF787 containing a KRAB domain[255] that leads to transcriptional repression[256]. Looking at ZNF787-correlated genes



(A) Quantification of ZNF787 OE.

(B) GO enrichment analysis.

FIGURE 7.6: Overexpression of ZNF787 and its GO enrichment: 7.6a Quantification of ZNF787 overexpression (OE) efficiency in WT neurons over time. m = 3 biological replicates. Significance was assessed with unpaired Student's t tests with Holm-Sidak correction for multiple comparisons with $***p-$value 0.001. Data are represented as mean ± SEM; 7.6b shows the GO term enrichment analysis of significantly downregulated transcripts ($p_{adj}-$value < 0.05, $\log_2$-fold change < 1) upon ZNF787 overexpression indicating its impact on repressing neuronal differentiation and maturation.

derived from our wTO analysis, corresponding expression levels massively differed between WT, ΔmiR-124 and WT-ZNF787-OE (Figure Figure 7.8). Specifically, 51 out of 78 ZNF787-associated genes showed a similar expression trend for ΔmiR-124 and WT-ZNF787-OE in comparison to WT (Spearman correlation, $\rho = 0.498$; $p-$value < 0.01; Pearson correlation, $\rho = 0.277$; $p-$value < 0.01). Hence, ZNF787 behaves as one mediator of miR-124 activity.

This highlights ZNF787's regulatory influence on the gene regulatory network when its expression was enhanced due to miR-124 depletion or when OE. These variations could lead to the detected downregulation of maturation-associated genes (Figure 7.6b) since known neurogenesis-regulating factors, such as PTBP1, SP1 and HES5, were upregulated[207;257–259].

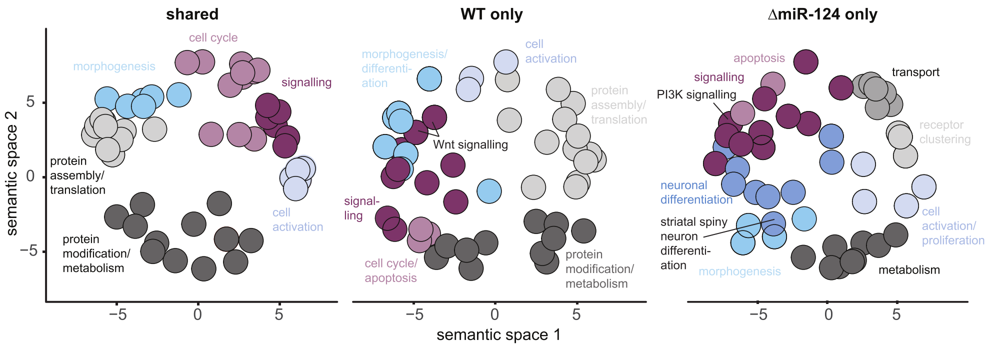


FIGURE 7.7: Semantic clustering of GO terms reveals shared and different biological processes at 3 dpi in WT and ΔmiR-124 cells: Co-expression differential network analysis of shared first graph WT-specific middle graph, or ΔmiR-124-specific last graph regulated genes at 3 dpi. Underlying biological processes are grouped and highlighted.

However, ZNF787 OE does not impede neurogenesis *per se* as the cells are still positive for MAP2. In summary, the wTO analysis suggested that the TF networks were globally altered and differentially connected, especially at 3 dpi upon miR-124 depletion. In addition, my analysis identified uncharacterised TFs - of which ZNF787 was experimentally validated - having regulatory functions during neurogenesis.

7.3.2 Enrichment analysis of wTO network nodes reveals indirect miR-124 functions

Later, I investigated which biological functions were controlled by the TF networks of WT and ΔmiR-124 cells at each timepoint, particularly which functions were common or different between the networks. For that, I used CoDiNA[52;260] and classified each $\omega_{i,j}$ value into common or specific networks (Figure Figure 7.7, Figure 7.4c), followed by GO enrichment tests for the categories to report the biological processes.

GO groups that were common between the WT and ΔmiR-124 networks at 0 dpi included, for example, *mRNA processing*, *cell division*, and *mitotic cell cycle* (Figure 7.7). In particular, WT and ΔmiR-124 networks shared the groups *regulation of asymmetric cell division* and *regulation of extracellular matrix disassembly* at 1 dpi, and *synapse assembly*, *regulation of synapse organization*, and *positive regulation of neurological system process* at 4 dpi, indicating that aspects of neurogenesis were also present in ΔmiR-124 cells.

Groups that were specific to WT network were also found. For example, the GO terms *layer formation in cerebral cortex* and *pyramidal neuron development* at 1 dpi, *regulation of exit from mitosis* and *positive regulation of long-term neuronal synaptic plasticity* at 2 dpi, *positive regulation of dendritic cell differentiation* at 3 dpi, and several *mRNA processing* groups at 4 dpi were different, detailed information can be found in S1 in the publication. Metabolic terms and cell signaling pathways (Wnt signaling at 3 dpi) were also specific to WT cells (Figure 7.7), indicating that functions related to cell-cycle control and neuronal differentiation started to be differentially controlled right from the beginning of neurogenesis.

ΔmiR-124 groups differed in ion incorporation and diverse metabolic processes at 1 and 2 dpi, and *signal transduction resulting in induction of apoptosis*, *positive regulation of long-term neuronal synaptic plasticity*, *regulation of dendritic spine morphogenesis*, and *striatal medium spiny neuron differentiation* at 3 dpi (Figure 7.7). In contrast to Wnt signaling in WT, *phosphatidylinositol 3-kinase signaling* was detected in ΔmiR-124 groups. In particular, the GO term *striatal medium spiny neuron differentiation* was in line with the BrainSpan analysis (more details in the paper). Our analysis further suggested that the differences in cell fate regulation were mediated by the altered miR$-124-$disregulated TF networks in ΔmiR-124 cells at 3 dpi. The regulatory network analysis of the TFs, which are direct miR-124 targets, revealed clear differences in network architecture between WT and ΔmiR-124 samples: these increased until the cells become post-mitotic. These complex relationships, i.e. indirect miR-124 functions, would have been impossible to detect by miR-124 target analysis alone. Furthermore, biological functions underlying these regulatory network differences are in line with the phenotypic differences observed in ΔmiR-124.

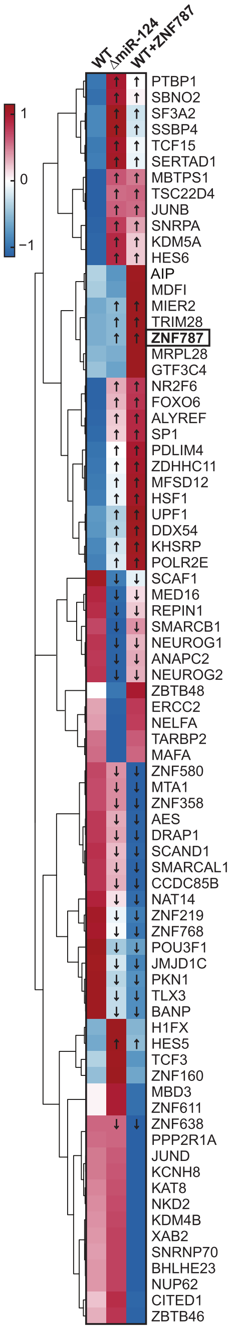


FIGURE 7.8: ZNF787-associated genes reappearing in target-wTO networks: Heatmap of ZNF787-associated genes reappearing in target-wTO networks for WT, ΔmiR−124, and WT-ZNF787-OE (RNA-seq, z scores from rlog-transformed counts, m = 3 biological replicates). Arrows indicate similar expression trends for ΔmiR−124 versus WT and WT-ZNF787 overexpression versus WT.

8

Understanding the human cognition using TF-networks

"My whole brain was out of tune
I don't know how to tune a brain, do you?
Went in to a brain shop
They said they'd have to rebuild the whole head
I said well, do what you gotta do
When i got my brain back, it didn't work right
Didn't have as many good ideas
Haven't really have a good idea since i got it fixed"

– My brain, *Morphine*

Cognition is a group of mental processes that include attention, memory, producing and understanding of language, learning, reasoning, problem-solving, and decision making. Many of these mental processes take place in the Prefrontal Cortex (PFC).

The closest living relatives of humans, the chimpanzees and bonobos, display a rich repertoire of higher cognitive functions, such as usage and production of tools, they teach others how to make tools, they communicate to hunt cooperatively, and they can deceive others. They further have some understanding of numbers and syntax, have a sophisticated memory, and seen to occasionally plan for the future[261–263]. Humans are distinct from other apes by having a more complex communication system and an extraordinary ability to learn from others. They acquire knowledge faster and maintain newly acquired knowledge over generations[264]. Genetic changes must have facilitated the evolution of human-specific cognitive skills. Several genes that are important for cognitive functions have been identified through research on cognitive disorders, such as Autism Spectrum Disorder (ASD), Schizophrenia (SCZ) or Alzheimer's Disease (AD). Many of these disorders share a common genetic basis, that is characterised by overlapping sets of genes implicated in these disorders, and overlapping phenotype and symptoms. Knowing that, this Chapter has two main objectives: i) define a Consensus Network (CN) wTO for the PFC of humans in some cognitive diseases, such as ASD, AD, Bipolar Disorder (BD), Major Depression Disorder (MDD), Parkinson Disease (PD) and SCZ and compare them with healthy individuals, and ii) compare the CN of healthy humans with healthy apes.

The results of this Chapter are being prepared for a publication.

- **D. M. Gysi** and K. Nowick. "Evolution of gene-co-expression networks implicated in cognitive functions in primates. *In preparation*. 2019.

8.1 Motivation

The adult human brain is not only composed by many types of neurons, as we saw in Chapter 7, but it is also arranged in many regions, each responsible for different functions, functional connectivity patterns and distinct distribution of cell types[59]. Even though there is a range of differences in connectivity between regions across individuals, a considerable part of transcripts is conserved across the human population[59]. However, changes on the levels of gene expression might lead to abnormal function[265] such as psychiatric illness, for example Autism Spectrum Disorder (ASD)[266–269],Bipolar Disorder (BD)[270–273], Major Depression Disorder (MDD)[46;47;272;274], Schizophrenia (SCZ)[50;188] or neurodegenerative disorders such as Alzheimer's Disease (AD)[180;275–277] and Parkinson Disease (PD)[182;278–280].

ASD is a developmental disorder and is characterised by deficits in social communications, restricted and repetitive behaviours[281;282]. It has a high familiar recurrence[283]. It is highly polygenic and many genes that increase the risk for ASD has been identified[284;285].

BD is a severe psychiatric disorder and is characterised by recurrent alternating episodes of mania and depression[270], neuropsychological shortfalls, immunological and physiological changes[286;287]. It is also associated with high rates of premature mortality from suicide and medical comorbidities[288]. BD is a complex disorder, with both environment and genetic components[270;289]. However, many genes have been implicated to this disorder it is still unknown much about it[270].

MDD is another important psychiatric illness[290]. This is a disease that affects how the person relates, feel and interacts with the world, to themselves and with other individuals. In general, they lose interest in their daily activities, on their lives, feel worthless and lose the capacity to feel pleasure[290;291]. Moreover, patients also show changes in biological processes that regulate sleep, appetite, sexual activity, autonomic function, and neuroendocrine activity[282;290]. As ASD and BD, MDD is also highly polygenic and a complex disorder, where the environment plays a big role[292].

SCZ is a severe mental disorder that affects 1% of the population, and similar to BD has also a high rate of premature mortality[188]. This disorder has different symptoms types, is multidimensional and heterogeneous. And its symptoms are divided as *positive* and *negative*. The positive symptoms include ideas of reference, delusions and magical thinking, while negative includes social withdraw[293;294]. It also is a polygenic complex phenotype.

The heritability for those mental disorders has been estimated as being around 80%[295–297], which is quite similar to type I diabetes[298], but is greater than breast cancer[299]. AD has an estimated heritability of 74%[300] and PD, on the other side, 41%[301;302].

ASD, BD, MDD and SCZ share signs, symptoms, clinical features[303] and some genes are similarly differentially expressed in those disorders[272;304]. AD and PD are also complex and polygenic[200;275;300–302;305;306].

In general, differently, from the previous disorders that start to be symptomatic in early age, AD and PD tend to appear later in life. AD is a severe disorder that is characterised by episodic memory loss accompanied by progressive impairment of other cognitive domains[305]. AD affects different areas of the brain differently and not simultaneously[275]. PD is a severe progressive neurodegenerative disorder and is a common cause of disability. There are no therapies that can decrease disease progression. This disorder is characterised by the presence of Lewy bodies and heavy damage of dopaminergic neurons in the substantia nigra[302].

8.1.1 A network of symptoms

It is common in psychometrics to explain a set of correlated measures, such as the joint occurrence of psychological symptoms, as caused by an unobservable variable, such as general intelligence or depression[307;308]. However, the causation structure assumed by these models implies that the unobservable variable is the *unique cause* of correlations between observations. This assumption might not be reasonable in clinical and psychiatric setting[309] in which symptoms might cause and modify other symptoms. As an example, lack of sleep and fatigue are associated with depression. However, it is reasonable to assume that sleep problems might cause some fatigue for physiological reasons not necessarily related to the presence of depression. Therefore, network approaches to investigate the interaction of symptoms have been proposed in the literature.

Those networks are assembled considering each symptom as a node and the association strength as the edge's weight[310]. For MDD and Generalised Anxiety Disorder this network resulted in two well-defined clusters, one associated with each disease with some *bridge symptoms* between both diseases, such as sleep problems, restlessness and concentration problems, which give a possible explanation for the comorbidity of both diseases. In addition, the strength of each node provided insights into which symptoms are more dominant to each separate diagnosis[310;311]. Similarly, the symptoms network of ASD and Obsessive-Compulsive disorder[312] was investigated.

Using similar methods, also the entire disease structure of the DSM-IV catalogue[282] has been investigated[311]. This is a comprehensive manual of 439 symptoms, criteria and language for the classification of mental disorders. Two symptoms were connected if they are used as diagnostic criteria for the same disease. This network revealed one large component comprising almost half (47.4%) of the symptoms, a high degree of clustering and high connectivity. These properties give the network a small world nature, predicting the observed pattern of multiple symptoms interacting in multiple disorders and the observed prevalence rate for comorbidity.

8.2 A network of genes involved in mental disorders reveals overlaps

Based on the same rationale as presented in the symptoms network, I wanted to understand the overlap of genes to mental disorders, using the Gene-to-Disorder (g2d) database presented in Chapter 6. For that, I constructed a bipartide network, $\mathbb{B}$, where the first set of nodes is given by the disorders and the second set is given by the genes. A disorder has a link to a gene only if the gene is associated with that particular disorder. The visual representation of this network is presented in Figure 8.1. It is quite interesting how the genetic basis overlap of those disorders can be verified. As in the symptoms networks that BD, MDD and SCZ shares symptoms, they also have a big overlap on the genes. On the other hand, ASD, impaired cognition and intellectual disability are also closer in the network since they have a higher gene overlap. Important to note here is that because ASD is a multidimensional disorder, there is a plurality of completely different symptoms, and probably, with a different genetic basis.

The bipartide network was transformed into a weighted *disease to disease network*, $\mathbb{D}$ based on how the gene amount two diseases share, this can be easily calculated by multiplying the adjacency $\mathbb{B}$ matrix by its transpose, i.e. $\mathbb{D} = \mathbb{B}\,\mathbb{B}^{\intercal}$. However, having the proportion of shared genes instead of the absolute number, give us a better insight of their overlap. The proportion of shared genes of any two disorders, o and p, can be easily calculated as

$$\mathbb{D}^* = [D^*_{o,p}] = \frac{D_{o,p}}{D_{o,o} + D_{p,p} - D_{o,p}}.$$

It gives us the adjacency matrix of the estimated proportion of shared genes in two disorders, represented by $\mathbb{D}^*$. The visual representation of this network is presented in Figure 8.2. This network enables us to describe better how the genetic basis of the mental disorders is related. The same pattern as seen in the bipartide graph can be visualised here. Moreover, a triangle from BD, SCZ and MDD appears. This means that those three disorders have, indeed, a higher set of overlapping genes than other disorders. ASD is the middle of the triangle, also sharing many connections with those three disorders.

However, because not all classes of genes are able to regulate the expression process, my interest lies on the ones that might be responsible for the rewiring in the networks, the Transcription Factors (TFs). With the result from this network, I hypothesise that the TF-TF network of the i) neurodegenerative disorders might have few similar patterns of deregulation compared to healthy controls. Because it has the highest overlap with brain development than the other disorders, I also want to compare it to the infant TF-TF network; ii) ASD, BD, MDD and SCZ have similar patterns of gene co-regulation. I expect to find genes that are highly specific for each one of those mental disorders, those genes are also referred to as *signature genes*.

An overlap of genes specific to humans that are also involved in those disorders was suggested[54]. Therefore, another question is to understand the patterns of co-expression in humans and other primates, such as Chimpanzee (CMP) and Rhesus macaque (RH).

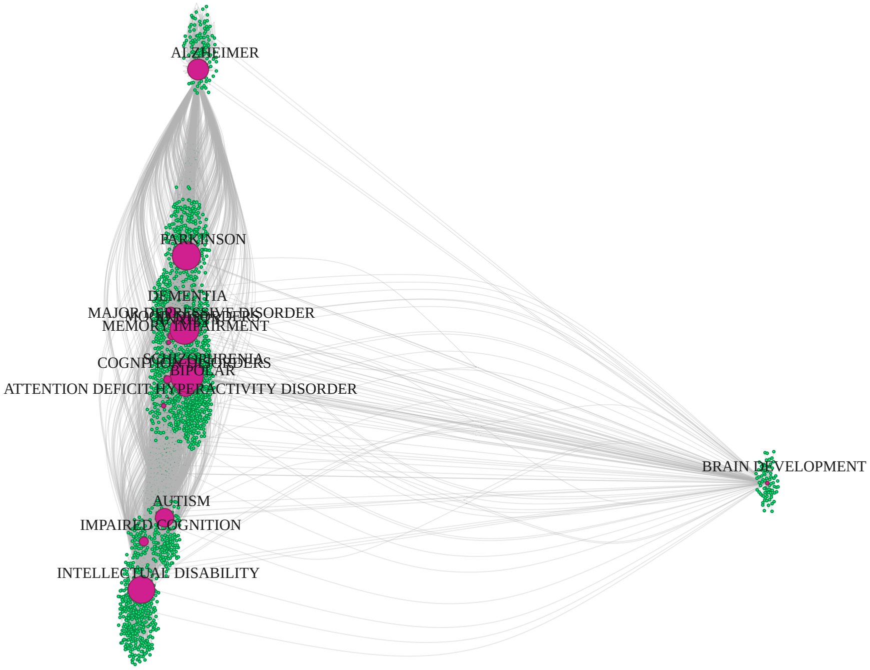


FIGURE 8.1: Gene and mental disorders bipartide network: Genes are represented in green and diseases in purple. The layout used for this visualisation is Distributed Recursive (Graph) Layout (DrL), that allows us to better visualise the clusters of gene-disease interactions. The size of the nodes is proportional to the node degree. It is clear that *brain development* does not share many genes with the other disorders, it is not surprising that *impaired cognition* and *intellectual disability* share genes. ASD seens to have a big overlap with other disorders. However, SCZ and BD have a higher overlap. Similarly, *dementia*, AD and PD have a higher overlap than the other disorders.

8.3 TF-TF co-expression network analysis

Because of this huge overlap in the symptoms and the genes that are involved with mental disorders I wanted to understand if the patterns of gene co-regulation were similar in those disorders. For that, I collected available data from the Gene Expression Omnibus (GEO). Performed quality control and normalised the datasets. Later, using the methodologies described previously in Chapter 4, Chapter 5 and Chapter 6, I constructed the independent weighted Topological Overlap (wTO) networks for each disorder, combined into one Consensus Network (CN), compared the networks using Co-expression Differential Network Analysis (CoDiNA) and finally performed a enrichment using Gene Ontology (GO) and `RichR`.

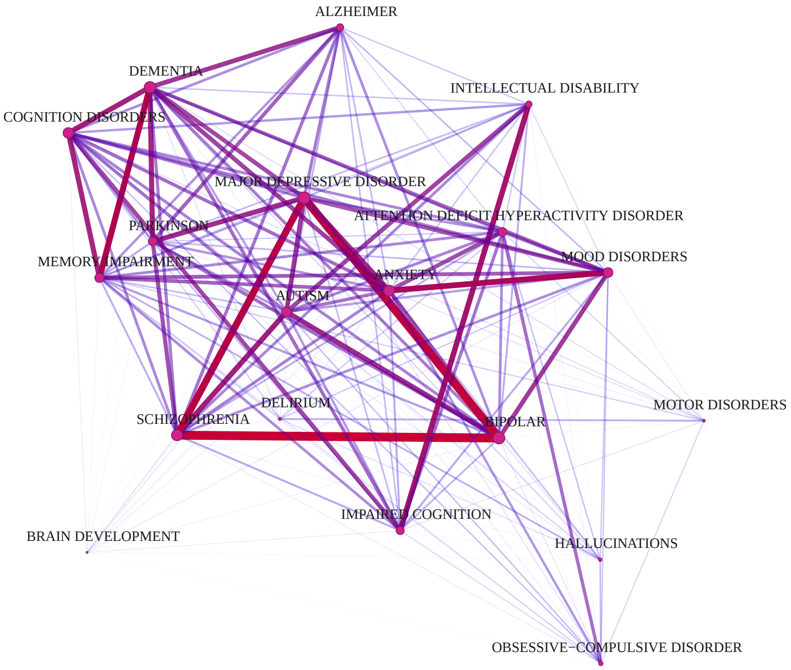


FIGURE 8.2: Weighted Disease-to-Disease network shows modules of diseases: Link thickness are proportional to the link weight, the stronger the red, the higher the proportion of shared genes the two disorders share, the lighter the blue, the fewer genes the two disorders share. As expected, BD, MDD and SCZ forms a triangle where some genes overlap; in the centre of this triangle, we find ASD, that also overlaps highly with those disorders, as discussed in the literature. On the other hand, the neurodegenerative disorders share a lot with dementia but not as much with each other.

8.3.1 Database selection

Data was collected from GEO. The data used originated from studies that contained at least 8 samples from the same brain tissue and disorder. The platform used for measuring the gene expression should not have been designed for a subset of genes. In total, I collected expression data from 4399 samples, divided into 39 studies, using 15 platforms, from 5 brain regions (Cerebellum (CER), Ganglia (GAN), Hippocampus (HIP), Prefrontal Cortex (PFC), Temporal Cortex (TC)) and 28, 901 different transcripts. A description of the datasets can be found on Appendix C. A table of the number of datasets per tissue can be found on Table 8.1. Not all diseases have data available for all tissues, and few do not have more than one dataset. However, because my interest lies in cognition, it is natural to only use the data that comes from the brain area responsible for that, that is the PFC. The brain areas that are not PFC are used only to help understand how the individuals' cluster in Subsection 8.3.2 but not carried on for the network analysis.

Table 8.1: Number of expression datasets for each tissue and phenotype used to construct the networks. The diseases collected for humans are: Autism Spectrum Disorder (ASD), Alzheimer's Disease (AD), Bipolar Disorder (BD), Major Depression Disorder (MDD), Schizophrenia (SCZ) and Parkinson Disease (PD). Moreover, data was collected for Chimpanzee (CMP) and Rhesus macaque (RH). Healthy human adults (CTR5) and healthy human infants are (CTR0) are also presented.

Phenotype	Cerebellum	Ganglia	Hippocampus	Prefrontal Cortex	Temporal Cortex
AD	0	0	1	3	1
ASD	1	0	0	2	1
BD	1	0	1	5	0
MDD	1	3	1	5	0
PD	0	3	0	2	0
SCZ	2	0	1	3	1
RH	1	2	4	2	0
CMP	1	0	0	2	0
CTR0	3	0	2	5	2
CTR5	12	12	8	28	7

Normalisation and quality control

After the selection of the datasets, all datasets went through a quality control step. Because each study is independent of each other this step has to be done separately. The controls samples were normalised among themselves, and the diseases were normalised against the controls. Some gene expressions were measured using microarrays, and those were analysed using the R environment[18] and Bioconductor[313] packages. For the microarray sets that were from *Illumina*, I used the package lumi[21], for the ones from *Affymetrix*, the package affy[19].

The probe expression levels (Robust Multi-array Average (RMA) expression values) and MicroArray Suite 5 (MAS5) detection $p-$values were computed and only probesets significantly detected in at least one sample ($p-$values < 0.05) were considered. Probes that were not specific for only one gene were deleted. Average expression was computed for those genes mapped by more than one probe. All datasets with low quality were removed. Only datasets that survived this step are presented here.

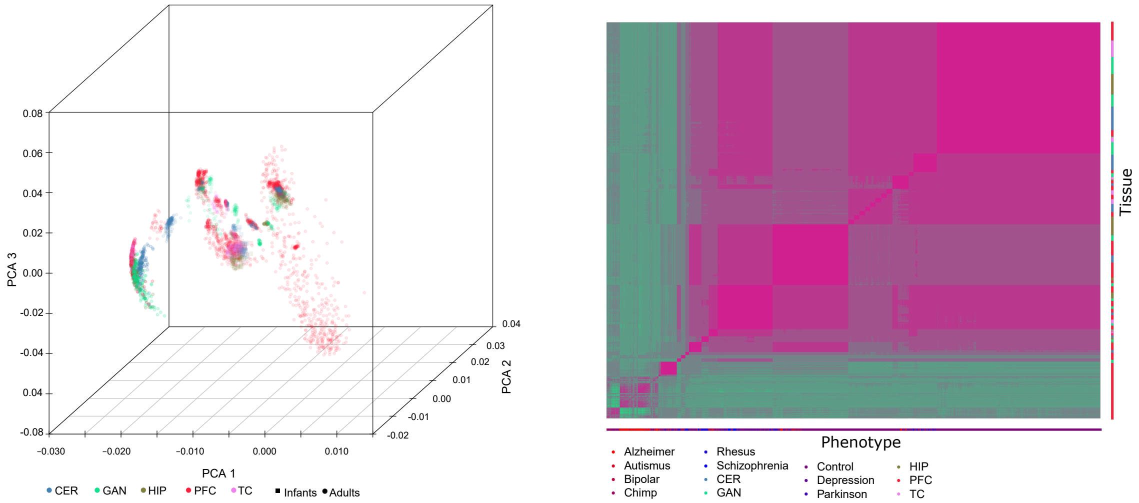
For the datasets that were measured using RNA sequencing (RNA-seq), I made the quality control using FASTQC[314], mapped the reads to the human genome assembly hg38/GRCh38[315] using Segemehl[34]. Uniquely mapped reads were extracted for further analysis and counted using rnacounter[38]. Genes with less than 5 Reads Per Kilobase Million (RPKM) in total were removed.

Transcription Factor list

Similarly to the study presented in Chapter 7, only TFs from the target list (assembled from the Gene Regulatory Factor (GRF) Catalogue[54] were considered for the network analysis.

8.3.2 Multivariate analysis

In order to verify how did the individual cluster, I used a Principal Component Analysis (PCA). The PCA was constructed using the function `princomp()` in R, using only TFs instead of the whole set of expressed genes. Similarly, a heatmap was also performed, using the cor-



(A) PCA.

(B) Heatmap.

FIGURE 8.3: PCA and correlation heatmap of the 4399 individuals: 8.3a is a 3D scatter-plot that shows that the individuals clusters into the study, platform, brain region, disease and age. Similarly, 8.3b represents the individuals correlations. The same colour code is used as before, negative correlations are represented in green and positive correlations are in purple.

relation of the individuals. In both analyses, five levels of disturbance can be detected on the data: Study and Platform are the stronger, however, the brain region, the disease and age also showed to cause differences on the levels of expression. Thus, it was necessary to construct the wTO networks from 129 different datasets, each containing data from the same: study, platform, disease and age group.

8.3.3 Constructing the networks

Networks were constructed independently for the PFC for each combination of study, platform, disease and age. I constructed the networks using the wTO methodology presented in Chapter 4. For each network, the parameters for the `wTO` used were Bootstrap, Pearson correlation and a 1000 replicates, using TFs for the Topological Overlap (TO).

8.3.4 Combining the networks

Networks were combined according to the combination of each disorder and age. For that, I used the CN methodology, presented in Chapter 4. Because not all transcripts were measured in all networks, I considered the interaction of all TFs that had not been measured to have a wTO value of 0 and the $p_{adj}-$value to be 1. This approach penalises all links that were not measured in all networks, however, if this link existed in the majority of the networks with low $p_{adj}-$value, it is still present in the CN.

For the final CN, links were considered to be significant only if its $p_{adj}-$value < 0.05. All links that did not fulfil this criterion had its wTO value set to 0. This allows all genes to be present in the comparison network.

8.3.5 Comparing the networks

In the previous sections, I discussed that the gene overlap of the mental disorders treated here is quite big. Therefore, it is of major importance to detect genes that can be identified as *signatures*. Those genes can be later used for improving and developing new treatments for those disorders. It is also important to understand if exists an overlap of TFs that are active during the brain development and the neurodegenerative disorders, however, are inactive during adulthood. Psychological disorders have an immense overlap of deregulation in the expression of the gene, however, it is not clear if there are patterns of deregulation in its co-expression. Humans and other primates share many similarities and differences. It is also of major interest to understand the specificities of each species.

In order to understand if there are patterns of co-expression on the neurodevelopmental disorders, I compared the CN derived from the PFC of patients with AD and PD with the adult (CTR5) and infant (CTR0) healthy controls. Similarly, to understand co-expression patterns in psychiatric disorders, I compared the co-expression network of ASD, BD, MDD and SCZ against healthy adult controls. To identify patterns on the co-expression on the PFC of primates, I compared the CN of healthy adult controls, chimpanzee (*Pan troglodytes*) and rhesus macaque (*Macaca mulatta*).

I filtered the CoDiNA network to contain only well defined, clustered and strong similarities or differences by filtering for the scores ratio, as proposed in Chapter 5 ($\Delta^{**}/\Delta_{\widetilde{\rho}}> 1$). Genes were categorised into a Φ category and $\widetilde{\Phi}$ subcategory using the `ClusterNodes()` function.

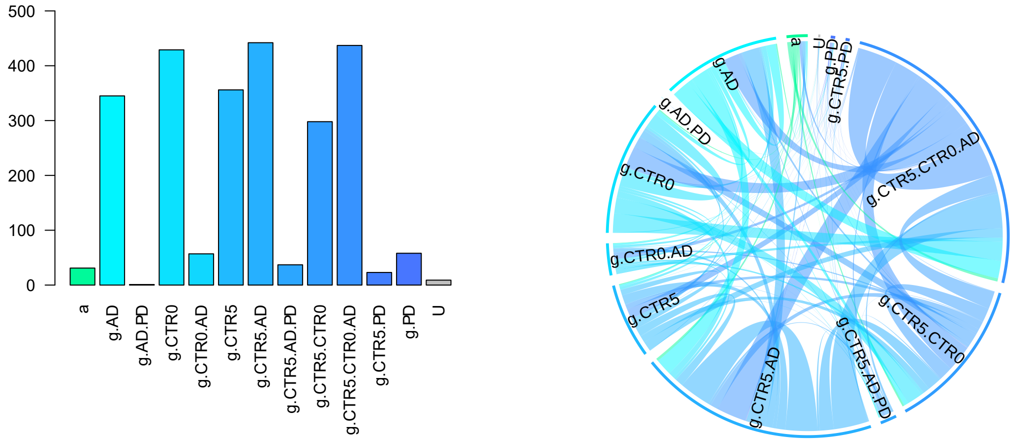
The Φ categories in `R` are shown as *a* for α, *b* for β and *g* for γ. Links classified as *a* are common to all networks under comparison, *b* had a signal change, by default, the first network is considered to be the reference level, in all comparisons shown here, the adult humans are taken as reference and are represented as `CTR5`. *g* links (or nodes) are specific to at least one category, the abbreviation *g.AD* means that that particular link (or node) exists only in AD, but not in other phenotypes, similarly *g.CTR5.CTR0* means that it exists only in the adult and young controls.

8.3.6 Disease enrichment

Each $\widetilde{\Phi}$ was tested for the disorder and GO enrichment, which was performed using `topGO`[170]. The geneset of the background was only TFs co-expressed in the networks. Filtering for this background allows that the enrichment is not biased towards this class function. The Gene-Disease Associations (GDA) enrichment was performed using `RichR` presented in Chapter 6 and filtered for both Fisher's exact test or proportion test $p_{adj}-$value< 0.05. I used only genes associated with mental disorders and the TF as the background of expressed genes.

8.4 The co-expression network of the prefrontal cortex in neurodegenerative disorders

To find specific genes to neurodegenerative disorders and test if there are genes involved in brain development and those disorders I compared the AD and PD networks with the controls from adult and infant individuals. In total, the differential network of AD, PD and the controls (adults, CTR5, and infants, CTR0), was able to classify $1,153,949$ interactions from $2,514$ TF. The link distribution according to the $\widetilde{\Phi}$ can be seen in Table 8.2 and the distribution of TFs classified in each $\widetilde{\Phi}$ is shown in Figure 8.4a. Interesting to note is that AD has more specific genes than PD, it also has genes that are shared with the infant's brain. Suggesting that there are TFs that behaves similarly to the ones found in brains that are still under development. Another important thing to note here is that AD has a higher part of the genes being shared with both controls when compared to PD. This is quite interesting and shows that PD has less TFs being co-expressed. Another important result is that there are many genes specific only to the controls, but absent in the disorders. This might indicate a lack of gene-regulation in the neurodevelopmental disorder.



(A) TFs distribution for neurodevelopmental disorders

(B) Chord graph of TF-TF interaction.

FIGURE 8.4: CoDiNA TFs specificity distribution and interaction for neurodevelopmental disorders Figure 8.4a displays the nodes distribution classified in each one of the CoDiNA $\widetilde{\Phi}$ subcategories. *g.* means specific TFs, *a* are genes with common pattern of co-expression. *U* are TFs that could not be classified. In Figure 8.4b we can see a chord graph of the $\widetilde{\Phi}$ TFs for neurodegenerative disorders. Colour code is the same as in the barplot. If a link starts in one category and moves to another, it *crosses* the circle. The Figure 8.4b shows that most of the interactions happens inside the same category instead of genes interacting with distinct $\widetilde{\Phi}$. This category dependent interactions might be interpreted as changes on big blocks of regulation.

Of note is how those genes interact in the CoDiNA network. In Figure 8.4b we can see that most of the TF-TF interactions occur in the same $\widetilde{\Phi}$ category, This is quite important to understand that the deregulation of a TF in these disorders disrupts a whole set of genes.

Table 8.2: Links distribution in the neurodegenerative co-expression network. α represents links that are common to all networks, γ are links that specific to at least one condition. The networks compared here are: Alzheimer's Disease (AD) and Parkinson Disease (PD) against healthy human adults (CTR5) and healthy human infants (CTR0).

$\widetilde{\Phi}$	# links
α	51086
β AD	1263
β AD.PD	43
β CTR0	3648
β CTR0.AD	105
β CTR0.AD.PD	6
β CTR0.PD	28
β PD	220
γ AD	146572
γ AD.PD	9356
γ CTR0	156835
γ CTR0.AD	68524
γ CTR0.AD.PD	3535
γ CTR0.PD	5845
γ CTR5	119330
γ CTR5.AD	166140
γ CTR5.AD.PD	64795
γ CTR5.CTR0	144458
γ CTR5.CTR0.AD	138364
γ CTR5.CTR0.PD	27373
γ CTR5.PD	30110
γ PD	16313

Regarding the enrichment analysis for each subcategory of the neurodegenerative disorders, I found 29 (out of 114) genes involved in the *brain development* for genes classified as specific to AD, showing enrichment for this ($p_{adj}-$value Fisher's exact test and proportion test < 0.05). Impaired cognition also presented itself as significant, with 5 out of 67 genes being present in AD, PD and adults. Meanwhile, the GO enrichment showed important terms for PD, for example, beta-catenin destruction complex, that is known to be affected in this disorder[316;317]. A full table of the enrichment can be found in Appendix D.

The genes found to be hubs in this network are ZMYM2, that is a gene classified as common to all phenotypes, it is associated to *regulation of transcription by RNA polymerase II*[318]. The TFs SLF1, SRSF2, ZFP90, SRSF10, ZNF772 and ZNF845 are specific to the controls (CTR5 and CTR0) and are involved with *RNA splicing*[318].

The set of genes HSF2, PIAS2, MYEF2, HCFC2, UBE3A, RBMX, ZMYM6, ZNF131, ZNF25, ZNF420 and ZNF84 are specific to both controls and AD, however, it is not to PD. Those hubs are enriched for *regulation of transcription, DNA-template, intracellular steroid hormone receptor signalling pathway* and *regulation of nucleic acid-template transcription.*

8.5 The co-expression network of the prefrontal cortex in psychiatric disorders

To find specific genes to psychiatric disorders that can be later used as signatures of those mental disorders, I compared the ASD, BD, MDD and SCZ networks with the healthy adults as a control. Although exists links that are classified as α to all conditions under comparison, no gene has enough connections to be considered significant to this category. In total, $2,379$ TFs were classified into the $\widetilde{\Phi}$ categories. There were $797,531$ links on the final CoDiNA network, those links were distributed according to Table 8.3.

Table 8.3: Links distribution in the psychiatric co-expression network. γ represents links that specific to at least one condition. The networks compared here are: Autism Spectrum Disorder (ASD), Bipolar Disorder (BD), Major Depression Disorder (MDD), Schizophrenia (SCZ) against healthy human adults (CTR).

$\widetilde{\Phi}$	# links
α	5001
β ASD	19
β ASD.SCZ	1
β BD	3
β BD.MDD.SCZ	2
β SCZ	3
γ ASD	83233
γ ASD.BD	6174
γ ASD.BD.MDD	374
γ ASD.BD.MDD.SCZ	64
γ ASD.BD.SCZ	316
γ ASD.MDD	1087
γ ASD.MDD.SCZ	78
γ ASD.SCZ	1310
γ BD	12881
γ BD.MDD	1213
γ BD.MDD.SCZ	198
γ BD.SCZ	1266
γ CTR	324849
γ CTR.ASD	218381
γ CTR.ASD.BD	23549
γ CTR.ASD.BD.MDD	7686
γ CTR.ASD.BD.SCZ	4153
γ CTR.ASD.MDD	11355
γ CTR.ASD.MDD.SCZ	2299
γ CTR.ASD.SCZ	8093
γ CTR.BD	26312
γ CTR.BD.MDD	9489
γ CTR.BD.MDD.SCZ	5174
γ CTR.BD.SCZ	4929
γ CTR.MDD	15073
γ CTR.MDD.SCZ	2988
γ CTR.SCZ	11154
γ MDD	2985
γ MDD.SCZ	374
γ SCZ	5465

As discussed previously, the overlap of genes found to be associated with psychiatric disorders is quite big. With this, we aim to find genes that are specific to each category individually, those genes could be useful in finding signatures for identifying psychiatric disorder and could potentially be helpful for identifying new specific treatments. However, in Figure 8.5a the majority of genes are specific to controls, it means that they are **not** regulated in the same way in the mental disorders, showing that, indeed, a big overlap on the patterns of co-expression can be found. ASD has also a big share on γ genes and some shared only with the controls. This is, again by exclusion, deregulated TFs in BD, MDD and SCZ.

Differently than the neurodevelopmental disorders, the mental ones have a big overlap also on the co-expression patterns and how genes interact (Figure 8.5b). The TFs classified

TABLE 8.4: RichR enrichment for mental disorders. γ represents links that specific to at least one condition. The categories presented here are: Autism Spectrum Disorder (ASD), Bipolar Disorder (BD), Major Depression Disorder (MDD), Schizophrenia (SCZ). Healthy human adults (CTR) were used as controls.

$\widetilde{\Phi}$	Disease	Genes Expected	Genes Observed	weight
γ SCZ	ATTENTION DEFICIT HYPERACTIVITY DISORDER	29	2	0.0132
γ SCZ	IMPAIRED COGNITION	61	3	0.0068
γ CTR.SCZ	ATTENTION DEFICIT HYPERACTIVITY DISORDER	29	1	0.0015
γ CTR.MDD	INTELLECTUAL DISABILITY	192	3	0.0000
γ CTR.BD	COGNITION DISORDERS	45	3	0.0000
γ CTR.BD	DEMENTIA	43	2	0.0247
γ CTR.ASD.BD	IMPAIRED COGNITION	61	1	0.0020
γ CTR	INTELLECTUAL DISABILITY	192	94	0.0153
γ BD	MAJOR DEPRESSIVE DISORDER	116	7	0.0049

as specific to the healthy humans mainly interact among themselves, it can indicate that there is a rewiring in the metabolic pathways and biological processes that underlie that networks. Part of the interactions occurs from one $\widetilde{\Phi}$ to another. It might indicate that there is not complete rewiring in the network, but only sub-parts of the pathways become active or inactive.

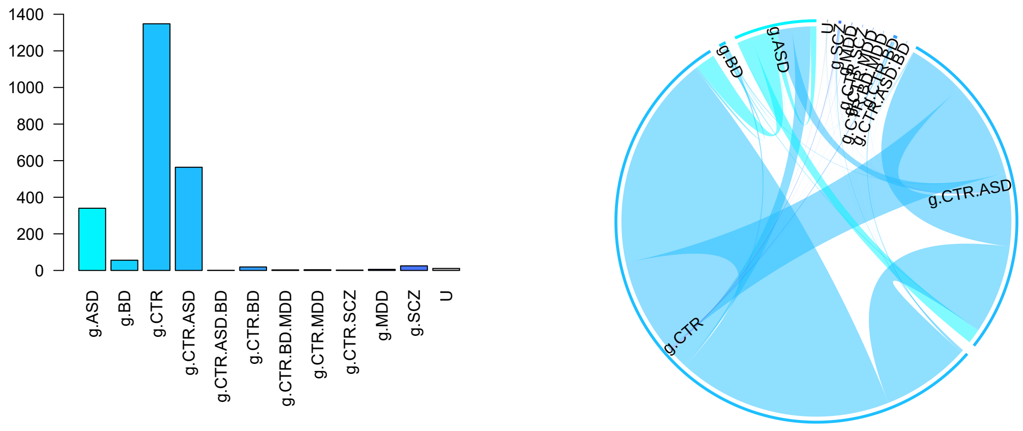
The CoDiNA network for the mental disorders showed that not many hubs appear in all $\widetilde{\Phi}$. The majority of genes with high connections are related to the controls and does not appear in the diseased brains. However, ASD shared specific genes with control, and it can be used as a signature to define the other two disorders. The genes that specific to controls and ASD are KMT5B, EXOC2, ORC2, GTF3C3, ZRANB2, ELP2, ZNF776 and SMARCAD1 and are involved with *5S class rRNA transcription from RNA polymerase III type 1 promoter, tRNA transcription* and *histone H4 deacetylation*. The controls specific hubs are ZNF597, ZFP90 and ZNF420. Those genes are enriched for *transcriptional repressor activity*.

The RichR enrichment, displayed in Table 8.4, shows that there is an enrichment for *cognition disorders* and *dementia* in controls and BD, meaning that those genes are absent in the other networks. Interestingly, BD also showed enrichment for *Major Depression Disorder*, this result was quite expected, mainly due to the huge overlap of symptoms and genes of both disorders. Of note, controls are enriched for *intellectual disability*, it indicates that those genes are being co-expressed in healthy controls, but lack regulation in the psychological disorders.

The GO enrichment for each category showed that there is an enrichment for cellular response to lithium ion in SCZ that is in line with literature[319], more terms can be found in Appendix E.

8.6 The evolution of the co-expression network of the prefrontal cortex

To find specific genes to humans, chimpanzees and rhesus macaque, I compared the networks of healthy human adults with the ones from healthy chimpanzees and healthy rhesus macaque.



(A) TFs specificity distribution for psychiatric disorders.

(B) Chord graph of TF-TF interaction.

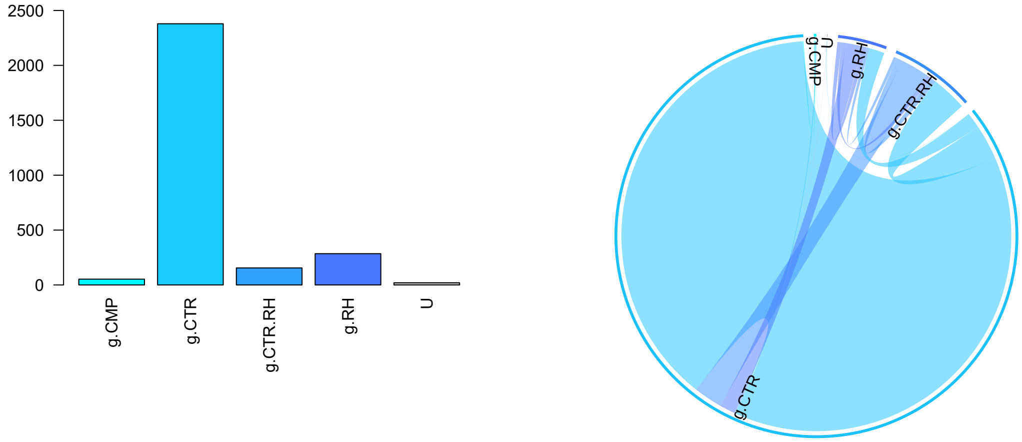
Figure 8.5: CoDiNA TFs specificity distribution and interaction for psychiatric disorders: Chord graph of the $\widetilde{\Phi}$ genes for psychiatric disorders: Figure 8.5a displays the nodes distribution classified in each one of the CoDiNA $\widetilde{\Phi}$ subcategories. *g.* means specific TFs, *U* are TFs that could not be classified. Controls have the biggest number of TFs that have an specific pattern of co-expression, however, ASD also shares many genes with the Controls. In Figure 8.5b we can see a chord graph of the $\widetilde{\Phi}$ TFs for psychological disorders. Colour code is the same as in the barplot. If a link starts in one category and moves to another, it *crosses* the circle. The Figure shows that most of the interactions of controls occurs inside the same category instead of with TFs from distinct $\widetilde{\Phi}$. This category dependency of interactions might indicate that there are big blocks TFs that are specific to each category.

The CoDiNA network for the primates had, in total, $906,735$ links, from $2,892$ TFs. The distribution of the links classification can be seen in Table 8.5. For this network, healthy adult humans are considered to be controls.

Table 8.5: Links distribution in the primates co-expression network. α represents links that are common to all networks, γ represents links that specific to at least one network. The networks tested are: Chimpanzee (CMP) and Rhesus macaque (RH). Healthy human adults (CTR) were used as controls.

$\widetilde{\Phi}$	# links
α	3027
β CMP	361
β CMP.RH	37
β RH	49
β CMP	7179
γ CMP.RH	916
γ CTR	664767
γ CTR.CMP	9143
γ CTR.RH	157559
γ RH	63697

When comparing the co-expression network of TFs of humans and other primates, we can see that humans have more TFs with human-specific links, thus have been characterised



(A) CoDiNA TFs specificity distribution for evolution in PFC.

(B) Chord graph of TF-TF interaction.

FIGURE 8.6: CoDiNA TFs specificity distribution and TF-TF interaction for the evolution in the PFC: Figure 8.6a displays the nodes distribution classified in each one of the CoDiNA $\widetilde{\Phi}$ subcategories. *g.* means specific TFs, *U* are TFs that could not be classified. Humans have the biggest number of TFs that have a specific pattern of co-expression, however, RH also has many specific TFs. The Figure 8.6b displays a chord graph of the $\widetilde{\Phi}$ TFs for evolution. Colour code is the same as in the barplot. If a link starts in one category and moves to another, it *crosses* the circle. This Figure shows that most of the interactions of controls occur inside the same category instead of genes interacting with distinct $\widetilde{\Phi}$. This category dependency of human-specific TFs shows that those TFs are interacting mostly with other human-specific TFs rather than primate-specific TFs.

in by CoDiNA as human-specific, when compared to CMP and RH (8.6a). This is quite surprising that humans have much more specific TFs when compared to other primates. Many TFs have a co-expression pattern specific to RH, is expected that humans and chimpanzees have a closer co-expression pattern of the PFC, due to their evolutionary distance. However, even that not many, humans and RH share some gene patterns that are not present in CMP.

The RichR enrichment for the primates did not find any disorder enriched in either TFs categorised as shared or specific. Which makes total sense, because the networks were constructed only using healthy individuals. However, GO terms were found to be enriched, for example, RH showed enrichment for *walking behaviour*. The hubs found in this network are all classified by CoDiNA as having a majority of links in the human-specific network, the hubs are: THAP12, ZNF280D, ZNF131, ZFP90, ZNF420, ZNF772 and ZNF534 and are enriched for *transcriptional repressor activity*.

8.7 Psychological disorders share patterns of co-expression with other primates

The combination of wTO, CN and CoDiNA revealed TFs and signatures for each of the disorders. At the beginning of this chapter, I discussed the gene and symptoms overlap of those disorders. The hypothesis was that the neurodevelopmental disorders would share its genetics basis with brains under development. For that, I compared the PFC co-expression network from healthy adult controls and infant brains with the one derived from patients diagnosed with AD and PD. In those networks, I identified TFs that are involved with each disorder, but mostly, two categories of TFs draw our attention: i) TFs that are present only in controls (independent of the age) might be TFs that plays an important role in the PFC functions and ii) TFs that are inactive in the adult PFC but active in the infants and the PFC with the disorders. The Venn diagram showing the amount of TFs in each one of these categories is displayed in Figure 8.7. Other studies already pointed out some evidence that neurodegenerative diseases and development of the brain might be regulated by some non-coding RNA (ncRNA)[320].

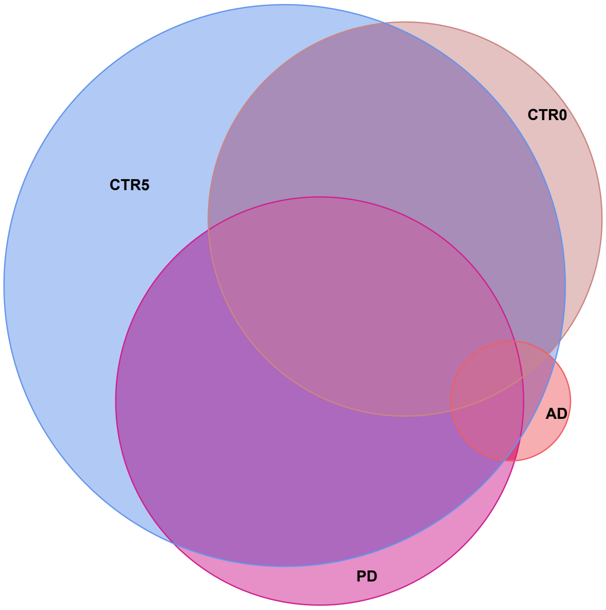


Figure 8.7: Venn diagram of the TFs overlap from the neurodegenerative disorders and healthy PFC shows that there are two important categories of TFs. In the first, some are present only in controls (independent of the age), can be related to TFs that play an important role in the PFC functions. The second is represented by TFs that are inactive in the PFC of adults (CTR5), however, are active in the infants (CTR0) and the PFC with the disorders, that might indicate that there are TFs that regulates other genes in a brain in development and in the neurodegenerative disorders.

Similarly, we wanted to understand the genetic overlap of the TFs in psychological disorders. Another study[321] showed that genes associated with ASD might also be involved with the primate evolution. In Figure 8.8 it is displayed the Venn diagram of TFs identified in each category and its overlap. It is possible to identify that ASD and RH have TFs identified as specific to both categories. It is also interesting to visualise that MDD and SCZ have the majority of their TFs also associated with BD. Of note is that TFs that specific to CMP is also associated with ASD, BD, MDD and SCZ.

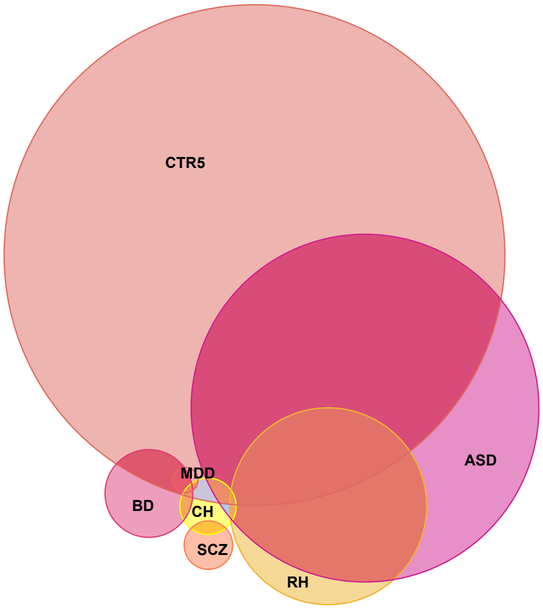


FIGURE 8.8: Venn diagram of the TFs overlap from the psychological disorders and healthy PFC: It is displayed the overlap of the TFs in the psychological disorders. MDD and SCZ have their majority of genes overlapping with BD. CMP and RH also have a part of genes specific to these three mental disorders.

9

Conclusion and future perspectives

In Chapter 4 I presented weighted Topological Overlap (wTO) method for constructing networks and how I improved this pre-existing method by attributing a probability to each link and allowing the method to be used to time series data when the interest lies in a single network for the whole time-frame. In the same chapter, I introduced the Consensus Network (CN), a method that can be used to find links that are common to a set of independent networks. The wTO method also showed to be more efficient than state-of-art methods in detecting experimentally validate genes.

In the next Chapter, I introduced a novel methodology, Co-expression Differential Network Analysis (CoDiNA), that allows for multiple comparisons of independent co-expression networks. CoDiNA applications clearly demonstrate the successful detection of genes associated with specific phenotypes. Moreover, CoDiNA can also be used for comparing any number of networks other than co-expression networks, for example, metabolic, protein-interaction, or ecological networks.

At the last methodological Chapter, I introduced RichR, an enrichment tool that enables to find enriched disorders in a set of genes. This is particularly useful in combination with the two previous methods. This methodology differs from the other Gene-Disease Associations (GDA) enrichment tools because the user can define both the background list of genes and also the Gene-to-Disorder (g2d) dataset.

The combination of wTO and CoDiNA shed light in the characterisation of genes involved in the neurogenesis, Chapter 7 and in mental disorders Chapter 8. In neurogenesis, CoDiNA pinpointed genes that were candidates to be involved in the miR-124 pathway. The Zinc Finger, ZNF787 tested showed affecting extensively the process of neuron differentiation.

In the second application showed here, it became more clear that the combination of the methods I developed can be helpful to identify novel genes involved in cognition and mental disorders. That can be used to find potential new therapies. This is an important topic, mainly because, as presented in this dissertation, there is a huge overlap in the network of genes associated with those disorders. It is of major importance to identify signature genes (and regulatory patterns) that can, in the future, enable better and more accurate diagnosis and treatment for those disorders. It is quite interesting to see that humans have a higher amount of TFs where the interactions occur specifically in the human network when compared to other primates. However, it is much more interesting and surprising that the psychological disorders that I presented here have an overlap with some chimpanzee-specific TFs.

The methodologies presented here can be used to help researches to understand other complex systems. For example, I used the combination of those tools to understand the co-infection of HIV and Tuberculosis in children and adults. CoDiNA was able to successfully identify enrichment of known genes associated with HIV infections among the specific nodes, providing support for the ability of CoDiNA to retrieve biological meaningful results. Importantly, I was also able to pinpoint modules of genes related to each one of the co-infections.

In addition, in another study, not presented here, but can be found on the publication list, the focus was to identify similarities and specificities of different cancer types. There I identified several Transcription Factors (TFs) specifically associated with Astrocytoma that were not previously linked to this type of cancer. Some of them were previously described as associated to *neoplasms* and *neoplasm metastasis*[164]. The most strongly differentiated TFs associated with oligodendroglioma that were not assosiated to this disease are associated to *neoplasm invasiveness* and *neoplasic cellular transformation*[164]. The new TFs CoDiNA identified for glioblastoma are described as associated with other *neoplams*[164].

This suggests that the methods that I developed have biologically meaningful results. More importantly, I was able to propose new candidate genes as associated with particular phenotypes or disorders and their interactions. I also expect that these tools will be helpful for many diverse studies for building, combing and comparing network data generated from multiple conditions, such as different diseases, tissues, species or experimental treatments. Furthermore, wTO and CoDiNA are not limited to the analysis of co-expression networks but can be applied to comparing any type of network.

In another application of wTO, the aim was to understand how the fungi in a marine environment interacted. Those interactions can be a fungi-fungi interaction, fungi and biotic factors (such as fish eggs) and abiotic factors (such as temperature and pH) through a whole year. For this aim, data was collected once a week, using metagenomics, the abundance of each Operational Taxonomic Unit (OTU) was measured. I was able to detect interactions of the fungi by using the time series version of the wTO. A publication with this analysis is being prepared.

In summary, my contribution to the scientific communities presented here was the development of two methods for high-quality network analysis and comparison. The first one, wTO, performs substantially better than other well-known methods. The second, CoDiNA, represents a completely novel method that enables the simultaneous comparison of as many networks as desired. It does not classify only links, but also nodes. The combination of those methods was used to better understand the role of miR-124 for the neuronal development, it allowed the identification of ZNF787, a therefore uncharacterised TF that was now shown to be involved in neuronal development. Also, combining the methods I developed gives support for a better understanding of complex brain disorders, including molecular relationships that explain similar symptoms in many disorders, but also TFs and links that could distinguish the disorders, and the potential relationship between the evolution of the human brain and its disorders

However, it is still necessary to construct models and approaches that enable the integration of multiple data types into one model. This could be achieved, for instance, by tighter integration of co-expression data with data on long non-coding RNAs, miRNA, transcription factor binding sites, taken from CHIP-seq data or databases, such as Jasper, Transfac and Encode. CHIP-seq has already been combined with TF expression data to predict the activation status of regulatory elements using a statistical Paired Expression and Chromatin Accessibility (PECA) model[322], those results should also be later experimentally validated[147]. Although some integration exists, it still has to be optimised to be able to include more data types. The integration of multiple data sources could be understood as a multilayer system, where each layer is one omic and the integration among layers can be given by experimental data. However, it is still challenging to integrate multiple omics data into one network[323]. Moreover, it is still necessary to have accessible platforms, software and models that can integrate different layers on those multilayer networks. Such integrated networks would allow for better prediction of diseases. In the sense of mental disorders, the data derived from the multi-omics multilayer network can also be enriched by the use of functional Magnetic resonance imaging (fMRI) for a better linking of symptoms to brain regions.

IV
Appendix

A
Computing Weighted Topological Overlaps (wTO) & Consensus wTO Network

A.1 wTO Manual

The wTO package computes the Weighted Topological Overlap with positive and negative signs (wTO) networks[55] given a data frame containing the mRNA count/ expression/ abundance per sample, and a vector containing the interested nodes of interaction (a subset of the elements of the full data frame).

It also computes the cut-off threshold or $p-$value based on the individuals bootstrap or the values reshuffle per individual[62]. It also allows the construction of a **consensus network**, based on multiple wTO networks. The package includes a visualisation tool for the networks.

The package can be downloaded from `CRAN` using:

```
install.packages('wTO')
```

A.1.1 Input data

The `wTO` package, can be used on any kind of count data, but we highly recommend to use normalised and quality controlled data according to the data type such as RMA, MD5 for microarray, RPKM, TPM or PKM for RNA-seq, sample normalised data for metagenomics.

As an example, the package contains three datasets, two from microarray chips (`Microarray_Expression1` and `Microarray_Expression2`), and one from abundance in metagenomics (`metagenomics_abundance`).

A.1.2 wTO

The wTO method is a method for building networks based on pairwise correlations normalised and corrected by all shared correlations. For this reason, the user can choose a set of factors of interest, called here *Overlaps*, those are the nodes that will be corrected and normalised by all other factors in the dataset. Those factors can be Transcription Factor, long non coding RNAs, a set of species of interest etc.

Genomic data

The `wTO` package contains 2 datasets that were obtained using expression arrays (`Microarray_Expression1` and `Microarray_Expression2`), they were previously normalised and the quality control was done. We will use it to build the wTO network using the different methods implemented in the package.

First we are going to inspect those datasets.

```
require(wTO)
#> Loading required package: wTO
require(magrittr)
#> Loading required package: magrittr
data("ExampleGRF")
data("Microarray_Expression1")
data("Microarray_Expression2")

dim(Microarray_Expression1)
#> [1] 268  18
dim(Microarray_Expression2)
#> [1] 268  18

Microarray_Expression1[1:5,1:5]
#>               ID1      ID2      ID3      ID4      ID5
#> FAM122B  5.325653 5.039814 5.099828 5.053185 5.213816
#> DEFB108B 2.038747 1.965599 1.925807 1.977435 2.079381
#> CCSER2   4.973347 4.865783 4.818910 5.024392 4.697314
#> GPD2     5.453287 5.595471 5.223886 5.130226 5.370672
#> HECW1    4.350837 4.279759 4.218375 4.472152 4.408025
Microarray_Expression2[1:5,1:5]
#>               ID1      ID2      ID3      ID4      ID5
#> FAM122B  5.532142 6.395654 5.159082 5.806040 5.339848
#> DEFB108B 2.456210 1.993044 2.251673 2.440018 2.493610
#> CCSER2   5.164945 4.923511 4.979691 5.080116 5.014569
#> GPD2     5.742455 5.649180 5.430411 6.007418 5.662126
#> HECW1    4.595407 5.243644 4.802716 4.957706 4.738554
```

```
head(ExampleGRF)
#>          x
#> 1    ACAD8
#> 2   ANAPC2
#> 3  ANKRD22
#> 4   ANKRD2
#> 5 ARHGAP35
#> 6    ASH1L
```

Please, note that the individuals are in the columns and the gene expressions are in the rows. Moreover, the `row.names()` are the names of the genes. The list of genes that will be used for measuring the interactions are in ExampleGRF. There should always be more than 2 of them contained in the expression set. If there are no common nodes to be measured, the method will return an error.

```
sum(ExampleGRF$x %in% row.names(Microarray_Expression1))
#> [1] 168
```

Running the wTO We can run the `wTO` package with 3 modes. The first one is running the wTO without resampling. For that we can use the `wTO()`. The second one, `wTO.Complete()`, gives you the whole diagnosis plot, hard-threshold on the $\omega_{i,j}$, the $\omega_{i,j}$, $|\omega_{i,j}|$ values and $p-$values. The last mode, `wTO.fast()`, just returns the $\omega_{i,j}$ values and $p-$value.

Using the `wTO()` function: To use the `wTO()` function, the first step is to compute the correlation among the nodes of interest using `CorrelationOverlap()` and then use it as input for the `wTO()`. In the first function the user is allowed to choose the method for correlation between Pearson ('p') or Spearman ('s'). The second function allows the choice between absolute values ('abs') or signed values ('sign'). Please, keep in mind that the result of the `wTO()` function is a matrix, and it can be easily converted to an edge list using the function `wTO.in.line()`.

```
wTO_p_abs = CorrelationOverlap(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = 'p') %>%
wTO(., sign = 'abs')

wTO_p_abs[1:5,1:5]
#>          ZNF333 ZNF28 ANKRD22   ZFR TRIM33
#> ZNF333    0.352 0.237   0.269 0.242  0.241
#> ZNF28     0.237 0.287   0.209 0.206  0.239
#> ANKRD22   0.269 0.209   0.299 0.199  0.252
#> ZFR       0.242 0.206   0.199 0.328  0.258
#> TRIM33    0.241 0.239   0.252 0.258  0.361
wTO_p_abs %<>% wTO.in.line()
```

```
head(wTO_p_abs)
#>     Node.1  Node.2   wTO
#> 1:  ZNF333   ZNF28 0.237
#> 2:  ZNF333 ANKRD22 0.269
#> 3:   ZNF28 ANKRD22 0.209
#> 4:  ZNF333     ZFR 0.242
#> 5:   ZNF28     ZFR 0.206
#> 6: ANKRD22     ZFR 0.199

wTO_s_abs = CorrelationOverlap(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = ’s’) %>%
wTO(., sign = ’abs’) %>%
wTO.in.line()
head(wTO_s_abs)
#>     Node.1  Node.2   wTO
#> 1:  ZNF333   ZNF28 0.236
#> 2:  ZNF333 ANKRD22 0.258
#> 3:   ZNF28 ANKRD22 0.215
#> 4:  ZNF333     ZFR 0.264
#> 5:   ZNF28     ZFR 0.187
#> 6: ANKRD22     ZFR 0.193

wTO_p_sign = CorrelationOverlap(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = ’p’) %>%
wTO(., sign = ’sign’) %>%
wTO.in.line()
head(wTO_p_sign)
#>     Node.1  Node.2    wTO
#> 1:  ZNF333   ZNF28 -0.099
#> 2:  ZNF333 ANKRD22 -0.185
#> 3:   ZNF28 ANKRD22  0.076
#> 4:  ZNF333     ZFR -0.117
#> 5:   ZNF28     ZFR -0.077
#> 6: ANKRD22     ZFR -0.036

wTO_s_sign = CorrelationOverlap(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = ’s’) %>%
wTO(., sign = ’sign’) %>%
wTO.in.line()
head(wTO_s_sign)
#>     Node.1  Node.2    wTO
#> 1:  ZNF333   ZNF28 -0.064
#> 2:  ZNF333 ANKRD22 -0.143
#> 3:   ZNF28 ANKRD22  0.029
#> 4:  ZNF333     ZFR -0.164
#> 5:   ZNF28     ZFR -0.011
#> 6: ANKRD22     ZFR  0.024
```

Using the `wTO.Complete()` function: The usage of the function `wTO.Complete()` is straight-forward. No plug-in-functions are necessary. The arguments parsed to the `wTO.Complete()` functions are the number *k* of threads that should be used for computing the $\omega_{i,j}$, the amount of replications, *n*, the expression matrix, *Data*, the *Overlapping* nodes, the correlation *method* (**Pearson** or **Spearman**) for the *method_resampling* that should be **Bootstrap**, **BlockBootstrap** or **Reshuffle**, the $p-$value correction method, *pvalmethod* (any from the p.adjust.methods), if the correlation should be saved, the δ is the expected difference, *expected.diff*, between the resampled values and the $\omega_{i,j}$ and also if the diagnosis *plot* should be plotted.

```
wTO_s_sign_complete = wTO.Complete(k = 5, n = 250,
Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = 'p',
method_resampling = 'Bootstrap', pvalmethod = 'BH',
savecor = TRUE, expected.diff = 0.2, plot = TRUE)
#> There are 168 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don't turn off the computer.
#> Simulations are done.
#> Computing p-values
#> Computing cutoffs
#> Done!
```

The diagnosis plot (Figure A.1) shows the quality of the resampling (first two plots). The closer the purple line to the black line, the better. The $\omega_{i,j}$ vs $|\omega_{i,j}|$ shows the amount of $\omega_{i,j}$ being affected by cancellations on the heuristics of the method, the more similar to a **smile plot**, the better. The last two plots show the relationship between $p-$values and the $\omega_{i,j}$. It is expected that higher ω's presents lower $p-$values.

The resulting object from the `wTO.Complete()` function is a list containing:

- wTO an edge list of information such as the signed and unsigned $\omega_{i,j}$ values its raw and adjusted $p-$values.
- Correlation values, also as an edge list.
- Quantiles, the quantiles from the empirical distribution and the calculated $omega$'s from the original data, for both signed and unsigned networks.

```
wTO_s_sign_complete
#> $wTO
```

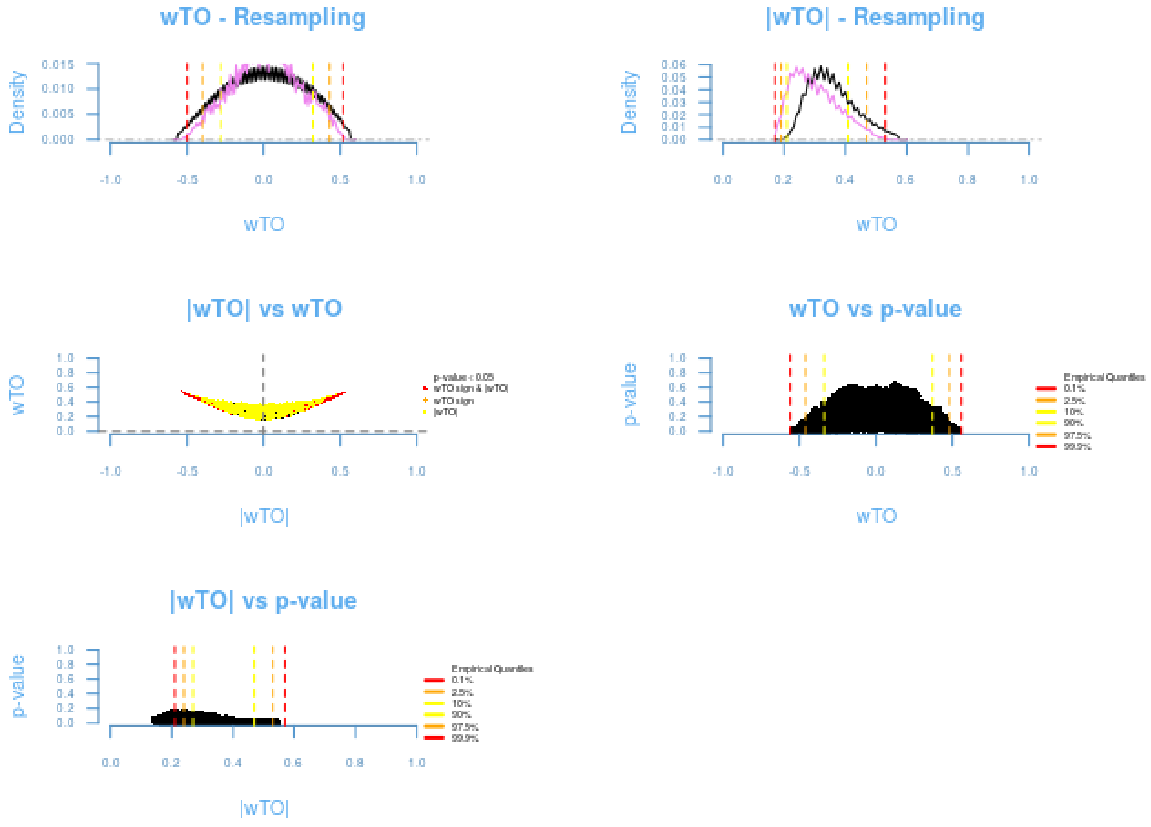


FIGURE A.1: Diagnosis of the wTO resampling

```
#>        Node.1  Node.2 wTO_sign wTO_abs pval_sig pval_abs  Padj_sig
#>     1: ZNF333   ZNF28   -0.099   0.237    0.168    0.004 0.3607366
#>     2: ZNF333 ANKRD22   -0.185   0.269    0.188    0.016 0.3607366
#>     3: ZNF333     ZFR   -0.117   0.242    0.180    0.012 0.3607366
#>     4: ZNF333  TRIM33    0.007   0.241    0.136    0.008 0.3607366
#>     5: ZNF333   RIMS3   -0.325   0.409    0.144    0.000 0.3607366
#>    ---
#> 14024: ANAPC2   SBNO2   -0.147   0.298    0.156    0.000 0.3607366
#> 14025: ANAPC2  ZNF528   -0.142   0.222    0.152    0.016 0.3607366
#> 14026:  TIGD7   SBNO2   -0.297   0.354    0.128    0.004 0.3607366
#> 14027:  TIGD7  ZNF528   -0.099   0.219    0.212    0.032 0.3607366
#> 14028:  SBNO2  ZNF528    0.141   0.311    0.368    0.024 0.4030531
#>          Padj_abs
#>     1: 0.01167541
#>     2: 0.02395390
#>     3: 0.02083624
#>     4: 0.01712559
#>     5: 0.00000000
#>    ---
#> 14024: 0.00000000
#> 14025: 0.02395390
#> 14026: 0.01167541
#> 14027: 0.03670450
#> 14028: 0.03016774
#>
#> $Correlation
#>          Node.1   Node.2          Cor
#>     1:  FAM122B DEFB108B  0.366857931
#>     2:  FAM122B   CCSER2  0.278870911
#>     3: DEFB108B   CCSER2 -0.252482453
#>     4:  FAM122B     GPD2 -0.005649124
#>     5: DEFB108B     GPD2 -0.107064848
#>    ---
#> 35774:   TRIM23   ZNF528  0.054249174
#> 35775:   ZNF559   ZNF528 -0.218309729
#> 35776:   ANAPC2   ZNF528 -0.013821370
#> 35777:    TIGD7   ZNF528  0.011807143
#> 35778:    SBNO2   ZNF528  0.092317502
#>
#> $Quantiles
#>                          0.1%  2.5%   10%  90% 97.5% 99.9%
#> Empirical.Quantile     -0.56 -0.46 -0.34 0.37  0.48  0.56
#> Quantile               -0.50 -0.40 -0.28 0.32  0.43  0.52
#> Empirical.Quantile.abs  0.21  0.24  0.27 0.47  0.53  0.57
#> Quantile.abs            0.17  0.19  0.21 0.41  0.47  0.53
```

Using the `wTO.fast()` function: The `wTO.fast()` function is a simplified version of the `wTO.Complete()` function, that doesn't return diagnosis, correlation, nor the quantiles, but allows the user to choose the method for correlation, the sign of the ω to be calculated and the resampling method should be one of the two **Bootstrap** or **BlockBootstrap**. The $p-$values are the raw $p-$values and if the user desires to calculate its correction it can be easily done as shown above.

```
fast_example = wTO.fast(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = 's',
sign = 'sign', delta = 0.2, n = 250,
method_resampling = 'Bootstrap')
#> There are 168 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don't turn off the computer.
#> Done!
```

```
head(fast_example)
#>     Node.1  Node.2    wTO  pval
#> 1:  ZNF333   ZNF28 -0.064 0.264
#> 2:  ZNF333 ANKRD22 -0.143 0.236
#> 3:   ZNF28 ANKRD22  0.029 0.188
#> 4:  ZNF333     ZFR -0.164 0.232
#> 5:   ZNF28     ZFR -0.011 0.256
#> 6: ANKRD22     ZFR  0.024 0.264

fast_example$adj.pval = p.adjust(fast_example$pval)
```

Metagenomic data

Along with the expression data, the `wTO` package also includes a metagenomics dataset that is the abundance of some OTU's in bacterias collected since 1997. More about this data can be found at https://www.ebi.ac.uk/metagenomics/projects/ERP013549.

The OTU (Operational Taxonomic Units) contains the taxonomy of the particular OTU and from Week1 to Week98, the abundance of that particular OTU in that week.

```
data("metagenomics_abundance")
metagenomics_abundance[2:10, 1:10]
#>
#> 2 Root;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;
#> 3 Root;k__Bacteria;p__Actinobacteria;c__Acidimicrobiia;
#> 4 Root;k__Bacteria;p__Actinobacteria;
#> 5 Root;k__Bacteria;p__Bacteroidetes;c__Cytophagia;
#> 6 Root;k__Bacteria;p__Bacteroidetes;
#> 7 Root;k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;
#> 8  Root;k__Bacteria;p__Bacteroidetes;
#> 9  Root;k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;
#> 10 Root;k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;
#>    Week1 Week2 Week3 Week4 Week5 Week6 Week7 Week8 Week9
#> 2      1     6     0     0     0     1     0     1     0
#> 3      0     0     0     0     0     0     0     5     0
#> 4      0     0     0     0     0     0     0     0     0
#> 5      0     1     0     0     0     0     0     0     0
#> 6      0     0     0     0     0     0     0     0     0
#> 7      0     1     0     0     0     0     0     1     0
#> 8      0     0     0     0     0     0     0     1     0
#> 9      0     0     0     0     0     0     0     1     0
#> 10     0     1     0     0     0     0     0     7     0
```

Before we are able to define the network, we have first to understand the patterns of autocorrelation of each species, and then define the lag, that will be used for the **BlockBootstrap** resampling in the `wTO.Complete()` or `fast.wTO()` functions. To define the lag, we use autocorrelation function `acf()`. The results of the acf() function are shown in Figure A.2.

```
row.names(metagenomics_abundance) = metagenomics_abundance$OTU
metagenomics_abundance = metagenomics_abundance[,-1]
par(mfrow = c(3,3))
for ( i in 1:nrow(metagenomics_abundance)){
  acf(t(metagenomics_abundance[i,]))
}
```

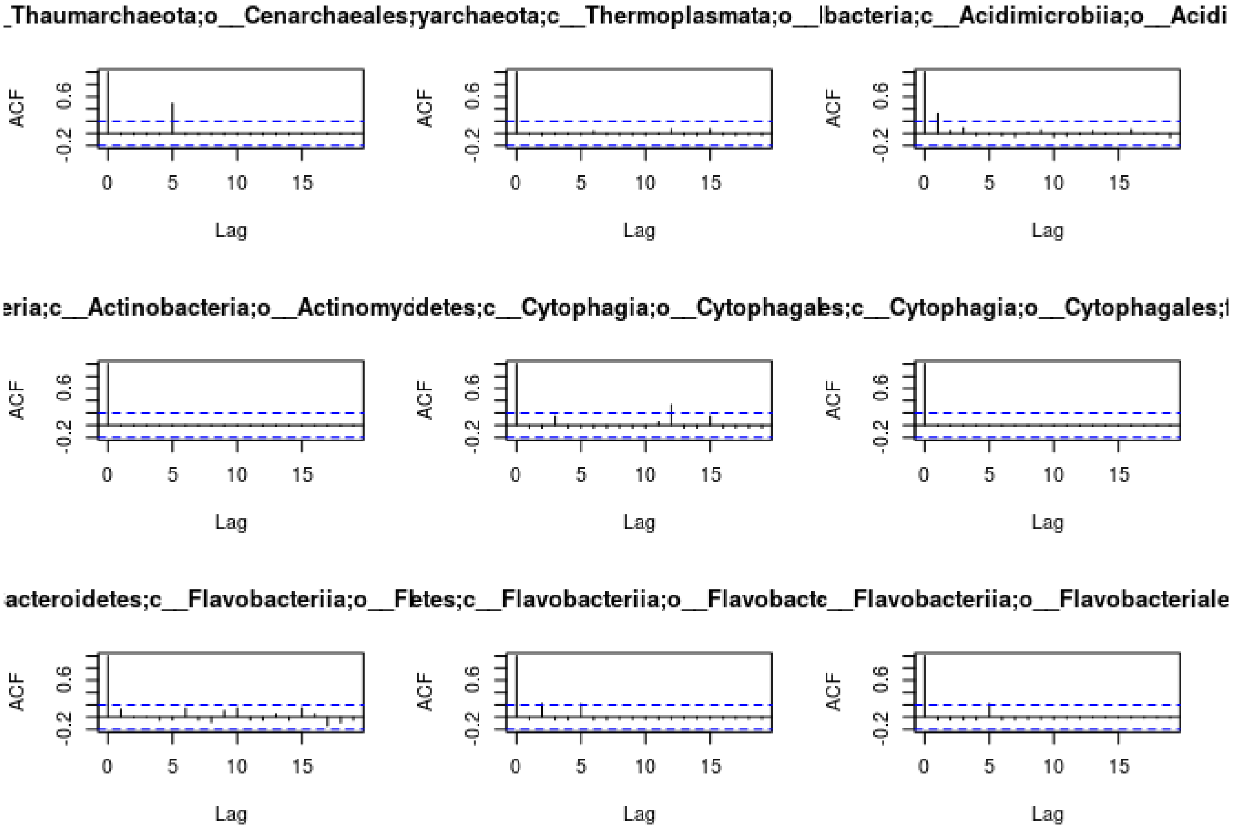


FIGURE A.2: Auto-correlogram of the OTUs

Because most of them have only a high autocorrelation with itself or maximum 2 weeks, we will use a lag of 2 for the blocks used in the bootstrap.

The functions `wTO.fast()` and `wTO.Complete()` are able to accommodate the lag parameter, therefore, they will be used here. Similarly to the previous analysis, the Figure A.3 represents the diagnostic of the resampling.

```
Meta_fast = wTO.fast(Data = metagenomics_abundance,
Overlap = row.names(metagenomics_abundance),
method = ’p’, sign = ’sign’, n = 250,
method_resampling = ’BlockBootstrap’, lag = 2)
#> There are 67 overlapping nodes, 67 total nodes and 98 individuals.
#> This function might take a long time to run. Don’t turn off the computer.
#> Done!

Meta_Complete = wTO.Complete(k = 1, n = 250, Data = metagenomics_abundance,
Overlap = row.names(metagenomics_abundance),
method = ’s’ , method_resampling = ’BlockBootstrap’, lag = 2 )
#> There are 67 overlapping nodes, 67 total nodes and 98 individuals.
#> This function might take a long time to run. Don’t turn off the computer.
#> Simulations are done.
#> Computing p-values
#> Computing cutoffs
#> Done!
```

A.1.3 Consensus Network

From the expression data-sets, we are able to draw a Consensus Network. For that, the function `wTO.Consensus()` can be used. This function works in a special way, that the user should pass a list of data.frames containing the Nodes names and the wTO and $p-$values. We show an example above.

Let’s calculate the networks the same way we did in the Subsection A.1.2.

```
wTO_Data1 = wTO.fast(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = ’p’, n = 250)
#> There are 168 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don’t turn off the computer.
#> Done!
wTO_Data2 = wTO.fast(Data = Microarray_Expression2,
Overlap = ExampleGRF$x, method = ’p’, n = 250)
#> There are 168 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don’t turn off the computer.
#> Done!
```

Now, let’s combine both networks in one Consensus Network.

```
CN_expression = wTO.Consensus(data = list (wTO_Data1 = data.frame
                                    (Node.1 = wTO_Data1$Node.1,
                                     Node.2 = wTO_Data1$Node.2,
                                     wTO = wTO_Data1$wTO,
```

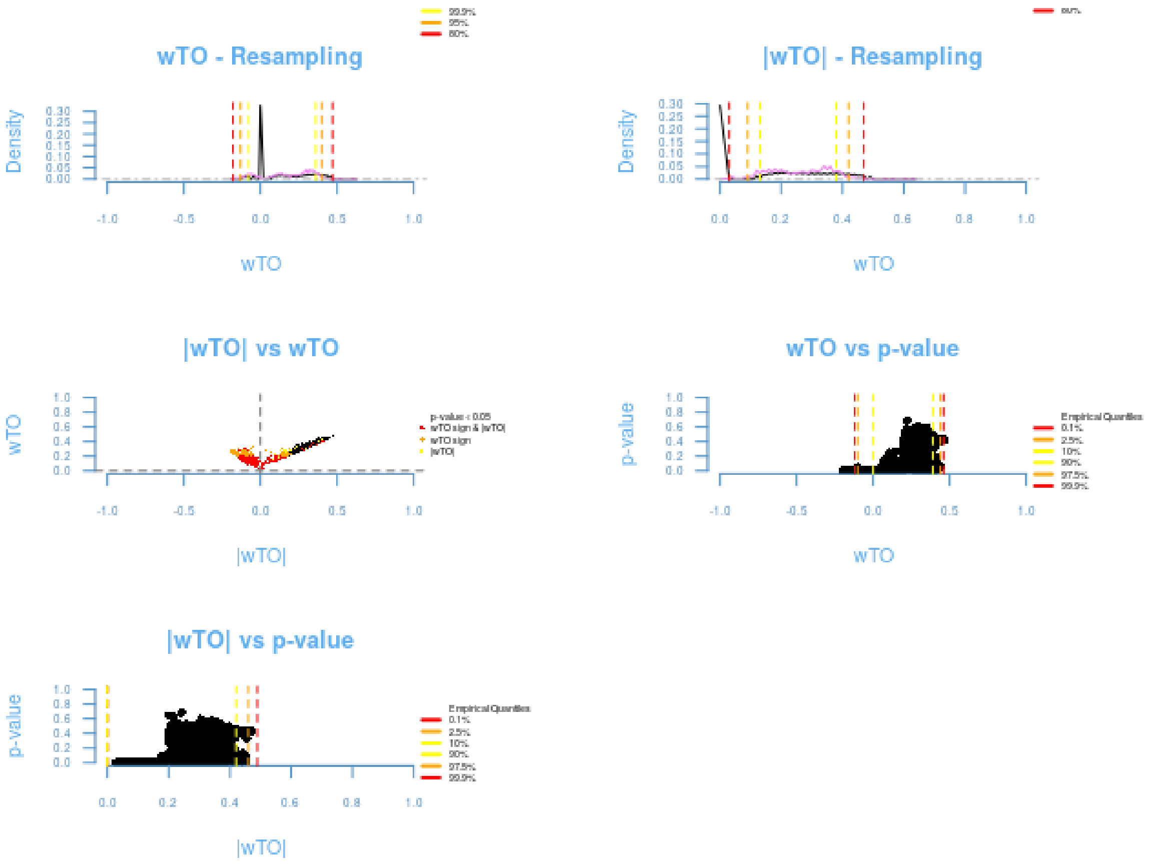


FIGURE A.3: Diagnostic plot of wTO in a time series approach

```
                                      pval = wTO_Data1$pval)
                                    , wTO_Data2C = data.frame
                                    (Node.1 = wTO_Data2$Node.1,
                                     Node.2 = wTO_Data2$Node.2,
                                     wTO = wTO_Data2$wTO,
                                      pval = wTO_Data2$pval)))
#> Joining by: Node.1, Node.2
#> Joining by: Node.1, Node.2
#> Joining by: ID
#> Total common nodes: 168
```

Or using the `wTO.Complete()`:

```
wTO_Data1C = wTO.Complete(Data = Microarray_Expression1,
Overlap = ExampleGRF$x, method = 'p', n = 250, k = 5, plot = F)
#> There are 168 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don't turn off the computer.
#> Simulations are done.
#> Computing p-values
#> Computing cutoffs
#> Done!
wTO_Data2C = wTO.Complete(Data = Microarray_Expression2,
Overlap = ExampleGRF$x, method = 'p', n = 250, k = 5, plot = F)
#> There are 168 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don't turn off the computer.
#> Simulations are done.
#> Computing p-values
#> Computing cutoffs
#> Done!
```

Now, let's combine both networks in one Consensus Network.

```
CN_expression = wTO.Consensus(data = list (wTO_Data1C = data.frame
                                    (Node.1 = wTO_Data1C$wTO$Node.1,
                                     Node.2 = wTO_Data1C$wTO$Node.2,
                                     wTO = wTO_Data1C$wTO$wTO_sign,
                                      pval = wTO_Data1C$wTO$pval_sig),

                                      wTO_Data2C = data.frame
                                    (Node.1 = wTO_Data2C$wTO$Node.1,
                                     Node.2 = wTO_Data2C$wTO$Node.2,
                                     wTO = wTO_Data2C$wTO$wTO_sign,
                                      pval = wTO_Data2C$wTO$pval_sig)))
```

```
#> Joining by: Node.1, Node.2
#> Joining by: Node.1, Node.2
#> Joining by: ID
#> Total common nodes: 168

head(CN_expression)
#>   Node.1  Node.2         CN pval.fisher
#> 1 ZNF333   ZNF28 -0.1191288  0.06006662
#> 2 ZNF333 ANKRD22 -0.1400000  0.13016905
#> 3 ZNF333     ZFR -0.1091659  0.10147819
#> 4 ZNF333  TRIM33 -0.0240000  0.11707440
#> 5 ZNF333   RIMS3 -0.2798101  0.10797662
#> 6 ZNF333  ZNF595  0.1542850  0.11529832
```

A.1.4 Visualisation

The wTO package also includes an interactive visualisation tool that can be used to inspect the results of the wTO networks or Consensus Network.

The arguments given to this function are the Nodes names, its wTO and $p-$values. Optional are the cutoffs that can be applied to the p-value or to the wTO value. We highly recommend using both by subsetting the data previous to the visualisation. The layout of the network can be also chosen from a variety that are implemented in igraph package, for the the Make_Cluster argument many clustering algorithms that are implemented in igraph can be used. The final graph can be exported as an html or as png.

```
Visualization = NetVis(Node.1 = CN_expression$Node.1,
       Node.2 = CN_expression$Node.2,
       wTO = CN_expression$CN,
       pval = CN_expression$pval.fisher,
       cutoff = list(kind = 'pval', value = 0.001),
       MakeGroups = 'louvain', layout = 'layout_components')
#> Joining by: id

CN_expression_filtered = subset(CN_expression,
abs(CN_expression$CN)> 0.4 &
CN_expression$pval.fisher < 0.0001)

dim(CN_expression_filtered)
#> [1] 45  4

Visualization2 = NetVis(
  Node.1 = CN_expression_filtered$Node.1,
       Node.2 = CN_expression_filtered$Node.2,
       wTO = CN_expression_filtered$CN,
       pval = CN_expression_filtered$pval.fisher,
       cutoff = list(kind = 'pval', value = 0.001),
       MakeGroups = 'louvain',
       layout = 'layout_components', path = 'Vis.html')
#> Joining by: id
#> Vis.html
```

B

Co-Expression Differential Network Analysis: CoDiNA

B.1 CoDiNA manual

The usage of the Co-expression Differential Network analysis has been growing by the Biological and Medical science for the analysis of complex systems or diseases. We have developed a method that is able to compare as many networks as desired, by characterising both links and nodes that are common, different or specific to each network.

You can download the package from CRAN using:

```
install.packages('CoDiNA')
```

B.1.1 Input data

The input data for CoDiNA is a list of data.frame, containing: `Node.1`, `Node.2` and `value`. It is important to mention here that the methodology should be employed only for **undirected graphs**. The `value` is the strength of the link between `Node.1` and `Node.2` and must any real number between -1 to 1. This value can be re-normalised by the package using the argument `stretch = TRUE` (by default the values are normalised).

As an example, the `CoDiNA` package contains 4 datasets from a Cancer study, GSE4290[163]. Each one of this datasets was previously normalised, the control quality was done for the genes and the networks were calculate using the `wTO` package[62]. Each dataset consists of the Gene names and the weight only for the significant interactions and filtered for a wTO value of |0.3|.

Using the wTO output for CoDiNA

The output from the `wTO` package can be easily used as input for `CoDiNA`.

```
require(wTO)
#> Loading required package: wTO
require(CoDiNA)
#> Loading required package: CoDiNA
require(magrittr)
#> Loading required package: magrittr

wTO_out = wTO.fast(Data = Microarray_Expression1, n = 100)
#> There are 268 overlapping nodes, 268 total nodes and 18 individuals.
#> This function might take a long time to run. Don't turn off the
#> computer.
#> Done!

wTO_filtered = subset(wTO_out, p.adjust(wTO_out$pval) < 0.05,
select = c('Node.1', 'Node.2', 'wTO'))
```

B.1.2 Creating the Differential Network

To generate the differential network one can use the `MakeDiffNet()` function.

This function will return the Φ and $\widetilde{\Phi}$ classification for each one of the links. Connections that are assigned to α (a) are in agreement in all networks and it means that all networks possess that particular link with the same sign. Links classified as β (b) are the ones that also exist in all networks but at least one network contains it with a different sign. The category γ (g) contains links that does not exist in all networks, meaning that they are specific to at least one network.

This function also assigns the link into a sub-category. It is important mainly for the β and γ links to understand its differences or specificities. It is important to note that **the first network is considered to be the reference for β and γ links**.

The output from this function is a data.frame containing the nodes, the original weights (or normalised), the Phi and Phi_tilde categories, a Group, which describes the sign or absence of the link, the Score_center (raw score), Score_Phi (normalised score by Φ), Score_Phi_tilde (normalised score by $\widetilde{\Phi}$, Score_internal (score of the link to its theoretical category). The first 3 scores, should be closer to 1, while for the last one, the closer to 0 the better.

```
DiffNet = MakeDiffNet(Data = list(CTR, OLI, AST),
Code = c('CTR', 'OLI', 'AST'))
#> Starting now.
#> CTR contains 17471 edges and 1022 nodes.
#> OLI contains 64791 edges and 1697 nodes.
#> AST contains 3384 edges and 1002 nodes.
#> Total of nodes: 442
#> Total of edges: 82558
#> Coding correlations.
#> Total of edges (inside the cutoff): 15950
#> Starting Phi categorization.
#> Coding the groups.
#> Recode is done!

DiffNet
#> Nodes 441
#> Links 15950

print(DiffNet) %>% head()
#> Nodes 441
#> Links 15950
#>   Node.1 Node.2        CTR        OLI AST Phi Phi_tilde            Group
#> 1   CTCF NKX6-3 -0.8861789 -0.7756813   0   g g.CTR.OLI  -CTR,-OLI,NoAST
#> 2   IRF3 NKX6-3 -0.8520325  0.0000000   0   g     g.CTR -CTR,NoOLI,NoAST
#> 3 NKX6-3    TDG -0.9040650 -0.8385744   0   g g.CTR.OLI  -CTR,-OLI,NoAST
#> 4  BUD31 NKX6-3 -0.8016260 -0.7484277   0   g g.CTR.OLI  -CTR,-OLI,NoAST
#> 5  HMGN3 NKX6-3 -0.8878049 -0.8364780   0   g g.CTR.OLI  -CTR,-OLI,NoAST
#> 6 NKX6-3  PUF60 -0.9479675  0.0000000   0   g     g.CTR -CTR,NoOLI,NoAST
#>   Score_center Score_Phi Score_Phi_tilde Score_internal
#> 1    0.8327648 0.5467547       0.5235928     0.17786809
#> 2    0.8520325 0.5989745       0.5937500     0.14796748
#> 3    0.8719348 0.6529143       0.6723478     0.13278127
#> 4    0.7754832 0.3915083       0.3060555     0.22654014
#> 5    0.8625233 0.6274069       0.6366059     0.14022695
#> 6    0.9479675 0.8589800       0.8571429     0.05203252
```

B.1.3 Clustering the nodes into Φ and $\widetilde{\Phi}$ categories

Because exclusively the information about the links is not enough to define a network, it is necessary to define the nodes accordingly to its Φ and $\widetilde{\Phi}$ categories. To do so, the function `ClusterNodes()` can be used. The input for this function is `DiffNet`, that is the output from the `MakeDiffNet()`, besides the external and internal cutoffs. The external cutoff is applied to the normalised $\widetilde{\Phi}$ Score, while the internal cutoff is applied to the internal Score.

The suggested values for the internal and external cutoffs are the median or the first and third quantiles of the internal and $\widetilde{\Phi}$ scores, depending on how conservative the network should be.

Using the median:

```
int_C = quantile(DiffNet$Score_internal, 0.5)
```

```
ext_C = quantile(DiffNet$Score_Phi, 0.5)

Nodes_Groups = ClusterNodes(DiffNet = DiffNet,
cutoff.external = ext_C, cutoff.internal = int_C)
table(Nodes_Groups$Phi_tilde)
#>
#>     g.AST     g.CTR g.CTR.OLI     g.OLI g.OLI.AST         U
#>        11       213         2       125         1        66
```

Using the first and third quantile:

```
int_C = quantile(DiffNet$Score_internal, 0.25)
ext_C = quantile(DiffNet$Score_Phi, 0.75)

Nodes_Groups = ClusterNodes(DiffNet = DiffNet,
cutoff.external = ext_C, cutoff.internal = int_C)
table(Nodes_Groups$Phi_tilde)
#>
#> g.AST g.CTR g.OLI     U
#>     8   188    64    80
```

B.1.4 Plotting the network

The visualisation of the final network can be quickly done with `plot`. The layout of the network can be also determined from a variety that is implemented in igraph package, the `Make_Cluster` argument allows the nodes to be clustered according to many clustering algorithms that are implemented in igraph can be used. The final graph can be exported as an `HTML` or as `png`. The argument `path` saves the network in the given path.

The plot returns the nodes and its information.

```
int_C = quantile(DiffNet$Score_internal, 0.25)
ext_C = quantile(DiffNet$Score_Phi, 0.75)

Graph = plot(DiffNet, cutoff.external = ext_C,
cutoff.internal = int_C,
layout = 'layout_components',
path = 'Vis.html')
#> Vis.html
```

The graph can also be exported as an `igraph` object, that can be further plotted.

```
g = as.igraph(Graph)

require(igraph)
```

```
#> Loading required package: igraph
#>
#> Attaching package: 'igraph'
#> The following objects are masked from 'package:CoDiNA':
#>
#>     as.igraph, normalize
#> The following objects are masked from 'package:stats':
#>
#>     decompose, spectrum
#> The following object is masked from 'package:base':
#>
#>     union
```

```
plot(g, layout = layout.fruchterman.reingold(g), vertex.label = NA)
```

C

Datasets for the mental disorders study

Table showing the studies used for the multivariate analysis. From the brain regions Cerebellum (CER), Ganglia (GAN), Hippocampus (HIP), Prefrontal Cortex (PFC) and Temporal Cortex (TC). The diseases collected for humans are: Autism Spectrum Disorder (ASD), Alzheimer's Disease (AD), Bipolar Disorder (BD), Major Depression Disorder (MDD), Schizophrenia (SCZ) and Parkinson Disease (PD). Moreover, data was collected for Chimpanzee (CMP) and Rhesus macaque (RH).

Table C.1: Datasets used to construct the networks for the mental disorders study.

Study	Platform	Disorder	Tissue	Age	Genes	TFs	m
GSE28521	GPL6883	ASD	CER	Adult	9005	2117	10
GSE35974	GPL6244	BD	CER	Adult	20252	3154	37
GSE22569	GPL6244	CMP	CER	Adult	20366	3154	5
GSE28521	GPL6883	Control (CTR)	CER	Adult	9005	2117	11
GSE2164	GPL91	CTR	CER	Adult	4840	1339	13
GSE22569	GPL6244	CTR	CER	Adult	20366	3154	15
GSE22570	GPL6244	CTR	CER	Adult	16563	2741	15
GSE2164	GPL8300	CTR	CER	Adult	5896	1617	17
GSE4036	GPL570	CTR	CER	Adult	15031	3042	18
GSE25219	GPL5175	CTR	CER	Adult	18345	2760	22
GSE25219	GPL5188	CTR	CER	Adult	18333	2753	22
GSE35974	GPL6244	CTR	CER	Adult	20252	3154	50
GSE45642	GPL17027	CTR	CER	Adult	9298	2426	79
GSE60862	GPL5175	CTR	CER	Adult	22134	3153	130
GSE60863	GPL5188	CTR	CER	Adult	22134	3153	130
GSE22570	GPL6244	CTR	CER	Infant	16563	2741	9
GSE25219	GPL5188	CTR	CER	Infant	18333	2753	24
GSE25219	GPL5175	CTR	CER	Infant	18345	2760	25
GSE35974	GPL6244	MDD	CER	Adult	20252	3154	13
GSE22569	GPL6244	RH	CER	Adult	20366	3154	8
GSE4036	GPL570	SCZ	CER	Adult	15031	3042	10
GSE35974	GPL6244	SCZ	CER	Adult	20252	3154	44
GSE11512	GPL6879	CTR	GAN	Adult	13408	3154	9
GSE7621	GPL570	CTR	GAN	Adult	13140	2979	9
GSE8397	GPL96	CTR	GAN	Adult	8137	2268	11
GSE8397	GPL97	CTR	GAN	Adult	4876	934	11
GSE44593	GPL570	CTR	GAN	Adult	13928	3003	14

Table C.1: Datasets used to construct the networks for the mental disorders study.

Study	Platform	Disorder	Tissue	Age	Genes	TFs	m
GSE54566	GPL570	CTR	GAN	Adult	14866	3070	14
GSE54564	GPL6947	CTR	GAN	Adult	18658	3083	21
GSE45642	GPL17027	CTR	GAN	Adult	9383	2480	62
GSE60862	GPL5175	CTR	GAN	Adult	22115	3153	101
GSE60863	GPL5188	CTR	GAN	Adult	22115	3153	101
GSE60862	GPL5175	CTR	GAN	Adult	22110	3153	129
GSE60863	GPL5188	CTR	GAN	Adult	22110	3153	129
GSE44593	GPL570	MDD	GAN	Adult	13928	3003	14
GSE54566	GPL570	MDD	GAN	Adult	14866	3070	14
GSE54564	GPL6947	MDD	GAN	Adult	18658	3083	21
GSE7621	GPL570	PD	GAN	Adult	13900	3029	16
GSE8397	GPL96	PD	GAN	Adult	8580	2353	24
GSE8397	GPL97	PD	GAN	Adult	5305	994	24
GSE42581	GPL3535	RH	GAN	Adult	10048	2201	12
GSE42581	GPL3535	RH	GAN	Adult	10363	2241	12
GSE1297	GPL96	AD	HIP	Adult	8854	2410	22
GSE53987	GPL570	BD	HIP	Adult	16008	3108	18
GSE1297	GPL96	CTR	HIP	Adult	8326	2324	9
GSE36980	GPL6244	CTR	HIP	Adult	16500	2752	11
GSE53987	GPL570	CTR	HIP	Adult	16008	3108	18
GSE25219	GPL5188	CTR	HIP	Adult	18287	2751	19
GSE25219	GPL5175	CTR	HIP	Adult	18243	2728	20
GSE45642	GPL17027	CTR	HIP	Adult	9678	2515	108
GSE60862	GPL5175	CTR	HIP	Adult	22112	3153	122
GSE60863	GPL5188	CTR	HIP	Adult	22112	3153	122
GSE25219	GPL5188	CTR	HIP	Infant	18287	2751	30
GSE25219	GPL5175	CTR	HIP	Infant	18243	2728	31
GSE53987	GPL570	MDD	HIP	Adult	16008	3108	17
GSE13824	GPL3535	RH	HIP	Adult	10336	2234	9
GSE13824	GPL3535	RH	HIP	Adult	10339	2228	9
GSE11697	GPL3535	RH	HIP	Adult	10183	2209	11
GSE11697	GPL3535	RH	HIP	Adult	10345	2226	12
GSE53987	GPL570	SCZ	HIP	Adult	16008	3108	15
GSE53697	RNAseq	AD	PFC	Adult	17133	2685	9
GSE36980	GPL6244	AD	PFC	Adult	16544	2753	15
GSE33000	GPL4372	AD	PFC	Adult	15348	2813	310
GSE28521	GPL6883	ASD	PFC	Adult	9005	2117	16
GSE59288	RNAseq	ASD	PFC	Adult	20684	3040	34
GSE5392	GPL96	BD	PFC	Adult	9179	2466	10
GSE12654	GPL8300	BD	PFC	Adult	6328	1757	11
GSE53987	GPL570	BD	PFC	Adult	15951	3105	15
GSE5388	GPL96	BD	PFC	Adult	9630	2515	30
GSE5392	GPL96	BD	PFC	Adult	9321	2482	30
GSE22521	GPL6244	CMP	PFC	Adult	20338	3154	7
GSE59288	RNAseq	CMP	PFC	Adult	6383	1261	39
GSE2164	GPL91	CTR	PFC	Adult	4861	1372	4
GSE53697	RNAseq	CTR	PFC	Adult	18720	2877	8
GSE5392	GPL96	CTR	PFC	Adult	8861	2405	11
GSE54570	GPL96	CTR	PFC	Adult	9230	2452	13
GSE54567	GPL570	CTR	PFC	Adult	14743	3071	14
GSE12654	GPL8300	CTR	PFC	Adult	5864	1664	15
GSE18069	GPL6244	CTR	PFC	Adult	16534	2747	15
GSE20168	GPL96	CTR	PFC	Adult	8704	2378	15
GSE22521	GPL6244	CTR	PFC	Adult	20338	3154	15
GSE22570	GPL6244	CTR	PFC	Adult	16534	2747	15
GSE54568	GPL570	CTR	PFC	Adult	14715	3072	15
GSE28521	GPL6883	CTR	PFC	Adult	9005	2117	16
GSE36980	GPL6244	CTR	PFC	Adult	16532	2751	18
GSE53987	GPL570	CTR	PFC	Adult	15951	3105	19
GSE17612	GPL570	CTR	PFC	Adult	12923	2942	22
GSE25219	GPL5175	CTR	PFC	Adult	18290	2746	23
GSE25219	GPL5188	CTR	PFC	Adult	18290	2769	23

TABLE C.1: Datasets used to construct the networks for the mental disorders study.

Study	Platform	Disorder	Tissue	Age	Genes	TFs	m
GSE11512	GPL6879	CTR	PFC	Adult	14296	3045	26
GSE2164	GPL8300	CTR	PFC	Adult	6009	1641	26
GSE5388	GPL96	CTR	PFC	Adult	9630	2515	31
GSE5392	GPL96	CTR	PFC	Adult	9104	2452	31
GSE51264	RNAseq	CTR	PFC	Adult	20551	3015	38
GSE53890	GPL570	CTR	PFC	Adult	14318	3052	41
GSE68719	RNAseq	CTR	PFC	Adult	19365	3012	44
GSE60862	GPL5175	CTR	PFC	Adult	22118	3153	127
GSE60863	GPL5188	CTR	PFC	Adult	22118	3153	127
GSE33000	GPL4372	CTR	PFC	Adult	15348	2813	157
GSE45642	GPL17027	CTR	PFC	Adult	9727	2507	161
GSE18069	GPL6244	CTR	PFC	Infant	16534	2747	10
GSE22570	GPL6244	CTR	PFC	Infant	16534	2747	10
GSE11512	GPL6879	CTR	PFC	Infant	14296	3045	18
GSE25219	GPL5175	CTR	PFC	Infant	18290	2746	26
GSE25219	GPL5188	CTR	PFC	Infant	18290	2769	26
GSE12654	GPL8300	MDD	PFC	Adult	6130	1717	11
GSE54570	GPL570	MDD	PFC	Adult	9230	2452	13
GSE54567	GPL570	MDD	PFC	Adult	14715	3072	15
GSE54568	GPL570	MDD	PFC	Adult	14715	3072	15
GSE53987	GPL570	MDD	PFC	Adult	15951	3105	17
GSE20168	GPL96	PD	PFC	Adult	9039	2433	14
GSE68719	RNAseq	PD	PFC	Adult	18846	2989	29
GSE22521	GPL6244	RH	PFC	Adult	20338	3154	12
GSE51264	RNAseq	RH	PFC	Adult	12220	2536	34
GSE12654	GPL8300	SCZ	PFC	Adult	6103	1728	13
GSE53987	GPL570	SCZ	PFC	Adult	15951	3105	17
GSE17612	GPL570	SCZ	PFC	Adult	13303	2981	28
GSE36980	GPL6244	AD	TC	Adult	16542	2755	10
GSE28521	GPL6883	ASD	TC	Adult	9005	2117	13
GSE28521	GPL6883	CTR	TC	Adult	9005	2117	13
GSE36980	GPL6244	CTR	TC	Adult	16522	2760	18
GSE21935	GPL570	CTR	TC	Adult	12508	2891	19
GSE25219	GPL5175	CTR	TC	Adult	18317	2794	20
GSE25219	GPL5188	CTR	TC	Adult	18268	2735	21
GSE60862	GPL5175	CTR	TC	Adult	22114	3153	119
GSE60863	GPL5188	CTR	TC	Adult	22114	3153	119
GSE25219	GPL5175	CTR	TC	Infant	18317	2794	31
GSE25219	GPL5188	CTR	TC	Infant	18268	2735	31
GSE21935	GPL570	SCZ	TC	Adult	13025	2943	23

D

GO enrichment for Neurodegenerative disorders

Table showing the GO enrichment for the neurodegenerative disorders, AD and PD. Humans are used as controls. Adults are represented as CTR5 and infants as CTR0.

Table D.1: GO enrichment for Neurodegenerative disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γPD	regulation of BMP signaling pathway	25	0.5500	0.0002	0.9918	0.0002
γPD	cardiac septum morphogenesis	33	0.7300	0.0006	0.7130	0.0006
γPD	negative regulation of oxidative stress-...	9	0.2000	0.0008	0.8769	0.0008
γPD	fat cell differentiation	75	1.6600	0.0011	0.2373	0.0011
γPD	adenohypophysis development	3	0.0700	0.0014	0.9988	0.0014
γPD	ureteric bud development	39	0.8600	0.0014	0.5675	0.0014
γPD	regulation of insulin-like growth factor...	4	0.0900	0.0028	0.9984	0.0028
γPD	neuron fate determination	4	0.0900	0.0028	0.9983	0.0028
γPD	positive regulation of astrocyte differe...	4	0.0900	0.0028	0.9984	0.0028
γPD	common bile duct development	4	0.0900	0.0028	0.9993	0.0028
γPD	energy homeostasis	4	0.0900	0.0028	0.9981	0.0028
γPD	Wnt signaling pathway involved in midbra...	4	0.0900	0.0028	0.9994	0.0028
γPD	regulation of phosphatase activity	16	0.3500	0.0003	0.9934	0.0036
γPD	endochondral bone growth	5	0.1100	0.0046	0.9986	0.0046
γPD	hepatocyte differentiation	5	0.1100	0.0046	0.9991	0.0046
γPD	negative regulation of alcohol biosynthe...	5	0.1100	0.0046	0.2958	0.0046
γPD	outflow tract morphogenesis	35	0.7700	0.0067	0.8138	0.0067
γPD	negative regulation of transcription by ...	6	0.1300	0.0068	0.9565	0.0068
γPD	Notch signaling pathway	57	1.2600	0.0076	0.0928	0.0076
γPD	negative regulation of epidermal cell di...	7	0.1500	0.0094	0.4619	0.0094
γPD	beta-catenin destruction complex disasse...	7	0.1500	0.0094	0.7607	0.0094
γCTR5.PD	muscle cell cellular homeostasis	4	0.0300	0.0004	0.7410	0.0004
γCTR5.PD	heart trabecula formation	6	0.0500	0.0010	0.9437	0.0010
γCTR5.PD	dorsal aorta morphogenesis	7	0.0600	0.0014	0.9430	0.0014
γCTR5.PD	eyelid development in camera-type eye	7	0.0600	0.0014	0.9995	0.0014
γCTR5.PD	regulation of cell shape	12	0.1000	0.0044	0.8527	0.0044
γCTR5.PD	cardiac muscle cell development	13	0.1100	0.0052	0.2880	0.0052
γCTR5.PD	associative learning	15	0.1300	0.0069	0.5018	0.0069
γCTR5.PD	regulation of heart rate by hormone	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	positive regulation of heart rate	1	0.0100	0.0086	1.0000	0.0086

TABLE D.1: GO enrichment for Neurodegenerative disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γCTR5.PD	uropod organization	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	nerve growth factor signaling pathway	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	natural killer cell degranulation	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	negative regulation of skeletal muscle t...	1	0.0100	0.0086	1.0000	0.0086
γCTR5.PD	negative regulation of actin nucleation	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	potassium ion homeostasis	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	early endosome to recycling endosome tra...	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	membrane repolarization during ventricul...	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	negative regulation of potassium ion exp...	1	0.0100	0.0086	0.9999	0.0086
γCTR5.PD	negative regulation of cardiac vascular ...	1	0.0100	0.0086	1.0000	0.0086
γCTR5.PD	negative regulation of myoblast prolifer...	1	0.0100	0.0086	1.0000	0.0086
γCTR5.PD	anterior/posterior axis specification	17	0.1500	0.0089	0.9900	0.0089
γCTR5.PD	regulation of transcription initiation f...	18	0.1500	0.0099	0.9596	0.0099
γCTR5.CTR0.AD	protein modification by small protein co...	219	38.2800	0.0000	0.7940	0.0001
γCTR5.CTR0.AD	positive regulation of cytoplasmic mRNA ...	4	0.7000	0.0009	0.9340	0.0009
γCTR5.CTR0.AD	neutrophil degranulation	27	4.7200	0.0010	0.7500	0.0010
γCTR5.CTR0.AD	nuclear-transcribed mRNA catabolic proce...	19	3.3200	0.0025	0.9020	0.0017
γCTR5.CTR0.AD	positive regulation of translational ini...	6	1.0500	0.0008	0.9910	0.0039
γCTR5.CTR0.AD	regulation of deoxyribonuclease activity	5	0.8700	0.0040	0.9950	0.0040
γCTR5.CTR0.AD	regulation of intracellular estrogen rec...	14	2.4500	0.0053	0.8450	0.0053
γCTR5.CTR0.AD	positive regulation of activin receptor ...	3	0.5200	0.0053	0.9900	0.0053
γCTR5.CTR0.AD	histone H4-K20 trimethylation	3	0.5200	0.0053	0.9450	0.0053
γCTR5.CTR0.AD	establishment of integrated proviral lat...	3	0.5200	0.0053	0.9980	0.0053
γCTR5.CTR0.AD	negative regulation of telomere maintena...	8	1.4000	0.0056	0.9530	0.0056
γCTR5.CTR0.AD	DNA-templated transcription, elongation	86	15.0300	0.0022	0.6270	0.0060
γCTR5.CTR0.AD	snRNA transcription by RNA polymerase II	42	7.3400	0.0091	0.6940	0.0091
γCTR5.CTR0.AD	iron ion homeostasis	12	2.1000	0.0099	0.8300	0.0099
γCTR5.CTR0.AD	positive regulation of NIK/NF-kappaB siγ..	12	2.1000	0.0099	0.9520	0.0099
γCTR5.CTR0	histone lysine demethylation	19	2.2800	0.0047	0.9370	0.0047
γCTR5.CTR0	mitotic G2/M transition checkpoint	7	0.8400	0.0053	0.8010	0.0053
γCTR5.AD.PD	ventricular cardiac muscle cell developm...	5	0.0700	0.0021	0.9990	0.0021
γCTR5.AD.PD	heart trabecula formation	6	0.0900	0.0031	0.4985	0.0031
γCTR5.AD	endocardium morphogenesis	6	1.0700	0.0009	0.9993	0.0009
γCTR5.AD	cellular response to ether	4	0.7100	0.0010	0.9679	0.0010
γCTR5.AD	positive regulation of cardiac muscle ce...	10	1.7800	0.0034	0.9802	0.0034
γCTR5.AD	cellular response to follicle-stimulatin...	3	0.5300	0.0056	0.9693	0.0056
γCTR5.AD	regulation of NMDA receptor activity	3	0.5300	0.0056	0.9613	0.0056
γCTR5.AD	post-translational protein modification	52	9.2600	0.0064	0.6060	0.0064
γCTR5.AD	7-methylguanosine mRNA capping	21	3.7400	0.0064	0.4703	0.0064
γCTR5	intracellular estrogen receptor signalin...	27	3.8000	0.0001	1.0000	0.0001
γCTR5	endocrine system development	54	7.6000	0.0002	0.9490	0.0006
γCTR5	visual perception	21	2.9600	0.0012	0.2960	0.0012
γCTR5	ear morphogenesis	34	4.7900	0.0004	0.2960	0.0016
γCTR5	dosage compensation by inactivation of X...	3	0.4200	0.0028	0.9570	0.0028
γCTR5	cell fate specification	33	4.6500	0.0039	0.8480	0.0066
γCTR5	epithelial cell maturation	7	0.9900	0.0095	0.7250	0.0095
γCTR5	regulation of macrophage derived foam ce...	7	0.9900	0.0095	0.8220	0.0095
γCTR5	replication fork processing	7	0.9900	0.0095	0.9720	0.0095
γCTR5	positive regulation of insulin secretion...	4	0.5600	0.0099	0.8980	0.0099
γCTR5	positive regulation of isotype switchinγ..	4	0.5600	0.0099	0.9160	0.0099
γCTR5	establishment of skin barrier	4	0.5600	0.0099	0.9630	0.0099
γCTR5	positive regulation of histone H4 acetyl...	4	0.5600	0.0099	0.8780	0.0099
γCTR0.AD	positive regulation of protein localizat...	15	0.3500	0.0045	0.5509	0.0045
γCTR0.AD	response to ischemia	5	0.1200	0.0051	0.6607	0.0051
γCTR0.AD	response to UV-B	5	0.1200	0.0051	0.9997	0.0051
γCTR0.AD	regulation of release of sequestered cal...	5	0.1200	0.0051	0.6504	0.0051
γCTR0.AD	cellular response to interleukin-1	17	0.4000	0.0065	0.8689	0.0065
γCTR0.AD	response to amino acid	18	0.4200	0.0077	0.0765	0.0077
γCTR0.AD	positive regulation of intracellular pro...	37	0.8600	0.0099	0.2942	0.0099
γCTR0	peripheral nervous system development	20	3.3700	0.0006	0.7687	0.0004
γCTR0	response to corticosterone	6	1.0100	0.0007	0.9755	0.0007
γCTR0	glial cell development	17	2.8700	0.0007	0.9815	0.0007

TABLE D.1: GO enrichment for Neurodegenerative disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γCTR0	glycogen biosynthetic process	4	0.6700	0.0008	0.9799	0.0008
γCTR0	regulation of gluconeogenesis by regulat...	4	0.6700	0.0008	0.9487	0.0008
γCTR0	positive regulation of T cell differenti...	3	0.5100	0.0048	0.9921	0.0048
γCTR0	spermatogenesis	89	15.0100	0.0048	0.1413	0.0053
γCTR0	oocyte maturation	7	1.1800	0.0021	0.8877	0.0089
γCTR0	histone H3-K27 trimethylation	6	1.0100	0.0090	0.9623	0.0090
γAD	regulation of GTPase activity	53	7.0500	0.0003	0.7828	0.0003
γAD	cellular response to mechanical stimulus	13	1.7300	0.0006	0.8581	0.0006
γAD	pallium development	41	5.4500	0.0005	0.7377	0.0008
γAD	regulation of I-kappaB kinase/NF-kappaB ...	48	6.3900	0.0009	0.7797	0.0009
γAD	defense response to virus	36	4.7900	0.0004	0.9892	0.0012
γAD	negative regulation of myoblast differen...	11	1.4600	0.0014	0.8479	0.0014
γAD	regulation of extrinsic apoptotic signal...	36	4.7900	0.0016	0.3743	0.0016
γAD	regulation of sodium ion transport	5	0.6700	0.0014	0.9946	0.0023
γAD	activation of transmembrane receptor pro...	3	0.4000	0.0023	0.9933	0.0023
γAD	branchiomotor neuron axon guidance	3	0.4000	0.0023	0.9683	0.0023
γAD	negative regulation of multicellular orγ..	3	0.4000	0.0023	0.9987	0.0023
γAD	regulation of small GTPase mediated sign...	23	3.0600	0.0003	0.7994	0.0026
γAD	negative regulation of cellular response...	20	2.6600	0.0026	0.7157	0.0026
γAD	negative regulation of T cell differenti...	14	1.8600	0.0062	0.8928	0.0032
γAD	regulation of germinal center formation	6	0.8000	0.0037	0.9202	0.0037
γAD	negative regulation of transcription by ...	483	64.2600	0.0038	0.2385	0.0038
γAD	SMAD protein signal transduction	30	3.9900	0.0039	0.9834	0.0039
γAD	regulation of ossification	60	7.9800	0.0039	0.4022	0.0039
γAD	response to wounding	84	11.1700	0.0001	0.9342	0.0040
γAD	pharyngeal system development	13	1.7300	0.0040	0.5586	0.0040
γAD	positive regulation of cell morphogenesi...	26	3.4600	0.0046	0.9799	0.0046
γAD	regulation of pri-miRNA transcription by...	22	2.9300	0.0053	0.9924	0.0053
γAD	semaphorin-plexin signaling pathway	11	1.4600	0.0014	0.9980	0.0057
γAD	hemostasis	31	4.1200	0.0050	0.9927	0.0060
γAD	regulation of transforming growth factor...	32	4.2600	0.0065	0.9868	0.0065
γAD	negative regulation of DNA-binding trans...	59	7.8500	0.0083	0.2537	0.0083
γAD	negative regulation of inflammatory resp...	19	2.5300	0.0083	0.7222	0.0083
γAD	negative regulation of cytokine producti...	4	0.5300	0.0084	0.9089	0.0084
γAD	common-partner SMAD protein phosphorylat...	4	0.5300	0.0084	0.9670	0.0084
γAD	G0 to G1 transition	4	0.5300	0.0084	0.9997	0.0084
γAD	regulation of axon extension involved in...	4	0.5300	0.0084	0.9579	0.0084
γAD	negative chemotaxis	4	0.5300	0.0084	0.9789	0.0084
γAD	response to cholesterol	4	0.5300	0.0084	0.9983	0.0084
γAD	head morphogenesis	11	1.4600	0.0094	0.9987	0.0094
α	nucleotide-excision repair, DNA duplex u...	13	0.1500	0.0004	0.9857	0.0004
α	nucleotide-excision repair, preincision ...	13	0.1500	0.0004	0.9515	0.0004
α	nucleotide-excision repair, DNA incision...	13	0.1500	0.0004	0.9515	0.0004
α	nucleotide-excision repair, DNA incision...	15	0.1800	0.0006	0.6690	0.0006
α	nucleotide-excision repair, preincision ...	16	0.1900	0.0008	0.8050	0.0008
α	global genome nucleotide-excision repair	16	0.1900	0.0008	0.9698	0.0008
α	transcription-coupled nucleotide-excisio...	37	0.4400	0.0008	0.1411	0.0008
α	protein monoubiquitination	19	0.2300	0.0013	0.6811	0.0013
α	UV-damage excision repair	7	0.0800	0.0027	0.8011	0.0027
α	post-translational protein modification	52	0.6200	0.0029	0.0445	0.0029
α	trophectodermal cell differentiation	8	0.0900	0.0036	0.8097	0.0036
α	DNA damage response, detection of DNA da...	10	0.1200	0.0058	0.5788	0.0058
α	response to arsenic-containing substance	10	0.1200	0.0058	0.6074	0.0058

E

GO enrichment for mental disorders

Table showing the GO enrichment for the psychiatric disorders Autism Spectrum Disorder (ASD), Bipolar Disorder (BD), Major Depression Disorder (MDD) and Schizophrenia (SCZ). Healthy humans adults are used as controls and are represented as CTR.

TABLE E.1: GO enrichment for mental disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γSCZ	thalamus development	2	0.0200	0.0001	1.0000	0.0001
γSCZ	cornification	2	0.0200	0.0001	1.0000	0.0001
γSCZ	bundle of His cell-Purkinje myocyte adhe...	2	0.0200	0.0001	1.0000	0.0001
γSCZ	regulation of ventricular cardiac muscle...	2	0.0200	0.0001	1.0000	0.0001
γSCZ	intermediate filament organization	3	0.0300	0.0003	0.9101	0.0003
γSCZ	canonical Wnt signaling pathway	66	0.6900	0.0005	0.4603	0.0005
γSCZ	desmosome organization	4	0.0400	0.0006	0.6617	0.0006
γSCZ	positive regulation of oligodendrocyte d...	4	0.0400	0.0006	0.8507	0.0006
γSCZ	cellular response to indole-3-methanol	4	0.0400	0.0006	0.6492	0.0006
γSCZ	regulation of heart rate by cardiac cond...	4	0.0400	0.0006	0.7206	0.0006
γSCZ	extracellular matrix disassembly	5	0.0500	0.0010	0.9999	0.0010
γSCZ	cellular response to lithium ion	6	0.0600	0.0015	0.9096	0.0015
γSCZ	melanocyte differentiation	7	0.0700	0.0021	0.9282	0.0021
γSCZ	negative regulation of T cell proliferat...	9	0.0900	0.0036	0.8306	0.0036
γSCZ	positive regulation of protein import in...	9	0.0900	0.0036	0.1752	0.0036
γSCZ	positive regulation of neuroblast prolif...	10	0.1000	0.0045	0.4949	0.0045
γSCZ	neural crest cell migration	12	0.1300	0.0064	0.8645	0.0064
γSCZ	spinal cord motor neuron differentiation	12	0.1300	0.0064	0.3359	0.0064
γSCZ	positive regulation of transcription, DN...	696	7.2600	0.0038	0.0009	0.0071
γSCZ	positive regulation of DNA-binding trans...	77	0.8000	0.0074	0.2091	0.0074
γSCZ	positive regulation of mesenchymal cell ...	13	0.1400	0.0076	0.0749	0.0076
γSCZ	regulation of osteoclast differentiation	13	0.1400	0.0076	0.9944	0.0076
γSCZ	positive regulation of fibroblast prolif...	13	0.1400	0.0076	0.3790	0.0076
γSCZ	ovarian follicle development	14	0.1500	0.0088	0.9768	0.0088
γMDD	DNA damage response, signal transduction...	29	0.0800	0.0022	0.2115	0.0022
γMDD	retinoic acid biosynthetic process	1	0.0000	0.0026	1.0000	0.0026
γMDD	mRNA localization resulting in posttrans...	1	0.0000	0.0026	1.0000	0.0026
γMDD	positive regulation of collateral sprout...	1	0.0000	0.0026	1.0000	0.0026
γMDD	platelet alpha granule organization	1	0.0000	0.0026	1.0000	0.0026
γMDD	positive regulation of DNA damage respon...	1	0.0000	0.0026	1.0000	0.0026
γMDD	negative regulation of transcription inv...	2	0.0100	0.0052	0.9970	0.0052
γMDD	regulation of fat cell differentiation	45	0.1200	0.0053	0.2174	0.0053

TABLE E.1: GO enrichment for mental disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γMDD	lens fiber cell apoptotic process	3	0.0100	0.0078	0.5336	0.0078
γCTR.SCZ	positive regulation of SMAD protein sign...	8	0.0100	0.0069	0.2160	0.0069
γCTR.MDD	regulation of epithelial cell proliferat...	79	0.1400	0.0001	0.0127	0.0001
γCTR.MDD	pharyngeal system development	13	0.0200	0.0002	0.0495	0.0002
γCTR.MDD	metanephros development	32	0.0600	0.0000	0.8997	0.0003
γCTR.MDD	regulation of cell division	17	0.0300	0.0003	0.1893	0.0003
γCTR.MDD	negative regulation of ossification	25	0.0400	0.0007	0.6504	0.0007
γCTR.MDD	neural tube closure	29	0.0500	0.0009	0.2812	0.0009
γCTR.MDD	cell fate specification	31	0.0500	0.0010	0.6003	0.0010
γCTR.MDD	response to estradiol	34	0.0600	0.0012	0.9235	0.0012
γCTR.MDD	embryonic organ development	124	0.2200	0.0000	0.0356	0.0013
γCTR.MDD	embryonic limb morphogenesis	35	0.0600	0.0013	0.1816	0.0013
γCTR.MDD	digestive tract development	38	0.0700	0.0016	0.1961	0.0016
γCTR.MDD	response to mechanical stimulus	40	0.0700	0.0017	0.1947	0.0017
γCTR.MDD	negative regulation of neuroblast prolif...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	defense response to fungus, incompatible...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	response to chlorate	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	negative regulation of extracellular mat...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	positive regulation of fibroblast migrat...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	negative regulation of macrophage cytoki...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	neural plate axis specification	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	hyaluronan catabolic process	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	inductive cell-cell signaling	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	positive regulation of superoxide anion ...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	positive regulation of vascular permeabi...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	positive regulation of phosphatidylinosi...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	ossification involved in bone remodeling	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	active induction of host immune response...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	positive regulation of isotype switchinγ..	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	negative regulation of release of seques...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	olfactory bulb mitral cell layer develop...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	ureteric bud invasion	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	negative regulation of hyaluronan biosyn...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	positive regulation of NAD+ ADP-ribosylt...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	transforming growth factor beta receptor...	1	0.0000	0.0017	1.0000	0.0017
γCTR.MDD	response to ionizing radiation	44	0.0800	0.0021	0.5380	0.0021
γCTR.MDD	epithelial cell differentiation	144	0.2500	0.0000	0.0043	0.0026
γCTR.MDD	connective tissue replacement involved i...	2	0.0000	0.0035	0.6196	0.0035
γCTR.MDD	tolerance induction to self antigen	2	0.0000	0.0035	0.4341	0.0035
γCTR.MDD	olfactory nerve development	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	positive regulation of exit from mitosis	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	regulation of interleukin-23 production	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	negative regulation of interleukin-17 pr...	2	0.0000	0.0035	0.4341	0.0035
γCTR.MDD	positive regulation of interleukin-17 pr...	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	receptor catabolic process	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	positive regulation of protein dephospho...	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	regulation of blood vessel remodeling	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	positive regulation of secondary heart f...	2	0.0000	0.0035	0.9364	0.0035
γCTR.MDD	negative regulation of protein localizat...	2	0.0000	0.0035	1.0000	0.0035
γCTR.MDD	positive regulation of receptor clusteri...	2	0.0000	0.0035	0.5467	0.0035
γCTR.MDD	protein localization	275	0.4800	0.0062	0.0000	0.0047
γCTR.MDD	establishment of mitotic spindle orienta...	3	0.0100	0.0052	0.5318	0.0052
γCTR.MDD	common-partner SMAD protein phosphorylat...	3	0.0100	0.0052	1.0000	0.0052
γCTR.MDD	germ cell migration	3	0.0100	0.0052	0.2957	0.0052
γCTR.MDD	evasion or tolerance of host defenses by...	3	0.0100	0.0052	0.6164	0.0052
γCTR.MDD	positive regulation of collagen biosynth...	3	0.0100	0.0052	0.2985	0.0052
γCTR.MDD	negative regulation of multicellular orγ..	3	0.0100	0.0052	0.9366	0.0052
γCTR.MDD	positive regulation of odontogenesis	3	0.0100	0.0052	0.7620	0.0052
γCTR.MDD	positive regulation of regulatory T cell...	3	0.0100	0.0052	0.3865	0.0052
γCTR.MDD	lymph node development	3	0.0100	0.0052	0.6209	0.0052
γCTR.MDD	negative regulation of phagocytosis	3	0.0100	0.0052	1.0000	0.0052
γCTR.MDD	frontal suture morphogenesis	3	0.0100	0.0052	0.4808	0.0052

TABLE E.1: GO enrichment for mental disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γCTR.MDD	branch elongation involved in mammary gl...	3	0.0100	0.0052	0.5677	0.0052
γCTR.MDD	regulation of branching involved in mamm...	3	0.0100	0.0052	0.9770	0.0052
γCTR.MDD	otic vesicle morphogenesis	3	0.0100	0.0052	0.5613	0.0052
γCTR.MDD	positive regulation of mononuclear cell ...	3	0.0100	0.0052	0.9260	0.0052
γCTR.MDD	negative regulation of production of miR...	3	0.0100	0.0052	1.0000	0.0052
γCTR.MDD	positive regulation of production of miR...	3	0.0100	0.0052	0.7011	0.0052
γCTR.MDD	regulation of sodium ion transport	4	0.0100	0.0069	0.8322	0.0069
γCTR.MDD	histone dephosphorylation	4	0.0100	0.0069	0.6266	0.0069
γCTR.MDD	mammary gland branching involved in thel...	4	0.0100	0.0069	0.8540	0.0069
γCTR.MDD	positive regulation of extracellular mat...	4	0.0100	0.0069	0.8482	0.0069
γCTR.MDD	membrane protein intracellular domain pr...	5	0.0100	0.0087	0.9859	0.0087
γCTR.MDD	semicircular canal morphogenesis	5	0.0100	0.0087	1.0000	0.0087
γCTR.MDD	smoothened signaling pathway involved in...	5	0.0100	0.0087	0.0359	0.0087
γCTR.BD.MDD	heart valve morphogenesis	16	0.0200	0.0001	0.4015	0.0001
γCTR.BD.MDD	negative regulation of cell proliferatio...	179	0.2300	0.0005	0.0698	0.0005
γCTR.BD.MDD	marginal zone B cell differentiation	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	chondrocyte hypertrophy	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	male germ-line sex determination	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	intrahepatic bile duct development	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	epithelial cell proliferation involved i...	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	regulation of cell proliferation involve...	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	Harderian gland development	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	ureter urothelium development	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	negative regulation of G1/S transition o...	1	0.0000	0.0013	1.0000	0.0013
γCTR.BD.MDD	Notch signaling pathway	51	0.0700	0.0014	0.1055	0.0014
γCTR.BD.MDD	chondrocyte differentiation involved in ...	2	0.0000	0.0026	0.6218	0.0026
γCTR.BD.MDD	otic vesicle formation	2	0.0000	0.0026	0.8552	0.0026
γCTR.BD.MDD	astrocyte fate commitment	2	0.0000	0.0026	0.3766	0.0026
γCTR.BD.MDD	retinal rod cell differentiation	2	0.0000	0.0026	0.3766	0.0026
γCTR.BD.MDD	bronchus cartilage development	2	0.0000	0.0026	0.7990	0.0026
γCTR.BD.MDD	regulation of branching involved in lunγ..	2	0.0000	0.0026	1.0000	0.0026
γCTR.BD.MDD	lung smooth muscle development	2	0.0000	0.0026	0.7990	0.0026
γCTR.BD.MDD	cellular response to heparin	2	0.0000	0.0026	0.8813	0.0026
γCTR.BD.MDD	renal vesicle induction	2	0.0000	0.0026	0.3766	0.0026
γCTR.BD.MDD	ureter morphogenesis	2	0.0000	0.0026	0.3766	0.0026
γCTR.BD.MDD	cell cycle process	245	0.3200	0.0012	0.0007	0.0027
γCTR.BD.MDD	stem cell population maintenance	74	0.1000	0.0030	0.1284	0.0030
γCTR.BD.MDD	cartilage condensation	3	0.0000	0.0039	0.5896	0.0039
γCTR.BD.MDD	negative regulation of bone mineralizati...	3	0.0000	0.0039	0.8247	0.0039
γCTR.BD.MDD	positive regulation of Ras protein signa...	3	0.0000	0.0039	0.9588	0.0039
γCTR.BD.MDD	trachea cartilage development	3	0.0000	0.0039	0.6561	0.0039
γCTR.BD.MDD	intestinal epithelial structure maintena...	3	0.0000	0.0039	1.0000	0.0039
γCTR.BD.MDD	ureter smooth muscle cell differentiatio...	3	0.0000	0.0039	0.9972	0.0039
γCTR.BD.MDD	metanephric nephron tubule formation	3	0.0000	0.0039	0.5255	0.0039
γCTR.BD.MDD	positive regulation of mesenchymal stem ...	3	0.0000	0.0039	0.7046	0.0039
γCTR.BD.MDD	regulation of cell cycle	245	0.3200	0.0012	0.0046	0.0047
γCTR.BD.MDD	lacrimal gland development	4	0.0100	0.0052	0.0526	0.0052
γCTR.BD.MDD	limb bud formation	4	0.0100	0.0052	0.8144	0.0052
γCTR.BD.MDD	positive regulation of extracellular mat...	4	0.0100	0.0052	0.4878	0.0052
γCTR.BD.MDD	positive regulation of intrinsic apoptot...	4	0.0100	0.0052	0.7186	0.0052
γCTR.BD.MDD	positive regulation of cell proliferatio...	4	0.0100	0.0052	0.9702	0.0052
γCTR.BD.MDD	regulation of epithelial cell proliferat...	4	0.0100	0.0052	0.6940	0.0052
γCTR.BD.MDD	negative regulation of chondrocyte diffe...	5	0.0100	0.0065	0.2310	0.0065
γCTR.BD.MDD	cellular protein-containing complex loca...	5	0.0100	0.0065	0.7051	0.0065
γCTR.BD.MDD	negative regulation of photoreceptor cel...	5	0.0100	0.0065	0.5067	0.0065
γCTR.BD.MDD	Sertoli cell development	5	0.0100	0.0065	0.0088	0.0065
γCTR.BD.MDD	atrial septum morphogenesis	5	0.0100	0.0065	0.7898	0.0065
γCTR.BD.MDD	positive regulation of male gonad develo...	5	0.0100	0.0065	1.0000	0.0065
γCTR.BD.MDD	notochord development	6	0.0100	0.0078	0.1834	0.0078
γCTR.BD.MDD	negative regulation of mesenchymal cell ...	6	0.0100	0.0078	0.6975	0.0078
γCTR.BD	secretory columnal luminar epithelial ce...	3	0.0200	0.0002	0.9880	0.0002
γCTR.BD	negative regulation of inflammatory resp...	16	0.1300	0.0002	0.6464	0.0002

TABLE E.1: GO enrichment for mental disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γCTR.BD	regulation of I-kappaB kinase/NF-kappaB ...	45	0.3700	0.0004	0.6903	0.0004
γCTR.BD	immune response	190	1.5700	0.0001	0.0944	0.0006
γCTR.BD	cardiac left ventricle morphogenesis	5	0.0400	0.0006	0.9999	0.0006
γCTR.BD	positive regulation of JAK-STAT cascade	5	0.0400	0.0006	0.9999	0.0006
γCTR.BD	mitral valve morphogenesis	6	0.0500	0.0009	0.7180	0.0009
γCTR.BD	pulmonary valve morphogenesis	6	0.0500	0.0009	0.9911	0.0009
γCTR.BD	response to muramyl dipeptide	6	0.0500	0.0009	0.8380	0.0009
γCTR.BD	negative regulation of ossification	25	0.2100	0.0010	0.7495	0.0010
γCTR.BD	cilium assembly	26	0.2100	0.0011	0.1179	0.0011
γCTR.BD	regulation of epithelial to mesenchymal ...	27	0.2200	0.0012	0.5340	0.0012
γCTR.BD	vasculogenesis	28	0.2300	0.0014	0.9438	0.0014
γCTR.BD	response to molecule of bacterial origin	58	0.4800	0.0011	0.0964	0.0018
γCTR.BD	artery development	34	0.2800	0.0001	0.0545	0.0019
γCTR.BD	foregut morphogenesis	9	0.0700	0.0022	0.9936	0.0022
γCTR.BD	regulation of tumor necrosis factor-medi...	9	0.0700	0.0022	0.4999	0.0022
γCTR.BD	embryonic axis specification	11	0.0900	0.0034	0.9997	0.0034
γCTR.BD	positive regulation of cell-substrate ad...	11	0.0900	0.0034	0.9276	0.0034
γCTR.BD	stem cell division	12	0.1000	0.0041	0.6553	0.0041
γCTR.BD	negative regulation of BMP signaling pat...	13	0.1100	0.0048	0.7002	0.0048
γCTR.BD	negative regulation of Wnt signaling pat...	52	0.4300	0.0082	0.2149	0.0082
γCTR.BD	coronary vein morphogenesis	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	Notch signaling pathway involved in regu...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	negative regulation of SMAD protein comp...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	detection of wounding	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of transcription by ...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	negative regulation of toll-like recepto...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of toll-like recepto...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of toll-like recepto...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of toll-like recepto...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	positive regulation of translational ini...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	embryonic ectodermal digestive tract mor...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	induction of positive chemotaxis	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	actin crosslink formation	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	branching involved in open tracheal syst...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	right lung morphogenesis	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	venous endothelial cell differentiation	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	negative regulation of nucleotide-bindin...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of nucleotide-bindin...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	response to interleukin-9	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	response to interleukin-11	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	death-inducing signaling complex assembl...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	protein deubiquitination involved in ubi...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	glomerular capillary formation	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	tolerance induction to lipopolysaccharid...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	mesenchyme migration	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of osteoclast prolif...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	apoptotic process involved in embryonic ...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	positive regulation of aorta morphogenes...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	regulation of membrane repolarization du...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	positive regulation of ceramide biosynth...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of CD40 signaling pa...	1	0.0100	0.0082	0.9999	0.0082
γCTR.BD	negative regulation of endothelial cell ...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	positive regulation of integrin-mediated...	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	positive regulation of neuron migration	1	0.0100	0.0082	1.0000	0.0082
γCTR.BD	ventricular cardiac muscle tissue morpho...	18	0.1500	0.0091	0.3658	0.0091
γCTR.BD	positive regulation of neural precursor ...	18	0.1500	0.0091	0.3461	0.0091
γCTR.ASD.BD	negative regulation of transcription fro...	2	0.0000	0.0009	0.6285	0.0009
γCTR.ASD.BD	negative regulation of histone H4 acetyl...	3	0.0000	0.0013	0.1847	0.0013
γCTR.ASD.BD	white fat cell differentiation	7	0.0000	0.0030	0.7312	0.0030
γCTR.ASD.BD	positive regulation of histone deacetyla...	9	0.0000	0.0039	0.2260	0.0039
γCTR.ASD	negative regulation of histone methylati...	16	3.8400	0.0002	0.9200	0.0002
γCTR.ASD	positive regulation of histone methylati...	20	4.8000	0.0006	0.7300	0.0006

Table E.1: GO enrichment for mental disorders.

$\widetilde{\Phi}$	Term	Annotated	Expected	Significant	Classic	Weight
γCTR.ASD	nuclear-transcribed mRNA catabolic proce...	19	4.5600	0.0063	0.9200	0.0020
γCTR.ASD	positive regulation of chromatin silenci...	8	1.9200	0.0033	0.9500	0.0033
γCTR.ASD	cellular macromolecule metabolic process	2122	508.8400	0.0008	1.0000	0.0050
γCTR.ASD	cytoplasmic mRNA processing body assembl...	9	2.1600	0.0079	1.0000	0.0079
γCTR	type B pancreatic cell differentiation	10	5.6700	0.0034	0.6300	0.0034
γCTR	cell fate commitment involved in formati...	17	9.6400	0.0009	0.8400	0.0040
γCTR	T cell receptor signaling pathway	23	13.0500	0.0083	0.9800	0.0083
γBD	negative regulation of epidermis develop...	5	0.1200	0.0001	0.9985	0.0001
γBD	angiogenesis	86	2.0200	0.0000	0.4399	0.0002
γBD	negative regulation of gliogenesis	14	0.3300	0.0002	0.8436	0.0002
γBD	regulation of bone mineralization	16	0.3800	0.0004	0.6916	0.0004
γBD	regulation of astrocyte differentiation	7	0.1600	0.0004	0.9187	0.0004
γBD	cellular response to organic cyclic comp...	175	4.1100	0.0000	0.0986	0.0005
γBD	response to folic acid	2	0.0500	0.0005	0.9999	0.0005
γBD	positive regulation of cell proliferatio...	176	4.1300	0.0000	0.3901	0.0006
γBD	GABAergic neuron differentiation	8	0.1900	0.0006	0.7876	0.0006
γBD	embryonic forelimb morphogenesis	9	0.2100	0.0009	1.0000	0.0009
γBD	negative regulation of transcription, DN...	628	14.7300	0.0008	0.1213	0.0011
γBD	positive regulation of transmembrane rec...	35	0.8200	0.0011	0.3452	0.0011
γBD	negative regulation of cell proliferatio...	179	4.2000	0.0002	0.4332	0.0013
γBD	positive regulation of apoptotic process	128	3.0000	0.0001	0.3174	0.0013
γBD	neuroendocrine cell differentiation	4	0.0900	0.0000	0.9993	0.0015
γBD	adenohypophysis development	3	0.0700	0.0016	0.9993	0.0016
γBD	negative regulation of aldosterone biosy...	3	0.0700	0.0016	0.9581	0.0016
γBD	regulation of T-helper 2 cell differenti...	3	0.0700	0.0016	0.9988	0.0016
γBD	cardiac neural crest cell development in...	3	0.0700	0.0016	0.9992	0.0016
γBD	negative regulation of cortisol biosynth...	3	0.0700	0.0016	0.9581	0.0016
γBD	negative regulation of neuron differenti...	48	1.1300	0.0000	0.7668	0.0016
γBD	odontogenesis	31	0.7300	0.0001	0.9345	0.0018
γBD	cardiac epithelial to mesenchymal transi...	16	0.3800	0.0004	0.8274	0.0022
γBD	cerebral cortex neuron differentiation	12	0.2800	0.0023	0.8560	0.0023
γBD	regulation of dendrite development	25	0.5900	0.0024	0.3066	0.0024
γBD	negative regulation of programmed cell d...	188	4.4100	0.0001	0.1859	0.0027
γBD	interferon-gamma-mediated signaling path...	26	0.6100	0.0028	0.3293	0.0028
γBD	blood vessel endothelial cell migration	20	0.4700	0.0010	0.5463	0.0028
γBD	olfactory bulb development	13	0.3000	0.0030	0.6804	0.0030
γBD	negative regulation of DNA damage respon...	4	0.0900	0.0031	0.8967	0.0031
γBD	regulation of insulin-like growth factor...	4	0.0900	0.0031	0.9997	0.0031
γBD	cranial suture morphogenesis	4	0.0900	0.0031	0.5092	0.0031
γBD	negative regulation of DNA binding	27	0.6300	0.0032	0.9678	0.0032
γBD	positive regulation of cellular componen...	65	1.5200	0.0036	0.0065	0.0036
γBD	SMAD protein signal transduction	29	0.6800	0.0042	0.6781	0.0042
γBD	epithelial cell morphogenesis	5	0.1200	0.0052	0.9852	0.0052
γBD	branching involved in salivary gland mor...	5	0.1200	0.0052	0.6756	0.0052
γBD	visual perception	16	0.3800	0.0055	0.0996	0.0055
γBD	positive regulation of lymphocyte prolif...	16	0.3800	0.0055	0.9799	0.0055
γBD	muscle tissue development	126	2.9600	0.0005	0.2032	0.0063
γBD	response to epidermal growth factor	17	0.4000	0.0066	0.8715	0.0066
γBD	negative regulation of DNA-binding trans...	53	1.2400	0.0072	0.4290	0.0072
γBD	subpallium development	8	0.1900	0.0006	0.7572	0.0074
γBD	positive regulation of meiotic nuclear d...	6	0.1400	0.0076	0.9614	0.0076
γBD	cell migration involved in heart develop...	6	0.1400	0.0076	0.9979	0.0076
γBD	positive regulation of NF-kappaB transcr...	35	0.8200	0.0083	0.4983	0.0083
γBD	roof of mouth development	35	0.8200	0.0083	0.8741	0.0083
γASD	IRE1-mediated unfolded protein response	8	1.1100	0.0020	0.9860	0.0020
γASD	oocyte differentiation	12	1.6700	0.0030	0.9900	0.0030
γASD	cell-matrix adhesion	18	2.5000	0.0076	0.9960	0.0076
γASD	positive regulation of response to endop...	7	0.9700	0.0091	0.6300	0.0091
γASD	myelination in peripheral nervous system	4	0.5600	0.0096	0.7550	0.0096
γASD	apoptotic cell clearance	4	0.5600	0.0096	0.9530	0.0096
γASD	negative regulation of ubiquitin-depende...	4	0.5600	0.0096	0.9900	0.0096

F

GO enrichment for genes involved in the primate evolution

Table showing the GO enrichment for the evolution of Transcription Factors (TFs). Humans were compared to Chimpanzee (CMP) and Rhesus macaque (RH).

Table F.1: GO enrichment for genes involved in the primate evolution.

$\widetilde{\Phi}$	Term	Annotated	Significant	Classic	Weight
γCMP	translational elongation	11	3	0.00085	0.00085
γCMP	regulation of long-term synaptic potenti...	3	2	0.00097	0.00097
γCMP	positive regulation of DNA-binding trans...	89	6	0.00497	0.00497
γCTR	transcription initiation from RNA polyme...	144	132	0.00085	0.0035
γCTR	mRNA processing	163	147	0.00281	0.0072
γCTR	transcription-coupled nucleotide-excisio...	36	35	0.00754	0.0075
γCTR.RH	endothelial cell activation	6	3	0.0028	0.0028
γCTR.RH	DNA double-strand break processing invol...	2	2	0.0030	0.0030
γCTR.RH	central nervous system neuron axonogenes...	7	3	0.0048	0.0048
γCTR.RH	G2 DNA damage checkpoint	7	3	0.0048	0.0048
γCTR.RH	double-strand break repair via homologou...	23	5	0.0069	0.0069
γCTR.RH	negative regulation of phagocytosis	3	2	0.0086	0.0086
γCTR.RH	positive regulation of mesenchymal stem ...	3	2	0.0086	0.0086
γCTR.RH	transforming growth factor beta receptor...	55	8	0.0090	0.0090
γRH	tube closure	37	11	0.00054	0.00047
γRH	negative regulation of aldosterone biosy...	3	3	0.00093	0.00093
γRH	negative regulation of cortisol biosynth...	3	3	0.00093	0.00093
γRH	regulation of GTPase activity	56	13	0.00225	0.00225
γRH	positive regulation of epithelial cell p...	51	12	0.00293	0.00293
γRH	negative regulation of extrinsic apoptot...	27	8	0.00322	0.00322
γRH	desmosome organization	4	3	0.00345	0.00345
γRH	regulation of cell shape	12	5	0.00385	0.00385
γRH	response to hypoxia	84	16	0.00624	0.00624
γRH	regulation of fibroblast proliferation	31	9	0.00210	0.00642
γRH	regulation of heart rate	9	4	0.00761	0.00761
γRH	positive regulation of MAP kinase activi...	37	11	0.00054	0.00768
γRH	adult walking behavior	5	3	0.00799	0.00799
γRH	negative regulation of insulin secretion	5	3	0.00799	0.00799
γRH	angiogenesis	106	20	0.00260	0.00820
γRH	BMP signaling pathway involved in heart ...	2	2	0.00955	0.00955

Table F.1: GO enrichment for genes involved in the primate evolution.

$\widetilde{\Phi}$	Term	Annotated	Significant	Classic	Weight
γRH	substrate-dependent cell migration	2	2	0.00955	0.00955
γRH	olfactory nerve development	2	2	0.00955	0.00955
γRH	regulation of hippo signaling	2	2	0.00955	0.00955
γRH	regulation of odontogenesis of dentin-co...	2	2	0.00955	0.00955
γRH	maintenance of epithelial cell apical/ba...	2	2	0.00955	0.00955
γRH	musculoskeletal movement	2	2	0.00955	0.00955
γRH	sensory perception of taste	2	2	0.00955	0.00955
γRH	regulation of branching involved in sali...	2	2	0.00955	0.00955
γRH	cornification	2	2	0.00955	0.00955
γRH	mesenchymal cell proliferation involved ...	2	2	0.00955	0.00955
γRH	bundle of His cell-Purkinje myocyte adhe...	2	2	0.00955	0.00955
γRH	regulation of ventricular cardiac muscle...	2	2	0.00955	0.00955
γRH	cellular response to drug	81	15	0.01040	0.00959

Bibliography

[1] S. Jiang and A. Mortazavi, "Integrating ChIP-seq with other functional genomics data," *Briefings in Functional Genomics*, vol. 17, pp. 104–115, 3 2018.

[2] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-Resolution Profiling of Histone Methylations in the Human Genome," *Cell*, vol. 129, pp. 823–837, 5 2007.

[3] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, pp. 553–560, 8 2007.

[4] D. S. Johnson, A. Mortazav, R. M. Myers, and B. Wold, "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[5] E. Taub, Floyd, J. M. Deleo, and E. B. Thompson, "Sequential Comparative Hybridizations Analyzed by Computerized Image Processing Can Identify and Quantitate Regulated RNAs," *DNA*, vol. 2, pp. 309–327, 12 1983.

[6] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, 12 1998.

[7] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell*, vol. 2, pp. 65–73, 7 1998.

[8] D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature*, vol. 405, pp. 827–836, 6 2000.

[9] R. K. Niedenthal, L. Riles, M. Johnston, and J. H. Hegemann, "Green Fluorescent Protein as a Marker for Gene Expression and Subcellular Localization in Budding Yeast," *Yeast*, vol. 12, pp. 773–786, 6 1996.

[10] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nature Genetics*, vol. 32, pp. 490–495, 12 2002.

[11] D. Shalon, S. J. Smith, and P. O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.," *Genome Research*, vol. 6, pp. 639–645, 7 1996.

[12] T. Tang, N. François, A. Glatigny, N. Agier, M. H. Mucchielli, L. Aggerbeck, and H. Delacroix, "Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment," *Bioinformatics*, vol. 23, pp. 2686–2691, 10 2007.

[13] A. Sánchez and M. C. R. D. Villa, “A Tutorial Review of Microarray Data Analysis,” *Bioinformatics*, pp. 1–55, 2008.

[14] A. B. Chetverin and F. R. Kramer, “Oligonucleotide arrays: New concepts and possibilities,” *Bio/Technology*, vol. 12, pp. 1093–1099, 11 1994.

[15] W. S. Cleveland and S. J. Devlin, “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting,” *Journal of the American Statistical Association*, vol. 83, pp. 596–610, 9 1988.

[16] W. S. Cleveland, “LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression,” *The American Statistician*, vol. 35, p. 54, 2 1981.

[17] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, “Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.,” *Nucleic acids research*, vol. 30, p. e15, 2 2002.

[18] R. D. C. Team and R. R Development Core Team, “R: A Language and Environment for Statistical Computing,” 2016.

[19] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, “affy—analysis of Affymetrix GeneChip data at the probe level,” *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[20] B. S. Carvalho and R. A. Irizarry, “A framework for oligonucleotide microarray preprocessing,” *Bioinformatics*, vol. 26, pp. 2363–2367, 10 2010.

[21] P. Du, W. A. Kibbe, and S. M. Lin, “lumi: A pipeline for processing Illumina microarray,” *Bioinformatics*, vol. 24, no. 13, pp. 1547–1548, 2008.

[22] P. Du, X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis,” *BMC Bioinformatics*, vol. 11, 2010.

[23] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, “Model-based variance-stabilizing transformation for Illumina microarray data,” *Nucleic Acids Research*, vol. 36, no. 2, 2008.

[24] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, “The transcriptional landscape of the yeast genome defined by RNA sequencing,” *Science*, vol. 320, pp. 1344–1349, 6 2008.

[25] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bähler, “Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution,” *Nature*, vol. 453, pp. 1239–1243, 6 2008.

[26] Y. Chu and D. R. Corey, “RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation,” *Nucleic Acid Therapeutics*, vol. 22, pp. 271–274, 8 2012.

[27] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, p. 10, 5 2011.

[28] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, “Quake: Quality-aware detection and correction of sequencing errors,” *Genome Biology*, vol. 11, no. 11, p. R116, 2010.

[29] J. Falgueras, A. J. Lara, N. Fernández-Pozo, F. R. Cantón, G. Pérez-Trabado, and M. G. Claros, “SeqTrim: A high-throughput pipeline for pre-processing any type of sequence read,” *BMC Bioinformatics*, vol. 11, p. 38, 1 2010.

[30] T. Lassmann, Y. Hayashizaki, and C. O. Daub, "TagDust - A program to eliminate artifacts from next generation sequencing data," *Bioinformatics*, vol. 25, pp. 2839–2840, 11 2009.

[31] S. Andrews, "FASTQC: A quality control tool for high throughput sequence data.."

[32] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, 3 2012.

[33] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: Discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, pp. 1105–1111, 5 2009.

[34] S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller, "Fast mapping of short sequences with mismatches, insertions and deletions using index structures," *PLoS Computational Biology*, vol. 5, p. e1000502, 9 2009.

[35] S. Hoffmann, C. Otto, G. Doose, A. Tanzer, D. Langenberger, S. Christ, M. Kunz, L. M. Holdt, D. Teupser, J. Hackermüller, and P. F. Stadler, "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection," *Genome Biology*, vol. 15, p. R34, 2 2014.

[36] C. Otto, P. F. Stadler, and S. Hoffmann, "Lacking alignments? The next-generation sequencing mapper segemehl revisited," *Bioinformatics*, vol. 30, pp. 1837–1843, 7 2014.

[37] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, 1 2013.

[38] Julien Delafontaine, "rnacounter," 2015.

[39] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology*, vol. 34, pp. 525–527, 5 2016.

[40] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," *Nature communications*, vol. 5, p. 3231, 2014.

[41] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[42] Q. You, L. Zhang, X. Yi, K. Zhang, D. Yao, X. Zhang, Q. Wang, X. Zhao, Y. Ling, W. Xu, F. Li, and Z. Su, "Co-expression network analyses identify functional modules associated with development and stress response in Gossypium arboreum," *Scientific Reports*, vol. 6, p. 38436, 12 2016.

[43] J. Huang, S. Vendramin, L. Shi, and K. M. McGinnis, "Construction and Optimization of a Large Gene Coexpression Network in Maize Using RNA-Seq Data.," *Plant physiology*, vol. 175, pp. 568–583, 9 2017.

[44] S. D. Gamboa-Tuz, A. Pereira-Santana, J. A. Zamora-Briseno, E. Castano, F. Espadas-Gil, J. T. Ayala-Sumuano, M. A. Keb-Llanes, F. Sanchez-Teyer, and L. C. Rodríguez-Zapata, "Transcriptomics and co-expression networks reveal tissue-specific responses and regulatory hubs under mild and severe drought in papaya (Carica papaya L.)," *Scientific Reports*, vol. 8, p. 14539, 12 2018.

[45] J. M. Lu, Y. C. Chen, Z. X. Ao, J. Shen, C. P. Zeng, X. Lin, L. P. Peng, R. Zhou, X. F. Wang, C. Peng, H. M. Xiao, K. Zhang, and H. W. Deng, "System network analysis of genomics and transcriptomics data identified type 1 diabetes-associated pathway and genes," *Genes and Immunity*, p. 1, 9 2018.

[46] L.-C. C. Chang, S. Jamain, C.-W. W. Lin, D. Rujescu, G. C. Tseng, and E. Sibille, "A conserved BDNF, glutamate- and GABA-enriched gene module related to human depression identified by coexpression meta-analysis and DNA variant genome-wide association studies," *PLoS ONE*, vol. 9, no. 3, p. e90980, 2014.

[47] T. T. Le, J. Savitz, H. Suzuki, M. Misaki, T. K. Teague, B. C. White, J. H. Marino, G. Wiley, P. M. Gaffney, W. C. Drevets, B. A. McKinney, and J. Bodurka, "Identification and replication of RNA-Seq gene network modules associated with depression severity," *Translational Psychiatry*, vol. 8, p. 180, 12 2018.

[48] C. Gaiteri, Y. Ding, B. French, G. C. Tseng, and E. Sibille, "Beyond modules and hubs: The potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders," *Genes, Brain and Behavior*, vol. 13, pp. 13–24, 1 2014.

[49] P. Jia, X. Chen, A. H. Fanous, and Z. Zhao, "Convergent roles of de novo mutations and common variants in schizophrenia in tissue-specific and spatiotemporal co-expression network," *Translational Psychiatry*, vol. 8, p. 105, 12 2018.

[50] S. de Jong, M. P. Boks, T. F. Fuller, E. Strengman, E. Janson, C. G. de Kovel, A. P. Ori, N. Vi, F. Mulder, J. D. Blom, B. Glenthøj, C. D. Schubart, W. Cahn, R. S. Kahn, S. Horvath, and R. A. Ophoff, "A gene co-expression network in whole blood of Schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes," *PLoS ONE*, vol. 7, p. e39498, 6 2012.

[51] S. G. Potkin, F. Macciardi, G. Guffanti, J. H. Fallon, Q. Wang, J. A. Turner, A. Lakatos, M. F. Miles, A. Lander, M. P. Vawter, and X. Xie, "Identifying gene regulatory networks in schizophrenia," *NeuroImage*, vol. 53, no. 3, pp. 839–847, 2010.

[52] D. Gysi, T. d. M. Fragoso, V. Busskamp, E. Almaas, and K. Nowick., "Comparing multiple networks using the Co-expression Differential Network Analysis (CoDiNA)," tech. rep., 2 2018.

[53] N. Slavov and K. A. Dawson, "Correlation signature of the macroscopic states of the gene regulatory network in cancer," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 4079–4084, 3 2009.

[54] S. Berto, A. Perdomo-Sabogal, D. Gerighausen, J. Qin, and K. Nowick, "A Consensus network of gene regulatory factors in the human frontal lobe," *Frontiers in Genetics*, vol. 7, p. 31, 3 2016.

[55] K. Nowick, T. Gernat, E. Almaas, and L. Stubbs, "Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 22358–22363, 12 2009.

[56] M. C. Oldham, S. Horvath, and D. H. Geschwind, "Conservation and evolution of gene coexpression networks in human and chimpanzee brains," *Proceedings of the National Academy of Sciences*, vol. 103, no. 47, pp. 17973–17978, 2006.

[57] S. P. Ficklin, F. Luo, and F. A. Feltus, "The Association of Multiple Interacting Genes with Specific Phenotypes in Rice Using Gene Coexpression Networks," *Plant Physiology*, vol. 154, no. 1, pp. 13–24, 2010.

[58] L. K. Kutsche, D. M. Gysi, J. Fallmann, K. Lenk, R. Petri, A. Swiersy, S. D. Klapper, K. Pircs, S. Khattak, P. F. Stadler, J. Jakobsson, K. Nowick, and V. Busskamp, "Combined Experimental and System-Level Analyses Reveal the Complex Regulatory Network of miR-124 during Human Neurogenesis," *Cell Systems*, pp. 1–15, 10 2018.

[59] M. Hawrylycz, J. A. Miller, V. Menon, D. Feng, T. Dolbeare, A. L. Guillozet-Bongaarts, A. G. Jegga, B. J. Aronow, C. K. Lee, A. Bernard, M. F. Glasser, D. L. Dierker, J. Menche, A. Szafer, F. Collman, P. Grange, K. A. Berman, S. Mihalas, Z. Yao, L. Stewart, A. L. Barabási, J. Schulkin, J. Phillips, L. Ng, C. Dang, D. R Haynor, A. Jones, D. C. Van Essen, C. Koch, and E. Lein, "Canonical genetic signatures of the adult human brain," *Nature Neuroscience*, vol. 18, pp. 1832–1844, 12 2015.

[60] M. Kitsak, A. Sharma, J. Menche, E. Guney, S. D. Ghiassian, J. Loscalzo, and A. L. Barabási, "Tissue Specificity of Human Disease Module," *Scientific Reports*, vol. 6, 2016.

[61] A. Saha, Y. Kim, A. D. Gewirtz, B. Jo, C. Gao, I. C. McDowell, B. E. Engelhardt, and A. Battle, "Co-expression networks reveal the tissue-specific regulation of transcription and splicing," *Genome Research*, vol. 27, no. 11, pp. 1843–1858, 2017.

[62] D. M. Gysi, A. Voigt, T. d. M. Fragoso, E. Almaas, and K. Nowick, "wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool," *BMC Bioinformatics*, vol. 19, p. 392, 12 2018.

[63] A. L. Barabási, *Network science*, vol. 371. 2013.

[64] M. E. J. Newman, *Networks*. 2018.

[65] J. A. Bondy and U. S. R. Murty, *Graph Theory*. Springer, 2008.

[66] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005.

[67] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 3747–3752, 3 2003.

[68] A. Li and S. Horvath, "Network module detection: Affinity search technique with the multi-node topological overlap measure," *BMC Research Notes*, vol. 2, p. 142, 7 2009.

[69] A.-L. Barabási and Z. N. Z. N. Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, pp. 101–113, 2 2004.

[70] Q. Li, J. A. Lee, and D. L. Black, "Neuronal regulation of alternative pre-mRNA splicing," *Nature Reviews Neuroscience*, vol. 8, pp. 819–831, 11 2007.

[71] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo, "How to infer gene networks from expression profiles," *Molecular Systems Biology*, vol. 3, no. 1, p. 78, 2007.

[72] K. Dempsey, I. Thapa, C. Cortes, Z. Eriksen, D. K. Bastola, and H. Ali, "On mining biological signals using correlation networks," in *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*, pp. 327–334, IEEE, 2013.

[73] L. I. Furlong, "Human diseases through the lens of network biology," *Trends in Genetics*, vol. 29, no. 3, pp. 150–159, 2013.

[74] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Research*, vol. 18, pp. 644–652, 4 2008.

[75] A. Wagner, "The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes," *Molecular Biology and Evolution*, vol. 18, pp. 1283–1292, 7 2001.

[76] N. Tomar and R. K. De, "Comparing methods for metabolic network analysis and an application to metabolic engineering," *Gene*, vol. 521, 2013.

[77] A. Wagner and D. A. Fell, “The small world inside large metabolic networks.,” *Proceedings. Biological sciences*, vol. 268, pp. 1803–10, 9 2001.

[78] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, “Towards a proteome-scale map of the human protein–protein interaction network,” *Nature*, vol. 437, pp. 1173–1178, 10 2005.

[79] J. H. Fong, A. E. Keating, and M. Singh, “Predicting specificity in bZIP coiled-coil protein interactions.,” *Genome Biology*, vol. 5, no. 2, p. R11, 2004.

[80] E. Eisenberg and E. Y. Levanon, “Preferential Attachment in the Protein Network Evolution,” *Physical Review Letters*, vol. 91, p. 138701, 9 2003.

[81] G. D. Amoutzias, D. L. Robertson, and E. Bornberg-Bauer, “The evolution of protein interaction networks in regulatory proteins.,” *Comparative and functional genomics*, vol. 5, no. 1, pp. 79–84, 2004.

[82] A.-L. Barabási, “Emergence of {Scaling} in {Random} {Networks},” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[83] A. Wagner, “Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization,” *Proceedings of the National Academy of…*, vol. 91, no. 10, pp. 4387–4391, 1994.

[84] A. Wagner, “Does evolutionary plasticity evolve?,” *Evolution*, vol. 50, pp. 1008–1023, 6 1996.

[85] S. Light, P. Kraulis, and A. Elofsson, “Preferential attachment in the evolution of metabolic networks,” *BMC Genomics*, vol. 6, p. 159, 11 2005.

[86] J. D. Orth, I. Thiele, and B. O. Palsson, “What is flux balance analysis?,” *Nature biotechnology*, vol. 28, pp. 245–8, 3 2010.

[87] J. I. Castrillo, P. Pir, and S. G. Oliver, “Yeast Systems Biology: Towards a Systems Understanding of Regulation of Eukaryotic Networks in Complex Diseases and Biotechnology,” *Handbook of Systems Biology*, pp. 343–365, 1 2013.

[88] C. R. Haggart, J. A. Bartell, J. J. Saucerman, and J. A. Papin, “Whole-Genome Metabolic Network Reconstruction and Constraint-Based Modeling,” *Methods in Enzymology*, vol. 500, pp. 411–433, 1 2011.

[89] E. Simeonidis and N. D. Price, “Genome-scale modeling for metabolic engineering.,” *Journal of industrial microbiology & biotechnology*, vol. 42, pp. 327–38, 3 2015.

[90] I. Thiele and B. O. Palsson, “A protocol for generating a high-quality genome-scale metabolic reconstruction.,” *Nature protocols*, vol. 5, pp. 93–121, 1 2010.

[91] T. Kumelj, S. Sulheim, A. Wentzel, and E. Almaas, “Predicting strain engineering strategies using iKS1317: a genome-scale metabolic model of Streptomyces coelicolor,” *Biotechnology Journal*, p. 1800180, 12 2018.

[92] M. MacGillivray, A. Ko, E. Gruber, M. Sawyer, E. Almaas, and A. Holder, “Robust analysis of fluxes in genome-scale metabolic pathways,” *Scientific Reports*, vol. 7, p. 268, 12 2017.

[93] B. O. Palsson, *Systems Biology: Constraint-based Reconstruction and Analysis.* 2015.

[94] J. L. Payne, F. Khalid, and A. Wagner, "RNA-mediated gene regulation is less evolvable than transcriptional regulation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, pp. E3481–E3490, 4 2018.

[95] J. L. Payne and A. Wagner, "The Robustness and Evolvability of Transcription Factor Binding Sites," *Science*, vol. 343, pp. 875–877, 2014.

[96] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene-disease predictions," *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592, 2018.

[97] R. Dai, Y. Xia, C. Liu, and C. Chen, "csuWGCNA: a combination of signed and unsigned WGCNA to capture negative correlations," *bioRxiv*, p. 288225, 2018.

[98] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[99] K. Basso, C. Wiggins, A. A. Margolin, A. Califano, I. Nemenman, G. Stolovitzky, and R. Favera, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," in *BMC Bioinformatics*, vol. 7, p. S7, BioMed Central, 2006.

[100] A. Li and S. Horvath, "Network neighborhood analysis with the multi-node topological overlap measure," *Bioinformatics*, vol. 23, pp. 222–231, 1 2007.

[101] A. Voigt, K. Nowick, and E. Almaas, "A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma," *PLoS Computational Biology*, vol. 13, no. 9, p. e1005739, 2017.

[102] S. Berto and K. Nowick, "Species-specific changes in a primate transcription factor network provide insights into themolecular evolution of the primate prefrontal cortex," *Genome Biology and Evolution*, vol. 10, pp. 2023–2036, 8 2018.

[103] A. Barberán, S. T. Bates, E. O. Casamayor, and N. Fierer, "Using network analysis to explore co-occurrence patterns in soil microbial communities," *ISME Journal*, vol. 6, no. 2, pp. 343–351, 2012.

[104] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C. E. T. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, J. M. Rose, I. Hewson, A. Patel, F. Sun, D. A. Caron, and J. A. Fuhrman, "Marine bacterial, archaeal and protistan association networks reveal ecological linkages," *ISME Journal*, vol. 5, pp. 1414–1425, 9 2011.

[105] N. Gotelli and A. Ellison, "EcoSim," 2013.

[106] D. M. Griffith, J. A. Veech, and C. J. Marsh, "<b>cooccur</b> : Probabilistic Species Co-Occurrence Analysis in <i>R</i>," *Journal of Statistical Software*, vol. 69, no. Code Snippet 2, pp. 1–17, 2016.

[107] J. A. Veech, "A probabilistic model for analysing species co-occurrence," *Global Ecology and Biogeography*, vol. 22, pp. 252–260, 2 2013.

[108] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology*, vol. 14, pp. 283–291, 6 2004.

[109] M. J. Mason, G. Fan, K. Plath, Q. Zhou, and S. Horvath, "Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells," *BMC Genomics*, vol. 10, no. 1, p. 327, 2009.

[110] J. Peng, P. Wang, N. Zhou, and J. Zhu, “Partial correlation estimation by joint sparse regression models,” *Journal of the American Statistical Association*, vol. 104, pp. 735–746, 6 2009.

[111] P. Langfelder and S. Horvath, “Fast {R} Functions for Robust Correlations and Hierarchical Clustering,” *Journal of Statistical Software*, vol. 46, no. 11, pp. 1–17, 2012.

[112] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano, “Reverse engineering cellular networks,” *Nature Protocols*, vol. 1, no. 2, pp. 662–671, 2006.

[113] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, “Comparing statistical methods for constructing large scale gene networks,” *PLoS ONE*, vol. 7, no. 1, pp. 1–9, 2012.

[114] L. Kaufman and P. J. Rousseeuw, *Finding groups in data : an introduction to cluster analysis.* Wiley-Interscience, 1990.

[115] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, “Hierarchical organization of modularity in metabolic networks,” *Science*, vol. 297, pp. 1551–1555, 8 2002.

[116] X. Liu, D. Han, M. Somel, X. Jiang, H. Hu, P. Guijarro, N. Zhang, A. Mitchell, T. Halene, J. J. Ely, C. C. Sherwood, P. R. Hof, Z. Qiu, S. Pääbo, S. Akbarian, and P. Khaitovich, “Disruption of an Evolutionarily Novel Synaptic Expression Pattern in Autism,” *PLoS Biology*, vol. 14, pp. 1–23, 9 2016.

[117] K. Murphy and S. Mian, “Modelling Gene Expression Data using Dynamic Bayesian Networks,” tech. rep., Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

[118] B. Zhang and S. Horvath, “A General Framework for Weighted Gene Co-Expression Network Analysis,” *Statistical Applications in Genetics and Molecular Biology*, vol. 4, p. Article17, 1 2005.

[119] M. R. J. Carlson, B. Zhang, Z. Fang, P. S. Mischel, S. Horvath, and S. F. Nelson, “Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks,” *BMC Genomics*, vol. 7, no. 1, p. 40, 2006.

[120] D. G. Altman, *Practical statistics for medical research.* CRC press, 1990.

[121] E. McCrum-Gardner, “Which is the correct statistical test to use?,” *British Journal of Oral and Maxillofacial Surgery*, vol. 46, no. 1, pp. 38–41, 2008.

[122] M. M. Mukaka, “A guide to appropriate use of correlation coefficient in medical research,” *Malawi Medical Journal*, vol. 24, no. 3, pp. 69–71, 2012.

[123] A. J. Bishara and J. B. Hittner, “Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches.,” *Psychological Methods*, vol. 17, no. 3, p. 399, 2012.

[124] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap.* CRC press, 1994.

[125] A. Valera, M. Pelegrin, G. Asins, C. Fillat, J. Sabater, A. Pujol, F. G. Hegardt, and F. Bosch, “Overexpression of mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase in transgenic mice causes hepatic hyperketogenesis,” *Journal of Biological Chemistry*, vol. 269, pp. 6267–6270, 6 1994.

[126] S. S. Fong, A. R. Joyce, and B. O. Palsson, “Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states,” *Genome Research*, vol. 15, pp. 1365–1372, 9 2005.

[127] S. S. Fong, A. Nanchen, B. O. Palsson, and U. Sauer, “Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes,” *Journal of Biological Chemistry*, vol. 281, pp. 8024–8033, 3 2006.

[128] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, “Integrating high-throughput and computational data elucidates bacterial networks,” *Nature*, vol. 429, pp. 92–96, 5 2004.

[129] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. Del Moral-Chavez, F. Rinaldi, and J. Collado-Vides, “RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond,” *Nucleic Acids Research*, vol. 44, pp. D133–D143, 1 2016.

[130] S. Horvath, B. Zhang, M. Carlson, K. V. Lu, S. Zhu, R. M. Felciano, M. F. Laurance, W. Zhao, S. Qi, Z. Chen, Y. Lee, A. C. Scheck, L. M. Liau, H. Wu, D. H. Geschwind, P. G. Febbo, H. I. Kornblum, T. F. Cloughesy, S. F. Nelson, and P. S. Mischel, “Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17402–17407, 2006.

[131] P. E. Meyer, F. Lafitte, and G. Bontempi, “Minet: an open source R/Bioconductor package for mutual information based network inference,” *BMC Bioinformatics*, vol. 9, p. 461, 2008.

[132] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[133] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[134] R. A. Fisher, “Statistical Methods for Research Workers.,” *The American Mathematical Monthly*, vol. 37, no. 10, p. 547, 1930.

[135] G. Csárdi, T. T. Nepusz, G. Csardi, and T. T. Nepusz, “The igraph software package for complex network research,” *Journal of Computer Applications*, vol. Complex Sy, no. 5, p. 9, 2014.

[136] B. K. Kuntal, A. Dutta, and S. S. Mande, “CompNet: a GUI based tool for comparison of multiple biological interaction networks,” *BMC Bioinformatics*, vol. 17, p. 185, 4 2016.

[137] M. Chen and R. Hofestädt, “PathAligner: metabolic pathway retrieval and alignment,” *Appl Bioinformatics*, vol. 3, 2004.

[138] Z. Liang, M. Xu, M. Teng, and L. Niu, “NetAlign: a web-based tool for comparison of protein interaction networks,” *Bioinformatics*, vol. 22, 2006.

[139] M. Watson, “CoXpress: Differential co-expression in gene expression data,” *BMC Bioinformatics*, vol. 7, no. 1, p. 509, 2006.

[140] Y. Tian, R. C. McEachin, C. Santos, D. J. States, and J. M. Patel, “SAGA: a subgraph matching tool for biological graphs,” *Bioinformatics*, vol. 23, 2007.

[141] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Mäkelä, and S. Hautaniemi, “Integrated network analysis platform for protein-protein interactions,” *Nature Methods*, vol. 6, no. 1, pp. 75–77, 2009.

[142] B. Dost, T. Shlomi, N. Gupta, E. Ruppin, V. Bafna, and R. Sharan, "QNet: A Tool for Querying Protein Interaction Networks," *Journal of Computational Biology*, vol. 15, no. 7, pp. 913–925, 2008.

[143] P. D. Lena, G. Wu, P. L. Martelli, R. Casadio, and C. Nardini, "MIMO: an efficient tool for molecular interaction maps overlap," *BMC Bioinformatics*, vol. 14, 2013.

[144] J. I. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. Walhout, "Using networks to measure similarity between genes: Association index selection," *Nature Methods*, vol. 10, no. 12, pp. 1169–1176, 2013.

[145] A. Fukushima, "DiffCorr: An R package to analyze and visualize differential correlations in biological networks," *Gene*, vol. 518, no. 1, pp. 209–214, 2013.

[146] Q. Yang, B. Guo, H. Sun, J. Zhang, S. Liu, S. Hexige, X. Yu, and X. Wang, "Identification of the key genes implicated in the transformation of OLP to OSCC using RNA-sequencing," *Oncology Reports*, vol. 37, no. 4, pp. 2355–2365, 2017.

[147] P. Wang, L. Gao, Y. Hu, and F. Li, "Feature related multi-view nonnegative matrix factorization for identifying conserved functional modules in multiple biological networks," *BMC Bioinformatics*, vol. 19, p. 394, 12 2018.

[148] Y. Lichtblau, K. Zimmermann, B. Haldemann, D. Lenze, M. Hummel, and U. Leser, "Comparative assessment of differential network analysis methods," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 837–850, 2017.

[149] A. Fukushima and K. Nishida, "Using the DiffCorr Package to Analyze and Visualize Differential Correlations in Biological Networks," *Computational Network Analysis with R: Applications in Biology, Medicine and Chemistry*, vol. 518, no. 1, pp. 1–34, 2016.

[150] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3733 LNCS, no. 2, pp. 284–293, 2005.

[151] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 172–188, 2008.

[152] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 16110, 2006.

[153] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 26113, 2004.

[154] V. A. Traag and J. Bruggeman, "Community detection in networks with positive and negative links," *Physical Review E*, vol. 80, no. 3, p. 36115, 2009.

[155] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[156] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[157] A. Clauset, M. E. J. Newman, C. Moore, N. Clauset A, and M. C. MEJ, "Finding community structure in very large networks.," *Physical Review E.*, vol. 70, no. 6, p. 66111, 2004.

[158] M. Rosvall and C. T. Bergstrom, “Maps of information flow reveal community structure in complex networks,” *arXiv preprint physics.soc-ph/0707.0609*, 2007.

[159] M. Rosvall, D. Axelsson, and C. T. Bergstrom, “The map equation,” *European Physical Journal: Special Topics*, vol. 178, no. 1, pp. 13–23, 2009.

[160] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[161] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” *Physical review E*, vol. 76, no. 3, p. 36106, 2007.

[162] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Physical review E*, vol. 74, no. 3, p. 36104, 2006.

[163] L. Sun, A. M. Hui, Q. Su, A. Vortmeyer, Y. Kotliarov, S. Pastorino, A. Passaniti, J. Menon, J. Walling, R. Bailey, M. Rosenblum, T. Mikkelsen, and H. A. Fine, “Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain,” *Cancer Cell*, vol. 9, no. 4, pp. 287–300, 2006.

[164] J. F. Fontaine and M. A. Andrade-Navarro, “Gene Set to Diseases (GS2D): disease enrichment analysis on human gene sets with literature data,” *Genomics and Computational Biology*, vol. 2, p. 33, 10 2016.

[165] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes.,” *Nucleic acids research*, vol. 28, pp. 27–30, 1 2000.

[166] D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. DIgles, F. Ehrhart, P. Giesbertz, M. Kalafati, M. Martens, R. Miller, K. Nishida, L. Rieswijk, A. Waagmeester, L. M. Eijssen, C. T. Evelo, A. R. Pico, and E. L. Willighagen, “WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research,” *Nucleic acids research*, vol. 46, pp. D661–D667, 1 2018.

[167] T. E. P. ENCODE Project Consortium, “The ENCODE (ENCyclopedia Of DNA Elements) Project.,” *Science*, vol. 306, pp. 636–40, 10 2004.

[168] X. Zhou and Z. Su, “EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species,” *BMC Genomics*, vol. 8, p. 246, 7 2007.

[169] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, “GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists,” *BMC Bioinformatics*, vol. 10, p. 48, 12 2009.

[170] A. Alexa and J. Rahnenfuhrer, “topGO: enrichment analysis for gene ontology,” 2010.

[171] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15545–50, 10 2005.

[172] J. Piñero, A. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, “DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants,” *Nucleic Acids Research*, vol. 45, pp. D833–D839, 1 2017.

[173] A. Gutierrez-Sacristan, S. Grosdidier, O. Valverde, M. Torrens, A. Bravo, J. Piñero, F. Sanz, and L. I. Furlong, “PsyGeNET: A knowledge platform on psychiatric disorders and their genes,” *Bioinformatics*, vol. 31, pp. 3075–3077, 9 2015.

[174] G. Babbi, P. L. Martelli, G. Profiti, S. Bovo, C. Savojardo, and R. Casadio, “eDGAR: A database of disease-gene associations with annotated relationships among genes,” *BMC Genomics*, vol. 18, no. Suppl 5, p. 554, 2017.

[175] S. Banerjee-Basu and A. Packer, “SFARI Gene: an evolving database for the autism research community,” *Disease Models & Mechanisms*, vol. 3, pp. 133–135, 3 2010.

[176] G. Barnby, A. Abbott, N. Sykes, A. Morris, D. E. Weeks, R. Mott, J. Lamb, A. J. Bailey, and A. P. Monaco, “Candidate-Gene Screening and Association Analysis at the Autism-Susceptibility Locus on Chromosome 16p: Evidence of Association at GRIN2A and ABAT,” *The American Journal of Human Genetics*, vol. 76, pp. 950–966, 6 2005.

[177] S. N. Basu, R. Kollu, and S. Banerjee-Basu, “AutDB: A gene reference resource for autism research,” *Nucleic Acids Research*, vol. 37, pp. D832–D836, 1 2009.

[178] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, “Transcriptomic analysis of autistic brain reveals convergent molecular pathology,” *Nature*, vol. 474, pp. 380–386, 6 2011.

[179] N. C. Allen, S. Bagade, M. B. McQueen, J. P. Ioannidis, F. K. Kavvoura, M. J. Khoury, R. E. Tanzi, and L. Bertram, “Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: The SzGene database,” *Nature Genetics*, vol. 40, pp. 827–834, 7 2008.

[180] L. Bertram, “Neurobiology of Dementia,” in *International Review of Neurobiology*, vol. 84, ch. Alzheimer’, pp. 167–184, Elsevier, 2009.

[181] P. Jia, J. Sun, A. Y. Guo, and Z. Zhao, “SZGR: A comprehensive schizophrenia gene resource,” *Molecular Psychiatry*, vol. 15, pp. 453–462, 5 2010.

[182] C. M. Lill, J. T. Roehr, M. B. McQueen, F. K. Kavvoura, S. Bagade, B. M. M. Schjeide, L. M. Schjeide, E. Meissner, U. Zauft, N. C. Allen, T. Liu, M. Schilling, K. J. Anderson, G. Beecham, D. Berg, J. M. Biernacka, A. Brice, A. L. DeStefano, C. B. Do, N. Eriksson, S. A. Factor, M. J. Farrer, T. Foroud, T. Gasser, T. Hamza, J. A. Hardy, P. Heutink, E. M. Hill-Burns, C. Klein, J. C. Latourelle, D. M. Maraganore, E. R. Martin, M. Martinez, R. H. Myers, M. A. Nalls, N. Pankratz, H. Payami, W. Satake, W. K. Scott, M. Sharma, A. B. Singleton, K. Stefansson, T. Toda, J. Y. Tung, J. Vance, N. W. Wood, C. P. Zabetian, P. Young, R. E. Tanzi, M. J. Khoury, F. Zipp, H. Lehrach, J. P. Ioannidis, and L. Bertram, “Comprehensive research synopsis and systematic meta-analyses in Parkinson’s disease genetics: The PDgene database,” *PLoS Genetics*, vol. 8, p. e1002548, 3 2012.

[183] J. K. Inlow and L. L. Restifo, “Molecular and Comparative Genetics of Mental Retardation,” *Genetics*, vol. 166, pp. 835–881, 2 2004.

[184] H. H. Ropers, “Genetics of intellectual disability,” *Current Opinion in Genetics and Development*, vol. 18, pp. 241–250, 6 2008.

[185] J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. Chi, D. D. Licatalosi, J. D. Richter, and R. B. Darnell, “FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism,” *Cell*, vol. 146, pp. 247–261, 7 2011.

[186] H. van Bokhoven, “Genetic and Epigenetic Networks in Intellectual Disabilities,” *Annual Review of Genetics*, vol. 45, pp. 81–104, 12 2011.

[187] H. A. Lubs, R. E. Stevenson, and C. E. Schwartz, “Fragile X and X-linked intellectual disability: Four decades of discovery,” *American Journal of Human Genetics*, vol. 90, pp. 579–590, 4 2012.

[188] S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. De Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julià, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kähler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lönnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Müller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietiläinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. C. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Söderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jönsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nöthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, and M. C. O'Donovan, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, pp. 421–427, 7 2014.

[189] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 12 2005.

[190] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt," *Nature Protocols*, vol. 4, pp. 1184–1191, 8 2009.

[191] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber, "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis," *Bioinformatics*, vol. 21, pp. 3439–3440, 8 2005.

[192] M. Akerblom, R. Petri, R. Sachdeva, T. Klussendorf, B. Mattsson, B. Gentner, and J. Jakobsson, "microRNA-125 distinguishes developmentally generated and adult-born olfactory bulb interneurons," *Development*, vol. 141, no. 7, pp. 1580–1588, 2014.

[193] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl, "Identification of Tissue-Specific MicroRNAs from Mouse," *Current Biology*, vol. 12, pp. 735–739, 4 2002.

[194] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl, "A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing," *Cell*, vol. 129, pp. 1401–1414, 6 2007.

[195] J. Silber, D. A. Lim, C. Petritsch, A. I. Persson, A. K. Maunakea, M. Yu, S. R. Vandenberg, D. G. Ginzinger, C. D. James, J. F. Costello, G. Bergers, W. A. Weiss, A. Alvarez-Buylla, and J. G. Hodgson, "miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells," *BMC medicine*, vol. 6, p. 14, 6 2008.

[196] K. Taniguchi, N. Sugito, M. Kumazaki, H. Shinohara, N. Yamada, N. Matsuhashi, M. Futamura, Y. Ito, Y. Otsuki, K. Yoshida, K. Uchiyama, and Y. Akao, "Positive feedback of DDX6/c-Myc/PTB1 regulated by miR-124 contributes to maintenance of the Warburg effect in colon cancer cells," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1852, pp. 1971–1980, 9 2015.

[197] R. Dutta, A. M. Chomyk, A. Chang, M. V. Ribaudo, S. A. Deckard, M. K. Doud, D. D. Edberg, B. Bai, M. Li, S. E. Baranzini, R. J. Fox, S. M. Staugaitis, W. B. Macklin, and B. D. Trapp, "Hippocampal demyelination and memory dysfunction are associated with increased levels of the neuronal microRNA miR-124 and reduced AMPA receptors," *Annals of neurology*, vol. 73, pp. 637–645, 5 2013.

[198] Q. Hou, H. Ruan, J. Gilbert, G. Wang, Q. Ma, W.-D. Yao, and H.-Y. Man, "MicroRNA miR124 is required for the expression of homeostatic synaptic plasticity," *Nature Communications*, vol. 6, p. ncomms10045, 12 2015.

[199] P. Rajasethupathy, F. Fiumara, R. Sheridan, D. Betel, S. V. Puthanveettil, J. J. Russo, C. Sander, T. Tuschl, and E. Kandel, "Characterization of small RNAs in Aplysia reveals a role for miR-124 in constraining synaptic plasticity through CREB," *Neuron*, vol. 63, pp. 803–817, 9 2009.

[200] M. Fang, J. Wang, X. Zhang, Y. Geng, Z. Hu, J. A. Rudd, S. Ling, W. Chen, and S. Han, "The miR-124 regulates the expression of BACE1/β-secretase correlated with cell death in Alzheimer's disease," *Toxicology Letters*, vol. 209, pp. 94–105, 2 2012.

[201] W. J. Lukiw, "Micro-rna speciation in fetal, adult and Alzheimer's disease hippocampus," *Neuroreport*, vol. 18, pp. 297–300, 2 2007.

[202] E. Gascon, K. Lynch, H. Ruan, S. Almeida, J. M. Verheyden, W. W. Seeley, D. W. Dickson, L. Petrucelli, D. Sun, J. Jiao, H. Zhou, M. Jakovcevski, S. Akbarian, W.-D. Yao, and F.-B. Gao, "Alterations in microRNA-124 and AMPA receptors contribute to social behavioral deficits in frontotemporal dementia," *Nature Medicine*, vol. 20, no. 12, pp. 1444–1451, 2014.

[203] M. Akerblom, R. Sachdeva, I. Barde, S. Verp, B. Gentner, D. Trono, J. Jakobsson, M. Åkerblom, R. Sachdeva, I. Barde, S. Verp, B. Gentner, D. Trono, and J. Jakobsson, "MicroRNA-124 Is a Subventricular Zone Neuronal Fate Determinant," *Journal of Neuroscience*, vol. 32, pp. 8879–8889, 6 2012.

[204] X. Cao, S. L. Pfaff, and F. H. Gage, “A functional study of miR-124 in the developing neural tube,” *Genes & Development*, vol. 21, pp. 531–536, 3 2007.

[205] L.-C. Cheng, E. Pastrana, M. Tavazoie, and F. Doetsch, “miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche,” *Nature Neuroscience*, vol. 12, pp. 399–408, 4 2009.

[206] C. Conaco, S. Otto, J.-J. Han, and G. Mandel, “Reciprocal actions of REST and a microRNA promote neuronal identity,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 2422–2427, 2 2006.

[207] E. V. Makeyev, J. Zhang, M. A. Carrasco, and T. Maniatis, “The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing,” *Molecular Cell*, vol. 27, pp. 435–448, 8 2007.

[208] J. Visvanathan, S. Lee, B. Lee, J. W. Lee, and S.-K. Lee, “The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development,” *Genes & Development*, vol. 21, pp. 744–749, 4 2007.

[209] A. S. Yoo, A. X. Sun, L. Li, A. Shcheglovitov, T. Portmann, Y. Li, C. Lee-Messer, R. E. Dolmetsch, R. W. Tsien, and G. R. Crabtree, “MicroRNA-mediated conversion of human fibroblasts to neurons,” *Nature*, vol. 476, pp. 228–231, 8 2011.

[210] A. S. Yoo, B. T. Staahl, L. Chen, and G. R. Crabtree, “MicroRNA-mediated switching of chromatin-remodelling complexes in neural development,” *Nature*, vol. 460, pp. 642–646, 7 2009.

[211] K. Franke, W. Otto, S. Johannes, J. Baumgart, R. Nitsch, and S. Schumacher, “miR-124-regulated RhoG reduces neuronal process complexity via ELMO/Dock180/Rac1 and Cdc42 signalling,” *The EMBO Journal*, vol. 31, pp. 2908–2921, 6 2012.

[212] X. Gu, S. Meng, S. Liu, C. Jia, Y. Fang, S. Li, C. Fu, Q. Song, L. Lin, and X. Wang, “miR-124 Represses ROCK1 Expression to Promote Neurite Elongation Through Activation of the PI3K/Akt Signal Pathway,” *Journal of Molecular Neuroscience*, vol. 52, pp. 156–165, 12 2013.

[213] G. Li and S. Ling, “MiR-124 Promotes Newborn Olfactory Bulb Neuron Dendritic Morphogenesis and Spine Density,” *Journal of Molecular Neuroscience*, pp. 1–10, 12 2016.

[214] R. Sanuki, A. Onishi, C. Koike, R. Muramatsu, S. Watanabe, Y. Muranishi, S. Irie, S. Uneo, T. Koyasu, R. Matsui, Y. Chérasse, Y. Urade, D. Watanabe, M. Kondo, T. Yamashita, and T. Furukawa, “miR-124a is required for hippocampal axogenesis and retinal cone survival through Lhx2 suppression,” *Nature Neuroscience*, vol. 14, pp. 1125–1134, 9 2011.

[215] Q. Xue, C. Yu, Y. Wang, L. Liu, K. Zhang, C. Fang, F. Liu, G. Bian, B. Song, A. Yang, G. Ju, and J. Wang, “miR-9 and miR-124 synergistically affect regulation of dendritic branching via the AKT/GSK3β pathway by targeting Rap2a,” *Scientific Reports*, vol. 6, p. 26781, 5 2016.

[216] J.-Y. Yu, K.-H. Chung, M. Deo, R. C. Thompson, and D. L. Turner, “MicroRNA miR-124 Regulates Neurite Outgrowth during Neuronal Differentiation,” *Experimental cell research*, vol. 314, pp. 2618–2633, 8 2008.

[217] M.-L. Baudet, K. H. Zivraj, C. Abreu-Goodger, A. Muldal, J. Armisen, C. Blenkiron, L. D. Goldstein, E. A. Miska, and C. E. Holt, “miR-124 acts through CoREST to control onset of Sema3A sensitivity in navigating retinal growth cones,” *Nature Neuroscience*, vol. 15, pp. 29–38, 1 2012.

[218] X. Gu, A. Li, S. Liu, L. Lin, S. Xu, P. Zhang, S. Li, X. Li, B. Tian, X. Zhu, and X. Wang, “MicroRNA124 regulated neurite elongation by targeting OSBP,” *Molecular Neurobiology*, pp. 1–9, 11 2015.

[219] X. Gu, C. Fu, L. Lin, S. Liu, X. Su, A. Li, Q. Wu, C. Jia, P. Zhang, L. Chen, X. Zhu, and X. Wang, “miR-124 and miR-9 mediated downregulation of HDAC5 promotes neurite development through activating MEF2C-GPM6A pathway,” *Journal of Cellular Physiology*, pp. n/a–n/a, 3 2017.

[220] F. Higuchi, S. Uchida, H. Yamagata, N. Abe-Higuchi, T. Hobara, K. Hara, A. Kobayashi, T. Shintaku, Y. Itoh, T. Suzuki, and Y. Watanabe, “Hippocampal microRNA-124 enhances chronic stress resilience in mice,” *The Journal of Neuroscience*, vol. 36, pp. 7253–7267, 7 2016.

[221] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson, “Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs,” *Nature*, vol. 433, pp. 769–773, 2 2005.

[222] K. Liu, Y. Liu, W. Mo, R. Qiu, X. Wang, J. Y. Wu, and R. He, “MiR-124 regulates early neurogenesis in the optic vesicle and forebrain, targeting NeuroD1,” *Nucleic Acids Research*, vol. 39, pp. 2869–2879, 4 2011.

[223] J. Malmevik, R. Petri, T. Klussendorf, P. Knauff, M. Åkerblom, J. Johansson, S. Soneji, and J. Jakobsson, “Identification of the miRNA targetome in hippocampal neurons using RIP-seq,” *Scientific Reports*, vol. 5, 2015.

[224] M. C. T. Santos, A. N. Tegge, B. R. Correa, S. Mahesula, L. Q. Kohnke, M. Qiao, M. A. R. Ferreira, E. Kokovay, and L. O. F. Penalva, “MiR-124, -128, and -137 orchestrate neural differentiation by acting on overlapping gene sets containing a highly connected transcription factor network,” *Stem Cells (Dayton, Ohio)*, 9 2015.

[225] R. Weng and S. M. Cohen, “Drosophila miR-124 regulates neuroblast proliferation through its target anachronism,” *Development*, vol. 139, pp. 1427–1434, 4 2012.

[226] P. J. Yaworsky and C. Kappen, “Heterogeneity of neural progenitor cells revealed by enhancers in the nestin gene,” *Developmental Biology*, vol. 205, pp. 309–321, 1 1999.

[227] S. W. Chi, G. J. Hannon, and R. B. Darnell, “An alternative mode of microRNA target recognition,” *Nature Structural & Molecular Biology*, vol. 19, p. 321, 3 2012.

[228] M. J. Moore, T. K. H. Scheel, J. M. Luna, C. Y. Park, J. J. Fak, E. Nishiuchi, C. M. Rice, and R. B. Darnell, “miRNA–target chimeras reveal miRNA 3’-end pairing as a major determinant of Argonaute target specificity,” *Nature Communications*, vol. 6, p. 8864, 11 2015.

[229] V. Busskamp, N. E. Lewis, P. Guye, A. H. Ng, S. L. Shipman, S. M. Byrne, N. E. Sanjana, J. Murn, Y. Li, S. Li, M. Stadler, R. Weiss, and G. M. Church, “Rapid neurogenesis through transcriptional activation in human stem cells,” *Molecular Systems Biology*, vol. 10, no. 11, 2014.

[230] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, “Multiplex genome engineering using CRISPR/Cas systems,” *Science (New York, N.Y.)*, vol. 339, pp. 819–823, 2 2013.

[231] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church, “RNA-guided human genome engineering via Cas9,” *Science*, vol. 339, pp. 823–826, 2 2013.

[232] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, and F. Zhang, “Genome engineering using the CRISPR-Cas9 system,” *Nature Protocols*, vol. 8, pp. 2281–2308, 11 2013.

[233] R. Petri, K. Pircs, M. E. Jönsson, M. Åkerblom, P. L. Brattås, T. Klussendorf, and J. Jakobsson, “let-7 regulates radial migration of new-born neurons through positive regulation of autophagy,” *The EMBO Journal*, p. e201695235, 3 2017.

[234] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "GENCODE: The reference human genome annotation for The ENCODE Project," *Genome Research*, vol. 22, pp. 1760–1774, 9 2012.

[235] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, 2014.

[236] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, *SOM_PAK: The Self-Organizing Map Program Package*. Helsinki: Helsinki University of Technology, Laboratory of Computer and Information Science, 1996.

[237] J. Yan, "som: Self-Organizing Map," 2016.

[238] D. M. Gysi, A. Voigt, T. d. M. Fragoso, E. Almaas, and K. Nowick, "wTO: an R package for computing weighted topological overlap and consensus networks with an integrated visualization tool," 11 2017.

[239] D. N. Messina, J. Glasscock, W. Gish, and M. Lovett, "An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression," *Genome Research*, vol. 14, pp. 2041–2047, 10 2004.

[240] K. Nowick, C. Fields, T. Gernat, D. Caetano-Anolles, N. Kholina, and L. Stubbs, "Gain, loss and divergence in primate Zinc-Finger genes: A rich resource for evolution of gene regulatory differences between species," *PLoS ONE*, vol. 6, p. e21553, 6 2011.

[241] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J. H. Han, T. Hashimoto, W. Hide, O. Hofmann, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. Ogawa, R. D. Teasdale, J. Tegnér, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki, "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man," *Cell*, vol. 140, pp. 744–752, 3 2010.

[242] S. Tripathi, K. R. Christie, R. Balakrishnan, R. Huntley, D. P. Hill, L. Thommesen, J. A. Blake, M. Kuiper, and A. Lægreid, "Gene Ontology annotation of sequence-specific DNA binding transcription factors: Setting the stage for a large-scale curation effort," *Database*, vol. 2013, p. bat062, 2013.

[243] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: Function, expression and evolution," *Nature Reviews Genetics*, vol. 10, pp. 252–263, 4 2009.

[244] E. Wingender, T. Schoeps, M. Haubrock, and J. Dönitz, "TFClass: a classification of human transcription factors and their rodent orthologs.," *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D165–D170, 2015.

[245] E. Wingender, T. Schoeps, and J. Dönitz, "TFClass: An expandable hierarchical classification of human transcription factors," *Nucleic Acids Research*, vol. 41, no. D1, pp. D165–D170, 2013.

[246] C. T. Butts, "network: a Package for Managing Relational Data in R.," *Journal of Statistical Software*, vol. 24, no. 2, 2008.

[247] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Research*, vol. 32, pp. 115–119, 1 2004.

[248] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "Revigo summarizes and visualizes long lists of gene ontology terms," *PLoS ONE*, vol. 6, p. e21800, 7 2011.

[249] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes," *Journal of Biomedical Informatics*, vol. 44, pp. 839–847, 10 2011.

[250] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human MicroRNA Targets," *PLOS Biology*, vol. 2, p. e363, 10 2004.

[251] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, pp. 15–20, 1 2005.

[252] K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos, "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes," *Cell*, vol. 126, pp. 1203–1217, 9 2006.

[253] C. G. Hill, L. V. Matyunina, D. E. Walker, B. B. Benigno, and J. F. McDonald, "Transcriptional override: A regulatory network model of indirect responses to modulations in microRNA expression," *BMC Systems Biology*, vol. 8, p. 36, 3 2014.

[254] P. Langfelder and S. Horvath, "Fast R Functions for Robust Correlations and Hierarchical Clustering," *Journal of statistical software*, vol. 46, 3 2012.

[255] M. Imbeault, P.-Y. Helleboid, and D. Trono, "KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks," *Nature*, vol. 543, no. 7646, pp. 550–554, 2017.

[256] J. F. Margolin, J. R. Friedman, W. K. Meyer, H. Vissing, H. J. Thiesen, and F. J. Rauscher, "Krüppel-associated boxes are potent transcriptional repression domains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, pp. 4509–4513, 5 1994.

[257] S. Bansod, R. Kageyama, and T. Ohtsuka, "Hes5 regulates the transition timing of neurogenesis and gliogenesis in mammalian neocortical development," *Development*, vol. 144, no. 17, pp. 3156–3167, 2017.

[258] M. Mondanizadeh, E. Arefian, G. Mosayebi, M. Saidijam, B. Khansarinejad, and S. M. Hashemi, "MicroRNA-124 regulates neuronal differentiation of mesenchymal stem cells by targeting Sp1 mRNA," *Journal of Cellular Biochemistry*, vol. 116, pp. 943–953, 6 2015.

[259] X. Zhang, M. H. Chen, X. Wu, A. Kodani, J. Fan, R. Doan, M. Ozawa, J. Ma, N. Yoshida, J. F. Reiter, D. L. Black, P. V. Kharchenko, P. A. Sharp, and C. A. Walsh, "Cell type-specific alternative splicing governs cell fate in the developing cerebral cortex," *Cell*, vol. 166, pp. 1147–1162, 8 2016.

[260] D. M. Gysi, T. de Miranda Fragoso, E. Almaas, and K. Nowick, "CoDiNA: Co-Expression Differential Network Analysis," *https://cran.r-project.org/web/packages/CoDiNA/index.html*, 2018.

[261] S. Hirata, "Chimpanzee social intelligence: selfishness, altruism, and the mother–infant bond," *Primates*, vol. 50, no. 1, pp. 3–11, 2009.

[262] E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, and M. Tomasello, "Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis," *science*, vol. 317, no. 5843, pp. 1360–1366, 2007.

[263] L. Cosmides, H. C. Barrett, and J. Tooby, “Adaptive specializations, social exchange, and the evolution of human intelligence,” *Proceedings of the National Academy of Sciences*, vol. 107, no. Supplement 2, pp. 9007–9014, 2010.

[264] R. Boyd, P. J. Richerson, and J. Henrich, “The cultural niche: Why social learning is essential for human adaptation,” *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 2, pp. 10918–10925, 2011.

[265] M. A. Just, P. A. Carpenter, and S. Varma, “Computational modeling of high-level cognition and brain function,” *Human Brain Mapping*, vol. 8, pp. 128–136, 1 1999.

[266] M. Fontenot and G. Konopka, “Molecular networks and the evolution of human cognitive specializations,” *Current Opinion in Genetics and Development*, vol. 29, pp. 52–59, 12 2014.

[267] L. Liu, J. Lei, S. J. Sanders, A. J. Willsey, Y. Kou, A. E. Cicek, L. Klei, C. Lu, X. He, M. Li, R. A. Muhle, A. Ma’Ayan, J. P. Noonan, N. Šestan, K. A. McFadden, M. W. State, J. D. Buxbaum, B. Devlin, and K. Roeder, “DAWN: A framework to identify autism genes and subnetworks using gene expression and genetics,” *Molecular Autism*, vol. 5, p. 22, 3 2014.

[268] J. M. Berg and D. H. Geschwind, “Autism genetics: searching for specificity and convergence,” *Genome biology*, vol. 13, no. 7, p. 247, 2012.

[269] A. Willsey, S. Sanders, M. Li, S. Dong, A. Tebbenkamp, R. Muhle, S. Reilly, L. Lin, S. Fertuzinhos, J. Miller, M. Murtha, C. Bichsel, W. Niu, J. Cotney, A. Ercan-Sencicek, J. Gockley, A. Gupta, W. Han, X. He, E. J. Hoffman, L. Klei, J. Lei, W. Liu, L. Liu, C. Lu, X. Xu, Y. Zhu, S. Mane, E. Lein, L. Wei, J. Noonan, K. Roeder, B. Devlin, N. Sestan, and M. State, “Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism,” *Cell*, vol. 155, pp. 997–1007, 11 2013.

[270] M. M. Ryan, H. E. Lockstone, S. J. Huffaker, M. T. Wayland, M. J. Webster, and S. Bahn, “Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes,” *Molecular Psychiatry*, vol. 11, pp. 965–978, 10 2006.

[271] C. M. Nievergelt, D. F. Kripke, T. B. Barrett, E. Burg, R. A. Remick, A. D. Sadovnick, S. L. McElroy, P. E. Keck, N. J. Schork, and J. R. Kelsoe, “Suggestive evidence for association of the circadian genes PERIOD3 and ARNTL with bipolar disorder,” *American Journal of Medical Genetics - Neuropsychiatric Genetics*, vol. 141 B, pp. 234–241, 4 2006.

[272] K. Iwamoto, C. Kakiuchi, M. Bundo, K. Ikeda, and T. Kato, “Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders,” *Molecular Psychiatry*, vol. 9, pp. 406–416, 4 2004.

[273] N. Akula, J. R. Wendland, K. H. Choi, and F. J. McMahon, “An integrative genomic study implicates the postsynaptic density in the pathogenesis of bipolar disorder,” *Neuropsychopharmacology*, vol. 41, pp. 886–895, 2 2016.

[274] A. O. Cramer, C. D. Van Borkulo, E. J. Giltay, H. L. Van Der Maas, K. S. Kendler, M. Scheffer, and D. Borsboom, “Major depression as a complex dynamic system,” *PLoS ONE*, vol. 11, no. 12, p. e0167490, 2016.

[275] M. Ray and W. Zhang, “Analysis of Alzheimer’s disease severity across brain regions by topological analysis of gene co-expression networks,” *BMC Systems Biology*, vol. 4, no. 1, p. 136, 2010.

[276] R. S. Desikan, C. C. Fan, Y. Wang, A. J. Schork, H. J. Cabral, L. A. Cupples, W. K. Thompson, L. Besser, W. A. Kukull, D. Holland, C. H. Chen, J. B. Brewer, D. S. Karow, K. Kauppi, A. Witoelar, C. M. Karch, L. W. Bonham, J. S. Yokoyama, H. J. Rosen, B. L. Miller, W. P. Dillon, D. M. Wilson, C. P. Hess, M. Pericak-Vance, J. L. Haines, L. A. Farrer, R. Mayeux, J. Hardy, A. M. Goate, B. T. Hyman, G. D. Schellenberg, L. K. McEvoy, O. A. Andreassen, and A. M. Dale, "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score," *PLoS Medicine*, vol. 14, no. 3, pp. 1–17, 2017.

[277] M. Narayanan, J. L. Huynh, K. Wang, X. Yang, S. Yoo, J. McElwee, B. Zhang, C. Zhang, J. R. Lamb, T. Xie, C. Suver, C. Molony, S. Melquist, A. D. Johnson, G. Fan, D. J. Stone, E. E. Schadt, P. Casaccia, V. Emilsson, and J. Zhu, "Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases," *Molecular Systems Biology*, vol. 10, pp. 743–743, 7 2014.

[278] B. Zheng, Z. Liao, J. J. Locascio, K. A. Lesniak, S. S. Roderick, M. L. Watt, A. C. Eklund, Y. Zhang-James, P. D. Kim, M. A. Hauser, E. Grünblatt, L. B. Moran, S. A. Mandel, P. Riederer, R. M. Miller, H. J. Federoff, U. Wüllner, S. Papapetropoulos, M. B. Youdim, I. Cantuti-Castelvetri, A. B. Young, J. M. Vance, R. L. Davis, J. C. Hedreen, C. H. Adler, T. G. Beach, M. B. Graeber, F. A. Middleton, J.-C. J.-C. Rochet, C. R. Scherzer, Global PD Gene Expression (GPEX) Consortium, E. Grunblatt, L. B. Moran, S. A. Mandel, P. Riederer, R. M. Miller, H. J. Federoff, U. Wullner, S. Papapetropoulos, M. B. Youdim, I. Cantuti-Castelvetri, A. B. Young, J. M. Vance, R. L. Davis, J. C. Hedreen, C. H. Adler, T. G. Beach, M. B. Graeber, F. A. Middleton, J.-C. J.-C. Rochet, C. R. Scherzer, and Global PD Gene Expression (GPEX) Consortium, "PGC-1 , A Potential Therapeutic Target for Early Intervention in Parkinson's Disease," *Science Translational Medicine*, vol. 2, pp. 73–52, 10 2010.

[279] M. H. POLYMEROPOULOS, "Genetics of Parkinson's Disease," *Annals of the New York Academy of Sciences*, vol. 920, pp. 28–32, 1 2006.

[280] D. C. Duke, L. B. Moran, R. K. Pearce, and M. B. Graeber, "The medial and lateral substantia nigra in Parkinson's disease: mRNA profiles associated with higher brain tissue vulnerability," *Neurogenetics*, vol. 8, pp. 83–94, 4 2007.

[281] D. M. Werling, N. N. Parikshak, and D. H. Geschwind, "Gene expression in human brain implicates sexually dimorphic pathways in autism spectrum disorders," *Nature Communications*, vol. 7, pp. 1–11, 2016.

[282] S. B. GUZE, *Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV)*, vol. 152. American Psychiatric Association, 5 1995.

[283] J. Hallmayer, S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S. Ozonoff, C. Lajonchere, J. K. Grether, and N. Risch, "Genetic heritability and shared environmental factors among twin pairs with autism," *Archives of General Psychiatry*, vol. 68, pp. 1095–1102, 11 2011.

[284] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, J. D. Smith, B. Paeper, D. A. Nickerson, J. Dea, S. Dong, L. E. Gonzalez, J. D. Mandell, S. M. Mane, M. T. Murtha, C. A. Sullivan, M. F. Walker, Z. Waqar, L. Wei, A. J. Willsey, B. Yamrom, Y. H. Lee, E. Grabowska, E. Dalkic, Z. Wang, S. Marks, P. Andrews, A. Leotta, J. Kendall, I. Hakker, J. Rosenbaum, B. Ma, L. Rodgers, J. Troge, G. Narzisi, S. Yoon, M. C. Schatz, K. Ye, W. R. McCombie, J. Shendure, E. E. Eichler, M. W. State, and M. Wigler, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, pp. 216–221, 11 2014.

[285] S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. Di Lullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. A. Meyer, K. Bilguvar, S. M. Mane, N. Š estan, R. P. Lifton, M. Günel, K. Roeder, D. H. Geschwind, B. Devlin, and M. W. State, "De novo mutations revealed by whole-exome sequencing are strongly associated with autism," *Nature*, 2012.

[286] S. Marwaha, A. Durrani, and S. Singh, “Employment outcomes in people with bipolar disorder: A systematic review,” *Acta Psychiatrica Scandinavica*, vol. 128, pp. 179–193, 9 2013.

[287] T. A. Rowland and S. Marwaha, “Epidemiology and risk factors for bipolar disorder,” *Therapeutic Advances in Psychopharmacology*, vol. 8, p. 204512531876923, 9 2018.

[288] J. F. Hayes, J. Miles, K. Walters, M. King, and D. P. Osborn, “A systematic review and meta-analysis of premature mortality in bipolar affective disorder,” *Acta Psychiatrica Scandinavica*, vol. 131, pp. 417–425, 6 2015.

[289] J. Callicott, V. Mattay, H. R. Siebner, T. Sommer, J. Barnett, and J. Smoller, “The genetics of bipolar disorder,” *Neuroscience*, vol. 164, pp. 331–343, 11 2009.

[290] B. R. Rittberg, “Major depressive disorder,” *The Medical Basis of Psychiatry: Fourth Edition*, vol. 358, pp. 79–90, 1 2016.

[291] A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, and R. Poulton, “Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene.,” *Science (New York, N.Y.)*, vol. 301, pp. 386–9, 7 2003.

[292] K. S. Kendler, M. C. Neale, R. C. Kessler, A. C. Heath, and L. J. Eaves, “Major Depression and Generalized Anxiety Disorder: Same Genes, (Partly) Different Environments?,” *Archives of General Psychiatry*, vol. 49, pp. 716–722, 9 1992.

[293] A. Fanous, C. Gardner, D. Walsh, and K. S. Kendler, “Relationship between positive and negative symptoms of schizophrenia and schizotypal symptoms in nonpsychotic relatives,” *Archives of General Psychiatry*, vol. 58, pp. 669–673, 7 2001.

[294] N. C. Andreasen, “Negative Symptoms in Schizophrenia: Definition and Reliability,” *Archives of General Psychiatry*, vol. 39, pp. 784–788, 7 1982.

[295] A. G. Cardno and I. I. Gottesman, “Twin studies of schizophrenia: From bow-and-arrow concordances to star wars Mx and functional genomics,” *American Journal of Medical Genetics - Seminars in Medical Genetics*, vol. 97, pp. 12–17, 21 2000.

[296] P. McGuffin, F. Rijsdijk, M. Andrew, P. Sham, R. Katz, and A. Cardno, “The heritability of bipolar affective disorder and the genetic relationship to unipolar depression,” *Archives of General Psychiatry*, vol. 60, pp. 497–502, 5 2003.

[297] C. M. Freitag, “The genetics of autistic disorders and its clinical relevance: A review of the literature,” *Molecular Psychiatry*, vol. 12, pp. 2–22, 1 2007.

[298] V. Hyttinen, J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto, “Genetic Liability of Type 1 Diabetes and the Onset Age Among 22,650 Young Finnish Twin Pairs,” *Diabetes*, vol. 52, no. 4, pp. 1052–1055, 2003.

[299] I. Locatelli, P. Lichtenstein, and A. I. Yashin, “The Heritability of Breast Cancer: A Bayesian Correlated Frailty Model Applied to Swedish Twins Data,” *Twin Research*, vol. 7, no. 2, pp. 182–191, 2004.

[300] M. Gatz, N. L. Pedersen, S. Berg, B. Johansson, K. Johansson, J. A. Mortimer, S. F. Posner, M. Viitanen, B. Winblad, and A. Ahlbom, “Heritability for Alzheimer’s Disease: The Study of Dementia in Swedish Twins,” *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 52A, pp. M117–M125, 3 1997.

[301] T. H. Hamza and H. Payami, “The heritability of risk and age at onset of Parkinson’s disease after accounting for known genetic risk factors,” *Journal of Human Genetics*, vol. 55, pp. 241–243, 4 2010.

[302] J. R. Vaughan, M. B. Davis, and N. W. Wood, “Genetics of parkinsonism: A review,” *Annals of Human Genetics*, vol. 65, pp. 111–126, 3 2001.

[303] W. H. Berrettini, “Are schizophrenic and bipolar disorders related? A review of family and molecular studies,” 9 2000.

[304] L. S. Carroll and M. J. Owen, “Genetic overlap between autism, schizophrenia and bipolar disorder,” 10 2009.

[305] P. Tellechea, N. Pujol, P. Esteve-Belloch, B. Echeveste, M. García-Eulate, J. Arbizu, and M. Riverol, “Early- and late-onset Alzheimer disease: Are they the same entity?,” *Neurología (English Edition)*, vol. 33, pp. 244–253, 5 2018.

[306] Y. Zhang, M. James, F. A. Middleton, and R. L. Davis, “Transcriptional analysis of multiple brain regions in Parkinson’s disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms,” *American Journal of Medical Genetics - Neuropsychiatric Genetics*, vol. 137 B, pp. 5–16, 8 2005.

[307] F. B. Baker, *The basics of item response theory*. ERIC, 2001.

[308] R. A. Johnson and D. W. Wichern, “Applied multivariate statistics,” *Englewood Cliffs, NJ: Prentice Hall*, 1982.

[309] D. Borsboom and A. O. Cramer, “Network Analysis: An Integrative Approach to the Structure of Psychopathology,” *Ssrn*, vol. 9, pp. 91–121, 2013.

[310] A. O. Cramer, L. J. Waldorp, H. L. Van Der Maas, and D. Borsboom, “Comorbidity: A network perspective,” *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 137–150, 2010.

[311] D. Borsboom, A. O. Cramer, V. D. Schmittmann, S. Epskamp, and L. J. Waldorp, “The Small World of Psychopathology,” *PLoS ONE*, vol. 6, no. 11, p. e27407, 2011.

[312] L. Ruzzano, D. Borsboom, and H. M. Geurts, “Repetitive Behaviors in Autism and Obsessive–Compulsive Disorder: New Perspectives from a Network Analysis,” *Journal of Autism and Developmental Disorders*, vol. 45, no. 1, pp. 192–202, 2014.

[313] R. C. Gentleman, V. J. Carey, D. M. Bates, and others, “Bioconductor: Open software development for computational biology and bioinformatics,” *Genome Biology*, vol. 5, p. R80, 2004.

[314] S. Andrews, “FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/,” 2010.

[315] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, G. Hickey, A. S. Hinrichs, R. Hubley, D. Karolchik, K. Learned, B. T. Lee, C. H. Li, K. H. Miga, N. Nguyen, B. Paten, B. J. Raney, A. F. A. Smit, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent, “The UCSC Genome Browser database: 2015 update,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D670–D681, 2015.

[316] D. C. Berwick and K. Harvey, “The importance of Wnt signalling for neurodegeneration in Parkinson’s disease,” *Biochemical Society Transactions*, vol. 40, no. 5, pp. 1123–1128, 2012.

[317] D. C. Berwick and K. Harvey, “The regulation and deregulation of Wnt signaling by PARK genes in health and disease.,” *Journal of molecular cell biology*, vol. 6, pp. 3–12, 2 2014.

[318] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma’ayan, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” *Nucleic Acids Research*, vol. 44, pp. W90–W97, 7 2016.

[319] B. Watmuff, S. S. Berkovitch, J. H. Huang, J. Iaconelli, S. Toffel, and R. Karmacharya, “Disease signatures for schizophrenia and bipolar disorder using patient-derived induced pluripotent stem cells.,” *Molecular and cellular neurosciences*, vol. 73, pp. 96–103, 2016.

[320] P. Wu, X. Zuo, H. Deng, X. Liu, L. Liu, and A. Ji, “Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases,” 8 2013.

[321] X. Liu, D. Han, M. Somel, X. Jiang, H. Hu, P. Guijarro, N. Zhang, A. Mitchell, T. Halene, J. J. Ely, C. C. Sherwood, P. R. Hof, Z. Qiu, S. Pääbo, S. Akbarian, and P. Khaitovich, “Disruption of an Evolutionarily Novel Synaptic Expression Pattern in Autism,” *PLoS Biology*, vol. 14, p. e1002558, 9 2016.

[322] Z. Duren, X. Chen, R. Jiang, Y. Wang, and W. H. Wong, “Modeling gene regulation from paired expression and chromatin accessibility data.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, pp. E4914–E4923, 6 2017.

[323] M. De Domenico, “Multilayer modeling and analysis of human brain networks,” *GigaScience*, vol. 6, 5 2017.

Glossary

Bipartite network is a network where the nodes can be divided into two disjoint sets of nodes such that links connect nodes from the two sets to each other, but never inside the same set

Bootstrap The basic idea of bootstrapping is that inference about a population from sample data can be modelled by resampling the sample data and performing inference about a sample from resampled data. Because the population is unknown, the true error in a sample statistic against its population value is unknown. In bootstrap-resamples, the *population* is in fact the sample, and this is known; hence the quality of inference of the *true* sample from resampled data is measurable

CRISPR/CAS9 It is a genome editing technology, short for *clustered regularly interspaced short palindromic repeats and CRISPR-associated protein 9*. This method is faster, cheaper, more accurate, and more efficient than other existing genome editing methods, such as transfactions and vectors. it was adapted from a naturally occurring genome editing system in bacteria, where it capture snippets of DNA from invading viruses and use them to create DNA segments known as CRISPR arrays. This allow the bacteria to "remember" the viruses. And, if the viruses attack again, the bacteria produce RNA segments from the CRISPR arrays to target the viruses' DNA. The bacteria then use Cas9 or a similar enzyme to cut the DNA apart, which disables the virus. In the lab, the CRISPR-Cas9 system works similarly. A small piece of RNA is created. It contains a short *guide* sequence that binds to a specific target sequence of DNA in a genome. The RNA also binds to the Cas9 enzyme. and it is used to recognise the DNA sequence, and the Cas9 enzyme cuts the DNA at the targeted location. Once the DNA is cut, the cell's uses its own DNA repair machinery to add or delete pieces of genetic material, or to make changes to the DNA by replacing an existing segment with a customised DNA sequence

Degree is the number of nodes each node interacts with

Direction The direction of a link specifies the source (starting point) and a target (endpoint) where the interaction occurs

Edge Connection between two graph vertices

Gene The fundamental physical and functional unit of inheritance. They are parts of Deoxyribonucleic Acid (DNA) sequence. Genes can be a recipe for constructing proteins, but not all genes code for proteins. They vary a lot in size, in humans, genes can be few hundred DNA base pair (bp) to more than 2 million bases. The Human Genome Project estimated that humans have around $20,000$ and $25,000$ genes

Genome The complete set of genes or genetic material present in an organism

Global measures Are measures that describe the whole network, for example degree distribution; average clustering coefficient; path length; modularity index.

Hub nodes with the with a much large degree compared with the average degree value

Link Connection between two network nodes

Local measures Are characteristics of individual nodes of a network, such as their degree and centrality

Microarray A chip containing spots where DNA is attached. The amount of cDNA that hybridises to the DNA of the array can be measured by a fluorescence emission

Node The fundamental unity in a network

Pipeline A set of data processing tools plugged in series, where the output of one tool is the input of the next one

Replication The process that the DNA copies itself into a DNA molecule

Strength The strength of a node is the sum of the weights attached to links belonging to a node

Transcription The process that the DNA copies its information into an RNA molecule

Transcriptome The complete set of transcripts present in an cell. It is the major determinant of a cellular phenotype

Translation The process that mRNA information is translated into a chain of aminoacids, that later is processed into a protein

Vertex The fundamental unity in a graph

Weight is a measure of how strong a particular link interaction is

CV

Personal Information

Nationality: Brazilian
Date of birth: 13/06/1990

Address

Haertelstrasse, 16-18
04109 Leipzig
Germany

Tel & Skype

+49 15753489055
deisy.gysi

Mail

deisy@bioinf.uni-leipzig.de

Web

https://deisygysi.github.io/

Deisy Morselli Gysi

Statistician and Biotechnologist

Bioinformatician and Statistician with a solid background in Biotechnology. Extensive experience with Big Data, transcriptomics and developing methods for construction and comparison of biological networks. Passionate about complex systems biology. Independent Consultant for statistics in Medical Science. Developed and implemented public available R packages. Organized, communicative, creative and fast learner with outstanding writing skills. Works well in teams and alone. Fluent in English and Portuguese, proficient in Spanish an German.

Experience

10/2015 - Current Position
PhD Candidate Leipzig University, Leipzig, Germany
Candidate at the *Bioinformatics* and *Swarm Intelligence and Optimization* groups, studying the evolution of cognition in primates and the differences and specificity of gene-gene interaction in different brain tissues for multiple cognitive disorders mainly through expression networks. Supervisor: Prof. Dr. Katja Nowick, PhD; Prof. Dr. Martin Middendorf, PhD and Prof. Dr. Peter Stadler, PhD.

09/2017 - 12/2017
PhD Candidate Visitor Norwegian University of Science and Technology - NTNU, Trondheim, Norway
Visitor at the Almaas Lab Group, developing measures for multiple network comparison (Consensus, Differential and Specificity of links). Supervisor: Prof. Dr. Eivind Almaas, PhD.

07/2014 – 03/2015
Senior Statistician Hospital Israelita Albert Einstein, Sao Paulo, Brazil
Consultant to the research department of the hospital on the statistical analysis for researches.

03/2013 – 07/2014
Bioinformatician Heart Institute – Medical School University of São Paulo, Sao Paulo, Brazil
Development of pipelines for retrieving gene annotation for SNPs associated with phenotypes in GWAS studies, calculation of gene co-expression networks and comparison of those networks.

06/2012 – 02/2013
Credit Analyst HSBC Multiply Bank, Curitiba, Brazil
Development, validation and report of statistical models in the context of Basel II (Probability of Default, Loss of Given Default and Exposure at Default) and BAU (Business as usual). Responsible for modeling Auto Finance portfolio origination and maintenance models and related documentation. Behavioral variables used in the analysis of credit models, through studies of stability and performance, seeking improvements for new model redevelopments. Approval of databases for model development credit.

11/2011 – 05/2012
Internship HSBC Multiply Bank, Curitiba, Brazil
Development, validation and report of statistical origination credit models for Credit Cards. Use of high performance computing for implementation and validation for time-series models.

OS Preference

GNU/Linux	★★★★★
Unix	★★★★★
MacOS	★★☆☆☆
Windows	★★☆☆☆

Programming Languages

R	★★★★★
SAS	★★★☆☆
Python	★★★☆☆
Perl	★★★☆☆
Bash	★★★☆☆
C	★★☆☆☆

Languages

English	★★★★★
Portuguese	★★★★★
German	★★★★☆
Spanish	★★☆☆☆

08/2011 - 11/2011

Internship Advanced Group in Animal Breeding, Curitiba, Brazil
Development of statistical methods for breeding selection of horses from the Military Police, with better features related with equestrian activities such as running.

02/2011 - 06/2011

Internship Laboratory of extra-cellular matrix protein biochemistry and poison biotechnology - UFPR, Curitba, Brazil
Development of molecular biology skills such as extraction, purification of DNA and proteins, sequencing, vector preparation, strains transformation for expression assays using recombining proteins of the *Loxoceles intermedia* venom.

07/2010 - 12/2010

Internship Department of Statistics - UFPR, Curitiba, Brazil
Developments of statistical skills in quality control, population structures and linkage disequilibrium using databases for genomic studies using SNP chips for Genome Wide Association Analysis (GWAS).

Education

10/2015 - Now

PhD in Computer Science - Bioinformatics Leipzig University
Development of measures for comparing multiple co-expression networks build using transcriptomics data. Pre-processing, processing and quality control of transcriptome arrays and RNAseq. Supervisors: Prof. Dr. Katja Nowick - Freie Universitaet Berlin ; Prof. Dr. Martin Middendorf - Leipzig University and Prof. Dr. Peter Stadler - Leipzig University

2009 - 2013

Bachelor's degree in Statistics Federal University of Paraná – UFPR – Brazil
Bachelor thesis: Survival Analysis for cervical cancer patients. Grade (10/10). Supervisor: Prof. Dr. Suely Ruiz Giolo

2008 - 2011

Bachelor's Degree in Biotechnology Pontifical Catholic University – PUCPR – Brazil

2005 - 2007

Technological Degree in Tourism and Hotel Business Microlins – Brazil

Short courses lectured

2018

Teaching Assistant at 7th Programming for Evolutionary Biology Course Free University of Berlin

2018

Statistics and Inference using R at 7th Programming for Evolutionary Biology Course Free University of Berlin

2017	**Teaching Assistant at 6th Programming for Evolutionary Biology Course** Leipzig University
2017	**Basic Statistics at 6th Programming for Evolutionary Biology Course** Leipzig University
2016	**Teaching Assistant at 5th Programming for Evolutionary Biology Course** Leipzig University

Publications

Computer program

1. **Gysi, D. M.**; Voigt, A. ; Fragoso, T. M. ; Almaas, E. ; Nowick, K. **wTO: Computing Weighted Topological Overlaps (wTO) & Consensus.** *CRAN*. 2017. https://CRAN.R-project.org/package=wTO
2. **Gysi, D. M.**; Almaas, E. ; Nowick, K. **CoDiNA: Differential co-expression network analysis for n dimensions.** *CRAN*. 2018. https://CRAN.R-project.org/package=CoDiNA
3. **Gysi, D. M.**; Nowick, K. **RichR: Gene-to-Disease enrichment tool** *CRAN*. 2019. https://CRAN.R-project.org/package=RichR

Published papers

1. **Gysi, D. M.**; Voigt, A. ; Fragoso, T. M. ; Almaas, E. ; Nowick, K. **wTO: an R package for computing weighted topological overlap and consensus network with an integrated visualization tool**. *BMC Bioinformatics*, 2018.
2. Kutsche, L. K*;**Gysi, D. M.***; Lenk, K.; Fallmann, J; Petri, R; Klapper, S. D.; Jakobsson, J.; Nowick, K ; Busskamp, V. **Combined experimental and system-level analyses reveal the complex regulatory network of miR-124 during human neurogenesis**. *Cell Systems*, 2018.
3. Vaidotas, M.; Yokota, P. K. O.; Marra, A. R.; Sampaio Camargo, T. Z.; Victor, E. S.; **Gysi, D. M.**; Leal, F.; Dos Santos, O. F. P.; Edmond, M. B. **Measuring hand hygiene compliance rates at hospital entrances**. *American Journal of Infection Control*, v. 43, p. 694–696, 2015.

Manuscripts in preparation

1. **Gysi, D. M.**; Fragoso, T. M.; Busskamp, V.; Almaas, E. ; Nowick, K. **Co-expression Differential Network Analysis: How compare multiple networks simultaneously?**. *Submitted*. 2019.
2. **Gysi, D. M.**; Fragoso, T. M.; Nowick, K. **Construction, comparison and evolution of networks in biology, social sciences, economy, and humanities – or: what can we learn from other disciplines**. *In preparation*, Invited Review. 2019.
3. Banos, S.; **Gysi, D. M.**; Richter-Heitmann, T.; Nowick, K.; Friedrich, M.; Glöckner, F. O.; Boersma, M.; Wiltshire, K. H.; Wichels, A.; Gerdts, G.; Reich, M. **Dynamics observed in a pelagic marine fungal community: an interplay of oscillation types, stability, resilient, and biotic interactions**. *In preparation*, 2019.

4. **Gysi, D. M.**; Nowick, K. **Make me Rich: an R enrichment package**. *In preparation*. 2019.
5. Geffre, A. ;Gernat, T. ;Toth, A. ; Robinson, G.; Bonning, B. ; Hamilton, A.; Jones, B. ; **Gysi; D. M.**; Dolezal. A. **Pathogen manipulation in the Anthropocene: Viruses of managed honey bees alter host social behaviour**. *In preparation*, 2019.

Oral presentation

1. **Gysi, D. M.**; Fragoso, T. M.; Almaas E.; Nowick, K. **Co-expression Differential Network Analysis**, XXIX International Biometric Conference, 2018.
2. **Gysi, D. M.**; Fragoso, T. M.; Almaas E.; Nowick, K. **Comparing multiple co-expression networks**, 4th Summer School in Complex Networks, 2018.
3. **Gysi, D. M.**; Nowick, K. **wTO: an R package to calculate weighted topological overlap networks**, XIV Herbstseminar der Bioinformatik, 2016.
4. **Gysi, D. M. Evolution of gene co-expression networks implicated in cognitive functions in primates**, XIII Herbstseminar der Bioinformatik, 2015.

Posters

1. **Gysi, D. M.**; Fragoso, T. M. ; Almaas, E. ; Nowick, K. **How to build and compare co-expression networks**, BenGenDiv, Berlin, 2018.
2. **Gysi, D. M.**; Voigt, A. ; Fragoso, T. M. ; Almaas, E. ; Nowick, K. **An R package for calculating the Weighted Topological Overlap Network with a visualization tool**, CompleNet'18, Boston, 2018.
3. **Gysi, D. M.**; Voigt, A. ; Fragoso, T. M. ; Almaas, E. ; Nowick, K. **wTO, an R package for computing the weighted Topological Overlap and Consensus Networks**, NORBIS annual conference, Tromso, 2017.
4. Kutsche, L. K.; **Gysi, D. M.**; Lenk, K.; Petri, R.; Jakobsson, J.; Nowick, K.; Busskamp, V. **A systems level view on miR-124 function during neuronal differentiation from human iPS cells**, Intelligent Systems for Molecular Biology, Prague, 2017.
5. Kutsche, L. K.; **Gysi, D. M.**; Lenk, K.; Petri, R.; Jakobsson, J.; Nowick, K.; Busskamp, V. **A systems level view on miR-124 function during neuronal differentiation from human iPS cells**, Gene regulatory mechanisms in neural fate decisions, San Juan de Alicante, 2017.
6. Bertoli, W; **Gysi, D. M.**. **Bayesian Estimation of the Zero-Inflated quasi Poisson-Lindley Model**, 4^{th} Workshop on Probabilistic and Statistical Methods, 2016, Sao Carlos, Brazil.
7. **Gysi, D. M.**; Pilar, P. G.; Giolo, S. R. **Modelo de Sobrevivência com Fração de Cura Aplicado aos Dados de Pacientes com Câncer de Colo do Útero**, IV WASA - Workshop em Analise de Sobrevivência e Aplicações, 2015, Belo Horizonte.
8. Bertoli, W.; **Gysi, D. M. Métodos Estatísticos na Análise de Dados Genômicos**, Workshop on Probabilistic and Statistical Methods, São Carlos, 2013.
9. **Gysi, D. M.**; Pilar, P. G.; Giolo, S. R. **Análise da sobrevida de pacientes com câncer do colo do útero**, XIII Escola de Modelos de Regressão, Maresias, 2013.
10. **Gysi, D. M.**; Giolo, S. R. **Metodologias Estatísticas Aplicadas à Genética Quantitativa e Genômica**. Encontro de Iniciação Científica da UFPR, 2012, Curitiba. *Caderno de Resumos Evinci*, 2012.

11. Bertoli, W.; **Gysi, D. M. Análise de Componentes Principais Para Obtenção de Grupos de SNPs Informativos**. VI Bienal da Sociedade Brasileira de Matemática, 2012, Campinas. *Anais da VI Bienal da Sociedade Brasileira de Matemática*, 2012.
12. **Gysi, D. M.**; Giolo, S. R. **Estatística Computacional em Genética Quantitativa e Genômica**. Encontro de Iniciação Científica da UFPR, 2011, Curitiba. *Caderno de Resumos Evinci*, 2011.
13. **Gysi, D. M.**; Rakin, S.; Saez, R. **Elaboração de um banco de Dados de DNA de pacientes com fissura lábio palatina não sindrômica**. XVIII Seminário de Iniciação Científica - PUC–PR, 2010, Curitiba. XVIII Seminário de Iniciação Científica. *Editora Champanag*, 2010.
14. Novelino, A.; Rakin, S.; **Gysi, D. M.**; Saez, R.; Grabowski, M.; Souza, J. **Elaboração de banco de DNA de pacientes com fissuras labiopalatais**. XVII Seminário de Iniciação Científica - PUC–PR, V PIBIC Jr., XII Mostra de Pesquisa, II SPPGEM, Curitiba, 2009.
15. **Gysi, D. M.**; Rakin, S.; Saez, R.; Grabowski, M.; Souza, J.; Novelino, A. **Elaboração de banco de dados epidemiológico, clínico, e genético de pacientes com fissuras labiopalatais**. XVII Seminário de Iniciação Científica, V PIBIC Jr., XII Mostra de Pesquisa, II SPPGEM, Curitiba, 2009.

Undergraduate research

08/2010 - 07/2012 **Statistical Methodologies Applied to Quantitative Genetics and Genomics** Federal University of Parana
Development of statistical models for associating genes with diseases via robust statistical methodologies. Finding associated Single Nucleotide Polymorphims associated with diseases, correcting by population structures. Supervisor: Suely Ruiz Giolo, PhD.

08/2008 - 07/2010 **Preparation of a DNA bank from patients with Cleft-Lip** Pontifical Catholic University of Parana
Acquisition of knowledge in the laboratory, in the field of molecular biology. Extraction, sample preparation of DNA, PCR, preparation of reagents and chemical solvents. The computational part of this project focused on the creation, development and maintenance of a database (development in Access). Supervisor: Salmo Raskin, PhD.

Honors & Awards

04/2015 - 09/2015 **German Course** DAAD
Fellowship awarded for a 6 months course of German language.

10/2015 - now **PhD** Science without borders - CNPq
Fellowship awarded for a 36 months PhD in Germany.

2010- 2012 **Undergraduate Research fellowship** UFPR - CNPq
Fellowship for Undergraduate Research awarded by the Brazilian Government in partnership with Federal University of Parana

08/2012 **First-Place award: Undergraduate Research** 20º. Evento de Iniciação Científica da Universidade Federal do Paraná
Best undergraduate student award at Federal University of Parana. Thesis: Statistical Methodologies Applied to Quantitative Genetics and Genomics. *Supervisor: Suely Giolo, PhD*

08/2011 **First-Place award: Undergraduate Research** 19º. Evento de Iniciação Científica da Universidade Federal do Paraná
Best undergraduate student award at Federal University of Parana. Thesis: Computational Statistics Applied to Quantitative Genetics and Genomics. *Supervisor: Suely Giolo, PhD*

2008- 2010 **Undergraduate Research fellowship** PUCPR - CNPq
Fellowship for Undergraduate Research awarded by the Brazilian Government in partnership with Pontifical Catholic University of Parana

Selected Courses and Extensions

2018 - 2018 **Scientific Writing** Leipzig, Germany.

2018 - 2018 **Leadership and working in multiprofessional teams** Leipzig, Germany.

2018 - 2018 **4th Summer school in Complex Networks.** Como, Italy.

2018 - 2018 **Network Meta-Analysis with R.** Barcelona, Spain.

2018 - 2018 **Summer school of Bioinformatics.** USP, Sao Paulo, Brazil.

2017 - 2017 **sDIV-working group ``The genomic evolution of key adaptive traits - utilizing the potential of non-model organisms sGENEVA!.** Leipzig, Germany.

2017 - 2017 **Metabolic pathway analysis.** NORBIS, Trondheim, Norway.

2017 - 2017 **Large genetic studies in biobanks: GWAS and beyond.** NORBIS, Oslo, Norway.

2017 - 2017 **Algebraic Statistics Day.** Max Plank Institute fuer Matematik, Leipzig, Germany

2017 - 2017 **Workshop Big Data in Business.**Universität Leipzig, Leipzig, Germany

2016 - 2016 **A beginner's Guide to RNA-Seq Data Analysis Course.** ESeq Bioinformatics, Leipzig, Germany

2013 - 2013 **Regression Models with limited and censured response.** University of São Paulo, Sao Paulo, Brazil.

2013 – 2013 **Statistical Inference using bootstrap and application.**University of São Paulo, Sao Paulo, Brazil.

2012 - 2012	**Modelling Academy** HSBC Bank Brazil, Curitiba, Brazil.
2012 - 2012	**Statistical modelling for credit risk analysis.** Federal University of São Carlos, São Carlos, Brazil.
2012 - 2012	**Non Linear mixed models using R.** Fundação de Estudos Agrários Luiz de Queiroz, Piracicaba, Brazil.
2012 - 2012	**Computational Methods in Statistical Inference** Federal University of Parana, Curitiba, Brazil.
2012 - 2012	**Mixed Models using R** Fundação de Estudos Agrários Luiz de Queiroz, Piracicaba, Brazil.
2011 - 2011	**Introduction to Proteomics Analysis.** Federal University of Parana, Curitiba, Brazil.
2011 - 2011	**II Winter school of Biochemistry and molecular biology.** Federal University of Parana, Curitiba, Brazil.
2011 - 2011	**Population Genetics.** Federal University of Parana, Curitiba, Brazil.
2011 - 2011	**Association Studies of Genes with human diseases.** Federal University of Parana, Curitiba, Brazil.
2011 - 2011	**IV Winter school of genetics.** Federal University of Parana, Curitiba, Brazil.
2010 - 2010	**Pirosequencing: A molecular approach.** Federal University of Parana, Curitiba, Brazil.
2010 - 2010	**Genome Wide Association Studies.** Federal University of Santa Catarina, UFSC, Brazil.
2008 - 2008	**Research with human embrionary stem cell.** Pontifical Catholic University of Parana, Curitiba, Brazil.
2008 - 2008	**Electronic Microscopy.** Pontifical Catholic University of Parana, Curitiba, Brazil.
2008 - 2008	**Virus, vaccines and antivirals.** Federal University of Parana, Curitiba, Brazil.

Other activities

2018 - now	**Representative of Caucus women in Statistics** German representative of Women in Statistics.
2017 - now	**R-Ladies chapter Leader** Leipzig leader of the Global R-Ladies.