DOCTORAL THESIS

CregNET: Meta-Analysis of *Chlamydomonas reinhardtii* Gene Regulatory Network

Author:
Ngoc C. PHAM

Supervisor:
Prof. Patrick E. MEYER

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in

The Bioinformatics and Systems Biology (BioSys) Lab
Faculté des Sciences

August 12, 2020

Declaration of Authorship

I, Ngoc C. PHAM, declare that this thesis titled, "CregNET: Meta-Analysis of *Chlamydomonas reinhardtii* Gene Regulatory Network " and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"If I have seen further it is by standing on the shoulders of giants."

Isaac Newton

"All things are difficult before they are easy."

Dr. Thomas Fuller

UNIVERSITY OF LIÈGE

Abstract

Faculté des Sciences
The Research Unit InBioS

Doctor of Philosophy

CregNET: Meta-Analysis of *Chlamydomonas reinhardtii* Gene Regulatory Network

by Ngoc C. PHAM

Chlamydomonas reinhardtii is a well-known model organism used to investigate numerous biological processes, such as photosynthesis, starch metabolism, etc. Thanks to it's available genome, a high volume of high-throughput transcriptomic data has been made available during the last few years.

Gene regulatory network (GRNs) underlie all cellular phenomena; and thus, a comprehensive understanding of GRN maps are essential tools to elucidate gene function, thereby facilitating interpretations of biological processes Mochida et al., 2018. However, a GRN underpinning cellular processes of *C. reinhardtii* with biotechnological interest has not been addressed. As a result, a system biology approach for *reverse engineering* it's GRN from the abundance of data is of great interest.

In the thesis, we have evaluated various meta-analysis methods in the context of inferring GRN from multiple transcriptomic data sets, which has led to a new version of the R and Bioconductor minet package (P. E. Meyer, Lafitte, and Bontempi, 2008). Afterwards, we developed a novel meta-analysis computational framework for inferring a GRN of *C. reinhardtii* (called CregNET) from various RNA-seq studies collected from NCBI SRA (Benson et al., 2012). In the first step of our pipeline the mutual information (MI) based network inference algorithm namely MRNET (P. E. Meyer, Lafitte, and Bontempi, 2008) was performed on all the RNA-seq data, resulting in a compendium of GRNs. The GRNs were then aggregated to create the CregNET in the next step. Experimental results with both synthetic data and biological data (e.g. *E. coli*, *Saccharomyces cerevisiae* and *Drosophila*) prove that the meta-analysis approach can generate robust biological hypotheses of gene regulations from a bunch of gene expression data. Additionally, a set of benchmarks performed on CregNET demonstrates the robustness and predictive power of CregNET.

Acknowledgements

My sincere gratitude to my supervisor Professor Dr. Patrick E. Meyer for the constant support throughout my PhD, for his patience and motivation. The support I received from him during the writing and editing of each and every part of this thesis enriched my analytical thinking ability and improved my scientific writing skill.

Furthermore, I would like to thank the members of the jury who have accepted to review this work, namely Prof. Denis Baurain (Université de Liège, Belgium) who is the president of the jury, Prof. Bernard Peers (Université de Liège, Belgium) - secretary of the jury, Professors Pierre Geurts (University of Liège, Belgium), Prof. Philippe Salembier (Universitat Politecnica de Catalunya BarcelonaTECH, Spain) and Dr. Oliver Caelen (Orange, Belgium).

A special thank to my colleague, Manuel Noll, with whom I have discussed and learned a lot of scientific questions. Also, thanks to his contribution the final result in this work has been validated.

Finally, I would like to thank my wife who took off the family burden and has continuously supported me throughout my PhD as well as my lovely daughter and son who are always by my side. My sincere thanks go to my parents who always support and love me no matter what. Special thanks to my late father-in-law who I admire very much and my mother-in-law who is pretty much always here to help.

Contents

List of Figures

List of Tables

List of Abbreviations

GRN	**G**ene **R**egulatory **N**etwork
DM	**D**ata **M**erging
MA	**M**eta **A**nalysis
NE	**N**etwork **E**ensamble
MI	**M**utual **I**nformation
MIM	**M**utual **I**nformation **M**atrix
MAGE	**M**icroarray **G**ene **E**xpression
D1	**D**ata Merging BMC
D2	**D**ata Merging z-score
D3	**D**ata Merging COMBAT
N1	**N**etwork Ensemble RankSum
N2	**N**etwork Ensemble Internal Quality Control Index
N3	**N**etwork Ensemble Median
M1	**M**atrices of Coexpression Aggregation Random-effects
M2	**M**atrices of Coexpression Aggregation IQCI
M3	**M**atrices of Coexpression Aggregation Median

For/Dedicated to/To my beloved family

Chapter 1

Introduction

1.1 Motivation

1.1.1 What are gene regulatory networks (GRNs) ?

The genome encodes thousands of genes whose products enable cell survival and various cellular functions (Karlebach and Shamir, 2008). Moreover, it is now understood that in cells, genes do not operate in isolation but they are expressed and work in concert by numerous regulatory mechanisms. In such network of gene-gene interactions, gene regulation plays a key role in the control of fundamental processes in living organisms, ranging from development, to nutrition and metabolic coordination (Carré, Mas, and Krouk, 2017). To demonstrate, gene regulation at the transcriptional level in bacteria is essential for controlling metabolic flexibility and cellular integrity (De Smet and Marchal, 2010). Interestingly, it is observed that the regulation and expression of some genes are f other genes are more variable: their levels are differenthighly robust while that o from cell to cell and from individual to individual (MacNeil and Walhout, 2011). Usually, this regulation process requires a time lag, which involves the regulation of protein translation, folding, nuclear transport, turnover, and the extension of the target mRNA (Yang et al., 2018). Though genes are regulated at several levels, one key step is the control of gene transcription (Carré, Mas, and Krouk, 2017).

Typically complex regulation involving multiple regulators results in single genes showing highly specific expression behavior that is not shared with other genes. In addition, it was observed that many transcription factors (TFs) are active in similar conditions and thus trigger similar sets of genes, suggesting either redundancy in their function or an intricate cooperation between different TFs to mediate a common response (De Smet and Marchal, 2010). For years, the bindings of TFs to DNA sequences, forming gene regulatory networks (GRNs), have been a crucial aspect in systems biology. In essence, a GRN links TFs to their target genes and represents a map of transcriptional regulation (Banf and Rhee, 2017). They are composed of multiple sub-circuits and each of them accomplishes individual regulatory tasks (Eric H. Davidson, 2010). Understanding the dynamics of these networks, therefore, not only can we shed light on the mechanisms of diseases that occur when these cellular processes are dysregulated but also speed up biotechnological projects, as such predictions are quicker and cheaper than lab experiments (Karlebach and Shamir, 2008). Thanks to GRNs we are able to understand the intracellular physiological activity and function of biology, interaction in the pathway and thus gain knowledge of how to make the organism change (Yang et al., 2018).

Fox example, changes in plant transcriptional regulation led to many modern crops and enabled large yield increases (R. S. Meyer and Purugganan, 2013). In

another example developmental GRNs provide the specific causal links between genomic regulatory sequences and the processes of development (E H Davidson and Levine, 2008). They consist of the regulatory and signaling genes that drive any given process of development and the functional interactions among them. The design features of the GRN directly explain why the events of a given process of development occur; for example, why a given set of cells becomes specified to a given fate, why it emits particular signals to adjacent cells, and why it differentiates in a given direction (E H Davidson and Levine, 2008). Needless to say computational analysis contribute to basic biological research, for example, by explaining developmental mechanisms or new aspects of the evolutionary process (Karlebach and Shamir, 2008). In precision medicine, only a small fraction of patients respond to the drug prescribed to treat their disease, which means that most are at risk of unnecessary exposure to side effects through ineffective drugs (Van Der Wijst et al., 2018). And since many diseases are associated with mutations in transcriptional regulators (TRs) or in TF binding sequences (Banf and Rhee, 2017), their mechanisms that are characterized by dysfunction of regulatory processes can be elucidated by personalized GRNs (Karlebach and Shamir, 2008). Consequently, analyses of GRNs are key to identify disease mechanisms and possible therapeutic targets for the future (Hase et al., 2013). In the prospect, well-validated, context-specific, personalized GRNs will be essential to move from more traditional medicine towards precision medicine (see figure 1.1 for a better clarification), which will provide treatment or preventive measures that will be effective for patients based on their specific genetic, environmental, and lifestyle characteristics (Van Der Wijst et al., 2018).

In the last decades, innovations in experimental methods have enabled large scale studies of GRNs and can reveal the mechanisms that underlie them (Karlebach and Shamir, 2008). Specifically, GRNs can be constructed from gene expression data sets that implicitly contain gene regulation information in specific conditions (e.g., disease-specific, tissue-specific, or drug-specific GRNs) (Yu et al., 2013). However, inferring GRNs from high-throughput expression data is a fundamental but challenging task in computational systems biology (Madhamshettiwar et al., 2012).

1.1.2 GRNs in the systems biology context

The recent advances in omics technology, combined with computational analysis to form an emerging approach named systems biology, holds great promise owing to its capacity to extract valuable information from a large amount of data (Park, K. H. Lee, et al., 2007). In 'systems biology', one aims to model the physiology of living systems as a whole rather than as a collection of single biological entities (Hecker et al., 2009). The development of molecular profiling techniques nowadays enables the high-throughput and affordable acquisition of large omics data sets, such as for transcriptomics, proteomics and metabolomics (Serin et al., 2016). While substantial efforts are being made to generate large omics data sets, there is a growing need to develop platforms to integrate these data and derive models describing biological interactions (Serin et al., 2016). This is because such system-level approach can offer insights into how to control and optimize parts of a system while potential side effects are well addressed (Hecker et al., 2009). Indeed, to gain a better understanding of the observed complex global behavior and the underlying biological processes, it is necessary to model the interactions between a large number of components that make up such a biological system (Hecker et al., 2009). Since GRNs provide information that is essential for a global understanding of the logic of gene-gene interactions,

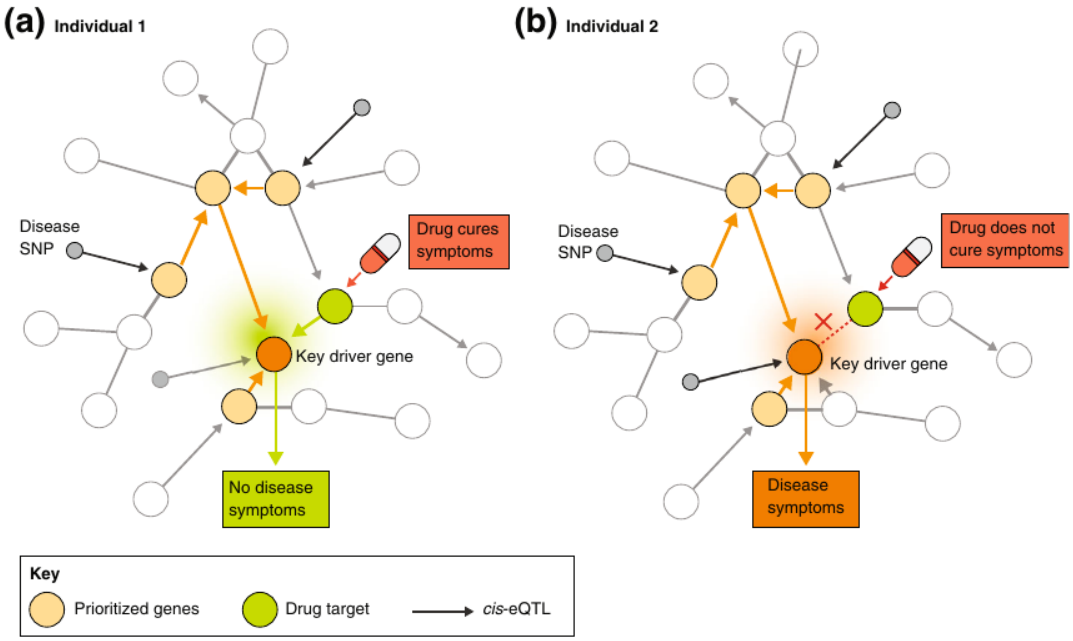

FIGURE 1.1: Implications of personalized gene regulatory networks for precision medicine. Depending on an individual's regulatory wiring, specific drugs may or may not be effective. Personalized GRNs will provide guidance for precision medicine in the future. In this example, GRNs of two hypothetical patients are shown in which the regulatory wiring between the drug target gene and the key driver gene is different. **a** In individual 1, the drug target gene activates the key driver gene. **b** In individual 2, the interaction between both genes is absent. Thus, in individual 1, the drug is effective, whereas in individual 2, the drug is ineffective. (Van Der Wijst et al., 2018)

inference of such networks has been one of the key challenges in system biology (Y. Wang et al., 2018).

Plants are fascinating and complex organisms. A comprehensive understanding of the organization, function and evolution of plant genes is essential to disentangle important biological processes and to advance crop engineering and breeding strategies. The ultimate aim in deciphering complex biological processes is the discovery of causal genes and regulatory mechanisms controlling these processes (Serin et al., 2016). Such networks can be utilized to identify target genes for both knockout and over-expression for the increased production of desired products such as biomass and resilience against pathogens. That is because the impact of genetic abnormality can spread through regulatory interactions in GRNs and alter the activity of other genes that do not have any genetic defects (Hase et al., 2013). In this context, random mutagenic technique without a full understanding of underlying regulatory mechanisms can lead to unwanted changes in physiology and growth retardation. In contrast, system-wide transcriptome analysis can reveal not only several target genes but also important regulatory circuits to increase the yield of products (Park, S. Y. Lee, et al., 2008). A transcription factor-based genetic engineering technique, for instance, has been proposed for crop improvement due to the fact that transcription factors (TFs) are master regulators of cellular processes (Century, Reuber, and Ratcliffe, 2008). However, expression levels of genes are the result of interactions between different biological processes and TFs usually regulate the expression of many downstream genes; hence, mutating TFs without taking into account their global effects on the system might cause unwanted phenotypes (Century, Reuber, and Ratcliffe, 2008).

Recently, systems biology has gained much success in modeling GRNs of unicellular organisms such as *Escherichia coli* (Gama-Castro et al., 2008). Moreover, systematic integration of DNA-binding data with genome-wide mRNA and miRNA expression data allowed determining key-regulatory properties in the GRN that controls early stages of flower development in *Arabidopsis thaliana*, including specification of floral whorls, and organ differentiation and growth (D. Chen et al., 2018). Additionally, in another project named Drosophila Model Organism Encyclopedia of DNA Elements (modENCODE) (Roy et al., 2010), network inference of a multicellular eukaryotic organism is shown to be feasible.

Typically, GRN inference methods primarily use gene expression data derived from microarrays or high-throughput sequencing (RNA-seq) technologies to identify a correlation between a pair of genes based on a significant interaction signal. However, it should be emphasized that high-throughput techniques are generating massive data sets to elucidate the behavior and interaction of thousands of genes across diverse conditions. Obviously, we cannot describe complex cellular systems by only focusing on one single control mechanism measured by a single experimental technique. GRNs reverse engineered from heterogeneous data sets are; hence, essential for explaining cellular response under various perturbations, though this is very challenging. This is because assembling expression data is no trivial task due to negative batch effects caused by even small differences in experimental preparation or the specific platform and analysis procedure used. Moreover, a single expression study is limited by small sample size (typically tens of thousands of probes are investigated in only tens or hundreds of biological samples (Ramasamy et al., 2008)), resulting in network inference algorithms' sensitivity to outliers. Thus, combining multiple experimental studies is also able to significantly increase the statistical power of network inference algorithms. Such data integration leverages dependencies that can be confidently uncovered thanks to the multitude of surveyed

conditions, but leads to context-agnostic wiring diagrams (Y. Wang et al., 2018).

1.1.3 Why GRNs for *Chlamydomonas reinhardtii*

Recently, there has been increased attention on algae, mostly due to their potential commercial applications in biofuels and nutritional supplements (R. L. Chang et al., 2011). The reason is because microalgal cells, characterized by high photosynthetic efficiency and rapid cell division, are an excellent source of neutral lipids as potential fuel stocks (Goncalves et al., 2016). The unicellular green algae *Chlamydomonas reinhardtii* is a well-known model organism used to investigate numerous biological processes, such as photosynthesis, starch metabolism, etc. (Siaut et al., 2011). Recent studies and literature on the species have recommended that it is a very important cell factory as well. For instance, it is presented in (Siaut et al., 2011) that high oil yields can be gained by blocking the starch synthesis. Moreover, approximately 235mg ethanol can be produced for every 1.0g of algal biomass i.e. a volume of ethanol equivalent to 24% of the biomass (Choi, M. T. Nguyen, and Sim, 2010). The very compelling results in (Kong et al., 2010), where *C. reinhardtii* was cultured in wastewater to not only produce bio-fuel but also remove nitrogen and phosphorus from wastewater, illustrate another potential application of the organism.

Although microalgae, such as *C. reinhardtii*, are highly promising cell factories for renewable biofuels, there is still an urgent need of a conceptual framework to make micro algal biofuels economically competitive (Chisti, 2007). Indeed a better understanding of the metabolic and regulatory networks can provide insights for increased TAG synthesis, with fewer drawbacks than for existing algal cells (Schaap et al., 2014). Indeed, microalgal cells naturally produce both desirable and undesirable metabolites; hence, novel target pathways for product formation should be amplified whereas biosynthetic pathways for byproducts should be removed or attenuated simultaneously (Park, S. Y. Lee, et al., 2008).

Thanks to the availability of its genome, a massive amount of transcriptomic data has been produced for *C. reinhardtii*. Unfortunately, a GRN underpinning its cellular processes with biotechnological interest has not been addressed. Above all, despite previous successes, system-level studies of *C. reinhardtii* in order to both enhance desired phenotype and reduce unwanted side effects for cost-effective bio-products are in high demand. To this end (Romero-campero et al., 2016) develop a first gene co-expression network and an associated web-based software tool that integrates RNA-seq data available for the *C. reinhardtii* transcriptome. It should be emphasized that constructing co-expression networks is generally straightforward. For instance, in the gene co-expression networks (GCNs), a node is a gene, and an edge is drawn between gene A and B if the correlation coefficient between these two genes is above a threshold. The main difference between GRNs and GCNs is GCNs treat TF and non-TF genes similarly whereas GRN involves sophisticated *reverse-engineering* algorithms that operate on TFs differently (Gupta and Pereira, 2019). The objective of the thesis; therefore, aims at delivering a first global-scale GRN of *C. reinhardtii*, underpinning it's responses to numerous stress conditions, from it's massive amount of transcriptomic data.

1.1.4 Data availability for reconstructing GRN of *C. reinhardtii*

Advances in microarray and, more recently, next-generation sequencing technologies provide a wealth of data for GRN inference (Madhamshettiwar et al., 2012), leading to the development of a community infrastructure for sharing data such as

NCBI Gene Expression Omnibus (Ron Edgar, Domrachev, and Lash, 2002), ArrayExpress (Brazma et al., 2003) or InSilicoDB (Taminau, Meganck, et al., 2012). The original goals of these databases was to make the data available to other researchers for independent analysis and, where appropriate, integration with their own data (Heider, 2013). Microarray is a well-established, cost-effective, high-throughput technology able to simultaneously measure the expression levels of thousands of genes and hereby offers an efficient way to generate a snapshot of the entire transcriptome (Larsen et al., 2014). It is a revolutionary tool for identifying genes or pathways whose expression changes in response to specific perturbations (C. Chen et al., 2011). Not surprisingly, transcriptome analysis by microarray technology has become a routine tool in many research areas ranging from basic cell biology to clinical research (Heider, 2013). Over time, different types and generations of microarrays have been produced by several manufacturers resulting in several hundred thousand microarray samples clustered by different species, manufacturers and chip generations (Heider, 2013). The integrative analysis of multiple microarray gene expression (MAGE) data sets has been acknowledged to be a crucial approach for extracting the maximum relevant biological information (Lazar et al., 2013). It should be noted that; however, those databases above still largely contain microarray data sets because they are relatively cheap and the pipeline of analysis is highly standardized.

Microarray data for *C. reinhardtii* is rich and can be easily found at NCBI Gene Expression Omnibus (figure 1.2). However, a big challenge lines on the fact that most of the microarray data has not been well annotated. One option is reannotating data by identifying which probes represent a given gene within and across the data sets using the BLAST algorithm (Ramasamy et al., 2008). Nevertheless, this is only possible when all probe sequences are provided for all platforms, which is not always a case for *C. reinhardtii*. Additionally, probe annotation and mapping is not a trivial task. For example, on the Affymetrix GeneChip U95A, 11% of the probes are non-specific and 9% of the probes are mismatched to the genome (Miao et al., 2011). Moreover, studies using the microarray technologies are only able to cover different subsets of all genes predicted for *C. reinhardtii* and the overlapping between these subsets are insignificant. That is to say even though there is a large number of microarray studies for *C. reinhardtii*, the question of how to integrate the data is fundamental but very challenging.

Dataset	Platforms	Perturbation	No of samples	Source
GSE24367	GPL9100 GPL10980	Nitrogen deprivation	8	Plant physiology (2010)
GSE42035	GPL9100	Absence of oxygen (acclimation responses)	12	Proc Natl Acad Sci (2013)
GSE40031	GPL15922	Dark to light transition	16	Proc Natl Acad Sci (2013)
GSE55253	GPL15922	Zn-resupply	8	Nature chemical biology (2014)
GSE56800	GPL18571	Different light intensity	20	Plant Cell (2014)
GSE58786	GPL15922	Nitrogen-starved	34	Eukaryotic cell (2014)
GSE48677	GPL13913	Various concentrations of silver	48	Proc Natl Acad Sci (2014)
GSE59629	GPL15922 GPL18983	Nitrogen and sulfur starvation	36	Nature Plants (2015)
...	...	...	...	...

FIGURE 1.2: C. reinhardtii's microarray data

In recent years, RNA-Seq is a developed approach to transcriptome profiling that uses next-generation high-throughput sequencing technologies. Moreover, RNA-Seq has proven to be a powerful tool for whole transcriptome profiling with enhanced sensitivity and enhanced specificity (Serin et al., 2016) over the microarrays. For example, it is shown that RNA-seq outperforms microarray (93% versus 75%) in DEG verification as assessed by quantitative PCR, with the gain mainly due to its improved accuracy for low-abundance transcripts (Megherbi et al., 2014). Another study also suggests that when using standard microarray and RNA-Seq protocols, RNA-Seq provides better estimates of absolute transcript levels (Y. Li et al., 2009). Consequently, RNA-seq is replacing cDNA microarrays as the dominant technology because it offers reduced cost, increased sensitivity, ability to quantify splice variants and perform mutation analyses, improved quantification at the transcript level, identification of novel transcripts, and improved reproducibility (Lachmann et al., 2018).

On the other hand, publicly available RNA-seq data is currently provided mostly in raw form, a significant barrier for global and integrative retrospective analyses (Lachmann et al., 2018). Furthermore like other high-throughput sequencing technologies, RNA-Seq faces several informatics challenges, i.e. the lack of standardization of pipelines using the RNA-seq technology (Z. Wang, Gerstein, and Snyder, 2009). Once these obstacles are overcome, it is clear that RNA-seq will become the predominant tool for expression analysis (Wan et al., 2015).

Public data repositories, such as the Sequence Read Archive (SRA) (Barrett et al., 2013), host > 50,000 human RNA-seq samples and it is estimated that these repositories are likely to double in size every 18 month (Collado-Torres et al., 2017). To address this growing presence of RNA-seq data, a large-scale integration of RNA-seq-based expression data is of great importance. However, it is worth mentioning that combining these data sets from different resources requires uniform alignment, quantification, and removal of batch effects, and several recent efforts have combined such uniformly processed bulk RNA-seq data sets in large repositories (Van Der Wijst et al., 2018).

Over the last few years most of studies for *C. reinhardtii* use RNA-seq technology, proving a rich compendium for system studying of it's large-scale GRN. A first attempt of utilizing such RNA-seq data can be found in the work of (Romero-campero et al., 2016). The authors collected more than 287 GigaBytes of information produced by seven different studies to construct a first gene co-expression network and an associated web-based software tool for *C. reinhardtii*. Nevertheless, the volume of data collected accounts for only a small part of all RNA-seq data could be found from SRA repository. Additionally, the problem of batch effects accompanied with integrating transcriptomics data is not well addressed.

1.2 Contributions

1.2.1 Meta-analysis for inferring GRNs from numerous gene expression datasets

In the thesis, we have proposed and evaluated various meta-analysis methods in the context of inferring GRN from multiple transcriptomic data sets (Ngoc C Pham, Haibe-Kains, et al., 2017). All the methods are then integrated in a new version of the R/Bioconductor minet package (Ngoc C Pham and P. E. Meyer, 2019). For the new version, the users can simply provide a list of gene expression datasets and their preference meta-analysis method for inferring a meta GRN.

1.2.2 CregNET - a first GRN of *C. reinhardtii*

Additionally, a main part of the thesis is the creation of CregNET - a first GRN of *C. reinhardtii* from various RNA-Seq datasets. To archive that we also propose a pipeline for collecting and pre-processing RNA-Seq data of*C. reinhardtii* from NCBA SRA. It should be noted that this pipeline can be extended for other model organisms as well. Experiment results then strongly suggest that CregNET outperforms the current co-expression network ChlamyNET in term of stability and predictive power for new GO discovery.

1.2.3 Publications

Here, we present a selection of the papers produced during the Ph.D. either as a first author or as a collaborator:

Articles

Pham, Ngoc C, Benjamin Haibe-Kains, et al. (2017). "Study of Meta-analysis strategies for network inference using information-theoretic approaches". In: *BioData mining* 10.1, p. 15.

Pham, Ngoc C and Patrick E Meyer (2019). "Minet version 4.0: meta-analysis methods to infer gene regulatory network from multiple data sets." In: *To be submitted in BMC Bioinformatics*.

Pham, Ngoc C, Manuel Noll, and Patrick E Meyer (2019). "CregNET: Meta-analysis of *Chlamydomonas reinhardtii* gene regulatory network." In: *To be submitted in Molecular Systems Biology*.

Books

Bellot, Pau, Philippe Salembier, et al. (2019b). *Unsupervised GRN Ensemble*. Springer, pp. 283–302.

1.3 Outline

The next chapter we explain why integrative analysis is important and how we can apply the technique for integrating transcriptomic data. In chapter 3 we will present the state-of-the-art MI-based data-merging approaches as well as our proposed MI-based meta-analysis methods for reverse engineering meta GRNs. In the fourth chapter we evaluate the methods with *in silico* and biological setups. Chapter 5 presents a pipeline for collecting, constructing and validating CregNET. Finally, the sixth chapter presents the conclusions of this thesis.

Chapter 2

Data-merging and meta-analysis for integrating transcriptomic data

Integrating multiple data sets is an inexpensive way to provide increased statistical power and thus help to gain valuable insights of the systems under study. This chapter provider an overview of the two most common approaches to infer GRNs from multiple data sets namely data-merging and meta-analysis.

2.1 Why integrative analysis?

Reverse engineering of GRNs remains a challenging task in systems biology. This is in part due to the large amount of experimental noise and the large number of genes relative to the small sets of conditions in gene expression of the data (Banf and Rhee, 2017). Since a single data set has typically a small sample size (usually less than 200 observations) and suffers from potential experimental biases, classical *reverse engineering* algorithms, which relies only on a standalone data set, show their limits in unraveling reliably the underlying interactions. It has been known for a long time that the small number of biological samples used per experimental study is a bottleneck in genomic analysis (Taminau, Lazar, et al., 2014). A clear application, where this limitation has been pointed out, is the prediction of disease outcome where thousands of samples are needed to generate robust gene/protein signatures (Lazar et al., 2013).

By contrast, integrative analysis of multiple studies is able to increase significantly the statistical power and thus is becoming a standard procedure in modern computational biology (Kugler et al., 2011). Another important beneficial aspect which naturally derives from developing and using integrative analysis tools is related to the cost of transcriptomics experiments. Recycling and reusing public available data would also considerably reduce the overall costs of experiments (Lazar et al., 2013). Additionally a different subset of biological information is captured by each data set. As a result integrating diverse biological data can improve functional description. Obviously, this is a relatively easy and inexpensive way of gaining new biological insights since it makes comprehensive use of already available data accumulated through the years by various groups all over the world (Taminau, Lazar, et al., 2014). For this reason, heterogeneous data integration methods have emerged to construct more reliable eukaryotic GRNs (Banf and Rhee, 2017). Nevertheless, the question of how to integrate data consistently and efficiently raises new challenges (Taminau, Meganck, et al., 2012).

Such integrative analysis could be performed in two ways: by data merging (DM), which is analyzing all the raw data coming from different studies with similar biological questions together (Bevilacqua et al., 2011) or by meta-analysis which

is the statistical analysis of a large collection of results from individual studies for the purpose of combining their findings to reach a common result (see figure 2.1 for more details).

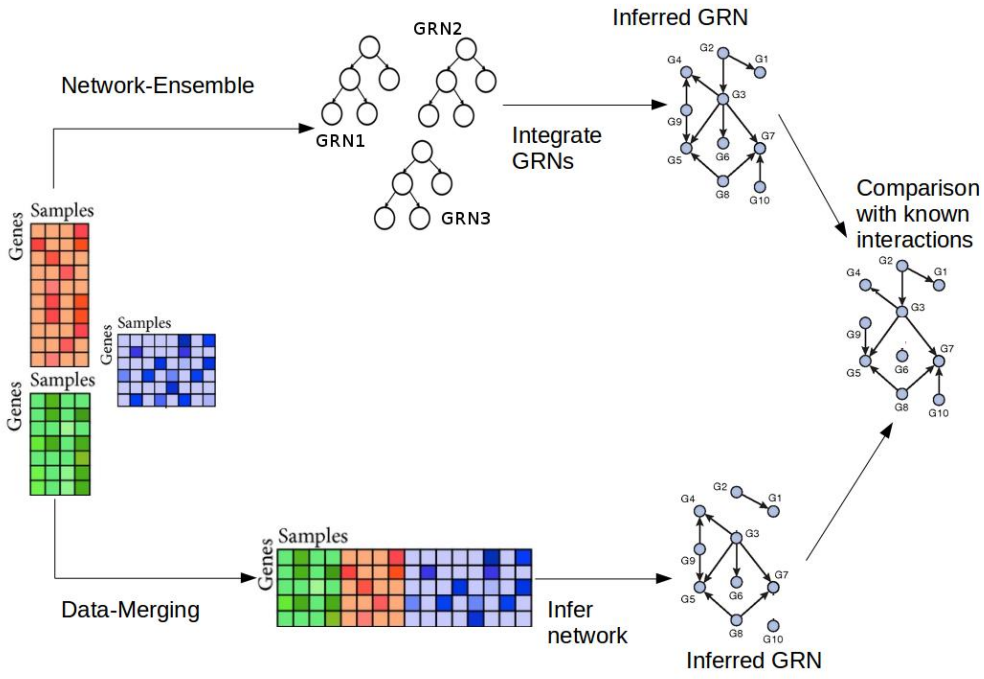

FIGURE 2.1: DM and MA for inferring GRNs from multiple data sets.

2.2 Data-merging (DM)

The DM approaches merge samples from different studies in a unique data set, on which subsequent analyses are performed. The main idea relies on integrating different studies in order to increase the sample size and thus increases the statistical power. Indeed, as more and more data sets are available on public repositories, merging different data sets appears as a simple solution to improve the relevance of the biological information extracted (Renard and Absil, 2017). Typically, the first transformation applied to expression data, referred to as normalization and summarization, removes non-biological variability between arrays and extracts gene level expression from probe intensities, respectively (Bevilacqua et al., 2011). Nevertheless, combining or merging data from different MAGE experiments for integrative analysis suffers from non-biological experimental variation or "batch effects" and it is still a challenging and difficult problem to be solved in computational biology (Lazar et al., 2013). In fact, the use of different chip types or platforms, and procedures in different research groups performed by different people introduce the statistical biases hindering downstream analysis (Leek, Scharpf, et al., 2010).

Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study (Leek, Scharpf, et al., 2010). For example, batch effects may occur if a subset of experiments was run on Monday and another set on Tuesday, if two technicians were responsible for different subsets of the experiments or if two different

lots of reagents, chips or instruments were used (Leek, Scharpf, et al., 2010). Some of the most common factors that can contribute to the generation of batch effects are pointed out by (Luo et al., 2010), namely chip type/platform, sites/laboratories, storage/shipment conditions, RNA isolation, etc.

Although a carefully designed experimental process can limit the impact of such effects, some are often unavoidable due to the large sample size requirements and potentially lengthy time required to complete a study (Stein et al., 2015; Renard and Absil, 2017). When combining data sets from different experiments, batch effects are carried over and therefore it is inappropriate to combine data sets without adjusting for batch effects (Bevilacqua et al., 2011) as this can confound true biological signals and lead to misinterpretation of the data (Larsen et al., 2014). Consequently it is important to identify and remove such effects before performing any downstream analysis.

In the work of (Heider, 2013) (see Figure 2.2) the authors introduced a procedure of 7 distinct steps to combine raw gene expression data sets in order to remove batch effects and generate a combined "ExpressionSet" object for downstream analysis. In the next subsection, various methods for batch-effect removal will be discussed.

2.2.1 Methods for batch effect removal

It has been acknowledged that batch effects is the main source of variation between different MAGE data sets (Lazar et al., 2013). The main issue is that batch effects introduce a new source of signal into the data that can be confused with the signal an analyst is looking for (Leek, 2014). Over the last few years, different approaches have been proposed to detect and remove batch effects from microarray data. However, it should be noted that over-adjusting can be more damaging, especially in the context of prediction tasks where the phenotype to predict is unequally distributed among the batches (Renard and Absil, 2017). Normalization is a data analysis technique that adjusts global properties of measurements for individual samples so that they can be more appropriately compared. Including a normalization step is now standard in data analysis of gene expression experiments (Leek, Scharpf, et al., 2010).

Batch Mean-Centering - BMC

In Sims et al., 2008 a simple normalization method based on mean centering, which is similar to z-score normalization, is applied for merging breast cancer data sets. This method transforms data by subtracting the mean of each gene over all samples (per batch) from its observed expression value, such that the mean for each gene becomes zero.

$$\hat{x}_{ij}^{k} = x_{ij}^{k} - \bar{x}_{i}^{k} \tag{2.1}$$

Gene standardization

Similarly, the z-score normalization (Cheadle et al., 2003) transforms all genes to have 0 mean and standard deviation σ of 1 by subtracting the mean $\bar{x}_i$ and dividing by the σ_i of each gene over all samples within a batch (Lazar et al., 2013).

$$\hat{x}_{ij}^{k} = \frac{x_{ij}^{k} - \bar{x}_{i}^{k}}{\sigma_{i}^{k}} \tag{2.2}$$

Quantile normalization

More complicated methods like the quantile normalization is widely used as pre-processing technique to remove technical noise presented in the microarray data generated by Affymetrix GeneChip platform (Qiu, H. Wu, and Hu, 2013). The motivaiton is that the quantile normalization makes the empirical distribution of gene expressions pooled from each array to be the same (Qiu, H. Wu, and Hu, 2013). To illustrate, scaling gene counts by a quantile of the gene-count distribution (Bullard et al., 2010) is used by (Romero-campero et al., 2016) to construct a gene co-expression network for *C. reinhardtii* using RNA-seq data.

Robust multi-array average - RMA

In another work the robust multi-array average (RMA) was adopted by (Henríquez-Valencia et al., 2018) in order to produce a sulfate co-expression network of *Arabidopsis*. Furthermore, (Van Parys et al., 2014) exploit the robust multiarray average method combined with quantile normalized to combine 45 series of experiments to produce an abiotic stress GRN of *Arabidopsis*.

Nevertheless, the normalization steps are ineffective in removing the batch effects, especially when combining data from different platforms (Lazar et al., 2013). This happens because the normalization steps take into account only few sources of batch effects unlike the more specialized methods for batch effect removal (Lazar et al., 2013). Fore instance, normalization methods only adjust the intensities at the global and they are not designed to remove artifacts presented only on a subset of probes or genes (Kocher et al., 2011). The same conclusion is also drawn by (Luo et al., 2010), in which significant batch effects still exist even after normalization for the majority of the data sets. In order to address the bottleneck of combining trancriptomics data sets, the system-level comparisons of normalization and batch effect removal algorithms have been performed over the last few years. For instance (Kocher et al., 2011) investigated and compared three common quantile normalization approaches, namely quantile normalization at average β value (QNβ), two step quantile normalization at probe signals (lumi) and quantile normalization of A and B signal separately (ABnorm). Interestingly, it was shown that normalization can reduce part but not all batch effects. Therefore, Empirical Bayes (EB) batch correction introduced by (Johnson, C. Li, and Rabinovic, 2007) along with normalization was recommended for effective batch effect removal.

COMBAT method

In (Johnson, C. Li, and Rabinovic, 2007) COMBAT, also known as Empirical Bayes (EB) method, is a method using estimations for the Location-Scale (LS) parameters (mean and variance) for each gene. COMBAT assumed that gene expression values of gene i in sample j in each batch can be depicted as:

$$x_{ij} = \alpha_i + C\beta_i + \gamma_i^X + \delta_i^X \epsilon_{ij}^X \tag{2.3}$$

where α_i is the real gene expression values for gene i, C is a design matrix for sample conditions (known covariates), β_i is the vector of regression coefficients corresponding to C, γ_i^X and δ_i^X are the additive and multiplicative batch effects for gene i respectively and ϵ_{ij}^X are noise terms. The noise term are assumed to follow a normal

distribution with mean zero and variance σ_i^2. Then the first step of COMBAT is to standardize the data using estimates $\widetilde{\alpha}_i$, $\widetilde{\beta}_i$, $\widetilde{\delta}_i^X$ and $\widetilde{\sigma}_i^2$:

$$z_{ij} = \frac{x_{ij} - \widetilde{\alpha}_i - C\widetilde{\beta}_i}{\widetilde{\sigma}_i^X} \tag{2.4}$$

The batch effects is adjusted as:

$$\hat{x}_{ij} = \frac{\widetilde{\sigma}_i}{\widetilde{\delta}_i^{X*}}(z_{ij} - \widetilde{\gamma}_i^{X*}) + \widetilde{\alpha}_i + C\widetilde{\beta}_i \tag{2.5}$$

with $\widetilde{\delta}_i^{X*}$ and $\widetilde{\gamma}_i^{X*}$ being estimates of batch effects parameters using equation 2.3 with parametric or nonparametric empirical priors. When using parametric priors it is assumed that
$gamma_i^X \sim N(\gamma^X, (\tau^X)^2)$ and $(\delta_i^X)^2 \sim InverseGamma(\lambda^X, \theta^X)$ where δ^X, $(\tau^X)^2$, λ^X and θ^X are estimated empirically.

Distance-weighted discrimination - DWD

Distance-weighted discrimination (DWD) (Benito et al., 2004) is among other popular methods for batch effects removal. DWD starts by searching for the optimal hyperplane $w \times x + b = 0$ separating samples from the different batches, with w the normal vector of the hyperplane. Next DWD remove bias by projecting the direction of the normal vector to this hyperplane by calculating the mean distance from all samples in each batch to the hyperplane $(\bar{d^X})$ and then subtracting the normal vector to this plane multiplied by the corresponding mean distance.

$$\hat{x}_{ij} = x_{ij} - \bar{d^X} w_i \tag{2.6}$$

2.2.2 Evaluations of batch effects removal methods and tools for batch effects removal

In (C. Chen et al., 2011) the authors evaluated six methods for adjusting microarray data for batch effects namely ComBat, Ratio_G, SVA, DWD, and PAMR using multiple measures of precision, accuracy and overall performance. ComBat, an Empirical Bayes method, outperformed the other five methods by most metrics. Also it can robustly manage high-dimensional data when sample sizes are small. In another work, (Larsen et al., 2014) further demonstrated how ComBat are suitable to successfully overcome systematic technical variations in order to unmask essential biological signals. Recently, a modification to ComBat called M-ComBat that centers data to the location and scale of a pre-determined, "gold-standard" batch was proposed by (Stein et al., 2015).

(Heider, 2013) provides the *virtualArray* software package to combine raw data sets, generating a combined "ExpressionSet" object and allowing further manipulation and analysis in R and other software. Interestingly, there are seven implemented methods to adjust for batch effects in the data namely quantile discretization (Warnat, Eils, and Brors, 2005), normal discretization normalization (Martinez, N. Pasquier, and C. Pasquier, 2008), gene quantile normalization (X. Q. Xia et al., 2009), median rank scores (Warnat, Eils, and Brors, 2005), quantile normalization (B. M. Bolstad et al., 2003), empirical Bayes methods (COMBAT) and mean centering (BMC).

In another work (Taminau, Meganck, et al., 2012) presents the *inSilicoMerging* R/Bioconductor package, which combines several of the most used methods to remove the unwanted batch effects to combine data sets in an intuitive and user-friendly manner, namely BMC , COMBAT, DWD, GENENORM (Z-score standardization), and XPN (Cross-Platform Normalization).

Recently, (Leek, 2014) has developed a version of the SVA approach specifically created for count data or FPKMs (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) from sequencing experiments based on appropriate data transformation.

2.3 Meta-analysis (MA)

2.3.1 What is MA?

In the mean time, MA of gene expression data sets is increasingly performed to help identify robust molecular signatures and to gain insights into underlying biological processes (J. Xia, Gill, and Hancock, 2015). In contrast to DM, in MA the results of individual studies (e.g., values, ranks, classification accuracy, etc.) are combined at the interpretative level. The main idea is combining data sets directly can be difficult because of inherent biases, i.e. batch effects, in the data that are not always removed with normalization (Steele and Tucker, 2008) or even with batch effect removal algorithms. In MA based approaches, it is assumed that if a result is found as being significant for a big number of individual studies, it will be significant for the particular problem the studies have been designed for. Moreover, if a finding is not significant in some studies, it could still be significant after meta-analysis if it appears as being significant in a big enough number of other individual studies, as the evidence will accumulate for this particular finding (Lazar et al., 2013).

Several strategies have been proposed in order to perform meta-analysis on gene expression data. For instance, a meta-analysis of public gene expression data and clinical data was conducted by using the concept of "coexpression" modules to reveal various results of previous gene expression studies in breast cancer (Wirapati et al., 2008; Desmedt et al., 2008). Moreover (Hong et al., 2006) developed a Bioconductor package named *RankProd* that allow researchers to do meta-analysis under two experimental microarray conditions to identify differentially expressed genes. However, when data sets containing few samples are studied, it is hard to derive rigorous inference upon the results issued from their analysis (Lazar et al., 2013). A direct consequence of combining the results issued from the analysis of data sets containing few samples is the fact that the statistical hypothesis tests used to make decisions using MAGE data are prone to high false-negative rates (Lazar et al., 2013). While the problem of detecting differential expressed genes across several studies has been intensively studied, it is, however, not yet the case when it comes to constructing GRNs.

2.3.2 MA for inferring GRNs

Recently “ensemble” methods of merging GRNs from different datasets, i.e. by weighting gene-gene interactions according to their average rank in each network (Marbach et al., 2012), have emerged as an alternative to the “data merging” approach. This approach rooted in the “wisdoms of crowds” concept, which was first introduced in the DREAM5 challenge and then further developed by (Hase et al.,

2013) with the Top*k*Net algorithms to produce consensus networks. Indeed, integration of 29 gene regulatory network inference methods in *yeast* and *E. coli* generated an ensemble prediction that outperformed all of the individual methods (Marbach et al., 2012). The consensus-based approach was afterwards increasingly applied to *reverse engineering* GRNs from multiple data sets. For instance, by integrating four different network inference methods, a recent study predicted an ensemble gene regulatory network in Arabidopsis (Van Parys et al., 2014). In addition, *miRsig*, an online tool for analysis and visualization of the disease-specific signature/core miRNA-miRNA interactions has been developed (Nalluri et al., 2017) using the consensus-based approach. Moreover, in the work of (Hansen et al., 2018) ensemble gene function predictions are also performed by integrating 10 co-function networks.

2.4 Data-merging or meta-analysis?

While MA methods implicitly take into account batch effects, DM require suitable batch-effect removal algorithms. The first study that try to directly compares the performances of the two approaches on finding differentially expressed genes can be found in (Taminau, Lazar, et al., 2014). Six batch effect removal methods were selected: NONE (no batch effect removal), BMC, COMBAT, DWD, and XPN for DM approach. However, for MA approach only the method of taking intersection of DEG lists estimated from each study is considered for comparisons with those DM methods. Interestingly, this study concludes that both approaches achieve comparable results. Moreover, (Silberberg et al., 2016) compared MA and DM methods in the context of retrieving gene-gene interactions in compendia of microarray studies, containing biological data with 11 studies on *Escherichia coli*, 7 studies on *Yeast* and synthetic data simulated across different networks, levels of systematic bias, number of considered studies and number of samples. It shows that batch effects should be carefully taken into account when retrieving gene-gene interactions, and researchers can adopt either a DM or MA approach depending on the specific application.

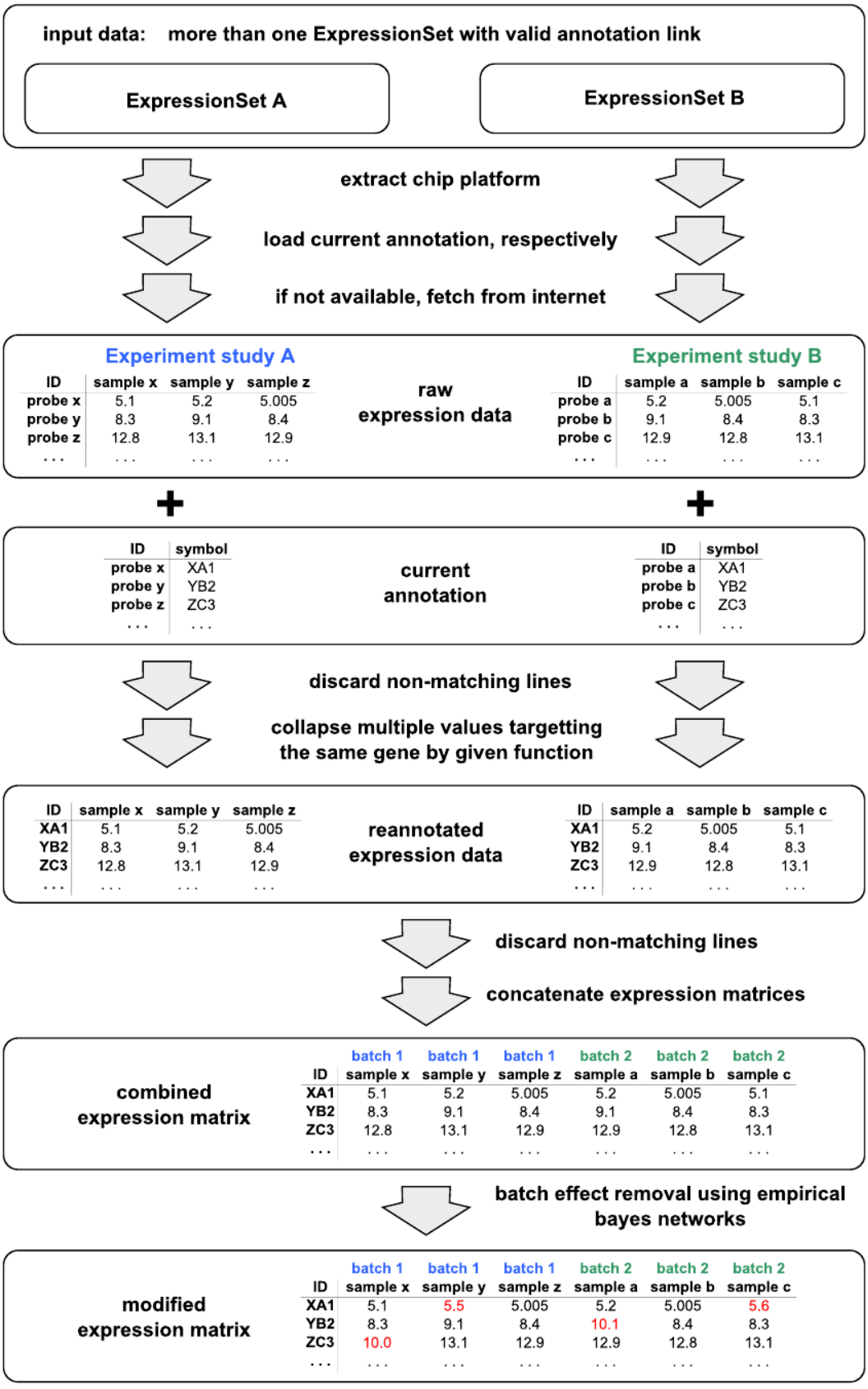

FIGURE 2.2: Distinct steps performed by virtualArray for removing batch effects when combining gene expression data sets (Heider, 2013).

Chapter 3

MI-based data-merging and meta-analysis

GRNs are typically represented as directed graphs in which nodes represent genes (for example, encoding a transcription factor or its target gene), and edges their regulatory interaction as Figure 3.1.

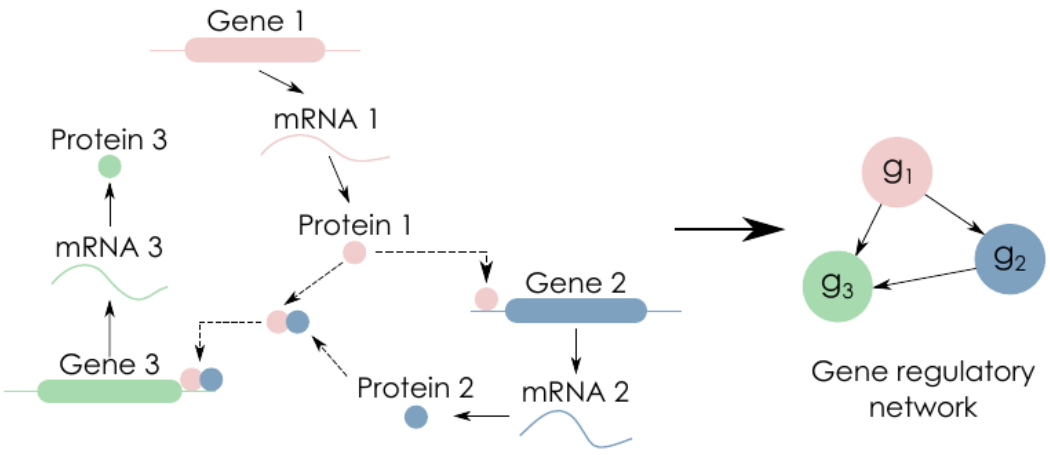

FIGURE 3.1: An example of a GRN representing the interaction between three genes, involving both direct regulation (gene 2 by gene 1) and combinatorial regulation via complex formation (gene 3 by genes 1 and 2). Image from Sanguinetti et al., 2019.

The identification of large-scale GRN has been a difficult and hot topic of system biology in recent years Yang et al., 2018. Network inference, which is the reconstruction of biological networks from high-throughput data, can provide valuable information about the regulation of gene expression in cells De Smet and Marchal, 2010. Various computational models have been developed for regulatory network analysis. A straightforward approach for performing integrative analysis of multiple studies is combining all data sets together and then analyzing the merged dataset. These method, named "data merging" and denoted here with the letter (D), were widely used in Wolfgang Huber et al., 2002; Belcastro et al., 2011; Adler et al., 2009 to reconstruct large-scale GRNs because of their simplicity. However, since high dimensional data often suffers from unwanted biases, a variety of techniques can be used to correct for these non-biological variations. We present in the following two classical scaling methods typically used to assemble data sets, and several widely-used batch-effect-removal methods.

In this work, we also introduce a new meta-analysis strategy to build consensus networks. The new strategy consists in aggregating matrices of pairwise mutual information with each being estimated from a gene expression dataset to produce a meta-matrix, from which a GRN is inferred using classical information-theoretic network inference algorithms.

3.0.1 Gene co-expression network (GCN)

A gene co-expression network (GCN) is an undirected graph with each node representing a gene, and a pair of nodes being connected with an edge if there is a significant co-expression relationship between them. To construct a GCN typically a two step approach is followed (Figure 3.2): calculating co-expression measure and selecting significance threshold. In the first step, a co-expression measure is selected (i.e. the absolute value of Pearson correlation coefficient) and a similarity score is calculated for all pairs of genes using this measure. Then when a threshold is determined all gene pairs which have a similarity score higher than the selected threshold are connected by an edge in the network.

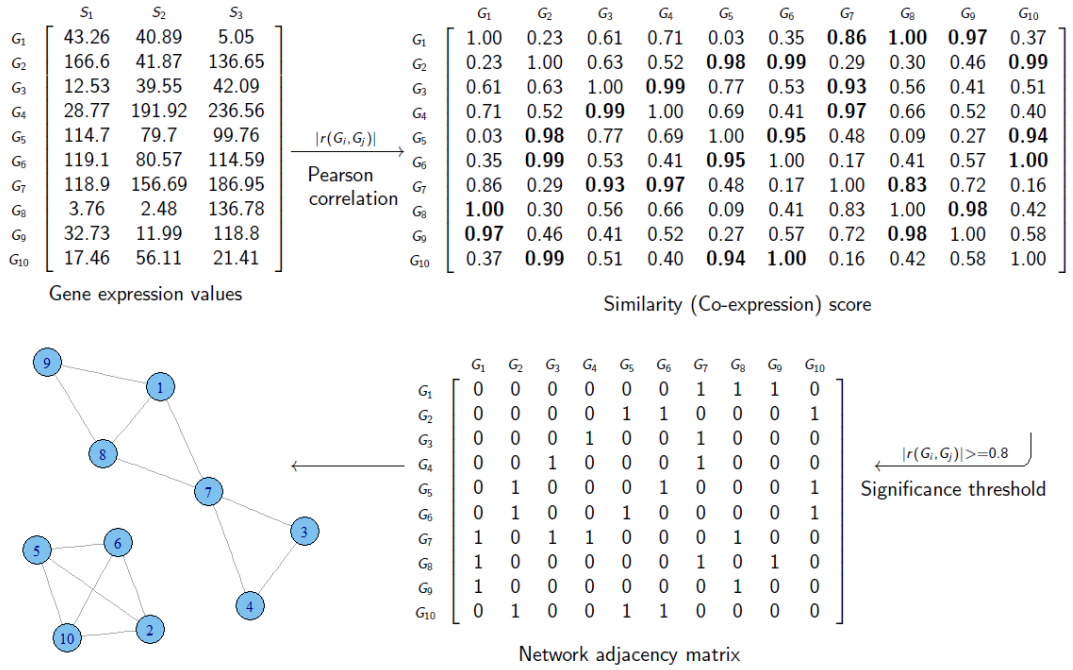

FIGURE 3.2: The two general steps for constructing a gene co-expression network. Image from Wikipedia

GRNs are of biological interest since co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex (Weirauch, 2011). Nevertheless, the direction and type of co-expression relationships can not be found in GCNs. Compared to a GRN, a GCN does not attempt to infer the causality relationships between genes and in a GCN the edges represent only a correlation or dependency relationship among genes (De Smet and Marchal, 2010).

3.1 Gene regulatory network (GRN) State-of-the-art

Innovations in experimental methods have enabled large-scale studies of gene regulatory networks and can reveal the mechanisms that underlie them (Karlebach and Shamir, 2008). Computational methods to infer GRNs typically combine evidence across different conditions to infer context-agnostic networks (Y. Wang et al., 2018). Therefore a spectrum of methods to construct GRNs from only gene expression data have been developed, counting on the relation between expression of TFs and expression of their target genes. In recent years, many methods that infer GRNs based on gene expression alone have been proposed. Early methods inferred regulatory relationships using mutual information between the expression levels of gene pairs (Y. Wang et al., 2018).

It should be stressed that state-of-the-art tools for network inference use specific assumptions and simplifications (i.e. linearity, independence or normality) to deal with the underdetermination, and these influence the inferences. As a result the final inferred network varies between tools and can be highly complementary (De Smet and Marchal, 2010).

Each network-inference algorithm generates a confidence score for a link between two genes from expression data and assumes that a predicted link with higher confidence score is more reliable. These algorithms can be classified into several categories (Marbach et al., 2012), such as: regression-based, pairwise similarity (mutual information, correlation, etc.), Bayesian networks or even ensemble approaches (that combine several different approaches). For instance, in regression models based methods regulators are inferred for each target gene. Thus, for every gene g, denoting by x_{gi} its expression level in sample i, one need to solve the regression problem (Sanguinetti et al., 2019):

$$x_{gi} = \sum_{j \neq g} w_j x_{ij} + \epsilon_i$$

with ϵ_i the bias, and w_j the weight associated with the network edge between gene j and gene g. This is the main idea for some algorithms like TIGRESS (Haury et al., 2012), LASSO (Omranian, J. M. Eloundou-Mbebi, et al., 2016), etc. GENIE3 (Huynh-Thu, Irrthum, et al., 2010) also follow this strategy, but replacing linear regression with an ensemble of regression trees. In the other category, Bayesian networks (BN) use genetic data as prior information with multiple testing and greedy search steps (Vignes et al., 2011). However their efficiency and accuracy in dealing with high dimensional transcriptomic data sets is still very limited (Vignes et al., 2011). Among those, mutual information (MI) based algorithms, such as CLR (Faith et al., 2007), ARACNE (Margolin et al., 2006), MRNET (P. E. Meyer, Kontos, and Bontempi, 2007; P. E. Meyer, Lafitte, and Bontempi, 2008) and so on, gather more and more attention owing to their capability to deal with up to several thousands of variables in the presence of a limited number of samples (P. E. Meyer, Lafitte, and Bontempi, 2008).

3.1.1 MI-based mehthods

Generally, MI-based algorithms start by estimating a pairwise mutual information (i.e. a non-linear dependency measure) between all pairs of genes, resulting in a mutual information matrix (MIM). Afterwards, indirect interactions are eliminated from the MIM by the different approaches and subsequently a GRN is inferred.

Mutual information (MI) is a non-linear measure of dependency between two variables (genes) X and Y, defined as follow

$$I(X,Y) = \sum_{x \in X, y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{3.1}$$

where $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

MI can also be defined in terms of entropy as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.2}$$

Figure 3.3 shows a simple example for calculating mutual information between two random variables (genes) using Equation 3.2. It should be noted that for using Equation 3.2 each mutual information calculus demands the estimation of three entropy terms. Interestingly, in (P. E. Meyer, Lafitte, and Bontempi, 2008) the authors presented four fastest and most used entropy estimators (i.e. empirical, Miller-Madow, Schurmann-Grassberger and shrink) for estimating MI. The four estimators were also implemented and made available by the authors in the minet Bioconductor package (P. E. Meyer, Lafitte, and Bontempi, 2008). It should be noted that as MI is a symmetric measure, MI-based network inference algorithms are not capable of deriving the direction of the gene-gene interactions. This dependency measure has been used for reconstructing networks by several methods such as Relevance network (Butte and Kohane, 2000), CLR faith2007large, ARACNE (Margolin et al., 2006) or MRNET (P. E. Meyer, Kontos, and Bontempi, 2007; P. E. Meyer, Lafitte, and Bontempi, 2008).

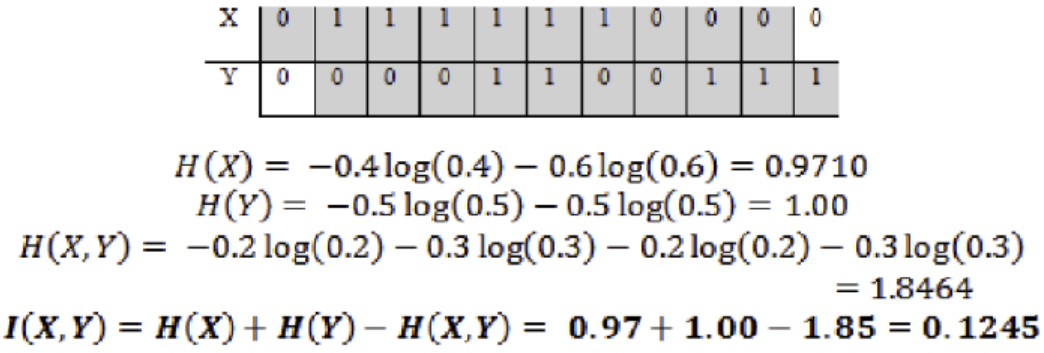

FIGURE 3.3: Estimation of MI between two random variables. Image from Chaitankar et al., 2010

3.1.2 Relevance network

The approach constructs a GRN where a pair of genes X_i, X_j is linked by an edge if the mutual information $I(X_i; X_j)$ is larger than a given threshold I_0. The complexity of the method is $O(n^2)$ since we need to consider all possible pairwise interactions within the network.

3.1.3 CLR Algorithm

The CLR method (The Context Likelihood or Relatedness network) is similar to the relevance network but applies an adaptive background correction step to eliminate false correlations and indirect influences (Faith et al., 2007). CLR creates an edge between each pair of genes i and j if the combined z-score of the mutual information between them (Figure 3.4) is above a given threshold, where the combined z-score is defined as:

$$z_{ij} = \sqrt{z_i^2 + z_j^2} \text{ with } z_i = \max(0, \frac{I_{ij} - \mu_{I_i}}{\sigma_{I_i}}) \tag{3.3}$$

in which, μ_{I_i} and σ_{I_i} are the mean and standard deviation of the empirical distribution of the mutual information of gene i.

This step removes many of the false correlations in the network by eliminating "promiscuous" cases, where one transcription factor weakly co-varies with a large numbers of genes, or one gene weakly co-varies with many transcription factors (Faith et al., 2007). The complexity of CLR is $O(n^2)$ given the computed MIM.

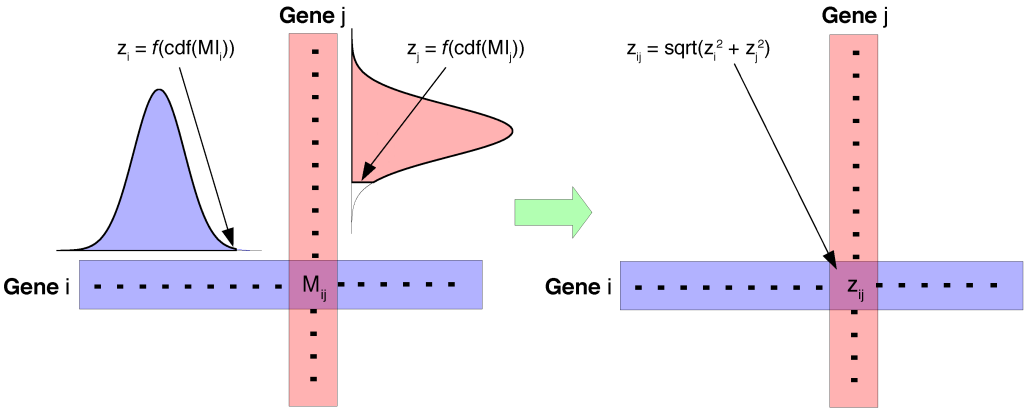

FIGURE 3.4: A schema of the CLR algorithm. The z-score of each regulatory interaction depends on the distribution of MI scores for all possible regulators of the target gene (z_i) and on the distribution of MI scores for all possible targets of the regulator gene (z_j). Image from Faith et al., 2007

3.1.4 ARACNE

The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Margolin et al., 2006) relies on the "Data Processing Inequality" (DPI) which removes the edge with the weakest mutual information, in every triplet of genes. For instance, if gene X_1 interact with gene X_3 through gene X_2, then

$$I(X_1; X_3) \leq min(I(X_1; X_2), I(X_2; X_3))$$

Given a threshold I_0 ARACNE assigns to each pair of nodes a weight equal to the MI and then all edges that $I(X_i; X_j) < I_0$ are removed. Afterwards, for each triplet the weakest edge is considered as an indirect interaction and is removed if the difference between the two lowest weights is above a threshold W_0. ARACNE's complexity is therefore $O(n^3)$ as all triplets of genes are considered.

3.1.5 MRNET

The Minimum Redundancy NETworks (MRNET) (P. E. Meyer, Lafitte, and Bontempi, 2008) method reconstructs a network using the feature selection technique known as Minimum Redundancy Maximum Relevance (MRMR) for every random variable $X_i \in X$ (Ding and Peng, 2005). The idea consists in performing a series of supervised MRMR gene selection procedures where each gene in turn plays the role of the target output (P. E. Meyer, Lafitte, and Bontempi, 2008). In order to select the predictor for a variable X_j the MRMR methods ranks a set $X_{S_j} \in X \setminus X_j$ of the predictor variables according to the difference between the mutual information of $X_i \in X_{S_j}$ with X_j (the relevance) and the average mutual information with the selected variables in X_{S_j} (the redundancy) (P. Meyer et al., 2010). The rationale is that variables with redundant information to the most relevant variables are indirect links. Then the MRNET infer a GRN using a forward selection strategy, which leads to subset selection that is strongly conditioned by the first selected variables.

Using these three information-theoretic network inference techniques, which are available from the Bioconductor Minet package, we will evaluate the performance of the three meta-analysis approaches depicted in figure 3.5.

3.2 Batch effects and batch effects removal methods

In the "data merging" (DM) approach, data sets are integrated at the expression level into a unique dataset, from which GRNs are inferred (Wolfgang Huber et al., 2002; Belcastro et al., 2011; Adler et al., 2009). In the next subsections we detail three DM methods that are commonly used in the literature.

3.2.1 Normalization: BMC batch mean-centering (D1) and gene standardization z-score (D2)

The two normalization methods explained in the previous chapter, namely BMC batch mean-centering (called D1 in the thesis) (Sims et al., 2008) and gene standardization z-score (called D2) (Cheadle et al., 2003) will be compared with other meta-analysis methods for reconstructing GRNs.

3.2.2 Batch effects removal with COMBAT (D3)

Gene expression data sets mostly come from different platforms and laboratories, causing the so-called batch effects. It is now known that unwanted noise and unmodeled artifacts such as batch effects can dramatically reduce the accuracy of statistical inference in genomic experiments (Leek, 2014). Consequently, batch removal methods, like COMBAT (also known as Empirical Bayes) (Johnson, C. Li, and Rabinovic, 2007), is often used to detect and remove this inevitable variation. COMBAT, which was shown to outperform other commonly used batch removal methods in some specific scenarios (C. Chen et al., 2011; Bevilacqua et al., 2011), uses estimations for the LS (location-scale) parameters (e.x. mean and variance) for each gene independently (Lazar et al., 2013). The gene, afterwards, is adjusted to meet the estimated model. In this thesis, combining data sets using the COMBAT algorithm will be included for comparison and referred as method D3.

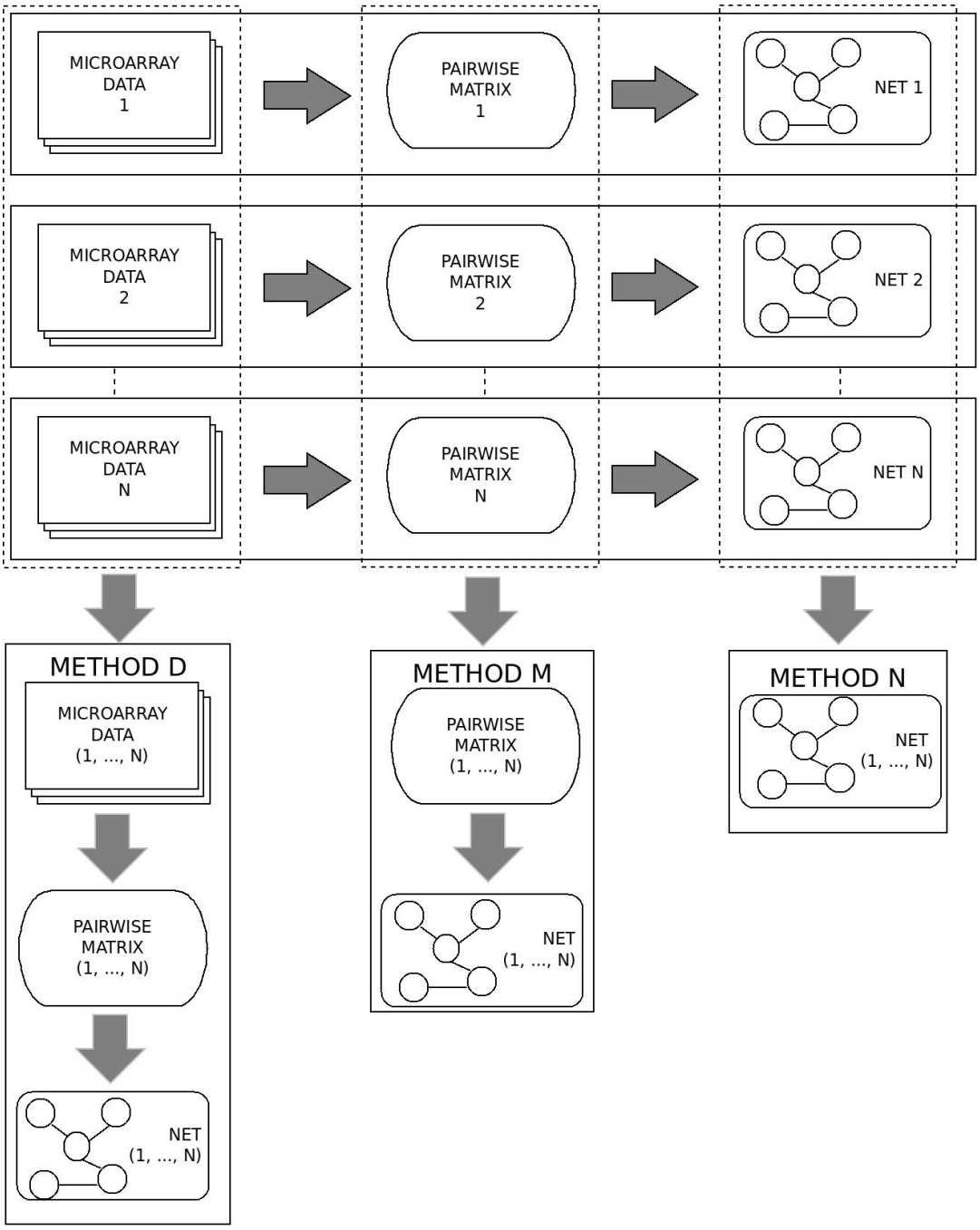

FIGURE 3.5: Meta-network strategies: Data Merging, Network Ensemble or Matrices of Coexpression based Aggregation.

3.3 Networks ensemble - N methods

As we presented in the previous subsection, one of the difficulties of the data-merging methods is how to handle the batch effects. Consequently, "networks ensemble" method (denoted with the letter (N) in the thesis) has been proposed as an alternative approach. In fact, by combining topologies of networks rather than data sets we are able to avoid dealing with batch effects implicitly. This method first constructs every single transcriptional networks independently before combining them to produce a so-called community network (Marbach et al., 2012).
Let e_{ij} be the edge between gene i and gene j in network n and $s_n(e_{ij})$ be the weight of

the edge. In the next subsections, we discuss three viable combinations of network transformation and aggregation.

3.3.1 RankSum method (N1)

Combining networks can consists in two distinct steps: transformation and aggregation (Bellot Pujalte et al., 2015). Indeed, before assembling networks, a network transformation step can be performed because it is common to observe networks that exhibit different distribution of edge weights (Bellot Pujalte et al., 2015). In this approach we replace the weight $s_n(e_{ij})$ in the network n by its rank $r_n(e_{ij})$ as such the most weighted edge gets the highest rank.

The RanSum method, which was introduced in (Marbach et al., 2012), is based on rank averaging: the final rank of the edge across N networks is computed by:

$$r(e_{ij}) = \sum_{n=1}^{N} r_n(e_{ij}) \tag{3.4}$$

3.3.2 Internal quality control index (N2)

In (Kang et al., 2012), six quantitative quality control measures have been proposed for the inclusion/exclusion of gene expression studies used for the meta-analysis. Among these measures, the internal quality control index will be included in this thesis, as method N2 for assembling networks. Let the similarity between two studies m and n be defined as

$$r_{mn} = spcor((t_n(e_{ij}); 1 \leq i \leq j \leq G), (t_m(e_{ij}); 1 \leq i \leq j \leq G)) \tag{3.5}$$

In which r_{mn} is the Spearmans rank correlation of the pairwise correlation structure between study m and n and G represents the total number of genes in the studies. The dissimilarity (or distance) between study m and n is defined as $d_{mn} = (1 - r_{mn})/2$. For a given study k, a weight -w_k will be granted as the fraction between the sum of distances between study k - D_k^* to all other studies and the sum of pairwise distances between all studies excluding the study k - $D_k^{\#}$:

$$w_k = \frac{D_k^*}{D_k^{\#}} \tag{3.6}$$

with

$$D_k^* = \{d_{kn}\}_{1 \leq n \leq N; n \neq k} \text{ and } D_k^{\#} = \{d_{mn}\}_{1 \leq m \neq n \leq N; m \neq k; n \neq k} \tag{3.7}$$

Afterwards, the weights over all studies are normalized such that the sum of all the weights is equal to 1. In the next step the weight of the edge between two variables (genes) X and Y is aggregated by the following equation:

$$\hat{e}_{IQC}(X;Y) = \frac{\sum_{k=1}^{N} w_k t_k(e_{XY})}{\sum_{k=1}^{N} w_k} \tag{3.8}$$

3.3.3 Median method (N3)

In (Hase et al., 2013) the median value was introduced for aggregating consensus networks. This method assigns the median value among N values representing the confidence score of a specific edge in N different networks.

$$a_M(e_{ij}) = median\{t_1(e_{ij}), ..., t_N(e_{ij})\} \tag{3.9}$$

3.4 Matrices of coexpression based aggregation approaches - M methods

Our new category of meta-analysis approaches (denoted with the letter (M) in this thesis) aggregates mutual information matrices rather than data or networks. The idea behind assembling pairwise matrices is that, although expression data typically shows high variability due to differences in technology, samples, labels, etc., pairwise dependency measures between genes should be much less variant (i.e. dependent variables, such as a regulating variable and its regulated counterpart, should remain dependent in every platform/experiment/dataset even if ranges of values differ greatly). Thus, to infer a network from various expression data, our approach consists in combining mutual information matrices (MIMs) estimated independently from each dataset. Then a GRN network is inferred from the aggregated MIMs. In the following subsections, we will demonstrate three feasible methods to assemble matrices of pairwise measure.

3.4.1 Random-effects model (M1)

It should be noted that the problem of combining MIMs across multiple data sets can be framed in the context of a meta-analysis of correlation coefficients (K. Wang, Narayanan, Zhong, Tompa, Eric E Schadt, et al., 2009). Hunter and Schmidt (F. L. Schmidt and Hunter, 2014) introduced a single random-effects method based on untransformed correlation coefficients, at which data sets are weighted simply by the sample sizes on which each effect size (the estimated MIM) is based. Our first weighting schema (method M1), described by equation 3.10, utilizes this random-effects method, but using MI instead of correlations.

$$\hat{I}_{RE}(X;Y) = \frac{\sum_{k=1}^{N} n_k I(X_k;Y_k)}{\sum_{k=1}^{N} n_k} \tag{3.10}$$

where $I(X_k;Y_k)$ is the MI between two variable X_k and Y_k in the study k and n_k is the number of samples of study k.

The idea is simply that effect sizes based on large samples will be more precise than those based on small samples.

3.4.2 Internal quality control index (M2)

Here, the internal quality control index measure was used again with some minor modifications. First, the similarity between two studies m and n was defined as

$$r_{mn} = spcor((I_{mij}; 1 \leq i \leq j \leq G), (I_{nij}; 1 \leq i \leq j \leq G)) \tag{3.11}$$

Then, the MI between two variables (genes) X and Y is aggregated by the following equation:

$$\hat{I}_{IQC}(X;Y) = \frac{\sum_{k=1}^{N} w_k I(X_k;Y_k)}{\sum_{k=1}^{N} w_k} \tag{3.12}$$

3.4.3 Median method (M3)

One of the major issue of M1 is that the quality of data sets used in meta-analysis is not explicitly taken into account. Indeed, inclusion of poor quality data sets is likely to weaken statistical power (Kang et al., 2012). Thus, an alternative schema for combining MIMs across heterogeneous studies namely method M3 can be proposed. Method M3 is explained by formula 3.13, in which the aggregated MI of a gene pair X and Y is the median value of all MI values between them across all studies.

$$\hat{I}_M(X,Y) = median(I(X_1,Y_1), I(X_2,Y_2), ...I(X_N,Y_N)) \tag{3.13}$$

A summary of all the methods is presented in Table 3.1. In the next chapter we present both *in silico* and biological experiments to evaluate the performance of the methods.

TABLE 3.1: Summary of meta-analysis methods used in the thesis

Method	Description
D1	Data Merging BMC
D2	Data Merging z-score
D3	Data Merging COMBAT
N1	Network Ensemble RankSum
N2	Network Ensemble Internal Quality Control Index
N3	Network Ensemble Median
M1	Matrices of Coexpression Aggregation Random-effects
M2	Matrices of Coexpression Aggregation IQCI

Chapter 4

Evaluation of methods with *in silico* and biological setups

The ultimate aim of the thesis is constructing a GRN for *Chlamydomonas reinhardtii*. And to this end we first need to evaluate the performance of the nine presented methods in Chapter 3 under different contexts. The best method then is selected in order to produce the CregNET.

4.1 Data

4.1.1 Simulated data sets

There are two tasks one needs to consider in order to validate networks: 1) defining a "gold standard" - which is a set of true regulations describing the underlying interaction network, 2) selecting quantitative measures to statistically assess the quality of inferred networks. Typically, the first task is addressed by collecting well-known regulations mined from literature with strong supporting evidences. However, those regulations just cover a small part of the underlying network and therefore cannot be an ideal reference network to thoroughly compare methods. Hence the latter approach is often completed by *in-silico* experiments.

In the work, *in silico* benchmarks are selected from every one of the 4 biological networks and artificially generated data sets coming from the Netbenchmark Bioconductor package (Bellot, Olsen, et al., 2015). The selected data sets are generated by two simulators namely GNW and SynTReN. The GNW simulator generates network structures by extracting parts of known real GRN structures from E. coli capturing several of their important structural properties while the SynTReN simulator generates the underlying networks by selecting sub-networks from *E. coli* and *Yeast* organisms (Bellot, Olsen, et al., 2015). The characteristics of the 4 biological networks are presented in more detail in Table 4.1.

TABLE 4.1: Networks used in the paper

Network	Name	Topology	Experiments	Genes	Edges
$SynTreN_{300}$	S1	E. coli	800	300	468
$SynTreN_{1000}$	S2	E. coli	1000	1000	4695
GNW_{1565}	G1	E. coli	1565	1565	7264
GNW_{2000}	G2	Yeast	2000	2000	10392

In the following step, each large data set will be split into 6 sub-data sets with a number of experiments ranging between 30 to 300 (a number chosen randomly in

order to simulate real case scenario) For example, in Figure 4.3, an original data set is split into 6 sub-data sets with the following number of samples: 50, 100, 150, 120, 70 and 190. Additionally, two extremely noisy studies are added, both with a large sample size for each (between 280 and 300). Those data sets allow to test the sensitivity of meta-network methods to data sets that should typically be excluded. Indeed, a few biological studies dating back to the beginning of the microarray technology have very little information and are typically excluded from meta-analysis studies.

4.1.2 *Saccharomyces cerevisiae*

The GRN of the *Saccharomyces cerevisiae* has been extensively studied (Kim et al., 2013) and in this work the current version of YeastNet(v3) (Kim et al., 2013) was used to test the DM and MA methods. Two microarray platforms based on Affymetrix chip designs GPL2529 and GPL90, which contain the majority of yeast gene expression profiling data, was targeted. However, we were only interested in gene expression series with no less than 8 samples. In total, we collected 44 series, which contained a total of 1,344 samples, from Gene Expression Omnibus (GEO) (Ron Edgar, Domrachev, and Lash, 2002). For gene expression data measured by microarray techniques, missing values are normally observed and hinder any downstream analysis . As a result, in the first step of preprocessing data, missing values were imputed using the KNN (Hastie et al., 2016). Furthermore, if multiple probes match a single gene, a similar method recommended by (X. Wang, Y. Lin, et al., 2012) was implemented, which is selecting the probe with the highest interquartile range (IQR). That is because larger IQR represents greater variability (and thus greater information content) in the data. In (X. Wang, Y. Lin, et al., 2012), two further sequential steps of gene filtering were then performed. In the first step, genes with very low gene expression that were identified with small average expression values across majority of studies were filtered out. Similarly, in the second step, non-informative (small variation) genes were removed by replacing mean intensity in the first step with standard deviation. However, in our research all genes that are present across the two platforms named above will remain for further analysis, reflecting the trade-off between increasing sample size and power versus decreasing the number of genes analyzed (Turnbull et al., 2012). In total, 5407 genes were retained for the downstream meta-analysis (covers 92 % of all yeast coding genes). Those genes were afterwards all log-transformed (base 2).

4.1.3 *Escherichia coli*

Similarly, *Escherichia coli* studies were also download from Gene Expression Omnibus, but we consider only studies using the Affymetrix E. coli Antisense Genome Array. Imposing a single microarray platform ensures that all data sets measure the same probesets (Silberberg et al., 2016). Probesets without annotations were excluded from the analysis, leaving a total of 4088 probesets, each corresponding to a specific gene (no gene was measured by multiple probesets) (Silberberg et al., 2016). Only data sets with no less than 8 samples were retained for downstream analysis. In total 29 data sets were collected, comprising nearly 900 samples. The latest version (version 9.4) of the transcriptional regulation in *Escherichia coli* K-12 was downloaded from RegulonDB (Gama-Castro et al., 2008) and used as the true network for validation.

4.1.4 Drosophila

The fruit fly *Drosophila melanogaster* provides an ideal model organism for the inference and study of functional regulatory networks in multicellular organisms (Bellot, Salembier, et al., 2019a). There exists a rich literature about regulatory relationships, which have resulted in small, but high-quality networks of known regulatory interactions such as REDfly (Bellot, Salembier, et al., 2019a). We have selected the data used in (Roy et al., 2010), since it provides a Heterogeneous scenario with networks of the same organism that comes from different kinds of data. There is a total of six networks that comes from both functional and physical regulatory interactions. For instance, experimental assays include high-throughput RNA sequencing (RNA-seq), capturing-small and large RNAs and splice variants and genomic DNA sequencing, measuring copy-number variation (Roy et al., 2010).

4.2 Evaluation metrics

In order to validate GRNs ones need to decide a suitable metric. A network inference problem can be seen as unsupervised learning problem where the inference algorithm acts like a binary classifier: for each pair of genes, the algorithm either predicts an edge or not depending on the confidence score of the edge (the higher the score the more certain there is an edge). And since in MI-based network inference algorithms a threshold is used in order to remove edges with low confidence score (weights) a confusion matrix can be computed for each value of threshold. For instance, a positive label predicted by the algorithm is considered either as a true positive (TP) or as a false positive (FP) depending on whether or not there is a corresponding edge in the reference network. Similarly, a negative label is counted as a true negative (TN) or a false negative (FN) when there is absence or presence of the corresponding edge in the underlying true network, respectively. Figure 4.1 gives an example of how the validation of a GRN as a binary classification task works. The network on the left is the true network, also known as gold standard. The network on the right is the inferred network. Edges colored in green of the inferred network are true positives while edges colored in pink are false positives.

A naive approach is thresholding the edges list of the network and calculating an average accuracy as a measure of the performance of an algorithm. However, this strategy is often misleading as GRNs are typically very sparse, and thus algorithms constantly predicting the absence of edges gain higher accuracy. Therefore, given the ground-truth knowledge of the underlying true network, traditional statistical error measures, such as F-score, AUCROC (Area Under the Receiver Operating Characteristic curve) or AUPR (Area Under the Precision-Recall curve) can be used to verify the quality of networks at the global-level (Emmert-Streib, Dehmer, and Haibe-Kains, 2014). Each of these measures is expressed as a single numerical value that integrates over all predicted interactions (Madhamshettiwar et al., 2012).

4.2.1 ROC curves and AUC

We define the false positive rate as:

$$FPR = \frac{FP}{TN + FP}$$

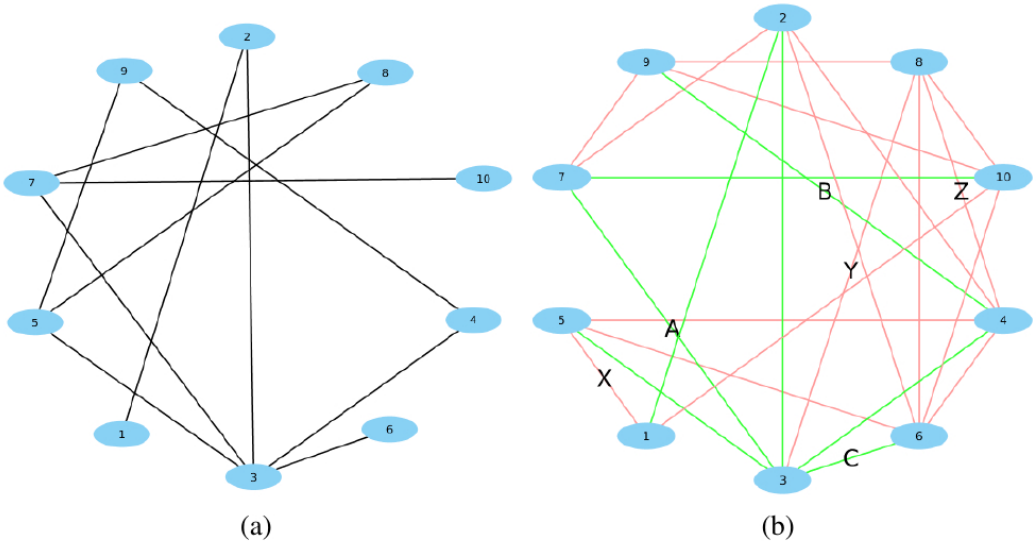

FIGURE 4.1: Example of validation of a GRN as a binary classification task. Image from Bellot Pujalte, 2017

, and the true positive rate as:

$$TPR = \frac{TP}{TP + FN}$$

, which is also known as recall or sensitivity.

A Receiver Operating Characteristic (ROC) curve is defined as the graphical plot of the TPR versus FPR for different values of the decision threshold by progressively lowering the threshold (J. Davis and Goadrich, 2006). Figure 4.2 illustrates how to make a ROC curve. On the left panel a ranked list of edges by weights is outputted by a GRN network inference algorithm. The true and false edges are colored in yellow and red respectively. By progressively changing the threshold we can create the corresponding ROC curve (in the middle) for the algorithm. The Receiver Operator Characteristic (ROC) Area Under the Curve (AUC, (Pintea and Moldovan, 2009)) can be then adopted as a global metric of performance for an algorithm. The AUC is in the range of $[0, 1]$, where one corresponds to perfect rank, 0.5 corresponds to random ordering. Thus, the closer the AUC is to 1 (and further away from 0.5) the better the overall performance of the network (Steele and Tucker, 2008).

4.2.2 PR curves and AUPRC

For validating GRNs, ROC curves, however, can present an overly optimistic view of the performance of an algorithm if there is a large skew in the class distribution (J. Davis and Goadrich, 2006), which is generally the case in network inference because of its spareness. Consequently, PR curves, which are often used in information retrieval, have been recommended as a preference measure of performance for GRNs(J. Davis and Goadrich, 2006). Let the precision defined as

$$p = \frac{TP}{TP + FP},$$

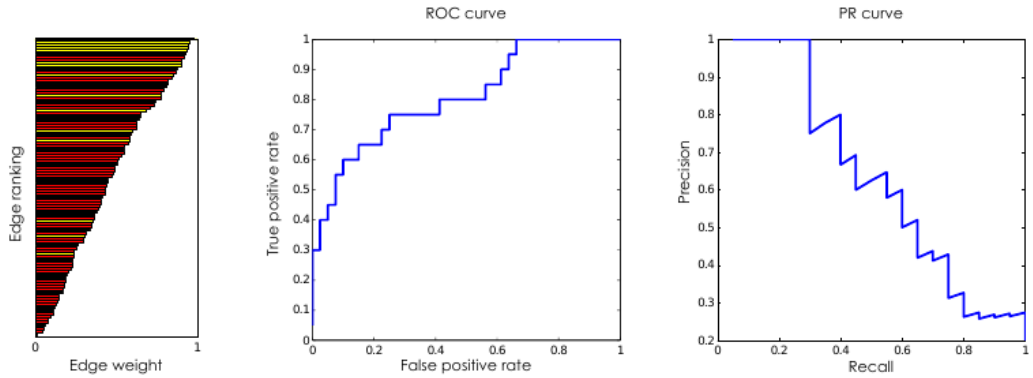

FIGURE 4.2: Evaluation of inferred networks with ROC curve and PR curve. Image from Sanguinetti et al., 2019

is the ability of the algorithm to correctly identify real edges among those classified as positive and the recall quantify

$$r = \frac{TP}{TP + FN},$$

also called true positive rate (TPR) is the fraction of real edges correctly identified.

An ideal algorithm will then have precision 1 for arbitrary recall between 0 and 1. To elucidate the effectiveness of an algorithm in handling the precision/recall trade-off, a plots of the precision versus recall for different values of the threshold (PR curves) is needed (see Figure 4.2 for illustration). And similar to the AUC, the Area Under the Precision Recall Curve (AUPRC, (J. Davis and Goadrich, 2006)) can be selected as the measure of global performance to summarize precision and recall information for varying . Universally the AUPRC is adopted as a measure to evaluate GRN inference algorithms (Sanguinetti et al., 2019). Thus, in the work, the AUPRC of each GRN is selected to report for each meta-analysis strategy.

4.3 Network prediction and validation with simulated data sets

4.3.1 Experimental setup

In order to make the network inference problem more challenging and realistic, noise and transformations of data are added. In particular, we define three levels of data-distortion: *i*) Level 1: An independent lognormal noise call "global" noise, with intensity between 20 and 50%, is added to the first 6 data sets. The standard deviation of this noise (σ_{Global}) is the same for the whole data set and is a percentage (κ_g%) of the mean variance of all the genes in the data set($\bar{\sigma_g}$). It is defined as follows: $\sigma_{Global;\kappa_g\%} = \bar{\sigma_g}\frac{U(0.8\kappa,1.2\kappa)}{100}$. *ii*) Level 2: In addition to the global noise, a normally distributed "local" noise with intensity also ranging between 20 and 50%, is added. This is an additive Gaussian noise with zero mean and a standard deviation ($\sigma_{Local(g)}$) that is around a percentage (κ%) of the gene standard deviation (σ_g). Therefore, the Signal-to-Noise-Ratio(SNR) of each gene is similar. The local noise standard deviation can be formulated as follows: $\sigma_{Local(g);\kappa\%} = \sigma_g\frac{U(0.8\kappa,1.2\kappa)}{100}$ where $U(a,b)$ is a uniform distribution between a and b. *iii*) Level 3: In addition to the

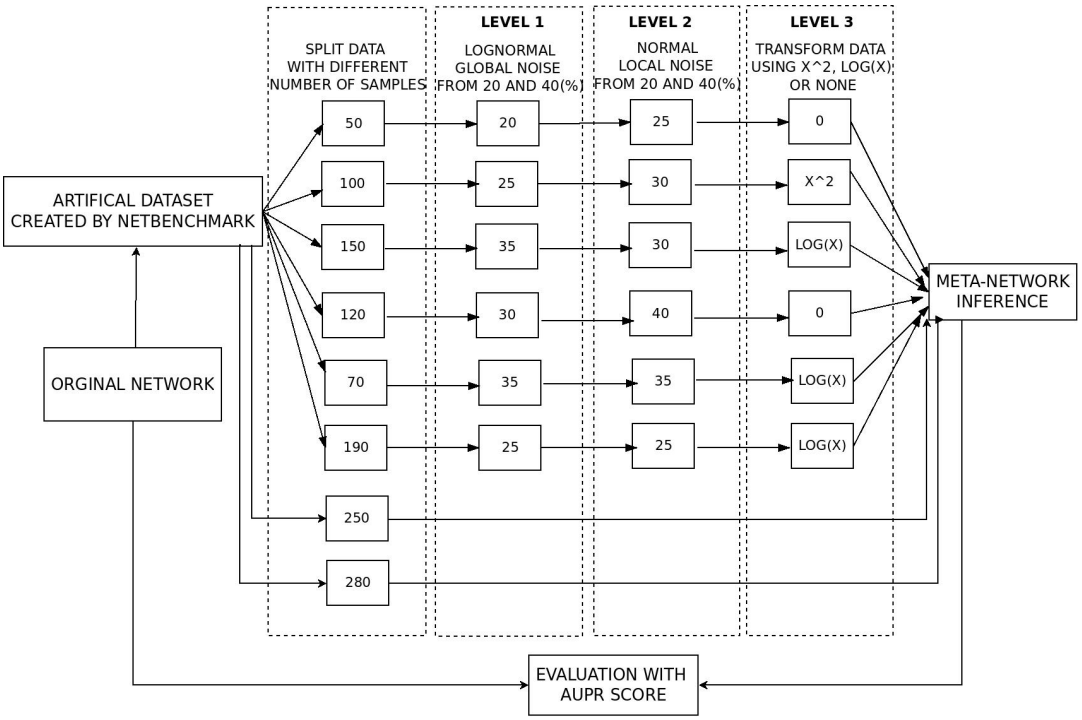

FIGURE 4.3: Framework for data collection, network prediction and validation

two previous noises, each sub-data set can be transformed using a randomly chosen non-linear transformation such as x^2 or $log(x)$. This random data transformation is not really meant to be realistic but rather to allow us to better assess the behaviour of each meta-method when faced with extreme distortion. It is worth emphasizing that the two non-informative studies remain unchanged across all experiments. A flowchart of this process is illustrated in figure 4.3.

The schema for network prediction and validation is also illustrated in figure4.3. Initially, all methods (three D, M and N, totaling nine) are used to construct a consensus GRN from the split data sets. All methods are assessed on 12 challenges (three levels of distortion for four data sets). Finally, the process is repeated for the three information-theoretic inference methods, hence totalling 36 challenges. This is done to make sure that our analysis is not method specific. The AUPR for each GRN is then selected to report for all methods in each challenge of the study.

Due to the randomization of various experimental parameters (noise intensity, number of samples), 10 repetitions are made. Finally, the average of the ten AUPR values, for each method on each challenge, is presented. Furthermore, in order to see how significantly better is the best method, a p-value using a Wilcoxon test (Cuzick, 1985) and adjusted, using a Bonferroni correction (Benjamini and Hochberg, 1995), between each approach and the best one is computed.

4.3.2 Experimental results

In this section, we present the experimental results of all presented methods for reconstructing GRNs from multiple expression data sets (Table 4.2). For the D family of methods, it can be observed that normalization using z-score transformation (D2) is better than BMC (D1). This conclusion is true for all three network inference algorithms used in this paper, namely MRNET, ARACNE and CLR. Another striking

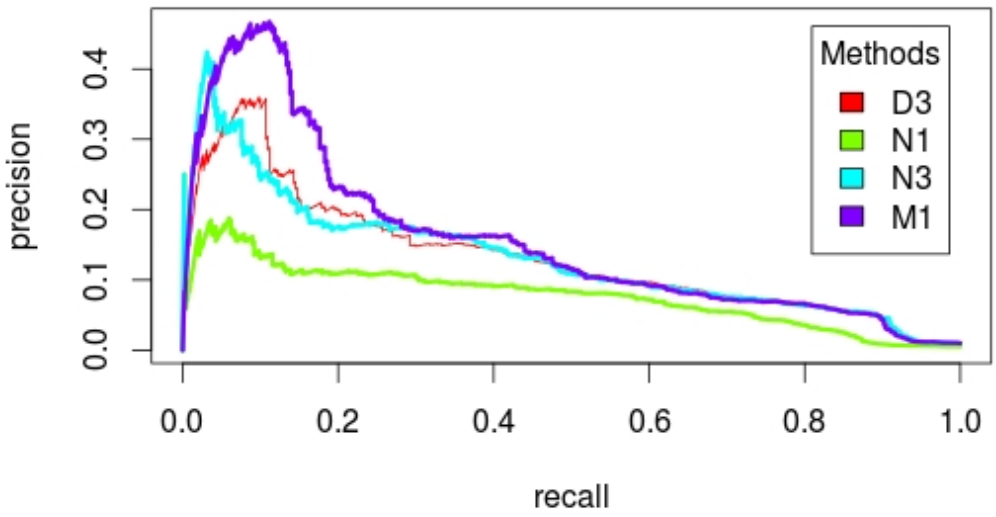

FIGURE 4.4: PR-Curves of method D3, N1, N3 and M1 on dataset S1 at level 1 of data distortion

feature is that batch effect removal methods like COMBAT (D3) is able to increase significantly the robustness of network inference algorithms. The results reinforce the idea that normalization alone can not remove batch effects, and therefore the removal of batch effects is essential when merging data sets. In the case of method N, N2 and N3 outperform N1 when MRNET or CLR used. However, in the case of using ARACNE, N1 is as good as N2 while poor results are observed for N3.

Interestingly, we can clearly observe that N2 outperforms all three D methods suggesting that assembling networks is better than merging data sets. This could be explained by the fact that gene expression values are very dissimilar in various experiments due to our simulated batch effects (i.e. data sets with different global and local noise). However, the particular combination CLR - D3 offers an exception to this observation. It also should be noted that assembling mutual information matrices (M methods) surpasses the two other well-known strategies (D and N) for all data sets under every different levels of distortion, in particular for MRNET (see figure 4.4 and figure 4.5) and CLR. Experimental results also show that MRNET benefits the most from meta-analysis and CLR appears to be the most robust. This suggests that while CLR might be a better strategy for analysing individual data sets, MRNET might be a better choice when multiple data sets are available. Although ARACNE appears to be much worse than the two other techniques, that is mainly due to a bad recall (though not visible with AUPR numbers, its precision remains quite competitive). Finally, in the M family of methods, it appears that combining MIM using random effect model (M1) is better than the two other strategies, the internal quality control index (M2) and the median method (M3).

4.3.3 Discussion

In the section, we proposed a framework for evaluating the different strategies for inferring GRNs from multiple expression data sets. To the best of our knowledge, this

TABLE 4.2: Area under PR-Curves (the higher the better) for 9 methods on 4 datasets with 3 levels of increasing data-distortion.

MRNET		D1	D2	D3	N1	N2	N3	M1	M2	M3
S1	Level 1	0.082	0.116	0.107	0.052	0.138	0.124	**0.141**	0.137	0.121
	Level 2	0.078	0.110	0.101	0.051	0.119	0.116	**0.120**	0.117	0.102
	Level 3	0.088	0.099	0.096	0.050	0.116	0.114	**0.120**	0.112	0.105
S2	Level 1	0.013	0.016	0.016	0.023	0.034	0.026	**0.046**	0.043	0.026
	Level 2	0.013	0.016	0.016	0.024	0.023	0.021	**0.026**	0.025	0.019
	Level 3	0.014	0.016	0.016	0.024	0.024	0.021	**0.027**	0.025	0.020
G1	Level 1	0.051	0.099	0.122	0.051	0.125	0.129	**0.156**	0.142	0.131
	Level 2	0.037	0.087	0.108	0.049	0.108	0.115	**0.138**	0.122	0.116
	Level 3	0.039	0.077	0.101	0.048	0.104	0.113	**0.133**	0.115	0.111
G2	Level 1	0.028	0.050	0.073	0.029	0.106	0.097	**0.131**	0.126	0.097
	Level 2	0.023	0.046	0.066	0.028	0.089	0.084	**0.116**	0.111	0.085
	Level 3	0.029	0.044	0.066	0.028	0.088	0.085	0.113	0.111	0.087
Mean		0.041	0.065	0.074	0.038	0.090	0.087	**0.106**	0.099	0.085
***p*-value**		.00195	.00195	.00195	.00195	.00195	.00195		.00195	.00195
ARACNE										
S1	Level 1	0.032	0.043	0.042	0.045	**0.101**	0.030	0.063	0.055	0.051
	Level 2	0.034	0.042	0.040	0.036	**0.080**	0.022	0.045	0.046	0.039
	Level 3	0.038	0.039	0.038	0.038	**0.083**	0.023	0.049	0.049	0.047
S2	Level 1	0.005	0.005	0.006	0.017	0.020	0.006	**0.025**	0.022	0.013
	Level 2	0.005	0.005	0.005	**0.015**	**0.015**	0.005	0.014	0.013	0.009
	Level 3	0.005	0.005	0.005	**0.015**	**0.015**	0.005	0.013	0.012	0.008
G1	Level 1	0.030	0.061	0.083	**0.126**	0.119	0.075	0.131	0.116	0.102
	Level 2	0.022	0.054	0.071	0.102	0.092	0.056	**0.105**	0.090	0.087
	Level 3	0.025	0.047	0.068	0.105	0.096	0.058	**0.109**	0.096	0.086
G2	Level 1	0.013	0.028	0.048	0.096	0.095	0.052	**0.124**	0.116	0.090
	Level 2	0.010	0.023	0.036	0.068	0.065	0.032	**0.081**	0.075	0.061
	Level 3	0.011	0.018	0.035	0.070	0.070	0.034	**0.087**	0.084	0.058
Mean		0.019	0.031	0.040	0.061	**0.071**	0.033	0.070	0.064	0.054
***p*-value**		.00195	.00195	.00195	.00977	1.0	.00195		.00586	.00195
CLR										
S1	Level 1	0.116	0.138	0.136	0.051	0.134	0.130	**0.137**	0.135	0.136
	Level 2	0.122	**0.140**	0.138	0.051	0.135	0.132	0.138	0.137	0.136
	Level 3	0.123	0.131	0.133	0.049	0.135	0.131	**0.138**	0.137	0.136
S2	Level 1	0.034	0.042	**0.043**	0.024	0.042	0.040	**0.043**	0.042	0.042
	Level 2	0.032	0.042	**0.043**	0.025	0.041	0.039	**0.043**	0.042	0.042
	Level 3	0.035	0.041	0.042	0.024	0.042	0.039	**0.043**	**0.043**	0.042
G1	Level 1	0.062	0.136	0.147	0.047	0.129	0.112	**0.147**	0.138	0.145
	Level 2	0.067	0.135	0.145	0.046	0.126	0.106	**0.142**	0.126	0.138
	Level 3	0.065	0.111	0.132	0.046	0.119	0.104	**0.135**	0.124	0.134
G2	Level 1	0.042	0.081	0.095	0.026	0.090	0.078	**0.105**	0.100	0.104
	Level 2	0.041	0.078	0.091	0.026	0.083	0.072	**0.097**	0.095	0.095
	Level 3	0.042	0.066	0.084	0.026	0.081	0.069	**0.094**	0.091	0.093
Mean		0.065	0.095	0.102	0.037	0.096	0.088	**0.105**	0.101	0.103
***p*-value**		.00195	.00195	.06446	.00195	.00195	.00195		.00195	.00195

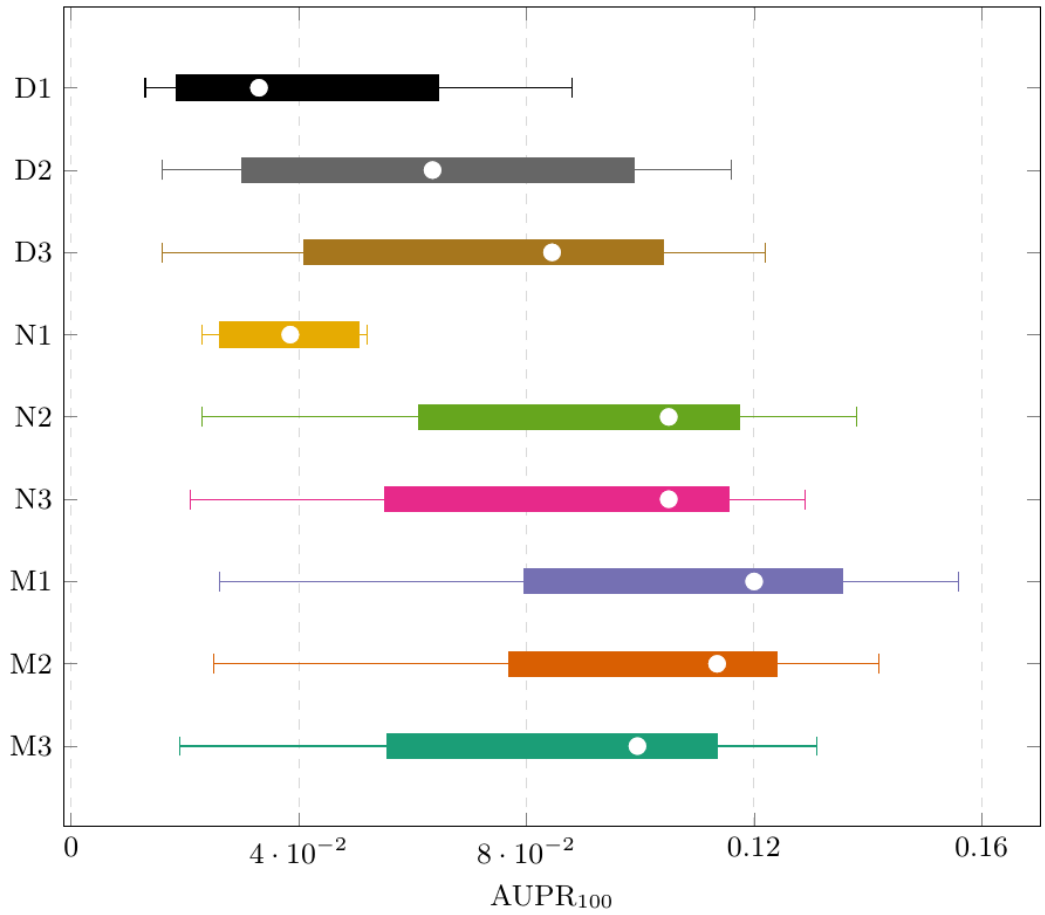

FIGURE 4.5: Boxplots for presented methods using MRNET

is the first systematic evaluation of the two state-of-the-art strategies for the problem of interest, namely "data merging" and "networks ensemble". Furthermore, we presented a new, but promising approach for methods based on coexpression matrices. Indeed, our set of experiments strongly suggest that assembling matrices of pairwise dependencies is a better strategy for network inference than the two commonly used ones. However, there exists many different methods of data and network assembly, as well as experimental conditions that have still to be tested in order to gain a complete understanding of the problem of meta-network inference. Moreover, as mentioned earlier, a large amount of under-exploited transcriptome data of model organisms is now available through public repositories. Thus, the use of the best strategy to reconstruct large-scale GRNs of these model organisms will be discussed in the next section.

4.4 Network prediction and validation with biological data sets

4.4.1 Results

Figure 4.6 demonstrates the AUPR score for all methods applied on different biological compendia. It could be seen that among D methods, batch effect removal using COMBAT is as effective as normalization using z-score, except for FlyNET

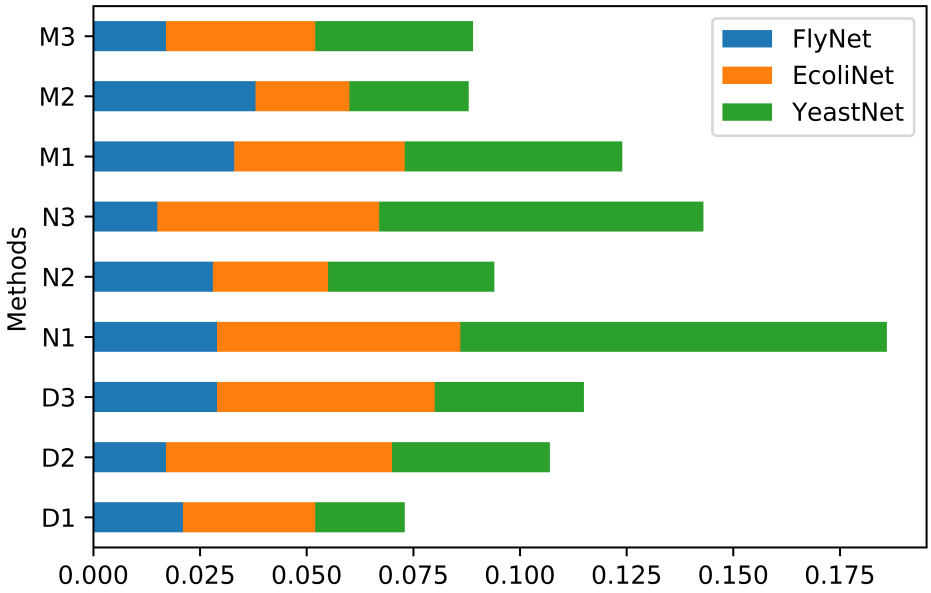

FIGURE 4.6: Bar plot of the AUPR scores of nine methods with biological data

where D3 outperforms D2. In the case of N methods, N1 and N3 are more robust than N2 while between M methods M1 is the most consistent. However, in terms of the best performance, the results illustrated in the table are very different from what we observed from synthetic data. For instance, in synthetic data, M family methods have better performance compared to those of N methods. However, N1 and N3 are the two best for EcoliNET and YeastNET while M2 is the best for FlyNET. Although N3 does not systematically yield the best performance on each and every network, it is competitive with other methods while preserving the scalability and thus N3 was selected to create the CregNET. Recently ChlamyNET has been presented by (Romero-campero et al., 2016) as the first co-expression network for *C. reinhardtii*. We will compare the performance of our CregNET with that of ChlamyNET in the next chapter.

Chapter 5

CregNET

The final goal of the thesis is to build a first GRN named CregNET for the model organism *C. reinhardtii*. In this chapter, first we present a pipeline to collect RNA-seq datasets of *C. reinhardtii* from Sequence Read Archive (SRA) (Benson et al., 2012). These are merely raw data and thus in the next step we need to quantify the data to create inputs for meta-analysis methods. A set of statistics measurement is then introduced to evaluate the performance of CregNET with ChlamyNET(Romero-campero et al., 2016) - the first co-expression network of *C. reinhardtii*. Experimental results strongly suggest that CregNET outperforms ChlamyNET for almost all measurement scores.

5.1 Pipeline

5.1.1 Essential Data Collection

In this study we used RNA-seq data of *C. reinhardtii* transcriptome publicly available at the Sequence Read Archive (SRA) (Benson et al., 2012), a database resource at the National Center for Biotechnology Information (NCBI) that stores more than 500 TeraBases of next-generation sequencing data. It is worth noting that The data within GEO/SRA is provided mostly in raw sequence form. This shortcoming makes it difficult to query and integrate this data at a global scale (Lachmann et al., 2018). To bridge the gap that currently exists between RNA-seq data generation and RNA-seq data processing, some pipelines have been developed. For example, (Lachmann et al., 2018) provides users with direct access to the data through a web-based user interface, while implementing a scalable and cost-effective solution for the raw data processing task. Furthermore, BioXpress (Wan et al., 2015) is a gene expression and cancer association database in which the expression levels are mapped to genes using RNA-seq data obtained from The Cancer Genome Atlas, International Cancer Genome Consortium, Expression Atlas and publications.

Admittedly, lack of efforts to integrate the expression profiles of *C. reinhardtii* genes obtained from RNA-seq prevent us from bettering our knowledge of the regulatory mechanisms for the organism. Thus, in the work we propose a pipeline for collecting and prepossessing transcriptomics data from RNA-seq technology for *C. reinhardtii*. Our pipeline is similar to (Lachmann et al., 2018) and depicted in Figure 5.1. Recently, SRAdb package (Zhu et al., 2013) has been developed to provide a convenient and integrated framework to query and access SRA metadata quickly and powerfully from within R and Bioconductor. Thus, in the first step of the pipeline, instead of GEOQuery package (Sean and Meltzer, 2007), we use SRAdb package for querying all available RNA-seq data associated with *C. reinhardtii* from the SRA database. It should be emphasized that when searching for the metadata, studies with less than 6 samples are not considered for further analysis, resulting in

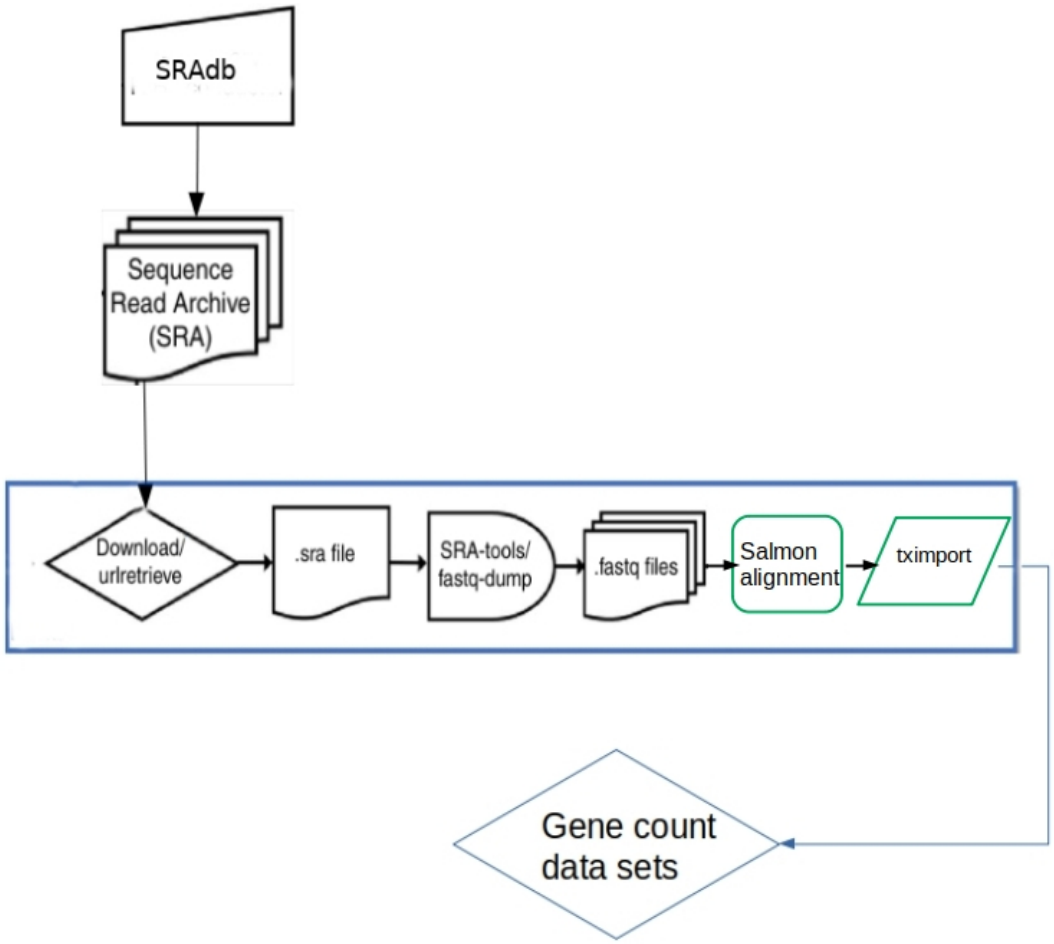

FIGURE 5.1: Schematic illustration of alignment pipeline workflow

20 series of 540 samples. The data provides a general overview of the *C. reinhardtii* transcriptome in numerous physiologically relevant context (Table 4.1). The data was then annotated using the available *C. reinhardtii* transcript version 5.5 downloaded from Phytozome (http://www.phytozome.net/) (Neupane et al., 2011), a web-based platform for green plant genomics.

The next step consists in all raw sequence reads in SRA format are downloaded using the *fastq-dump* from SRA Toolkit to detect single or paired reads file. This step resulted in more than 2.3 TB of raw data collected. Then, the SRA files is converted into FASTQ format and in case of a paired read file, the data is split into two FASTQ files.

5.1.2 Quantifying data

RNA-seq remains a great computational challenge: accurately aligning sequencing reads for inferring gene expression levels (D. C. Wu et al., 2018). The first step in quantifying transcription levels with RNA-seq is aligning reads, or pseudo-aligning parts of the read to transcripts. Specifically classical read-alignment tools start by aligning sequencing reads to a reference genome, at which gene expression levels can be inferred for the relevant genes. However, even for fast aligners this step can be time-consuming and computationally intensive (D. C. Wu et al., 2018). Over the last few years, alignment-free transcript quantification utilizing k-mer-based counting

SRA Number	Study	No of samples
DRP002675	RNA-seq for Chlamydomonas tar1-1 mutant	16
DRP003701	Genome-wide response to CO2 deficiency in Chlamydomonas reinhardtii revealed by RNA-seq analysis	27
ERP001997	Transcriptional profiling of Chlamydomonas reinhardtii	49
ERP005811	amit2	8
ERP006242	An evolutionarily conserved DOF-CONSTANS module controls photoperiodic signaling in plants	8
ERP011956	RNA-Seq analysis of parental (wild type) and dcl3 mutant lines of Chlamydomonas reinhardtii	12
SRP002284	RNA-seq analysis of the transcriptome from Sulfur Deprivation Chlamydomonas cells	8
SRP003630	Global Changes following N-deprivation in Chlamydomonas: Illumina sequencing	6
SRP010062	Three acyltransferases and a nitrogen responsive regulator are implicated in nitrogen starvation-induced triacylglycerol accumulation in Chlamydomonas	25
SRP031856	Systems-level analysis of N-starvation induced modifications of carbon metabolism in a Chlamydomonas starchless mutant	47
SRP037997	Chlamydomonas reinhardtii experimental evolution	8
SRP040308	Global Changes following N-deprivation and N-resupply in Chlamydomonas in the cht7 mutant and the wild-type: Illumina sequencing	24
SRP040659	Analysis of transcriptome of Chlamydomonas upon ClpP1 depletion and rapamycin treatment	42
SRP044681	Lineage-Specific Chromatin Signatures Reveal a Master Lipid Switch in Microalgae	142
SRP052618	Transcriptome response of Chlamydomonas reinhardtii exposed to inorganic or methylmercury	27
SRP058188	UV-B-induced gene expression changes in Chlamydomonas reinhardtii	8
SRP061735	Chlamydomonas diurnal transcriptome	56
SRP094014	The biosynthesis of nitrous oxide in the green algae Chlamydomonas reinhardtii	18
SRP094886	Global transcriptome analysis of heterodimeric homeobox-driven zygote developmental program in Chlamydomonas reinhardtii	8
SRP103581	Chlamydomonas reinhardtii strain:WT222+ Transcriptome or Gene expression	64

TABLE 5.1: Studies used in the work

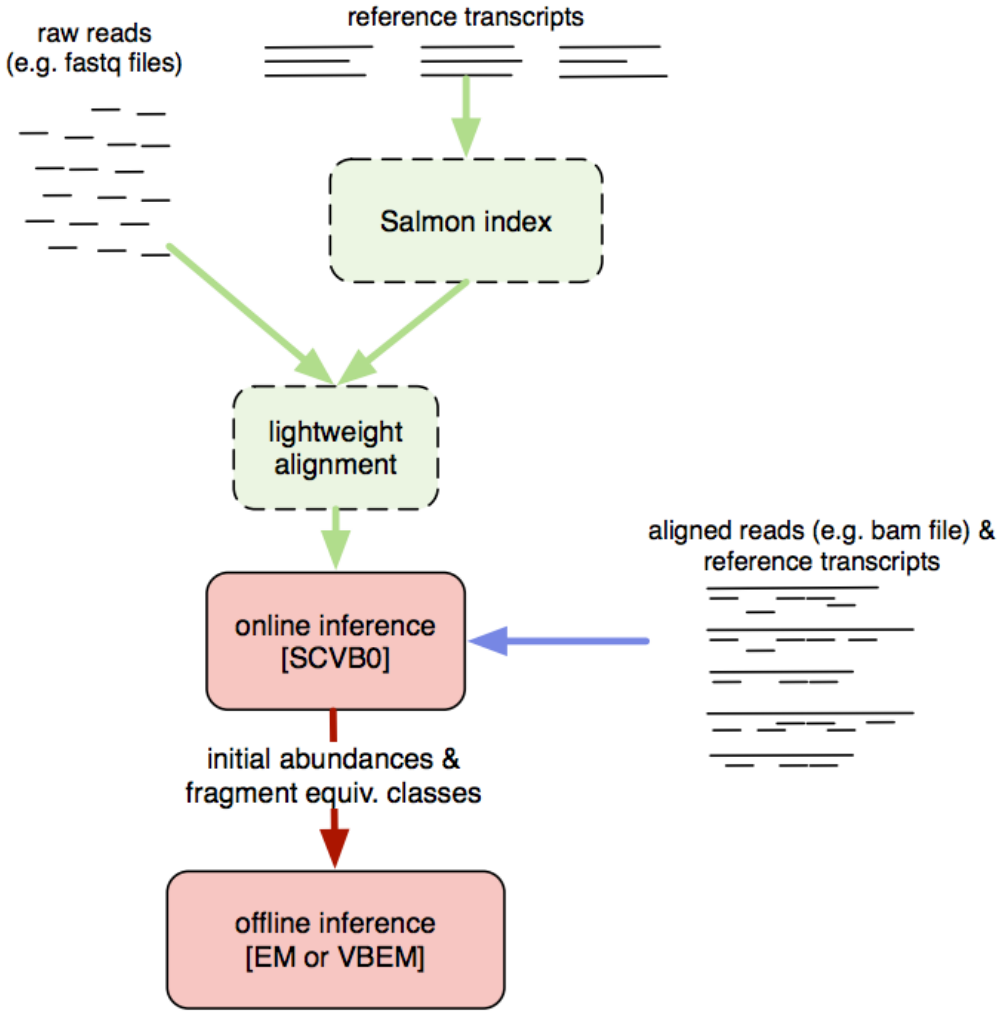

FIGURE 5.2: Overview of Salmon's method and component. Image from Patro et al., 2017

algorithms has been introduced (such as Salmon (Patro et al., 2017), Kallisto (Nicolas L. Bray et al., 2016), etc.) as a novel approach to replace classical aligners. Among those, *Salmon* (see Figure5.1) is the k-mer counting software that learns and corrects sequence-specific and GC biases on-the-fly, in addition to using quasi-mapping for further improvement in transcript quantification. Therefore, the core component of the pipeline in the thesis is using *Salmon* to estimate abundances without aligning reads, followed by the *tximport* package (**Soneson2016**) for assembling estimated count and offset matrices. This is also a pipeline recommended by (Michael I Love et al., 2015) for gene-level exploratory analysis and differential expression since the approach is newer and faster. The advantages of using the transcript abundance quantifiers in conjunction with tximport to produce gene-level count matrices and normalizing offsets are: 1) correction of any potential changes in gene length across samples (e.g., from differential isoform usage); 2) some of these methods are substantially faster and require less memory and disk usage compared to alignment-based methods; 3) it is possible to avoid discarding those fragments that align to multiple genes with homologous sequence (Michael I Love et al., 2015). The guideline and code for collecting and preprocessing the data can be found in Appendix A.

Another point of debate is which unit one should opt for as read counts need

to be properly normalized to extract meaningful expression estimates (Garber et al., 2011). Indeed, RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample (Garber et al., 2011) (Fig. 3a). Furthermore, the variability in the number of reads produced for each run causes fluctuations in the number of fragments mapped across samples (Garber et al., 2011)

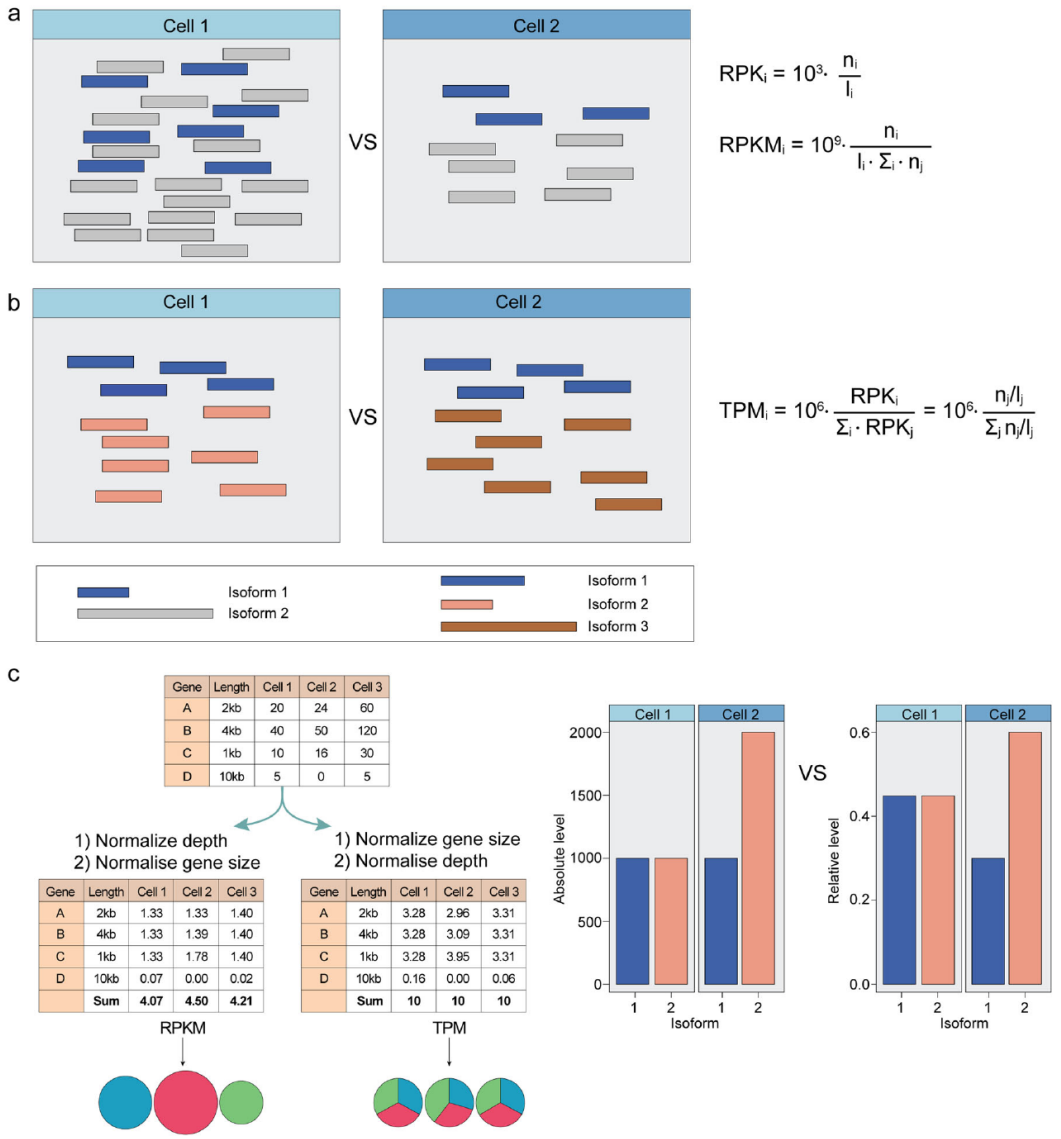

FIGURE 5.3: Methods for the quantification of expression. Image from Hwang, J. H. Lee, and Bang, 2018

The most commonly used approaches include R/FPKM (reads/fragments per kilobase per million reads) (Mortazavi et al., 2008) and TPM (transcripts per kilobase million) (Fig 5.3 a,b). RPKM can be calculated as

$$RPKM_i = 10^9 . \frac{n_i}{l_i . \sum_i . n_j}$$

. The only difference between RPKM and FPKM is that FPKM considers the read count in one of the aligned mates if paired-end sequencing is performed (Hwang, J. H. Lee, and Bang, 2018). TPM is a modification of RPKM in which the sum of

al TPMs in each sample is consistent across samples (Hwang, J. H. Lee, and Bang, 2018)

$$TPM_i = 10^6 . \frac{n_j / l_j}{\sum_i n_j / l_j}$$

This approach makes comparisons of mapped reads for each gene easier than PKM or FPKM based estimates because the sum of normalized reads in each sample is the same in TPM (Hwang, J. H. Lee, and Bang, 2018). Consequently, the traditional R/FPKM have been largely superseded by the TPM (Wagner, Kin, and Lynch, 2012) since the latter is more consistent across libraries (Michael I Love et al., 2015). Regardless, all these units attempt to "correct for" sequencing depth and feature length and thus do not reflect the influence of these on quantification uncertainty. In order to account for these aspects, most statistical tools for analysis of RNA-seq data operate instead on the count scale. Most of these tools were designed to be applied to simple read counts, and the degree to which their performance is affected by using fractional estimated counts resulting from portioning reads aligning to multiple transcripts is still an open question. The fact that the most common sequencing protocols provide reads that are much shorter than the average transcript implies that the observed read counts depends on a transcripts length as well as on its abundance; thus, simple counts are arguably less accurate measures than TPMs of the true abundance of RNA molecules from given genes (Michael I Love et al., 2015). The use of gene counts as input to statistical tools typically assumes that the length of the expressed part of a gene does not change across samples and thus its impact can be ignored for differential analysis. Furthermore, it was shown by (Michael I Love et al., 2015) that 1) gene-level estimation is considerably more accurate than transcript-level; 2) regardless of the level at which abundance estimation is done, inferences at the gene level are appealing in terms of robustness, statistical performance and interpretation; 3) taking advantage of transcript-level abundance estimates when defining or analyzing gene-level abundances leads to improved differential gene expression(DGE) results.

5.2 Network validation

Even though high-throughput technologies provides an abundance of biological data, be it of the realm of genomics, transcriptomics or bibliotomics, it remains a bioinformatics challenge to meaningfully transform those data into information. We are finding ourselves at the uprising edge of inferring GRNs for many model organisms. However, their validation is rather straightforward for GRN where extensive ground-truth values are known, but constitutes a major part of research, if there has not been any GRN established yet. And this has been the case for *C. reinhardtii*.

Prominent metabolic pathway databases like KEGG (Kanehisa, Minoru and Goto, 2000) provide only a part of the gold-standard and is outdated in the case for *C. reinhardtii*, as its current gene and protein refer to NCBI (**Barrett2013a**) version from 2007. In the meantime, many of those genes have been mapped to different genes and corresponding proteins do not longer exist. From internal communications we also know that certain pathways provided by KEGG for *C. reinhardtii* do not exist in the algae, but have been inferred from orthologs. There are only a few databases for model organisms like *E. coli* that provide metabolic or gene networks, which have been curated by experts. While there exist several repositories for *C. reinhardtii*, most prominently Phytozome (Neupane et al., 2011) there exist no GRN for *C.reinhardtii* so far. Therefore, it turned out to be a delicate challenge to validate our network. To this

end, we collected data and information from four sources: the protein-protein interaction database STRING (Szklarczyk et al., 2015), the Gene Ontology Consortium (Ashburner et al., 2000), KEGG (Kanehisa, Minoru and Goto, 2000) and PubMed (Doms and Schroeder, 2005). Those findings were used to validate our inferred network using methods similar to the modENCODE project (Roy et al., 2010).

5.2.1 ChlamyNET

In Chapter 3 we present what is a gene co-expression network (GCN) and how to construct a GCN using the correlation to measure the co-expression of a pair of genes. Recently ChlamyNET has been presented by (Romero-campero et al., 2016) as the first GCN for *C. reinhardtii*. Additionally, the authors developed a web-based tool, ChlamyNET, for the exploration of the *Chlamydomonas* transcriptome. ChlamyNET was constructed from 287 GigaBytes of collected RNA-seq data accounting for 50 samples and representing eight different genotypes under diverse physiological conditions. Moreover, ChlamyNET consists of 9171 genes exhibiting an overall of 139019 co-expression relationships. It was shown that ChlamyNET exhibits a scale-free and small world topology. Moreover, the authors identified nine gene clusters that capture the structure of the transcriptome under the analyzed conditions.

Nevertheless, the volume of data collected accounts for only a small part of all RNA-seq data could be found for *C. reinhardtii* from SRA repository. More importantly, the problem of batch effects accompanied with integrating transcriptomics data is not well addressed by the authors. By comparison, in our study we have collected 2.3 TB of raw data from 20 different series consisting of 540 samples under numerous physiological conditions (see Table 5.1) to create CregNET.

5.2.2 Permuted graph

Formally let $G = (V, E)$ be a graph where $V = \{v_1, v_2, ..., v_n\}$ denotes its node set and E a subset of $V \times V$ representing the edges. The permutation of a graph G is a swapping of the node set, while preserving the edge sets' topology. Figure 5.4 shows an example of permutation graphs of six vertices. Let $\sigma : V \to V$ be a permutation of V, i.e. $\sigma(v_1, v_2, ..., v_n) = (\sigma(v_1), \sigma(v_2), ..., \sigma(v_n))$, where $\sigma(v_i) = v_j$ is unique and all pairs $(\sigma(v_i), \sigma(v_j))$, $i \neq j$ are distinct. An edge $e = (v_i, v_j)$ under a permutation is denoted by $\sigma(e) = (\sigma(v_i), \sigma(v_j))$. Thus, an edge is mapped onto an edge, and the topological edge structure is preserved, while only the labeling of the nodes is permuted. We denote the permuted graph by $\sigma(G)$.

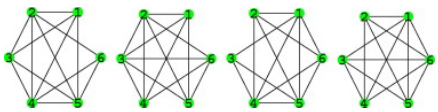

FIGURE 5.4: Example of permutation graphs of six vertices. Image from Seoud and Mahran, 2012

5.2.3 Network enrichment and scoring

A network scoring function S is any function taking a graph as input and returning a real value larger or equal to zero. If a higher score reflects a higher valued network, then the underlying scoring function is called an increasing score, a decreasing score

otherwise. The enrichment of a graph G wrt. a permutation σ is defined by $\frac{S(G)}{S(\sigma(G))}$ for increasing score, and $\frac{S(\sigma(G))}{S(G)}$ for a decreasing score. We call a network to be enriched wrt. σ, if its enrichment score > 1 for an increasing score, and < 1 for a decreased score.

The underlying idea of an enriched graph is that a good network is "truer" than its permuted, and somehow randomized, version. This is simply stating that a graphs score is higher than the score achieved by the graph permuted wrt. σ. On the other side, a score < 1 tells that the permuted graph is higher valued than the original graph. An equivalent meaning is introduced for a decreasing score by reversing the sign.

Even though a graph is enriched wrt. one permutation σ_i it might not be enriched wrt. a permutation σ_j. However, one expects for a network with good topology that its score is better than most of its permuted versions. Therefore, the enrichment of a graph G is defined as

$$\frac{1}{|\Pi|} \sum_{\sigma \epsilon \Pi} \frac{S(G)}{S(\sigma(G))} \tag{5.1}$$

where Π is the set of all possible permutations for G. The enrichment is hence the averaged enrichment value wrt. all permutations $\sigma \epsilon \Pi$. The number of permutations for a graph of size n is $n!$ and thus computationally not feasible for large graphs. In order to get an approximation, we took 100 randomly drawn permutations. In order to validate the accuracy of our network prediction, we considered the enrichment of the $top - k$ weighted edges at several percentages of the total network size, where the size of a network is defined as the number of its edges. However, the network of CREG-N3 and ChlamyNET have different sizes and provide also a different number of genes, i.e. ChlamyNET provides 13,446 genes, while our CREG-N3 has 17,441 genes.

In order to compare them consistently, we took the size of the smallest network as reference, i.e. GREG-N3 with a size of 72,532,463. However, the use of a single scoring function is prone to be biased towards a certain type of network. Therefore, it is indispensable to use several scores relying on different aspects of the graph, which we will treat in the next subsections.

5.2.4 Protein-Protein-Interaction-Network (PPI-network)

A PPI-network is a graph $G_{PPI} = (V, E)$ whose nodes V representing proteins and edges E representing a form of high specificity established between the corresponding proteins. The STRING database (Szklarczyk et al., 2015) provides a publicly available PPI-network of *C. reinhardtii*, containing 4,321,366 interactions of 13,307 proteins taken from all organelles. Each edge e comes with a confidence weight w_e, provided by the STRING community. The STRING annotation refers to the current NCBI annotation , wherefore each protein is uniquely mapped to one gene, and no gene exhibits any spliced forms. However, since the NCBI annotation dates back to 2007, we subsequently had to map the genes to the current genome annotation v5.5.

If there is a well-established protein-protein interaction in G_k, then the gene-gene network might have a corresponding gene-gene interaction, and vice versa. Of course, this is not a 1-1 mapping, however, we assume that a good network inference will not only have captured the top protein-protein interaction, but weighted them high too. Therefore, let G_k be the network consisting only of its $top - k$ edges,

then we define the PPI score by

$$S_{PPI}(G_k) = |G_k \cap G_{PPI}|\,, \tag{5.2}$$

where the cardinality of the intersection of two graphs G and H is

$$|G \cap H| = |\{E(G) \cap E(H) : V(G) \cap V(H)\}| \tag{5.3}$$

that is the number of edges the two graphs have in common on their shared node set. This score was computed for ChlamyNET and CREG-N3 for several different k (see Network enrichment and scoring for the choice of k).

5.2.5 Gene ontology

The Gene Ontology (GO, (Ashburner et al., 2000)) project is a major bioinformatics initiative to develop a computational representation of our evolving knowledge of how genes encode biological functions at the molecular, cellular and tissue system levels. GO provides GO-terms, which are machine readable, describing the functions of specific genes. Using GO-terms for the validation of associated genes underlies the hypothesis that specific inferred gene-gene interactions have many of the corresponding GO-terms in common. In information retrieval and text mining one common approach to compare the similarity between two documents is the Jaccard index. It is defined by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.4}$$

for two sets A and B, where in our scenario A and B will be the set of GO-terms of two genes G_A and G_B. The Jaccard index will be equal to 1 if the common GO-terms of G_A and G_B are exactly the union of their GO-terms. If the intersection is empty, then it will be 0.

We have retrieved 26,216 GO-terms (3,593 unique ones) for 7579 genes from NCBI and computed for each of those gene pairs its Jaccard index, resulting in a 7579×7579 gene-gene GO-matrix G_{GO}.

It was found that for each gene 3.45 Go-terms were assigned on average, a more confluence figuring of the GO-term distribution can be seen in Fig 5.5 Let G_k the network consisting only of its $top - k$ edges, then we define the GO score by

$$S_{GO}(G_k) = |G_k \cap G_{G0}| \tag{5.5}$$

Similar to the PPI score, the GO score reflects the amount of rediscovered gene-gene pairs that are meaningful in terms of GO annotation. The higher the Jaccard-index of two genes, the more plausible that they share a biological function, which should be reflected by sharing an edge in the network with a high score.

5.2.6 KEGG ontology

In the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, Minoru and Goto, 2000), molecular-level functions are stored in the KO (KEGG Orthology) database and associated with ortholog groups in order to enable extension of experimental evidence in a specific organism to other organisms. Genome annotation in KEGG is an ortholog annotation, assigning KO identifiers (K numbers) to individual genes in the GENES database. In general KO grouping of functional orthologs is defined in the context of KEGG molecular networks (KEGG pathway

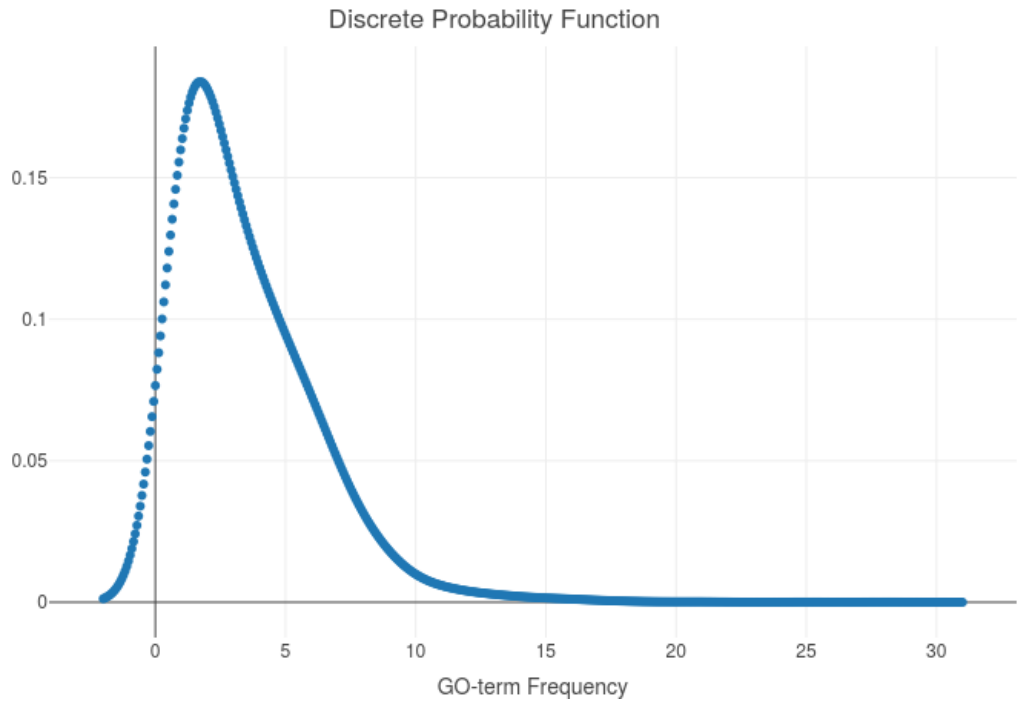

FIGURE 5.5: Frequency of GO-terms per gene

maps, BRITE hierarchies and KEGG modules), which are in fact represented as networks of nodes identified by K numbers. The fact that functional information is associated with ortholog groups is a unique aspect of the KEGG resource (Kanehisa, Minoru and Goto, 2000).

We have extracted 4910 KO-terms (3269 unique ones) for 4907 genes and equivalently to the GO-terms, we used the Jaccard-index to form a 4907×4907 gene-gene matrix. The KO-score is analogously computed to the GO-score

5.2.7 Literature

PUBMED (Doms and Schroeder, 2005), the central repository for references to life science articles, gives rise to a total number of 2816 abstracts about *C. reinhardtii* at the time when the research was conducted. We downloaded all available abstracts counted for each abstract, including its title, the occurrences of all gene entities, i.e. symbols, gene names and aliases. The basis of those entities was formed by the NCBI annotation for Chlamydomonas (e.g. CHLREDRAFT_106571) and the standard gene symbols, names and aliases. We have neither explored the articles for different genome annotations as jgi | Chlre3 | 107200 (v3.1), g17.t (v5.3) or Cre01.g000700 (v5.5), which all refer to the same gene, nor did we crawl the articles for protein entities. Exclude ambiguities and handle genes that have been mapped to different transcript or proteins due to a refined and updated genome annotation, is a research in its own right.

The standard and widely used annotation resulted in a high dimensional vector of 13,046 dimensions, reflecting 13,046 distinct and unique gene entities, for each document. Each gene-gene pair occurring in a document, i.e. two entries not equal to zero in one of the high dimensional vectors, is considered as a potential gene-gene interaction.

The frequency of each interaction is counted and stored as weight. We discovered this way 15,236 interactions, where 70% are singular interactions, i.e. have

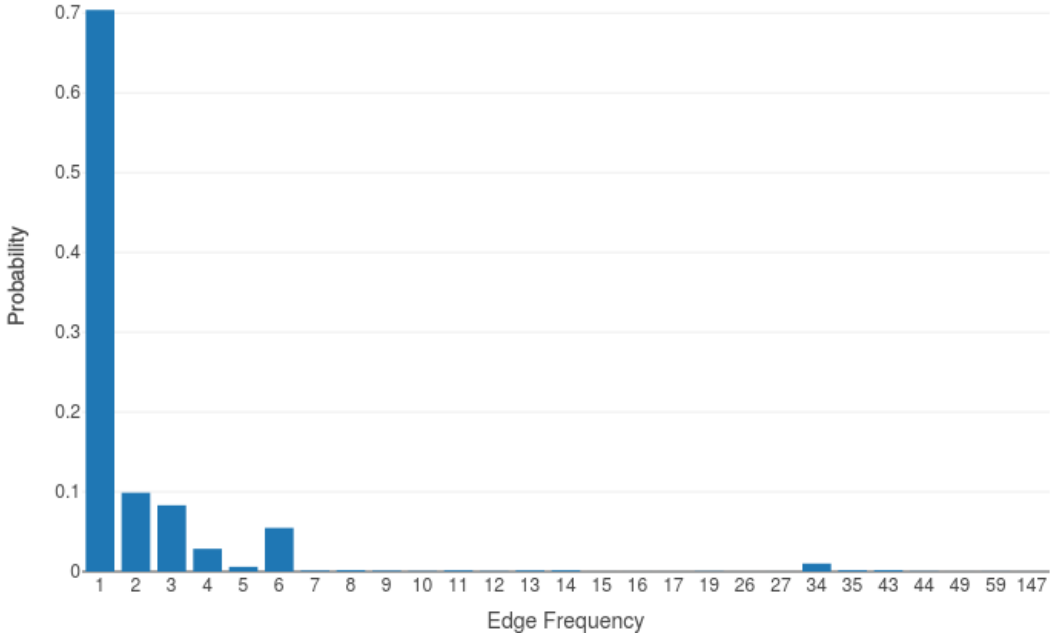

FIGURE 5.6: Gene-Gene interaction found in the literature

been found only once, while there have been also a significant amount of interactions found with a frequency of 34 occurrences. The latter consists mainly of membrane, antiporter or transporter genes, comprising the ATM, MTP and CAX gene family. The resulting edge-list was transformed into a matrix G_{Lit} and compared to the edges discovered by the network

$$S_{Lit}(G_k) = |G_k \cap G_{Lit}|$$

5.2.8 Average Shortest Path (ASP)

The power of gene A directly controlling gene B is related to their distance on a regulatory path. The more distant they are, the less influence A exhibits on B. A GRN reflects those various influence strengths by being a trade-off between a fully connected gene-gene network and a too short-ranged dependency network. In particular, the predicted length of the shortest path between A and B, i.e. the minimal amount of nodes necessary to reach B from A or vice versa, for two closely regulated genes should be small compared to a randomly assigned graph between them. Hence, the average shortest path of all genes should be smaller than in a random graph.

However, a complete gene-gene network (all gene pairs have a shortest path of exactly 1) has always the smallest average shortest path, namely 1. Nevertheless, if the inference of a complete gene-gene network is good, then restricting the network to the top weighted edges should still deliver good results. Contrary to the previous scores, the ASP-score is a decreasing score and therefore the enrichment can be computed with

$$\frac{S_{ASP}(\sigma(G))}{S_{ASP}(G)}$$

where the S_{ASP} is defined as the average of all shortest paths on $|G_k \cap G_{List}|$ i.e. the intersection with the graph derived from literature.

5.2.9 PPi-Triangles

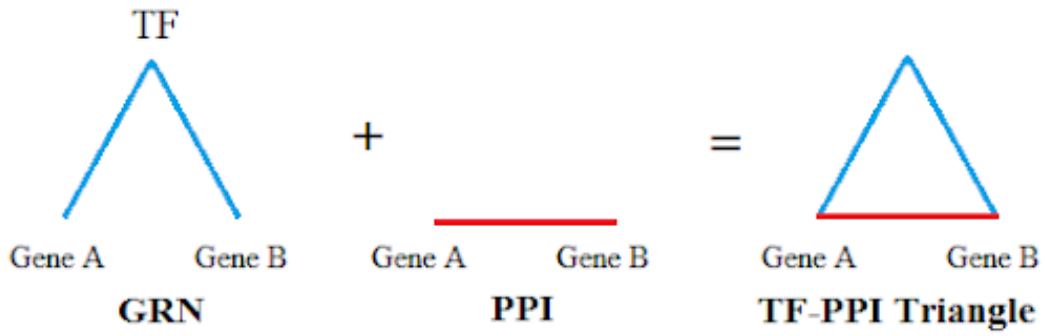

FIGURE 5.7: Left : TF-genes in the GRN, Middle : A protein interaction in PPi, Right: Combined triangle of GRN and PPI

A PPi-network reflects specificity between proteins, in particular interactions between them. An interaction of proteins A and B might be explained by a transcription factor regulating the genes corresponding to protein A and B. On the contrary, an inferred co-regulation of two genes A and B should be reflected by an interaction of their proteins in the PPi-network. The number of correctly assumed co-regulated genes in a GRN can be assessed by the number of closed TF-PPi triangles. That is the completion of GRN edge pairs sharing a TF by an edge from the PPi network (see Fig 5.7).

We have downloaded the most recent list of TF from the Plant Transcription Factor Database (Jingchu Luo et al., 2013), consisting of 234 genes comprising transcriptions factors and transcription regulators. They have been transformed into a TF-gene matrix. Counting the number of TF-PPi triangles means to count the true checks for every gene pair of the same TF if there is a PPi connection. So, a TF-gene matrix, where each row contains the genes regulated by a TF, is a suitable representation. Each gene pair of a TF row has to be checked if they can be completed by a PPi in the PPi network.

On the other side, as stated before, a PPi has not be present in the GRN, being the reason that one score the amount of PPis that were rediscovered by the GRN, but the closed TF-PPi triangles in the average for each TF. This counterbalances a complete network, where all gene pairs seem to be regulated by a TF.

5.3 Experimental results

In this section, we present the experimental results of all validation scores explained above computed for CregNET and ChlamyNET. And in order to validate the robustness of our network, we considered the enrichment of the $top - k$ weighted edges at several percentages of the total network size.

In Table 5.2 when top 1% of all edges were selected, CREG_N3 outperforms all the others in term of PPI, GO, KO, Literature and Averaged Shortest Path (ASP) scores. It is also interesting to note that CREG_N2 is also better than ChalmyNET with those scores. For PPI-Tri score, CREG_N2 is the top performer followed by ChamyNET, CREG_M1 and then CREG_N2. A similar pattern is observed in Table 5.3 measuring the scores for the top 5% of edges for each network. For instance, CREG_N3 leads with PPI, GO, KO, Literature and PPI-Tri scores while CREG_N2 is

the best with respect to ASP score. In general, CREG_N3 is the best and CREG_N2 is the second best network.

TABLE 5.2: Top 1%

	PPI	GO	KO	Literature	PPI-Tri	ASP
CREG_M1	1.033058	1.272837	2.851852	2.250000	1.2684910	0.9946715
CREG_N2	1.499628	3.098478	11.970588	4.808824	**1.9122458**	1.0912617
CREG_N3	**1.778624**	**5.712112**	**19.584906**	**8.666667**	1.2557466	**1.1275794**
ChlamyNET	1.005364	1.640209	5.250000	2.794872	1.4653919	1.0214104

A same conclusion can be reached by Table 5.4 when we includes top 10% of all weighted edges. This time, CREG_N3 is still the best network followed by CREG_N2. Even though ChlamyNET is better than CREG_M1, it is by far outperformed by CREG_N3 and CREG_N2.

TABLE 5.3: Top 5%

	PPI	GO	KO	Literature	PPI-Tri	ASP
CREG_M1	1.044536	1.179604	1.973494	1.462985	1.014224	0.9926273
CREG_N2	1.335464	1.713317	3.676115	2.633156	1.5974866	**1.0378010**
CREG_N3	**1.484548**	**2.402395**	**5.241573**	**3.149701**	**1.7151664**	0.8825367
ChlamyNET	1.039675	1.216215	2.120690	1.539602	1.2235342	0.9948883

TABLE 5.4: Top 10%

	PPI	GO	KO	Literature	PPI-Tri	ASP
CREG_M1	1.124379	1.210021	1.627936	1.376795	0.8163115	0.9950922
CREG_N2	1.409908	1.981277	4.841035	3.019938	**1.637731**	**1.0446838**
CREG_N3	**1.517054**	**3.081725**	**8.800000**	**3.695327**	1.669193	**1.0468796**
ChlamyNET	1.030407	1.310526	2.948473	1.690250	1.152630	0.9970119

In order to have a better view of the performance of CREG_N3 and ChlamyNET we compare their statistical scores with the variation of number of edges. Fig 5.8 shows that CREG_N3 is approximately 1.5 times better than ChlamyNET. Moreover, this ratio is higher when less edges are taken into consideration explaining the fact that CREG_N3 is better recovering the most significant edges. A similar figure for GO Enrichment is also observed with Fig 5.9 when CREG_N3 is at least 1.5 times better than ChlamyNET. Moreover Fig 5.10 and Fig 5.11 shows a same pattern when CREG_N3 is significantly better than ChlamyNET with different variation of $top - k$ weighted edges.

However, when compared with PPI-Tri Enrichment score in Fig 5.12 CREG_N3 is still better but only when k is large enough (more than 1% of edges). And it is not obvious by Fig 5.13 when with some specific range of values of k CREG_N3 has a better scores and vice versa. In conclusion, the figures strongly suggest that CREG_N3 outperforms ChlamyNET with almost all statistical scores under all different contexts, i.e. the number of edges considered.

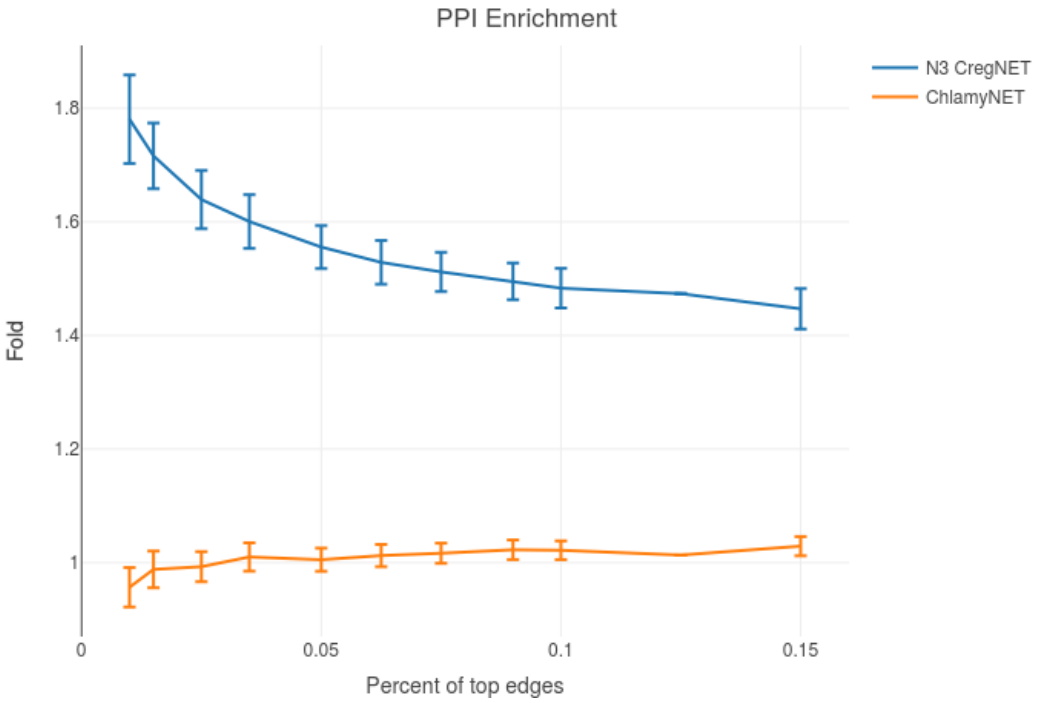

FIGURE 5.8: Fold change of PPI enrichment score with regard to random networks of CREG_N3 and ChlamyNET

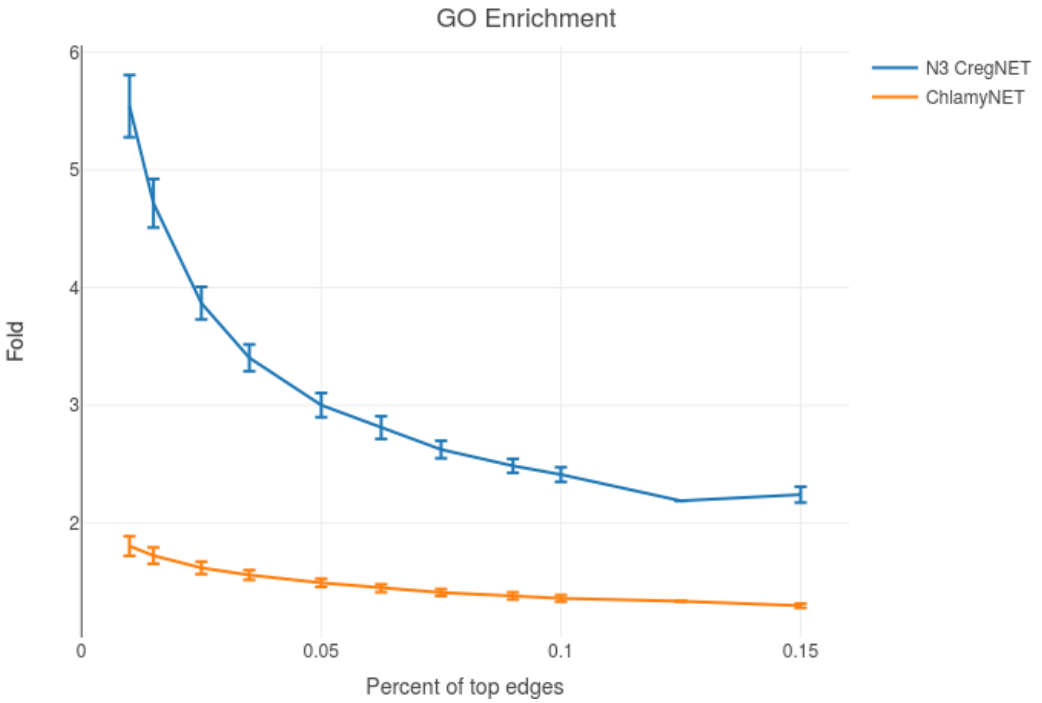

FIGURE 5.9: Fold change of GO enrichment score with regard to random networks of CREG_N3 and ChlamyNET

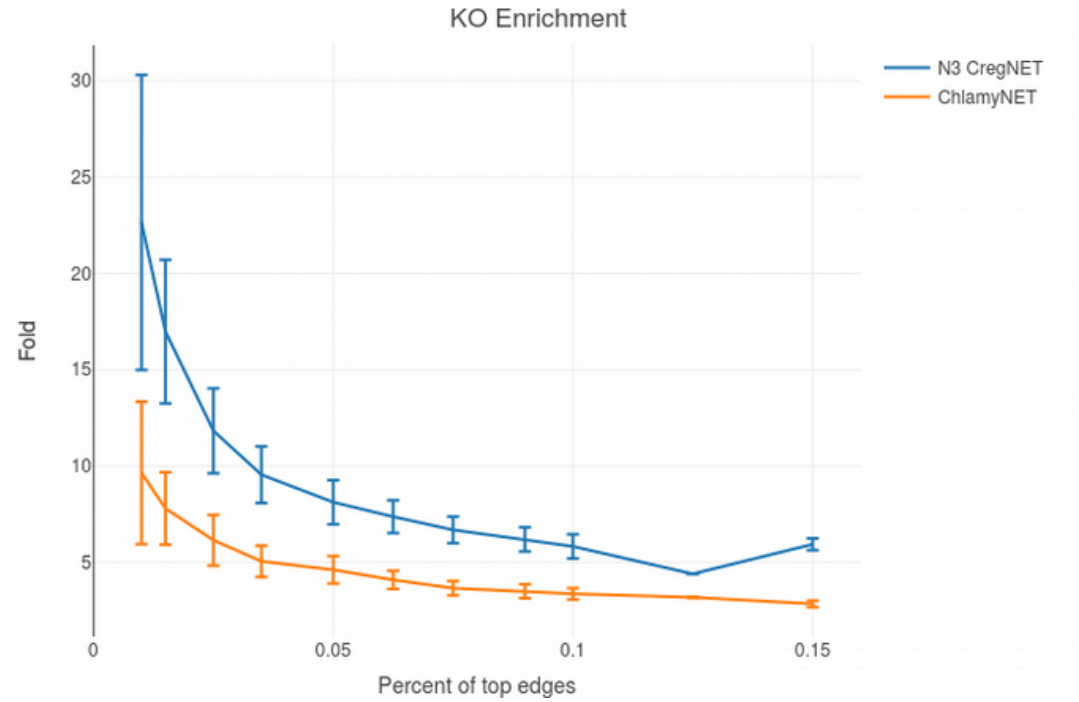

FIGURE 5.10: Fold change of KO enrichment score with regard to random networks of CREG_N3 and ChlamyNET

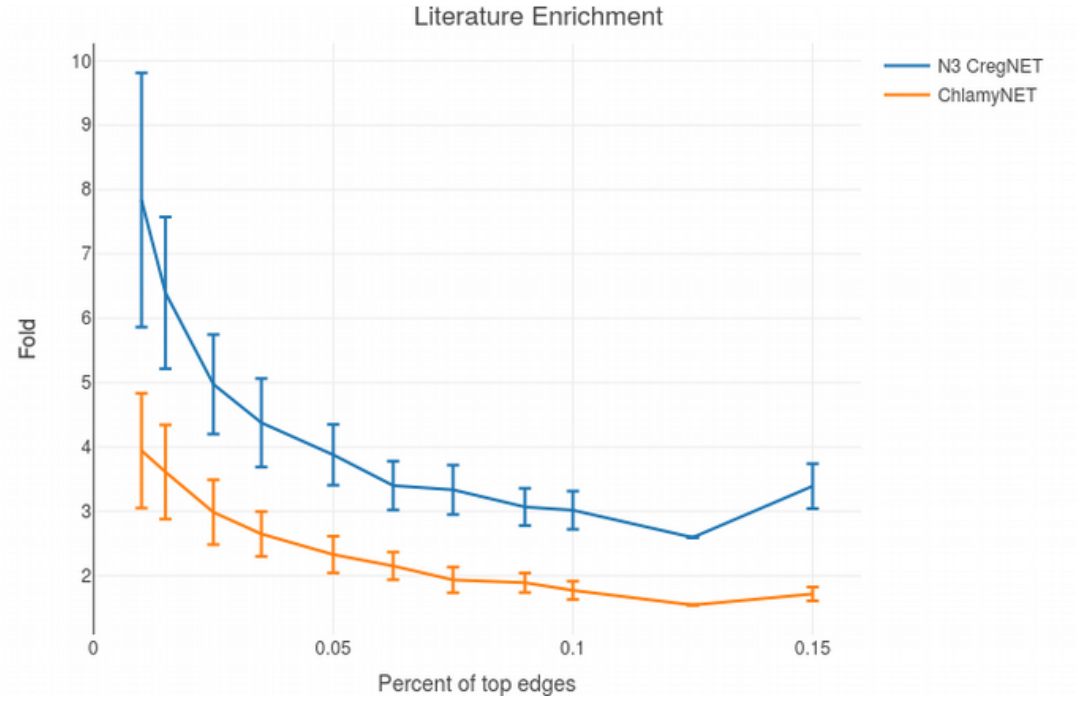

FIGURE 5.11: Fold change of Literature enrichment score with regard to random networks of CREG_N3 and ChlamyNET

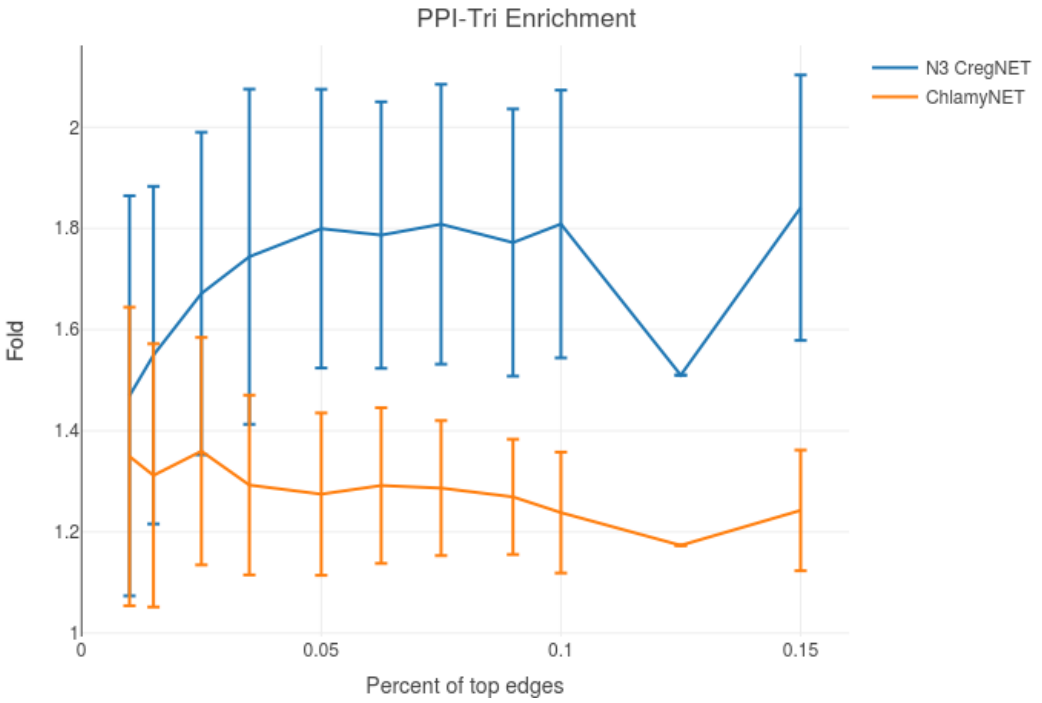

FIGURE 5.12: Fold change of PPI-Triangle enrichment score with regard to random networks of CREG_N3 and ChlamyNET

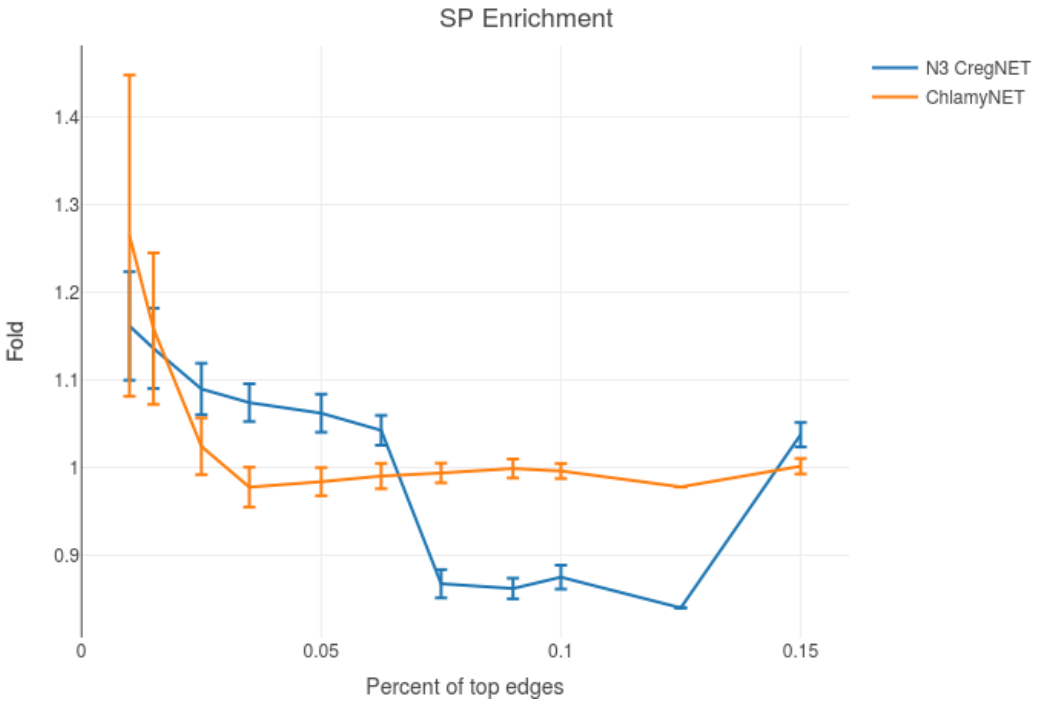

FIGURE 5.13: Fold change of Average Shortest Path enrichment score with regard to random network of CREG_N3 and ChlamyNET

Chapter 6

Conclusions and Future Work

This thesis has been intended to the analysis of various meta-analysis methods for reverse-engineering GRNs from multiple gene expression datasets. GRNs are of essential because by better understanding GRNs we could, for example, manipulate algae to generate biofuel or alternative foods. Furthermore, from a biomedical perspective, this enables us to understand causes of diseases in order to design a suitable drug for a particular disease without side effects accompanied with it. Interestingly, the no free lunch theorem, which is said that if an algorithm performs well on a certain class of problems then it necessarily will perform poorly on the set of all remaining problems, is seen for all benchmarking methods. For instance, with *in silico* data, M family methods have better performance compared to those of N methods. However, N1 and N3 are the two best for EcoliNET and YeastNET while M2 is the best for FlyNET. Consequently, N3 which is competitive with the other methods while preserving the scalability is selected to create CregNET - a first GRN of *C. reinhardtii*. Experiment results using various measures suggest that CregNET perform better than ChlamyNET - a first GCN of *C. reinhardtii*. The main key explaining for CreNET's superior is more data. The amount of raw RNA-seq data we have collected is about 10 times as large as that for ChlamyNET (2.3 TB of 540 samples compared to 287 GB of 50 samples). Secondly, in the thesis we provide a more systematic and robust meta-analysis method (N3) to address the batch effects accompanied with transcriptomic data.

6.1 Accomplished work

6.1.1 Propose and Comparing meta-networks

Reverse-engineering of GRNs from multiple studies has been a normal routine in systems biology since such approach can lead to more accurate results compare to the traditional approach working with a single study. In this work we presented different methods namely data merging, networks ensemble and then propose MI-based meta-analysis to construct GRNs for model organisms from their transcriptome data. First, we evaluate the effectiveness of the collection of methods on *In silico* data and then with *E. coli* , *Saccharomyces cerevisiae* and *Drosophila compendia*. To the best of our knowledge this is the first and thorough analysis of different meta-analysis approaches for reconstructing GRNs from various gene expression datasets.

6.1.2 Create and Validate CregNET - a first GRN for *C. reinhardtii*

After comparing different meta-analysis methods, we selected the most consistent one from the collection to reconstruct a GRN for the model organism *Chlamydomonas*

reinhardtii (CregNET). It should be noted that we also provide a pipeline for collecting and pre-processing RNA-Seq data of *C. reinhardtii* from NCBA SRA. This pipeline can be extended for other model organisms as well. Experiment results then strongly suggest that CregNET outperforms the current co-expression network ChlamyNET in term of stability and predictive power for new GO discovery.

6.2 Future Direction

Apart from regulations at transcription level, there are other molecular levels, such as metabolites or proteins, which should be included in GRNs to capture the full complexity of cellular processes. In other words, it is of great important to integrate different types of networks with GRNs, including protein-protein interactions, microRNA, RNA binding protein, and metabolic and signaling networks. Additionally, we merely validate CregNET using some statistics measurements namely PPI, GO, KO, etc. Therefore for the future work, biological validation is necessary to discover potential important biological processes.

Appendix A

Pipeline for collecting and preprocessing RNA-Seq data of *C. reinhardtii* from NCBI SRA

Guideline for collecting and preprocessing RNA-Seq data of *C. reinhardtii* from NCBI SRA

1. Preparation
   - Packages
   - Reference directories
2. Search for all RNA-Seq data of *C. reinhardtii*
3. Download raw data from NCBI SRA
4. Quantify raw data
5. Transform data to gene count summary

1. Preparation

Packages

```
suppressMessages(library(GEOquery))
suppressMessages(library(GEOmetadb))
suppressMessages(library(SRAdb))
suppressMessages(library(RCurl)) # for getURL
suppressMessages(library(dplyr)) # manipulating matrices
```

Reference directories

```
basedir <- "/home/biosys/Working/doctoral_thesis"
```

2. SRA search

```
searchSRA <- function(dir=getwd()) {
   # define the sql file
   sqlfile = "SRAmetadb.sqlite"
   file.info(sqlfile)

   # create a connection for later queries
   sra_con <- dbConnect(SQLite(), sqlfile)

   # cread a query for full text search
   rs <- getSRA(search_terms = "Chlamydomonas rna-sequencing",
                out_types=c("sra"),
                acc_only=F,
                sra_con=sra_con)
   dim(rs)
   head(rs)
```

```
    # filter out the results to retain only studies with at least 6 samples
    rsFiltered <- as.data.frame(rs %>%
                                group_by(study, study_title,
                                         library_strategy, library_source) %>%
                                select(study, study_title,
                                       library_strategy, library_source) %>%
                                summarise(NoOfSamples = n()) %>%
                                filter(NoOfSamples>=6 &
                                       library_strategy=="RNA-Seq" &
                                       library_source=="TRANSCRIPTOMIC"))

    # remove Small RNA study
    rsFiltered <- rsFiltered[-9, ]

    # get all sra files for every study
    sra.files <- lapply(rsFiltered[,1],
                    function(x) listSRAfile(x, sra_con)$run)
    names(sra.files) <- rsFiltered[,1]

    # disconnect to the database
    dbDisconnect(sra_con)

    # return sra files
    sra.files
 }

# now use the above function to
# search for all sra files of Chlamydomonas
rs <- searchSRA()
```

```
## `summarise()` regrouping output by 'study', 'study_title', 'library_strategy' (override with `.groups
```

```
rs[0:2]
```

```
## $DRP002675
##  [1] "DRR021709" "DRR021703" "DRR021713" "DRR021704" "DRR021710" "DRR021707"
##  [7] "DRR021706" "DRR021715" "DRR021717" "DRR021702" "DRR021705" "DRR021716"
## [13] "DRR021708" "DRR021711" "DRR021712" "DRR021714"
##
## $DRP003701
##  [1] "DRR039907" "DRR039893" "DRR039892" "DRR039903" "DRR039905" "DRR039899"
##  [7] "DRR039908" "DRR039896" "DRR039900" "DRR039901" "DRR039906" "DRR039909"
## [13] "DRR039894" "DRR039904" "DRR039898" "DRR039897" "DRR039895" "DRR039902"
## [19] "DRR039913" "DRR039912" "DRR039910" "DRR039911" "DRR039915" "DRR039917"
## [25] "DRR039916" "DRR039918" "DRR039914"
```

3. Download raw data from NCBI SRA

- SRA provides tools (*fastq-dump*) to download and covert from SRA format to fastq format that can be passed in to *Salmon*
- It is important to use *–split-files* option since some of the reads in SRA are paired-end reads
- *–skip-technical* is to retain only biological reads by removing technical reads

```
DownloadSRA <- function (sraFile,
                         dir=getwd(),
                         destdir="temp") {
    l <- length(sraFile)
    setwd(paste(dir, destdir, sep=""))
    for (i in 1:l) {
        if (file.exists(sraFile[i])) next
        command <- paste("mkdir -p ", sraFile[i])
        system(command)
        setwd(paste(dir, destdir, sraFile[i], sep=""))
        command <- paste("prefetch", sraFile[i])
        #print(command)
        system(command)
        command <- paste("fastq-dump --gzip --skip-technical --readids",
                         "--dumpbase --split-files --clip", sraFile[i])
        #print(command)
        system(command)
        setwd(paste(dir, destdir, sep=""))
    }
    command <- "rm -f ~/ncbi/public/sra/*"
    system(command)
}
```

4. Quantify raw data

The first step of using *Salmon* is preparing transcriptome indices

```
# make index
transcript_fasta <- paste0(basedir, "/Creinhardtii_281_v5.5.transcript.fa")
chlamy_index <- paste0(basedir, "/chlamy_index/")
command <- paste0("salmon index -t ", transcript_fasta, " -i ", chlamy_index )
if (!file.exists(chlamy_index)) {
  system(command)
} else {
  message(paste0("Index already exists."))
}
```

```
## Index already exists.
```

After that the next step is quantifying

```
# function to quantify count
QuantSRA <- function (sraFile, sraDir, indexFile="chlamy_index") {
    l <- length(sraFile)
    for (i in 1:l) {
        outDir <- paste0(sraDir, "/", sraFile[i])
        inDir <- paste0(sraFile[i], "/")
        print(paste0("Processing sample", sraFile[i]))
        if (length(list.files(inDir) == 1)) {
            command <- paste0("salmon quant -i ", indexFile, " -l A -r ", inDir,
                              sraFile[i], "_1.fastq.gz -p 8 -o quants/",
```

```
                                outDir, "_quant")
        } else {
            command <- paste0("salmon quant -i ", indexFile, " -l A -1 ", inDir,
                                sraFile[i], "_1.fastq.gz -2 ", inDir, sraFile[i],
                                "_2.fastq.gz -p 8 -o quants/", outDir, "_quant")
        }
        # print(command)
        system(command)
    }
}
```

5. Gene count summary

```
importCountFile <- function(sraName, sraFiles, tx2gene) {
    print(paste0("Importing count file for ", sraName))
    files <- file.path(paste0(basedir, "/ChlamyGRN"),
                       "chlamydata",
                       sraName,
                       paste0(sraFiles, "_quant"),
                       "quant.sf")
    if (all(file.exists(files))) {
        txi.salmon <- tximport(files,
                               type = "salmon",
                               tx2gene = tx2gene)
        txi.salmon
    }
}
```

Bibliography

Aderem, Alan (2005). "Systems Biology: Its Practice and Challenges". In: *Cell* 121.4, pp. 511–513.

Adler, Priit et al. (2009). "Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods". In: *Genome biology* 10.12, R139.

Anders, Simon et al. (2013). "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor". In: *Nature Protocols* 8.9, pp. 1765–1786.

Anna, C. and Y. Yee (2010). "Comparison study of microarray meta-analysis methods". In: *BMC Bioinformatics* 11.

Arloth, Janine et al. (2015). "Re-Annotator: Annotation pipeline for microarray probe sequences". In: *PLoS ONE* 10.10, pp. 1–13.

Ashburner, Michael et al. (2000). "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1, pp. 25–29.

Ballouz, S., W. Verleyen, and J. Gillis (2015). "Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers". In: *Bioinformatics* 31.13, pp. 2123–2130.

Banf, Michael and Seung Y. Rhee (2017). "Enhancing gene regulatory network inference through data integration with markov random fields". In: *Scientific Reports* 7.October 2016, pp. 1–13.

Barrett, Tanya et al. (2013). "NCBI GEO: Archive for functional genomics data sets - Update". In: *Nucleic Acids Research* 41.D1, pp. 991–995.

Barzel, Baruch and Albert-László Barabási (2013). "Network link prediction by global silencing of indirect correlations." In: *Nature biotechnology* 31.8, pp. 720–5.

Belcastro, Vincenzo et al. (2011). "Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function". In: *Nucleic acids research* 39.20, pp. 8677–8688.

Bellot Pujalte, Pau (2017). "Study of gene regulatory networks inference methods from gene expression data". In:

Bellot Pujalte, Pau et al. (2015). "Study of normalization and aggregation approaches for consensus network estimation". In: *IEEE SSCI 2015: 2015 IEEE Symposium Series on Computational Intelligence; 7-10 December 2015, Cape Town, South Afrika*. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–6.

Bellot, Pau, Catharina Olsen, et al. (2015). "NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference". In: *BMC bioinformatics* 16.1, p. 1.

Bellot, Pau, Philippe Salembier, et al. (2019a). *Unsupervised GRN Ensemble*, pp. 283–302.

— (2019b). *Unsupervised GRN Ensemble*. Springer, pp. 283–302.

Benito, Monica et al. (2004). "Adjustment of systematic microarray data biases". In: *Bioinformatics* 20.1, pp. 105–114.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Benson, Dennis A. et al. (2012). "GenBank". In: *Nucleic Acids Research* 41.1, pp. D36–D42.

Betel, Doron et al. (2018). "Unifying cancer and normal RNA sequencing data from different sources". In: *Scientific Data* 5, p. 180061.

Bevilacqua, Vitoantonio et al. (2011). "Comparison of data-merging methods with SVM attribute selection and classification in breast cancer gene expression". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6840 LNBI.Suppl 7, pp. 498–507.

Bhargava, A. et al. (2013). "Identification of Cytokinin-Responsive Genes Using Microarray Meta-Analysis and RNA-Seq in Arabidopsis". In: *Plant Physiology* 162.1, pp. 272–294.

Bittner, Anton et al. (2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells". In: *PLoS ONE* 9.1, e78644.

Bolstad, B M et al. (2003). "Gene Expression Omnibus\rA Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance". In: *Bioinformatics* 19, pp. 185–193.

Bolstad, B. M. et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics*.

Borenstein, Michael et al. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Bray, Nicolas L. et al. (2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5, pp. 525–527.

Brazma, Alvis (2009). "Minimum Information About a Microarray Experiment (MIAME) Successes, Failures, Challenges". In: *The Scientific World JOURNAL* 9, pp. 420–423.

Brazma, Alvis et al. (2003). "ArrayExpress - A public repository for microarray gene expression data at the EBI". In: *Nucleic Acids Research* 31.1, pp. 68–71.

Bruggeman, Frank J. and Hans V. Westerhoff (2007). "The nature of systems biology". In: *Trends in Microbiology* 15.1, pp. 45–50.

Bullard, James H et al. (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". In: *BMC bioinformatics* 11, p. 94.

Burnett, Jonathan E. et al. (2017). "Batch effects and the effective design of single-cell gene expression studies". In: *Scientific Reports* 7.1, pp. 1–15.

Butte, Atul J and Isaac S Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". In: *Pac Symp Biocomput*. Vol. 5, pp. 418–429.

Carré, Clément, André Mas, and Gabriel Krouk (2017). "Reverse engineering highlights potential principles of large gene regulatory network design and learning". In: *npj Systems Biology and Applications* 3.1, pp. 1–14.

Century, Karen, T Lynne Reuber, and Oliver J Ratcliffe (2008). "Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products". In: *Plant physiology* 147.1, pp. 20–29.

Cestarelli, Valerio et al. (2015). "CAMUR: Knowledge extraction from RNA-seq cancer data through equivalent classification rules". In: *Bioinformatics*, btv635.

Chaitankar, Vijender et al. (2010). "A novel gene network inference algorithm using predictive minimum description length approach". In: *BMC systems biology* 4.1, pp. 1–12.

Chang, Lun Ching et al. (2013). "Meta-analysis methods for combining multiple expression profiles: Comparisons, statistical characterization and an application guideline". In: *BMC Bioinformatics* 14.1.

Chang, Roger L et al. (2011). "Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism". In: *Molecular systems biology* 7.1, p. 518.

Cheadle, Chris et al. (2003). "Analysis of microarray data using Z score transformation". In: *The Journal of molecular diagnostics* 5.2, pp. 73–81.

Chen, Chao et al. (2011). "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods". In: *PLoS ONE* 6.2.

Chen, Dijun et al. (2018). "Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana". In: *Nature communications* 9.1, p. 4534.

Chisti, Yusuf (2007). "Biodiesel from microalgae". In: *Biotechnology advances* 25.3, pp. 294–306.

Choi, Seung Phill, Minh Thu Nguyen, and Sang Jun Sim (2010). "Enzymatic pretreatment of Chlamydomonas reinhardtii biomass for ethanol production". In: *Bioresource Technology* 101.14, pp. 5330–5336.

Collado-Torres, Leonardo et al. (2017). "Reproducible RNA-seq analysis using recount2". In: *Nature Biotechnology* 35.4, pp. 319–321.

Conesa, Ana et al. (2016). "A survey of best practices for RNA-seq data analysis". In: *Genome Biology* 17.1, pp. 1–19.

Cramer, Grant R. et al. (2011). "Effects of abiotic stress on plants: A systems biology perspective". In: *BMC Plant Biology* 11.

Cui, Jian et al. (2008). "AtPID: Arabidopsis thaliana protein interactome database - An integrative platform for plant systems biology". In: *Nucleic Acids Research* 36.SUPPL. 1, pp. 999–1008.

Cuzick, Jack (1985). "A wilcoxon-type test for trend". In: *Statistics in medicine* 4.4, pp. 543–547.

Davidson, E H and M S Levine (2008). "Gene Networks in Development and Evolution Special Feature Sackler Colloquium: Properties of developmental gene regulatory networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.51, pp. 20063–20066.

Davidson, Eric H. (2010). "Emerging properties of animal gene regulatory networks". In: *Nature* 468.7326, pp. 911–920.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 233–240.

De Smet, Riet and Kathleen Marchal (2010). "Advantages and limitations of current network inference methods". In: *Nature Reviews Microbiology* 8.10, pp. 717–729.

Desmedt, Christine et al. (2008). "Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes". In: *Clinical Cancer Research* 14.16, pp. 5158–5165.

Dhanasekaran, Saravana M. et al. (2014). "Transcriptome meta-analysis of lung cancer reveals recurrent aberrations in NRG1 and Hippo pathway genes". In: *Nature Communications* 5.

Dillies, Marie Agnès et al. (2013). "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Briefings in Bioinformatics*.

Ding, Chris and Hanchuan Peng (2005). "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of bioinformatics and computational biology* 3.02, pp. 185–205.

Doms, Andreas and Michael Schroeder (2005). "GoPubMed: Exploring PubMed with the gene ontology". In: *Nucleic Acids Research* 33.SUPPL. 2, pp. 783–786.

Dyrskjøt, Lars et al. (2004). "Gene Expression in the Urinary Bladder A Common Carcinoma in Situ Gene Expression Signature Exists Disregarding Histopathological Classification". In: *Cancer Research* 64.11, pp. 4040–4048.

Edgar, R. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Research* 30.1, pp. 207–210.

Edgar, Ron, Michael Domrachev, and Alex E Lash (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic acids research* 30.1, pp. 207–210.

Ellis, Paul D (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Emmert-Streib, Frank and Matthias Dehmer (2018). "Inference of Genome-Scale Gene Regulatory Networks: Are There Differences in Biological and Clinical Validations?" In: *Machine Learning and Knowledge Extraction* 1.1, pp. 138–148.

Emmert-Streib, Frank, Matthias Dehmer, and Benjamin Haibe-Kains (2014). "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks." In: *Frontiers in cell and developmental biology* 2.August, p. 38.

Espín-Pérez, Almudena et al. (2018). "Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data". In: *PLoS ONE* 13.8, pp. 1–19.

Faith, Jeremiah J et al. (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles". In: *PLoS biology* 5.1, e8.

Field, Andy P (2001). "Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods." In: *Psychological methods* 6.2, p. 161.

Forst, Christian V. et al. (2017). "Integrative gene network analysis identifies key signatures, intrinsic networks and host factors for influenza virus A infections". In: *npj Systems Biology and Applications* 3.1, pp. 1–16.

Gama-Castro, Socorro et al. (2008). "RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation". In: *Nucleic acids research* 36.suppl_1, pp. D120–D124.

Garber, Manuel et al. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq". In: *Nature Methods* 8.6, pp. 469–477.

Goncalves, Elton C. et al. (2016). "Metabolic regulation of triacylglycerol accumulation in the green algae: identification of potential targets for engineering to improve oil yield". In: *Plant Biotechnology Journal* 14.8, pp. 1649–1660.

Gonzalez-Angulo, Ana Maria, Bryan T.J. Hennessy, and Gordon B. Mills (2010). "Future of personalized medicine in oncology: A systems biology approach". In: *Journal of Clinical Oncology* 28.16, pp. 2777–2783.

Gupta, Chirag and Andy Pereira (2019). "Recent advances in gene function prediction using context-specific coexpression networks in plants". In: *F1000Research* 8.0, p. 153.

Han, Leng et al. (2014). "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types". In: *Nature Communications* 5.1, pp. 1–9.

Hansen, Bjoern Oest et al. (2018). "Ensemble gene function prediction database reveals genes important for complex I formation in Arabidopsis thaliana". In: *New Phytologist* 217.4, pp. 1521–1534.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney (2013). "GSVA: Gene set variation analysis for microarray and RNA-Seq data". In: *BMC Bioinformatics* 14.

Hase, Takeshi et al. (2013). "Harnessing Diversity towards the Reconstructing of Large Scale Gene Regulatory Networks". In: *PLoS Computational Biology* 9.11.

Hastie, T et al. (2016). *Impute: Imputation for microarray data. R package version 1.38. 1*.

Haury, Anne-Claire et al. (2012). "TIGRESS: trustful inference of gene regulation using stability selection". In: *BMC systems biology* 6.1, p. 145.

Hecker, Michael et al. (2009). "Gene regulatory network inference: Data integration in dynamic models-A review". In: *BioSystems* 96.1, pp. 86–103.

Heider, Andreas (2013). "virtualArray : A R / Bioconductor package to merge raw data from different microarray platforms Supplementary information Detailed explanation to set up example data". In: pp. 1–6.

Henríquez-Valencia, Carlos et al. (2018). "Integrative Transcriptomic Analysis Uncovers Novel Gene Modules That Underlie the Sulfate Response in Arabidopsis thaliana". In: *Frontiers in Plant Science* 9.April, pp. 1–20.

Herrgård, Markus J. et al. (2008). "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology". In: *Nature Biotechnology* 26.10, pp. 1155–1160.

Hong, Fangxin et al. (2006). "RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis". In: *Bioinformatics* 22.22, pp. 2825–2827.

Howe, Eleanor A. et al. (2011). "RNA-Seq analysis in MeV". In: *Bioinformatics* 27.22, pp. 3209–3210.

Hubbard, T. (2002). "The Ensembl genome database project". In: *Nucleic Acids Research* 30.1, pp. 38–41.

Huber, Wolfgang et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". In: *Bioinformatics* 18.suppl 1, S96–S104.

Huynh-Thu, Vân Anh and Pierre Geurts (2018). "DynGENIE3: Dynamical GENIE3 for the inference of gene networks from time series expression data". In: *Scientific Reports* 8.1, pp. 1–12.

— (2019). "Unsupervised Gene Network Inference with Decision Trees and Random Forests". In: 1883, pp. 195–215.

Huynh-Thu, Vân Anh, Alexandre Irrthum, et al. (2010). "Inferring regulatory networks from expression data using tree-based methods". In: *PloS one* 5.9, pp. 1–10.

Huynh-Thu, Vân Anh and Guido Sanguinetti (2018). "Gene regulatory network inference: an introductory survey". In: pp. 1–23.

Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang (2018). "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental and Molecular Medicine* 50.8.

Iancu, Ovidiu D. et al. (2012). "Utilizing RNA-Seq data for de novo coexpression network inference". In: *Bioinformatics* 28.12, pp. 1592–1597.

Irizarry, Rafael A et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". In: *Biostatistics* 4.2, pp. 249–264.

J., Ruan, Dean A.K., and Zhang W. (2010). "A general co-expression network-based approach to gene expression analysis: Comparison and applications". In: *BMC Systems Biology* 4.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127.

Kanehisa, Minoru and Goto, Susumu (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1, pp. 27–30.

Kang, Dongwan D. et al. (2012). "MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis". In: *Nucleic Acids Research* 40.2, pp. 1–14.

Karlebach, Guy and Ron Shamir (2008). "Modelling and analysis of gene regulatory networks". In: *Nature Reviews Molecular Cell Biology* 9.10, pp. 770–780.

Khatri, P et al. (2007). "A systems biology approach for pathway level analysis". In: *Genome Research* 17.10, pp. 1537–1545.

Kiemeney, Lambertus A et al. (2013). "GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer". In: *Nature Genetics* 45.4, pp. 362–370.

Kim, Hanhae et al. (2013). "YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae". In: *Nucleic acids research*, gkt981.

Kirschner, Marc W. (2005). "The Meaning of Systems Biology". In: *Cell* 121.4, pp. 503–504.

Kitano, Hiroaki (2002a). "Nature01254". In: *Nature* 420.6912, pp. 206–210.

— (2002b). "Systems Biology: A brief overview". In: *Aaas* 295.March, pp. 1662–1664.

Kleinman, J E et al. (2014). "RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder". In: *Molecular Psychiatry* 19.11, pp. 1179–1185.

Kocher, Jean-Pierre A et al. (2011). "Batch effect correction for genome-wide methylation data with Illumina Infinium platform". In: *BMC Medical Genomics* 4, p. 84.

Kodama, Yuichi, Martin Shumway, and Rasko Leinonen (2012). "The sequence read archive: Explosive growth of sequencing data". In: *Nucleic Acids Research* 40.D1, pp. 2011–2013.

Kong, Qing-xue et al. (2010). "Culture of microalgae Chlamydomonas reinhardtii in wastewater for biomass feedstock production". In: *Applied biochemistry and Biotechnology* 160.1, p. 9.

Kugler, Karl G et al. (2011). "Integrative network biology: graph prototyping for co-expression cancer networks". In: *PLoS One* 6.7, e22843.

Kunkel, Eric J (2006). "Systems biology in drug discovery." In: *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 1.10, p. 37.

Lachmann, Alexander et al. (2018). "Massive mining of publicly available RNA-seq data from human and mouse." In: *Nature communications* 9.1, p. 1366.

Larsen, Martin J. et al. (2014). "Microarray-based RNA profiling of breast cancer: Batch effect removal improves cross-platform consistency". In: *BioMed Research International* 2014.

Lazar, Cosmin et al. (2013). "Batch effect removal methods for microarray gene expression data integration: A survey". In: *Briefings in Bioinformatics* 14.4, pp. 469–490.

Leek, Jeffrey T. (2014). "Svaseq: Removing batch effects and other unwanted noise from sequencing data". In: *Nucleic Acids Research* 42.21, e161.

Leek, Jeffrey T., W. Evan Johnson, et al. (2012). "The SVA package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6, pp. 882–883.

Leek, Jeffrey T., Robert B. Scharpf, et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10, pp. 733–739.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway (2011). "The sequence read archive". In: *Nucleic Acids Research* 39.SUPPL. 1, pp. 2010–2012.

Li, Yixue et al. (2009). "Estimating accuracy of RNA-Seq and microarrays with proteomics". In: *BMC Genomics* 10.1, p. 161.

Lin, Shoukai et al. (2018). "Construction and Analysis of Gene Co-Expression Networks in Escherichia coli". In: *Cells* 7.3, p. 19.

Liu, Ganqiang, John S. Mattick, and Ryan J. Taft (2013). "A meta-analysis of the genomic and transcriptomic composition of complex life". In: *Cell Cycle* 12.13, pp. 2061–2072.

Liu, Shikai et al. (2015). "Claudin multigene family in channel catfish and their expression profiles in response to bacterial infection and hypoxia as revealed by meta-analysis of RNA-Seq datasets". In: *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* 13, pp. 60–69.

Lohse, Marc et al. (2012). "RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics". In: *Nucleic Acids Research* 40.W1, pp. 622–627.

Love, MI, S Anders, and W Huber (2017). "Analyzing RNA-seq data with DESeq2". In: *Bioconductor* 2.January, pp. 1–63.

Love, Michael I., Charlotte Soneson, and Rob Patro (2018). "Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification". In: *F1000Research* 7, p. 952.

Love, Michael I et al. (2015). "RNA-Seq workflow: gene-level exploratory analysis and differential expression". In: *F1000Research* 4.

Lun, Aaron T.L., Davis J. McCarthy, and John C. Marioni (2016). "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor". In: *F1000Research* 5, p. 2122.

Luo, Jingchu et al. (2013). "PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors". In: *Nucleic Acids Research* 42.D1, pp. D1182–D1187.

Luo, J et al. (2010). "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data." In: *The pharmacogenomics journal* 10.4, pp. 278–291.

MacNeil, Lesley T. and Albertha J.M. Walhout (2011). "Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression". In: *Genome Research* 21.5, pp. 645–657.

Madhamshettiwar, Piyush B. et al. (2012). "Gene regulatory network inference: Evaluation and application to ovarian cancer allows the prioritization of drug targets". In: *Genome Medicine* 4.5.

Mao, Linyong et al. (2009). "Arabidopsis gene co-expression network and its functional modules". In: *BMC Bioinformatics* 10, pp. 1–24.

Marbach, Daniel et al. (2012). "Wisdom of crowds for robust gene network inference". In: *Nature methods* 9.8, pp. 796–804.

Margolin, Adam A et al. (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context". In: *BMC bioinformatics* 7.Suppl 1, S7.

Marino, Simeone et al. (2008). "A methodology for performing global uncertainty and sensitivity analysis in systems biology". In: *Journal of Theoretical Biology* 254.1, pp. 178–196.

Marquardt, Jens U. et al. (2012). "RNA-Seq Atlasa reference database for gene expression profiling in normal tissue by next-generation sequencing". In: *Bioinformatics* 28.8, pp. 1184–1185.

Martinez, Ricardo, Nicolas Pasquier, and Claude Pasquier (2008). "GenMiner: Mining non-redundant association rules from integrated gene expression data and annotations". In: *Bioinformatics*.

Mawhood, Rebecca et al. (2016). "Production pathways for renewable jet fuel: a review of commercialization". In: *Biofuels, Bioprod. Bioref.* Pp. 431–440.

Megherbi, Dalila et al. (2014). "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance". In: *Nature Biotechnology* 32.9, pp. 926–932.

Menand, Benoît et al. (2013). "Recruitment and remodeling of an ancient gene regulatory network during land plant evolution". In: *Proceedings of the National Academy of Sciences* 110.23, pp. 9571–9576.

Meyer, Patrick E, Benjamin Haibe-Kains, and Gianluca Bontempi (2009). "Meta-Analysis in Transcriptional Network Inference". In: *Recomb Satellite 09*. MIT.

Meyer, Patrick E, Kevin Kontos, and Gianluca Bontempi (2007). "Biological network inference using redundancy analysis". In: *Bioinformatics Research and Development*. Berlin Heidelberg: Springer, pp. 16–27.

Meyer, Patrick E, Frederic Lafitte, and Gianluca Bontempi (2008). "minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information". In: *BMC bioinformatics* 9.1, p. 461.

Meyer, Pe et al. (2010). "Information-Theoretic Inference of Gene Networks Using Backward Elimination". In: *Biocomp*, pp. 700–705.

Meyer, Rachel S. and Michael D. Purugganan (2013). "Evolution of crop species: Genetics of domestication and diversification". In: *Nature Reviews Genetics* 14.12, pp. 840–852.

Miao, Ruoyu et al. (2011). "Large-scale prediction of long non-coding RNA functions in a codingnon-coding gene co-expression network". In: *Nucleic Acids Research* 39.9, pp. 3864–3878.

Mining, Transcriptomic Data et al. (2019). "Chapter 5 for Computational Drug Discovery". In: 1903, pp. 73–95.

Mochida, Keiichi et al. (2018). "Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets". In: *Frontiers in Plant Science* 9.November, pp. 1–7.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628.

Nalluri, Joseph J. et al. (2017). "MiRsig: A consensus-based network inference methodology to identify pan-cancer miRNA-miRNA interaction signatures". In: *Scientific Reports* 7.January, pp. 1–14.

Neupane, Rochak et al. (2011). "Phytozome: a comparative platform for green plant genomics". In: *Nucleic Acids Research* 40.D1, pp. D1178–D1186.

Nygaard, Vegard, Einar Andreas Rødland, and Eivind Hovig (2016). "Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses". In: *Biostatistics* 17.1, pp. 29–39.

Omranian, Nooshin, Jeanne M. O. Eloundou-Mbebi, et al. (2016). "Gene regulatory network inference using fused LASSO on multiple data sets". In: *Scientific Reports* 6, p. 20533.

Omranian, Nooshin, Jeanne MO Eloundou-Mbebi, et al. (2016). "Gene regulatory network inference using fused LASSO on multiple data sets". In: *Scientific reports* 6, p. 20533.

Oshlack, Alicia, Mark D. Robinson, and Matthew D. Young (2010). "From RNA-seq reads to differential expression results". In: *Genome Biology* 11.12, pp. 1–10.

Pachter, Lior et al. (2012). "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nature Biotechnology* 31.1, pp. 46–53.

Park, Jin Hwan, Kwang Ho Lee, et al. (2007). "Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation". In: *Proceedings of the National Academy of Sciences* 104.19, pp. 7797–7802.

Park, Jin Hwan, Sang Yup Lee, et al. (2008). "Application of systems biology for bioprocess development". In: *Trends in biotechnology* 26.8, pp. 404–412.

Patro, Rob et al. (2017). "Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference". In: *Nature Methods* 14.4, pp. 417–419.

Pham, Ngoc C, Benjamin Haibe-Kains, et al. (2017). "Study of Meta-analysis strategies for network inference using information-theoretic approaches". In: *BioData mining* 10.1, p. 15.

Pham, Ngoc C and Patrick E Meyer (2019). "Minet version 4.0: meta-analysis methods to infer gene regulatory network from multiple data sets." In: *To be submitted in BMC Bioinformatics*.

Pham, Ngoc C, Manuel Noll, and Patrick E Meyer (2019). "CregNET: Meta-analysis of *Chlamydomonas reinhardtii* gene regulatory network." In: *To be submitted in Molecular Systems Biology*.

Pham, Ngoc Cam et al. (2017). "Study of Meta-analysis Strategies for Network Inference Using Information-Theoretic Approaches". In: *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, pp. 76–83.

Pintea, Sebastian and Ramona Moldovan (2009). "The Receiver-Operating Characteristic (ROC) analysis: Fundamentals and applications in clinical psychology". In: *Journal of Cognitive and Behavioral Psychotherapies* 9.1, pp. 49–66.

Poilblanc, Didier and Yasumasa Hasegawa (1990). "Numerical study of flux phases in the t-J model". In: *Physica B: Physics of Condensed Matter* 163.1-3, pp. 538–540.

Qiu, Xing, Hulin Wu, and Rui Hu (2013). "The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis". In: *BMC bioinformatics* 14.1, p. 124.

Querec, Troy D. et al. (2009). "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans". In: *Nature Immunology* 10.1, pp. 116–125.

Ramasamy, Adaikalavan et al. (2008). "Key issues in conducting a meta-analysis of gene expression microarray datasets". In: *PLoS Medicine* 5.9, pp. 1320–1332.

Rampasek, Ladislav and Anna Goldenberg (2016). "TensorFlow: Biology's Gateway to Deep Learning?" In: *Cell Systems* 2.1, pp. 12–14.

Rau, Andrea, Guillemette Marot, and Florence Jaffrézic (2014). "Differential meta-analysis of RNA-seq data from multiple studies". In: *BMC Bioinformatics* 15.1, pp. 1–10.

Renard, Emilie and P Absil (2017). "Gene expression Comparison of batch effect removal methods in the presence of correlation between outcome and batch". In: *Technical report*, pp. 1–7.

Romero-campero, Francisco J et al. (2016). "ChlamyNET : A Chlamydomonas gene co-expression network reveals global properties of the transcriptome and the early setup of key co-expression patterns in the green lineage". In: *BMC Genomics*, pp. 1–41.

Roy, Sushmita et al. (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE". In: *Science*, p. 1198374.

Sanguinetti, Guido et al. (2019). "Gene regulatory network inference: an introductory survey". In: *Gene Regulatory Networks*. Springer, pp. 1–23.

Schaap, Peter J. et al. (2014). "Green genes: bioinformatics and systems-biology innovations drive algal biotechnology". In: *Trends in Biotechnology* 32.12, pp. 617–626.

Schmidt, Florian et al. (2018). "An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets". In: *Bioinformatics* 34.17, pp. i908–i916.

Schmidt, Frank L and John E Hunter (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Los Angeles, Washington DC, London, New Delhi, and Singapore: Sage publications.

Sean, Davis and Paul S. Meltzer (2007). "GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor". In: *Bioinformatics* 23.14, pp. 1846–1847.

Seoud, MA and AEA Mahran (2012). "On permutation graphs". In: *Journal of the Egyptian Mathematical Society* 20.2, pp. 57–63.

Serin, Elise A. R. et al. (2016). "Learning from Co-expression Networks: Possibilities and Challenges". In: *Frontiers in Plant Science* 7.April, pp. 1–18.

Sherman, P. M. et al. (2008). "NCBI GEO: archive for high-throughput functional genomic data". In: *Nucleic Acids Research* 37.Database, pp. D885–D890.

Siahpirani, Alireza Fotuhi, Deborah Chasman, and Sushmita Roy (2019). *Integrative Approaches for Inference of Genome-Scale Gene Regulatory Networks*. Vol. 1883, pp. 161–194.

Siaut, Magali et al. (2011). "Oil accumulation in the model green alga Chlamydomonas reinhardtii: characterization, variability between common laboratory strains and relationship with starch reserves". In: *BMC biotechnology* 11.1, p. 7.

Silberberg, Gilad et al. (2016). "A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions". In: *BMC Bioinformatics* 17.S5.

Sims, Andrew H et al. (2008). "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis". In: *BMC medical genomics* 1.1, p. 1.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences". In: *F1000Research* 4.

Steele, Emma and Allan Tucker (2008). "Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets". In: *Journal of Biomedical Informatics* 41.6, pp. 914–926.

Stein, Caleb K. et al. (2015). "Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat". In: *BMC Bioinformatics* 16.1, pp. 1–9.

Su, Shian et al. (2018). "analysis is easy as 1-2-3 with limma , Glimma and edgeR [version 3 ; referees : 3 approved] Referee Status :" in: 1, pp. 1–29.

Sudmant, Peter H., Maria S. Alexis, and Christopher B. Burge (2015). "Meta-analysis of RNA-seq expression data across species, tissues and studies". In: *Genome Biology* 16.1, pp. 1–11.

Szklarczyk, Damian et al. (2015). "STRING v10: Protein-protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43.D1, pp. D447–D452.

Taminau, Jonatan, Cosmin Lazar, et al. (2014). "Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis." In: *ISRN bioinformatics* 2014, p. 345106.

Taminau, Jonatan, Stijn Meganck, et al. (2012). "Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages." In: *BMC bioinformatics* 13.1, p. 335.

Tan, Shi Dong, Xian W. Chen, and Liang Gong (2010). "Investigation of clinical, imaging features and olfactory function of vascular parkinsonism". In: *Journal of Clinical Neurology* 23.3, pp. 181–184.

Tibshirani, R. et al. (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression". In: *Proceedings of the National Academy of Sciences*.

Turnbull, Arran K et al. (2012). "Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis". In: *BMC medical genomics* 5.1, p. 1.

Van Der Wijst, Monique G.P. et al. (2018). "An integrative approach for building personalized gene regulatory networks for precision medicine". In: *Genome Medicine* 10.1, pp. 1–15.

Van Parys, Thomas et al. (2014). "Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress ". In: *The Plant Cell Online* 26.12, pp. 4656–4679.

Vignes, Matthieu et al. (2011). "Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis". In: *PloS one* 6.12, e29165.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch (2012). "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples". In: *Theory in Biosciences*.

Wan, Quan et al. (2015). "BioXpress: An integrated RNA-seq-derived gene expression database for pan-cancer analysis". In: *Database* 2015.11, pp. 1–13.

Wang, Kai, Manikandan Narayanan, Hua Zhong, Martin Tompa, Eric E. Schadt, et al. (2009). "Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases". In: *PLoS Computational Biology* 5.12.

Wang, Kai, Manikandan Narayanan, Hua Zhong, Martin Tompa, Eric E Schadt, et al. (2009). "Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases". In: *PLoS Comput Biol* 5.12, e1000616.

Wang, Xingbin, Dongwan D. Kang, et al. (2012). "An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection". In: *Bioinformatics* 28.19, pp. 2534–2536.

Wang, Xingbin, Yan Lin, et al. (2012). "Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder". In: *BMC bioinformatics* 13.1, p. 1.

Wang, Yijie et al. (2018). "Reprogramming of regulatory network using expression uncovers sex-specific gene regulation in Drosophila". In: *Nature Communications* 9.1.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10, p. 57.

Warnat, Patrick, Roland Eils, and Benedikt Brors (2005). "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes". In: *BMC Bioinformatics*.

Weirauch, Matthew T (2011). "Gene coexpression networks for the analysis of DNA microarray data". In: *Applied statistics for network biology: methods in systems biology* 1, pp. 215–250.

Wirapati, Pratyaksha et al. (2008). "Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures". In: *Breast Cancer Res* 10.4, R65.

Wu, Douglas C. et al. (2018). "Limitations of alignment-free tools in total RNA-seq quantification". In: *BMC Genomics* 19.1, pp. 1–14.

Xia, Jianguo, Erin E. Gill, and Robert E.W. Hancock (2015). "NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data". In: *Nature Protocols* 10.6, pp. 823–844.

Xia, Xiao Qin et al. (2009). "WebArrayDB: Cross-platform microarray data analysis and public data repository". In: *Bioinformatics*.

Yang, Bin et al. (2018). "Inference of Large-scale Time-delayed Gene Regulatory Network with Parallel MapReduce Cloud Platform". In: *Scientific Reports* 8.1, p. 17787.

Yi, Lynn, Nicolas L Bray, and Lior Pachter (2018). "Gene-level differential analysis at transcript-level resolution". In: pp. 1–24.

You, Qi et al. (2016). "Co-expression network analyses identify functional modules associated with development and stress response in Gossypium arboreum". In: *Scientific Reports* 6.December, pp. 1–15.

Young, Matthew D et al. (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias GOseq GOseq is a method for GO analysis of RNA-seq data that takes into account the length bias inherent in RNA-seq". In: *Genome Biology* 11.

Yu, Donghyeon et al. (2013). "Review of Biological Network Data and Its Applications". In: *Genomics & Informatics* 11.4, p. 200.

Zhang, Yuqing et al. (2018). "Alternative empirical Bayes models for adjusting for batch effects in genomic studies". In: *BMC Bioinformatics* 19.1, pp. 1–15.

Zhu, Yuelin et al. (2013). "SRAdb: Query and use public next-generation sequencing data from within R". In: *BMC Bioinformatics* 14.1, pp. 2–5.