

GENE REGULATORY NETWORK INFERENCE USING
MACHINE LEARNING TECHNIQUES

STEPHANIE KAMGNIA WONKAP

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE)
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

AUGUST 26 2020

© STEPHANIE KAMGNIA WONKAP, 2020

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Miss. Stephanie Kamgnia Wonkap

Entitled: Gene Regulatory Network Inference using Machine Learning
Techniques

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards
with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Liangzhu Wang

_____ External Examiner
Dr. Mathieu Blanchette

_____ Examiner
Dr. Leila Kosseim

_____ Examiner
Dr. Malcolm Whiteway

_____ Examiner
Dr. Volker Haarslev

_____ Supervisor
Dr. Gregory Butler

Approved _____
Dr. Leila Kosseim,
Graduate Program Director

August 26th, 2020

Date of Defence

Dr. Amir Asif, Dean

Faculty of Engineering and Computer Science

Abstract

Gene Regulatory Network Inference using Machine Learning Techniques

Stephanie Kamgnia Wonkap, Ph.D.
Concordia University, 2020

Systems Biology is a field that models complex biological systems in order to better understand the working of cells and organisms. One of the systems modeled is the gene regulatory network that plays the critical role of controlling an organism's response to changes in its environment. Ideally, we would like a model of the complete gene regulatory network. In recent years, several advances in technology have permitted the collection of an unprecedented amount and variety of data such as genomes, gene expression data, time-series data, and perturbation data. This has stimulated research into computational methods that reconstruct, or infer, models of the gene regulatory network from the data. Many solutions have been proposed, yet there remain open challenges in utilising the range of available data as it is inherently noisy, and must be integrated by the inference techniques. The thesis seeks to contribute to this discourse by investigating challenges of performance, scale, and data integration.

We propose a new algorithm **BENIN** that views network inference as feature selection to address issues of scale, that uses elastic net regression for improved performance, and adapts elastic net to integrate different types of biological data. The **BENIN** algorithm is benchmarked on a synthetic dataset from the DREAM4 challenge, and on real expression data for the human HeLa cell cycle. On the DREAM4 dataset **BENIN** out-performed all DREAM4 competitors on the size 100 subchallenge, and is also competitive with more recent state-of-the-art methods. Moreover, on the HeLa cell cycle data, **BENIN** could infer known regulatory interactions and propose new interactions that warrant further experimental investigation.

Keys words: gene regulatory network, network inference, feature selection, elastic net regression.

Acknowledgments

First of all, I would like to thank my academic supervisor, Dr. Gregory Butler, for his precious advice that helped me accomplish this project and become a better researcher. Moreover, I would like to thank him for providing me with financial support.

I would like to thank my mother, Bernadette, and my father, Emmanuel, for showing the path through my Ph.D. Your prayer, love, and support helped go through these tough times. Thank you for your education that made me the strong woman I am today.

I would like to thank my sister Nathalie who is my model since our childhood. I made it I am a doctor like you. Thank you to my sister Helene who was always there, our phone calls, and our discussion helped during all these years. I will not forget my sister Diana, my nephews, my older brother Armand and two little brothers Joan and Gracien. Every one of you plays an essential role in this journey.

To you my beloved husband Samir, you were my rock through this journey. Thank you for holding my back, for being such a good confident and my motivation. I think that if you were not there, I would not be able to accomplish this. I thank God for having you in my life.

Last but not the less, my colleagues from office 11.411. I will never forget all these good times: the sushi time, our potluck dinner, our late discussions and all our fun time. I will miss these times and I wish you all the best in your life.

Contents

List of Figures	vii
List of Tables	ix
List of Terms and Abbreviations	1
1 Introduction	1
1.1 Gene Regulation	2
1.2 Gene Regulatory Network	5
1.3 Problem Statement	9
1.4 Motivation	10
1.5 Challenge in Gene Regulatory Network Inference	12
1.6 Limitation of State-of-the-Art	14
1.7 Contribution	14
1.8 Organization of the Thesis	18
2 Background	19
2.1 Background for Network Inference	19
2.2 Feature Selection	30
2.3 Resources Available for Network Inference	33
2.4 Assesment and Validation of Network Inference	39
2.5 Computational Methods	45
2.6 Conclusion	78
3 BENIN	80
3.1 The BENIN Algorithm	81

3.2	Experimental Validation	91
3.3	Computational Complexity	97
3.4	Results and Discussion	98
3.5	Conclusion	119
4	BENIN: Application to the HeLa Cell cycle	121
4.1	Introduction	121
4.2	Background	124
4.3	Building a gold-standard	132
4.4	Material	137
4.5	Method	146
4.6	Results and Discussion	158
4.7	Conclusion	186
5	Conclusion	188
5.1	Recap	188
5.2	Contributions	189
5.3	Limitations	192
5.4	Future Work	192
	Bibliography	193
A	Background	227
A.1	IUPAC degenerate base symbols	227
B	BENIN	229
B.1	BENIN parameters setting	229
B.2	BENIN results	229
C	BENIN: Application to Human HeLa Cell Cycle GRN	235
C.1	Data	235
C.2	Other	345

List of Figures

1	Organization of an operon in prokaryotes.	3
2	Tryptophan regulation in <i>E. coli</i>	4
3	Eukaryotic gene structure	6
4	Chromatin in eukaryotic cells	7
5	Gene regulatory network abstraction	8
6	Different representations of binding sites	29
7	DNA microarray experiment	35
8	RNA-seq experiment	37
9	Confusion matrix	41
10	Procedure to identify regulon	51
11	Step for regulatory network inference.	58
12	Example DREAM4 Input for BENIN	84
13	Effect of the noise in location data	99
14	Influence of BENIN parameters	100
15	A subnetwork from 100-nodes network 4	109
16	The Eukaryotic cell cycle	125
17	From nucleus to DNA sequence	126
18	Steps for retrieving knockdown data	140
19	Steps for collecting promoter sequences	142
20	Steps for collecting protein sequences	144
22	Snapshot of FIMO output	153
24	Inference of GRN controlling HeLa cell cycle through BENIN	157
21	BED file and BETA-minus output	160
23	Differential Expression analysis output	161
25	Effect of τ on BENIN performance.	161

26	Precision-recall curves for BENIN	163
27	ROC curves for BENIN	165
28	Precision-recall curves for BENIN +orthology	167
29	ROC curves for BENIN +orthology	167
30	Orthologous Regulatory Network From mouse	180
31	Edge Distribution	181
32	Global score Distribution for the DREAM4 size 100 subchallenge. . .	233
33	Global score Distribution for the DREAM4 size 10 subchallenge. . .	234

List of Tables

1	Motifs Finding Methods	52
2	Reverse-Engineering Methods	71
3	Description of DREAM4 size 10 and size 100 networks	91
4	Motifs and errors type	94
5	BENIN execution time on the DREAM4	96
6	DREAM4 size 100 performance with KO expression	103
7	DREAM4 size 100 performance with Location data	104
8	Global score on the DREAM4 size 100 subchallenge	105
9	DREAM4 size 10 performance with Location data	106
10	DREAM4 size 10 performance with KO expression	106
11	Global score on the DREAM4 size 10 subchallenge	107
12	Motif prediction confidence (median rank)	113
13	Cell cycle Transcription Factors	127
14	Characteristics of our Human “gold-standard” network	136
15	Missing transcription factors in our “gold-standard network”	137
16	Mouse gene regulatory network	146
17	BENIN execution time	159
18	BENIN performance	162
19	Transcription factor and target gene	168
20	Inference from BENIN +combined+max	174
21	Inference from BENIN +orthology	182
22	List of Degenerate IUPAC base symbols	227
23	BENIN General Parameter setting	230
24	BENIN +KO parameters on size 100 subchallenge	230
25	BENIN +Location parameters setting on size 100 subchallenge	231

26	BENIN +KO parameters on size 10 subchallenge	231
27	BENIN +Location data parameters on size 10 subchallenge	232
28	List of HeLa Peak Files	238
29	List of knockdown datasets	239
30	Information Motif and Transcription Factor	240
31	Cell cycle genes	249
32	Cell Cycle Transcription factor	251
33	Knockdown Data from Gene Expression Omnibus	268
34	Edges repetition in networks from HumanBase	270
35	Edges repetition in Garcia networks	286
36	HeLa “gold-standard” network - Positive links	302
37	HeLa “gold-standard” network- Negative links	319
38	Duplicate regulatory interaction from TRRUST and RegNetwork	340
39	Edges duplicated in our mouse regulatory network	340
40	BENIN execution time on different network sizes	345

List of Terms and Abbreviations

ACC Accurary.

AUPR Area Under the Precision Recall curve.

AUROC Area under Receiver Operating Characteristic.

Biological pathway series of actions among molecules in a cell that leads to a certain product or a change in a cell

BLAST Basic Local Alignment Search Tool. It is a sequence database searching program which compares a nucleotide or protein query sequence against all sequences in a database.

ChIP Chromatin ImmunoPrecipitation

DNA DeoxyriboNucleic Acid. It is a long double stranded molecule made of nucleotides A, C, G and T that contains the genetic information necessary for the development, functioning and the reproduction of all known living thing.

DREAM Dialogue for Reverse Engineering Assessments and Methods.

EM Expectation Maximisation.

ENet Elastic Net.

Enzyme Macro molecule that accelerates, or catalyzes, chemical reactions.

E-value Expected value, is the number of different alignments with scores equivalent to or better than threshold that are expected to occur by chance in a database search. The lower the E-value, the more significant the score.

FASTA Text-based format for representing either nucleotide sequences or amino acid, in which base pairs or amino acids are represented using single-letter codes.

FDR False Discovery Rate.

FFL Feed Forward Loop.

FN False Negative.

FP False Positive.

Gap Refer to substitution or indel in a sequence, where indel can be insertion or deletion in the sequence.

Gene expression profile Describes the expression levels of a gene across a set of samples obtained for a particular array experiment design.

GENIE3 GEne Network Inference with Ensemble of trees.

GO Gene Ontology. It is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.

GRN Gene Regulatory Network.

IUPAC International Union of Pure and Applied Chemistry. It is the universally-recognized authority on chemical nomenclature and terminology.

LASSO Least Absolute Shrinkage and Selection Operator.

MEME Multiple EM for Motif Elicitation.

NPV Negative Predictive Value.

Non-coding DNA Components of DNA that do not encode protein sequences or RNA.

ODE Ordinary Differential Equation.

OLS Ordinary Least Square.

Ontology Formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of interest.

Operon Set of genes situated next to each other and under the control of the same promoter and operator.

Organelle Specialized subunit within the cell that has a specific function.

Ortholog Orthologous sequences are sequences occurring in different species that diverge from a common ancestral sequence after speciation, the evolutionary process in which new species arise.

Paralogy Sequences are paralogous if they were created by a duplication event within the genome.

Phenotype Observable characteristics of the organism such as eye's color/shape, the hair's color and so on.

Phylogenetic tree Diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.

PK Prior Knowledge.

PPI Proteins Proteins Interaction.

PPV Positive Predictive Value.

PWM Position Weight Matrix.

Regulatory sequence Segment of non-coding DNA that is capable to control the increase or decrease of the expression of specific genes within an organism.

Regulon Set of genes or operons regulated by the same transcription factor.

RNA Ribonucleic acid, is a single stranded molecule made of nucleotides A, C, G and U. It plays a major role in protein synthesis as it is involved in the transcription, decoding, and translation of the genetic code to produce proteins.

SPC Specificity.

System Biology Computational and mathematical modeling of complex biological network.

TG Target Gene. It is a gene that is regulated by a transcription factor is called targeted gene of this transcription factor.

TF Transcription Factor.

TFBS Transcription Factor Binding Site.

TIGRESS Trustful Inference of Gene REgulation using Stability Selection.

TN True Negative.

TP True Positive.

TPR True Positive Rate.

TFBS Transcription Factor Binding Site.

VAR Vector AutoRegressive.

Enhancer TO DO

Silencer TO DO

Histone TO DO

Chromatin TO DO

ANOVA Analysis Of Variance

MRMR Minimum Redundancy Maximum Relevance

LASSO Least Absolute Shrinkage and Selection Operator

DEG Differentially Expressed Gene

ChiP-seq Chromatin immunoprecipitation followed by sequencing

DBN Dynamic Bayesian Network

DAG Directed Acyclic Graph

DBD DNA Binding Domain

CDK Cyclin Dependent Kinase

Sister Chromatid copies of a chromosome held at the centromere

Spindle fibers aggregate of microtubules that are formed during the cell cycle and that move the chromosomes.

Microtubule protein filament that resembles hollow tube.

KO Knockout

KD Knock down

KNN K nearest neighbor

Peak calling computational method used to identify areas in a genome that have been enriched with aligned reads after performing ChIP-sequencing experiment

AWG ENCODE Analysis Working Group

IC Information Content

MICA Most Informative Common Ancestor

TSS Transcription Start Site

FDR False Discovery Rate

EM Expectation Maximization

k-mer sequence of length k

MCMC Markov Chain Monte Carlo

BMA Bayesian Model Averaging

GGM Graphical Gaussian Model

FPR False Positive Rate

DPI data processing inequality

GEO Gene Expression Omnibus

Chapter 1

Introduction

All living organisms on the earth interact with other organisms and are regularly exposed to environmental factors in their habitats. These factors are varied, encompassing temperature, oxygen level, nutrient and water availability, and in some cases, the presence of toxic elements. In response to variations in these factors, organisms need to develop features to survive. These features take the form of gene expression and regulation. The following thesis engages with the complexities of this process. The thesis will be grounded on three key questions: “What is gene expression?” “What does it mean to regulate the expression of a gene?” and finally, “How do these processes work?”

System Biology, a discipline that is deeply rooted in biology, physics, chemistry, as well as in computer science and mathematics, provides a mechanism for modeling the complex networks of biologically relevant entities (DNA, RNA, proteins, or cells) and in so doing, provides an avenue for answering questions such as “How does a biological component interact with other components and its environment?”, “What regulates its function and in what manner?”, “What kind of properties emerge from these interactions?” and so on [240]. The **gene regulatory network (GRN)** is an example of these complex networks. The GRN offers a path to understand parameters that contribute to a properly functioning cell. Moreover, GRN helps understanding interactions between different organisms as well as the interaction with their habitats. Thus, there is a strong need to model such a complex network for scientists to have abstract reasoning about its dynamics. Even though we now witness high-throughput experiments that produce a plethora of data, the question of modeling

and reconstructing a GRN remains largely unsolved and a big challenge in Systems Biology. The following thesis will contribute to the discourse on the problems of GRN reconstruction.

1.1 Gene Regulation

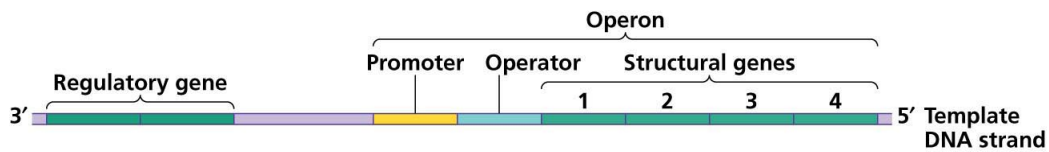
1.1.1 Prokaryotic Gene Regulation

A gene is a portion of DNA responsible for the physical and inheritable characteristics or the phenotype (e.g., the shape, the color, or the size) of all living organisms. It is the way biological information is transmitted through generations and the basis of heredity. Each organism has a certain number of genes, *e.g.* *E. coli*, a bacteria, has between 4,000 and 5,500 known genes. Inside the cells of every living thing, after receiving a signal triggered by distinct factors, each gene is transcribed into mRNA, a kind of RNA, by an enzyme called RNA-polymerase through a process known as **transcription**. Through another process known as **translation**, the mRNA is then transformed into a polypeptide chain, a component of proteins responsible for the observable characteristics of the organism. Gene expression is the process (transcription + translation) in which the biological information contained in a gene is used to synthesize the gene products, which are principally proteins.

To better present and understand the gene regulation process, we will consider the prokaryotes' case, as it is the easiest to comprehend. A eukaryote is an organism whose cells contain a nucleus and other organelles enclosed within membranes, *e.g.* human. A prokaryote is a single-celled organism that lacks membrane-bound organelles such as bacteria. Usually, an organism does not produce all proteins simultaneously because different proteins are involved in different cellular processes. It is important to control how much a gene is expressed at any given time and when a gene is needed. Any disruption to this control can yield serious consequences. For example, it is important for *E. coli* to control the levels of tryptophan (Trp), an essential amino acid for its survival. Hence, if its environment is lacking tryptophan, *E. coli* needs to synthesize the proteins necessary to produce the tryptophan. One can thus define **gene regulation** as the set of mechanisms used by the cell to control (increase or decrease)

the products of gene expression. Figure 2 shows an example the Trp regulation in *E. coli*. In this figure, there are two types of genes. There are transcription factors (TFs) or regulatory genes, which are genes whose products control other genes' expression. When those proteins increase the gene expression, they are called **activators**. Alternatively, when proteins inhibit the expression of genes, we call them **repressors**. In Figure 2, there are also target genes (TGs) that are structural genes that encode proteins not involved in regulation. Gene regulation will manifest differently depending on whether the organism is a prokaryote or a eukaryote, as discussed in Section 1.1.2.

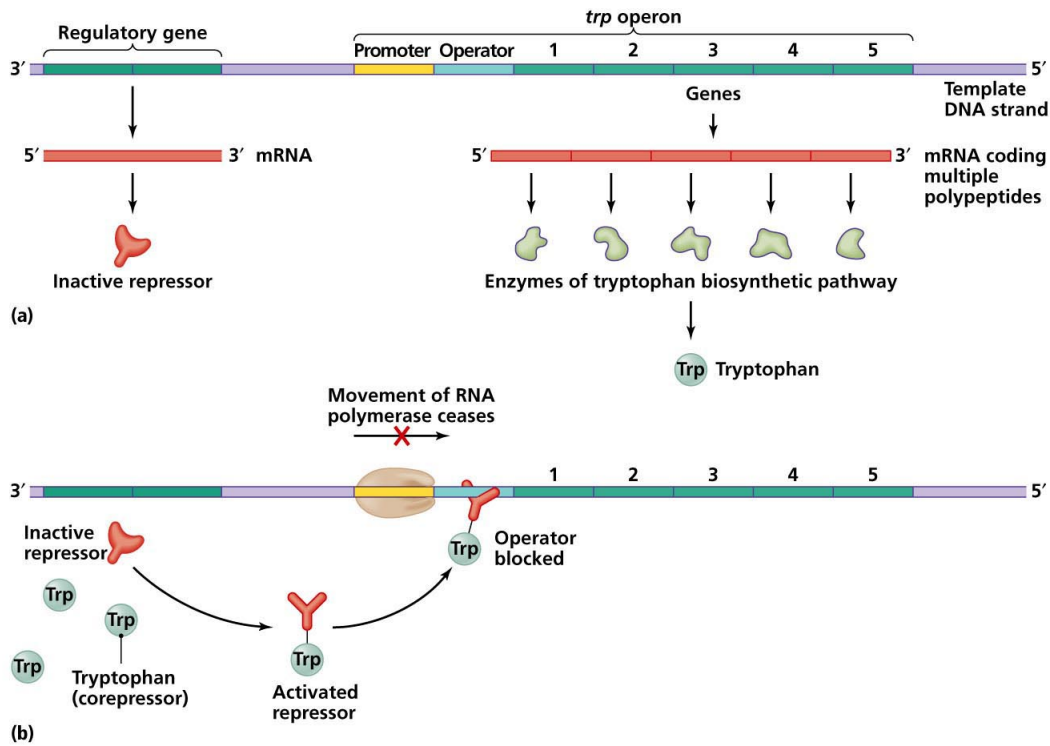
In prokaryotes we identify specific regions of genes: **operons**, **promoter regions** and **transcription factor binding sites (TFBS)**. Genes that produce proteins involved in the same process and are controlled by the same regulatory genes are located next to each other in clusters called operons. RNA polymerase will bind to a promoter region, a sub-region of the non-coding region upstream in an operon. A transcription factor binding site (TFBS), also known as an operator, is another non-coding sub-region where TFs will bind to allow gene regulation. Figure 1 summarizes the structure of a typical operon within prokaryotes.



Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

Figure 1: Organization of an operon in prokaryotes.

Organization of genes in prokaryotes: related structural genes are situated next to each other, forming a cluster called an operon. The operon is under the control of a single promoter—where the RNA polymerase binds—and a single operator—where the TF will bind to control the expression of genes within the operon. This TF comes from the expression of the regulatory gene. The set formed by the promoter, the operator, and the structural genes is called operon [189].



Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

Figure 2: Tryptophan regulation in *E. coli*

The tryptophan regulation in *E. coli*. In (a), the tryptophan is absent in the environment of *E. coli*. A repressor is made from a regulatory gene. However, as the environment lacks trp, it is inactive; thus, it does not bind to the operator. The RNA polymerase can thus transcribe the genes (structural genes) in the operon, and enzymes (here proteins) for the synthesis of tryptophan will be produced. (b) The environment of *E. coli* contains tryptophan, the repressor is active and can thus bind to the operator and block the activity of the RNA polymerase [189].

1.1.2 Eukaryotic Gene Regulation

As in prokaryotes, the process of gene regulation is controlled by proteins which at specific region allow or block the activity of RNA polymerase. However, in eukaryotic cells, gene regulation is far more complicated than in prokaryotic cells. First of all, eukaryotes have more genes than prokaryotes. Nearly all the cells of eukaryotes have the same DNA sequence. However, cell specialization is a result of the difference in gene regulation in these cells.

Another divergence is the organization of genes within the genome. Unlike prokaryotic cells, operons are generally not found in eukaryotes. Instead, each gene is associated with its promoter element where the RNA polymerase and the regulatory protein will bind. The promoter is almost always situated upstream to the coding genes. Most of the time, transcription factor binding sites (TFBS) are located within promoter regions. However, in some cases, TFBS are located far from the promoter, either upstream or downstream from the coding region; they are called enhancers. It worth mentioning that in prokaryotic cells, the expression of genes may be controlled by the action of several TFs [144, 142]. In eukaryotes, gene expression is regulated at different levels, during transcription, and both before and after translation. It contrasts with prokaryotes, where gene regulation happens primarily at the transcription level. Furthermore, a significant difference between the gene regulation in eukaryotic and prokaryotic cells is that, in eukaryotic cells, the DNA sequence is compacted around a protein called a histone, forming the nucleosome. Nucleosomes are assembled into a compact structure called chromatin. The chromatin can either promote or prevent genes regulation. TFs and RNA polymerase cannot access the target gene when the DNA is compacted around the histone. Figure 4 summarizes how the DNA is packed in the eukaryote genome.

1.2 Gene Regulatory Network

A gene regulatory network is a set of all elements (transcription factors, genes, or RNA) that interact together directly or indirectly to control genes' expression. In this thesis, we will only consider the transcriptional level of regulation. Accordingly,

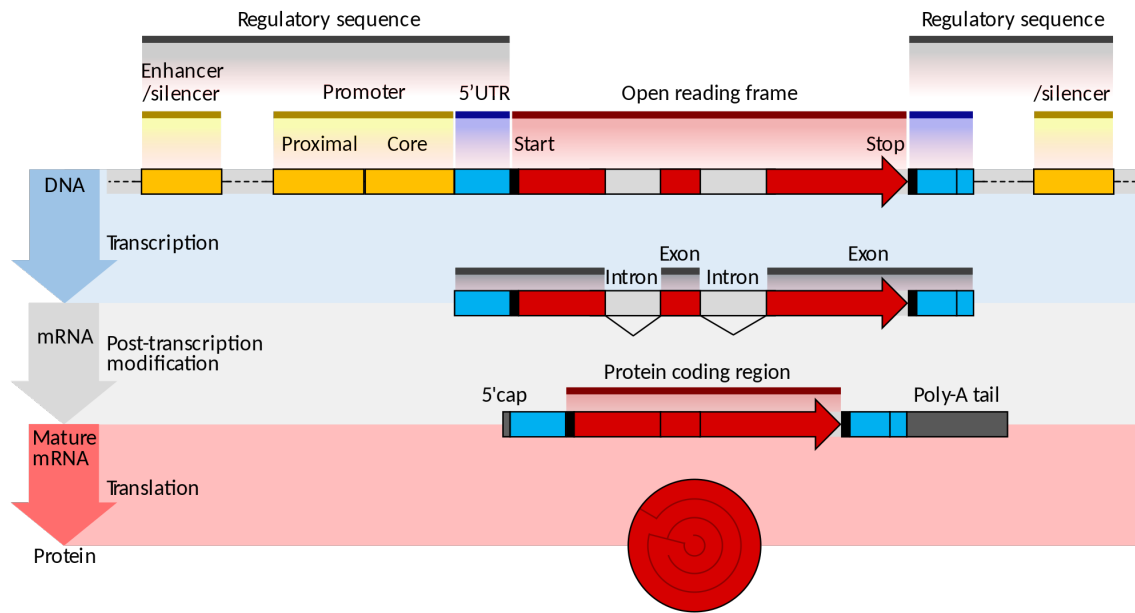


Figure 3: Eukaryotic gene structure

Organization of a gene within the genome of a eukaryote. The open reading frame contains the DNA sequence (target gene) transcribed by RNA polymerase. The promoter contains regions where a variety of TFs may bind, allowing the RNA polymerase to transcribe the adjacent gene: this is **gene expression**. Note that the RNA polymerase also binds in the promoter region, particularly in the core promoter region. Furthermore, the TFs can also bind in distant regions called enhancer or silencer regions, which also control gene expression.

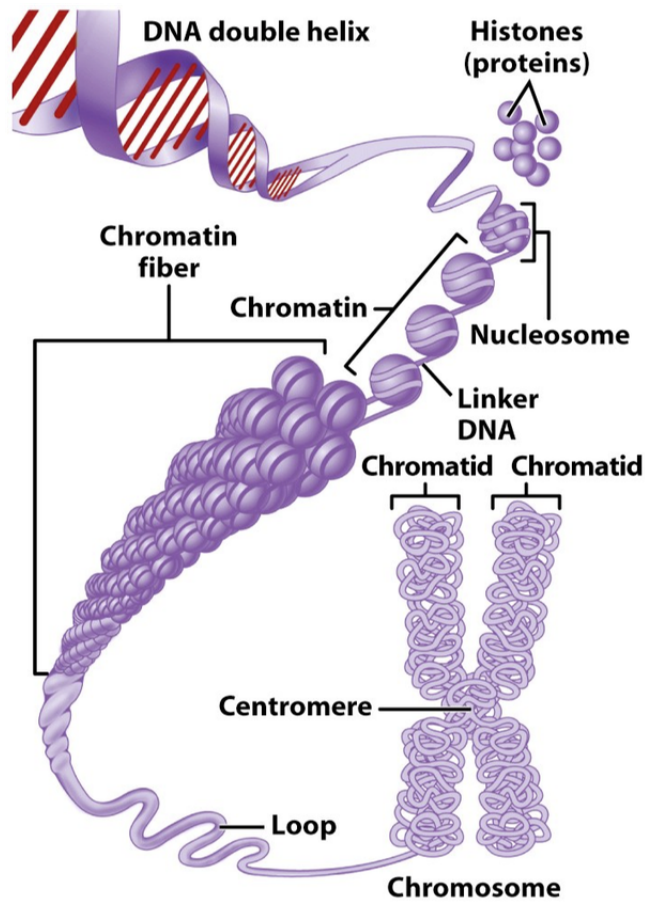


Figure 3-23 Principles of Anatomy and Physiology, 11/e
© 2006 John Wiley & Sons

Figure 4: Chromatin in eukaryotic cells

The figure show different scale how a chromosome in a eukaryotic the cell [231].

the gene regulatory network (GRN) will be the set of target genes (TGs) and transcription factors (TFs) that interact together through relations called **regulatory links**. Figure 5 shows a simplifying picture of the gene regulatory network consisting of a set of target genes and transcription factors and their regulatory interactions.

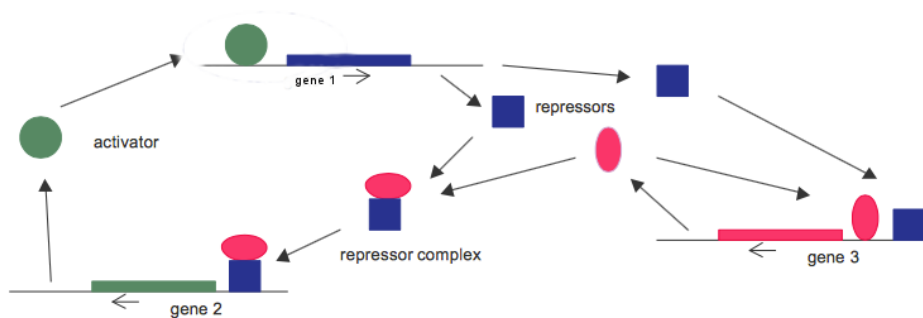


Figure 5: Gene regulatory network abstraction

The figure presents an abstraction of the gene regulatory network [193]. It consists of a set of genes, their expression products, and the regulatory interactions that exist between them.

Several studies [3, 4, 5] have demonstrated that, like many real networks, the out-degree of genes in the GRN follows a scale-free distribution. Following a scale-free distribution indicates that most of the TFs are connected to a small number of genes, while only a few TFs regulate many genes. TFs that regulate a multitude of genes are called hub genes. This particular organization of the GRN ensures its connectivity and integrity [4, 5]. The presence of hub TFs in the GRN make it robust against random disruption [3], as they will generally affect non-hub genes, and, will consequently not lead to a loss of connectivity. Hubs are essential for the GRN and are generally the target of diseases like cancer. Given their importance, researchers have hypothesized that hub genes are subject to strict evolutionary constraints.

Apart from the gene connectivity distribution, the GRN has long been thought to have a modular organization that is a critical feature for the cell to coordinate its complex functions (the different tasks are split over the modules which can either interconnected or be insulated from) [95, 185]. Albert-László *et al* have defined a module as a set of physically or functionally linked molecules that work together

to achieve distinct functions [11]. Given the GRN, a module will refer to a set of genes involved in a joint elementary function, sharing the same behavior (expression pattern) and under the control of a set of regulators that controls their expression. A gene can be part of multiple modules at a time, which implies that the functional modules overlap each other.

1.3 Problem Statement

Networks are omnipresent in biology and widely used to represent different kinds of information and most likely interactions. There exist several types of networks. For example, Protein-Protein networks that model the physical interactions of proteins or metabolic networks that comprehensively describe all possible biochemical reactions for an organism.

Gene expression regulation differs between eukaryotes and prokaryotes. In prokaryote, the regulation is much simpler and happens at the transcription. However, in eukaryotes, gene expression regulation is more complex and happen at several levels:

- At the epigenetic level: i.e., when the DNA is unwound and loosened from the nucleosome to allow the transcriptional machinery to start the transcription
- At the transcriptional level, i.e., when the DNA is transcribed into RNA
- At the post-transcriptional level, i.e., after the transcription but before the RNA is translated into protein
- At the post-translational level, i.e., after the RNA is translated into proteins.

In this work, we restrict the GRN at the transcriptional level where most of the genes are regulated [20]: it is the transcriptional gene regulatory network (TRN). The TRN offers a condensed view of the regulation. In what follows, the TRN represents the GRN. Restricting the expression to the transcriptional level. Restricting our model to transcription will ignore other types of regulation.

The GRN is generally represented as a graph. In this graph, the nodes are all the genes acting in the regulation or even modules of co-expressed genes. The graph can be directed or not. In this graph, a directed edge communicates the direct causal

relationship from a transcription factor (the source) to its target gene (the sink). Note that the edges can be signed, with a positive sign denoting activation and negative sign repression.

Our research focuses on reverse-engineering the directed unsigned graph of the interacting genes at the transcription level, forming the GRN. Our problem is a binary classification problem in which we seek to infer whether or not there is an interaction between each TF and the TGs. Our model does not report other information about regulation, such as the interaction type (enhance or repress), the TF's influence degree on a TG, or the way TFs associate together.

Given that the GRN graph structure is unknown, the computational problem of GRN inference amounts to reverse-engineering the graph structure (i.e., the list of the edges) between all the TFs and genes. One uses as input for this computational problem the available high-throughput omics data, such as expression data or sequence data. The output is the graph of the interactions between the TFs and the TGs.

1.4 Motivation

A model is anything that one uses as a substitute for a system we wish to understand [21]. GRN modeling is an iterative process in which available high-throughput data is used to build and refine a model (the links within the graph), representing a GRN. Roughly speaking, the goal of GRN modeling is to answer the following four principal questions:

1. Why do cells in organisms have different properties even though they all have the same genetic information: the same DNA?
2. How does a cell in an organism know which genes to express at a particular time?
3. What is the full range of behavior that the system will exhibit if some parts stop functioning, or if the organism is exposed to different conditions?
4. How robust is the system under extreme conditions?

In a nutshell, modeling and reconstructing a GRN is essential for understanding, visualizing, exploring, and analyzing the regulatory process [173, 21, 98].

Understanding. Modeling a GRN provides scientists with a framework and an abstraction at the genome-scale for understanding the principles behind gene regulation. It allows automatic interpretation and greater scrutiny of a GRN, thus revealing the hidden properties of the GRN. Furthermore, modeling a GRN is a way to link cellular processes and states to physical states, thus helping to understand why, given some conditions, we observe a particular phenotype. The different phenotypes that an organism adopts originate from complex molecular processes occurring within the cell, making it challenging to decipher simply through lab experiments. For example, modeling facilitates an analysis of which cellular states lead to complex diseases such as cancer. In a sense, modeling will help to underline or define the states associated with the observed disease. Moreover, modeling the GRN can serve as scaffold information to extract local or global properties that, once demonstrated to be statistically different from random networks, can be related to a better understanding of biological processes.

Analyzing and reasoning. By modeling a GRN, scientists have a mechanism for examining the actions of many genes simultaneously under different given conditions, thus enabling them to predict how cells behave under new conditions automatically. Also, it has the potential to facilitate experiments conducted at a large scale, such as simulations, that would alternatively need to be conducted in a wet lab experiment at a much higher cost. Hence, lab scientists will benefit from engaging in modeling as a part of their work. They will be better able to derive novel biological hypotheses about how those conditions affect the molecular interactions that can be later investigated in wet-lab experiments such as gene expression experiments. Moreover, scientists will have a view of the GRN as a whole rather than a collection of single biological entities, offering insights on how to optimize and control parts of the network while having global knowledge of how it will affect the whole network. Finally, modeling and reconstructing a GRN will facilitate information transfer from well-studied organisms to unknown organisms.

Visualizing. Modeling a GRN will provide scientists a way to visualize extremely large-scale complex relationships among elements operating in the GRN, thus serving as a map or a blueprint of molecular interactions within the cells.

1.5 Challenge in Gene Regulatory Network Inference

GRN inference is a daunting problem in Systems Biology. Scientists face several difficulties. The following list gives an overview of the problems they face:

- The data obtained from high-throughput experiments are noisy. If we consider microarray data, they contain a noise magnitude of 20–30% [2]. This noise has several origins, such as measurement errors. The difficulty here lies in dissociating real gene expression values (real signal) from experimental noise [183]. In Chapter 2, we present reverse-engineering methods that use various strategies to infer a GRN from noisy expression data.
- The amount of experimental data available is minimal, as it is mainly the case for expression data. Data availability restriction seems paradoxical with current high-throughput facilities. Although it is now possible to experimentally investigate a considerable number of genes simultaneously, the number of samples available has not and cannot be expanded in the same way because of limitations such as cost. The results are datasets, where the number of genes is far higher than the number of samples. It is known as the high dimension, low sample problem [91]. When the number of dimensions increases, the amount of data needed to represent the data accurately increases exponentially. This phenomenon is known as the curse of dimensionality problem [40]. As such, data obtained from gene expression experiments is sparse, compounding the problem of the GRN model complexity stemming from the innate complexity of gene regulation itself. Furthermore, the GRN model is very complex due to the complexity of gene regulation itself. There is a strong relation between model complexity, the amount of data required to construct the model, and the constructed model's quality. Due to this connection, the development of an accurate and complex genome-scale GRN model is difficult. Some computational methods break down when data is sparse [98]. In section Chapter 2, we will discuss some statistical methods and the strategies they use to deal with the problem of data sparsity. In Chapter 3, a new solution is proposed to cope with

the data's limited availability.

- It is challenging to distinguish direct from indirect regulation [82]; gene regulation is a complex process. For example, at a certain time a gene (name it *genea*) within the cell may be activated by a TF (name it *TFa*) that we know is a protein which originates from expression from another gene (name it *geneb*) which in turn is activated by another TF (name it *TFb*). Consequently, *TFb* will indirectly influence the expression of the former gene. Looking at the expression profile, it becomes difficult to recognize that the *TFb* does not directly interact with the *genea*.
- High dimension data that is available today represent only a snapshot of a particular cell state and time interval of the cell's life. So we miss several cell states. Thus, data obtained is incomplete, resulting in a limited understanding of how all functional units are put together in the cell [200]. Moreover, most lab measurements (gene expression, proteins-DNA interactions) are on cell populations. Even though they have the same genetic information, cells can exhibit a significant difference in the amount of gene expression products. These measurements result in an averaging of the behavior of the cells that may cause a loss of relevant information such as relevant events that may occur in a particular cell but may not be present at the global view [54].
- It is challenging to identify regulatory sequences because they are short sequences in the midst of a lot of noise. Moreover, those sequences are highly variable, and they are repeated frequently in the genome. Some of those repetitions do not represent regulatory sequence at all [230, 46, 30]. Several algorithms that try to overcome this problem using different strategies to find TFBS in a set of sequences have been proposed in the literature. In Chapter 2, we will present some state of the art solutions.
- Our knowledge of the encoding regulatory elements in genomes remains elementary [218, 22]. It results in myriads of available sequences, of which only a small fraction have been functionally annotated [30].
- The limited number of available well-studied organisms remains a significant problem in the research. Thus, the number of well-reconstructed gold standard

GRN remains limited. This constraint causes a problem, particularly when scientists want to assess the inferred networks or assess the performances of the methods used to infer the network. A solution to this problem is presented in Section 2.4.3.

1.6 Limitation of State-of-the-Art

Gene regulatory network inference is a long-standing problem in systems biology. Many solutions have been proposed in the literature, but they still present some limitations that render the inference an unresolved problem. Among the limitations we can list:

- The use of only one type of data for the inference of the network. In effect, the rapid technological advances have led to the production of different types of biological data that carry on complementary but incomplete knowledge about the regulation; the GRN inference of networks using only one type of biological data leads to incomplete and less accurate GRNs.
- Most existing algorithms for GRN inference based on expression profiles assume a linear dependency among genes. However, the dependencies involved in regulation are too complex to explain using a simple linear model.
- The majority of existing studies that reconstruct the GRN have focused on inferring individual regulatory links. These algorithms try to elucidate all the regulatory links between all the candidates' genes, given the limited availability of data, leading to many more false positives than true positives.

1.7 Contribution

The contributions of this research project are summarized as below:

- Implementation of a GRN inference method that uses Elastic Net for feature selection.
- Implementation of a method that integrates several types of omics data with expression data for GRN inference.

- Reconstruction of the gene regulatory network that controls the cell cycle in a model organism: human.

1.7.1 BENIN: Network Inference as Feature Selection using Elastic Net

Gene regulatory network inference is one of the central problems in computational biology. Researchers have developed computational methods to reverse-engineer the GRN using varied mathematical models, ranging from Boolean networks [146], Information theory [272], correlation [248], Bayesian networks [258] and differential equations [36]. In this thesis, we introduce **BENIN: Biologically Enhanced Network INference**. **BENIN** is a simple and intuitive inference method for integrating any prior knowledge data with time-series expression data. **BENIN** states GRN inference as a feature selection problem: finding the direct regulators of each gene. It assumes that a target gene's expression profile is a linear function of its direct regulators' expression profiles. **BENIN** applies a regression technique called *Elastic Net*, combined with a resampling technique to perform feature selection.

1.7.2 BENIN: Integration of Prior Knowledge data

The advent of high-throughput technologies such as DNA microarray, RNA-seq, or ChIP-seq has triggered the production of a large variety of data that is stored in diverse curated databases. This data drives machine learning challenges, particularly for systems biology, such as GRN inference. Common problems in GRN inference include the poor knowledge of cell function, the limited number of samples compared to the number of genes being studied, and the data's noisy nature.

Data integration is a common approach to improve inference. Researchers have proposed several ways to combine expression data with prior knowledge available in data such as pathways [216], protein-protein interactions [271], gene annotation data [177], sequence data [80], literature [140] or functional association [223]. Most use the Bayesian network framework to include prior information into GRN inference. However, the Bayesian approach has many drawbacks when applied to high-dimensional data and requires deep knowledge of the prior for good integration. Moreover, many existing methods are designed for a specific type of prior knowledge.

In this work, we used the *Adaptive Elastic Net*, a modified version of the *Elastic Net*, to include prior knowledge. In this work, we consider different types of prior knowledge data:

- Knockout (KO) and Knockdown (KD) gene expression data. They are expression data measured in an organism where a transcription factor is made inoperative (KO expression data), or its expression is reduced (KD expression data). This data type is integrated either through the z-score (for KO data) or the probabilistic framework (for KD data).
- ChIP-seq data. They report regions in the genome where a specific transcription factor (TF) will physically bind to the DNA to, for example, control the expression of proximal genes. They are obtained through *in vivo* experiments. These kinds of data are integrated through the computation of a score that measures potential binding between each TF and all the genes in the genome.
- Functional annotation, which reports the gene ontology (GO) annotation for a gene’s function. For a specific gene, the annotation is a set of terms that captures the gene’s current biological knowledge. We consider the functional similarity between genes by comparing their functional annotations and computing a similarity score, which will be integrated into BENIN.
- TFBS, which are reported in term matrices, which store binding specificity for a specific TF. We used this data to scan the genome’s region of interest, and the result of the scanning process is integrated through a probabilistic framework into BENIN to boost the network inference.
- Genome-wide location data use p-values to report physical interactions between TFs and genes of the organism of interest. We integrated genome-wide location into BENIN in a probabilistic manner.

The probabilistic framework is defined through the Bayes formula. BENIN allows for control of the impact of the prior on the model. BENIN is generic enough to integrate any type of data.

BENIN allows the integration of regulatory information across species. Comparative studies have demonstrated that GRNs from closely related species may share

conserved topological properties known as kernel components [64, 100, 227]. GRN inference in an organism can thus leverage knowledge and findings of regulatory networks from other well-known organisms. The key idea behind information transfer among related species is the conservation of biological function among orthologous genes. Hence, the assumption is that orthologous transcription factors regulate orthologous genes. The challenge here is to define “True” orthologous genes for a reliable transfer of information. Orthology should be distinguished from paralogy in which the biological function is not preserved. Many existing algorithms infer the GRN either based on the expression data alone or through comparative evolution solely. However, integrating both strategies may help refine GRNs inferred from expression data and, besides, will enrich the network with new potential regulatory interaction. We extended **BENIN** to include orthologous regulatory information from model organisms, through orthology-based information transfer.

1.7.3 Application of **BENIN** to Human cell Cycle

The cell cycle is a fundamental biological process that occurs in all living cells and is essential for their survival. Cell division is a highly regulated process. Proper regulation of gene activities during the cell cycle is critical for the well functioning of several cellular processes and accurate transmission of the genetic information. A disruption to this regulation may lead to complex and irreversible phenotypes. Therefore, it is crucial to unravel the network of interacting molecules controlling the cell cycle to get insights into both normal and abnormal cell divisions related to diverse pathological phenotypes.

We used **BENIN** to infer the GRN that controls the cell cycle of the HeLa cell cycle. The HeLa cell line is a cancerous human cell line. We integrate prior knowledge from diverse sources: ranging from TFBS information, knock-down gene expression data, functional annotation, and ChIP-seq data. Several studies have suggested conservation of the general mechanism of cell cycle regulation among vertebrates [18, 55]. Hence, we refined the regulatory network inferred from expression data and prior biological knowledge with regulatory information from orthologous genes in the mouse model organism through sequence orthology detection.

1.7.4 List of publications

- Kamgnia, S., & Butler, G. (2019, December). BENIN: combining knockout data with time-series gene expression data for the gene regulatory network inference. In Proceedings of the Tenth International Conference on Computational Systems-Biology and Bioinformatics (pp. 1-9). [123].
- Wonkap, S. K., & Butler, G. (2020). BENIN: Biologically enhanced network inference. *Journal of Bioinformatics and Computational Biology*, 18(03), 2040007 [250]

1.8 Organization of the Thesis

The thesis is structured as follows:

Chapter 2 details the background notions needed to comprehend this dissertation. It then follows an analysis of the data available to overcome this challenge and the strategies available to evaluate GRN inference algorithms. Then it explores the different methods that have been undertaken to reconstruct the gene regulatory network.

Chapter 3 introduces BENIN, a GRN inference algorithm for multiple data integration, and details its results on the DREAM4 challenge.

Chapter 4 presents the results of applying BENIN to infer the gene regulatory network that controls the Human HeLa cell cycle. It also offers an extension of BENIN to integrate regulatory information from other model organisms through sequence homology for the gene regulatory network inference.

Chapter 5 concludes this thesis by highlighting our different results, findings, and points for future work.

Chapter 2

Background

With the availability of a deluge of genomic data, we now witness many algorithms' emergence to tackle the GRN modeling. The chapter covers the mathematical background notions such as Bayesian networks, feature selection, and regression. The chapter gives an overview state of the art methods for gene regulatory network inference. Hence, Section 2.1 defines machine learning and statistical notions. Section 2.5 presents the three main methodologies introduced in the literature for regulatory network inference. We give for each methodology some state-of-the-art works proposed in the literature.

2.1 Background for Network Inference

This section highlights critical aspects of statistics and machine learning relevant to this thesis: Bayesian networks; the notion of mutual information; Elastic Net and regression; the vector autoregressive model; the Granger causality; the stationary bootstrap; a position weight matrix; a consensus sequence and finally a DNA motif.

2.1.1 Bayesian Network

Graphical models are robust and extremely popular tools to model uncertainty[134]. They allow us to deal with uncertainty with the use of probability theory and cope with complexity through graph theory. The most common type of graphical model is the Markov network and the Bayesian Network, also known as the causal network.

In this thesis, we only consider Bayesian Network; the Markov network is out of the thesis's scope.

Let consider a set $U = \{X_1, X_2, \dots, X_n\}$ of discrete variables, where each X_i may take values from a finite set. A Bayesian Network is a representation of the joint probability distribution of a set of random variables U . More formally, a Bayesian network is defined as a pair $N = \langle G, \Theta \rangle$. G is a directed acyclic graph whose vertices are the random variables X_i , and the edges represent the direct probabilistic dependencies between the variables. G encodes an independence assumption, which states that each variable X_i is independent of the variables in $\{X_1, X_2, \dots, X_{i-1}\}$ given its parents $Pa^G(X_i)$ (set of variables connected to X_i in G) in G . The second component, Θ , describes the conditional distribution for each X_i given $Pa^G(X_i)$. The overall model defines an unique joint probability distribution on X_1, X_2, \dots, X_n such that:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa^G(X_i)) \quad (1)$$

Bayesian networks are suitable for modeling and learning causal relationships. An extension of Bayesian Network was introduced, which allows handling time series or sequential data: the Dynamic Bayesian Network (DBN) [167, 75]. It allows representing dynamic processes that evolve through time. It extends the set of random variables in the model (in the graph). Now, each node in the graph represents a variable at a specific time point t . In this new graph, a node can only be connected to another node in subsequent time points. This restriction is to ensure the DAG nature of the graph. In a DBN, the state of variable at time time $T = t + 1$ is conditionally dependent on the values of its parents through the interval $T = 1$ to $T = t$. More formally, let X_i^{t+1} a random variable X_i at time $T = t + 1$; let $Pa^G(X_i)^{[1,t]}$ the set of X_i parent variables through the time interval $[1, t]$, the new joint distribution is defined as:

$$P(X_1^{t+1}, X_2^{t+1}, \dots, X_n^{t+1}) = \prod_{i=1}^n P(X_i^{t+1} | Pa^G(X_i)^{[1,t]}) \quad (2)$$

2.1.2 Mutual information

Mutual information is a positive quantity that measures how much a random variable X tells us about another Y and vice versa: it measures the information shared by both variables. It is generally used as a powerful tool to measure the nonlinear dependency between two variables. Let X with alphabet \mathcal{X} a random variable with

probability distribution $p(x) = Pr\{\mathcal{X} = x\}$. Let Y with alphabet \mathcal{Y} a random variable with probability distribution $p(y) = Pr\{\mathcal{Y} = y\}$. The mutual information $I(X; Y)$ between X and Y is defined as:

$$I(X; Y) = \sum_{x \in X, y \in Y} P(x, y) \log \frac{P(x, y)}{P(x) P(y)} \quad (3)$$

where $P(x, y)$ the joint distribution of X and Y . The mutual information is a symmetric measure. We have:

$$I(X; Y) = I(Y; X) \quad (4)$$

A value of $I(X; Y) = 0$ indicates that the two variables are independent, and a high value indicates a high correlation between the variables.

2.1.3 Regression Technique

Linear regression is a statistical method for modeling the linear relationship between a dependent variable and a set of predictor variables. This linear relationship takes the form $\vec{y} = \mathbf{X}\vec{\beta} + \vec{\xi}$, where $\vec{y} = (y_1, \dots, y_N)^T$, is an N vector representing the dependent variable with $y_i \in \mathbb{R}$. $\mathbf{X} = (\vec{x}_1, \dots, \vec{x}_N)^T$, $\vec{x}_i \in \mathbb{R}^M$, is the $N \times M$ matrix of explanatory variables, and, $\vec{\beta} = (\beta_0, \beta_1, \dots, \beta_M)^T$ is the M coefficients vector and finally, $\vec{\xi}$ is the error vector of size N . For simplicity we will assume that \mathbf{X} is standardized, i.e. $\sum_{i=1}^N x_{ij} = 0$, $\frac{1}{N} \sum_{i=1}^N x_{ij}^2 = 1$ for $j = 1, 2, \dots, N$

Usually, an estimation $\vec{\beta}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y}$ of $\vec{\beta}$ is obtained by minimizing the residual sum of square (RSS) defined in Equation 5

$$RSS(\vec{\beta}) = \sum_{i=1}^N (y_i - \vec{x}_i^T \vec{\beta})^2. \quad (5)$$

However, when the number of variables M becomes very large compared to the number of samples N , i.e., $M \gg N$ (high dimensional problem), many of these variables may be irrelevant to the output, and a large number of them are highly correlated (multicollinearity problem). Therefore, the matrix $\mathbf{X}^T \mathbf{X}$ will be singular (the matrix is not invertible), and the estimated $\vec{\beta}_{OLS}$ will no longer exist [69]. Moreover, the multicollinearity in the input matrix causes the OLS estimation not to be robust [160]. In fact, in this setting, the problem becomes ill-posed and small changes

in the input matrix may lead to big changes in the OLS estimate. Hence, we can no longer use the vector that minimizes Equation 5 as an estimation of $\vec{\beta}$ [174, 69]. All these may suggest a parsimonious coefficient vector $\vec{\beta}$, such as keeping the model a smaller set of the most relevant predictors, leading to a more relevant and meaningful model.

Several solutions have been proposed in the literature to tackle the problem by introducing a penalty to the residual sum of square. Thus, instead of minimizing Equation 5 we minimize Equation 6,

$$RSS_P = RSS(\vec{\beta}) + P_\lambda(\vec{\beta}) \quad (6)$$

where $P_\lambda(\vec{\beta})$ is a function that penalizes the values of the parameters we are looking for (here $\vec{\beta}$), and λ is a parameter that controls the trade-off between penalization and likelihood. Different penalties have been introduced in the literature, but we will only consider three of them. Interested reader can refer to [163, 39, 68] for a detailed description of other penalization techniques.

2.1.3.1 Ridge Regression

The *Ridge* regression was introduced by Andrey Tikonov [101]. It minimizes the l_2 penalized RSS described in Equation 7.

$$\begin{aligned} \vec{\beta}_{ridge} &= \underset{\vec{\beta}}{\operatorname{argmin}} \quad RSS(\vec{\beta}) + \lambda \|\vec{\beta}\|_2^2 \\ &= \underset{\vec{\beta}}{\operatorname{argmin}} \quad RSS(\vec{\beta}) + \lambda \sum_{j=1}^M \beta_j^2 \end{aligned} \quad (7)$$

The parameter $\lambda \geq 0$ controls the strength of the penalty, which increases with the values of λ . λ is dependent on the data, and it is generally estimated with data-driven methods like cross-validation.

The **Ridge** penalization is ideal when dealing with many predictors variables, each having a small effect on the dependent variable. It prevents the low prediction of the regression coefficients when many of the predictors are correlated. The **Ridge** shrinks the coefficients of the correlated predictors equally towards zero [72, 169] without setting them to zero. As a consequence, **Ridge** regression does not select the most

informative predictors. Instead, it minimizes their impact on the model, which may still be uninterpretable.

2.1.3.2 LASSO

The limitation of the **Ridge** has led to the introduction of the **LASSO** of Tibshirani [228]. The **LASSO** uses $L1$ -norm to penalize the coefficients vector $\vec{\beta}$ and minimizes the optimization problem describes in Equation 8.

$$\begin{aligned}\vec{\beta}_{Lasso} &= \underset{\vec{\beta}}{\operatorname{argmin}} \operatorname{RSS}(\vec{\beta}) + \lambda \|\vec{\beta}\|_1 \\ &= \underset{\vec{\beta}}{\operatorname{argmin}} \operatorname{RSS}(\vec{\beta}) + \lambda \sum_{j=1}^M |\beta_j|\end{aligned}\tag{8}$$

The **LASSO** shrinks many unimportant predictors coefficients exactly to zero, with only a small subset of nonzero coefficients. Since it selects some variables among the set of predictors, the **LASSO** can be regarded as a feature selection method. λ controls the sparsity of the model. The **LASSO** regularization allows shrinking unimportant variables to zero. The obtained model is thus more interpretable. Like with **Ridge** regression, **LASSO** is good at dealing with many input variables. However, it presents some drawbacks. The **LASSO** is not efficient when many of the predictors are correlated. In this situation, it will randomly choose one of the predictors amongst the correlated predictors that will be included in the model. Hence, if all the predictors are correlated, the **LASSO** will break down. Furthermore, when $M \gg N$, **LASSO** selects at most N variables before it saturates.

2.1.3.3 Elastic Net

More recently, a new regularization has been proposed to solve the **LASSO**'s limitations: the **Elastic Net** of Zou and Hasti [274]. It combines the idea of the **Ridge** and **LASSO** regression and solves the optimization problem described in Equation 9.

$$\begin{aligned}
\vec{\beta}_{ENet} &= \underset{\vec{\beta}}{\operatorname{argmin}} \operatorname{RSS}(\vec{\beta}) + \lambda_1 \|\vec{\beta}\|_2^2 + \lambda_2 \|\vec{\beta}\|_1 \\
&= \underset{\vec{\beta}}{\operatorname{argmin}} \operatorname{RSS}(\vec{\beta}) + \lambda \left[(1 - \alpha) \|\vec{\beta}\|_2^2 + \alpha \|\vec{\beta}\|_1 \right] \\
&= \underset{\vec{\beta}}{\operatorname{argmin}} \operatorname{RSS}(\vec{\beta}) + \lambda \left[(1 - \alpha) \sum_{j=1}^M \beta_j^2 + \alpha \sum_{j=1}^M |\beta_j| \right]
\end{aligned} \tag{9}$$

where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ and $\lambda = \lambda_1 + \lambda_2$. As previously, λ controls the degree of regularization while α controls the tradeoff between ridge and lasso regression. **Elastic Net** is equivalent to **Ridge** regression for $\alpha = 0$ and to **LASSO** when $\alpha = 1$. By combining both regularizations, the **Elastic Net** integrates the advantages of both techniques and overcomes the drawbacks of each regularization taken separately. The l_1 part performs the variable selection, while the l_2 part favors the grouped selection and stabilizes the solutions path with respect to random variable selection therefore, improving the solution. With the grouping effect, the **Elastic Net** ensures that the group of correlated variables will get approximately the same magnitude of coefficients. When $M \gg N$ the **Elastic Net** is capable of selecting more than N variables[169]. However, the **Elastic Net** lacks the oracle property. From the work of Fan and Li [67], a method is said to have the oracle property if it can asymptotically estimates the zero coefficients of the true parameter vectors as exactly zero with a probability close to one, as if the true zero coefficients were known beforehand; and it remains consistent with the estimate of the nonzero coefficients.

2.1.3.4 Adaptive Elastic Net

Several efforts have been made to extend the *Elastic Net* to remedy the lack of oracle property. The **Adaptive Elastic Net** was introduced by Zou *et al* Hastie [273, 72] which solve the optimization problem in Equation 10:

$$\lambda \sum_{j=1}^M \nu_j P_\alpha(\beta_j) = \lambda \sum_{j=1}^M \nu_j (1 - \alpha) \beta_j^2 + \alpha \sum_{j=1}^M |\beta_j|, \tag{10}$$

where ν_j ($j = 1, 2, \dots, M$) are the adaptive data driven weights. These weights allow applying different levels of shrinkage to the predictors variables regarding the prior

knowledge or bias over these variables [72]. The idea is to give large weights ν_j to unimportant variables, and thus to heavily shrink their corresponding coefficient; on the other hand give small weights ν_j to important variables to slightly shrink their associated coefficients. Therefore, the larger is ν_j the more penalized will be β_j .

2.1.4 The p -order Vector Autoregressive Model

The vector autoregressive model (VAR) is one of the easiest models and the most used to analyze and capture interdependencies among multiple time series. In a VAR(p) model, each variable is expressed as a linear combination of a constant c , the p lags of its own values as well as the p lags of the other variables in the model and finally, an error term $\vec{\xi}$. Let $\vec{x}_t = (\vec{x}_{1,t}, \vec{x}_{2,t}, \dots, \vec{x}_{M,t})^T$ be an M -dimensional multiple time series data vector; \vec{x}_t is assumed to be generated from a VAR(p) if it can be written as in Equation 11.

$$\begin{bmatrix} \vec{x}_{1,t} \\ \vec{x}_{2,t} \\ \vdots \\ \vec{x}_{M,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{bmatrix} + \begin{bmatrix} a_{1,1}^1 & a_{1,2}^1 & \cdots & a_{1,M}^1 \\ a_{2,1}^1 & a_{2,2}^1 & \cdots & a_{2,M}^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_{M,1}^1 & a_{M,2}^1 & \cdots & a_{M,M}^1 \end{bmatrix} \begin{bmatrix} \vec{x}_{1,t-1} \\ \vec{x}_{2,t-1} \\ \vdots \\ \vec{x}_{M,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a_{1,1}^p & a_{1,2}^p & \cdots & a_{1,M}^p \\ a_{2,1}^p & a_{2,2}^p & \cdots & a_{2,M}^p \\ \vdots & \vdots & \ddots & \vdots \\ a_{M,1}^p & a_{M,2}^p & \cdots & a_{M,M}^p \end{bmatrix} \begin{bmatrix} \vec{x}_{1,t-p} \\ \vec{x}_{2,t-p} \\ \vdots \\ \vec{x}_{M,t-p} \end{bmatrix} + \begin{bmatrix} \vec{\xi}_{1,t} \\ \vec{\xi}_{2,t} \\ \vdots \\ \vec{\xi}_{M,t} \end{bmatrix} \quad (11)$$

or equivalently

$$\vec{x}_t = \vec{c} + \mathbf{A}_1 \vec{x}_{t-1} + \cdots + \mathbf{A}_p \vec{x}_{t-p} + \vec{\xi}_t, \quad (12)$$

where p denotes the lag length or the order of the VAR model; \mathbf{A}_i is a $M \times M$ matrix of coefficients, M represents the number of variables in the time series; $\vec{\xi}_t$ is a M -dimensional white noise vector, *i.e.* $E(\vec{\xi}_t) = 0$, $E(\vec{\xi}_t, \vec{\xi}_t) = \Sigma$ and $E(\vec{\xi}_t, \vec{\xi}_{t-k}) = 0$. From the system of equations in Equation 11, each variable in the time series can be separately written as follows:

$$\begin{aligned} \vec{x}_{1,t} &= c_1 + a_{1,1}^1 \vec{x}_{1,t-1} + a_{1,2}^1 \vec{x}_{2,t-1} + \cdots + a_{1,M}^1 \vec{x}_{M,t-1} + \cdots + a_{1,1}^p \vec{x}_{1,t-p} + a_{1,2}^p \vec{x}_{2,t-p} + \cdots + a_{1,M}^p \vec{x}_{M,t-p} + \vec{\xi}_{1,t} \\ \vec{x}_{2,t} &= c_2 + a_{2,1}^1 \vec{x}_{1,t-1} + a_{2,2}^1 \vec{x}_{2,t-1} + \cdots + a_{2,M}^1 \vec{x}_{M,t-1} + \cdots + a_{2,1}^p \vec{x}_{1,t-p} + a_{2,2}^p \vec{x}_{2,t-p} + \cdots + a_{2,M}^p \vec{x}_{M,t-p} + \vec{\xi}_{2,t} \\ &\vdots \\ \vec{x}_{M,t} &= c_M + a_{M,1}^1 \vec{x}_{1,t-1} + a_{M,2}^1 \vec{x}_{2,t-1} + \cdots + a_{M,M}^1 \vec{x}_{M,t-1} + \cdots + a_{M,1}^p \vec{x}_{1,t-p} + a_{M,2}^p \vec{x}_{2,t-p} + \cdots + a_{M,M}^p \vec{x}_{M,t-p} + \vec{\xi}_{M,t} \end{aligned} \quad (13)$$

Equation 12 can be solved by any regression algorithms: either *OLS* or penalized regression algorithms.

2.1.5 Granger Causality

The notion of Granger causality [84] is a widely used concept introduced by the Nobel prize-winning economist Clive Granger, to analyze the relationship between time series. It is based on the intuition that a cause always comes before its effects. Hence, a time series variable \vec{y}_t is said to Granger cause another \vec{x}_t , if the prediction of \vec{x}_t in term of its own lagged values and the lagged values of \vec{y}_t are better than the prediction of \vec{x}_t based only on its own lagged values. This means that, in the general VAR(p) process described in Equation 12, a variable $\vec{x}_{i,t}$ is called a Granger cause of another $\vec{x}_{j,t}$ if at least one element of $\mathbf{A}_{\tau=1,\dots,p}(j, i)$ is different from zero.

2.1.6 The Stationary Bootstrap

Bootstrapping is a powerful statistical method introduced by Efron [60] for estimating the distribution of an estimator or statistic test from resampled independently and identically distributed data (iid). However, the method no longer works when considering more complex dependent data such as time-series data as the iid assumption breaks down. The situation is more complicated when considering the time series because the bootstrap samples should be built in a way that captures the dependencies in the data. The work of Efron [60] has been extended to account for dependencies in the data when performing bootstrapping. Several algorithms have been proposed in the literature. However, in this thesis, we will only consider one of them, which preserves the stationarity of the original time series: the stationary bootstrap [180]. Interested readers may refer to review papers [137, 102] to have a deeper knowledge about existing algorithms for bootstrapping time series. Note that a time series is stationary if it fulfills the following conditions: the mean, variance, and autocorrelation are constant over time. It is an important property to preserve as it is an assumption underlying many statistical procedures used in time series analysis.

The general idea of the stationary bootstrap is that a pseudo time series is generated by resampling with replacement from the original data and blocks of random size. The blocks sizes follow a certain distribution. In the original version of the algorithm the authors chose the geometric distribution. The algorithm assumes that the original time series is stationary and weakly dependent. A time series \vec{x}_t is said to be weakly dependent if we have $corr(x_t, x_{t+h}) = 0$, for $h \rightarrow \infty$. To better

explain the algorithm, let $\vec{x}_{t=1,2,\dots,N} = (x_1, x_2, \dots, x_N)$ the original time series and $B_{il} = \{x_i, x_{i+1}, \dots, x_{i+l-1}\}$ a block of observations starting from x_i . The algorithm samples with replacement a sequence of blocks of random length $B_{i_1 l_1}, B_{i_2 l_2}, \dots$ until the final pseudo time $\vec{x}_t^* = x_1^*, x_2^*, \dots, x_N^*$ has N observations. The first l_1 -observations are determined using the first block $B_{i_1 l_1}$ the next l_2 -observations by $B_{i_2 l_2}$ and so on. Assuming a geometric distribution for iid random variables l_1, l_2, \dots, l_m representing the blocks lengths, we have $\Pr(l_i = m) = (1 - p)^{m-1}p$, for $m = 1, 2, \dots$ and p a fixed number in $[0, 1]$. The sequence i_1, i_2, \dots, i_m is a sequence of iid variables with uniform distribution over $[1, n]$ representing the starting position for a block. The following lines summarize the stationary algorithm:

1. Choose p uniformly from $[0, 1]$.
2. Assign to i a random number from 1 to N and pick the i th element in the original time series and add it to the pseudo time series.
3. Randomly pick a number from a uniform distribution over $[0, 1]$ and assign it to j .
 - (a) if $j > p$, then pick the next element of the original time series as the next one in the pseudo time series. Note that the algorithm wraps around the original time series. Thence, if $i = N$ then we pick the 1st element of the original time series as our next element.
 - (b) if $j \leq p$ then go to step 2.
4. Repeat from step 3 until the pseudo time series has N observations.

2.1.7 Representation of Sites

2.1.7.1 Consensus Sequence

A consensus sequence is a string over the nucleotides alphabet A, C, G, T and an extended alphabet (generally from the IUPAC alphabet [44]), which shows variable degenerate or conserved nucleotides at each position of a motif representing the binding sites of a transcription factor. Note that degenerate base symbols are IUPAC symbols used to represent the DNA position that can have several alternatives. They

are used to report positional variation in situations such as DNA sequencing errors, consensus sequences, or single-nucleotide polymorphisms. Table 22 gives the list of IUPAC degenerate symbols. An example of consensus is the sequence depicted in Figure 6a. It describes the consensus sequence for the TrpR transcription factor.

2.1.7.2 Position Weight Matrix

A position weight matrix (PWM) is a model widely used to depict the DNA binding preferences (motifs) of a transcription factor. The model is a matrix \mathbf{W} . In the matrix, each row corresponds to a letter in an alphabet, e.g., amino acids or nucleic acids, over the sequences, and each column corresponds to a position in the motif. This matrix defines the probability of each letter in the alphabet to occur at a specific position of the motif. The coefficient $\mathbf{W}[i, j]$ gives the score of having i th letter of the alphabet at position j of the motif. This representation of a biological motif was introduced by American geneticist Gary Stormo and colleagues in 1982 [222] as an alternative to consensus sequences (to overcome their limitations). Figure 6b shows an example of the PWM for the TrpR transcription factor that regulates the trp regulon's expression.

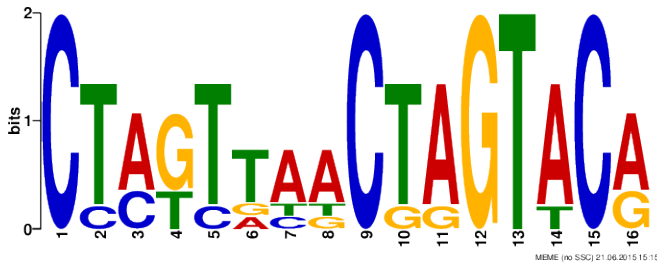
2.1.7.3 Sequence Logo

The sequence logo is a graphical technique for summarizing the alignment of a set of sequences. These sequences can be, for example, protein sequences, RNA sequences, or DNA sequences. The sequence logo is a series of stacks of letters. Each stack shows how well a letter is conserved at a position. This conservation is computed through a score based on Shannon entropy [205]. At each position, individual letters' height is proportional to its frequency at the specific position of the alignment. Sequence logos are used to represent TFs DNA binding. Figure 6c shows an example of a sequence logo representing the binding site of the TrpR TF in *E.coli*. They are mainly used to visualize a large number of sequences that share a common conserved pattern.

		a	c	g	t
>site_0	CTAGTAACTAGTACG	0.000000	1.000000	0.000000	0.000000
>site_1	CTAGTAACTAGTTCG	0.000000	0.166667	0.000000	0.833333
>site_2	CTAGTTCTCTAGTACA	0.666667	0.333333	0.000000	0.000000
>site_3	CTCTTTAGCGAGTACA	0.000000	0.000000	0.666667	0.333333
>site_4	CTCGTGTACTGGTACA	0.000000	0.166667	0.000000	0.833333
>site_5	CCATCAAACCTAGTACA	0.166667	0.000000	0.166667	0.666667
Consensus	CtaGttaaCtaGTaCa	0.000000	1.000000	0.000000	0.000000
IUPAC consensus	CtmKttaaCtaGTaCr	0.000000	0.000000	0.000000	1.000000
		0.833333	0.000000	0.166667	0.000000
		0.000000	0.000000	0.000000	0.166667
		0.000000	1.000000	0.000000	0.000000
		0.666667	0.000000	0.333333	0.000000

(a) Consensus sequences

(b) Position specific probability matrix



(c) Sequence logo

Figure 6: Different representations of binding sites

(a) Alignment of TrpR binding sites in *E. coli* and the derived consensus sequence: the nucleotides consensus sequence and the IUPAC consensus sequence obtained with MEME. In the latter, a letter 'm' means the presence of 'A' or 'C', a letter 'K' means the presence of 'G' or 'T', and finally, a letter 'r' means the presence of 'G' or 'A' at the considered position in the motif. (b) Sequence logo representation obtained with MEME web tool. The relative height of the letters indicates their frequency at each position measured in bits. (c) Position specific probability matrix (PSPM) that is MEME's motif representation. For each position in the motif, it gives the observed frequency ("probability") of each possible letter.

2.2 Feature Selection

Given data with many variables, feature selection is defined as a method that selects the maximal subset of most important features to the output, i.e., the subset of variables that conveys information about the output. The objectives of feature selection are manifolds:

- Reduce the model's complexity and improve its quality to make it easier to interpret by removing redundant and noninformative variables.
- Understand the process underlying the data.
- Reduce overfitting.
- To speed up computation and make a more cost-effective model.

Feature selection methods are split up into four categories depending on how they are combined with the model learning process [194]: filter methods, wrapper methods, embedded methods, or ensemble methods.

2.2.1 Filter Methods

Filter methods consider the intrinsic properties and statistical characteristics of the data to assess their relevance. They are independent of the learning algorithm. In this category, weights are assigned to each variable based on their dependency on the problem/ class label. These weights are generally computed using correlation-based methods or information theory-based methods. Then, generally, the features are ranked regarding the computed weights, and a threshold is applied to get the subset of selected features. Otherwise, a cost function is optimized to find the subset of relevant features. The simplicity of these methods makes them scalable to the data. Filter methods are divided into two categories: univariate and multivariate methods. In univariate methods, the relevance of each variable is evaluated separately according to the selection criterion. There are methods like *t-statistics*, correlation methods, fold change ratio, *B-statistics*. In multivariate methods, the interaction between the features is considered when evaluating the relevance of features. These methods are

among others, Analysis of the Variance (*ANOVA*), mutual information, or Minimum Redundancy Maximum Relevance (*MRMR*).

2.2.1.1 Differential Expression Analysis:

Differentially expressed genes (DEG) analysis consists of comparing the expression profiles of genes among several groups or conditions in designed experiments. This problem is challenging and important in gene expression analysis. It allows filtering informative genes, which is valuable for drug discovery, biomarker identification, or even inference of gene regulatory networks. DEG analysis is performed in two main steps: ranking and selection. In the ranking, a filter-based feature selection method (statistic) is defined to capture the variability of the expression per gene (between the conditions). The statistics are used to compute a score that measures the degree of differential expression. The higher the score, the more the gene is differentially expressed. In selection, a methodology needs to be defined (e.g., setting a threshold) to describe what are “significant” differentially expressed genes. Several feature selection techniques have been proposed for DEG analysis[117], among which we can list:

- **Fold Change:** it is the simplest method for DEG analysis, in which we compute the ratio between the expression mean of the two compared groups. Thus we have

$$FC = \log_2(\mu_T(g)) - \log_2(\mu_C(g)) \quad (14)$$

where $\mu_T(g)$, respectively $\mu_C(g)$, is the average expression of gene g in condition T , respectively in condition C .

- **t-statistic:** which compares the distribution of expression values of genes in two conditions through the means of expression data in the two conditions/groups. It is computed as :

$$\frac{\mu_T(g) - \mu_C(g)t}{\sqrt{\frac{\sigma_T(g)^2}{N_1} + \frac{\sigma_C(g)^2}{N_2}}} \quad (15)$$

where $\sigma_T(g)$ (respectively $\sigma_C(g)$) is the standard deviation of the expression of gene g in condition T (respectively in condition C); N_1 (respectively N_2) is the number of samples in condition T (respectively in condition C).

- The *Empirical Bayes Statistic* [215]: in this method, statistical tests like the above *t-test* is defined within a Bayesian framework, and the empirical Bayes is used to estimate the error in differential expression. This method results in more stable estimations in case of low samples. This method is used in R-package for DEG analysis like `Limma` [188].
- Other statistics tests like the Wilcoxon rank sum test [235], the *F-statistic* have also been used for DEG analysis.

2.2.2 Wrapper Methods

Wrapper methods consider the selection of a subset of features as a search problem. Hence, different subsets of features are built and tested iteratively. Evaluating a specific subset of features is obtained by training a model with only the subset of features and testing its performances. The main advantage of these methods is that they interfere with the model learning. Moreover, they consider interaction with other variables. However, the computational cost severely impedes these methods as the number of features increases, so the search space. Wrapper methods are not widely used for expression analysis as they are prone to overfitting due to the low sample size of expression data.

2.2.3 Embedded Methods

Embedded methods include the selection of the features while the model is learned. This category's advantage is that the selection interacts with the model learning and takes into account interaction with other features. They offer a good compromise between filter and wrapper methods. They are far less computationally expensive than wrapper methods and overcomes the limitation of filter methods. A popular method in this category is the support vector machine method combined with recursive features elimination (SVM-RFE) [90]. As presented in Section 2.1.3, penalized regression can be seen as an embedded feature selection method. In effect, some penalization methods such as *Elastic Net* allow for shrinking coefficients precisely to zero, and in this way, features with zero coefficients are removed from the model.

2.2.4 Other Methods

Recently, researchers have started to combine several types of features selection methods (embedded, filter, and wrapper methods) using hybrid or ensemble methods. Hybrid methods sequentially combine several features selection methods that use different concepts. Ensemble methods are based on the principle that combining several experts' performances is better than the performance of a single expert taken separately. The aim is to combine different feature selection technologies' strengths, as they may perform differently on variable datasets.

Feature selection is a crucial task, especially for high dimensional data, i.e., where the number of variables is very high compared to the number of observations. In fact, in this situation, it is difficult to look at the variables and say which are relevant and which are not. On the other hand, it is difficult to build and interpret a model that will consider all the variables.

2.3 Resources Available for Network Inference

With the advancement of high-throughput experiments, a disparate type of biological data from diverse sources is now available for GRN inference (modeling).

Gene expression data. In gene expression measurements, one determines the level at which a particular gene is expressed within the cell or tissue. It can be done at two main levels:

- **mRNA level:** at this level, gene expression is determined by the amount of mRNA. It is the transcript abundance.
- **Protein level:** the gene expression level, corresponds to the quantity of protein present in the cell. It is protein abundance.

However, the protein abundance measurement is much more challenging to perform than transcript abundance measurements. Thus, gene expression via mRNA abundance is widely used. The transcript abundance is generally measured with high-throughput technology such as microarray experiments.

The microarray experiment is a widely used high-throughput technique for measuring the transcript abundance of a thousand genes simultaneously. Therefore, it enables researchers to establish expression profiles of the genes of a cell. A microarray is a collection of spots attached to a solid surface (a chip). Each spot corresponds to a gene and comprises a million copies of a single-stranded fragment of DNA (gene) called a probe. Each DNA fragment (probe) is designed to uniquely complement an mRNA. The mRNAs are extracted from the genome of interest and then labeled with a fluorescent label and finally spread over the chip. The complementary sequences bind together (hybridize), and the unbound sequences are washed away. The hybridized probes produce a fluorescent signal whose intensity is proportional to the number of copies of probes hybridized on the spot. The gene's expression level can be determined by its corresponding spot's fluorescence intensity, with a bright spot analogous to high expression and a dark spot to low expression. Figure 7 summarizes the microarray experiment for gene expression measurement.

There exist two main types of expression data depending on when the microarray experiment is carried out after the cell has been subject to some perturbations. Hence, we discern:

- The steady-state microarray data that are acquired when the cell reaches the steady-state.
- The time series data measured at different equally spaced time points after perturbations are applied to the cell, and before the cell goes back to its steady-state.

There exist several databases that store gene expression data, but the most important database is **GEO** (Gene Expression Omnibus) [59]. It is publicly accessible and stores different types of gene expression data for many organisms obtained from different sources.

As microarray analyses are costly, the number of samples in datasets is far smaller than the number of genes, causing significant difficulties for GRN inference, as discussed in Section 1.5. This situation has encouraged researchers to adopt some strategies that generate realistic *in silico* expression data from a simulated GRN. The simulated network can either be random graphs models or a part of the right network. Several methods have been proposed to generate simulated expression data,

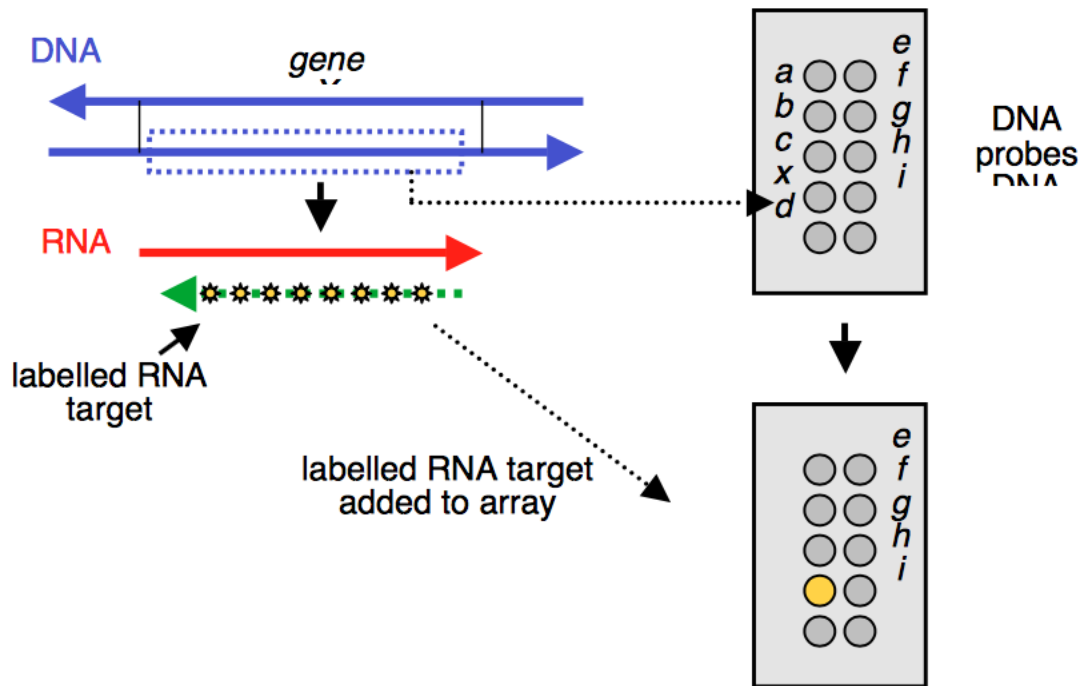


Figure 7: DNA microarray experiment

This figure summarizes the process of measurement of the level of the genes' expression within the cell using a DNA microarray experiment. First, the mRNA is isolated from a sample of interest. The next step consists of labeling the transcripts. To do so, one performs reverse transcription to produce the complementary DNA (cDNA) fragment of the mRNA. The cDNA is then labeled with a fluorescent color. In the next step, the labeled cDNAs are placed onto the microarray, where they will hybridize with their complementary sequences attached to the microarray. The cDNAs that do not hybridize are washed away. In the last step, the fluorescence's intensity (which corresponds to the proportion of cDNAs that hybridized to the probe) of each probe is measured and reported as genes' expression level.

but ordinary differential equations (ODE) are widely used. Many softwares have been developed exploiting the proposed methodology, among which we can mention SynTReN [238] or GeneNetWeaver [202].

Nowadays, with the advent of high throughput technologies that have allowed the sequencing of the genome of many species, new methodologies have emerged that enable deep and rapid investigation of the transcriptome. One of these methods is the RNA-seq (RNA sequencing), which uses sequencing to measure the mRNA level present in biological samples. A typical RNA-seq experiment works as follows: the mRNAs present in the medium are transformed into cDNA (complementary DNA). Then Tags are added to these cDNA fragments to allow later sequencing using short-read sequencing. It results in millions of short sequences (read) that correspond to each cDNA. The reads are then mapped to the original genome. The expression values are the normalized count that have been mapped to genes in the genomes. Figure 8 summarizes an RNA-seq experiment. Note that RNA-seq data offers several advantages like measurement of expression in any species, even in non-model organisms, detecting novel genes.

Protein-DNA interaction sequences data: As presented before, protein-DNA interactions occur when a protein (TF) binds to a DNA sequence (regulatory sequence) located upstream to the gene(s) it controls. Protein-DNA interaction preferences are transcription factors binding sites (TFBSs). They can be determined either through expensive wet-lab experiments or through computational methods. In Section 2.5.1 we will present computational methods for identifying the TFBS. TFBS are generally modeled as matrices. For now, we will focus on experimental techniques. Several lab techniques exist to identify TFBSs. Chromatin immunoprecipitation (ChIP) coupled with either microarray (ChIP-chip) or with sequencing (ChIP-seq) is the most used method. Several databases exist for experimentally reported TFBS sequences and TFBS motifs (see Section 2.5.1 on what has been done to define the TFBS motifs). Amongst them, there is for example **RegTransBase** [130], which is a publicly available database that store TFBS sequences on prokaryotes, and **TRANSFAC** [162] a private store of TFBS sequences and motifs about human, **cis-BP** [245] which stores information about TFBS for several species (≈ 700 species) and gather binding information from several other curated database like **JASPAR** [198].

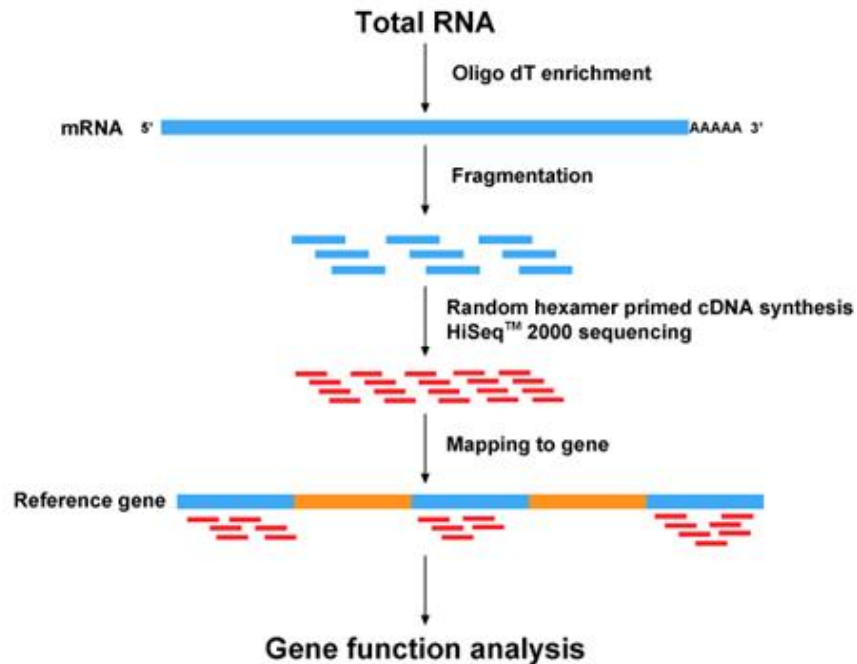


Figure 8: RNA-seq experiment

This figure summarizes the process of measurement of the level of the genes' expression within the cell using an RNA-seq experiment. This image is from <http://bio.lundberg.gu.se/courses/vt13/rnaseq.htm>. A typical RNA-seq experiment works as follows. The mRNAs are collected from the cells. They are then fragmented. Then the mRNAs fragments are converted into double-stranded DNA. Sequencing adaptors are added to the sequences. These adaptors will help the sequencing machine to recognize the fragment. In the next step, the DNA fragments with sequencing adaptors are amplified. Then the library is verified to check, for example, for the size of the fragments. The fragments/read are subsequently sequenced. Finally, the reads are aligned to a genome, and then one counts the number of aligned reads per genes.

These databases generally store information about the TFs associated with each binding information data.

Genomic Data: We choose to divide this type of data into two main categories. In the first category, we have the nucleotide sequences from the DNA of organisms. **GenBank** [15] from NCBI (National Center for Biotechnology Information) is the publicly available reference database for DNA sequences. Other well-curated databases store information about the genome, such as, **UCSC genome browser** [131], which is a web tool for displaying user-defined parts of the genome. It stores information about several organisms like *Human*, *mouse*, *yeast*. It also allows retrieving diverse data related to genes such as their sequences, their promoter regions, their symbols. Another example of such databases is **Ensembl** [113], which is dedicated to vertebrates. Like the **UCSC genome browser**, it allows genome annotation, sequence alignment, regulatory function prediction. In the second category, we have protein sequences. **UniProt** (Universal Protein resource) [237] is a freely available database and a reference for proteins sequences. Protein sequences are out of the scope of this thesis

Gene perturbation data. This data can be obtained from different techniques:

- Through gene knockout (KO), which is a technique in which one or more of an organism's genes are made inoperative or deleted, and genes expression is next measured to capture changes in the system. There are several methods to inactivate a gene, such as mutation. The gene knockout is used to determine gene function, and genes targets if the gene knocked out is a TF.
- Through gene knockdown, which is a technique in which the expression of one or more of an organism's genes is reduced. It is performed through experiments in which RNA interference (RNAi) is used to reduce a gene's expression. Like with gene KO, it allows determining genes function and target of a TF in case it is knocked down.

Organism specific database: Researchers have put the effort in gathering diverse biological information about model organisms such as *Escherichia coli*, and *Human*, *Saccharomyces cerevisiae* into curated databases. This information can be

functional annotations, sequences, regulatory associations, and expression datasets. These databases help scientists in their everyday work. An example of such databases is **SGD** (*Saccharomyces* Genome Database) [99] which is the database for *Saccharomyces*. We also have **MGI** [29], which is the official resource database for the laboratory mouse, providing information such as genomic data, to facilitate the research on human health.

Other type of data: Available data for network modeling are interactome data. An organism's interactome is formed by the full set of the interactions (physical, biochemical, or functional) that can occur among all its macromolecules and metabolites such as proteins, RNA molecules, or even gene sequences. Those interactions include, for example, protein-protein, DNA-protein, RNA-protein interactions. Many databases exist that gather known or predicted interactions. Some of these databases provide information about regulatory proteins and their regulated genes (an example is YeastTract [166]). Others give information about direct or indirect association among proteins (PPI) (an example is the STRING database [118]).

Another type of data relevant to the study genes and their regulatory interactions are gene functional annotations. Many projects have been proposed to manage concepts/classes used to describe gene and gene products' properties. A significant project is the Gene Ontology (GO) [6]. The functional annotations in the GO database (GO terms) are hierarchically organized in a way that groups together subsets of genes sharing common biological functions. This type of information alleviates the functional interpretation of genes participating in a GRN.

This section does not give a complete list of available data but instead introduces the potential usable for the GRN inference.

2.4 Assessment and Validation of Network Inference

Many methods have been proposed to further the task of engaging in analysis of GRN inference. However, the methods need to be reliable to obtain a useful and accurate model of the GRN. Thus, it becomes vital to have a fair assessment and

comparison of existing methods. Several measures have been used throughout the literature to evaluate the accuracy of the GRN inference methods. As in [63], we categorize the most used methods in two main types: statistical-based measures and ontology-based measures. The last category, which is instead a challenge to fairly compare GRN inference methods, is also presented.

2.4.1 Statistical Measures

When we use statistical measures to assess GRN inference algorithms, GRN inference is considered as a binary classification. In essence, the aim is to classify each inferred interaction as either a correct regulatory link or not. The inferred network (the model) is then compared to a gold standard network, and standard evaluation metrics such as ROC curves and Precision-Recall curves are computed. A confusion matrix is first built, as described in Figure 9.

In the context of GRN inference, **TP**, **TN**, **FP**, **FN** are defined in terms of inferred edges. Therefore:

- **TP** are edges occurring in the reconstructed network, and that also occur in the gold standard network.
- **FP** are edges occurring in the inferred network, but that do not appear in the gold standard network.
- **TN** refer to edges that neither belong to the inferred network nor the gold standard network.
- **FN** refer to edges in gold-standard network that are missing in the predicted network.

Here the gold standard network is generally built from wet lab experiments (for some model organisms). The statistical metrics used to assess algorithms for GRN inference are the following:

- The positive predictive value that is obtained with the the following formula:

$$PPV = Precision = P = \frac{TP}{TP + FP}$$

		Actual(Gold standard)	
		Positive	Negative
Predicted	Positive	True Positive(TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 9: Confusion matrix

The figure represents a confusion matrix. In the case of GRN inference, the Actual is the gold standard network, and the predicted is the inferred network. The true positives are the edges that occur both in the reconstructed network and the gold standard network. The false positives are the edges present in the inferred network but absent in the gold standard network. True negative refers to edges absent in the gold standard network and the inferred network. False negatives are edges that are absent in the inferred network but present the gold standard network.

- The negative predictive value that is computed as follow:

$$NPV = \frac{FN}{FN + TN}$$

- The accuracy (ACC) computed with the following formula:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- The sensitivity or recall or true positive rate (TPR) that is obtained as follow :

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

- The specificity (SPC) or true negative rate computed as follow:

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

- The false discovery rate (FDR), which is obtained with:

$$FDR = \frac{FP}{FP + TP}$$

From the above statistical metrics, three pairwise measures are widely used in the literature to assess GRN inference algorithms:

1. The area under the receiver operating curve (**AUROC**). The **ROC** curve represents a plot of the sensitivity (y-axis) against the true positive rate (x-axis) when varying the threshold the algorithm depends on. The AUC is the area between the ROC curve and the x-axis.
2. The AUC of the precision-recall curve (**AUPR**), which plots the precision (y-axis) against the recall (x-axis). The AUC is obtained as previously. Note that, the **AUPR** score is mostly adopted as a metric to evaluate GRN [115] as it is suitable for class imbalance problem: i.e., when the number of positive is much lower than the number of negatives, which is the case for GRN inference [47].
3. The F-measure that is computed with the following formula

$$F_{\beta} = (1 + \beta^2) \frac{PR}{\beta^2(P + R)}$$

A particular case of this measure that is widely used is the F_1 score, obtained when we set $\beta = 1$ so :

$$F_1 = 2 \frac{PR}{P + R}$$

In the process of evaluating the network inference methods, scientists combine the above measures with statistical tests in order to assess the statistical significance of the results obtained when comparing with random networks. The two main statistics used are:

1. the **p-value** that is the probability of occurrence of a given finding by chance alone in comparison with the known distribution of possible findings (the actual finding) considering the number of observations, the kind of data, and the technique of analysis [94]; and

2. The **Z-score** is a number indicating how many standard deviations an element is from the mean.

Note that the above statistical metrics are also used to assess the accuracy of the reconstructed network. For example, in the case where there is no gold standard network against which we can compare the reconstructed network, the statistical test (p-value and Z-score) can be used to assess the statistical significance of the characteristics of the inferred network, like functional annotation.

Evaluation of GRN inference methods using the above metrics is a daunting task owing to the limited availability of the gold standard networks. Only a few organisms are well known and have a set of biologically verified regulatory links, due to our limited current knowledge of the cell. Instead, researchers have put some effort to generate realistic simulated data based on biochemically plausible interaction models. These efforts are discussed in Section 2.4.3.

2.4.2 Ontology Measures

To assess the inference algorithms' performances using ontology-based measures, one uses biological information to quantify the reconstructed network's biological relevance. One uses the idea that in a GRN, genes regulated by the same TF are more likely to be involved in the same biological processes. Thus, one uses the Gene Ontology [6] (GO), to test that it holds in the reconstructed network [249, 63]. This methodology is called functional enrichment. The principle is as follows: given the set of target genes for a particular TF, one maps each gene in the set to its associated biological annotation and then using statistical methods, including Chi-square, Fisher's exact test, Binomial probability, and Hypergeometric distribution, one finds which GO terms are statistically over-represented (or under-represented) in the set, by comparing the distribution of the terms within a target genes set with the background distribution of these terms (e.g., annotation term of all genes in the network). Many softwares exist that automate the process, among which we can list DAVID [108, 107], `g:Profiler` [184], `GO::TermFinder` [25], BiNGO [154]. Interested readers may refer to [107, 229] for details about functional enrichment analysis.

2.4.3 The DREAM Challenge Measures

More recently, the need of a fair comparison of strengths and weaknesses of the network inference methods as well as a clear sense of the reliability of the network models they produce, have to lead to a community effort to catalyze discussion about the design, application, and assessment of systems biology models through annual reverse-engineering challenges: the DREAM challenges. DREAM challenges are a series of projects designed to evaluate model predictions and pathway inference algorithms in systems biology, organized around annual challenges. The challenges data are widely used as gold standard datasets for a fair comparison of many GRN inference algorithms' performances. Each challenge provides the participants with curated datasets, imposes a specific format for the submissions, and defines standard evaluation metrics. For example, in a network inference challenge from expression profiles, the challenge's organizers provide the participants with gold-standard networks, gene expression profiles, and the evaluation metrics. The output is generally an adjacency list L , used to assess the method's performances.

The assessment of the methods works as follow: for each submitted list L , series of subnetworks of increased size $k = 1, 2, \dots, |L|$ corresponding to the top k prediction of L is built by sequentially adding on entries of L at a time. Next, for each subnetwork, a confusion matrix is constructed with regard to the gold standard network. It gives the number of true positives (TP(k)), true negatives (TN(k)), false positives (FP(k)), and finally, the number of false negatives (FN(k)). A true positive is a correct prediction of an edge, while a false positive occurs when the prediction is not actually in the gold standard network. On the other hand, a true negative represents an edge that neither belongs to the prediction nor the gold standard. Finally, a false negative is an edge that belongs to the gold standard, but that is missed by the prediction. Afterward, as previously described, usual metrics are computed, such as precision-recall, AUROC, or AUPR. Note that, generally, a challenge is made up of several networks. Each participant has to infer all the networks to participate in the challenge. The final evaluation of a method is a combination of the performances of the method on each network. Hence, if a challenge is made up of n networks, the usual metrics are computed for the n inferred networks.

Apart from these metrics, p-values are computed for each of the n AUPR and

AUROC scores to evaluate their statistical significance. The p-value describes the probability that a given or larger area under the curve is obtained by a random ordering of the $|L|$ potential network links. The n p-values are combined into two unique p-values (one for AUROC and AUPR). They are computed as the geometric mean of the n individual p-values (c.f. Equation 16).

$$p = (p_1 * p_2 * \dots * p_n)^{1/n} \quad (16)$$

Finally, a global score S_G , that combines *AUPR* and *AUROC* scores is computed as the log-transformed “average” of the two overall AUROC and AUPR p-values, the formula is presented in Equation 17.

$$S_G = -0.5 \log_{10} (P_AUROC * P_AUPR) \quad (17)$$

Larger global score indicates greater statistical significance of the prediction. The scoring metrics really depend on the challenge. Here we describe the metrics of the challenge we consider in Chapter 3.

2.5 Computational Methods

Many efforts have been undertaken to unravel the gene regulatory network. For this purpose, researchers have developed various methods that use different strategies. We can divide the existing methods into three main categories: methods that infer a GRN by identifying the binding sites of the transcription factors on the regulatory regions of genes, methods that infer a GRN from expression data, and finally, methods that use a template to reconstruct a GRN. In the following section, for each category, we will present its general idea and some state of the art algorithms that use the specific strategy.

2.5.1 Methods for Transcription Factor Binding Sites

In general, **cis-regulatory** (or regulatory) elements are regions of non-coding DNA that serve as the DNA-binding sites for transcription factors. The prefix *cis* specifies that the regulatory elements are situated in the vicinity of the gene(s) they control. In GRN inference methods via prediction of cis-regulatory elements, one uses experimentally well-characterized data about regulation (if available) such as transcription

factors (TFs) and TFBSs models (e.g., PWM, motifs) from the genome of interest (target genome) or a model organism like *Escherichia coli*, to infer regulatory links in the target genome. The aim is to identify regions recognized by TFs. Thus, one scans the regulatory regions of genes in the genome of interest with known specific binding sites weight matrices of experimentally well-characterized TFs to determine the genes that have the TFBS in their regulatory regions. These genes are then hypothesized to be regulated by the corresponding TF. In this category, the inferred regulatory links are physical TF-TG binding interactions. Note that several genes hypothesized to be regulated by the same TF are said to be **co-regulated** genes.

In [191], Rodionov grouped the principles behind motif-based GRN inference methods in two main axes, which differ by the availability of experimental data about the regulation of genes. In the first strategy, one has access to known TFs. Hence, the general procedure of this strategy is as follow:

- Step 1: all available information of TFBSs of the well-characterized TFs in model genomes are gathered and constitute the training set for the TFBS profile construction. However, in the case where the TF TFBSs are unknown, one collects the TF known co-regulated genes from the reference genome and their orthologs in the analyzed genome. Then, we build the training set to construct the TFBS's profile with the upstream regions of the known TF-regulated genes in the model genome along with the upstream regions of their orthologous genes in the analyzed genome.
- Step 2: One constructs a TFBS profile with the obtained training set.
- Step 3: The profile is used to scan the whole genome of interest to recruit additional binding sites.
- Step 4: One checks the predicted binding sites' consistency using the principle that co-regulated genes tend to be conserved between genomes that contain orthologous TFs. Thus, one scans the regulatory regions upstream of orthologous genes. If one finds the same TFBSs, then it is considered a true regulatory site; otherwise, if the TFBSs matches are scattered across the genome, then the prediction is false.

The second axis is considered when there is no knowledge about the data of genes regulation. In this case, one can adopt two possible options. In the first option, the assumption is that genes on the same biological pathway may be co-regulated by the same TF. Thus, one gathers genes that belong to the same pathway from closely phylogenetically related organisms to the genome of interest. The regulatory regions of co-regulated genes are then used to build the TFBS model, and then one adopts steps 3 and 4 of the first strategy. Another option is to use phylogenetic footprinting, in which one identifies highly conserved regions of the upstream regions of orthologous genes from a set of closely related species. The TFBS profile is built with a set of conserved regions for orthologous genes. Then one adopts steps 3 and 4 of the first strategy. Note that the TFBS profile can be either position weight matrices (PWM) or consensus sequences. Figure 10 summarizes the two strategies used to reconstruct the GRN via the identification of cis-regulatory elements.

The main difficulty with TFBS data-based methods is that they require high-quality data. The use of divergent organisms may cause the discovery of many false positives.

One of the essential steps of GRN reconstruction via prediction of cis-regulatory elements is identifying TFBSs and the construction of their models. Thus, we choose to present the state of the art algorithms for the construction of TFBS profiles. Generally, the user feeds the algorithms with the set of regulatory regions of genes that are believed to be co-regulated. The algorithm identifies DNA motifs that are overrepresented in the regulatory regions provided. The difficulty is that motifs are short signals in the midst of a vast amount of noise [230]. Another difficulty arises from our poor understanding of the variability in the binding sequences of a given TF.

The existing algorithms differ in their representation of the motifs, their definition of motif “statistical over-representation,” and the method for finding the statistically overrepresented motif. For the motif representation, we observe two main categories of algorithms: PWM based methods and consensus-based methods. Note that the mentioned methods can input other sequences data than DNA sequences (e.g., proteins sequences), but we restrict this section’s scope to DNA sequences. We chose two state-of-the-art algorithms from each category:

- **MEME** [9] and **AlignACE** [112] for the PWM based methods.
- **YMF** [213] and **Weeder** [175] for consensus-based methods.

MEME, which stands for Multiple EM (expectation-maximization) for Motif Elicitation, is a popular tool for motifs discovering in a set of related proteins or DNA sequences. It uses expectation maximization (EM) for motif finding. EM-based motif finding methods work as follows: they alternate between an “Expectation step” and a “Maximization step.” In the “Expectation step”, the scores of all possible motif positions in the input sequences are computed using entries in the PWM. In the “Maximization step,” the high scoring positions are used to refine the PWM. More precisely, **MEME** works as follows: it starts with a random motif. It tries to improve the motif with the EM algorithm until the values in the PWM do not improve, or the algorithm reaches a maximum number of iterations. The EM alternates between the scoring of motif matching positions in the sequences and using the k-mers at the matching positions to refine the PWM (the motif). Note that the algorithm builds the initial PWMs by choosing a single position in all sequences and extracts all k-mers at that position, then it performs one iteration of the EM. It does this for all possible k-mers. Only the best initial motifs are chosen to run EM to convergence. The advantages of **MEME** are the following: it allows multiple motifs to be learned; it does not assume that there is exactly one motif occurrence per sequence, and it is not restricted to short motifs. However, the main limitation is that computation time depends on the length and number of input sequences. Furthermore, it does not return gapped motifs. **MEME** has been recently improved using suffix trees (**STEME** [186]) or with an online version of the EM (**EXTREME** [226]) that allows handling large datasets. Note that the **MEME-suite** is available online and offers a list of different tools for motifs finding.

AlignACE is based on the Gibbs sampling method. More precisely, it uses a Markov Chain Monte Carlo (MCMC) approach to derive the motifs. Markov Chain because the result on the current step depends only on the result at the previous step. Monte Carlo, because the next step is chosen by random sampling. The Gibbs sampling works as follows:

1. Takes as input N sequences.
2. Randomly initializes the motif position in the N sequences, assuming a one motif

occurrence per sequence. The background probabilities are computed from the non-motif position in the N-1 sequences.

3. Compute the probability of all possible motif locations using the previously obtained PWM and the background probabilities.
4. Find new motif starting position in the excluded sequence from step 3.
5. Iterate steps 2-4 until the values in the PWM do not improve, or the algorithm reaches a maximum number of iterations.

AlignACE uses an improved version of the Gibbs sampling method. First of all, it checks both strands of the input sequences. It uses an improved sampling method and allows for discovering multiple motifs. The main advantage of **AlignACE** is that it is not restricted to short motifs. Moreover, it can detect several motifs. Nevertheless, it is susceptible to the initial parameter setting, and like **MEME**, the computation time depends on the number of input sequences.

Weeder is an enumerative-based motif finding method. The general idea of enumerative approaches is to generate all possible words up to a given length. Then determine those occurring with potential substitutions in a significant fraction of the input sequences. The discovered motifs are then ranked using statistical measures. Enumerative approaches perform an exhaustive search of the whole search space and generally find a global optimum. However, they are computationally demanding. **Weeder** uses this principle to find motifs. It uses a suffix tree to optimize the search time. Hence, it preprocesses all the input sequences into a suffix tree. It uses a recursive suffix tree search with pruning to find the pattern that occurs with at most a certain number of substitutions in at least a certain number of the input sequences. The advantage of this method is that, compared to other enumerative methods, the execution time depends on the substitution number rather than the input sequences' length.

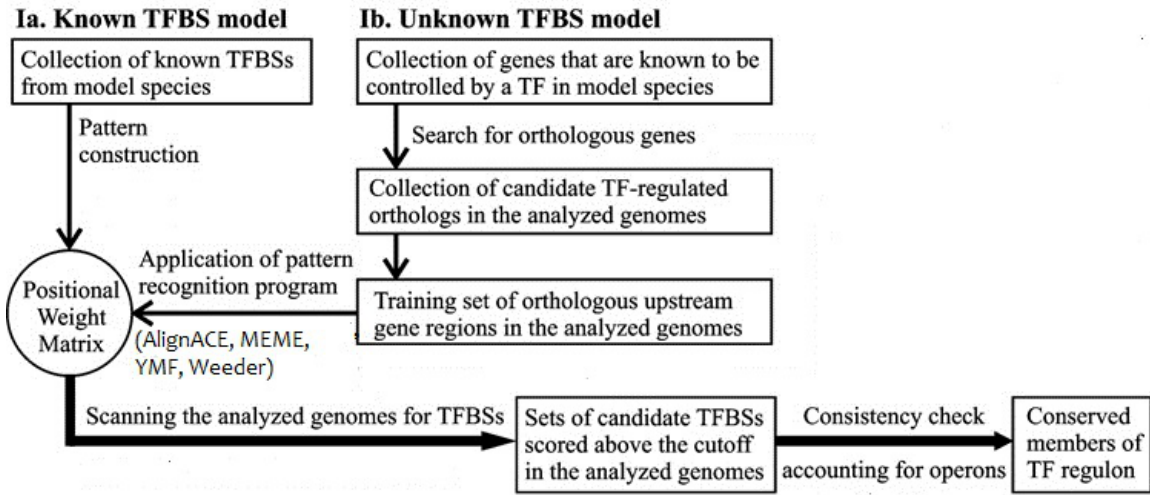
YMF stands for Yeast Motif Finder as the model was derived from the study of known TFBS in *Saccharomyces cerevisiae*. It is based on an enumerative strategy, as described previously. It enumerates all motifs in the search space, and it guarantees to find the motif with the greatest Z-score. The Z-score is the number of standard deviations by which the observed number of occurrences in the input sequences exceeds

the expected number of occurrences if the input sequences were random. YMF detects short motifs with a small number of degenerate symbols. The main advantage of this method is that it returns gaped motifs and ensures that it returns the best motif. However, it is limited to retrieve pretty short and simple motifs that do not vary too much (a small number of degenerate symbols).

The algorithms listed here are the most popular algorithms for motif finding. Of course, there exist other algorithms with different strategies. For example, researchers have proposed combining several motifs finding algorithms (ensemble method) since they generally exhibit complementary outputs [105, 106]. We refer the reader to survey papers on motif discovery methods for a more in-depth comparison of the existing methods [46, 230, 96].

Table 1 presents the selected algorithms in terms of their principle, their output model, their advantages, and their limitations. Figure 6 presents the different representations Tryptophan (Trp) TFBS, which is an *E. coli*'s TF that regulates the *trp* operon presented in Section 1.2. The PWM in the figure has been obtained using MEME.

A. Strategy I: Known TFs



B. Strategy II: Putative TFs

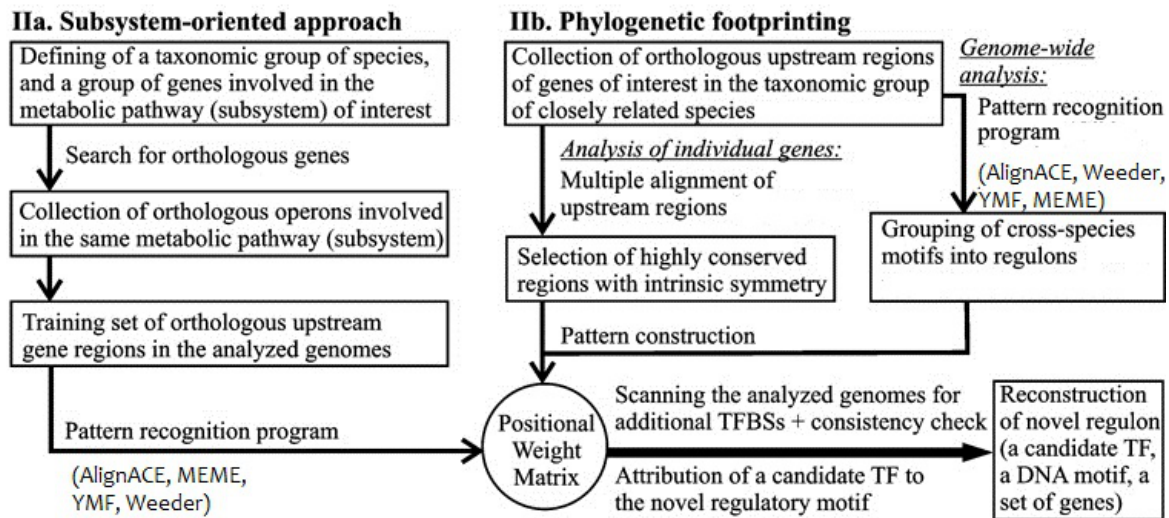


Figure 10: Procedure to identify regulon

The figure summarizes the strategies used to reconstruct the GRN via the identification of regulons [191]. Broadly there are two strategies. In the first strategy, one uses information about experimentally-determined TFs to infer the GRN from position weight matrices that build either with known binding sites of TFs or with genes' promoters. The second axis is considered when there is no knowledge about the data of genes regulation. In this axis, one uses methods such as phylogenetic footprinting to collect promoter of genes that belong to the same pathway from closely phylogenetically related organisms to the genome of interest. A PWM will then be constructed from these promoters.

Table 1: Motifs Finding Methods

Algorithms	Brief description	Output	Advantages	Drawbacks
MEME [9]	Uses expectation and maximisation for motifs finding. It starts with a random motif and tries to improve the motif with the EM algorithm until the values in the PWM do not improve, or the algorithm reaches a maximum number of iterations. The EM alternates between the scoring of motif, matching positions in the sequences, and using the k-mers at the matching positions to refine the PWM (the motif).	Return a set of motifs as position weight matrices.	<ul style="list-style-type: none"> + Can deal with sequences containing reasonable noise. + Can find several distinct motifs in the same set of sequences. + The assumption made by other EM-based algorithms that each sequence contains exactly one occurrence of the shared motif is removed. 	<ul style="list-style-type: none"> - Performance decreases significantly as the length of sequences increases. - Not suitable for whole-genome TFBS motifs discovery.

Table 1 continued from previous page

Algorithms	Brief description	Output	Advantages	Drawbacks
	The 0-order model consists of the frequencies of the letters in the training set.		+ Able to adapt motif length.	- No gaps allowed in the motifs. - Assumes that the positions in the motifs are independent, which is not valid in reality. - Sensitive to initial parameters.

Table 1 continued from previous page

Algorithms	Brief description	Output	Advantages	Drawbacks
AlignACE [112]	Uses Gibbs sampling to find the motifs and the maximum a posteriori (MAP) score to measure the degree to which a motif is overrepresented. The MAP score is combined with another score that measures how well a given motif targets the gene whose upstream regions were used to find the motif. This score allows the selection of functional motifs.	Set of motifs as position weight matrices.	+ Can find long motifs. + Several distinct motifs can be found in the same set of sequences. + Input sequences may not exhibit the motif.	- Performance decreases significantly as the length of sequences increases. - Assumes that the positions in the motifs are independent, which is not valid in reality. - Has difficulty in modeling gapped motifs. - Sensitive to initial parameters.

Table 1 continued from previous page

Algorithms	Brief description	Output	Advantages	Drawbacks
YMF [213]	Uses an exhaustive search approach to discover the over-represented motifs: i.e., motifs with the greatest z-score.	Consensus sequence motifs.	<ul style="list-style-type: none"> + It allows gaps in the motifs. + Allows mismatches within the motifs. + Easy for a human to interpret and visualize the result. + Considers both DNA strands. 	<ul style="list-style-type: none"> - Limited size of motifs. - Time-consuming. - Suitable only when all instances of motifs do not vary too much. - Only suitable for short motifs.
Weeder [175]	The algorithm uses enumeration to find motifs with limited size and a maximum fixed number of mismatches within the input sequences. It uses a suffix tree to optimize the search time. It preprocesses all the input sequences into a suffix tree.	Consensus sequence motifs.	<ul style="list-style-type: none"> + Easy for a human to interpret and visualize the results. + Allows mismatch within the motifs. 	<ul style="list-style-type: none"> - Suitable only when all motifs instances do not vary too much. - Time-consuming. - Only suitable for short motifs.

Table 1 continued from previous page

Algorithms	Brief description	Output	Advantages	Drawbacks
	It uses a recursive suffix tree search with pruning to find the pattern that occurs with at most a certain number of substitutions in at least a certain number of the input sequences.			

The table summarizes some state of the art methods that perform motif finding. We consider consensus-based methods and PWM based methods. We report the most popular algorithm in each category. The 1st column gives the name of the algorithm. The 2nd column gives a short description of the algorithm. The 3rd column gives the type of output the algorithm produces. The 4th column provides the advantages of the algorithm. Finally, the 5th column provides the limitations of the algorithm.

2.5.2 Reverse-Engineering Methods.

In this section, we will present algorithms that reverse-engineer the GRN from gene expression data.

Reverse engineering is the process of unraveling a system’s design by studying its structure, function, and operation. The goal of reverse engineering is typically to understand the target system to the point where it can be rebuilt (copied) or re-engineered (modified) [156].

In reverse engineering of a GRN, the aim is to infer its graph structure (i.e., the interactions between the genes) and parameters (e.g., type/strengths of these interactions) from the expression of all its genes by developing models and algorithms. One scan for patterns underlying the data measurement (time-series or steady-state gene expression data) to learn the interactions and parameters. Expression data is generally represented as a matrix \mathbf{X} (Equation 18), whose rows represent the genes in the GRN, and the columns are either the set of experimental conditions, time points, or tissue samples. More precisely an entry $x_{i,j}$ of \mathbf{X} is a real value representing expression level of the i^{th} gene under j^{th} experimental condition, time point or tissue sample. The regulatory network is represented by a matrix \mathbf{A} , where an entry $a_{i,k}$ is the regulatory interaction between the i^{th} and k^{th} genes. Note that $a_{i,k}$ can either be discrete ($a_{i,k} \in \{0, 1\}$), signed (“+” for activation and “-” for repression) or a real value (to determine strength of interaction).

The reverse engineering methods make the following assumption: if a gene G_1 is linked to another gene G_2 (respectively other genes G_2, G_3, \dots, G_k) then the expression of G_1 influences the expression of G_2 (respectively those of genes G_2, G_3, \dots, G_k). Hence, one has a network of an unknown structure \mathbf{A} and its list of genes. One measures the expression level of the list of genes. We obtain a matrix \mathbf{X} of gene expression profiles of the considered GRN. Finally, one uses the information in the \mathbf{X} to infer connections among genes by quantifying the dependencies among their expression profiles. Figure 11 presents a summary of the steps towards reverse engineering of a gene regulatory network.

$$\mathbf{X}_{N,M} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M,1} & x_{M,2} & \cdots & x_{N,M} \end{pmatrix}, \quad \mathbf{A}_{N,M} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,M} \end{pmatrix} \quad (18)$$

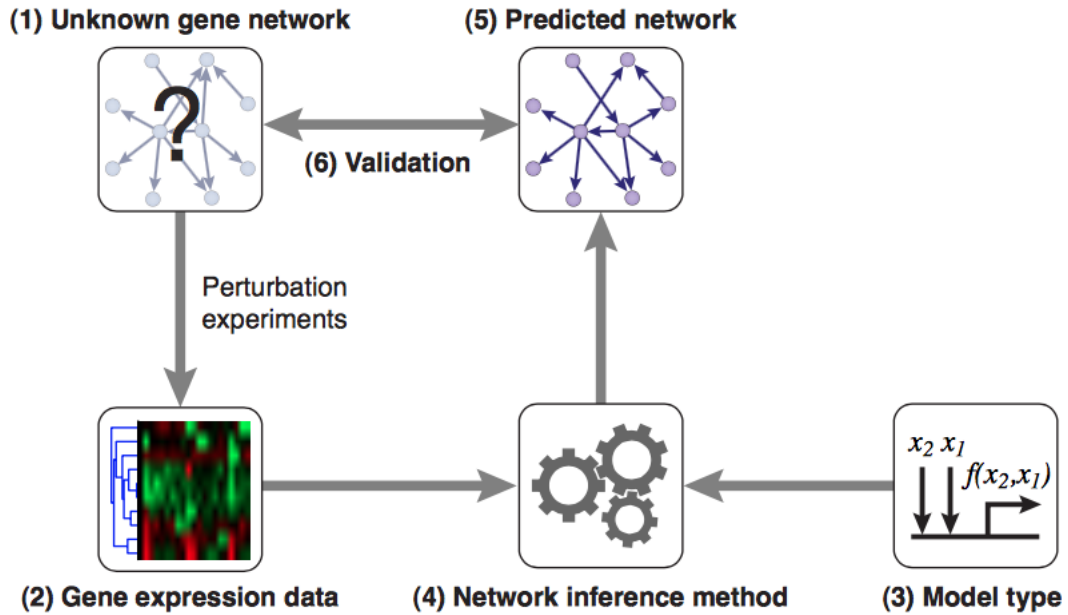


Figure 11: Step for regulatory network inference.

This figure summarizes the steps toward reverse-engineering the gene regulatory network from expression data [156]. (1) Gene network of unknown structure (the so-called target network). (2) Gene expression levels are measured. (3) A modeling framework (model type) for the gene network needs to be defined. (4) The inference method predicts one or several networks that are consistent with the available gene expression data. (5) Depending on the model type, only the structure or a quantitative model of the network can be inferred. (6) The predicted gene network is validated with additional experiments.

There exist several methods that exploit this idea to unravel the GRN. These methods differ in the strategies and the model used to obtain the set of regulatory links. In this thesis, we will emphasize the models used and give details about them, but first, it is essential to talk about the strategies adopted to cope with the problem of GRN inference.

In their paper [49], De Smet and Marchal have proposed to organize GRN reverse engineering methods regarding different strategies they used. First, we distinguish supervised learning from unsupervised network inference. Supervised and semi-supervised methods view the inference problem as a classification problem and use experimentally verified or literature-based interactions to train a machine learning classifier. On the other hand, there are unsupervised methods that neither rely on classification nor assume any a priori knowledge of the network to infer. Furthermore, we have integrative versus non-integrative methods. Non-integrative methods use expression data alone to infer the GRN. They assume that information about regulation is entirely given by the expression activity of the genes.

On the other hand, integrative methods complement the information in expression data with other data such as known TFBS, information on molecular interactions, PPI, and literature. Finally, we distinguish direct methods from module-based methods. Direct methods consider each gene individually and infer all its interactions with other genes. In contrast, modules based methods take advantage of the modular nature of the GRN, and instead of working at the level of the genes, consider the network as modular. A module here is a set of genes regulated in concert by the same regulator(s) under a shared regulatory program, which specifies the behavior of the genes in a module as a function of the module regulators expression. Module-based methods consider the GRN as a set of nested modules obtained with any clustering methods then the regulatory program has to be learned for the modules.

Aside from the strategies used to overcome the problems that arise from GRN, many models have been used for GRN inference. In this thesis, our categorization is based on the different models proposed over the literature for GRN inference. Therefore, we summarize existing efforts into the following five categories. Namely: (i) Probabilistic graphical model-based methods (ii) Correlation-based methods (iii) Partial correlation-based methods, (iv) Information theory-based methods, (v) Regression-based method and finally (vi) ODE based methods.

2.5.2.1 Probabilistic Graph Methods

In modeling the GRN, the aim is to capture both the entities involved in (genes) and their different attributes (e.g., expression data). Probabilistic graphical models

treat different attributes as random variables [73]. The defined model represents the description of the joint probability distribution of all random variables, which is a product of terms involving only a few expressions. A graph is thus used to specify the structure of the product. It shows dependencies between variables and provides tools to reason about the properties entailed by the product. The aim of modeling here is to find the model that closely represents the distribution of the data. It can be done in two ways. Either by parameter estimation through the maximum likelihood problem or by selecting among different model structures, the one that best represents the data, using a scoring measure. The common model of this category is the Bayesian Network, which is among the first models used to infer the GRN from expression data, with the work of Friedman *et. al* [74].

To represent this category, we choose **Banjo** [260]. **Banjo** uses a dynamic Bayesian network to infer the GRN from time-series expression data. The expression data are first discretized using either quantile or interval discretization. The algorithm uses the 1st order Markov DBN, which assumes that gene expression at time t is only dependent on the expression data of its parent genes and the gene itself at time $t-1$. Then, the Bayesian Dirichlet equivalence (BDe) scoring metric is applied to evaluate the goodness of each possible network G in the search space. In the next step, the algorithm searches the top N networks with the highest score using either a greedy strategy or simulated annealing. The top N networks are then averaged to obtain a consensus network. The algorithm outputs a weighted signed directed network. The advantage of **Banjo** is that it can infer the directionality of the data. Furthermore, the algorithm is specially designed to work with data with a limited amount of samples. However, in the initial version, the algorithm had difficulty inferring combinatorial links (targeted by many TFs) that are common in GRN.

Other algorithms such as **scanBMA** [258], **G1DBN** [141] have also been proposed.

scanBMA is an unsupervised algorithm that uses a Bayesian network and incorporates prior knowledge data to improve the accuracy of the inferred GRN from time-series gene expression data. It poses the GRN inference problem as a series of feature selection problems for each TG. In each problem, a list of TFs is inferred for a specific TG. It uses BMA (Bayesian Model Averaging) to account for uncertainty in the model selection, by averaging different models to derive the posterior density

on model parameters. It uses a greedy method to explore the search space and eliminates improbable models using the Ocam widow principle [152]. The prior knowledge data is used to compute prior probabilities of regulatory interactions. These probabilities are used to compute posterior probabilities of regulatory interactions. They defined Zellner’s g-prior [265] on the prior distribution of the model parameters and used EM to find g. Furthermore, the method uses a faster implementation of BMA, which allows an efficient search of the model space. The faster implementation of BMA permits `scanBMA` to have a running time comparable to that of `LASSO`. `scanBMA` runs in a couple of minutes for a network of thousands of genes on a regular laptop. The method has been tested on simulated data from DREAM4 challenge (with networks of size 10 and size 100) and experimental data from *Saccharomyces cerevisiae* [256] that consists of 3556 genes. The authors compared their performance to a dynamic Bayesian network, `LASSO` and mutual information-based methods. They used AUPR and AUROC scores to evaluate their performance and the performance of the competing methods. For the DREAM4 data, the authors considered only time-series expression data and did not include any prior knowledge data. On the simulated data, `scanBMA` performed comparably to the competing methods. However, it outperformed the competing methods on the yeast dataset.

`G1DBN` uses the dynamic Bayesian network as defined in Section 2.1.1. From Section 2.1.1, when dealing with time-series expression data, the DAG may be huge and impossible to infer with input data where the data number of variables is larger than the number of samples. In `G1DBN`, the authors have proposed approximating the DAG to infer using the q^{th} order conditional dependency DAG. More precisely, the authors use the 1^{st} order conditional dependence to approximate the DAG of the dynamic Bayesian network. Under some conditions demonstrated by the authors, the 1^{st} order conditional dependence graph contains the full DAG to be recovered. The algorithm proceeds into two steps. In the first step, the algorithm learns the DAG of 1^{st} order conditional dependence assuming linear dependencies. In the second step, the DBN’s real DAG structure is inferred from the coefficients learned in the previous step. The authors benchmarked their method on both simulated and real expression data. As simulated data, they generated 100 random time-series expression data using a multivariate autoregressive model of order 1. They used two experimental datasets: one on the yeast cell cycle [217] with 786 genes expressed in the cell cycle

and the other on *Arabidopsis Thaliana* [214] with 800 genes. The method has been compared to LASSO and the autoregressive model. They used precision-recall curves to report the performance. G1DBN presents superior results on both simulated and experimental data. Furthermore, it was able to infer biologically validated regulatory links as well as new potential regulatory links. The authors have shown that the performance of G1DBN may depend on the size of the network since they observed the degradation of their performance on real network data as the size increases. The main advantage of this method is the fact that it can infer the direction of regulatory links. However, it assumes linear dependency among genes, and it is computationally demanding.

2.5.2.2 Correlation Methods

Correlation-based methods are the most straightforward way to investigate a GRN using gene expression data since regulatory links among genes imply a correlation between their expression profiles. Thus, in this category, a matrix of gene expression similarity, $\mathbf{S} = [s_{ij}]$ is defined using the matrix \mathbf{X} (see Equation 18); where s_{ij} is the pairwise correlation coefficient between expression profile $\mathbf{X}_{i\cdot}$ of gene i and $\mathbf{X}_{j\cdot}$ of gene j . The coefficients are computed using a correlation measure, such as the Pearson correlation coefficient. From the matrix \mathbf{S} , regulation links are inferred using a threshold τ : a regulation link is established between genes i and j if and only if $s_{ij} \geq \tau$. The threshold τ is generally obtained from randomization of the data allowing statistical significance assessment. The inferred GRN is an undirected graph since we cannot infer causality from correlation. To represent this category, we consider the WGCNA [138] algorithm, which is available as an R package. The WGCNA proposes several correlation measures, which can be used to construct the correlation similarity matrix \mathbf{S} from expression data matrix \mathbf{X} . The correlation similarity matrix is then used to compute the adjacency matrix by thresholding the entries of \mathbf{S} . The package offers a function in which the scale-free topology of the inferred network criterion [266] is used to choose the threshold τ . The algorithm can provide either a weighted or an unweighted network as output, in which weights represent the confidence of the regulation links. The inferred graph is undirected.

2.5.2.3 Partial Correlation Methods

Partial correlated based methods (Gaussian graphic model) are also known as the covariance selection problem. In this approach, the observed data matrix \mathbf{X} (see Equation 18) is assumed to be drawn from a multivariate normal distribution $\mathcal{N}(\vec{\mu}, \Sigma)$, with $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ the mean vector, and Σ the covariance matrix. A partial correlation matrix $\mathbf{C} = [c_{ij}]$ is computed from the inverse of the covariance matrix. It is used to describe dependency between any pair of genes conditioned on the rest of the genes. The normality assumption allows determining conditional independence between two genes from the zero entries of the inverse of the covariance matrix and the contrary from non-zero entries. The general step of the algorithms in this category are :

1. Estimate the covariance matrix from the data \mathbf{X} .
2. Invert the covariance matrix and compute the partial correlation matrix $\mathbf{C} = [c_{ij}]$.
3. Use a statistical test to determine entries in the partial correlation matrix that significantly differ from zero.
4. Infer regulatory links from non-zero entries in the partial correlation matrix.

As a representative of the category we choose **GeneNet** [201]. The major problem faced in this category is that the number of gene expression samples is much smaller than the number of genes. It makes the covariance matrix impossible to invert since, in those conditions, the obtained input data matrix \mathbf{X} loses the characteristics of an invertible matrix. The **GeneNet** algorithm uses the Moore-Penrose pseudoinverse to compute the inverse of the covariance matrix and uses bagging (bootstrap aggregation) to stabilize the estimator. The Moore-Penrose pseudo inverse is a generalization of the matrix inverse that is based on singular value decomposition. Moreover, they computed the p-value as well as the posterior probability for each edge. They used FDR to correct for multiple testing and select the edges to be included in the GGM based on adjusted p-values. **GeneNet** was initially tested on both simulated and experimental data from Human breast cancer [246] the dataset covers 7129 genes. The authors use statistical measures such as FPR, specificity, or FDR to report their

performances. This method’s main advantage is that it is particularly designed for high dimensional data (low samples compared to the number of variables). However, the method cannot infer combinatorial links, i.e., regulatory links targeted by many TFs.

2.5.2.4 Information Theoretical Methods

Information theory-based methods use mutual information to infer correlation coefficients among expression profiles of pairs of genes. Thus, for each pair of genes, mutual information is computed then compared to a threshold τ . If the mutual information of the pair of genes is greater or equal to the threshold τ , then a regulation link is inferred between the pair of genes. Mutual information is an interesting measure of correlation since a mutual information value of 0 between two variables implies that the two variables are independent. Furthermore, information theory allows identifying any correlation that can exist between the two variables: either linear or nonlinear. As a representative of this category, we choose ARACNE [159]. It infers the GRN from steady-state expression data. It computes the mutual information between all possible pairs of genes and uses randomization of the data (e.g., bootstrap) to select the threshold τ . In a second step, the algorithm considers all gene triplets (pairwise mutual information of the genes in the triplet) and uses DPI (data processing inequality) to reduce the number of false positives regulatory links. The algorithm may accidentally consider a direct interaction between two genes when, in reality, there is a 3rd gene involved. DPI will remove such a direct link. DPI states that if two genes g_1 and g_3 interact only through a 3rd gene g_2 (i.e. there exists no alternative path between g_1 and g_3 than through g_2) then $I(g_1, g_3) \leq \min(I(g_1, g_2); I(g_2, g_3))$. $I(\cdot)$ is the mutual information. The algorithm has been tested on both simulated and experimental data. For simulated data, the authors used different topologies proposed in [164] to simulate their expression data. The networks were either random or scale-free and consisted of 100 nodes with 200 interactions. For the experimental data, they used a dataset from *Human* B cells. They evaluated their performances against the Bayesian network and Relevance Network using precision-recall curves. ARACNE showed superior performance compared to concurrent methods. Furthermore, ARACNE was able to infer validated targets of some known *Human* oncogenes. The algorithm also inferred

other biologically validated regulatory links. The main limitation of **ARACNE** is that it cannot infer combinatorial links (i.e., links targeted by several TFs). Moreover, it cannot infer the edge directionality.

Recently, an extension of **ARACNE** to time series data has been introduced: **Time delay-ARACNE** [272] (**TD-ARACNE**). **TD-ARACNE** proceeds in three steps. In the first step, the algorithm detects for each gene, the time point where its expression will initially change. It will help in computing the mutual information in the next step. Secondly, the network is obtained upon the mutual information computed from each pair TF-TG and for different time shifts regarding the information obtained from the first step. In the last step, the algorithm uses the same strategy as **ARACNE** to prune false positives in the network. **TD-ARACNE** has been evaluated on both simulated and experimental expression data. For the simulated dataset, the authors have tested different data sizes: different number of genes (10 and 20 genes) and a different number of time points. The point was to evaluate how **TD-ARACNE** performance depends on the input data size. They worked on three datasets from *Saccharomyces cerevisiae* [217] with 11 genes, from *Escherichia coli* [192] with 8 genes and from the IRMA dataset [33] made up of 5 genes. The IRMA dataset is obtained by extracting a subnetwork of 5 genes from the *Saccharomyces cerevisiae* GRN. The dataset contains both time-series and steady-state gene expression. It includes two sub-datasets: one switch-on data and one switch off data. The switch-on data covers five experiments. The switch-off data covers four experiments. The whole dataset contains 142 measured samples. The performance was compared to dynamic Bayesian network methods, ODE based method, and the original **ARACNE** using measures such as the PPV, the recall, or the F-score. **TD-ARACNE** outperformed the concurrent methods. They have demonstrated that **Time delay-ARACNE** was able to recover the true structure of the GRN more reliably compared to the concurrent methods. Furthermore, **TD-ARACNE** was also able to infer several known interactions. The main advantage of this method is that it can infer the direction of the edges.

To represent mutual information-based methods in Table 2, we will only consider *ARACNE* as it has been proven to be state of the art on many data sets.

2.5.2.5 Regression Methods

In this category, algorithms use the genome-wide expression profiles of genes to infer the network of regulation. Here, the expression profile of a target gene is modeled as a linear/nonlinear combination of its regulator's expression levels. Hence, the network inference amounts to finding for each gene, the small subset of transcription factors whose expression profile is sufficient to predict its expression. The problem is thus recast as a series of variable selection problem. In each problem, a regression model is used to rank the variables. However, the high dimensionality low samples problem of expression data seriously impedes regression techniques. This situation has caused researchers to employ different strategies to overcome this difficulty. Hence, some authors have used regression trees for each target gene, using a compact set of regulators at each node [116, 168, 207]. Others, have adopted a concept which consist in penalizing the regression model using either LASSO [97, 262] or Elastic net [210, 147].

Two state of art methods of this category are GENIE3[116] and TIGRESS [97].

GENIE3 uses a set of random trees to model the dependencies between the expression levels of the TFs and their TGs. The algorithm decomposes the inference of network of p genes into p different regression problems, in which the steady-state expression pattern of each gene of interest (TG) is predicted from those of other genes (TFs) using an ensemble of random trees (Random forest or Extra Tree). The importance of a potential TF in the inference of the TG gene expression serves as an indicator of putative regulatory links. The weight of a regulator is the sum of the mean decrease in the impurity of all the tree nodes where it is used to split. Note that the mean decrease in impurity computes, at each test node in the tree, the reduction of the variance of the output due to the split. The algorithm aggregates the putative regulatory links over all genes (the p subproblems) to provide a final ranking of interactions from which the whole network is reconstructed. The method has been tested on both simulated multifactorial data from the DREAM4 and experimental data from *Escherichia coli*. The simulated data set is made up of 100 genes. On the other hand, *Escherichia coli* data is made up of 4297 genes. They used the DREAM4 scoring methodology described in Section 2.4.3 to evaluate their performance on the simulated data. For real data, the performance was

reported in terms of precision-recall curves. The method was compared to **GENIE3** combined with different tree-based methods (random forest, ensemble tree), mutual information-based methods, and the Gaussian graphical model. It was competitive with concurrent methods on the *E.coli* dataset assuming that information about potential TFs is provided to the algorithm. **GENIE3** was the best performer of the size 100 multifactorial DREAM4 subchallenge and the best performer of DREAM5. This method’s main advantage is that it does not make any assumption about the nature of gene regulation; it can deal with combinatorial and nonlinear interaction; it is fast and scalable.

Several extension of **GENIE3** has been introduced in the literature. One of them, **iRafNet** [178], integrates heterogeneous prior knowledge data such as knockout genes expression, TFBS, or protein-protein interaction to improve the accuracy of the reconstructed network. The prior knowledge is used to construct weights to sample potential regulators during the tree construction. The method has been rigorously tested on simulated data from the DREAM4 and the DREAM5 challenges. They used knockout and time-series gene expression data as prior knowledge. They used two measures to evaluate their performances: the AUPR and the AUROC. The method has demonstrated superior performance compared to original random forest based GRN inference, **GENIE3**. Furthermore, the authors have demonstrated that **iRafNet** performance on simulated data is comparable to the ensemble learning method, i.e., a network obtained by combining results from different models. The authors have further evaluated their method on *in vivo* data from the *Saccharomyces cerevisiae* cell cycle. The method has demonstrated that it provides functional insights to the inferred regulatory links. This method’s main advantage is that it includes different types of available biological data for the regulatory network inference. Furthermore, as **GENIE3**, it is fast and scalable.

More recently, the authors of **GENIE3** have extended their work and introduced the dynamical version of their algorithm: **dynGENIE3** [81]. It extends **GENIE3** to handle both steady-state and time-series expression data. Initially, **GENIE3** was designed to work only on steady-state expression data. **dynGENIE3** assumes that the transcription rate of a gene is a function (potentially nonlinear) of the expression of other genes and, potentially, itself, plus a parameter specifying its decay rate. The algorithm

combines time-series and steady-state gene expression data to learn the ordinary differential equation defining each gene’s transcription rate. As in **GENIE3**, the method uses the mean decrease in impurity to compute the importance of each regulator. Note that the decay value is computed either with the data assuming an exponential decay or obtained from the literature. The method was tested against different inference technologies: dynamic Bayesian network, ordinary differential based methods, Granger causality based methods, and nonlinear dynamical model. The performance was tested on simulated data from the DREAM4 challenge (using their evaluation methodology), and on three real-world datasets: a *Saccharomyces cerevisiae* [172], *Drosophila melanogaster* [103] and *Escherichia coli* [120]. **dynGENIE3** consistently outperforms **GENIE3** on simulated dataset. However, the same result is not observed on experimental data as the datasets and organisms exhibit many differences. These results show that **dynGENIE3** performance is very data-dependent. Apart from the scalability and speed, this method’s main advantage is that it integrates time-series data, which allows modeling the network dynamics. Moreover, the authors have extended the method to allow the user to specify the list of potential TFs, which is not available in the original work. However, the main drawback is that the method does not consider the myriads of other data that exist, such as TFBS to supplement expression data.

TIGRESS combines stability selection with *LASSO* regression (implemented with the **LARS** [61]) to infer the GRN from expression data. Stability selection consists of running a feature selection method several times on perturbed data and computing the score of a feature as the number of times it was selected. As with **GENIE3**, the problem for p genes network is made up of p regression subproblems fitted on the bootstrapped randomized expression level of the TFs of the network. A modified measure of selection frequency for each potential TF is used as evidence of possible regulatory links. In summary, the weight of each potential TF is based on the frequency with which the TF is selected by the **LARS** in the top features and the area under each curve up to a fixed number of **LARS** steps. The method was mainly compared to mutual information-based methods. The method was benchmarked against the DREAM4 and DREAM5 challenge datasets for simulated data. They used the DREAM5 scoring methodology, which is the same as the DREAM4 methodology. Furthermore, the method was evaluated on experimental data from *Saccharomyces*

cerevisiae [66]. When tuned optimally, TIGRESS shows similar performances to GENIE3 on simulated data but not as good on experimental data. This method’s main limitations are the linearity assumption and the fact that it considers only expression data as input.

To represent regression-based methods in Table 2, we will only consider GENIE3 in as it has been proven to be state of the art on many data sets.

2.5.2.6 Differential Equation Methods

Differential equations allow modelling the change in expression level of a gene as a function of the change in other genes expression plus some external factors. The function is time dependent. Hence, it is adequate for capturing the dynamic of a system. More precisely we have:

$$\frac{d\vec{x}}{dt} = f(\vec{x}, p, u) \quad (19)$$

Where $\vec{x} = (x_1, x_2, \dots, x_N)$ is the expression level of genes g_1, g_2, \dots, g_N ; N is the total number of genes in the network; p is the model parameter set and u is the external perturbation factor. Inferring the GRN amounts to identify the function f and the model parameter set p , using the measured signals \vec{x} and u . There exist many solutions to Equation 19 when the problem is unconstrained. However, a solution exists when an assumption is made upon the nature of $f()$. Many GRN inference algorithms assume that $f()$ is linear. However, this assumption may be too simplistic to model the complex nature of regulatory interactions. Other functions exist, such as piecewise linear, continuously linear, or nonlinear, each of them models different levels of complexity of the model. The most accurate being the nonlinear function. However, estimating the parameters of a nonlinear with low sample data may prevent getting reliable results. A popular method in this category is the **Inferelator** [23]. It uses regression and variable selection to infer the set of transcriptional influences on each gene of a GRN based on the integration of genome association and gene expression data. The algorithm uses ODE to define the expression level of a gene or the mean expression of a set of functionally related genes as a function of the TFs transcriptional level plus some external stimuli. The point is then to select, for each gene or set of genes, the subset of factors that influence its expression level. They assume that $f()$ in Equation 19 is truncated linearly. $f()$ is then fitted with LASSO

to strictly enforce parsimony. The model allows fitting time-series and steady-state gene expression data simultaneously. They also extended the model to account for pairwise interaction between the predictors (TFs and external stimuli). The method was tested on an experimental dataset from *Halobacterium*, which is made up of 2404 genes. The **Inferelator** was able to infer new interactions that were experimentally tested and verified. Moreover, the algorithm was able to predict *Halobacterium* global expression after perturbing the inferred network.

Having an overview of each category and the algorithms we choose to represent them, Table 2 presents a comparison of the selected algorithms in terms of their input data type, their complexity when available, their advantages and limitations. Note that this is not an exhaustive list of the methods that exist in the literature that use expression data as the main input to infer the GRN. We refer the reader to reviews paper [98, 158, 157, 165]

Table 2: Reverse-Engineering Methods

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
GeneNet [201]	Uses Moore-Penrose pseudoinverse to compute the inverse of the covariance matrix from which the partial correlation matrix is computed. Edges are added between pairs of genes if their common entry in the partial correlation matrix is non-zero.	Time series and steady-state expression data	<ul style="list-style-type: none"> + Can deal with high dimensional data + Few number of parameters are computed + Can infer the putative direction of regulatory links + Works well to construct a GRN at large scale 	$O(m^3 + nm^2)$ [73]	<ul style="list-style-type: none"> - Can only detect pairwise regulation links. - Assumes linear relation.

Table 2 continued from previous page

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
Aracne [159]	Works in two steps. In the first step, the algorithm computes the mutual information of all the pairs of genes in the network. Then only statistically significant pairs are considered as being regulation links in the output network. In a second step, the algorithm considers all gene triplets and uses DPI to reduce the number of false positives regulatory links.	Steady-state gene expression data.	+ Works well with high-dimensional data. + Can infer a network of any dimension size (scalable).	$O(m^3 + n^2m^2)$ [73]	- Inability to infer direction of regulations links. - Cannot infer combinatorial links (links targeted by many TFs).

Table 2 continued from previous page

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
Banjo [260]	<p>Uses dynamic Bayesian network to infer the GRN from time-series expression data. The expression data are first discretized. Then the algorithm evaluates all possible networks with a Bayesian-based score. In the next step, the algorithm searches the top N networks with the highest scores using either a greedy strategy or simulated annealing. Finally, output a consensus of the top N networks.</p>	Time-series genes expression data	<ul style="list-style-type: none"> + Deals with uncertainty due to the use of probability. + Can infer the type (inhibition or activation) of regulation links. + Infers direction of the regulation links between genes. 	–	<ul style="list-style-type: none"> - Requires many samples for the estimation of the density distribution. - Loss of information due to gene expression discretizing - The quality of the result depends on the gene expression discretizing.

Table 2 continued from previous page

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
WGCNA [138]	Regulatory links between genes are inferred using correlation measures on which a threshold is applied.	Steady-state gene expression data	+ Works well to reconstruct large GRN. + Can construct weighted networks where each weight shows the significance of the regulation links.	[7]	- Inability to infer the direction of regulation links. - Assumes linearity
GENIE3 [116]	Uses a set of randomized trees to infer the GRN from expression data. For a p genes network, the algorithm decomposes the network prediction into p different regression problems.	Steady-state expression data	+ No assumption on the type of regulatory interaction; thus, it can handle either linear or combinatorial interactions + Simple to interpret	$O(TKmn \log n)$ T: number of trees K: number of selected variables at each node of the trees	- Consider only one type of data (static data).

Table 2 continued from previous page

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
	In each sub-problem, a set of randomized trees (random forest or extra-trees) is used to predict the expression pattern of one gene based on the expression profiles of all the other genes. Input genes importance in the prediction of the target gene expression pattern indicates putative regulatory links.		+ Able to predict edges direction + Fast and scalable		

Table 2 continued from previous page

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
Inferelator [23]	Uses regression and variable selection to infer the set of transcriptional influences on each gene of a GRN based on the integration of genome association and gene expression data. The algorithm uses ODE to define the expression level of a gene or the mean expression of a set of functionally related genes as a function of the TFs transcriptional level plus some external stimuli.	Time-series and steady-state gene expression data	+ Consider both steady-state and time-series expression data + Allows incorporation of other regulatory information + Infer edge direction.	–	

Table 2 continued from previous page

Algorithms	Brief description	Input data type	Advantages	Complexity	Limitations
	The algorithm assumes that $f()$ Equation 19 is a truncated linear function. $f()$ is then fitted with LASSO to strictly enforces for sparsity				

The table summarizes some state of the art methods that reverse engineer the GRN from gene expression data. We consider the probabilistic graphical-based methods, correlation-based methods, partial correlated based methods, information theory-based methods, regression-based methods, and ODE based methods. We report one algorithm per category. The 1st column gives the name of the algorithm. The 2nd column gives a short description of the algorithm. The 3rd column gives the type of input the algorithm is expecting. The 4th column provides the advantages of the algorithm. The 5th column gives the complexity of the algorithm. In this column, we used the following notation: m = number of genes in the dataset; n = number of samples in the dataset (typically, $n \gg m$). Finally, the 6th column provides the limitations of the algorithm.

2.5.3 Template Methods

Template-based methods exploit the idea that orthologous TFs regulate orthologous genes. Thus, in this category, one starts with the well reconstructed GRN of a well-known organism (the template) and then transfers information about regulation to orthologous genes in the genome of interest. This methodology requires the entire template genome and its GRN *i.e* the set of its TF-gene interactions. The genome can either be represented by its nucleotides sequence (DNA sequence) or its proteins sequences. These sequences are then used to determine their representatives (orthologs) in the genome of interest. Orthologs are detected using sequences alignment tools. To present this category, we consider the works of Babu *et.al* [8], in which they used one the most well-characterized bacterial network, *E. coli*, as a template to reconstruct networks of 175 prokaryotic genomes. The orthology is detected using a hybrid method combining sequence alignment and the Bidirectional Best Hit method(BBH). BBH consists of finding the pairs of genes in two different genomes that are more similar to each other than either is to any other gene in the other genome. Research has recently demonstrated that detecting homology with DNA is a challenging task [176] as they are rapidly evolving. Hence, it will be almost impossible to identify homology sequences after many years of divergence. Nowadays, homology is detected using protein sequences.

Even though methods in his category are relatively simple, they present some drawbacks. In effect, they necessitate a template that should be complete in order for the reconstructed network to be as well. Nevertheless, most existing template GRNs are far from complete, and the number of template genomes that exist is very small. Moreover, the template should be close enough in the phylogenetic tree in order for the conservation to be significant.

2.6 Conclusion

In this chapter, we mentioned the mathematical background notions necessary to comprehend the thesis. Furthermore, we summarized the state-of-the-art regulatory network modeling in the following three categories:

1. Model-based methods: These methods use the principle of evolutionary conservation and exploit the idea that orthologous transcription factors regulate orthologous target genes. Hence, in this category, one uses a model organism (i.e., an organism for which the GRN is well known); information about regulation among orthologous genes is transferred from the model network to the network of interest.
2. Reverse engineering methods using gene expression data: These approaches use the fact that a target gene's expression profile is influenced by its direct regulators' expression profile. Hence, one chooses an appropriate type of model architecture that is a mathematical function that describes the general behavior of a TG depending on the activity (expression profile) of its TFs; then, the model parameters are learned from data. Several different model architectures for reverse engineering GRNs from gene expression data have been proposed ranging from the Boolean network, Bayesian Network, information theory model to regression models.
3. Network inference by prediction of cis-regulatory elements: These approaches make use of experimentally well-characterized transcription factor binding sites (TFBSs) for inferring regulatory links. Hence the promoter regions of all the genes in the genome are scanned with the known TFBSs. The genes are hypothesized to be regulated by the TF if they possess the TFBS in their regulatory region.

We also pinpoint the advantages of the proposed solutions as well as their drawbacks.

Chapter 3

BENIN

GRN inference is a challenging problem due to the task’s combinatorial nature and the limitation of available data. With technological advances, we are now witnessing the accumulation of a large variety of data that carry on an incomplete but complementary picture of the regulatory process. Hence, taken together, they form a complete picture of the regulatory circuit. This complementarity created a need for the development of GRN inference methods that integrate this diversity to circumvent the use of each data separately. Sophisticated methods integrating diverse biological knowledge with expression data have thus been proposed. This integration is generally done in the form of prior knowledge, i.e., a subjective belief of how the network should resemble. The majority of these methods uses a Bayesian Network (BN) framework for combining prior knowledge and data as it reflects both causal and probabilistic semantic. However, due to the complexity of learning BN, these methods can only be applied to small networks (with a minimal number of nodes). In this work, we aim to contribute to data integration discourse by proposing an elegant and easy method to incorporate several biological knowledge to guide the inference of GRN of any size.

This chapter will present BENIN, a new GRN inference algorithm that incorporates biological knowledge with time-series expression data. The objective is to infer a directed graph $G = (V, E)$ representing the GRN from gene expression data guided by prior knowledge of possible edges. In this graph, the nodes set V represents the network genes and, the edges set E , the regulatory links between the *TFs* (the sources) and the *TGs* (the sinks). We formulate the challenge as a features selection problem.

Details about the formulation of the problem will be given subsequently. The chapter is organized as follows: Section 3.1 details the methodology of **BENIN**; Section 3.2 presents the software used to implement **BENIN** as well as the data employed to evaluate its performances; Section 3.4 shows the performances of **BENIN** on the DREAM 4 challenge data.

3.1 The **BENIN** Algorithm

This section presents our method **BENIN**. In what follows, we will use the following notation: \vec{x} for vector, boldface upper case letter \mathbf{X} for matrix representation, uppercase calligraphic font \mathcal{S} for sets, and $\mathbb{1}$ to represent the unit function. TF will designate the transcription factor, TG the target gene, and finally, GRN will correspond to the gene regulatory network, and KO stands for Knockout.

3.1.1 Overview

BENIN is a regression-based method that uses feature selection combined with stability selection to reverse engineer the GRN from expression data. **BENIN** uses a simple but efficient method to integrate any prior knowledge data with time-series expression data to boost the GRN inference. Moreover, **BENIN** integrates regulatory interactions from other model organisms into the studied model through orthology sequence transfer (c.f. Chapter 4).

In this part of the thesis, we will summarize **BENIN** functioning on a simple example from size 10 DREAM4 subchallenge. We will reverse engineer network 1 using knockout gene expression data as prior knowledge combined with time-series gene expression data.

BENIN takes as input the prior knowledge which can either be a matrix \mathbf{A} of association strengths or probabilities of interaction between each TF and the TGs; the matrix of time series expression data \mathbf{X} , the set of regulators \mathcal{R} , a power γ , the number of bootstrap R and an optional threshold τ . The following major steps summarize **BENIN**. In this example we set τ to 0.5 and R to 1000.

Step 1: If the prior knowledge is not in the form of association strengths or probabilities, it is first transformed into probabilities or association strengths $\{\mathbf{A}_{r_j \rightarrow g_i}\}$

for $i = 1, \dots, M$; and $j = 1, \dots, M'$ of the likelihood of the regulatory interactions between each TF and the TGs. In our example, $M = 10$ and $M' = 8$. In our example the prior knowledge is not in the required form.

Step 2: The association strengths or probabilities are then transformed into weights: $\{w_{r_j \rightarrow g_i}\}$ with $i = 1, \dots, M$; $j = 1, \dots, M'$. These weights are utilized into **BENIN** to build the model.

Step 3: For each TG $g_i, i = 1, \dots, M$, we model its expression profile as a linear combination of the expression profile of its direct TFs, using **Elastic net**. The weights $w_{r_j \rightarrow g_i}$ are fed into **Elastic net** to guide the selection of more plausible TFs. At this step, we generate R bootstraps and compute a score $s_{r_j \rightarrow g_i} \in \mathbb{R}$ for each edge $(g_i, r_j) \in \mathcal{E}$, which provides the strength of the potential interaction. The scores $s_{r_j \rightarrow g_i}$ are such that true interactions get the highest scores. The whole process is summarized in Figure 24.

Step 4 All the scores $\{s_{r_j \rightarrow g_i}\}_{i=1, \dots, M; j=1, \dots, M'}$ are put together and sorted in decreasing order. A threshold τ can then be applied to this sorted list to obtain the final network.

Step 4 This step is not part of **BENIN**, but the final network is evaluated against the true structure using different statistical measures such as the area under the precision-recall curve or area under the ROC curve.

Here we give **BENIN** general overview; we will provide details about each step in subsequent sections.

3.1.2 Problem Specification

The GRN is a collection of molecules such as genes, non-coding RNAs, proteins, and metabolites that interact together to control genes' expression to ensure proper cell functioning. The gene's expression involves many steps, and regulation may occur at each of these steps. We restrict the scope of our research to the transcriptional level, where most of the genes are regulated [20]. In what follows, the GRN will refer to the transcriptional regulatory network (TRN) and represents the graph of direct interactions between the set of transcription factors (TFs) \mathcal{R} and their target

genes (TGs). Figure 12d shows an example of such a graph from the DREAM4 challenge. For illustrative purposes, we changed the original naming of some genes to discriminate against the set of TFs from all the other genes.

We focus on inferring a weighted directed graph of the GRN using time series gene expression data coupled with prior evidence of interactions. In this graph, the edges represent the set of direct regulatory interactions between the set of transcription factors (TFs) \mathcal{R} and their target genes (TGs). We assume that the sources and sinks of each edge should be different: i.e genes do not directly regulate themselves. In what follows, let \mathcal{R} the set of TFs and \mathcal{G} the set of all genes in the network, we have $\mathcal{R} \subseteq \mathcal{G}$. A time series gene expression data matrix $\mathbf{X}_{\mathcal{G},t}^{TS}$ over a set of genes $\mathcal{G} = \{g_1, g_2, \dots, g_M\}$ is defined as follow:

$$\mathbf{X}_{\mathcal{G},t}^{TS} = [\vec{x}_{g_1,t}, \vec{x}_{g_2,t}, \dots, \vec{x}_{g_M,t}] \in \mathbb{R}^{N \times M},$$

where the $\vec{x}_{g_i,t}$ are column vectors of expression values of the i -th gene g_i measured at N discrete time points (cf Figure 12b). The matrix $\mathbf{P}_{\mathcal{G},\mathcal{R}}$ of the p-values of binding interactions among the set transcription factors \mathcal{R} and the set of genes \mathcal{G} , is defined as follow:

$$\mathbf{P}_{\mathcal{G},\mathcal{R}} = [\vec{p}_{\mathcal{G},r_1}, \vec{p}_{\mathcal{G},r_2}, \dots, \vec{p}_{\mathcal{G},r_{M'}}] \in \mathbb{R}^{M \times M'},$$

where $M' = |\mathcal{R}|$, and $\vec{p}_{\mathcal{G},r_i}$ is a vector representing r_i binding location profile regarding all the TGs in the network. Figure 12a shows an example of a genome-wide location data matrix. And finally, the matrix of knockout gene expression data $\mathbf{X}_{\mathcal{R},\mathcal{G}}^{KO}$ is defined as follow:

$$\mathbf{X}_{\mathcal{R},\mathcal{G}}^{KO} = \begin{bmatrix} \vec{x}_{\Delta r_1, \mathcal{G}}^{KO} \\ \vdots \\ \vec{x}_{\Delta r_{M'}, \mathcal{G}}^{KO} \end{bmatrix},$$

where $\vec{x}_{g_j \Delta r_i}^{KO}$ is the vector of expression values of all the genes in the strain where r_i has been knocked out.

Our aim here, is to uncover the set of weighted direct links:

$$\mathcal{E} = \{(g_i, r_j), g_i \in \mathcal{G}, r_j \in \mathcal{R}\}$$

	G1	G3	G4	G6	G7
G1	0.000	0.530	0.861	0.084	0.864
G3	0.059	0.000	0.061	0.875	0.048
G4	0.007	0.022	0.000	0.339	0.007
G6	0.482	0.477	0.071	0.000	0.961
G7	0.600	0.145	0.099	0.347	0.000
G8	0.494	0.732	0.316	0.334	0.435
G9	0.186	0.693	0.519	0.476	0.713
G10	0.827	0.478	0.662	0.892	0.400
G2	0.038	0.789	0.438	0.027	0.390
G5	0.007	0.023	0.245	0.839	0.777

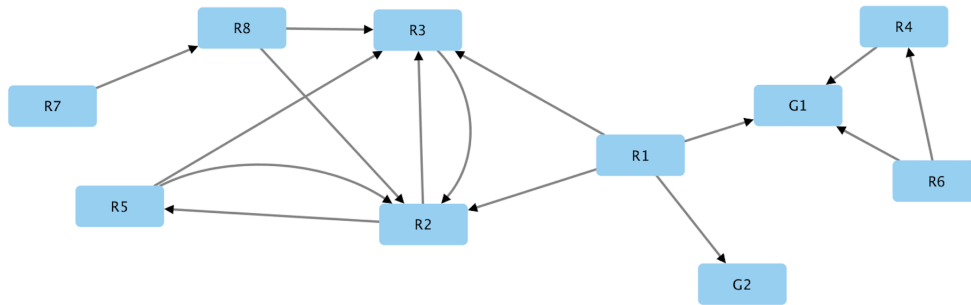
(a) Simulated location data

Time	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
0	0.67	0.13	0.36	0.78	0.1	0.28	0.61	0.74	0.67	0.7
50	0.33	0.12	0.35	0.72	0.19	0.31	0.61	0.76	0.56	0.73
100	0.18	0.04	0.57	0.59	0.23	0.36	0.47	0.67	0.72	0.67
150	0.18	0.06	0.49	0.63	0.41	0.34	0.42	0.73	0.64	0.7
200	0.09	0.14	0.34	0.54	0.56	0.3	0.42	0.68	0.55	0.75
250	0.07	0.09	0.31	0.53	0.64	0.37	0.58	0.8	0.6	0.61
300	0.15	0.1	0.41	0.46	0.54	0.3	0.6	0.62	0.73	0.67
350	0.09	0.09	0.38	0.42	0.56	0.34	0.68	0.52	0.77	0.74
400	0.14	0.12	0.52	0.5	0.67	0.23	0.51	0.61	0.68	0.84
450	0.17	0.08	0.38	0.48	0.67	0.38	0.5	0.6	0.75	0.72
500	0.12	0.08	0.43	0.54	0.74	0.33	0.51	0.72	0.67	0.63

(b) Simulated expression data

G1	G2	G3	G4	G5	G6	G7
0.00	0.37	0.26	0.36	0.80	0.38	0.83
0.73	0.00	0.38	0.60	0.15	0.34	0.52
0.71	0.14	0.00	0.37	0.08	0.40	0.66
0.85	0.11	0.62	0.00	0.15	0.28	0.11
0.88	0.14	0.43	0.56	0.00	0.33	0.54
0.62	0.40	0.46	0.63	0.09	0.00	0.46
0.69	0.09	0.32	0.27	0.08	0.30	0.00
0.80	0.28	0.38	0.66	0.12	0.63	0.55
0.79	0.12	0.24	0.40	0.14	0.30	0.69
0.79	0.14	0.42	0.36	0.12	0.34	0.62

(c) Simulated knockout expression data



(d) An example of size 10 network from the DREAM4 challenge

Figure 12: Example DREAM4 Input for BENIN

The figures shows example of input for BENIN from DREAM4 size 10 subchallenge. (a) sub-matrix of simulated genome wide location for the network 1 from DREAM4 size 10 sub-challenge. (b) sub-matrix of time-series expression data for the network 1 from DREAM4 size 10 sub-challenge. (c)Sub-matrix of knockout gene expression data matrix for size 10 network 1 from DREAM4 challenge. (d) The network 1 from the DREAM4 size 10 sub-challenge.

3.1.3 Network inference as Feature Selection

The basic idea of our method is to decompose the inference of the GRN into as many sub-problems as the number of genes in the network. Hence, for a network of M genes, we decompose the problem into M sub-problems, in which one considers each gene at a time, and the aim then amounts to finding the set of its direct regulators. We assume that the expression profile $\bar{x}_{g_i}^{TS}$ of a gene g_i is a linear function of the expression values $\mathbf{X}_{\mathcal{R}_{g_i}}^{TS}$ of its direct regulators, plus some noise. For each $g_i \in \mathcal{G}$ we can then write its expression profile $\bar{x}_{g_i}^{TS}$ as in Equation 20.

$$\bar{x}_{g_i}^{TS} = f(\mathbf{X}_{\mathcal{R}_{g_i}}^{TS}) + \epsilon \quad (20)$$

The problem is to find, for each gene g_i , the subset of its direct regulators $\mathcal{R}_{g_i} \subseteq \mathcal{R}$ whose expression is predictive of its expression profile. This is the well-known problem of feature selection in machine learning [208].

To model the dynamics from time series expression data, we consider the vector autoregressive model (VAR) [212, 221]. The p -lag vector autoregressive model (VAR(p)) captures linear dependencies between variables in a time series. Particularly, each variable is expressed as a linear combination of the p lags of its own values as well as the p lags of the other variables in the model and, finally, an error term. More formally, let $\bar{x}_t^{TS} = (\bar{x}_{g_1,t}^{TS}, \bar{x}_{g_2,t}^{TS}, \dots, \bar{x}_{g_M,t}^{TS})$ be an M -dimensional multiple time series expression data vector; \bar{x}_t is assumed to be generated from a VAR(p) if it can be written as in Equation 21.

$$\bar{x}_t^{TS} = \vec{c} + \mathbf{B}_1 \bar{x}_{t-1}^{TS} + \dots + \mathbf{B}_p \bar{x}_{t-p}^{TS} + \vec{\xi}_t, \quad (21)$$

where p denotes the lag length or the order of the VAR model; \mathbf{B}_i is an $M \times M$ matrix of coefficients for the i -th lag, M represents the number of genes (variables) in the time series; $\vec{\xi}_t$ is an M -dimensional noise vector. We restrain the scope of this work to the first order of the model, i.e. $p = 1$.

From Equation 21, setting $p = 1$, the expression profile of each gene at time t can be written as follow:

$$\bar{x}_{g_i,t}^{TS} = c_i + \mathbf{X}_{\mathcal{R}',t-1}^{TS} \vec{\beta}_{i,\cdot} + \xi_t \quad (22)$$

Note that $\vec{\beta}_{i,\cdot}$ is the transpose of a row-vector of \mathbf{B} and $\mathcal{R}' = \mathcal{R} \cup g_i$. To find the subset of regulators for each gene, the problem amounts to retrieve the vector $\vec{\beta}_{i,\cdot}$.

which can be obtained by any regression method.

3.1.4 Feature Selection using Elastic Net

One of the major problems of time series expression data is that they are measured over a short period; which results in datasets where the number of genes is far greater than the number of time points (high-dimensionality problem) [255, 239]. Furthermore, many of these variables may be irrelevant to the output and a large number of them highly correlated (multicollinearity problem). To deal with those problems, Zhou and Hastie [274] have proposed a regularization method: the **Elastic net**. It combines two well-known regularizations techniques: the **LASSO** [228] and the **Ridge** [101]. **LASSO** uses L1-norm, it tends to produce a sparse model but is limited by the number of samples in the learning dataset. **Ridge** uses the L2-norm and is good at retrieving correlated variables, but does not produce sparse models. By combining both regularization methods, **Elastic net** integrates the advantages of both techniques while overcoming the drawbacks of each regularization taken separately.

$$\vec{\beta}_i^{Enet} = \underset{\vec{\beta}_i}{\operatorname{argmin}} \|\vec{x}_{g_i,t}^{TS} - \mathbf{X}_{\mathcal{D}',t-1}^{TS} \vec{\beta}_i\| + \lambda_{Enet} \left[(1 - \alpha) \|\vec{\beta}_i\|_2^2 + \alpha \|\vec{\beta}_i\|_1 \right] \quad (23)$$

3.1.5 Bootstrapping the Elastic Net to Score Regulatory Links

One approach to compute the scores of the edges could be to use the absolute values of the regression coefficients stored in the vector $\vec{\beta}_{\dots}^{Enet}$. However, this can be problematic since our data are high-dimensional. In effect, performing feature selection with this type of data may produce unstable results [180]. To remedy this problem, we propose combining bootstrap with the **Elastic net**. The general idea is to generate several bootstraps of the original time series data. Our resampling algorithm is based on stationary bootstrap [180], which resamples time series by consecutive blocks of varying length, ensuring that dependencies between the variables are preserved. Afterward, the **Elastic net** is applied to the bootstraps. The non-zero components of $\vec{\beta}_{\dots}^{Enet}$ are used to select the potential regulators in each bootstrap. Then, the final score of each link corresponds to the frequency with which the regulator of the interaction is chosen by the **Elastic net** within each of the R bootstrap samples, as reported in Equation 24.

Different sub-problems yield different possible links in the final network. Those links are then combined into a single list and ranked according to their scores $\{s_{r_j \rightarrow g_i}\}$ for $j = 1, \dots, M'$ and, $i = 1, \dots, M$. Finally, a user-defined threshold τ can be applied to this list to get the final list of regulatory interactions of the reconstructed network.

$$s_{r_j \rightarrow g_i} = \frac{1}{R} \sum_{k=1}^R \mathbb{1}_{\vec{\beta}_{i,j}^{Enet,k} \neq 0}, \text{ where } \mathbb{1}_{\vec{\beta}_{i,j}^{Enet,k} \neq 0} = \begin{cases} 1, & \text{if } \vec{\beta}_{i,j}^{Enet,k} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

3.1.6 Integrating Prior Knowledge

The limited availability of expression data and the quantity of noise they contain, have made the inference of the GRN from expression data alone, a challenging problem. One way to overcome the difficulties and improve the reconstructed network is to supplement the expression data with other data types to take advantage of the wealth of complementary information about the regulation they offer. This information can be used to design informative priors, to boost the network inference. In this work, we use TF binding location data and knockout gene expression data. We consider an extended version of `Elastic net` (the adaptive `Elastic net` [275]), which modifies the regularization term in Equation 23 by using different degrees of shrinkage on the regression coefficients $\vec{\beta}_{i,j}$ depending on which predictors we want to keep in the model. The new regression problem is defined in Equation 25. Note that the vector $\vec{\nu}$ modifies both the l1 and l2 norm as in the implementation of `glmnet` R package [72]. For the `Elastic net`, finding the subset of regulators for gene g_i is equivalent to solving Equation 25 for the variables $\vec{x}_{g_i,t}^{TS}$ and $\mathbf{X}_{\mathcal{R}',t-1}^{TS} = [\vec{x}_{r_1,t-1}^{TS}, \vec{x}_{r_2,t-1}^{TS}, \dots, \vec{x}_{r_{M'},t-1}^{TS}]$.

$$\vec{\beta}_i^{Enet} = \underset{\vec{\beta}_i}{\operatorname{argmin}} \|\vec{x}_{g_i,t}^{TS} - \mathbf{X}_{\mathcal{R}',t-1}^{TS} \vec{\beta}_i\| + \vec{\nu} \lambda_{Enet} \left[(1 - \alpha) \|\vec{\beta}_i\|_2^2 + \alpha \|\vec{\beta}_i\|_1 \right] \quad (25)$$

3.1.6.1 Prior knowledge from Location Data

Genome-wide location data provide evidence of physical interaction between TFs and TGs within the genome, through the identification of the region in the upstream region

of the genes where the TF will bind: the TFBS. This evidence is generally reported as p-values, which suggests the statistical significance of the binding event. The smaller the p-value, the more significant is the existence of the physical interaction between the TF and the considered TG. Integrating gene expression data with location data allows extracting reliable and useful information about regulation, as they provide complementary information about regulation. However, genome-wide location data are very noisy [17, 206]. To tackle the noise inherent in location data, we integrate such data through a probabilistic framework, as suggested in [17]. The aim is to match the p-values to the corresponding probabilities of edges being present in the final GRN.

Let $P_{r_j \rightarrow g_i}$ be a random variable over $[0, 1]$ which represents the p-value of the location data of the regulatory link $E_{r_j \rightarrow g_i}$ in the graph G of the GRN. In a previous study [206], $P_{r_j \rightarrow g_i}$, has been assumed to be exponentially distributed if $E_{r_j \rightarrow g_i} \in G$, and uniformly distributed if $E_{r_j \rightarrow g_i} \notin G$. More formally, we have:

$$\Pr(P_{r_j \rightarrow g_i} = p | E_{r_j \rightarrow g_i} \in G) = \lambda e^{-\lambda p} / (1 - e^{-\lambda}), \quad (26)$$

where λ is the parameter that controls the scale of the truncated exponential distribution. And:

$$\Pr(P_{r_j \rightarrow g_i} = p | E_{r_j \rightarrow g_i} \notin G) = 1 \quad (27)$$

We now define the probability of having the edge $E_{r_j \rightarrow g_i}$ in G , knowing the p-value of the binding event. Let $\Pr(E_{r_j \rightarrow g_i} \in G) = \beta$ be the probability that an edge $E_{r_j \rightarrow g_i}$ is in the graph without any prior knowledge. Using the Bayes formula we have:

$$\Pr(E_{r_j \rightarrow g_i} \in G | P_{r_j \rightarrow g_i} = p) = \frac{\lambda e^{-p\lambda} \beta}{\lambda e^{-p\lambda} \beta + (1 - e^{-\lambda}) (1 - \beta)} \quad (28)$$

In [17], using Equation 28, the authors have demonstrated that λ acts as a tunable parameter indicating the degree of confidence in the evidence provided by the location data. Therefore, λ models the belief level of noise inherent in location data, and at the same time, it weights the evidence we are giving to it. A suitable weighting of the prior could be to choose the appropriate value of λ ; instead, as proposed in [17], we adopt a more robust method and marginalize Equation 28 over λ . We assume λ is uniformly distributed over the interval $[\lambda_{min}, \lambda_{max}]$ and we integrate Equation 28 over that interval. The new equation to compute the conditional probability on an edge $E_{r_j \rightarrow g_i}$ is given in Equation 29.

$$\Pr(E_{r_j \rightarrow g_i} \in G | P_{r_j \rightarrow g_i} = p) = \frac{1}{\lambda_{max} - \lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} \frac{\lambda e^{-p\lambda} \beta}{\lambda e^{-p\lambda} \beta + (1-e^{-\lambda})(1-\beta)} d\lambda \quad (29)$$

Equation 29 can be easily computed numerically for fixed values of $P_{r_j \rightarrow g_i}$. Using Equation 29, we precompute the probabilities associated with each p-value and store them in a matrix \mathbf{A} for later use. \mathbf{A} is then transformed into weight. The intuition is that the weights are defined so that small probabilities are associated with high weights and vice versa. We thus compute the weight matrix \mathbf{W} as the inverse component-wise of the elements of the matrix \mathbf{A} raise to the power γ . More formally we have:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{A}_{r_j \rightarrow g_i})^\gamma} \quad (30)$$

3.1.6.2 Prior knowledge from Knockout Expression Data

Knockout (KO) expression data are expression data measured in an organism where one of its genes is made inoperative (“knocked out” of the organism). We consider KO data measured at a steady state. KO data represents valuable prior information to boost network inference. KO data informs about possible direct interaction between a TF and a TG. We compute the z-scores of each association $r_j \rightarrow g_i$. The z-score assumes that knocking out a TF directly affects the expression of its direct target genes more strongly than the other genes [182]. We calculate the z-score of a regulatory link $r_j \rightarrow g_i$ as in Equation 31 and store it into a matrix \mathbf{Z} .

$$z_{r_j \rightarrow g_i} = \frac{\bar{x}_{\Delta r_j, g_i}^{KO} - \mu_{g_i}}{\sigma_{g_i}} \quad (31)$$

where $\bar{x}_{\Delta r_j, g_i}^{KO}$ is the expression value of the gene g_i in the strain where r_j has been knocked out, μ_{g_i} is the mean expression value of the gene g_i in all the strains (wild type and deleted strains) and σ_{g_i} is its standard deviation in all the strains.

We then transform these z-scores into weights to feed `elastic net`. Note that the higher the absolute value of the z-score, the more affected is the expression value of the target gene by the TF knocked out. Since we aim to penalize the TF with low *a priori* binding potential, the intuition is that the weights are defined so that small absolute z-scores are associated with high weights and vice versa. Thus, we compute

the weight matrix \mathbf{W} as the component-wise inverse of the elements of the matrix \mathbf{Z} raise to the power $\gamma \geq 0$. More formally, we have:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\text{abs}(\mathbf{Z}_{r_j \rightarrow g_i}))^\gamma} \quad (32)$$

The function $\text{abs}()$ computes the absolute value.

Algorithm 1 summarizes BENIN.

Algorithm 1 The BENIN algorithm

Input: list of genes \mathcal{G} , list of TFs \mathcal{R} , time series expression matrix $\mathbf{X}_{\mathcal{G},t}^{TS}$, the associations strengths matrix \mathbf{A} , power γ , threshold τ

- 1: Transform the probabilities $\{\mathbf{A}_{r_j \rightarrow g_i}\}_{i=1, \dots, M; j=1, \dots, M'}$ into weights $\{\mathbf{W}_{r_j \rightarrow g_i}\}_{i=1, \dots, M; j=1, \dots, M'}$ as:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{A}_{r_j \rightarrow g_i})^\gamma}$$

- 2: **for** each gene g_i , $i = 1, \dots, M$ **do**

- 3: Generate the learning sample:

$$LS := (\bar{x}_{g_i,t}^{TS}, \mathbf{X}_{\mathcal{R},t-1}^{TS}), \text{ for } t = 0, \dots, T$$

- 4: Generate R samples of LS with stationary bootstrap.
5: Compute R elastic net vectors $\vec{\beta}^{Enet,k}$, $k = 1, \dots, R$
6: Compute the scores $\{s_{r_j \rightarrow g_i}\}_{i=1, \dots, M; j=1, \dots, M'; i \neq j}$ as

$$s_{r_j \rightarrow g_i} = \frac{1}{R} \sum_{k=1}^R \mathbb{1}_{\vec{\beta}_{i,j}^{Enet,k} \neq 0}$$

where,

$$\mathbb{1}_{\vec{\beta}_{i,j}^{Enet,k} \neq 0} = \begin{cases} 1 & \text{if } \vec{\beta}_{i,j}^{Enet,k} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- 7: **end for**

- 8: Aggregate the $s_{r_j \rightarrow g_i}$, $i = 1, 2, \dots, |\mathcal{G}|$, $j = 1, 2, \dots, |\mathcal{R}|$, $i \neq j$ and rank them in decreasing order.

- 9: Apply the threshold τ to select links in the inferred network.

Output: Ordered links $r_j \rightarrow g_i$ with scores $s_{r_j \rightarrow g_i}$.

3.2 Experimental Validation

3.2.1 Data

3.2.1.1 The DREAM4 Challenge Dataset

The DREAM4 dataset is a widely used benchmark dataset to evaluate network inference methods. We have worked with two datasets: the DREAM4 *in silico* size 100 and size 10 sub-challenges. Each sub-challenge provide time series expression data as well as other types of data such as perturbation data (knockdown or knockout data) for five networks. The networks differ in their structure which mimics either *E. coli* or *Saccharomyces cerevisiae* regulatory network. Table 3 summarizes the characteristics of the five networks in terms of the numbers of TFs and the number of regulatory links for both sub-challenges. The topologies were obtained by extracting subnet-

	size 10		size 100	
Network	# TF	# Regulatory links	# TF	# Regulatory links
Net 1	8	15	41	176
Net 2	9	16	36	249
Net 3	9	15	44	195
Net 4	9	13	41	211
Net 5	9	12	34	193

Table 3: Description of DREAM4 size 10 and size 100 networks

The table presents the number of regulators and regulatory links for each of the five networks in the 10-nodes and 100-nodes in DREAM4 sub-challenge. The character “#” stands for “number of”. Columns 2-3 provides the numbers of TFs and regulatory links for the size 10 networks. Columns 4-5 provides the numbers of TFs and regulatory links for the size 100 networks

works of either *E. coli* or *Saccharomyces cerevisiae* regulatory network, notably part of the network with cycles. However, self-interactions are omitted. Their dynamics were obtained by using a kinetic model of gene regulation. The expression data were generated using `GeneNetWeaver` version 2.0. Time series for size 100 sub-challenge

(respectively size 10 sub-challenge) consist of 10 (respectively 5) different experiments with 21 time points each. Knockout data include wild type expression data as well as steady state expression data obtained after knocking out each of the M genes in the network. We considered only the single knockout expression data.

3.2.1.2 Simulated Location Data

Genome-wide location data provide direct evidence of physical interactions amongst genes within the genome. Different databases exist that gather information about location data, for example, the *Young Lab*, which gathers different works on genome-wide location data for organisms such as *Saccharomyces Cerevisiae*. Simulated genome-wide location data are obtained by generating p-values for the TFs of the networks in both sub-challenges. We use a uniform distribution $\mathcal{U}[0, 1]$ for the edges that do not belong to the gold-standard network. In counterpart, we use exponential distribution over the interval $[0, 1]$ with scale λ , for edges from each TF that are present in the gold-standard network [206]. The scale λ controls the level of noise in the generated generated dataset.

- For each pair (r_j, g_i) of regulator and target gene:

$$p\text{-value} = \begin{cases} \text{random number in the interval } [0, 1] \text{ using exponential distribution with parameter } \lambda, & \text{if } (r_j, g_i) \in G. \\ \text{random number over the interval } [0, 1], & \text{otherwise.} \end{cases} \quad (33)$$

We generated eleven location datasets for each of the ten networks in both the sub-challenges. More specifically, location data are generated using the R functions `rexp` for the exponential distribution and `runif` for the uniform distribution. Both are implemented in the R `stats` package. The data are generated using the following R code:

```
> lambda=20
> ifelse(gold_standard_network[i,3],rexp(n = 1, rate =lambda),
runif(1, min = 0, max = 1))
```

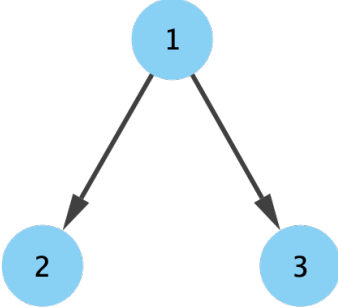
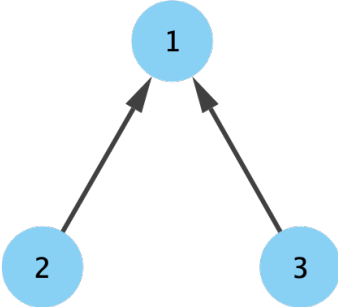
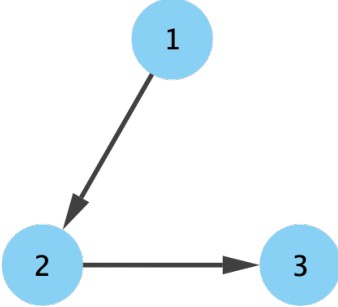
3.2.2 Performance Metrics

We use the DREAM4 challenge scoring methodology for a fair evaluation. We compute the AUROC, and AUPR as well as their respective p-values, p_{AUPR} and p_{AUROC} . The p-values are probabilities that random predictions would have the same or larger scores. As we inferred five networks for each subchallenge, we combined all the p_{AUPR} and p_{AUROC} into two global p-values (one for each score), which are used to compute a global score as in Equation 34:

$$S_G = -0.5 \log_{10} (p_{AUROC} * p_{AUPR}). \quad (34)$$

The global score is used to rank all the participants in the challenge. The larger the global score then, the more statistically significant is the prediction. More details can be found on the DREAM4 page <https://www.synapse.org/#!Synapse:syn3049712/wiki/74628>. We computed all these scores with `DREAMTools` version 1.3.0[41], the standalone application provided by the DREAM challenge team. We further assess the errors each method is making. We analyze how well each method predicts network edge motifs. We compute the motifs edges' prediction confidence, which is the edges' median rank in the final ordered edges list. The first edge in the list has 100% prediction confidence, and the last edge has 0% (we scaled the prediction confidence to the interval $[0, 1]$). For each inferred method, we extracted all instances of the three motifs. We then get the rank of all the edges in each motif. The point is to see how each edges motif is ranked in the output list from all concurrent methods. Note that we add missing links at the end of the inferred list if some links are omitted. We consider 3 types of motif: the **Fan-in**, the **Fan-out** and the **Cascade** motifs. These motifs are illustrated in Table 4. We use `GeneNetWeaver` [202] to perform the network motifs analysis.

Table 4: Motifs and errors type

	Network motif	Error types
Fan-out		Fan-out error: incorrect prediction of edges between coregulated genes ($2 \rightarrow 3$ and $3 \rightarrow 2$)
Fan-in		Fan-in error: low prediction of the edges $2 \rightarrow 1$ and $3 \rightarrow 1$ of the fan-in motif.
Cascade		Cascade error: incorrect prediction of indirect edge $1 \rightarrow 3$ in cascade motifs.

The table presents the three types of motifs we considered (represented in 2nd column) as well as the three types of error possibly ensued from network inference (represented in 3rd column). The nodes are the genes in the GRN. An arrow indicates that there is regulatory interaction between a transcription factor (source) and a target gene (sink). A non-arrow indicates that the genes are not interacting.

3.2.3 BENIN Parameters

BENIN is controlled by three main parameters: the number of bootstraps R , the elastic net mixing parameter α , and the power γ controlling the weight of the prior. We evaluated the importance of each of these parameters on BENIN’s performance using one independent 100-node network generated with `GeneNetWeaver`. We use the DREAM4 default setting. Note that we used location data as prior knowledge.

We proceed as follows, we fix two parameters, and we vary the third one. Starting with the default parameters $\alpha = 0.3$ and $R = 1000$ and $\gamma = 1$ we vary each of the parameters at a time as follows: $R \in \{5, 55, 105, \dots, 10000\}$, $\gamma \in \{0.1, 0.2 \dots, 1.5\}$ and $\alpha \in \{0.1, 0.2 \dots, 0.9\}$. Note that, we set the parameter λ_{Enet} with cross-validation as implemented in `glmnet` package. We chose the λ_{Enet} that yields the minimum mean squared error. We set the number of folds in cross-validation to 10.

3.2.4 BENIN Implementation

We implemented BENIN with R libraries: `glmnet` [71] version 2.0-13 (<https://cran.r-project.org/web/packages/glmnet/index.html>), `boot` version 1.3-20 [34] (<https://cran.r-project.org/web/packages/boot/index.html>). All computations were performed on server `Salus` with an Intel(R) Xeon(R) processor, 768GB of RAM and 56 cores. The execution time (elapsed time) for each network size is depicted in Table 5. The results reported in Table 5 are obtained setting the number of bootstraps to 1000 ($R = 1000$). All the other parameters are set to default.

Table 5: BENIN execution time on the DREAM4

Method	Network size	
	10	100
BENIN-non-optimized	602s	7200s
BENIN-parallel	50s	368s
BENIN +all-parallel	51s	899s

The table shows the elapsed time when using BENIN to reconstruct different size networks from the DREAM4 challenge. BENIN +all represents BENIN considering all potential genes as TFs. We specified whether or not we use parallel programming to optimize BENIN. The results reported here are obtained setting the number of bootstraps to 1000 ($R = 1000$). All the other parameters are set to default.

3.2.5 Comparison with the State-of-the-Art

First of all, we compared the performance of BENIN with three top-ranked teams of each DREAM4 sub-challenge. We named the top three methods as follows: **DREAM4 Winner** for the winner, **DREAM4 2nd** for the first runner up and **DREAM4 3rd** for the third-place finisher. Note that the winner of the size 100 sub-challenge is different from the winner of the size 10 sub-challenge. However, we do not have information about the first and the second runner up teams for both sub-challenges. At the time of the DREAM4 challenge, only information about the winners of the sub-challenges was made available. The winner of the size 10 sub-challenge [136] applied Petri nets to all provided datasets (knockout, time course, steady-state expression data, and knockdown expression data) to infer the networks. The winner of size 100 network [179] used z-scores combined with graph methodology to infer the networks from knockout expression data. We consider their scores as reported on the official website of the DREAM4 challenge: <https://www.synapse.org/#!/Synapse:syn3049712/wiki/74631>.

We further rigorously compared BENIN’s performance with the existing state of art methods, which have claimed to perform well on the DREAM4 challenge. They

use different methodologies for the GRN inference, and some of them integrate prior knowledge data. We have ensemble trees based methods (`dynGENIE3` [81], `iRafNet` [178]), pairwise mutual information (`TD-ARACNE` [272]), dynamic Bayesian network (`G1DBN` [141], `scanBMA` [258]), linear regression-based method (`gelNet` [216]). We used existing R packages for these methods. Whenever the implementation allows it, we specify the list of TFs (`dynGENIE3`). For methods that output regression coefficients, we rank the regulatory links using the absolute values of the coefficients. We use default parameters for `G1DBN`, `scanBMA` and `gelNet`. For the others, we set their parameters as specified in their papers. For integrating the KO expression data into the results of `dynGenie3`, we take the product of the scores `dynGenie3` and the Z-scores as suggested by the authors [81]. For combining the two priors into `BENIN`, we averaged the output scores of `BENIN +Location` (`BENIN` with location data as a prior) and `BENIN +KO` (`BENIN` using KO expression data as prior).

3.3 Computational Complexity

Investigating the complexity of `BENIN` amounts to investigate the complexity of the Elastic Net. As mentioned above, our method was implemented using `glmnet`. We particularly used the package function `cv.glmnet` to build our model. `cv.glmnet` uses cyclical coordinate descent to find the optimal $\vec{\beta}$. Cyclical coordinate descent successively optimizes the penalized regression equation over each parameter (β_i) while keeping others fixed, and cycles repeatedly until convergence. Through a cycle, two main types of variables update are used depending on the number of covariates. The naive update requires $O(Nd)$, where N is the number of samples and d the number of candidates covariates. It is used if the number of covariates is less than 500. The second type of update is the covariance update. When using this type of update, with m nonzero coefficients (β_i) in the model, a complete cycle costs $O(md)$ operations if no new variables become nonzero, and costs $O(Nd)$ for each new variable entered. The algorithm builds a grid of closely spaced λ -values $\{\alpha_l\}_{l=0}^L$. For each λ -value in the optimization path, the cyclical coordinate descent is repeated until the algorithm converges, to compute the coefficient vector β . The complexity deeply depends on the convergence rate of the cyclic coordinate descent. The convergence rate of coordinate descent minimization for solving linear systems is a classic topic.

Beck and L. Tetruashvili [12] have studied the cyclic coordinate descent for smooth function in general and have shown that it achieves a convergence rate of $O(1/\epsilon)$ under Lipschitz gradient condition and a rate of $O(\log(1/\epsilon))$ under strong convexity; where ϵ is a pre-specified accuracy of the target. Tseng *et.al* [234] have also studied the convergence of cyclic coordinate descent Note that the general case of smooth and separable function is not well understood. In summary, the worst-case complexity:

- Assuming a smooth function and Lipschitz gradient condition:

$$O\left(\frac{1}{\epsilon}Nd\right) \tag{35}$$

- Assuming a smooth function and strong convexity condition:

$$O\left(\log\left(\frac{1}{\epsilon}\right)Nd\right) \tag{36}$$

- More generally let s the number of steps till convergence, we have:

$$O(sNd) \tag{37}$$

A detailed analysis of the coordinate descent convergence rate can be found in [252], and for elastic net in [72].

The package makes use of techniques to fasten the convergence such as warm start (i.e., the solution $\beta(\lambda_l)$ is a warm start for the solution $\vec{\beta}(\lambda_{l+1})$), and the active-set convergence (which cause the algorithm to iterate only on variables which have nonzero coefficient). The algorithm uses k-fold cross-validation to select the best λ . In our GRN learning, the algorithm is repeated L bootstrap times.

3.4 Results and Discussion

3.4.1 Effect of the Noise in Prior Knowledge

We have investigated the effect of noise inherent in location data on the accuracy of BENIN. We generated several location datasets with varied level of reliability, by fluctuating λ . Note that the larger is λ , the more reliable will be the location data in the sense the p-values of regulatory links will be close to zero. In our experiment

we chose $\lambda = \{1, 10, 20, 100\}$ leading to 4 different location datasets that we name: *completely noisy location data*, *reasonably noisy location data*, *fair location data* and *perfect location data*. Figure 13 shows the result of BENIN when varying the noise in location. We plot it only for 100-nodes networks and principally for the easiest network to infer (network 1) and the most difficult to infer (network 5). As expected, we observe that as the prior becomes perfect, BENIN gets better performance.

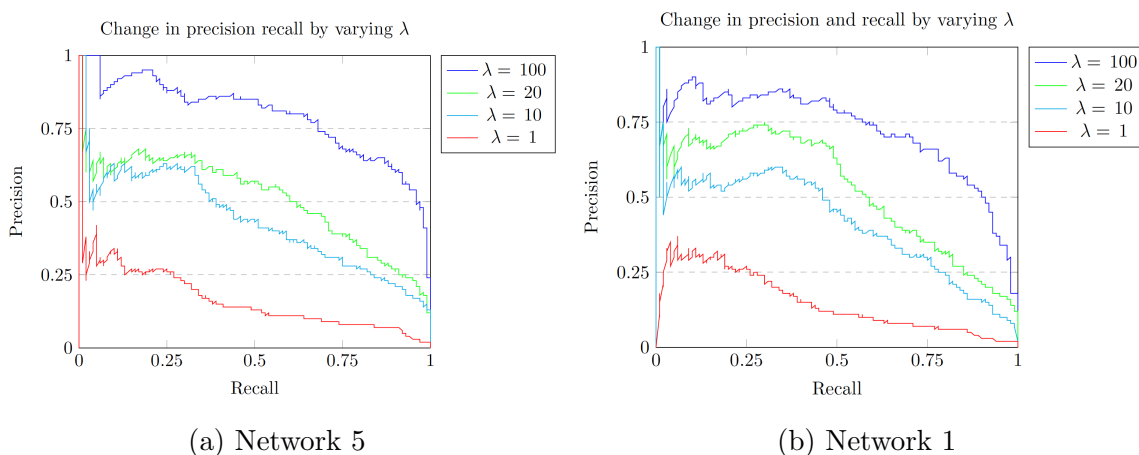


Figure 13: Effect of the noise in location data

The figures show the AUPR when learning 100-nodes DREAM 4 with different types of location data as prior knowledge. We report data with three different levels of noise: completely noisy ($\beta = 1$), reasonably noisy ($\beta = 10$) and fair location data ($\beta = 20$) and, perfect ($\beta = 100$) location data. The graph shows that as the level of noise decrease in the data, the performances of BENIN increase.

3.4.2 Influence of BENIN Parameters

Figure 14 shows that the quality of BENIN prediction is less sensitive to α than to the two other parameters. In fact, for different values of α , the AUPR score does not vary much from 0.5. On the other hand, as γ increases, BENIN yields higher AUPR scores (Figure 14c) but after $\gamma = 1$ the performance starts to decrease. Furthermore, we can also observe that as R increases, there is an improvement in performance that stabilizes for $R \geq 5000$ (Figure 14b). In effect, increasing the number of bootstraps improves the chance to select the true TFs in the model. From these tables, the most important parameters are R and γ .

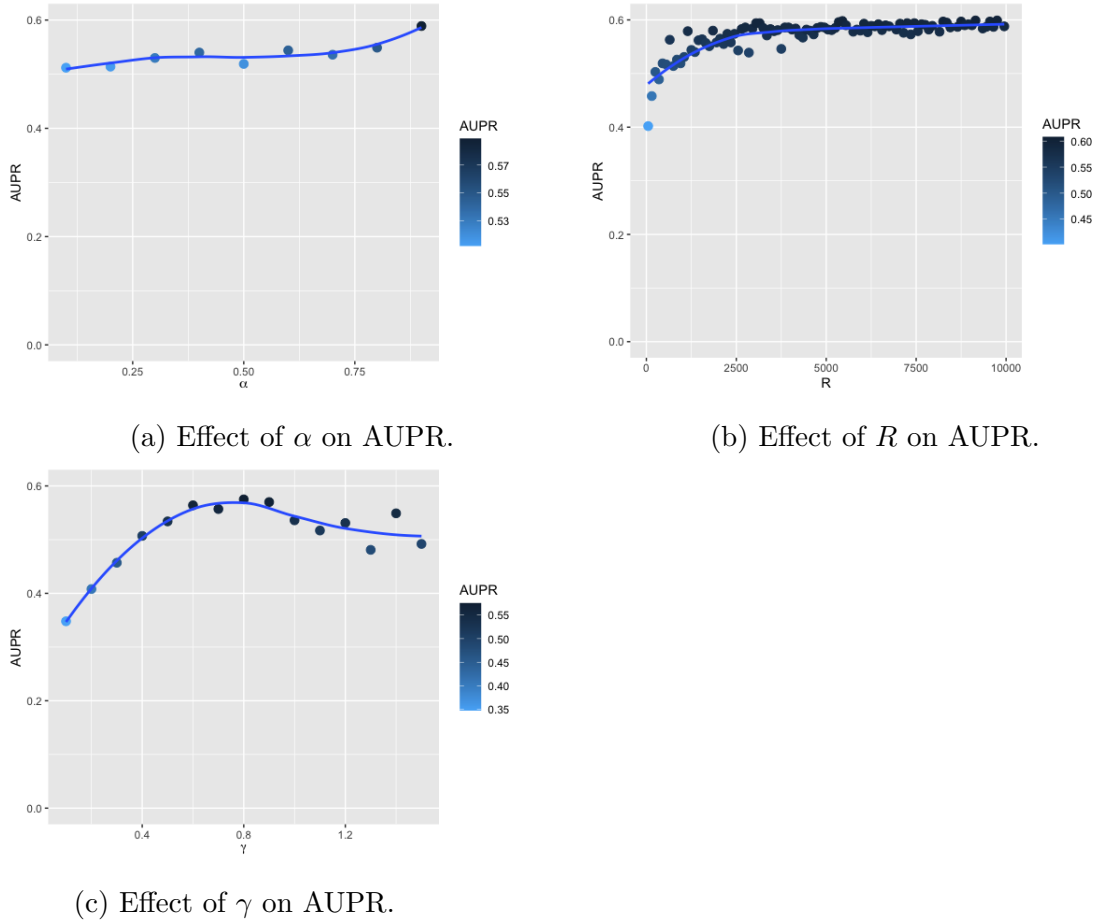


Figure 14: Influence of BENIN parameters

We consider the AUPR score of BENIN on a 100-node network when varying each of the parameters: α which controls the penalization strength in the Elastic net, the number R of bootstrap samples and the parameter γ that weights the influence of the prior. We used location data as our prior. We vary $R \in \{5, 55, 105, \dots, 10000\}$, $\gamma \in \{0.1, 0.2 \dots, 1.5\}$ and $\alpha \in \{0.1, 0.2 \dots, 0.9\}$.

3.4.3 Effect of Prior Knowledge

The global scores, when considering each prior separately (or combined), are reported in Table 8 for size 100 network and Table 11 for size 10 networks. The associated AUPR and AUROC scores are detailed in Table 6 (respectively Table 10) when knock-out expression data is considered as prior knowledge and in Table 7) (Table 9) when genome-wide location data is considered as prior knowledge for the GRN inference

of size 100 (respectively size 10) DREAM4 networks. From Table 8 and Table 11, we first observe that the inclusion of prior knowledge into **BENIN** drastically improves its performances. We further notice in Table 8 and Table 11 that each type of prior knowledge data yields different performances. Including Location data as prior knowledge yields better results compared to KO expression data on both sub-challenges. Location data are more informative than KO expression data. In Section 3.4.6, we dug up into the results on size 100 networks to see the contribution of each data type. However, not surprisingly, the combination of KO expression data and genome-wide location yields superior results compared to **BENIN**'s performance using each prior separately and **BENIN** when we do not consider any prior. These results confirm the benefit of the integration of prior knowledge into a model for GRN inference.

We evaluated **BENIN**'s performances without restricting the set of potential TFs on the DREAM4 challenge. From Table 5 and Table 8, we observe that restricting the input TFs to the list of known TFs improved **BENIN**'s performance in two directions. First, we observe from Table 5 that, when we consider all potential genes as TFs, the execution time increases. On the other hand, from Table 8, we further observe an increase in the global score on the DREAM4 challenge when we restrict the potential TFs to known TFs. These two observations confirm the need to use and invest in methods for identifying TFs using techniques such as sequences annotation, homology, identification of DNA binding domain (DBD), and wet-lab experiments. Many resources help predict TFs from the protein sequences and, several databases store information about the TFs. We can list **AnimalTFDB** [104] or **JASPAR** [198]. There is a need to integrate this information as prior knowledge into the GRN inference for scaling up when inferring large network, but also in order to obtain more biologically meaningful GRN.

3.4.4 Performance on the DREAM4 Challenge

On the first hand, when considering only KO expression as prior knowledge, from Table 6, we observe that **BENIN** gets a better score than the winner for the size 100 sub-challenge, which uses knockout gene expression data alone to infer all five networks in the sub-challenge. We notice that our AUROC score is the highest on almost all the five networks (except for network 2 and 3). Moreover, our AUPR scores

are far better than the performances of the size 100 sub-challenge participants. We principally care about the AUPR score and the final global score. The AUPR score is more informative than the AUROC score in the case of imbalanced datasets [196]. Regulatory networks are such an imbalanced case, as the number of true links is far less than the number of non-links (sparse network). Our final global score on size 100 networks considering KO expression data as prior knowledge indicates that our performance is more statistically significant than those of all other participants. However, when considering KO expression data as prior knowledge, BENIN gets the 2nd best score for size 10 sub-challenge. This result is not surprising as the winner of this challenge integrates all the data that were made available in the challenge, proving the power of data integration.

On the other hand, when we combine KO expression data and location data with time-series expression, we notice from Table 8 and Table 11 that BENIN gets a better score than the winners of both sub-challenges. This result first testifies that location data are very informative and confirm that the integration of several data with time-series expression data improves BENIN’s performance.

3.4.5 Comparison with the State-of-the-Art

From Table 7, Table 6, Table 9 and, Table 10 we observe that, for both size 10 and size 100 networks, BENIN significantly outperforms the state of the art methods, particularly when considering genome-wide location data as prior knowledge. However, when we consider KO expression data, we observe in Table 6 that, for size 100 network 2 and 4, dynGENIE3+KO gets better results than BENIN but in average, BENIN’s performance is superior to dynGENIE3+KO. From Table 8 and Table 11, BENIN overall performance when considering both prior knowledge data confirms the statistical significance of our results.

Table 6: DREAM4 size 100 performance with KO expression

Algorithm	Net 1	Net 2	Net 3	Net 4	Net 5
BENIN + KO	0.611(0.964)	0.455(0.925)	0.442 (0.923)	0.496 (0.932)	0.403 0.927)
BENIN + all + KO	0.516 (0.913)	0.322(0.783)	0.373(0.835)	0.384(0.831)	0.250(0.765)
gelNet	0.042(0.695)	0.047(0.631)	0.096(0.669)	0.051(0.647)	0.056(0.682)
BENIN- no prior	0.306 (0.904)	0.218 (0.872)	0.275 (0.860)	0.279 (0.880)	0.279 (0.911)
TDARACNE	0.063(0.656)	0.066(0.613)	0.077(0.642)	0.073(0.618)	0.069(0.651)
scanBMA	0.119(0.685)	0.064(0.625)	0.146(0.658)	0.116(0.662)	0.099(0.693)
G1DBN	0.058(0.789)	0.064(0.7)	0.057(0.728)	0.051(0.727)	0.064(0.771)
dynGENIE3+ KO	0.559(0.964)	0.483(0.933)	0.409(0.933)	0.528(0.938)	0.340(0.922)
dynGENIE3+ all + KO	0.481(0.920)	0.352(0.807)	0.350(0.849)	0.458(0.857)	0.283(0.788)
iRafNet+ KO	0.476(0.888)	0.295(0.791)	0.383(0.829)	0.356(0.839)	0.237(0.789)
KO z-score	0.521(0.962)	0.453(0.930)	0.412(0.924)	0.404(0.932)	0.214(0.913)
DREAM4 Winner	0.536(0.914)	0.377(0.801)	0.390(0.833)	0.349(0.842)	0.213(0.759)
DREAM4 2 nd	0.512(0.908)	0.396(0.797)	0.380(0.829)	0.372(0.844)	0.178(0.763)
DREAM4 3 rd	0.490(0.870)	0.327(0.773)	0.326(0.844)	0.400(0.827)	0.159 (0.758)

The table reports the AUPR and AUROC (in brackets) for each of the five networks of the size 100-node DREAM4 subchallenge for different algorithms with KO expression data as prior information. The highest score is shown in bold.

Table 7: DREAM4 size 100 performance with Location data

Algorithm	Net 1	Net 2	Net 3	Net 4	Net 5
BENIN +Location	0.599(0.983)	0.562(0.979)	0.534(0.972)	0.580(0.982)	0.615(0.985)
BENIN +all+Location	0.356(0.955)	0.318(0.943)	0.352(0.940)	0.353(0.953)	0.357(0.952)
BENIN-no prior	0.306 (0.904)	0.218(0.872)	0.275(0.860)	0.279(0.880)	0.279(0.911)
ge1Net+Location	0.033(0.648)	0.038(0.601)	0.088(0.636)	0.043(0.642)	0.048(0.677)
TDARACNE	0.063(0.656)	0.066(0.613)	0.077(0.642)	0.073(0.618)	0.069(0.651)
scanBMA+Location	0.149(0.833)	0.093(0.7611)	0.175(0.8276)	0.144(0.787)	0.131(0.829)
G1DBN	0.058(0.789)	0.064(0.7)	0.057(0.728)	0.051(0.727)	0.064(0.771)
iRafNet+Location	0.328(0.943)	0.327(0.941)	0.408(0.953)	0.344(0.946)	0.400(0.956)
dynGENIE3	0.251(0.8918)	0.225(0.889)	0.165(0.883)	0.270(0.888)	0.207(0.903)
dynGENIE3+all	0.196(0.761)	0.111(0.664)	0.106(0.723)	0.194(0.725)	0.124(0.730)

The table presents the AUPR and AUROC (in brackets) for each of the five networks in the 100-nodes DREAM4 subchallenge for different algorithms with location data as prior information. They are the geometric mean of the scores obtained on the eleven generated location datasets. The highest score is shown in bold.

Table 8: Global score on the DREAM4 size 100 subchallenge

Algorithm	Methods	Global Score	Prior
BENIN + Both		129.563	KO+Location Data
BENIN + Location		122.716	Location Data
BENIN + KO	Regression	100.383	KO
BENIN + all+location data		82.200	Location data
BENIN- no prior		61.431	None
gelNet+ KO		11.078	KO
gelNet+ Location		8.626	Location Data
dynGENIE3+ KO		99.917	KO
iRafNet+ Location	Tree Ensemble	84.193	Location Data
dynGENIE3+ all+KO		73.748	KO
iRafNet+ KO		66.071	KO
dynGENIE3		56.695	None
dynGENIE3+ all		26.662	None
scanBMA+ Location		33.207	Location Data
scanBMA	Dynamic Bayesian Network	17.476	None
G1DBN		16.922	None
TDARACNE	Mutual Information	11.084	None
DREAM4 Winner		71.589	None
DREAM4 2 nd	Other	71.297	No information
DREAM4 3 rd		64.715	No information
KO z-score		90.291	None

The table reports the global scores of different inference methods combined with or without prior knowledge for inferring the five networks of the DREAM4 size 100 subchallenge. The Prior column specifies the type of prior information used. See Table 6 and Table 7 for more details of this table.

Table 9: DREAM4 size 10 performance with Location data

Algorithm	Net1	Net2	Net3	Net4	Net5
BENIN +Location	0.817(0.952)	0.726 (0.910)	0.804(0.949)	0.856(0.957)	0.915(0.975)
BENIN +allgenes+Location	0.805 (0.944)	0.693 (0.897)	0.799 (0.947)	0.840 (0.953)	0.891 (0.974)
BENIN +no prior	0.502 (0.847)	0.465 (0.666)	0.441 (0.722)	0.752 (0.924)	0.205 (0.567)
gelNet+Location	0.363 (0.723)	0.234 (0.606)	0.215 (0.621)	0.324 (0.740)	0.319 (0.737)
TDARACNE	0.379 (0.756)	0.270 (0.684)	0.313 (0.620)	0.308 (0.638)	0.409 (0.687)
scanBMA	0.453 (0.633)	0.433 (0.615)	0.325 (0.567)	0.470 (0.654)	0.483 (0.667)
G1DBN	0.507(0.772)	0.416 (0.664)	0.418 (0.750)	0.499 (0.760)	0.652 (0.824)
iRafNet+Location	0.714 (0.935)	0.630 (0.904)	0.711 (0.910)	0.669 (0.911)	0.805 (0.955)
dynGENIE	0.612 (0.876)	0.484 (0.702)	0.765 (0.854)	0.686 (0.922)	0.595 (0.842)
dynGENIE+allgenes	0.483 (0.743)	0.419 (0.636)	0.512 (0.758)	0.484 (0.734)	0.669 (0.834)

The table reports the AUPR and AUROC (in brackets) for each of the five networks of the size 10-node DREAM4 sub-challenge for different algorithms with Locations data as prior information. The highest score is shown in bold.

Table 10: DREAM4 size 10 performance with KO expression

Algorithm	Net1	Net2	Net3	Net4	Net5
BENIN +KO	0.799 (0.928)	0.572 (0.742)	0.649 (0.919)	0.796 (0.944)	0.709 (0.902)
BENIN +allgenes+KO	0.738 (0.908)	0.533 (0.659)	0.728 (0.936)	0.849 (0.959)	0.626 (0.853)
BENIN +no prior	0.502 (0.847)	0.465 (0.666)	0.441 (0.722)	0.752 (0.924)	0.205 (0.567)
gelNet+KO	0.363 (0.723)	0.234 (0.606)	0.215 (0.621)	0.324 (0.740)	0.319 (0.737)
TDARACNE	0.379 (0.756)	0.270 (0.684)	0.313 (0.620)	0.308 (0.638)	0.409 (0.687)
scanBMA	0.453 (0.633)	0.433 (0.615)	0.325 (0.567)	0.470 (0.654)	0.483 (0.667)
G1DBN	0.507(0.772)	0.416 (0.664)	0.418 (0.750)	0.499 (0.760)	0.652 (0.824)
iRafNet+KO	0.646 (0.879)	0.272 (0.708)	0.657 (0.847)	0.563 (0.790)	0.573 (0.873)
dynGENIE+KO	0.612 (0.876)	0.484 (0.702)	0.765 (0.854)	0.686 (0.922)	0.595 (0.842)
dynGENIE+KO+allgenes	0.611 (0.865)	0.475 (0.667)	0.751 (0.834)	0.694 (0.927)	0.593 (0.830)
DREAM4 Winner	0.916(0.972)	0.547(0.841)	0.968(0.990)	0.852(0.954)	0.761(0.928)
DREAM4 2 nd	0.881(0.967)	0.382(0.796)	0.682 (0.916)	0.698 (0.902)	0.424 (0.822)
DREAM4 3 rd	0.623 (0.864)	0.301 (0.567)	0.646 (0.824)	0.693(0.820)	0.673(0.776)
KO z-score	0.638(0.835)	0.262(0.666)	0.701(0.840)	0.776(0.942)	0.405(0.723)

The table presents the AUPR and AUROC (in brackets) for each of the five networks in the 10-nodes DREAM4 sub-challenge for different algorithms with location data as prior information. They are the geometric mean of the scores obtained on the eleven generated location datasets. The highest score is shown in bold.

Table 11: Global score on the DREAM4 size 10 subchallenge

Algorithm	Method	Global score	Prior
BENIN +Both		7.481	Location+KO
BENIN +Location		7.324	Location
BENIN +all+Location		7.139	Location
BENIN +KO	Regression	5.802	KO
BENIN +all+KO		5.535	KO
BENIN +no prior		3.272	None
gelNet+Location		1.696	Location
gelNet+KO		1.626	KO
dynGENIE3+KO		4.814	KO
dynGENIE3+all+KO		4.657	KO
iRafNet+Location	Tree Ensemble	4.965	Location
iRafNet+KO		4.140	KO
dynGENIE3		3.222	None
dynGENIE3+all		3.206	None
G1DBN	Dynamic Bayesian Network	3.222	None
scanBMA		2.022	None
TDARACNE	Mutual Information	1.859	None
DREAM4 Winner		7.127	KO
DREAM4 2 nd	Other	5.290	No information
DREAM4 3 rd		3.968	No information
KO z-score		4.120	None

The table reports the global scores of different inference methods combined with or without prior knowledge for inferring the five networks of the DREAM4 size 10 subchallenge. The Prior column specifies the type of prior information used. See Table 10 and Table 9 for more details of this table. If a method uses all the genes as potential TFs we specify it with “+all”.

We dug up into the inferred networks and Figure 15 shows an example of a sub-network from the 4th network of the size 100 sub-challenge. Figure 15 shows how the subnetwork is inferred by each method. The subnetwork is anchored in a critical/hub transcription factor “G64”: i.e., a transcription factor linked to many other genes. In this figure, the gray links are the links missed by the method, the red links are the false positives, and the green links are the true links. The subnetwork is inferred with different accuracy by the different methods. Note that we restricted the subnetwork to the top 20 edges for each method. As expected, methods that do not consider prior knowledge miss many links and have the highest false positives. Location data are most informative than KO expression data. We can observe that methods that consider location data as prior knowledge can infer the true edge with fewer false-positive links. From the network inferred by **BENIN**, when we combine both KO expression and location data, we can observe that the prior are complementary. We observe that **BENIN** infers less number of false-positive links than when we consider KO. However, we are still missing some links. We further perform network motif analysis to highlight the types of error each method is doing.

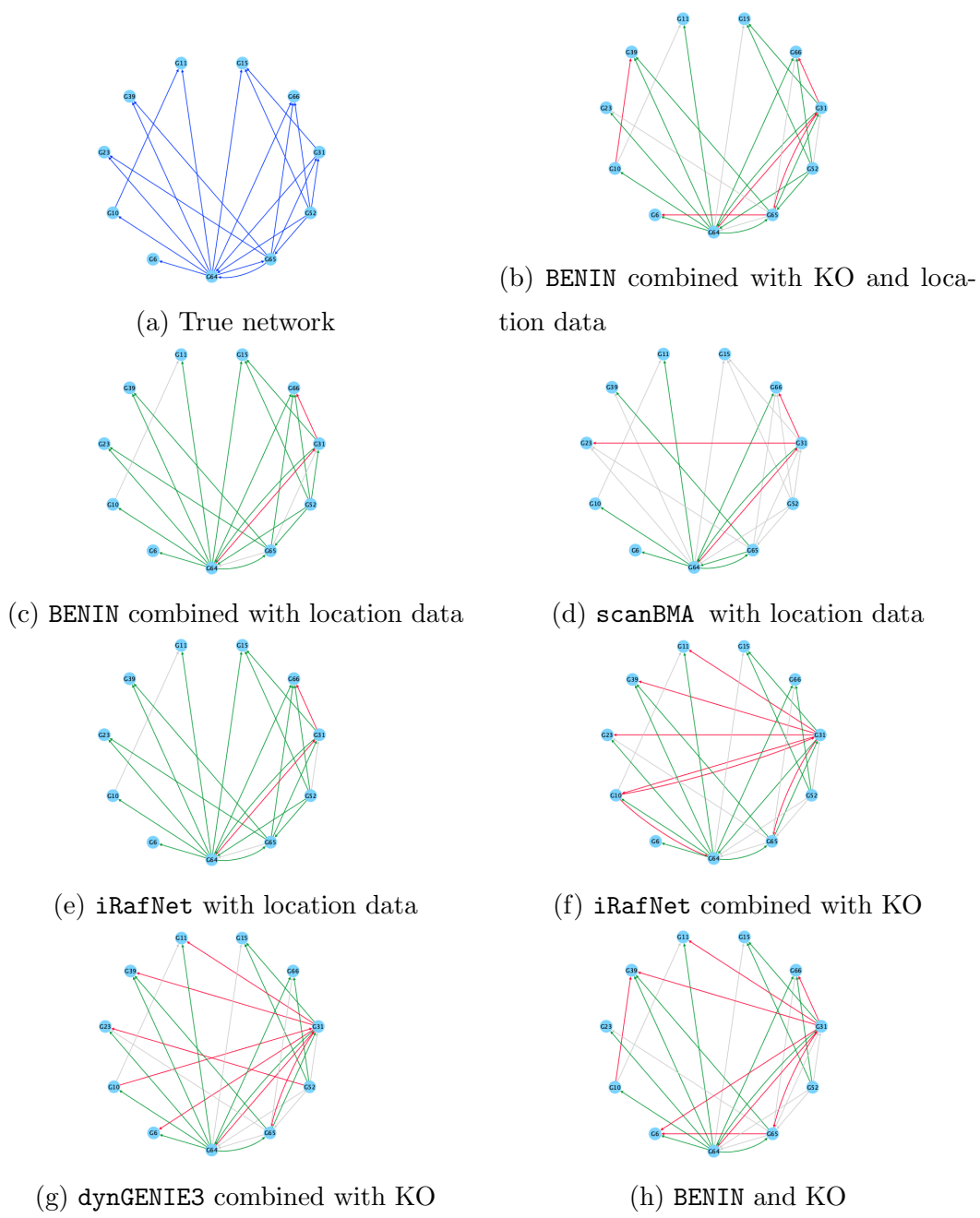
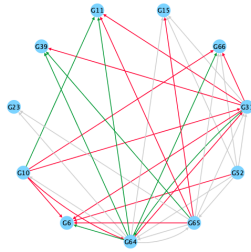
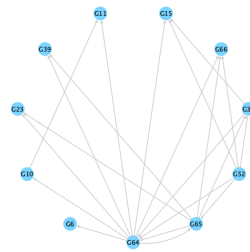


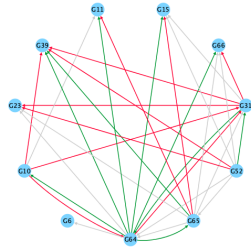
Figure 15: A subnetwork from 100-nodes network 4
The figure reports how each of the methods infers a subnetwork (sub-figure a) from the 4th network in the size 100 DREAM4 subchallenge. We consider a subnetwork anchored on a *key* transcription factor, i.e., a TF linked to many other genes. In the figures, green links represent the true positives, red links represent the false positives, and finally, gray links are edges missed by the method.



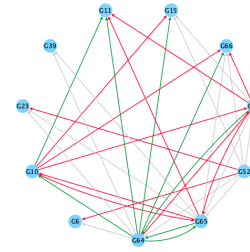
(i) gelNet and KO



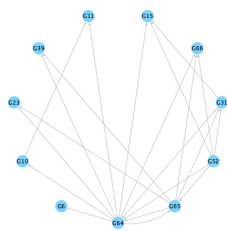
(j) TDARACNE without prior



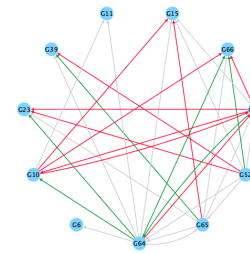
(k) BENIN without prior



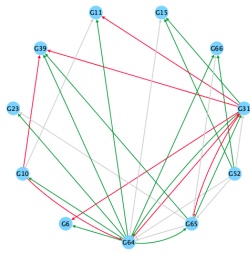
(l) dynGENIE3 without prior



(m) scanBMA without prior



(n) G1DBN without prior



(o) KO z-score

A subnetwork from 100-nodes network 4

The figure reports how each of the method infers a subnetwork (sub-figure a) from the 4th network in the size 100 DREAM4 subchallenge. We consider a subnetwork anchored on a *key* transcription factor i.e. a TF linked to many other genes. In the figures green links represent the true positives, red links represent the false positives and finally, gray links are edges missed by the method.

3.4.6 Network Motif Analysis

For each method, we only report the motif prediction confidence on the 4th network from the size 100 DREAM4 subchallenge, as it is the one where we perform less than the state-of-art (more specifically when combining BENIN with KO expression data). Furthermore, the difference in the error profiles between all methods is remarkable on this network. The network motifs analysis will help us pinpoint where the errors are being made by each method and the influence of each type of prior knowledge data on BENIN prediction. We extracted 642 fan-out motifs, 250 fan-in motifs, and 187 cascade motifs from these networks. We use **GeneNetWeaver** to analyze how well the edges of these motifs are inferred by BENIN, **iRafNet**, **dynGENIE3**, **gelNet**, **scanBMA**, **TDARACNE**, **G1DBN** and **z-score**, when they consider or not prior knowledge data (KO and/or Location data). Table 12 presents the error profile for each method. The first row stores the true structure. Here, the black edges are those we want to infer. The intensity of the edge color is proportional to its prediction confidence (median rank).

Different methods, different error profile: From Table 12, we can observe that each method is affected to a different degree by each error. Therefore each method has different error profile, demonstrating that various method has different strength and weakness.

Considering first the Fan-out motif, we observe from Table 12 that almost all methods except BENIN and G1DBN have the tendency to confuse co-regulation and regulation, and infer regulatory links between co-regulated genes. The most affected methods are TDARACNE, and scanBMA+noprior. We observe that the median rank of the true edges ($1 \rightarrow 2$ and $1 \rightarrow 3$) is very close to the median rank of the false edges ($2 \rightarrow 3$ and $3 \rightarrow 2$): these methods rank edges between co-regulated genes on average as good as the true regulatory links. The other affected methods (**dynGENIE3**, **iRafNet** and **gelNet**) although affected by the error, rank the true regulatory links at the top of their inferred list of regulatory links. Moreover we observe that some of these methods (**dynGENIE3**, **iRafNet**, **gelNet**, **scanBMA+Location** and **TDARACNE**) have difficulty to infer the directionality of the edges. On the other hand, we can observe that BENIN can clearly distinguish co-regulation and regulation, but also can distinguish the directionality of the edges.

Looking up at the Fan-in motif, we observe that `dynGENIE+noprior`, `gelNet+KO`, `scanBMA+KO`, and `scanBMA+noprior` have the difficulty to rank edges targeted by many TFs at the top of the inferred list of regulatory links. We also notice that methods that do not incorporate prior knowledge with expression data are mostly affected by this error. It is principally the case for `scanBMA`, which is the most affected by this error. The inclusion of prior knowledge data into the network inference helps the method to rank combinatorial links among the top edges.

Finally, observing the Cascade motif, we can see that methods that integrate KO expression data as prior knowledge are the most affected by this error: `BENIN+KO` is the most affected by this error. They give higher rank to the indirect edge $1 \rightarrow 3$ compared to the true edges ($1 \rightarrow 2$ and $2 \rightarrow 3$). It is not surprising since, as if we look at the prediction confidence of motif edges with KO expression data alone (`KO-zscore-alone`), we can see that the median rank of the indirect edge is 0.91. It is normal since KO expression data helps to infer direct links and indirect interactions as perturbing a TF will affect not only its direct TGs but also its indirect TGs. On the other hand, we observe that methods that consider location data (`iRafNet+Location` and `scanBMA+Location`) rank the true edges of the motif on average at the top of their inferred list of regulatory links, demonstrating that they are not affected by the cascade error.

Table 12: Motif prediction confidence (median rank)

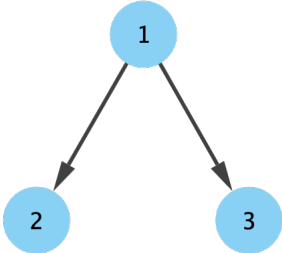
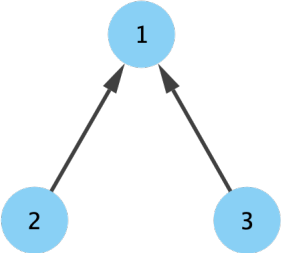
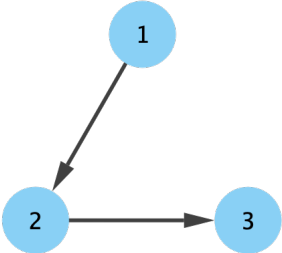
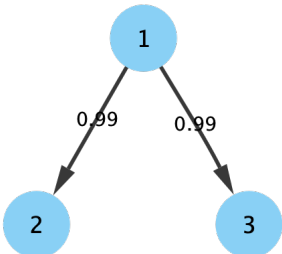
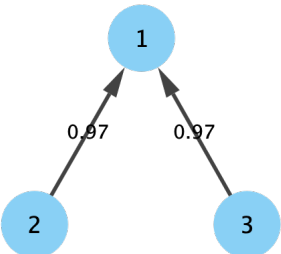
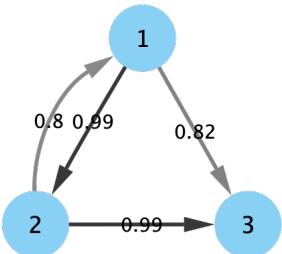
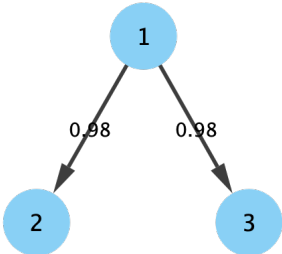
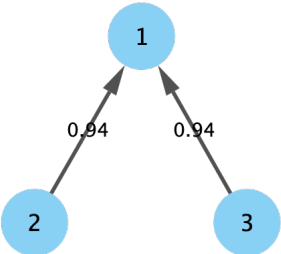
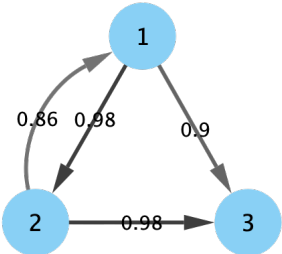
		Fan-out	Fan-in	Cascade
Methods	Motifs			
	BENIN-combined			
	BENIN +KO			

Table 12 continued from previous page

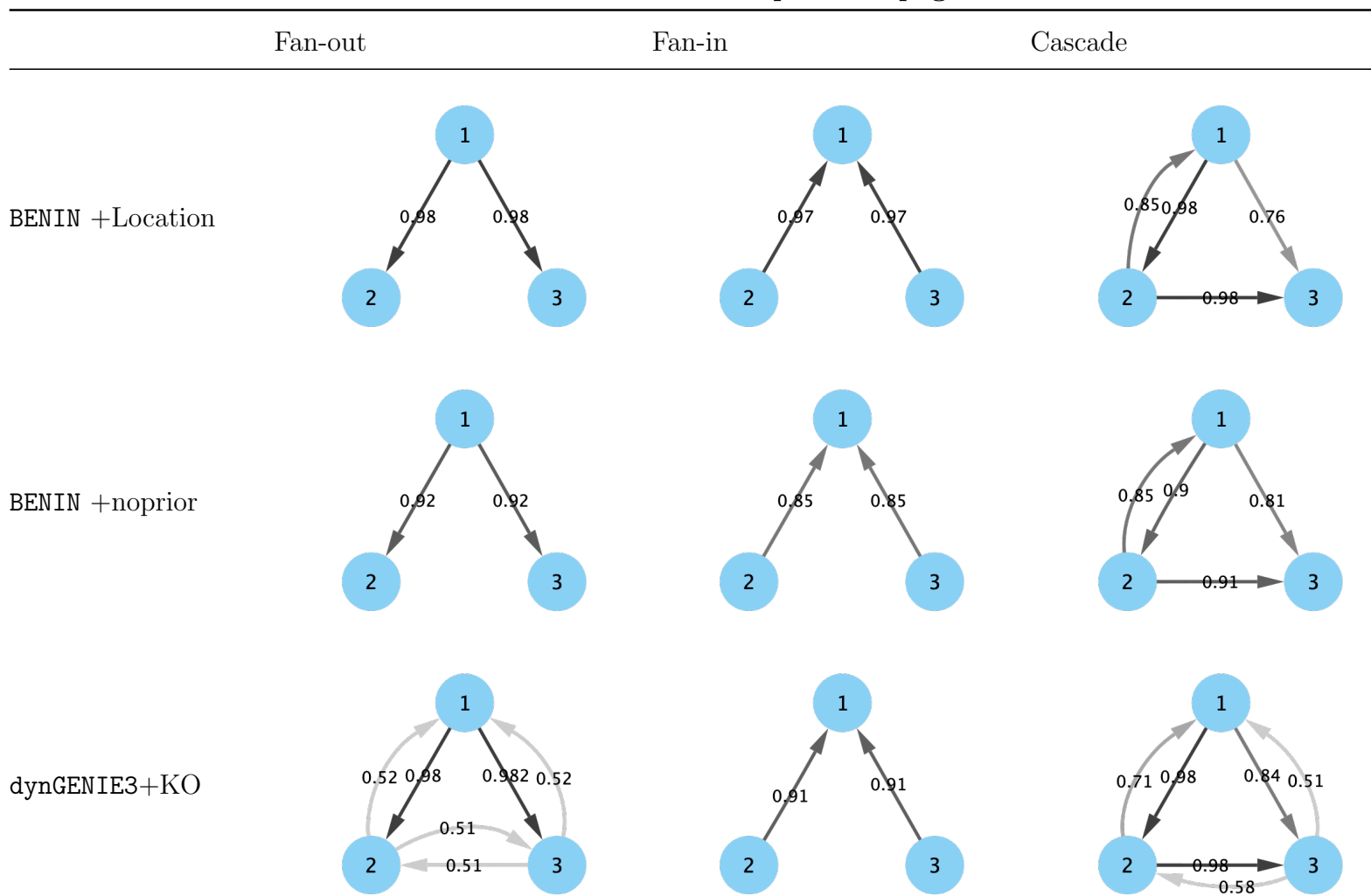


Table 12 continued from previous page

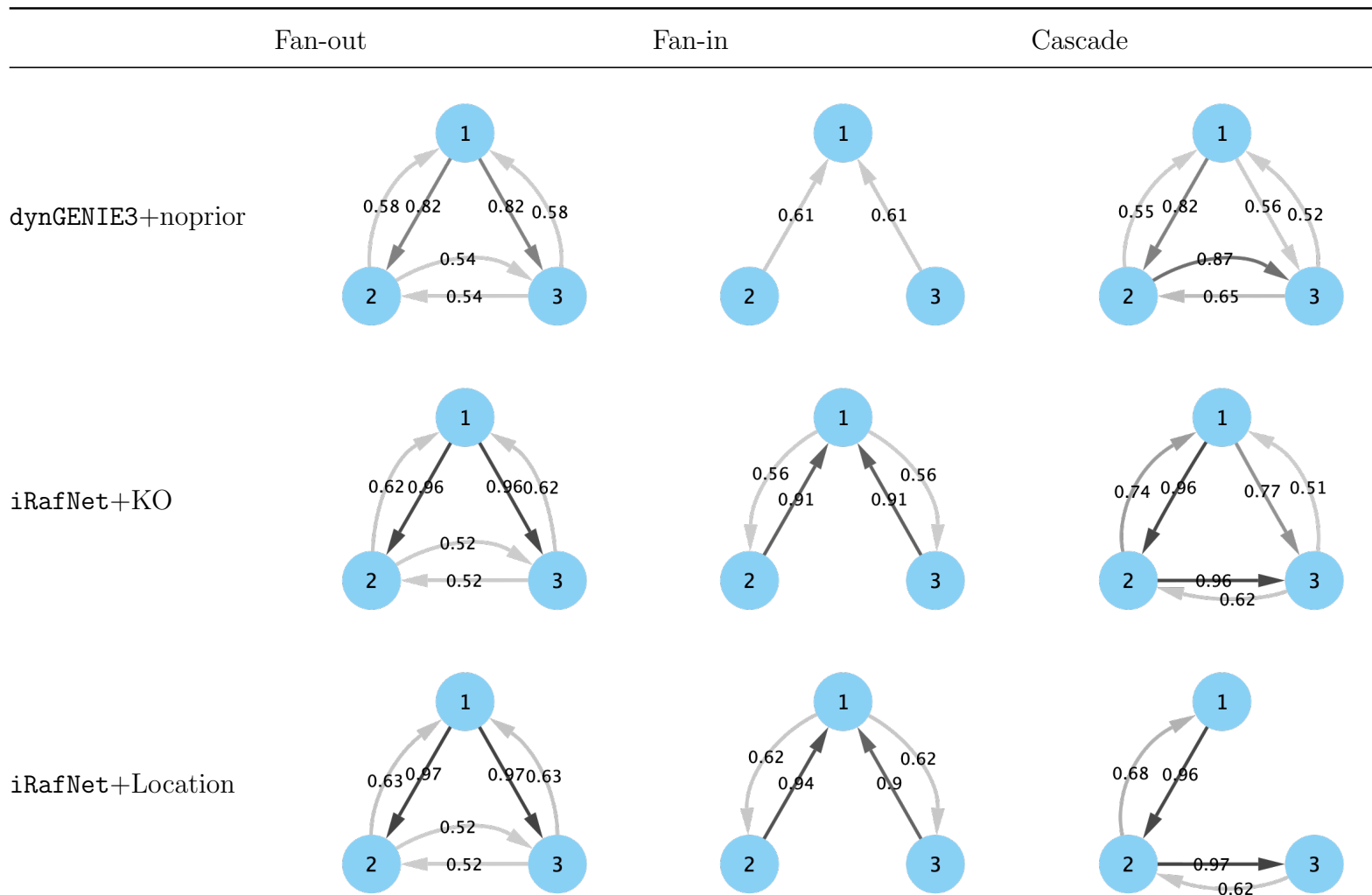


Table 12 continued from previous page

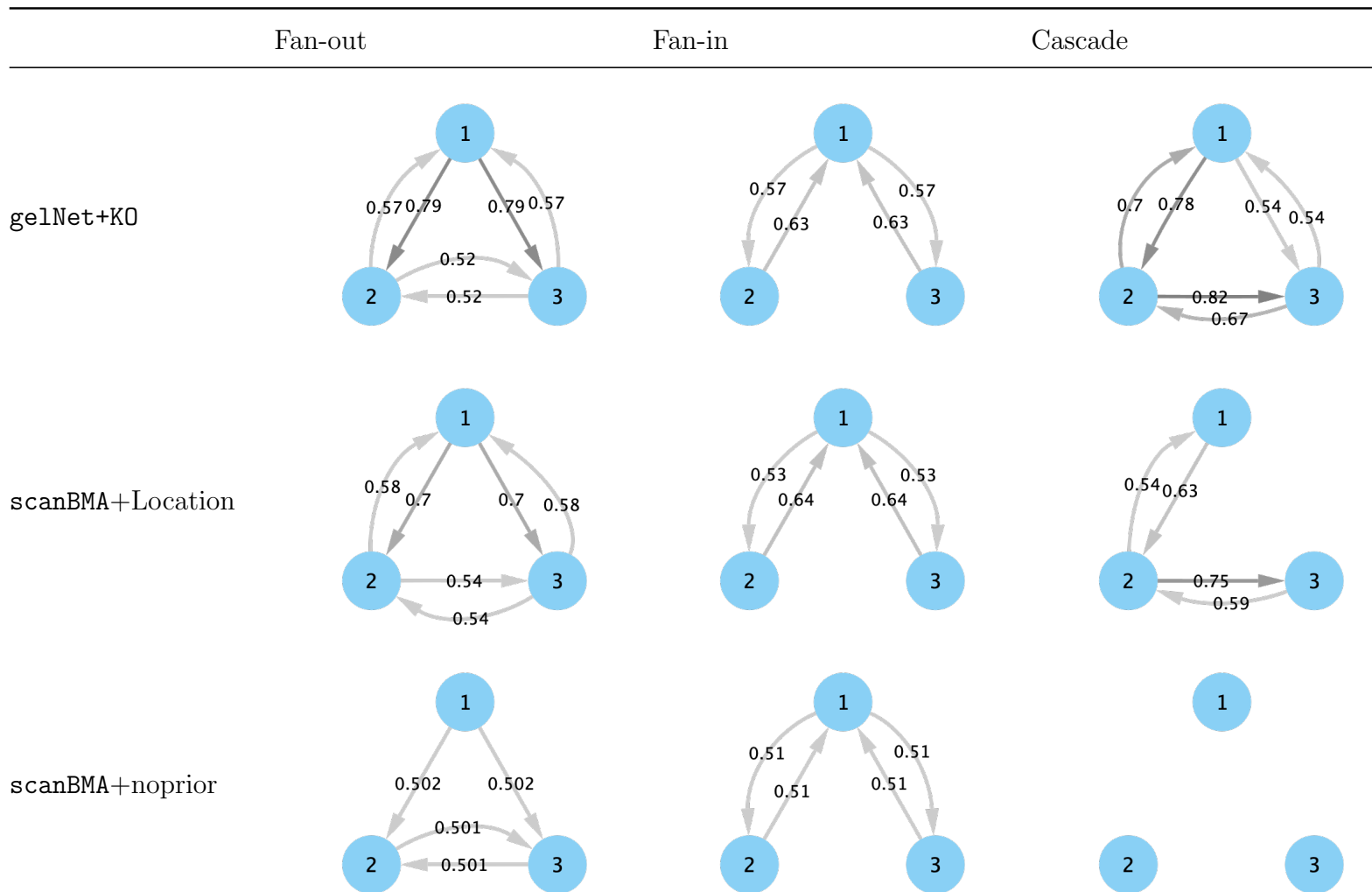
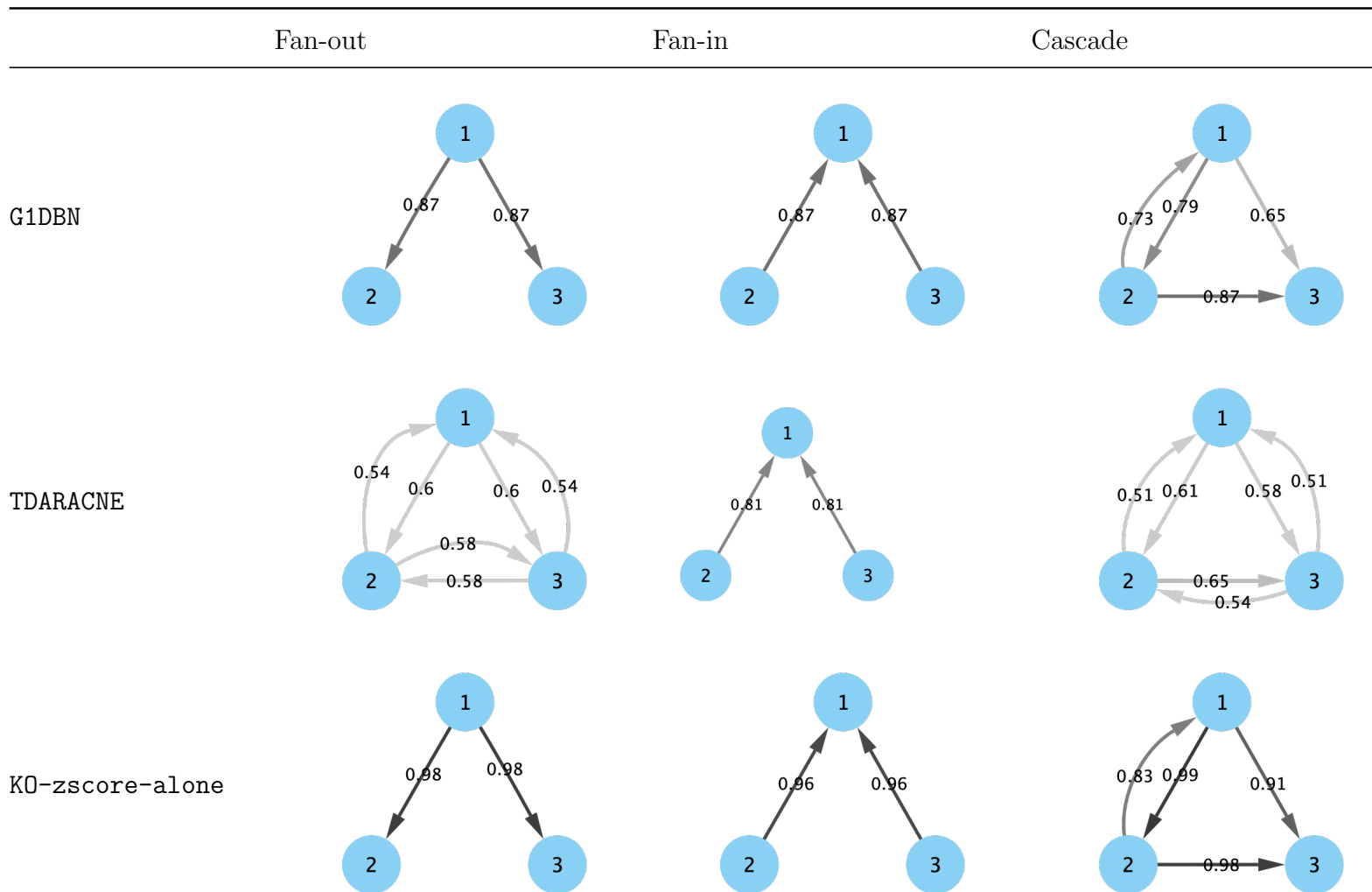


Table 12 continued from previous page



The table shows the median prediction (median rank) of edges motifs on the network 4 from the size 100 DREAM4 subchallenge. Using **GeneNetWeaver**, we extracted 642 fan-out motifs, 250 fan-in motifs and 187 cascade motifs from the networks inferred by each method. The first row stores the true structure of the motifs with regulatory the true links shown in black. The first column is the Fan-out motif. 2^{nd} column is the Fan-in motif and finally, 3^{rd} column is the Cascade motif. Each row shows the motifs as inferred by each method. In these motifs, the intensity of the color is proportional to the median prediction confidence (median rank). The labels of the edges are the median prediction confidence values.

3.5 Conclusion

In this chapter, we introduced **BENIN**, a framework that infers regulatory networks by jointly learning from time-series expression data and prior knowledge. The prior knowledge serves to derive weights that are then used to penalize non-potential interactions and thus lead toward a more intuitive solution. The proposed method utilizes a popular algorithm the **Elastic net**, which permits a direct and simple integration of prior knowledge while the model is learned. In this chapter, we report **BENIN**'s performance on a widely used benchmark dataset for network inference assessment: the DREAM4 challenge dataset. We combined time-series expression data with simulated KO and genome-wide location data. We compared our performance to state-of-the-art methods.

A simple but efficient method. Compared with existing integration models, which mainly rely on the Bayesian network framework, the advantage of **BENIN** is the simplicity of the model and its simplicity to integrate the prior knowledge. Bayesian-based methods are computationally demanding [38, 148]; they generally require many samples to learn the model. Above all, they require knowledge of the prior data to choose the right prior distribution that will fit the knowledge we want to integrate. Our results on simulated data demonstrate that even a simple model with proper integration of prior knowledge can be competitive with sophisticated methods. Care should be taken with the quality of the prior knowledge data because very noisy data may worsen the algorithm's performance. In our algorithm, this problem is handled at two levels. The first level is the adoption of a probabilistic model to define the prior. In that way, we use prior knowledge to guide network inference without making a strong assumption about their accuracy. The second level is in the model building itself. Our algorithm offers the possibility to control the feature penalization.

Prior knowledge boosts the network inference In this chapter, we have also demonstrated that joint learning from expression data and informative prior knowledge is beneficial. Not surprisingly, the inclusion of prior evidence in the network inference substantially increases our algorithm's performance. When we compare the error profile of several state-of-the-art methods and **BENIN** when they integrate or

not knockout data and the transcription factor binding location data, we notice a complementary in their performance. Different methods are robust against different errors. For example, **BENIN** is robust for inferring edge targeted by several TFs and distinguishing co-regulation and regulation. On the other hand, **iRafNet** is robust against the cascade error. We further observe that different methods are affected by different error types depending on the type of prior knowledge data integrated. This complementarity in the performances was expected because different data sources will tell different parts of the story about the regulatory network and have different noise levels.

What is next? Although the results of **BENIN** are encouraging, a lot still needs to be done. In this chapter, we only presented preliminary results on simulated data. In the next chapter, we will confirm our result on real data on human expression data and consider other types of prior knowledge, such as ChIP-seq/ChIP-chip data, functional similarity, or protein-protein interactions. The method presented here for combining results from different priors is very simplistic. Alternatively, the integration could be done with ensemble methods. From the motif analysis, we can observe that **BENIN** is mostly affected by the cascade error. The reason for this failure needs further investigation.

Chapter 4

BENIN: Application to the HeLa Cell cycle

4.1 Introduction

In Chapter 3, we introduced a method that integrates any type of prior information with time-series expression data to infer the GRN: BENIN. In Chapter 3, we tested BENIN on simulated data from the DREAM4 and considered simulated knockout gene expression data and simulated genome-wide location. In this chapter, we propose applying BENIN on real data using a variety of real prior knowledge data. More specifically, we applied BENIN to human data. In particular, we used the HeLa cell line [203]. The HeLa cell line is an immortal human cancer cell line, which has allowed several medical research breakthroughs. Because of its immortality, HeLa cells have become the model cancer cell in cancer research [161].

We consider the human organism for several reasons. First of all, it is among the multicellular Eukaryotes of interest in nowadays researches. Hence, scientists have produced a variety of data to understand the complexity of human cell functioning. Furthermore, it has a complex regulatory network. Our point is to show that BENIN can infer complex GRN of higher organisms. Our goal is to infer the GRN that controls the cell cycle of the HeLa cell line. The cell cycle is a series of coordinated stages that allow cells to grow, replicate, and create new cells, permitting them to stay alive. It is an essential process by which the genetic material is transmitted through cells. This transmission should be accurate to prevent the transmission of genetic

mutations. Because of its importance for every living cells, the cell cycle is a highly controlled process. Some genes control the passage from one phase of the cycle to another. Other genes are responsible for holding the cell at specific points of the cell cycle. Any malfunctioning of this complex regulation may lead to the development of cancer. The regulation of the cell cycle happens at different levels. However, for our research, we restrict the regulation at the transcriptional level. Our main goal is to show that **BENIN** can infer not only interactions supported in the literature but also new high scoring interactions. We believe that reconstructing the GRN in cancers cell may help scientists identifying critical factors that may have led to a cancer state.

In this chapter, we propose integrating several prior knowledge data, ranging from TFBS, knockdown gene expression data, ChIP-seq data, or even functional annotation. We describe step by step how we transform the data into prior knowledge weights that are later integrated into **BENIN** (described in Chapter 3) to infer a list of regulatory links. The final GRN is obtained by applying a threshold τ on the inferred list of interactions. Our results demonstrate that the integration of diverse prior knowledge may improve **BENIN** performance, helps **BENIN** inferring interactions that are missed when we do not consider any prior knowledge.

We extended **BENIN** to include regulatory interaction from other closely related organisms. This integration will enrich the GRN inferred from time-series expression data with new regulatory links. We use orthology information transfer through sequence alignment to transfer known regulatory interactions from closely related model organisms into the studied model. The orthology mapping is based on the assumption that ortholog genes preserve their function. In our study, we consider the mouse as our model organism to study the GRN in human.

Mouse or *Mus Musculus* has several similarities to human in terms of genetics, physiology, and anatomy. These similarities make the mouse genomic research particularly insightful to gain knowledge on how human functions. Furthermore, the ease with which the mouse genome can be analyzed and manipulated has to lead to a production of a large variety of data available on different platforms. We use **eggno-mapper** [111] to get the 1:1 orthologous human genes into the mouse. Our results demonstrate that **BENIN** can infer several documented regulatory links and interactions supported by the literature and other potential regulatory interactions that necessitate further investigations.

Different strategies have been proposed to infer the human GRN in general and the GRN controlling the HeLa cell cycle. Ranging from computational methods that use statistical models to infer the GRN from mainly expression data [211, 247, 197, 77]; *in vivo* based methods that use wet-lab experiments to identify binding sites of TFs of interest [247, 197, 267, 267] and finally, hybrid methods that combine both strategies. The main limitation of these methods is that most of them use only a specific data type to infer the GRN. Some methods do not infer the whole GRN but rather a network anchored at the TF targeted by a specific experiment.

BENIN contributions to the inference of GRN controlling the HeLa cell cycle are the following:

- We propose integrating ChIP-seq, functional annotation, TFBS, and KD expression data with time-series expression data to infer the GRN controlling the HeLa cell cycle.
- We further integrate regulatory information from mouse through orthology information transfer, to confirm the inferred network from expression data and enrich the inferred network with potential interaction.
- **BENIN** can infer not only known interactions but also new potential regulatory interactions with high confidence that are supported to some extent with the literature that necessitates further investigation.

The chapter is organized as follows: in Section 4.2.1, we introduce the cell cycle and the cell cycle regulation paradigm. Section 4.2 gives the list of different strategies that have inferred the GRN controlling the human cell cycle in general and the HeLa cell cycle in particular. Section 4.3 describes our methodology to build the gold-standard for evaluating **BENIN** performances. Section 4.4 provides detail on data collection. Section 4.5 gives details on reverse-engineering the GRN controlling the HeLa cell cycle using time series combined with different prior knowledge information. We provide details on the different steps for transforming the prior knowledge information into prior weights. We further detail our methodology to transfer regulatory information from mouse using sequence similarity and discuss the data collection. In Section 4.6.1, we present the results of applying **BENIN** on Whitfield data [247], to

infer the GRN controlling the HeLa cell cycle. Finally, in Section 4.6.2, we discuss our findings.

4.2 Background

4.2.1 The Eukaryotic Cell Cycle

The cell cycle is an important phenomenon that occurs in all organisms to allow them to survive. It is the story of all living cells. It is an important sequence of stages by which a cell will go through to replicate its genetic material and divide to produce new cells. For the rest of this chapter, we will concentrate on the eukaryotic cell cycle. The eukaryotic cell cycle consists of two main phases:

- The mitosis or M-phase, which is the shortest phase of the cycle. In this phase, the cell will perform division to produce daughter cells with the same genetic material.
- The anaphase, which is the longest part of the cycle. It is in this phase where the cell will undergo most of its processes. The anaphase is divided into three discrete phases: one synthesis phase or S-phase in which the DNA is replicated. Two gap phases: the G1-phase (gap 1) that is the gap phase immediately after the mitosis and finally, the G2-phase (gap 2) in which the cell continues to grow, and the proteins are synthesized.

In summary, the eukaryotic cell cycle is divided into four phases: the M-phase, the G1-phase, the S-phase, and the G2-phase. Figure 16 gives an overview of the cell cycle in a eukaryotic cell.

The regulation of the cell cycle is essential for several reasons. First of all, it is important to control the cell division; otherwise, cells will undergo division infinitely, leading to cancer growth. Furthermore, regulation is important to ensure proper coordination and signal passage through the different cell cycle stages. Through the cell cycle, the cell considers several factors to decide whether it will progress from one stage to another. These factors are internal, e.g., DNA damage, or external, e.g., nutrient availability or cell size. These cues trigger the activities of key regulators at checkpoints. A checkpoint is a stage in the cell cycle where internal and external

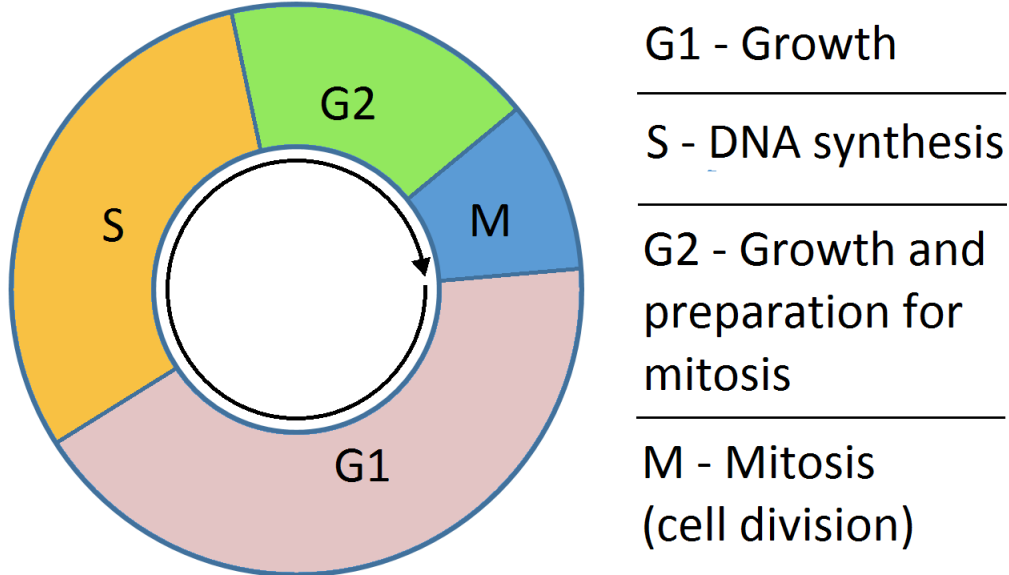


Figure 16: The Eukaryotic cell cycle

cues are checked to decide whether or not progression toward another step in the cycle should be halted. Hence, a checkpoint's general purpose is to ensure that all conditions are met before the cell proceeds to the next stage, hence ensuring that the complete genome is transmitted to daughter cells. For example, all the genome must be synthesized before moving to the mitosis phase. Otherwise, it will result in daughter cells having mutations that will be transmitted to following new cells. There are three main checkpoints:

- The G1-checkpoint: it happens during the transition from the G1 to the S-phase. It is at this step that major regulation occurs. At this stage, factors such as cell size, DNA integrity, nutrient resources are assessed.
- The G2-checkpoint: it happens at the transition from G2 to M-phase. At this checkpoint, the cell checks if the DNA is completely replicated and not damaged.
- The M-checkpoint or spindle checkpoint: it occurs during the mitosis. Here, the cell makes sure that sisters chromatid are properly attached to the spindle.

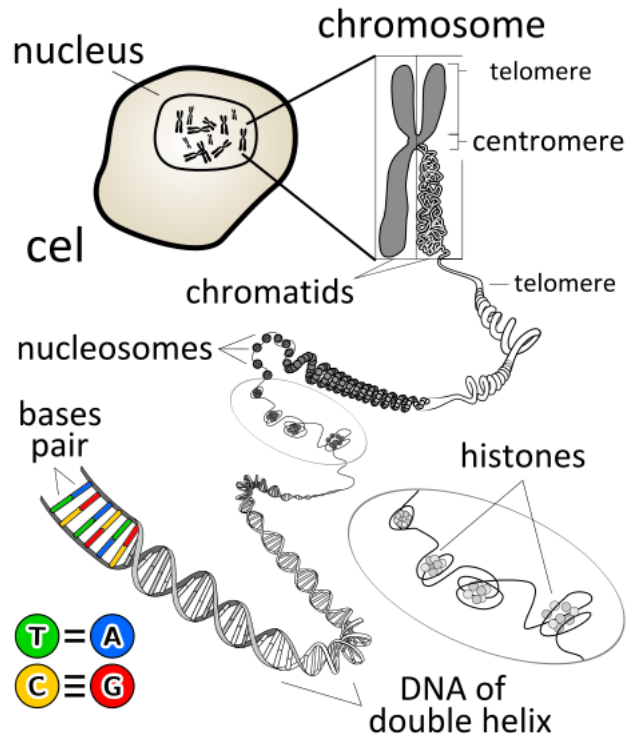


Figure 17: From nucleus to DNA sequence

The figure shows how the DNA is wrapped inside the cell of eukaryotic cells. The DNA length is far greater than the size of the nucleus in which it is stored inside the cell. Hence, the DNA needs to be condensed. The double helix of the DNA sequence is compacted around a protein called a histone, forming the nucleosome. Several nucleosomes are coiled together and stacked on top of each other, forming a chromatin fiber. The chromatin fiber is then looped. The chromatin fiber loops are compressed and folded to produce a fiber tightly coiled into the chromosome's chromatid. A chromosome is made up of two chromatids called sister chromatids.

The cell cycle regulation is controlled by different molecules, such as cyclin-dependent kinases (CDKs), which are enzymes that phosphorylate (add phosphate group) to other proteins for activating or repressing their activity. Note that CDKs are only activated when associated with cyclin. The other regulators are transcription factors. In this work, we focus on TFs. They can either repress or activate the activity of their target genes. In Table 13, we give a list of some of the TFs in the cell cycle. We considered almost all the reported TFs in our analysis, except TP53. We considered some members of the E2F family (E2F2, E2F3, E2F4, E2F6, and E2F7).

Furthermore, we consider only two members of the KLF family (KLF6 and KLF9) and two members of the STAT family (STAT4, STAT5B). The excluded TFs were not identified as HeLa cell cycle genes in the original work of Whitfield *et al* [247].

Table 13: Cell cycle Transcription Factors

HGNC symbol	DBD	Cell cycle Phase	Function	source
E2F-family				
(e.g. E2F1, E2F2, E2F4, E2F8 etc)	E2F	G2/M, G1/S	cell cycle progression, proliferation, DNA replication, DNA damage checkpoint DNA repair, chromatin assembly/condensation, Chromosome segregation, mitotic spindle checkpoint	[187, 7, 233, 50]
FOXM1	Forkhead	G1/S, G2/M	G1/S transition, mitotic progression, cell proliferation	[37, 45, 242]
TP53	p53		Control cell cycle progression, apoptosis, controls G2/M and G1 checkpoints, DNA damage response, cell growth	[128, 1, 263]
BRCA1	unknown	G2/M	DNA repair	[52, 257]
KLF-family (e.g. KLF9, KLF6, etc)	C2H2 ZF		cell proliferation, differentiation, development, and apoptosis.	[19]

Table 13 continued from previous page

HGNC symbol	DBD	Cell cycle Phase	Function	source
SP1	C2H2 ZF	G1	Cell differentiation, cell growth, apoptosis	[53, 87]
NF-Y family (e.g. NFYA, NFYB)	CBF/NF-Y		Apoptosis	[151, 89]
STAT-family (e.g. STAT1, STAT5, STAT4 etc)	STAT		differentiation, proliferation, cell survival, apoptosis, and angiogenesis	[32, 27, 253, 132]

The table shows the description of some important TFs that regulates the cell cycle in Eukaryotes, particularly in human. The first column is the official gene name or the TF family name. The second column provides the DNA-binding domain of the TF/family. The third column gives information about the function of the TF/family, and finally, in the fourth column, we provide the source for function description. We report in bold the keys TFs in the human cell cycle. The reported annotations are obtained from *in vivo* experiments (the 4th column reports the work related to the annotations). We considered almost all the reported TFs in our analysis, except TP53. We considered some members of the E2F family (E2F2, E2F3, E2F4, E2F6, and E2F7). Furthermore, we consider only two members of the KLF family (KLF6 and KLF9) and two members of the STAT family (STAT4, STAT5B). The excluded TFs were not identified as HeLa cell cycle genes in the original work of Whitfield *et al* [247].

4.2.2 HeLa Cell Line

The inference of the GRN that controls the human cell cycle general and the HeLa cell cycle, in particular, have been explored in the literature using different strategies. In this section, we will highlight the different strategies that have been proposed by the researchers to elucidate the cell cycle GRN. We will split them into three main

categories: Computationally-based methods, biologically based methods, and hybrid methods. Different Human cell lines are studied in the literature. For example, we can list the Fibroblast cell line, the Epstein-Barr virus (EBV) transformed lymphoblastoid cell line (LCL), or the U2OS cell line. In this study, we will mainly report works on the Human HeLa cell line.

Computationally-based methods mainly use mathematical models to infer the GRN controlling the Human cell cycle. In the literature, several models have been proposed for the GRN inference and tested on the Human cell cycle gene expression data. The main obstacle of methods in this category is the lacking of a gold-standard network against which the inferred network could be evaluated. Different methods have adopted a different strategy to evaluate their performance. Hence, Ali Shojae *et.al.* have proposed a lasso-based penalty method to infer causal interaction from time-series gene expression data [211]. They have tested their method on the HeLa cell cycle. They considered a subnetwork of nine genes for which the true regulatory network has been extracted from BioGRID. They used the Whitfield HeLa dataset [247], which original work consists of identifying genes that are periodically expressed in the HeLa cell cycle. To evaluate the inferred network, they considered the sub-network extracted from BioGRID by Sambo *et al.* [197]. They used statistical measures such as F1, or recall. As the BioGRID network is not complete, they considered edges that were absent in the gold-standard network as potential edges and compared these links to the literature to see potentially valid interactions that were not included in the BioGRID network. Other authors have used a different model to infer the GRN controlling the HeLa cell cycle. Fujita *et.al* have used the first order sparse autoregressive model to infer the GRN from time-series expression data [77]. They further evaluated the statistical significance of the inferred interactions and used the FDR to control for false positives. They used the Whitfield HeLa cell cycle dataset [247] and consider only a subset of 94 genes based on their association with cell cycle and tumor development. The inferred network was evaluated using literature. They were able to identify several interactions confirmed to be part of three pathways related to cell transformation and tumor progression, namely the P53, STAT3, and NFkB pathways. Other researchers have proposed an integrative framework that infers the GRN by incorporating diverse biological data. Zhang *et.al.* have proposed a modular network strategy that integrates information from time-series gene expression data,

protein-protein interactions (PPI), protein-DNA interactions and functional annotation [267]. The functional annotation was used to define the number of modules obtained with fuzzy clustering. The PPI and PDI data were used to extract network motifs. The idea is to assign TFs to at least one motif and then assign each module to a TF motif. The algorithm was tested the Whitfield HeLa cell cycle dataset [247]. They considered 846 genes that were demonstrated by Whitfield to be expressed in the cell cycle. They validated their result using functional enrichment and by comparing the inferred link with the literature. Zhengli *et.al.* have proposed integrating ODE with a dynamic Bayesian network to infer GRN from time-series expression data. They also validated their method on the Whitfield HeLa cell cycle dataset [247]. They considered 1009 clone IDs that were shown to be expressed during the cell cycle. Note that several clone IDs can correspond to the same unique gene. To validate their performance, they particularly focused on evaluating how the subnetwork routed at BRCA1 was inferred by their method. They evaluated the inferred subnetwork based on literature and functional coherence of the BRCA1 neighborhood. We provided above a non-exhaustive list of research works that have mainly considered time-series gene expression data to infer the GRN controlling the HeLa cell cycle.

In this category, we will also list methods that use statistical tests to infer GRN from perturbation expression data (KO or KD expression data). Here methods perform differential expression analysis as described in Section 2.2.1.1 (c.f. Chapter 2) to infer the TGs of a specifically screened TF. In [170] Oleaga *et.al.* have knocked-down SP1 to determine its TGs and particularly those involved in proliferation and cancer. The authors determined the TG from differential expression analysis. They used unpaired t-Test combined with Benjamini–Hochberg FDR correction for multiple testing. They also computed the fold change as the ratio of the expression value compared to the control condition. They obtained a large list of SP1 TGs that were validated using promoter scanning with known SP1 PWMs. Furthermore, they selected a subset of TGs for further validation using a ChiP experiment and other independent *in vivo* experiments.

Computational based methods present the following drawbacks:

- The perturbation experiments generally target one TF in a specific cell cycle. So on only the sub-network related to the screened TF can be inferred.

- Although studies have demonstrated the need to integrate diverse data to cope with noise in expression data, the dimensionality (data insufficiency), and to obtain more reliable results, most of the methods in this category consider only one type of data and do not integrate other omics data.
- Finally, generally, the inferred network is very limited.

In vivo based methods generally use chromatin immunoprecipitation (ChiP) experiments. The aim here is to find genomic loci bound by a specific TF of interest: the TFBS. A ChiP experiment can either be combined with DNA microarrays (ChiP-ChiP) or ultra-high-throughput sequencing (ChIP-seq). Ren *et.al* [187] have performed genome-wide location analysis of E2F TFBS using ChiP-ChiP experiment. The method has allowed us to identify cell cycle-regulated genes in mammalian cell lines. Hence, they identified previously unknown E2F TGs (target promoters) that were independently experimentally validated. Chen *et.al* [37] have used ChIP-seq experiment to elucidate genome-wide binding sites recognized by the forkhead TF FOXM1. They identified a group of cell cycle genes bound by FOXM1. Gordon *et.al* [190] have used the same strategy for identifying regions within the genome of the HeLa cell line bound by the STAT1 transcription factor. Nowadays, ChIP-seq experiments have become an indispensable and preferred in vivo method to detect DNA interaction between a gene a TF of interest, because of its signal to noise ratio. ChIP-seq data are deposited in database such as ENCODE or *Chip-Atlas*. Some of these databases offer the possibility to predict target genes bound by a given TF. The main limitation of *in vivo* methods is that experiments are generally restricted to one or a few TFs of interest and specific cell lines. Hence only part of the GRN can be inferred with these methods. Furthermore, they infer only physical interactions. However, physical binding does not necessarily imply functional association.

Hybrid methods generally combine perturbation experiments (gene knockout or knockdown) with ChiP experiments. Generally, target genes inferred from ChiP experiments are validated through perturbation experiments targeting the TF screened in the ChiP experiments.

4.3 Building a gold-standard

An important step in the GRN is the evaluation of the reconstructed network. Different strategies have been proposed. They are experimentally based or *in insilico* based methods. Experimentally-based methods consist, for example, on performing perturbation experiments to validate the finding. For our research, we are focusing on an *in silico* evaluation. The strategy here is to define the GRN inference as a binary classification problem, which consists in predicting an edge as being present or absent in the final network. Then one uses statistical methods as defined in Section 2.4.1 to evaluate the inferred network. To achieve this, one needs to have a defined gold-standard network with positive and negative interactions.

One difficulty in evaluating computational methods for the inference the GRN is the lack of a proper and manually curated list of regulatory interactions that will serve as the truth. Some efforts have been put together to define databases storing regulatory links for well-studied organisms such as *saccharomyces cerevisiae* with the `yeasttract` database [166]. If we compare the number of existing databases of regulatory interactions with existing organisms, we can observe a significant discrepancy. Another difficulty is the lack of curated nonregulatory links. Thus many existing curated databases consist only in positive links. It is difficult to define a negative link as our knowledge of the transcriptional regulation is very limited. The non-existence of interaction in the literature does not mean that the two genes are not interacting together.

One challenge of applying BENIN to human data is the construction of our “gold-standard” network for performance evaluation. Unfortunately, no repository provides a complete, curated gold-standard list of human regulatory interactions. Nevertheless, for our study, we use the “gold-standard” networks from Garcia Alonso work [78] that we combined with interaction from the `HumanBase` database [86, 135, 269, 270]

We are conscious that the “gold-standard” network is not complete, but it represents, to the best of our knowledge, the human GRN. As preliminary results, we did not consider the possibility that the network may differ for each cell type. Instead, we consider the regulatory network to be the same for all the cell types.

4.3.1 Material

We collected two gold-standard networks from Garcia’s work [78]. One for cancer cell line and the other for normal cell lines. The networks are obtained from the supporting tables S3 and S4 of [78] (GarciaAlonso_supplemental_table_S3_regulons Normal.xlsx and GarciaAlonso_supplemental_table_S4_regulons Cancer). It gathers signed regulatory interactions. However, as we are not interested in the type of regulatory interaction, we ignored the sign of the interactions. Their “gold-standard” network combines information from diverse curated databases. More precisely it gathers regulatory interactions from 13 databases: HTRidb [24], Oreganno [143], KEGG [124, 126, 125], Fantom4, TRRUST [93], reviews, TFact [65], IntAct [171], NPISRegulomeDB, TRRD [133], TRED [268], PAZAR [181], TFe [264].

We also collected regulatory interaction networks for 132 cell lines from the HumanBase database <https://hb.flatironinstitute.org/download>.

4.3.2 Method

We build our “gold-standard” network by merging the different networks from Garcia *et al* and the 132 networks from HumanBase database. Our challenge here is to define the negative example (i.e., absence interaction). It is a very tricky and challenging task since our knowledge of the human regulatory network is limited. Furthermore, as specified above, existing databases that store regulatory interaction provide only positive links. We define our negative interactions from the 132 HumanBase database networks do. We follow the idea of Huttenhower *et al* [114]. They have proposed to used as negative examples gene pairs not co-annotated to any terms in a set of 433 Gene Ontology (GO) [42] biological processes terms selected by their experts. These negatives interactions are included in the 132 HumanBase database networks. Note that we considered all of their set interactions (both positive and negative).

From the two networks collected from Garcia *et al* paper, we considered only literature-curated interactions and coexpression based interactions. We made sure that they do not come from ChIP-seq experiments, and they are not obtained from TFBS motif analysis. We want to avoid any bias in the performances since we consider both ChIP-seq and TFBS as prior knowledge to infer the GRN controlling the HeLa cell cycle. We merge the two networks using a simple merge function. We merged

based on the TF- TG combination. Note that this merge takes care of the duplicate edges. The obtained links represent part of our positive interactions.

We concatenate the 132 networks on the command line using the “cat” command. We then proceed to analyze and remove duplicated edges. Table 34 gives the result of our analysis of the edges repetition. We consider two cases to remove the duplicates.

- Case 1: An edge is marked absent in a certain cell line but present in at least one of the other cell lines. In this case, the positive occurrence is kept in the merged network, and the other occurrences are discarded.
- Case 2: All the occurrences are positive links. In this case, one of the occurrences is kept in the merged network, and the other occurrences are discarded.
- Case 3: All the occurrences are negative links. In this case, one of the occurrences is kept in the merged network, and the other occurrences are discarded.

In the last, we merge the two big networks (from Garcia and from the HumanBase) to build our final “gold-standard” network. Here, we also need to remove the duplicates edges. There are different cases to consider:

- Case 1: An edge is marked absent in the HumanBase’s network but present in Garcia’s network. In this case, the edge is added as a positive link in the final gold-standard network.
- Case 2: An edge is marked present in the HumanBase’s network and Garcia’s network. In this case, one occurrence of the edge is added to the final regulatory network.

In any other case, edges that belong either to HumanBase’s network or to Garcia’s network are directly added to the final “gold-standard” network.

Algorithm 2 summarizes our methodology to build our “gold-standard” from existing “gold-standard” networks:

Algorithm 2 Steps for Building the “gold-standard network”

- 1: Collect the 132 cell line networks available from HumanBase <https://hb.flatironinstitute.org/download>.
- 2: Concatenate the 132 networks on the command line using the cat command

- 3: Analyze the network obtained in Step 2 to remove duplicated edges. Table 34 gives the result of our analysis of the edges repetition.
- 4: Remove duplicated edges. There are different cases to deal with repeated edges:
 - Case 1: An edge is marked absent in a cell line but present in at least one of the other cell lines. In this case, the positive occurrence is kept in the merged network, and the other occurrences are discarded.
 - Case 2: All the occurrences are positive links. One of the occurrences is kept in the merged network, and the other occurrences are discarded.
 - Case 3: All the occurrences are negative links. One of the occurrences is kept in the merged network, and the other occurrences are discarded.
- 5: The `HumanBase` database stores the genes with their Entrez ID. We converted the Entrez ID to official gene names using the human genome-wide annotation R package `org.Hs.eg.db` [35].
- 6: Collect the two networks from Garcia’s work [78]. The networks are obtained from the supporting tables S3 and S4 of [78] (`GarciaAlonso_supplemental_table.S3_regulonsNormal.xlsx` and `GarciaAlonso_supplemental_table.S4_regulons Cancer`)
- 7: For each network, we subset the edges and consider only those that are from curated databases. We make sure that they do not come from ChIP-seq experiments, and they are not obtained from TFBS motif analysis. We want to avoid any bias in the performances since we consider both ChIP-seq and TFBS as prior knowledge to infer the GRN controlling the HeLa cell cycle.
- 8: We merge the two networks using a simple merge function. We merged based on the TF- TG combination. Note that this merge takes care of the duplicate edges. For each edge in the merged network, Table 35 gives the number of times it appeared before removing the duplicates.
- 9: We merge the network from Step 4 with the network obtained in Step 7. Here we need to deal with the repeated edges. Let $gs1$ the network obtained Step 4 and $gs2$ the network obtained from Step 7. There are different cases to take into consideration:
 - Case 1: An edge is marked absent in $gs1$ but present in $gs2$. In this case, the edge is added as a positive link in the final gold-standard network.

- Case 2: An edge is marked present in *gs1* and in *gs2*. One occurrence of the edge is added to the final regulatory network.

In any other case, edges that belong either to *gs1* or *gs2* are directly added to the final “gold-standard” network. The list edges in the final “gold-standard network” are depicted in Table 36 and Table 37.

4.3.3 Results

Table 35 gives the list of edges that were duplicated after merging the two Garcia networks. The table also reports the number of times each duplicated edges appeared before removing the duplicates in the merged network (from the two Garcia’s networks). In Table 34, we report the result of our analysis of the edges repetition after concatenating the 132 networks from the **HumanBase** database.

We ended up with a gold-standard network whose characteristics are summarized in Table 14. The detailed list edges in the final “gold-standard network” are depicted in Table 36 and Table 37. Note that not all of our considered genes are part of our “gold-standard” GRN. From Table 14, we observe that we are missing information for 43 genes, among which 3 TFs: ZNF207 GTF2B and BRCA1.

Table 14: Characteristics of our Human “gold-standard” network

# Regulatory links	#TFs	#Genes	# Positive links	# Negative links
3333	39	585	1463	1870

The table provides the characteristics of our gold-standard network in term of the number of regulatory links (see column 1), the number of TFs (see column 2), the total number of genes (column 3), the number of positive links (column 4) and the number of nonlinks (column 5).

Table 15: Missing transcription factors in our “gold-standard network”

#TF	TFs Name
15	MZF1, MNT, DMTF1, CIC, ZNF414, ZNF587, HMG20B, ZNF521 ZNF207, TSC22D1, ZNF281, ZBTB7A, ZNF217, ZBED5, GTF2B

4.4 Material

4.4.1 Data

In this section we will present in details how we collected the diverse data use for inferring the GRN controlling the HeLa cell cycle.

4.4.1.1 HeLa Time-Series Expression Data

The HeLa cell cycle time-series gene expression data were generated by Whitfield *et al* [247]. We downloaded the Whitfield HeLa cell cycle time-series gene expression from <http://genome-www.stanford.edu/Human-CellCycle/HeLa/>. It is a well-known time-series gene expression dataset. It consists of five different time-series experiments with different synchronization methods (double thymine block, Thymidine-nocodazole block, or mitotic shake-off). These synchronization methods arrest the cell at either the S-phase or the M-phase (see section Materials and Methods on <http://genome-www.stanford.edu/Human-CellCycle/HeLa/>). We used only a part of the microarray dataset for our experiments: we considered the third time-series named “Thy-Thy 3” by Whitfield *et al*. In the “Thy-Thy 3” experiment, a double thymine block is used to arrests cells at the G1/S boundary. It is the most extended time series of the experiment. Gene expression values were measured at 1h intervals from 0 to 46h. Note that there is an extra time point at t=0, where the expression values are the average of the same measurement obtained from two biological replicates. In total, we considered 48 times points. In their work, Whitfield *et al* [247] identified a list of 1132 IMAGE clones ID that are periodically expressed during the cell cycle. From the annotated IMAGE clone IDs, 777 out of the 1132 IDs have a Gene IDs. They correspond to 632 different genes: 82.3% mapped to one

unique clone ID, 14.1% mapped to two clone IDs, and the rest (3.6%) are mapped to up to six different clone IDs. Out of these 632 different genes, we excluded four genes because they do not have a GO annotation. We summarized the duplicated genes (genes represented by several probes IDs) by averaging their expression profile. In summary, we considered a list of 628 unique genes that are periodically expressed in the HeLa cell cycle. We imputed the missed values in the dataset using K -nearest neighbor (KNN). We set the number of neighbors K to 12, as suggested by Whitfield *et al* [247].

In summary we proceed as follow to collect our time-series gene expression data

Step 1: We collected the raw time-series expression matrix from <http://genome-www.stanford.edu/Human-CellCycle/HeLa/>

Step 2: Input missing value using knn with in the following R code. We set $k=12$. Let *exprdata* the original expression data matrix and *exprdataimputed* the imputed expression data matrix.

```
exprdataimputed<-knnImputation(exprdata , k=12)
```

Step 3: We remove probes that map to the same gene by averaging their expression profile. See Table 31 for the complete list of considered genes.

Step 4: Save the imputed matrix for later use.

4.4.1.2 ChIP-seq data

We downloaded the peak files on the **UCSC Genome Browser website**: <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/>. The peak files report regions on the genome that have been enriched with aligned reads as a consequence of performing a ChIP-sequencing experiment. These areas are reported in terms of genomic coordinates. For each region, the file also reports a measurement of the overall enrichment and the statistical significance of this enrichment (with p-value and q-value). The ENCODE Analysis Working Group (AWG) generated the files using a uniform processing pipeline. The whole dataset covers 91 cell lines with various treatments. We restricted the files to those reporting analysis on the HeLa cell line. We further restricted them to the peak files of the TFs expressed in

our HeLa cell cycle expression dataset. Out of the 54 TFs expressed in our cell cycle time-series expression data, we got the peak files for only eight TFs: *BRCA1*, *CTCF*, *E2F1*, *NFYA*, *NFYB*, *STAT1*, *TFAP2A*, and *ZNF143* (c.f. Table 28).

In summary we proceed as follow for collecting our ChIP-seq data:

Step 1: Collect the peak files from **UCSC Genome Browser website**: <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/>

Step 2: Subset the collected file to those concerning the HeLa cell line

Step 2: Subset the list of files to those concerning the TF expressed in the cell cycle (c.f Table 32).

4.4.1.3 Knockdown Gene Expression data

The raw data were downloaded from **GEO-NCBI**. We also downloaded analyzed data from *knockTF* [70] (<http://www.licpathway.net/KnockTF/>), which gathers data from **GEO-NCBI**, **ENCODE** or others databases. Note that we considered both RNA-seq and microarray gene expression data. We downloaded data from different cell types because we do not have enough data for the HeLa cell type. We assume that TFs may bind to the same genes in different cells, depending on the biological process. We gather KD data for approximately 20 TFs (c.f. Table 29 and Table 33).

In summary we used Algorithm 3 to collect our KD expression datasets.

4.4.1.4 Transcription Factors and Binding Sites

We get the set of PWM representing the TFBS from **CisBP** [245] database <http://cisbp.cabr.utoronto.ca> an online database of the TFs and their PWMs. Note that we restricted the potential TFs to those which have at least one PMW in the database. **CisBP** gathers the matrices from diverse database such as **JASPAR**. We ended up with data for 41 unique TFs (c.f. Table 30).

We downloaded the promoters sequences of our genes from **UCSC Genome Browser website**, under the table browser <https://genome.ucsc.edu/cgi-bin/hgTables>.

Algorithm 3 Steps for collecting the KD gene expression data

- 1: Query the GEO database accessible from <https://www.ncbi.nlm.nih.gov/gds> to get the the list of HeLa knockdown expression dataset. Enter the following query on the search bar:
 - *(HeLa knockdown) AND "Homo sapiens"[porgn:__txid9606]*
 - *(knock down HeLa) AND "Homo sapiens"[porgn:__txid9606]*
 - 2: Scan the list of results to collect GEO datasets with at least three biological replicate samples. This minimum number of replicates is necessary for our later differential expression analysis. Refer to Table 33 for the list of collected datasets.
 - 3: Collect the whole dataset from knockTF: <http://www.licpathway.net/KnockTF/download.php>. We did not restricted the cell line. But rather restricted the TFs to those expressed in the HeLa cell cycle. Table 29 reports the list of TFs and their corresponding KD dataset IDs.
-

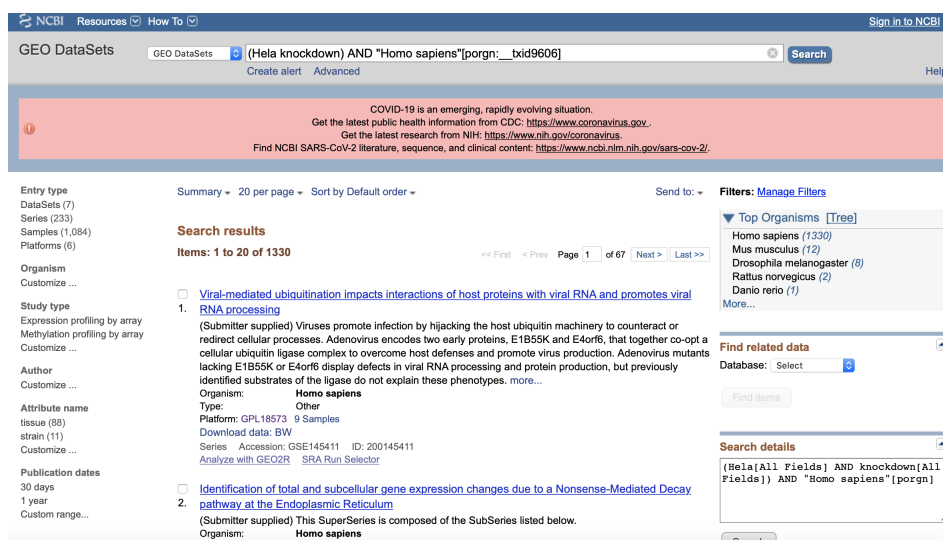


Figure 18: Steps for retrieving knockdown data

The figure shows a snapshot of the query we performed to retrieve the KD expression data.

The nucleotide sequences are 1000bp long as recommended on the FIMO web page. We considered the version Genome Reference Consortium Human Build 38 (hg38/GRCh38) Algorithm 4 summarizes the promoter data collection as well the TFBS collection.

Algorithm 4 Steps for collecting the TFBS and promoter regions

- 1: Reach the **UCSC** table browser <https://genome.ucsc.edu/cgi-bin/hgTables>
 - 2: Set the clade to **mammal**, the genome to **human**, the assembly to **Dec 2013 (GRch38/hg38)**.
 - 3: Choose the group **Genes and Predictions** and set the track to **GENCODE v32**.
 - 4: Choose the table **knowGene**. Set the region to **genome**. We let the other parameters to their default values see (<https://genome.ucsc.edu/cgi-bin/hgTables>).
 - 5: Paste the list genes identifiers.
 - 6: Choose the output format and choose the output file name if needed.
 - 7: Select **genomic** as sequence type
 - 8: Select **Promoter/Upstream** by **1000** bases for the retrieval region options.
 - 9: Set the formatting to **all lower case**.
 - 10: On the CisBP database, select the bulk download option (<http://cisbp.cabr.u-toronto.ca/bulk.php>) to collect the whole database for an organism of interest. Set the organism to *Homo sapiens*
-

4.4.1.5 Proteins Sequences

We downloaded the proteins sequences of the cell cycle genes from UniProt [43]. We considered only manually annotated sequences from Swiss-Prot. Out of the 628 considered genes, we got sequences for 624 genes. Among the four missing genes, two mapped to one gene already included (HIST1H4C, HIST1H4B, HIST1H4E), and the two other (*SETD8P1*, *LINC00339*) do not have sequences in UniProt and do not have orthologous genes in EggNOG-DB [111]. As UniProt contains redundant sequences, we downloaded a total of 636 sequences. To remove redundant sequence, we used CD-hit [76, 145], which is a fast incremental clustering algorithm that uses heuristic to cluster similar sequences. Sequences are compared based on k-mers. We set the similarity threshold to 70% to remove sequences that are 70% similar and allow a maximum redundancy of 1. It helps to remove ten sequences. We ended up

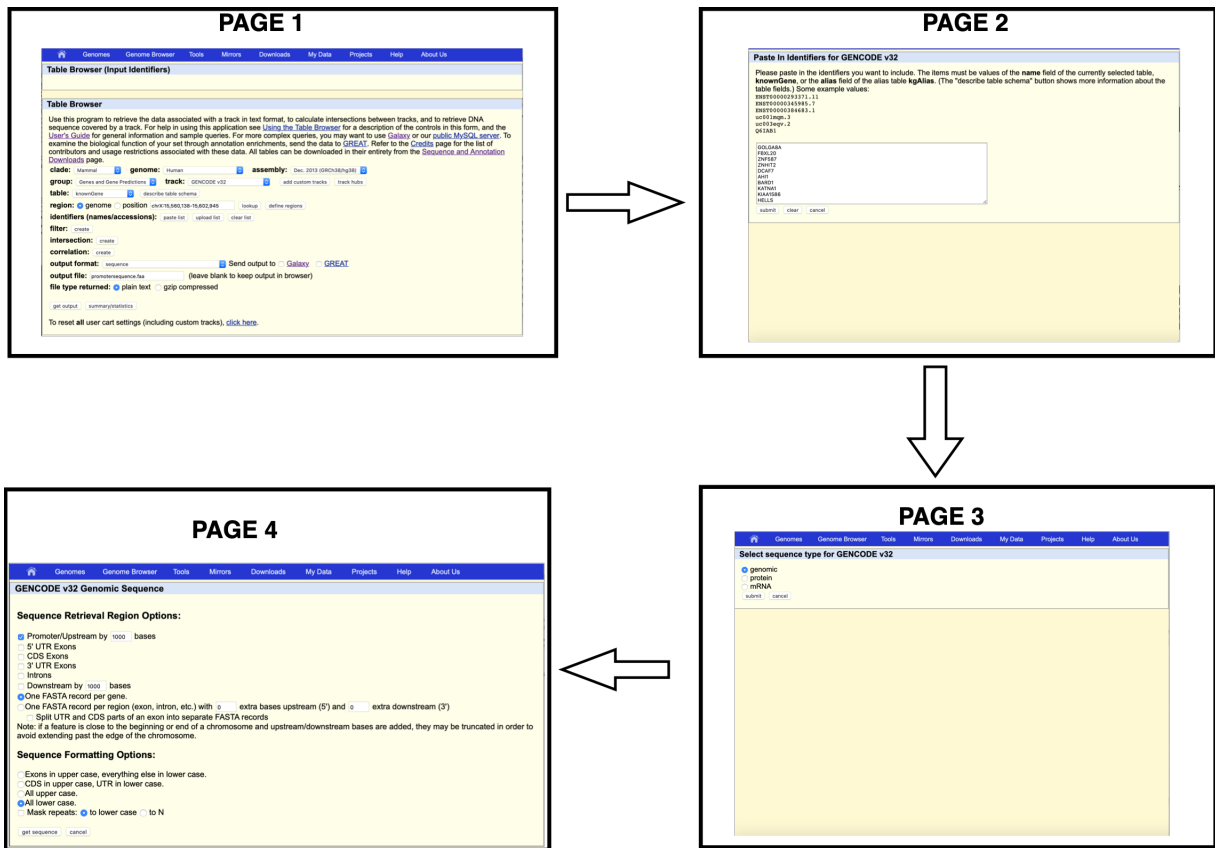


Figure 19: Steps for collecting promoter sequences

with 626 sequences. Algorithm 5 summarizes the steps for collecting and cleaning the proteins sequences (Figure 20).

4.4.1.6 Model Organism Regulatory Links

For our orthology-based regulatory network inference, we consider the mouse (*Mus musculus*) as our model organism. We downloaded the set of curated regulatory interactions from diverse curated databases. Our first database is STRINGDB [224], a database that stores known and predicted Protein-Protein interactions for more than 5000 organisms. The interactions can be physical or functional. The interactions are obtained from different sources such as literature, knowledge databases, or High-throughput lab experiments. The database attributes a score for each interaction. This score is computed as a combination of the probabilities of interaction from different evidence. The score is multiplied by 1000. A score of 800 corresponds

Algorithm 5 Collecting and cleaning proteins sequences

- 1: On UniProt website, select the “Retrieve/ID Mapping” tool <https://www.uniprot.org/uploadlists/>. Upload the list of cell cycles genes that you need to be converted.
- 2: We chose the ID we want to map (gene name) to the UniProt ID. Then we chose the organism, which is human in our case.
- 3: Filter the results to **Swiss-Prot** sequences, which are manually curated. Download the selected sequences.
- 4: Cluster duplicates sequence using `cd-hit` as follow:

```
$ ./cd-hit -i uniprot-human-cellcycle.fasta  
-o cluster_proteins_human_seq.clstr -c 0.9  
-T 2 -t 1 -sf 1 -sc 1
```

Where `cluster_proteins_human_seq.clstr` is the output file containing the sequences cluster. Note that similar sequences are clustered together. The parameter `-c` controls sequence identity threshold; `-T` controls the number of threads for parallel computing; `-t` controls tolerance for sequences redundancy; the parameter `-sf` allows to the obtained clusters regarding their size; finally, `-sc` output the sequences by decreasing cluster size.

- 5: Collect the list of nonredundant sequences for next steps analysis.
-

STEP1

Retrieve/ID mapping

How to use Retrieve/ID mapping tool

1. Enter identifiers, separated by spaces or new lines, into the form field, for example: P13943, P62558, A8UJ, HUMAN, P7FU, ECOLI
2. If you need to convert to another identifier type, select the source and target type from the dropdown menus.
3. Click the Submit button.

1. Provide your identifiers

2. Select options

From: To: Operation:

Clear Submit

STEP2

UniProtKB results

How to use Retrieve/ID mapping tool

623 out of 628 Gene name identifiers were successfully mapped to 636 UniProtKB IDs in the table below.

Click here to download the 5 unmapped identifiers.

Filter by:

Download selected (0)	Download all (636)	Columns	Sort	Filter	Export	Print
<input type="checkbox"/> GOLG8A	<input type="checkbox"/> GOLG8A	Family A member	GOLG8A	631	Human	631
<input type="checkbox"/> FBXL20	<input type="checkbox"/> FBXL20	Link-repeat protein	FBXL20	436	Human	436
<input type="checkbox"/> ZNF587	<input type="checkbox"/> ZNF587	Zinc finger protein	ZNF587	175	Human	175
<input type="checkbox"/> ZNF12	<input type="checkbox"/> ZNF12	Zinc finger HET domain-containing	ZNF12	403	Human	403
<input type="checkbox"/> DCAF7	<input type="checkbox"/> DCAF7	DDR1- and CSL4-associated	DCAF7	342	Human	342

STEP5

```

>Cluster 0
0 130aa, >sp|P0C808|H2A1_HUMA... at 91.54%
1 130aa, >sp|O717L|H2A3_HUMA... at 91.54%
2 143aa, >sp|P1E104|H2AX_HUMA... *
3 130aa, >sp|O93077|H2A1C_HUM... at 92.31%
>Cluster 1
0 445aa, >sp|P68371|TBB4A_HUM... *
1 445aa, >sp|O13B85|TBB2A_HUM... at 96.48%
>Cluster 2
0 453aa, >sp|O71U3G|TBA1A_HUM... *
1 450aa, >sp|P0DPH7|TBA3C_HUM... at 97.78%
>Cluster 3
0 356aa, >sp|O9H301|BORG4_HUM... *
>Cluster 4
0 356aa, >sp|O14558|KPRA_HUMA... *
>Cluster 5
0 356aa, >sp|O9NU0G|OTULL_HUM... *
>Cluster 6
0 364aa, >sp|O15906|VP572_HUM... *
>Cluster 7
0 365aa, >sp|O15264|MK13_HUMA... *
>Cluster 8
0 363aa, >sp|P35249|RFCA_HUMA... *
>Cluster 9
0 360aa, >sp|O95149|SPN1_HUMA... *
>Cluster 10
0 360aa, >sp|O96HE9|PRR11_HUM... *
>Cluster 11
0 351aa, >sp|O98586|ZPBP1_HUM... *
>Cluster 12

```

STEP4

```

(base) MacBook-Pro:cd-hit-v4.8.1-2019-0228 Stephanie@ /cd-hit -i uniprot-human-celcycle.fasta -o cluster_proteins_human_seq-new.clstr -c 0.9 -t 2 -l 1 -sf 1 -sc 1
Program: CD-HIT, V4.8.1 (+OpenMP), May 21 2020, 17:06:17
Command: /cd-hit -i uniprot-human-celcycle.fasta -o cluster_proteins_human_seq-new.clstr -c 0.9 -t 2 -l 1 -sf 1 -sc 1
Started: Wed Jul 15 20:57:46 2020
Output
-----
total seq: 636
longest and shortest : 4678 and 79
Total letters: 484687
Sequences have been sorted

Approximated minimal memory consumption:
Sequence      : 0M
Buffer        : 2 X 11M = 22M
Table         : 2 X 60M = 120M
Miscellaneous : 0M
Total         : 154M

Table limit with the given memory limit:
Max number of representatives: 78373
Max number of word counting entries: 80725583

# comparing sequences from      0 to 636
----- new table with      631 representatives

636 finished      631 clusters

Approximated maximum memory consumption: 157M
writing new database
writing clustering information
program completed !

Total CPU time 0.28
(base) MacBook-Pro:cd-hit-v4.8.1-2019-0228 Stephanie@

```

STEP6

```

>>sp|A7E2F4|GOLG8A_HUMAN Golg1n subfamily A member 8A OS=Homo sapiens OX=9606 GN=GOLG8A PE=2 SV=3
MLPQGEERKTEGSDGRTSPCAVSATLKLEVGSGSRCDPAGPAGNLLPQRLG
APLPAETHTPTPNDRSLVLSFKSSASSLHARSPCEQAAVNLNRSIKTSLRINDTI
KSLKDDKQVHEHLEKANKKAKKRELIGQURLATEKINLTLVYKKSLSYFFE
EESKLAGRLQKSSQRIEGLWSLCAVATOKKPKPGRSRSKALKRLEQSTREIILL
KQNTLKEKSLKLEVLQERDVAEDKAGEAAGHQRHRRPQVEVTLKEEKQDTRHEEL
ERSLRSLQDMATLPPAPASSEVELODKLELVAAGLQGVENQCCSLINRQK
ERLREDEERLEQDEERLEREKRLDQLEPQSDLEELHENSALQLEQVKELEKLGQ
MNETLTSAREPEANPKASQSGESGLRMLLEKADLREKVELLEGFQYRERKQK
VHRLLEPQDSAKASPGGGHAGPQGGEGEAGAAGAGGVAACGSYEGHOKFLAA
RNPAAEPPSAPAPGLGADNDIGLCEASLNSVEPAGGAREGSSDNPATQPVVPL
GENQHQHPLGSLNCVPCFWMLPRRR
>>sp|O96IG2|FBXL20_HUMAN F-box/LRR-repeat protein 20 OS=Homo sapiens OX=9606 GN=FBXL20 PE=1 SV=2
RRRDVNGVTKSRFEMFNSGDEAVINMKLPKELLRIEFSFLDVTLCRAQVSRANVIAL
DGSNWDIDLFDYRQDIEGRVNESSKCGFLKLSLRGLDIGNALRFFANQWIE
VNLNMGCTKTDTATCSLTKFCKLRHLDLASCTSLTINMSKALSEGCPLEQLINSDQ
OVRKDGIALVGGCGKALFLKGLCTOLEDALRYIGHCFELVTLNLTLOTLOLDEGLT
TICRGGKRLSLCSGKSLTALNALGMPRLLELEKRCQLTQVGTTLRQKHE
LEKMLDEEVQITDSTLITQLS1HCPRLQVLSLHCELITDQGRPLGNGACHQDLEVE
LQNCPLITDASLHLKAGLSLELTYDQCTTRAGKRLHLLPHEVNYAYAPVPPV
SVGGSQRQRCRCIIL
>>sp|O965D5|ZNF587_HUMAN Zinc finger protein 587 OS=Homo sapiens OX=9606 GN=ZNF587 PE=1 SV=1
MAKAVRPPFTQOYTFEDVAVNFSDKEKLLSEAKNLYVDVNLLENALISSLCGCG
KDEAPKQRISVQRESQRTFRAGVSPKAKHRCMGLILEVFFADHQETRHQKLN
RSGACQNLDDTYLHQRQKQKGEKPKKSKVRESFVKKKLNKSEKPPVRFQGVLL
PSSGLCQEAIVKTSDETHMPPFQEGKTNVSCQRKTFKSTHVSIPKHLFTFDGCV
VCSGKASFSRYFSNHRDHTAKAPRQKCGECSYSRKSLQHRWRVGTGTATPCEE
CKSFQKGLSLSHLYTGEQPYTEKCGKSGKQKLNLDHQDGTGERAVYCGECSK

```

Figure 20: Steps for collecting protein sequences

to 0.8 probability. For each edge, the score is the probability that the interaction exists. We downloaded the set interactions for each organism. We considered only binding and expression interactions. We set a threshold to 500 on the scores for selecting interactions. To further ensure that we extract only regulatory interactions, we subset binding interactions for which the direction is known.

We then consider two other databases to get our mouse regulatory interactions: **TRRUST** [93], **RegNetwork** [149]. They both store regulatory interactions for human and mouse. Table 16 summarizes the characteristics of the regulatory networks obtained from all the databases.

Algorithm 6 summarizes our mouse regulatory network construction.

Algorithm 6 Building the mouse Regulatory network

- 1: On the **TRRUST** database website, we downloaded the regulatory interactions for the mouse organism <https://www.grnpedia.org/trrust/downloadnetwork.php>.
 - 2: On the **RegNetwork** website <http://regnetworkweb.org/download.jsp> download the regulatory directions.
 - 3: We concatenated the networks from Step 1 and Step 2. Then proceed to analyze the duplicated edges (c.f Table 38). It is important to highlight that the edges downloaded are only positive links.
 - 4: On the **STRINGDB** website <https://string-db.org/> download the mouse proteins actions. Subset the list of proteins links to those for which directionality is mentioned and that have a score of at least 500(0.5).
 - 5: Let $gs1$ the network obtained in Step 4 and $gs2$ the network collected in Step 5. In this step, the aim is merging $gs1$ and $gs2$. We first mapped the genes' name to their corresponding **ENSEMBL** protein IDs. In fact, genes are accessed with their **ENSEMBL** protein IDs **STRINGDB**. We used the R library **biomaRt** version 2.38 [57, 56]. We then merged $gs1$ and $gs2$. Note that a gene can map to several **ENSEMBL** protein IDs. The list of replicated edges can be found in Table 39.
 - 6: Remove the duplicated edges by choosing one occurrence per repeated edge.
-

Table 16: Mouse gene regulatory network

Databases	# TFs	# Genes	# Interactions
TRRUST	827	2456	7057
RegNetwork	1902	3805	323636
STRINGDB	N/A	1136	2636

The table reports details about mouse regulatory information collected from the TRRUST, the RegNetwork and the STRINGDB. For each data we report the total number of collected interaction (3rd column), the number of genes covered by the interactions (2nd column) and finally if applicable the number of TFs (1st column).

4.4.2 Evaluation

As a primary evaluation, we considered AUPR and AUROC scores to evaluate the performance of BENIN for inferring known interactions in the human cell cycle GRN. Note that we removed inferred interaction from the TFs that are not part of our “gold-standard” network for the evaluation. We further removed self-interactions. We also evaluated the algorithm based on the functional annotation of groups of coregulated genes. The point is to evaluate the coherence between a transcription factor and the set of its inferred target genes. We also performed a literature review to assess the inferred links and potential new interactions.

4.5 Method

4.5.1 Integrating Prior Knowledge

We applied BENIN to infer the GRN controlling the cell cycle of the HeLa human cancer cell line. It is the oldest and the most extensively used human cell line for scientific researches. The line is derived from cervical cancer cells. We considered the gene expression data from Whitfield *et al* [247] work, which is made up of five time-series experiments. Their experiment’s objective was to identify genes that are periodically expressed in the human HeLa cell cycle. Our goal is to consider the

genes expressed in the cell cycle to decipher the transcriptional regulatory network controlling the cell cycle. We combined four types of prior knowledge data with time-series expression data: functional annotation, ChIP-seq data, TFBS, and knockdown (KD) gene expression. In this section, we will describe in detail how different prior knowledge data are integrated into BENIN with time-series expression data for the GRN inference.

4.5.1.1 Integrating Functional Annotation

The first data we considered as prior data is the functional annotation from the Gene Ontology. We considered the ‘‘Biological process’’ (BP) annotation. Our idea is that, if a TF r_j and gene g_i participate in the same BP, then it is most likely that r_j controls the expression of g_i . Hence we want to check for each TF-TG pair how similar is their BP annotation profile. Thus, for each pair (TF-TG), we compute the semantic similarities of their lists of BP GO terms with the R package `GoSemSim` [259]. Different measures are available in the package to compute the semantic similarity among GO terms, set of GO terms, and among genes. Here we considered the Relevance (Rel) method to compute the similarity between term. The method was introduced by Schlicker [204] and defines the similarity as follows:

$$sim_{Rel}(t1, t2) = \frac{2IC(MICA)(1 - p(MICA))}{IC(t1) + IC(t2)} \quad (38)$$

Where IC stands for information content, MICA stands for most informative common ancestor. We then chose Best Match Average technique to compute the semantic similarity between genes. It is defined as following: let gene g_1 annotated by GO terms sets $GO1 = \{go_{11}, go_{12} \dots, go_{1m}\}$ and g_2 annotated by $GO2 = \{go_{21}, go_{22} \dots, go_{2n}\}$, we have :

$$sim_{BMA}(g_1, g_2) = \frac{\sum_{i=1}^m \max_{1 \leq j \leq n} sim(go_{1i}, go_{2j}) + \sum_{j=1}^n \max_{1 \leq i \leq m} sim(go_{1i}, go_{2j})}{m + n} \quad (39)$$

After computing the similarity scores from the GO annotation, we stored them into a matrix $\mathbf{Sm} = \{sim_{BMA}(r_j, g_i)\}$. Afterwards, we transformed the similarities into weights $\mathbf{W}_{r_j \rightarrow g_i}$ to feed the elastic net. The weights are computed as follow:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{Sm}_{r_j, g_i})^\gamma} \quad (40)$$

Algorithm 7 outline the steps for computing the prior weights from functional association scores.

Algorithm 7 Steps for computing the functional prior weights

- 1: Let N the total number of genes
- 2: **for** each pair $(r_j, g_i), i = 1, \dots, N_{cl_i}, j = 1, \dots, N_{TF}$ compute the functional similarity as: **do**

$$sim_{BMA}(r_i, g_j) = \frac{\sum_{i=1}^m \max_{1 \leq j \leq n} sim(go_{1i}, go_{2j}) + \sum_{j=1}^n \max_{1 \leq i \leq m} sim(go_{1i}, go_{2j})}{m + n}$$

- 3: **end for**
- 4: Store the functional similarity between all pairs of genes into a matrix $\mathbf{Sm} = \{sim_{BMA}(r_j, g_i)\}$
- 5: Transform the similarities into weights $\mathbf{W}_{r_j \rightarrow g_i}$ as:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{Sm}_{r_j, g_i})^\gamma}$$

4.5.1.2 Integrating ChIP-seq

We also considered ChIP-seq data as prior information. The ChIP-seq methodology is very effective at investigating genome-wide protein–DNA interactions; therefore, identifying regions in the genome where a TF will bind to control the expression of its target genes.

Our aim here is to compute a score of potential binding between each TF and all the genes considered. We use the BETA [243] software of the Cistrome database (<http://cistrome.org>). Cistrome offers an integrative pipeline to help analyzing publicly available high-throughput data.

A simple method to get the TF-TG associations from peaks in the ChIP-seq data is to assign each TF to the proximal gene or the gene containing the TF peaks in its promoter region. Nevertheless, this will result in unreliable results. In fact, for most ChIP-seq data, only a small percentage of binding is found at the genes' promoters [243]. Also, assigning a TF to a gene only based upon the presence of the peak at a promoter of genes will produce a binary vector that is not the type of

input BENIN is expecting for the moment. Instead, we decided to consider a metric, the regulatory potential from BETA software, that is computed as the sum of the individual contribution of the binding sites.

The regulatory potential reports the likelihood of a gene to be regulated by a TF. It is computed as in Equation 41

$$\mathbf{Sch}_{r_j \rightarrow g_i} = \sum_{l=1}^k e^{-(0.5+4\Delta_l)} \quad (41)$$

, where k is the number of the binding sites of the TF r_j near the transcription start site (TSS) of the gene g_i . Only binding sites within a user-defined region length are considered. We set region length to the default value on BETA software (100Kb). Δ is the exact distance between a binding site and the TSS. It is proportional to 100Kb (note that $\delta = 0.1$ is equivalent to 10Kb). We can also restrict the number of binding sites that will contribute to computing the binding potential. We run the BETA software on Galaxy <http://cistrome.org/ap/> with the default parameters: the number of peaks considered is 10000, and the distance from gene TSS within which peaks will be selected is 100Kb.

Finally we integrate the regulatory potential into BENIN as in Equation 42.

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{Sch}_{r_j \rightarrow g_i})^\gamma} \quad (42)$$

Algorithm 8 summarizes the steps for getting the BENIN ChIP-seq prior weight from the the input BED files.

4.5.1.3 Integrating TFBS

We considered data from position weight matrices (PWMs) and promoter sequences to get an apriori information of potential binding between each TF and the TGs. These matrices are obtained from different technologies, such as Chip-Chip.

Our aim here is to scan the promoters of the genes for occurrences of each PWM. We used FIMO [85] which is a tool of the MEME-suite [10]. It scans the promoter region of each gene for individual matches to each provided input PWM. The only parameter that we set is the background file, which is the 0-order background model and the q-value threshold. We set the q-value threshold to 0.05. The 0-order background model is used to convert a frequency matrix into a log-odds score matrix and estimate the

Algorithm 8 Steps for transforming ChIP-seq data into association scores

Input: A list of BED files obtained from ChIP-seq experiments.

- 1: **for** each BED files obtained in Section 4.4.1.2 **do**
- 2: Upload the file into the Cistrome-galaxy server <http://cistrome.org/ap/root>, using the import tab. Figure 21a gives an overview of BED file for the *BRCA1* TF.
- 3: Select the Integrative analysis, then **BETA** and finally **BETA-minus** as we want to infer TF target genes only ChIP-seq data.
- 4: Set the input parameters of the **BETA-minus** software. For our experiment, we use the default parameters.
- 5: Run the **BETA-minus** on the uploaded ChIP-seq file (BED) file and collect your output output. Figure 21b gives an overview of the **BETA-minus** output file. The binding potential score for each edge $r_j \rightarrow g_i$ is computed as :

$$\mathbf{Sch}_{r_j \rightarrow g_i} = \sum_{i=1}^k e^{-(0.5+4\Delta_i)}$$

- 6: Transform each score into BENIN weight using:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{Sch}_{r_j \rightarrow g_i})^\gamma}$$

- 7: **end for**
-

p-values of match scores. We build the background file considering all the promoters sequences. We use the `fasta-get-markov` tool from `MEME-suite`. We run the tool with the default parameters. `FIMO` outputs a file containing the scores, the p-values, and the q-value of each motif occurrence. The q-values are adjusted p-values following the Benjamini and Hochberg method. Note that one PWM can have several matches at the promoter region of a gene. To assign a score to TF-TG pair, we considered the occurrence with the lowest q-value.

The challenge here is to transform the q-values into corresponding probabilities of edges being present in the final network. Let $P_{r_j \rightarrow g_i}$ be a random variable over $[0, 1]$ which represents the q-value of the binding occurrence of the TF r_j at the promoter region of g_i ($E_{r_j \rightarrow g_i}$). We assume here that it is exponentially distributed if $E_{r_j \rightarrow g_i} \in G$, and uniformly distributed if $E_{r_j \rightarrow g_i} \notin G$. More formally we have:

$$\Pr(P_{r_j \rightarrow g_i} = p | E_{r_j \rightarrow g_i} \in G) = \lambda e^{-\lambda p} / (1 - e^{-\lambda}), \quad (43)$$

where λ is the parameter controlling the scale of truncated exponential distribution, and:

$$\Pr(P_{r_j \rightarrow g_i} = p | E_{r_j \rightarrow g_i} \notin G) = 1. \quad (44)$$

We use the Bayes formula to define the probability of the edge $E_{r_j \rightarrow g_i}$ in G , knowing the binding q-value as follow:

$$\Pr(E_{r_j \rightarrow g_i} \in G | P_{r_j \rightarrow g_i} = p) = \frac{\lambda e^{-p\lambda} \beta}{\lambda e^{-p\lambda} \beta + (1 - e^{-\lambda})(1 - \beta)}, \quad (45)$$

where $\beta = \Pr(E_{r_j \rightarrow g_i} \in G)$ is the probability that an edge $E_{r_j \rightarrow g_i}$ is in the graph without any prior knowledge. We further assume that λ is uniformly distributed over the interval $[\lambda_{min}, \lambda_{max}]$ and integrate Equation 45 over that interval. The new equation for computing the conditional probability on an edge $E_{r_j \rightarrow g_i}$ is:

$$\Pr(E_{r_j \rightarrow g_i} \in G | P_{r_j \rightarrow g_i} = p) = \frac{1}{\lambda_{max} - \lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} \frac{\lambda e^{-p\lambda} \beta}{\lambda e^{-p\lambda} \beta + (1 - e^{-\lambda})(1 - \beta)} d\lambda \quad (46)$$

Equation 46 can be easily computed numerically for fixed values of $P_{r_j \rightarrow g_i}$. We pre-compute the probabilities associated with each q-value and store them in a matrix \mathbf{A} which is then transformed into weights. We then compute the weight matrix \mathbf{W} as the component-wise inverse of the elements of the matrix \mathbf{A} raised to the power $\gamma > 0$:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{A}_{r_j \rightarrow g_i})^\gamma} \quad (47)$$

Algorithm 9 summarizes the steps for computing the binding prior weight from the TFBS and promoter regions.

Algorithm 9 Step to compute prior weight from position weight matrice

- 1: Transform each Cis-BP PWMs into MEME input format using the R library `universalmotif` version 1.0.22 [232]

```
> lapply(seq(1,nbmotif),writemotif, allmotifsfilename=  
allmotifsfilename , matallmotif=sub_description_motif)  
> cisbpmotifs<-read_cisbp(allmotifsfilename)  
> memecisbpmotifsfilename=" ../data/data_human/final_data_  
hum_reg_network/Hela_data/Homo_sapiens_2020_02_24_4-34_pm/  
Homo_sapiens.meme"  
> write_meme(cisbpmotifs , memecisbpmotifsfilename)
```

- 2: create the 0-order background file for motif scanning with the `fasta-get-markov` tool from the `MEME-suite`. We use all the promoter sequences all together to create our background as with the following command:

```
$ fasta-get-markov -dna -m 0 promoter_sequence_all_fa  
backgroundpromoter
```

The background model gives the frequencies of the four bases (A, C, G, T) since we are working with DNA sequences.

- 3: Perform promoter motif scanning with the `FIMO` from `MEME-suite` with the following bash command:

```
fimo --bfile backgroundpromoter --qv-thresh  
--thresh 0.05 --verbosity 1 --oc res_promoter_  
scanning/humanpromoterseq Homo_sapiens.meme humanpromo  
terseq.fa
```

- 4: Collect `FIMO` output files. Figure 22 gives a snapshot of the `FIMO` output file after scanning promoter sequences.
- 5: Transform the q-values (adjusted p-values) into corresponding probabilities using Equation 46 and store them in a matrix **A**

6: Compute the weight matrix \mathbf{W} as the component-wise inverse of the elements of the matrix \mathbf{A} raised to the power $\gamma > 0$ as:

$$\mathbf{W}_{r_j \rightarrow g_i} = \frac{1}{(\mathbf{A}_{r_j \rightarrow g_i})^\gamma}$$

motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence
T095041_2.00		ENST000006	646	667	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000006	668	689	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000005	679	700	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000004	693	714	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000006	775	796	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000004	777	798	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000003	859	880	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000004	879	900	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000003	881	902	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000006	884	905	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000003	884	905	-	28.5055	6.94E-11	1.02E-05	GGACTACAAGTCCCAGAATCCC
T095041_2.00		ENST000004	23	44	+	28.2088	9.83E-11	1.32E-05	aaactacaatcccagaatcct
T095233_2.00		ENST000004	458	479	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000004	459	480	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000004	460	481	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000004	466	487	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000003	479	500	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000004	529	550	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000006	652	673	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000003	652	673	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000006	652	673	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000004	832	853	-	26.8193	4.09E-10	6.25E-05	CCCGCCTCGGGCCCCGCCCT
T095233_2.00		ENST000004	737	758	-	26.494	5.68E-10	7.89E-05	CCCGCCTCGGGCCCCGCCCT
T094868_2.00		ENST000005	827	841	-	23.0723	1.27E-09	0.000671	GGCCACGCCCTCC
T094868_2.00		ENST000003	857	871	-	23.0723	1.27E-09	0.000671	GGCCACGCCCTCC
T094868_2.00		ENST000002	879	893	-	23.0723	1.27E-09	0.000671	GGCCACGCCCTCC
T095233_2.00		ENST000004	46	60	+	20.494	1.43E-09	0.000105	cccccccccccc
T095233_2.00		ENST000004	47	61	+	20.494	1.43E-09	0.000105	cccccccccccc
T095233_2.00		ENST000004	48	62	+	20.494	1.43E-09	0.000105	cccccccccccc

Figure 22: Snapshot of FIMO output

The file gives an overview of the FIMO output file after scanning genes promoter sequences.

4.5.1.4 Integrating Knockdown Expression Data

Knockdown gene expression data are expression data measured in an organism where the expression of one or more of its genes is reduced. KD expression data help to infer the direct target genes of the perturbed TF. Our idea here is to get the probabilities of interactions between the perturbed TF and all the genes in the genome (the considered genes).

We analyzed the raw data with R. We performed differential expression analysis using either `Limma` [188] (for microarray expression data) or `DESeq2` [150] (for RNA-seq expression data). Our objective is to get the adjusted p-values from which we will derive the probabilities of the TF-TG interactions. Note that the p-values of differential expression analysis are adjusted with the False Discovery Rate approach

(FDR). Hence the adjusted p-values are q-values. For TFs investigated in several KD datasets, we combined them using the following idea: if we have several different q-values for the same edge $r_j \rightarrow g_i$, we considered the minimum q-value. We then follow the methodology described in Section 4.5.1.3 to get the probabilities that will then be integrated into BENIN.

In summary, we proceeded as follow to transform KD expression data into BENIN prior weights:

Step 1 For each file downloaded manually from GEO database perform differential expression analysis with `Limma` R library if for microarray experiment or with the R library `DeSeq2` for RNA-seq data. The data downloaded from `KnockTF` are obtained from differential expression analysis performed by the author of the database.

Step 2 Combine result in Step1 with data from `KnockTF` differential expression analysis. There are two cases:

- In the first situation, the TFs KD data are analyzed twice (our analysis and the `KnockTF` analysis). Each edge concerning the TF will appear twice. In this case, for each edge, we attributed the minimum of all the reported q-values.
- In the second situation, each TF is analyzed once. In this case, we add the edges and their reported q-values to the final set of potential prior interactions from KD gene expression analysis.

Figure 23 shows a snapshot of the data obtained after performing differential expression analysis and combining our results with the data from `knockTF`

Step 3 Transform the q-values obtained from differential expression analysis into probabilities using Equation 46 and store the obtained probabilities into a matrix **A**.

Step 4 Use Equation 47 to transform **A** into the prior weight to feed BENIN

4.5.2 Orthology Information Transfer

Another new functionality of **BENIN** is the integration of knowledge and discoveries about regulations from other organisms into the organism of interest. We exploit the idea that orthologous TFs regulate orthologous genes. Orthologous genes are genes from different species that evolve from a common ancestral gene and that preserve the same function. Thus, our idea is to transfer information about regulation from several well-known organisms into the genome we are currently studying. Using information from orthologous genes enriches the studied organism from expression data and prior knowledge data with new TF-TG regulatory links.

We detect orthologous genes in other organisms using sequence similarity at the protein level [176]. We got the human orthologous genes into other organisms from **eggNOG** [111]. More specifically, we run **eggNOG-mapper** [110], as it offers a quick and easy way to get the list of orthologs for several genes in parallel. Note that **eggNOG-mapper** is mainly a tool for functional sequences annotation based on orthology assignments. However, it also allows retrieving the orthologues considered to perform the functional annotation. In this work, we only consider the mouse as our model. First of all, because it is a well-studied organism (model organism). Also, it is the only mammal organism we have access to enough regulatory interactions. We consider mammal organisms principally as we are working on the uterine cervix. After collecting the orthologs in mouse, we mapped the interactions from mouse to human. We ended with a network with 545 regulatory links, 27 TFs, and 341 out of 602 orthologs.

More formally we proceed as described in Algorithm 10

The whole process to infer the HeLa cell cycle data using **BENIN** is summarized in Figure 24.

Algorithm 10 Ortholog Information transfer

- 1: Collect the proteins sequences of the studied organism from UniProt databses
- 2: Remove duplicated sequences.
- 3: Collect different model organism regulatory interactions.
- 4: Find othologs proteins in the model organisms using eggNOG-mapper with the following bash command:

```
> emapper.sh --data_dir /datasets -i uniprot-human-cell
cycle.fasta --predict_ortho --output_dir
human_orth_eggnog --target_orthologs
one2one -o hum-cellcycle-ouput -m diamond
--seed_ortholog_evalue 0.001
--seed_ortholog_score 60 --query-cover 30
--subject-cover 30 --go_evidence
non-electronic --override
```

- 5: For each model organism considered: let r_j^{model} a TF in the current model and r_j it ortholog in the studied organism. Let g_i^{model} a TG in the current model and g_i it ortholog in the studied organism. For each interaction $r_i^{model} \rightarrow g_i^{model}$: infer an interaction $r_i \rightarrow g_i$ the studied organism.
 - 6: Combine the new inferred regulatory links with those inferred with expression data. Either with max or with average
-

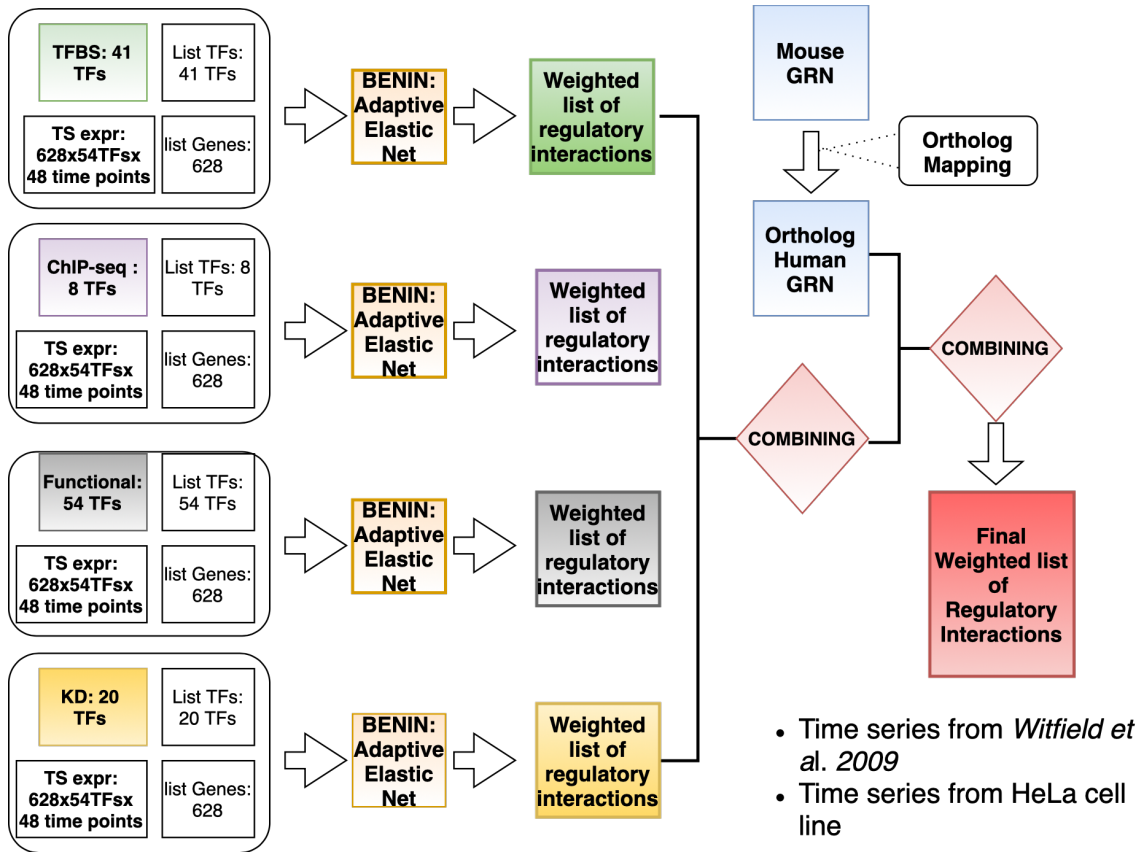


Figure 24: Inference of GRN controlling HeLa cell cycle through BENIN

The figure summarizes the whole process for inferring the GRN controlling the HeLa cell cycle. Different prior knowledge data will be integrated independently. We considered TFBS, ChIP-seq, functional annotation, and KD. Each prior knowledge data is combined with time-series expression data and will produce a weighted list of regulatory interactions. The four weighted lists will be combined. Then we get regulatory interactions from mouse through orthology mapping and combine them with those from expression and other prior knowledge data.

4.5.3 Experiments

We perform all the computations on the ENCS speed cluster. It has sixteen, 32-core nodes, each with 512 GB of memory and approximately 1 TB of volatile-scratch disk space. The results presented in Section 4.6.1 are obtained with the following BENIN parameters: the elastic net parameter $\alpha = 0.9$, the exponent $\gamma = 1.5$, and the number of bootstraps $R = 5000$. We set the parameters to the same values when running

BENIN with different prior knowledge data. The ensemble network is obtained either using the average or max score. Note that we ignore missing values when using the **mean** to combine results from different prior knowledge data. Some prior knowledge data have missing information about some TFs and their TGs. We set the threshold $\tau = 0.5$ on the final regulatory links weights to get the final inferred network. As we are dealing with imbalanced data (more negative than positive examples), the AUPR is more representative of the model performance, as it does not account for true negatives.

An important BENIN parameter is the threshold τ that allows selecting the final list of edges present in the inferred regulatory network. To select τ , we vary its value in the $[0, 1]$ interval and we record the AUPR score. We considered the AUPR score as we are working with imbalanced classes. AUPR is the most informative in this case of imbalanced classes. The number of true edges is less than the false edges. We also need to set τ so that we have a good compromise between high-scoring edges and good AUPR.

4.6 Results and Discussion

4.6.1 Results

BENIN execution time Table 17 reports BENIN execution time. These results are obtained setting the number of bootstraps R to 1000. The results are reported for BENIN with prior and without prior. We requested 25 cores on the cluster server to measure the computation time. Time is the elapsed time measured in seconds. From Table 17 we can observe that for the whole network of size 628 edges, BENIN takes 7399s ($\approx 2h$) when we integrate all the prior knowledge, including the orthology from mouse data and 5335s ($\approx 1h$) when we do not integrate prior knowledge.

Integrating Prior knowledge improves BENIN performance Figure 26 (respectively Figure 27) shows the precision-recall (respectively the ROC) curve when BENIN is combined or not with prior knowledge data. Table 18 reports the AUPR and AUROC scores when we do not consider prior knowledge data, and we combine BENIN with different prior knowledge data. Figure 28 (respectively Figure 29) shows

Table 17: BENIN execution time

Prior	All genes (628)
TFBS	1414s
ChIP-seq	1420s
KD	1474s
Functional	2955s
All priors	7399s
None	5335s

BENIN execution time on a 628 genes network when we consider different prior knowledge data separately, integrate all the prior knowledge data and do not consider prior knowledge data. The time is the elapsed time in seconds.

the BENIN performance when we combine the output network from time-series gene expression data (and other prior) with the network from orthology mapping. These scores report how well BENIN performs on what is known about the regulatory interactions in the human. As we are working with imbalanced data, the AUPR is more informative than the AUROC. From Table 18, we observe that regarding the nested confidence intervals, it is difficult to distinguish the performances of BENIN when we integrate the prior knowledge data separately and when we combine them. However, we observe two groups: the red group and the blue group.

chr19	49999221	49999664	1000	565.801079	-1	3.10071509	243
chr14	102414332	102414670	1000	562.279321	-1	3.10071509	169
chr12	123237192	123237501	1000	541.320919	-1	3.10071509	151
chr17	4843383	4843762	1000	525.270115	-1	3.10071509	197
chr3	49131395	49131698	1000	524.808057	-1	3.10071509	150
chr12	49351180	49351491	1000	513.533546	-1	3.10071509	155
chr17	37617417	37617831	1000	510.331406	-1	3.10071509	205
chr9	123605108	123605466	1000	506.823221	-1	3.10071509	178
chr9	98637701	98638053	1000	506.241484	-1	3.10071509	179
chr11	13484708	13484993	1000	501.150918	-1	3.10071509	146
chr20	34330089	34330557	1000	494.440202	-1	3.10071509	242
chr17	53045936	53046217	1000	494.050645	-1	3.10071509	144
chr10	70091603	70091871	1000	493.172988	-1	3.10071509	139
chr6	35995280	35995658	1000	493.16754	-1	3.10071509	178
chr9	6413014	6413290	1000	491.231375	-1	3.10071509	138
chr15	22833255	22833574	1000	487.991004	-1	3.10071509	165
chr9	88555786	88556082	1000	481.500613	-1	3.10071509	147
chr12	29534022	29534324	1000	480.66344	-1	3.10071509	153
chr16	90088879	90089153	1000	478.981445	-1	3.10071509	141
chr6	44355100	44355495	1000	476.895956	-1	3.10071509	172
chr1	156252514	156252855	1000	476.170969	-1	3.10071509	166
chrX	77154735	77155025	1000	475.730522	-1	3.10071509	154
chr17	4167132	4167453	1000	474.314332	-1	3.10071509	169
chr10	51565011	51565317	1000	473.985442	-1	3.10071509	155
chr1	207226192	207226488	1000	473.89962	-1	3.10071509	142
chr1	45987463	45987756	1000	473.589863	-1	3.10071509	146
chr1	110576991	110577320	1000	473.122573	-1	3.10071509	165
chr7	75677203	75677502	1000	472.766147	-1	3.10071509	155
chr7	129845194	129845478	1000	471.6041	-1	3.10071509	142
chr18	47018763	47019043	1000	466.855044	-1	3.10071509	145

(a) A snapshot of a BED file

```
# Argument List:
# Name = BRCA1_chipseq_bindingscore
# peak file = /project/Cistrome/CistromeAP/galaxy_database/files/001/633/dataset_1633721.dat
# distance = 100000 bp
```

Chromosome	TSS	TTS	RefseqID	Score	Strand	GeneSymbol
chr11	65190268	65194003	NR_028272	4.723	+	NEAT1
chr17	8090262	8090322	NR_039746	4.219	+	MIR4521
chr17	8076296	8079714	NM_032354	4.083	-	TMEM107
chr17	8076296	8079714	NM_183065	4.083	-	TMEM107
chr17	8023907	8027410	NM_032580	4.064	-	HES7
chr17	8023907	8027410	NM_001165	4.064	-	HES7
chr17	8091650	8093564	NM_017622	3.976	-	C17orf59
chr12	125400092	125400205	NR_049820	3.923	+	MIR5188
chr17	7999217	8022234	NM_001165	3.923	-	ALOXE3
chr12	125396190	125399587	NM_021009	3.911	-	UBC
chr17	7999217	8021860	NM_021628	3.866	-	ALOXE3
chr17	8043787	8055753	NM_002616	3.852	-	PER1
chr17	8048311	8048389	NR_106943	3.802	-	MIR6883
chr17	8123947	8127361	NR_026951	3.78	-	LINC00324
chr17	8062464	8066293	NM_014232	3.745	-	VAMP2
chr17	8108048	8113944	NM_004217	3.718	-	AURKB
chr17	8108048	8113944	NM_001284	3.718	-	AURKB
chr17	8108048	8113944	NM_001256	3.718	-	AURKB
chr5	180618045	180618908	NR_108031	3.59	-	LOC102577426
chr5	180620923	180632293	NM_203293	3.587	-	TRIM7
chr5	180630121	180632293	NM_033342	3.587	-	TRIM7
chr5	180620923	180631340	NM_203295	3.554	-	TRIM7
chr5	180620923	180631340	NM_203294	3.554	-	TRIM7
chr5	180620923	180631340	NM_203296	3.554	-	TRIM7
chr5	180620923	180627930	NM_203297	3.478	-	TRIM7
chr5	180649565	180649633	NR_039781	3.391	-	MIR4638
chr5	180650363	180650388	NR_039781	3.283	-	TRIM41

(b) BETA-minus output

Figure 21: BED file and BETA-minus output

Overview of a BED from a ChIP-seq experiment for the BRCA1 TF and a BETA-minus output file (a) Snapshot of a BED file the BRCA1 transcription. (b) factor Snapshot of the BETA-minus output file after analyzing the *BRCA1* BED file on Cistrome-galaxy server.

TF	Gene	P.value	adj.P.Val	Log2FC
YY1	IFI44L	0.02191	0.34464119	-5.28455
YY1	C7orf57	0.0846	0.45850327	4.43378
YY1	KLHL32	0.06895	0.43188269	3.93168
YY1	TREM2	0.01334	0.31852849	3.3117
YY1	EPHX2	2.96E-06	0.04907872	-3.28387
YY1	LRRC25	0.028	0.35490052	3.15417
YY1	IGJ	0.047	0.39789132	3.08196
YY1	PXDNL	0.00437	0.30728877	3.06759
YY1	TECRL	0.00691	0.30922751	3.05896
YY1	TTC29	0.00296	0.29330698	3.05737
YY1	CNN1	0.00057	0.29330698	3.04694
YY1	RAB25	0.00062	0.29330698	3.01797
YY1	MSTN	0.05825	0.4166842	3.00867
YY1	MYH1	0.00041	0.29330698	2.93958
YY1	FU16779	0.01816	0.33224645	2.91672
YY1	DEFB129	0.00059	0.29330698	2.87318
YY1	YJEFN3	0.0042	0.30436245	-2.85574
YY1	OLFM4	0.01947	0.33309758	-2.83596
YY1	FAM196A	0.08703	0.46084373	2.8087
YY1	DEFB114	0.00025	0.29330698	2.78306
YY1	ZNF404	0.05061	0.40456308	2.75973
YY1	GIMAP7	0.00906	0.30922751	2.74925
YY1	IGSF11	0.06296	0.42482662	2.70465
YY1	NCKAP5	0.18218	0.57464601	2.68144
YY1	KAAG1	1.12E-05	0.06170298	2.67766
YY1	C11orf96	0.03124	0.363653	2.66136
YY1	GPR22	0.01785	0.33224645	2.63021
YY1	LOC253573	0.00599	0.30922751	2.60181
YY1	TNNC1	0.00039	0.29330698	-2.59059
YY1	ID2B	0.00031	0.29330698	2.57819
YY1	LOC286083	0.00012	0.27658333	2.57361
YY1	MYBPC2	0.12176	0.50768778	2.55998

Figure 23: Differential Expression analysis output

The figures gives a snapshot of the combined data after performing differential expression analysis and combining our results with data from knockTF.

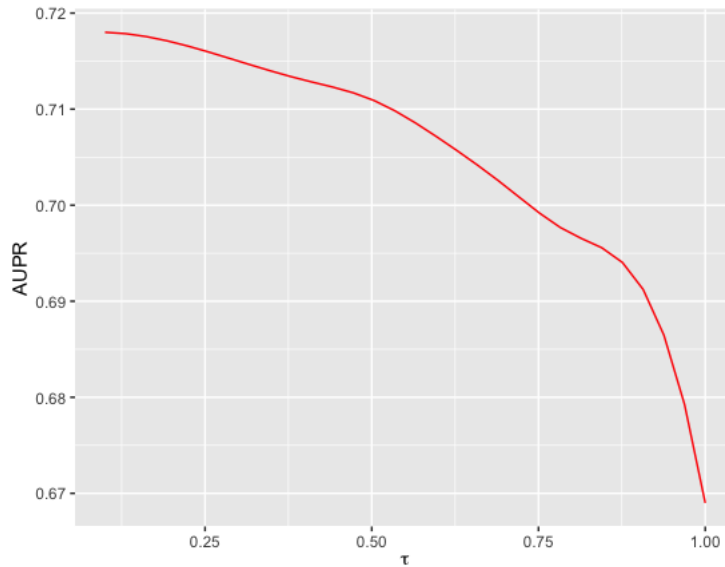
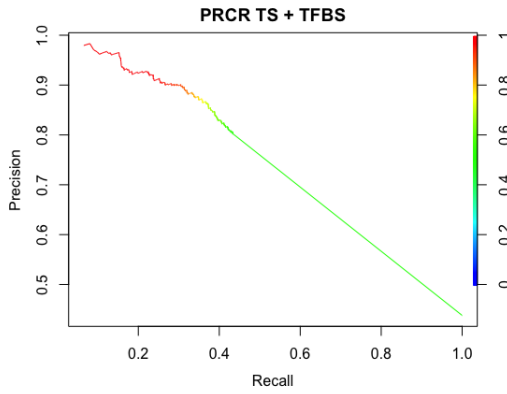


Figure 25: Effect of τ on BENIN performance.

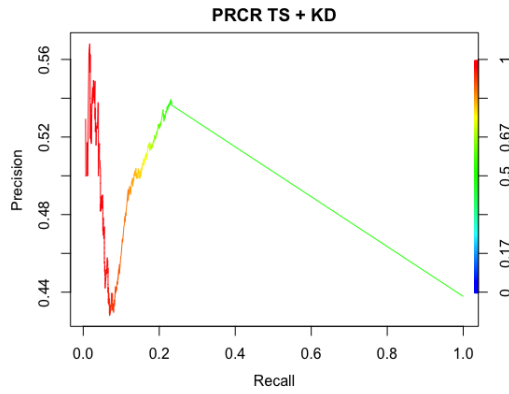
Table 18: BENIN performance

Method	AUPR	AUROC
BENIN+none	0.440 [0.398; 0.483]	0.501 [0.488; 0.514]
BENIN+KD	0.490 [0.446; 0.534]	0.684 [0.670; 0.698]
BENIN+TFBS	0.733 [0.692; 0.771]	0.755 [0.739; 0.779]
BENIN+Chipseq	0.732 [0.693; 0.769]	0.686 [0.672; 0.700]
BENIN+functional	0.479 [0.438; 0.521]	0.527 [0.513; 0.540]
BENIN+combined+max	0.711 [0.682; 0.737]	0.767 [0.751; 0.783]
BENIN+combined+mean	0.547 [0.511; 0.583]	0.580 [0.565; 0.594]
BENIN+combined+max+orth+mean	0.715 [0.687; 0.741]	0.771 [0.756; 0.787]
BENIN+combined+max+orth+max	0.702 [0.673; 0.730]	0.775 [0.759; 0.790]

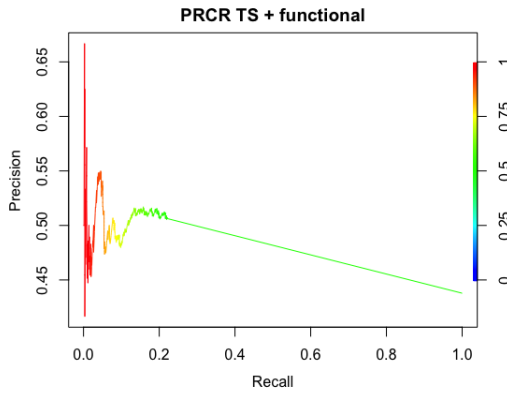
The table reports AUPR and AUROC scores when we BENIN run with or without prior knowledge data to infer the GRN that controls the human HeLa cell cycle. We also provide in bracket the confidence interval of these scores. The highest score is marked in bold. The results are obtained setting the BENIN parameters as following: $R = 5000$, $\alpha = 0.9$ and $\gamma = 1.5$



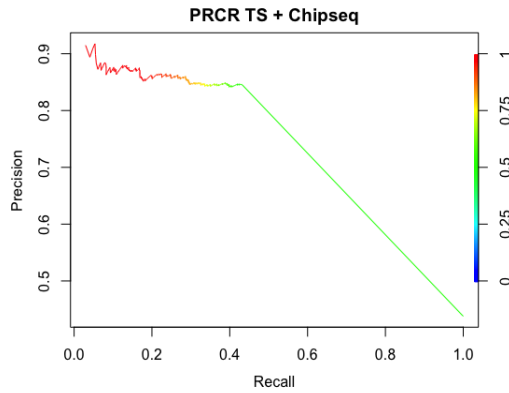
(a) BENIN +TFBS



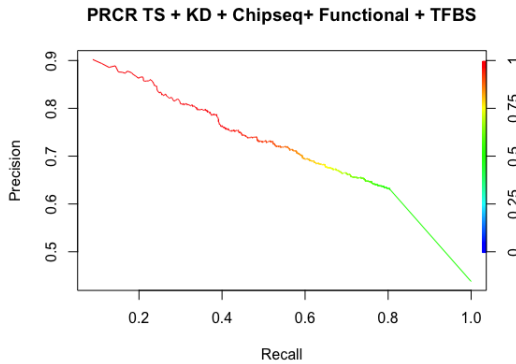
(b) BENIN +KD



(c) BENIN +functional



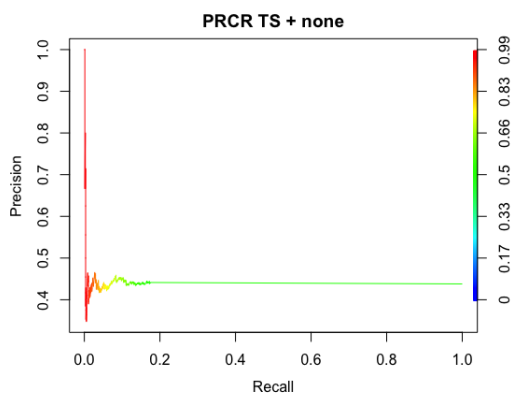
(d) BENIN +chipseq



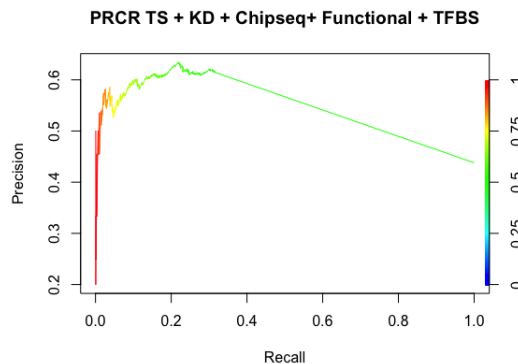
(e) BENIN +combined+max

Figure 26: Precision-recall curves for BENIN

The figure shows the precision-recall curves when using BENIN combined with different prior knowledge data to infer the GRN controlling the HeLa cell cycle.



(f) BENIN +noprior



(g) BENIN +combined+mean

Precision-recall curves for BENIN

BENIN can infer real regulatory networks We dig up the results to perform gene annotation analysis of coregulated genes. Firstly, we consider known cell-cycle transcription factors [62, 18], and analyze the GO annotation of their target genes. Then we perform a manual literature analysis of some selected inferred interactions. We considered a total of 62 edges.

We consider annotations that have at least five genes and which have an adjusted p-value $\leq 5e^{-2}$. In Table 19 we report a non exhaustive list of TG annotations. For each group of coregulated genes, we report the annotations related to the annotation of its TF. In Table 19, we report the annotation of inferred targets genes of E2F1 (a main regulator in the cell cycle that binds many important targets genes in the cell cycle [119]), SP1, NFYA, YY1, FOXM1 and, KLF6. Some of these TFs are members of a family/complex (i.e., E2F1 member of the E2F or KLF6 member of the KLF family, NFYA member of NFY) and control (activate or repress) approximately the same genes and participate approximately in the same biological process. So we chose to analyze one member of each protein family. Table 19 shows that the annotations of coregulated genes are consistent with the annotation of their TF. For example, the literature has reported SP1 as a key transcription factor in regulating cell proliferation [48, 236]. Out of 613 inferred edges with a score of at least 0.5, 111 genes are annotated as part of the cell proliferation. Another interesting finding is that several edges in the vicinity of the FOXM1 transcription factor are annotated as Mitotic genes (77), and we know from the literature that FOXM1 is an essential transcription

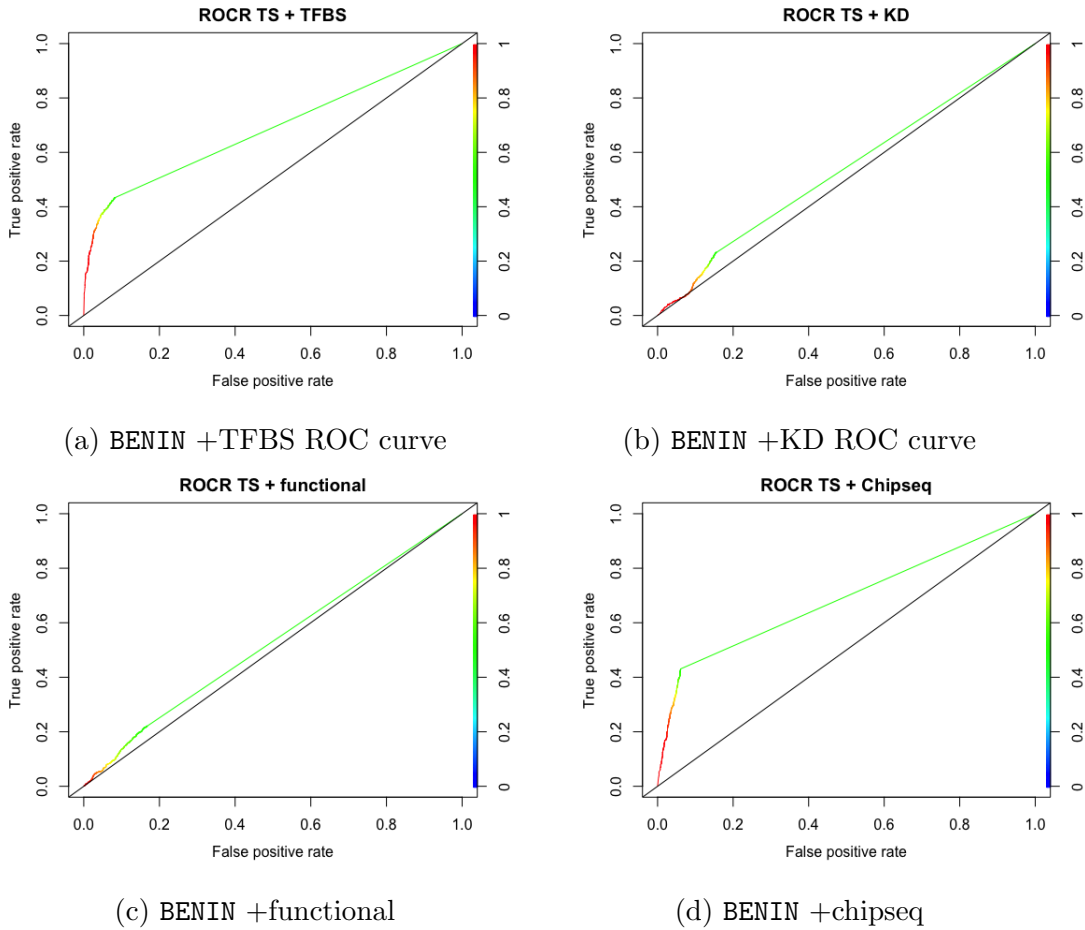
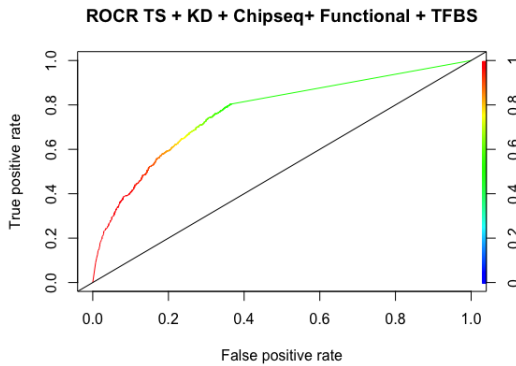


Figure 27: ROC curves for BENIN

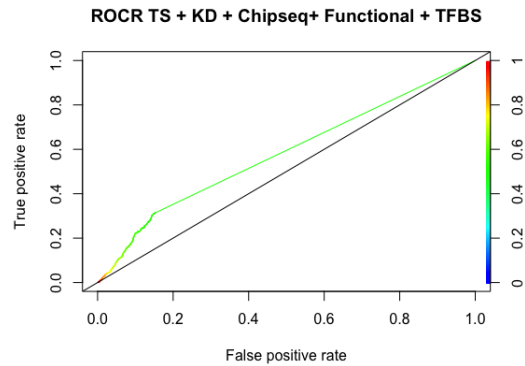
The figure shows the ROC curves when using BENIN combined with different prior knowledge data to infer the GRN controlling the HeLa cell cycle. The results are obtained setting the BENIN parameters as following: $R = 5000$, $\alpha = 0.9$ and $\gamma = 1.5$

factor for the progression of through the Mitotic phase of the cell cycle [242], and it has its peaks expression at the S and G2/M phases. It is a master regulator of genes that ensure the transition from G2 to M phase and the progression through mitosis. From our functional annotation, some of the FOXM1 inferred TGs are mitotic cell cycle genes (77/227). It is coherent with the function of FOXM1 as it a key role in progression through Mitosis [242, 139].

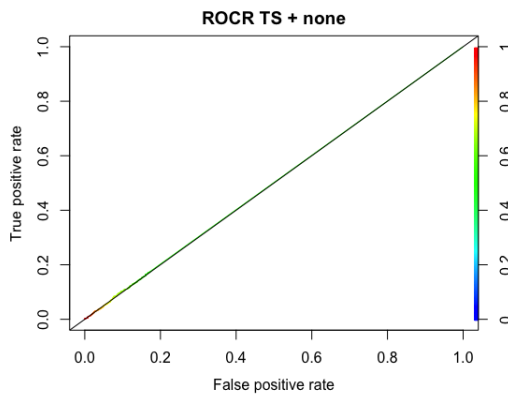
Moreover, BENIN was able to infer CDC25B as a target of FOXM1. CDC25B is essential for progression into mitosis [58]. Another interesting finding is that BENIN



(e) BENIN +combined+max



(f) BENIN +combined+mean



(g) BENIN +combined+mean

ROC curves for BENIN

inferred six out the seven interactions between the FOXM1 and its direct target genes that are involved in regulating G1/S and G2/M progression [242] (AURKB, CENPA, CKS1B, CDC25B, PLK1, CDC25A, BIRC5). Specifically, we inferred interaction between FOXM1 and AURKB, CENPA, CKS1B, CDC25B, PLK1, CDC25A, and BIRC5. Four of these interactions were confirmed with the orthology mapping. It worth mention that these links are not part of our “gold-standard” network

Another interesting transcription factor is the E2F1 that is a member of the E2F family of TFs. It plays a crucial role in cell cycle regulation. It targets several proteins that regulate the transition from the G1 phase to the S phase and controls genes that play a role in DNA repair and apoptosis [187]. Out the 122 genes that are expressed in HeLa cell cycle and that have been inferred as E2F1 target gene by Ren *et.al* [187], BENIN inferred 33. Out of the 38 genes expressed in the HeLa cell cycle and that

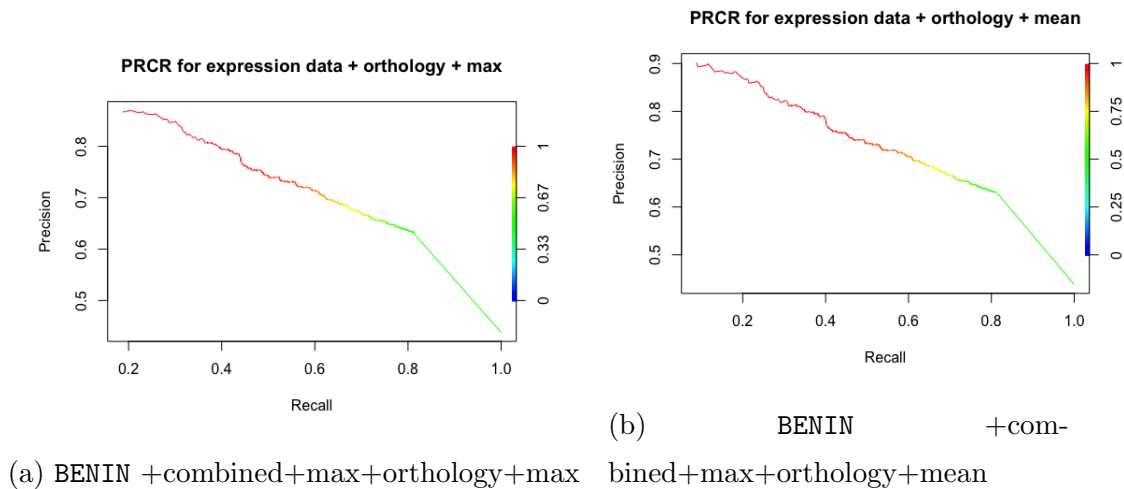


Figure 28: Precision-recall curves for BENIN +orthology

The figure shows the precision-recall curves when using BENIN combined results from orthology mapping to infer the GRN controlling the HeLa cell cycle. The results are obtained setting the BENIN parameters as following: $R = 5000$, $\alpha = 0.9$ and $\gamma = 1.5$

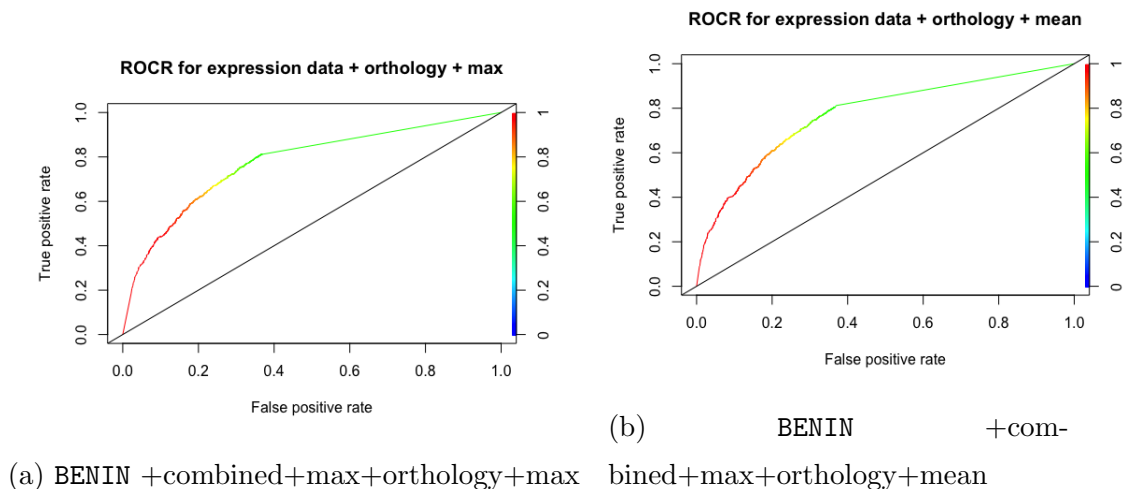


Figure 29: ROC curves for BENIN +orthology

The figure shows the ROC curves when using BENIN combined results from orthology mapping to infer the GRN controlling the HeLa cell cycle. The results are obtained setting the BENIN parameters as following: $R = 5000$, $\alpha = 0.9$ and $\gamma = 1.5$

Adrien *et.al.* have demonstrated to be potential E2F1 TG [26], BENIN has inferred 36 TGs. Some studies have demonstrated that E2F1 induces G1/S-phase genes [51],

which is consistent with the annotation of the E2F1 inferred TGs.

Table 19: Transcription factor and target gene

TFs	TF annotation	# Targets	Biological process category	adjPval
SP1	Linked to cell proliferation[19]	613	GO:0008283: Cell proliferation(111)	$4.0e^{-7}$
	Regulates apoptosis[129]		GO:0008219: Cell death(102)	$5.5e^{-5}$
	Positive regulation of transcription by RNA polymerase II[254]		GO:0034645: Cellular macromolecule biosynthetic process(229)	$2.4e^{-12}$
	DNA damage response pathway [28, 13]		GO:0006259: DNA metabolic process (130)	$2.5e^{-38}$
	Involve in regulation of transcription, DNA-templated [48, 236]		GO:0006355: Regulation of transcription, DNA-templated(182)	$6.9e^{-7}$
Fundamental player in the regulation of cell proliferation[14]		GO:0008283: Cell proliferation(111)	$4.0e^{-7}$	
NFYA	Induces Apoptosis [88]		GO:0012501: Programmed cell death(93)	$2.2e^{-4}$
	control the expression of several key regulators of the cell cycle [263, 122]		GO:0051726: Regulation of cell cycle (110)	$5.3e^{-30}$

Table 19 continued from previous page

TFs	TF annotation	# Targets	Biological process category	adjPval
			GO:0000278: Mitotic cell cycle (142)	$5.8e^{-54}$
	DNA metabolism [62]	558	GO:0006259: DNA metabolic process(116)	$3.9e^{-33}$
	involves in regulation of transcription, DNA-templated (UniProtKB:P23511)		GO:0006355: Regulation of transcription, DNA-templated(164)	$5.1e^{-6}$
	positive regulation of transcription from RNA polymerase II promoter in response to iron [79]		GO:0001079: Regulation of transcription from RNA polymerase II promoter(81)	$2.1e^{-2}$
	positive regulation of apoptotic process [219]		GO:0006915: Apoptotic process(85)	$2.6e^{-4}$
E2F1	Regulation of G1/S transition of mitotic cell cycle [195]		GO:0044843: Cell cycle G1/S phase transition(45)	$9.3e^{-20}$
			GO:1903047: Mitotic cell cycle process(143)	$9.0e^{-62}$
	DNA damage response [220]		GO:0006974: Cellular response to DNA damage stimulus(92)	$3.8e^{-26}$

Table 19 continued from previous page

TFs	TF annotation	# Targets	Biological process category	adjPval
	regulation of transcription, DNA-templated [220]	536	GO:0006351: transcription, DNA-templated (161)	$2.0e^{-6}$
	[187]		GO:0008283:cell proliferation(100)	$4.9e^{-7}$
	[187]		GO:0006281: DNA repair(72)	$2.3e^{-24}$
	[187]		GO:0006260: DNA Replication(64)	$1.1e^{-33}$
	[187]		GO:0000075: Cell cycle checkpoint(47)	$1.1e^{-22}$
	[187]		chromosome segregation(69)	$1.5e^{-34}$
	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest (UniProtKB:Q0109)		GO:0006977: DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest(11)	$3.5e^{-4}$
YY1	Many YY1-regulated genes have crucial roles in cell proliferation, differentiation, apoptosis, and cell cycle regulation[127]		GO:0006915:Apoptotic process(44)	$1.6e^{-3}$
			GO:0010564: regulation of cell cycle process (39)	$1.2e^{-13}$
			GO:0006974: cellular response to DNA damage stimulus(54)	$6.0e^{-20}$

Table 19 continued from previous page

TFs	TF annotation	# Targets	Biological process category	adjPval
	YY1 has been found to activate DNA repair[83, 209, 225]	231	GO:0006281: DNA repair(44) GO:0006260: DNA replication(33) GO:0051726: regulation of cell cycle(53)	$3.3e^{-19}$ $6.3e^{-18}$ $6.1e^{-16}$
FOXM1	FOXM1 regulates genes involved in transcription and cell cycle regulation[251] Regulates the Transcriptional Network of Genes Essential for Mitotic Progression[242] Play a key role chromosomal segregation maintenance [139]	227	GO:0051726: Regulation of cell cycle(64) GO:0000278: Mitotic cell cycle(77) GO:0044772: Mitotic cell cycle phase transition(53) GO:0044839: Cell cycle G2/M phase transition(24) GO:0007059:Chromosome segregation(36)	$4.6e^{-27}$ $5.9e^{-36}$ $1.4e^{-28}$ $2.1e^{-13}$ $2.4e^{-19}$
KLF6	KLF6 regulator of cell apoptosis [109]		GO:0006915: Apoptotic process(91) GO:0008219: Cell death (99)	$2.2e^{-4}$ $2.7e^{-4}$

Table 19 continued from previous page

TFs	TF annotation	# Targets	Biological process category	adjPval
	KLF expression was shown to mediate growth inhibition[31]	582		
	KLF6 also directly interacts with cyclin D1 to suppress cyclin-dependent kinase 4 and causes cell cycle arrest[109, 16]		cell cycle arrest(27) GO:0006974: Cellular response to DNA damage stimulus (93)	$2.8e^{-6}$ $3.6e^{-24}$
	others		GO:0033554: cellular response to stress (138) GO:0010556: Regulation of macromolecule biosynthetic process(196) GO:0016070: RNA metabolic process(211)	$2.0e^{-20}$ $1.4e^{-7}$ $1.4e^{-6}$

Table 19 continued from previous page

TFs	TF annotation	# Targets	Biological process cat- egory	adjPval
-----	---------------	-----------	----------------------------------	---------

The table summarizes the annotation of core TFs and the annotation of their inferred TGs. The 1st column gives the name of the TFs. The 2nd column provides the TFs functions, as reported in referee papers. They are from *in vivo* experiments. The 3rd column gives the total number of inferred TGs for a specific TF. The 4th column report the functional annotation of all the TGs for a specific TF (biological process). We performed the functional annotation using DAVID online functional annotation tool. The number in parenthesis represents the number of TGs associated with the GO term. We filter the GO Biological Processes terms related to their reported TF functions. We selected annotations that have at least ten of the TGs and that have the smallest adjusted p-value. In the 5th column, we provide the adjusted p-values of each GO term. It shows the statistical significance of the annotation.

Table 20 enumerates the regulatory interactions that are not part of our “gold-standard” network but are supported to some extent in the literature. We performed a manual literature analysis of the selected edges. We proceed as follow: for each TF and TG considered, we scan the PubMed papers, if any, for specific word/sentences to classify the edges as either:

- **Supportive** if there is an explicit and direct experimental evidence demonstrating the presence of such a regulatory relationship. We were looking at words that explicitly suggest regulation, such as “*binding*” and “*regulates*”.
- **Predictive** if previously documented evidence implies the possibility of the regulatory interaction between the genes, but remains to be experimentally verified. We were looking at words like “*potential binding*” “*potentially regulates*”.
- **Hypothetical** if the biological knowledge for the regulation lacks so far. We were checking if the TF and the TG share the same potential annotation.

From Table 20, we notice that BENIN infers several news interactions that are missing in our gold-standard network and interactions that need further investigation.

Table 20: Inference from BENIN +combined+max

Regulations	Category	Original	Description
$E2F1 \rightarrow MCM5$ (0.9716)	Supportive	[187, 26]	
$E2F1 \rightarrow PCNA$ (1.00)			
$E2F1 \rightarrow MCM6$ (0.999)			
$E2F1 \rightarrow TMPO$ (0.839)	Predictive	[187]	
$E2F1 \rightarrow NEK2$ (0.970)			
$E2F1 \rightarrow CKS2$ (0.949)			
$E2F1 \rightarrow BRCA1$ (0.996)	Supportive	[241, 244, 26]	<i>“E2F1 transcriptional activity leads to high expression of several DNA repair genes, including BRCA1, RAD51 and RAD52”</i>
$E2F1 \rightarrow MCM2$ (1.00)	Supportive	[26]	see Table 1 of [26, 92]
$E2F1 \rightarrow BUB3$ (0.995)			
$E2F1 \rightarrow BARD1$ (1.00)			
$E2F1 \rightarrow CASP3$ (0.999)			
$E2F1 \rightarrow BMP2$ (0.998)			

Table 20 continued from previous page

Regulations	Category	Original	Description
$E2F5 \rightarrow CDC25A$ (0.515)			
$E2F5 \rightarrow E2F1$ (0.672)			
$E2F5 \rightarrow PRC1$ (0.768)			
$E2F5 \rightarrow CDC6$ (0.668)	Predictive	[26]	It is not clearly mentioned that they are targets of E2F5 but E2F in general see Table 1 of [26]
$E2F5 \rightarrow BUB1$ (0.565)			
$E2F5 \rightarrow BUB1B$ (0.706)			
$E2F5 \rightarrow CENPE$ (0.756)			
$E2F5 \rightarrow MAD2L1$ (0.659)			
$E2F8 \rightarrow CCNE2$ (0.715)			
$E2F8 \rightarrow CDC6$ (0.898)			
$E2F8 \rightarrow MCM5$ (0.861)			
$E2F8 \rightarrow RFC2$ (0.632)			
$E2F8 \rightarrow RPA2$ (0.854)	Predictive	[26]	It is not clearly mentioned that they are targets of E2F8 but E2F in general see Table 1 of [26]
$E2F8 \rightarrow CDKN2C$ (0.874)			
$E2F8 \rightarrow BUB3$ (0.995)			
$E2F8 \rightarrow MSH2$ (0.918)			
$E2F8 \rightarrow RAD51$ (0.994)			
$E2F8 \rightarrow BMP2$ (0.658)			

Table 20 continued from previous page

Regulations	Category	Original	Description
$FOXM1 \rightarrow CENPA$ (0.935)	Supportive	[251]	<i>“FOXM1 regulates genes that are essential for proper chromosome segregation and mitosis, such as NEK2, KIF20A, and CENPA”</i>
$FOXM1 \rightarrow CDC25B$ (0.982)	Supportive	[242]	<i>“FOXM1 target genes include CDC25B and PLK1, which are important for activating CDK1 for mitosis”</i>
$FOXM1 \rightarrow CDC25C$ (0.956)		[153]	<i>“These results showed that unusual expression of FOXM1 increased the expression levels of the FOXM1 targets PLK and CDC25C”</i>

Table 20 continued from previous page

Regulations	Category	Original	Description
$NFYB \rightarrow CDC25C$ (0.894) $NFYB \rightarrow CDC25B$ (0.900)	Supportive	[155]	“ <i>NF-Y transcription factor plays a central role in cellular proliferation by controlling the expression of genes required for cell-cycle progression such as cyclin A, cyclin B1, cyclin B2, CDC25A, CDC25C, and CDK1</i> ”
$NFYA \rightarrow CDC25B$ (0.884) $NFYA \rightarrow CDC25C$ (0.99)		[155]	“ <i>NF-Y mediates the transcriptional inhibition of the mitotic cyclins and the CDC25C genes during p53-dependent G2 arrest induced by DNA damage</i> ”
$SP1 \rightarrow YY1$ (0.995)	Predictive	[83]	See Table 2 in [83]
$CENPA \rightarrow BUB1$ (1.00)	Hypothetical		Potential binding from STRINGDB [224]

The table reports the list of interactions inferred by **BENIN** but that are not part of our gold-standard network. These links are obtained when combining **BENIN** with TFBS, KD expression data, functional annotation, and ChIP-seq data. We used the max function to combine the output from different prior knowledge data. The 1st column reports the interactions as well as their score as inferred by **BENIN**. The number in parenthesis is the score returned by **BENIN** +expression. The 2nd reports the type of evidence about the interaction. It can be **supportive** if there is an explicit and direct experimental evidence demonstrating the presence of such a regulatory relationship; **predictive** if previously documented evidence implies the possibility of the regulatory interaction between the genes, but remains to be experimentally verified, or **hypothetical** if the biological knowledge for the regulation lacks so far. The 3rd column gives reference papers/works that support the evidence, if any. The 4th column gives more details from the paper that support the evidence of the interaction.

Orthology mapping confirms interactions and potential links Figure 30 shows the network inferred with orthology mapping from the mouse regulatory network. We compared the inferred network with our gold-standard network and the network inferred from the time-series expression data combined with all the prior knowledge data. For the inferred edges from expression data, we consider those whose weights are ≥ 0.5 . The point here is to highlight the extent to which the network from orthology agrees with the network from time-series. Figure 31 shows the distribution of the inferred edges with **BENIN** +orthology compared to the gold-standard and the sub-network from **BENIN** +expression (we considered only the edges shared with network from orthology). We are mainly interested in the links shared by both **BENIN** +expression and **BENIN** +orthology but that are missing in the gold-standard network (345/545) and that are false edge in the gold-standard network (28/545). We can observe that almost half of the edges in the inferred network with **BENIN** +orthology are new interactions confirmed by expression data. In Table 21 we report some of these new interactions. Note that in Table 21, we provide the edges that are not already part of Table 20. Among the new links, some of them have been reported in the literature. For example, with orthology information transfer, we can infer CNA2, FAN1, GCLM and MEPCE as target genes of CTCF. In [121], CTCF has been found to bind the promoter of these genes.

However, we are also interested in those not reported in the literature as they may suggest new interactions that will need further investigation in wet labs. It is the case of the regulatory link between FOXM1 and NCAPH, which was inferred with a score of 0.99. However, there is no literature that supports a direct interaction between FOXM1 and NCAPH. We consider the links that have high confidence (score of at least 0.80)

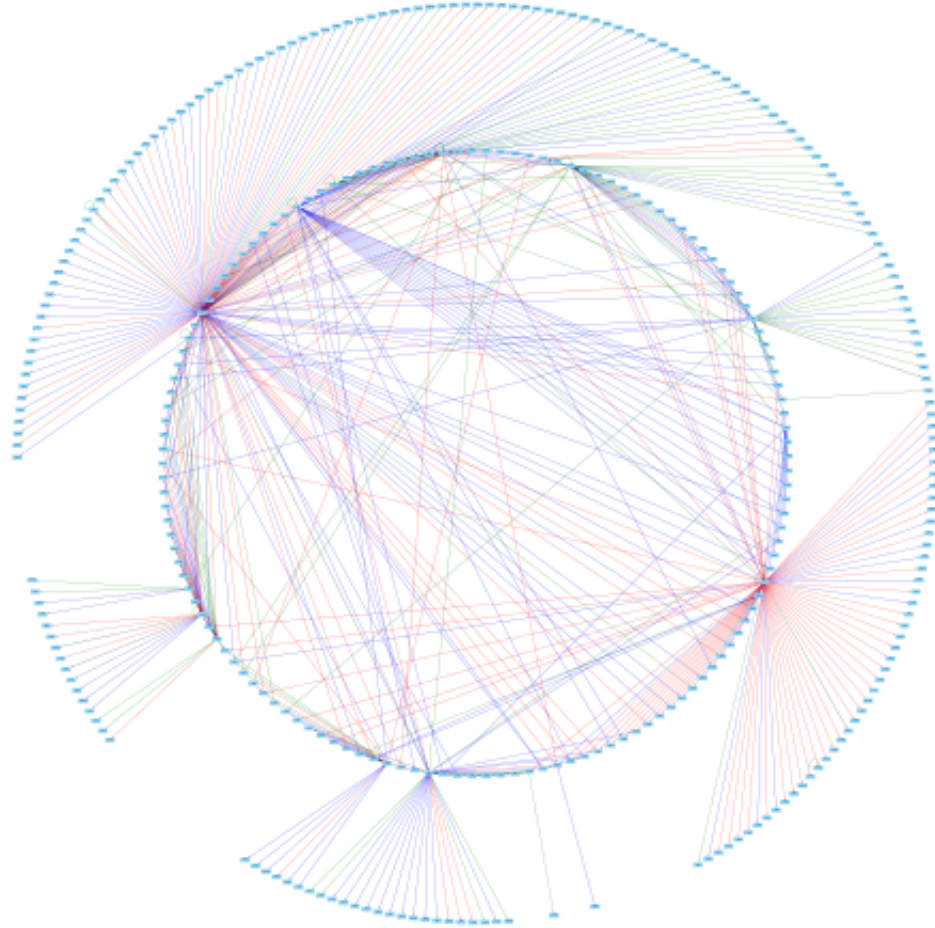


Figure 30: Orthologous Regulatory Network From mouse

The figure represents GRN controlling the HeLa cell cycle network inferred with BENIN combined with sequence orthology information transfer. We use the mouse as the model organism. Green edges are edges obtained only from orthology mapping. Red edges are shared between our gold-standard network, expression-based inferred network and orthology-based inferred network. Blue edges are those shared among the expression-based inferred network and the orthology-based inferred network. The results are obtained setting the BENIN parameters as following: $R = 5000$, $\alpha = 0.9$ and $\gamma = 1.5$

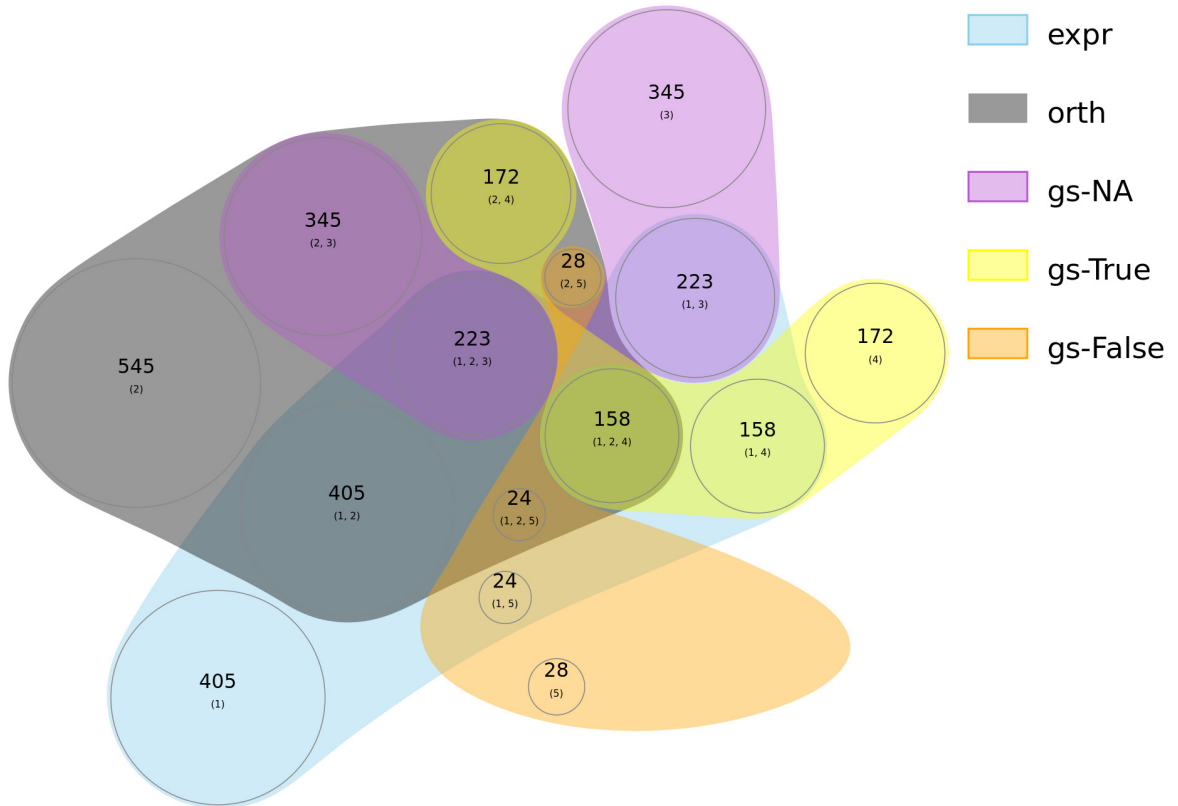


Figure 31: Edge Distribution

Distribution of the edges in the inferred Network regulatory network controlling the human HeLa cell cycle. Here BENIN has been combined with results from orthology mapping using the mouse as our model organism. We compared the obtained network to the gold-standard and the network inferred from expression and prior knowledge. “expr” represent the edges inferred from time series expression data. “orth” are the edges inferred through orthology mapping. “gs-NA” are the missing link in the gold-standard network and “gs-False” are the false edges.

Table 21: Inference from BENIN +orthology

Regulations	Category	Original	Description
<i>CTCF</i> → <i>CCNA2</i> (1)			
<i>CTCF</i> → <i>FAN1</i> (1)			
<i>CTCF</i> → <i>GCLM</i> (0.992)			
<i>CTCF</i> → <i>MEPCE</i> (0.986)	Supportive	[121]	Have found to be bounded by CTCF
<i>CTCF</i> → <i>MBD4</i> (0.996)			
<i>CTCF</i> → <i>GADD45A</i> (0.992)			
<i>CTCF</i> → <i>STAG1</i> (0.992)			
<i>CTCF</i> → <i>ANTXR1</i> (0.996)			
<i>CTCF</i> → <i>RFC2</i> (0.997)	Hypothetical	RA	RAS
<i>CTCF</i> → <i>TIPIN</i> (0.955)			
<i>FOXM1</i> → <i>CCNB2</i> (0.984)	Supportive		
<i>FOXM1</i> → <i>NCAPH</i> (0.997)	Hypothetical	RAS	RAS
<i>FOXM1</i> → <i>DLGAP5</i> (0.904)			

Table 21 continued from previous page

Regulations	Category	Original	Description
<i>FOXM1</i> → <i>UHRF1</i> (0.946)	Supportive	[261]	“ <i>FOXM1</i> and <i>UHRF1</i> are highly correlated in prostate cancer cells and tissues. <i>FOXM1</i> regulates CSCs by regulating <i>UHRF1</i> gene transcription in an <i>E2F</i> -independent manner and <i>FOXM1</i> protein directly binds to the <i>FKH</i> motifs at the <i>UHRF1</i> gene promoter” [261]
<i>NFYB</i> → <i>CENPF</i> (0.927) <i>NFYB</i> → <i>TTK</i> (0.998)	Hypothetical	RAS	RAS
<i>STAT1</i> → <i>FYN</i> (0.809)	Supportive	See Table 3 in [199]	[199]
<i>STAT1</i> → <i>LPP</i> (0.999)	Hypothetical	RAS	RAS

The table reports the list of interactions inferred by **BENIN** +orthology and confirmed by **BENIN** +expression, but that are missing in our gold-standard network. The 1st column reports the interactions. The number in parenthesis is the score returned by **BENIN** +expression. The 2nd reports the type of evidence about the interaction. It can be **supportive** if there is an explicit and direct experimental evidence demonstrating the presence of such a regulatory relationship; **predictive** if previously documented evidence implies the possibility of the regulatory interaction between the genes, but remains to be experimentally verified, or **hypothetical** if the biological knowledge for the regulation lacks so far. The 3rd column gives reference papers/works that support the evidence, if any. The 4th column gives more details from the paper that support the evidence of the interaction.

4.6.2 Discussion

BENIN can integrate several types of Prior knowledge Results presented in Section 4.6.1 demonstrate that BENIN can integrate a diverse type of prior knowledge to deal with the limitation of the data. We saw that the inclusion of prior knowledge data might increase BENIN performance. Moreover, we notice that integrating different prior data into BENIN may lead to different results. These results confirm the fact that different prior data may have different potential. A close observation of Table 18 shows that TFBS seems to be the most informative prior data. They store direct binding information. On the other hand, functional data seems to be less informative. However, when we average the score from all the prior data, we notice that BENIN performance may not be better than its performance with either TFBS or ChIP-seq data. This performance may result from the fact that we adopted the same parameters for all the data types; however, BENIN performance is very data-dependent. Moreover, integrating all the prior knowledge data through a simple average implies considering the different prior knowledge data are equally informative. The performance on the less informative prior will have a big impact on the combined performance. It may suggest a weighted integration.

The results reported in Section 4.6.1 shows the power of integrating regulatory interactions from closely related organisms into an organism of interest. In fact, including regulatory information from the mouse genome has allowed us to add around 300 edges and confirmed around 200 inferred links.

BENIN can infer both known and potential regulatory links In Section 4.6.1, we demonstrated that BENIN could retrieve links that are part of our gold-standard network. For example, when setting $\tau = 0.5$, we observe that BENIN infers around 1171 out of the 1463 interactions present in our gold-standard network. We demonstrated that BENIN was able to enrich the inferred network with new high scoring interactions that are biologically relevant. Some of these interactions were confirmed by orthology information transfer. They constitute interesting candidate interactions that will necessitate further investigation.

BENIN can scale to realistic problem In this chapter, we have demonstrated that BENIN can infer a realistic network. From the execution time presented in Section 17, we observe that BENIN runs in about 2h to infer a size 628 network integrating all the five different prior knowledge data.

4.7 Conclusion

In this chapter, we presented the result of applying BENIN to infer the GRN that controls the cell cycle of the HeLa cell line. We considered four different prior knowledge data: ChIP-seq, TFBS, KD expression, and functional annotation. We evaluated BENIN performances using our “gold-standard network.”

Comparing BENIN results when we integrate prior evidence of regulatory interactions to when we do not, we observe that prior knowledge data integration may improve BENIN performances. Testifying the importance of prior biological information. A close analysis of the returned edges shows that many inferred links were missing in our gold-standard network. BENIN can infer new interactions. Some of these interactions get support to some extent with the literature. In contrast, others that were not supported in literature may suggest potential research to confirm their existence. We also presented an extension of BENIN that integrates regulatory interaction from other organisms into the studied organism, through sequence orthology transfer. We tested this extension on the HeLa cell cycle using the mouse as our model organism. We were able to add more than 300 interactions that were or were not supported by the expression data. Some of these links were absent in our gold-standard network or marked as non-edges. These links may be subject to further investigation. Mainly those supported both by the expression and obtained through orthology transfer.

Even though our results on the HeLa cell line are encouraging, there is still much work to do. First of all, it will be interesting to consider other organisms for orthology mapping. We observed that our genes get orthologues into many other model organisms, such as the *zebrafish*, *rat*, or the *saccharomyces cerevisiae*. Some of these organisms (*zebrafish*, *rat* or *frog*) do not have an explicit database that stores their regulatory interactions or existing database lack this information. We need to automatically or manually scan the literature to get the list of potential regulatory

interactions of their GRN. A next extension will be to infer the GRN for several cell lines in human.

Chapter 5

Conclusion

5.1 Recap

The gene regulatory network, which designates the set of genes that interact together within the cell to control specific biological processes, is essential to understand how the cell functions and how it responds to its environment. The advancement in high-throughput instruments has allowed the generation of a high volume of a variety of omics data, that each may provide a complementary part of the picture of regulation.

This thesis had three goals for regulatory network inference:

1. Develop a method that integrates diverse data;
2. Develop a method that scales to handle a real dataset; and
3. Develop a method that can integrate information across organisms.

The thesis was that using `Elastic Net` regression for feature selection would lead to a method for network inference that met these goals, and also had a state-of-the-art performance.

Chapter 3 presented `BENIN` as a method that viewed network inference as a feature selection and applied adaptive `Elastic Net` regression to solve the feature selection problem. The adaptive `Elastic Net` allowed data integration. `BENIN` was evaluated on synthetic data from the `DREAM4` challenge, and with our own synthetic data. On the `DREAM4` dataset `BENIN` out-performed all `DREAM4` competitors on the size 100 subchallenge, and is also competitive with more recent state-of-the-art methods.

Chapter 4 applied **BENIN** to real data for the cell cycle of the Human HeLa cell line to demonstrate scalability and the integration of a range of types of data. We developed a gold standard network for evaluation purposes, and compared the effect of each prior and combination of priors on the predictive performance of **BENIN**. Furthermore, **BENIN** proposed new interactions. These were reviewed for support in the literature, as a preliminary validation of **BENIN**'s practicality, and whether the proposed new regulatory links might warrant further experimental investigation.

5.2 Contributions

This thesis addresses open challenges in computational reconstruction, or inference, of gene regulatory networks of performance, scale, and data integration.

The thesis presents a new algorithm **BENIN** that views network inference as feature selection to address issues of scale, that uses **Elastic Net** regression for improved performance, and adapts **Elastic Net** to integrate different types of biological data.

The **BENIN** algorithm is benchmarked on a synthetic dataset from the DREAM4 challenge, and on real expression data for the *Human* HeLa cell cycle. On the DREAM4 dataset **BENIN** out-performed all DREAM4 competitors on the size 100 sub-challenge, and is also competitive with more recent state-of-the-art methods. Moreover, on the HeLa cell cycle data, **BENIN** could infer known regulatory interactions and propose new interactions that warrant further experimental investigation.

The three contributions of the thesis are

1. The **BENIN** algorithm, addressing scale and performance issues, by viewing Network inference as the feature selection problem, and solving feature selection using **Elastic Net** regression;
2. **BENIN** addressing the integration of prior knowledge by adapting the **Elastic Net** regression technique; and
3. The application of **BENIN** to real data for the cell cycle of the *Human* HeLa cell line.

5.2.1 BENIN: Network Inference as Feature Selection using Elastic Net

In this thesis, we introduce **BENIN**: Biologically Enhanced Network INference. **BENIN** is a simple and intuitive inference method for integrating any prior knowledge with time-series expression data. **BENIN** states GRN inference as a feature selection problem: finding the direct regulators of each gene. It assumes that a target gene's expression profile is a linear function of its direct regulators' expression profiles. **BENIN** applies a regression technique called **Elastic Net** combined with a resampling technique to perform feature selection.

5.2.2 BENIN: Integration of Prior Knowledge

Data integration is a common technique to improve inference in computational biology. Yet the successful integration of a variety of types of data remains a challenge. In this work, we used a modified version of the **Elastic Net**: the adaptive **Elastic Net** to include prior knowledge. We developed ways to incorporate each prior into the mathematical formulation of the adaptive **Elastic Net** for each type of data:

- Knockout (KO) expression data;
- Knock-down (KD) expression data;
- ChIP-seq data;
- Functional annotations;
- Transcription factor binding sites; and
- Genome-wide location data.

The probabilistic framework for the integration is defined by the Bayes formula. This allows **BENIN** the possibility to control the impact of the prior on the model.

We have demonstrated that **BENIN** can integrate many types of data.

Moreover, **BENIN** allows the integration of regulatory information across species through the use of orthology.

Knockout(KO) and knock-down(KD) gene expression data are expression data measured in an organism where a transcription factor is made inoperative (KO expression data), or its expression is reduced (KD expression data). The data is integrated either through the z-score (for KO data) or a probabilistic framework (for KD data).

ChIP-seq data reports the regions in the genome where a specific transcription factor (TF) will physically bind to the DNA. We integrate ChIP-seq by a score measuring potential binding between each TF and each gene in the genome.

Functional annotation is given as a set of terms in the Gene Ontology (GO). We use a similarity measure of sets of terms to integrate functional annotation into BENIN.

Transcription factor binding sites are given as matrices storing binding specificities for a specific TF. We used this data to scan the region of interest in the genome. The result of the scanning process is integrated through a probabilistic framework into BENIN.

Genome-wide location data is given as p-values of physical interactions between a TF and a gene. The p-values are integrated into BENIN using a probabilistic framework.

5.2.3 Application of BENIN to Human Cell Cycle

To study BENIN on real data, where there was a range of data types available for integration, we applied BENIN to the cell cycle of the *Human* HeLa cell line.

This showed the effect on the performance of each prior and each combination of priors, and demonstrated BENIN at a realistic scale of the problem.

We integrate prior knowledge from transcription factor binding sites, knock-down gene expression data, functional annotation, and ChIP-seq data.

Data integration across organisms was demonstrated using orthology between genes of *mouse* and *human* to transfer information about regulatory links in *mouse* to the network inferred for the cell cycle of the *Human* HeLa cell line.

5.3 Limitations

One of the key limitations to BENIN, as with most regulatory network inference approaches, is the difficulty of distinguishing between direct and indirect regulatory links. This shows up clearly in the analysis of network motifs.

BENIN may be too simplistic in how it weighs each regulatory link. BENIN weighs the link by the number of bootstraps in which the link is selected during the feature selection by the adaptive Elastic Net. We notice that many links have the same weight, so the final rank does not show a clear preference between the links.

Our gold standard network in Chapter 5 did not distinguish cell lines, so it was not specific to the cell cycle of the *Human* HeLa cell line. We integrated information from different cell lines. The impact of this is unknown. It may be minor, as the data used for the network inference was for the *Human* HeLa cell line.

The use of orthology information was restricted to *mouse* in our case study. However, each of the model vertebrate organisms is a good candidate as a source of information, as indeed may be all the model organisms.

5.4 Future Work

Future work should definitely address the limitations above. Besides, our techniques should be made part of a widely-used suite of tools for the complete systems biology workflow, that works not simply with one organism at a time, but fully exploits orthology with model organisms, and accommodates recent needs for single-cell genomics, and microbial communities.

Our work takes a simple binary view of the regulatory link between transcription factors and target genes: it is either on or off. Even then, the algorithms have difficulty distinguishing direct regulation from indirect regulation. A model that considers positive (enhancement) and negative (repression) the behavior of regulation is the first step towards a more realistic model. Furthermore, transcription factors work in combinations, or as complexes, in the regulatory regions of a gene. This thesis did not consider the task of determining these so-called regulatory program of the transcription factors working together.

This thesis has considered only transcriptional regulation. This is but one of

the regulatory mechanisms in a cell. Systems biology, in the long term, will require dynamic models including all the regulatory mechanisms,

Bibliography

- [1] Munna L Agarwal, Archana Agarwal, William R Taylor, and George R Stark. P53 controls both the G2/M and the G1 cell cycle checkpoints and mediates reversible growth arrest in human fibroblasts. *Proceedings of the National Academy of Sciences*, 92(18):8493 – 8497, 1995.
- [2] Fadhl M Alakwaa. Modeling of gene regulatory networks: a literature review. *Journal of Computational Systems Biology*, 1(1):1, 2014.
- [3] Réka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947 – 4957, 2005.
- [4] Réka Albert and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378 – 382, Jul 2000. doi:10.1038/35019019.
- [5] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47 – 97, 2002. doi:10.1103/RevModPhys.74.47.
- [6] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25 – 29, 2000.
- [7] Claire Attwooll, Eros Lazzerini Denchi, and Kristian Helin. The E2F family: specific functions and overlapping interests. *The EMBO Journal*, 23(24):4709 – 4716, Nov 2004. doi:10.1038/sj.emboj.7600481.

- [8] M Madan Babu, Benjamin Lang, and L Aravind. Methods to reconstruct and compare transcriptional regulatory networks. In *Computational Systems Biology*, pages 163 – 180. Springer, 2009.
- [9] Timothy Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, 1994.
- [10] Timothy L. Bailey, James Johnson, Charles E. Grant, and William S. Noble. The MEME Suite. *Nucleic Acids Research*, 43(W1):W39 – W49, May 2015. doi:10.1093/nar/gkv416.
- [11] Albert-Laszlo Barabasi and Zoltan Oltvai. Network Biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101 – 113, 2004.
- [12] Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037 – 2060, 2013.
- [13] K. Beishline, C. M. Kelly, B. A. Olofsson, S. Koduri, J. Emrich, R. A. Greenberg, and J. Azizkhan-Clifford. Sp1 facilitates DNA Double-Strand Break Repair through a Nontranscriptional Mechanism. *Molecular and Cellular Biology*, 32(18):3790 – 3799, Jul 2012. doi:10.1128/mcb.00049-12.
- [14] Paolo Benatti, Diletta Dolfini, Alessandra Viganò, Maria Ravo, Alessandro Weisz, and Carol Imbriano. Specific inhibition of NF-Y subunits triggers different cell proliferation defects. *Nucleic Acids Research*, 39(13):5356 – 5368, Mar 2011. doi:10.1093/nar/gkr128.
- [15] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36 – D42, 2013.
- [16] Sharon Benzeno, Goutham Narla, Jorge Allina, George Z. Cheng, Helen L. Reeves, Michaela S. Banck, Joseph A. Odin, J. Alan Diehl, Doris Germain,

- and Scott L. Friedman. Cyclin-Dependent Kinase Inhibition by the KLF6 Tumor Suppressor Protein through Interaction with Cyclin D1. *Cancer Research*, 64(11):3885 – 3891, Jun 2004. doi:10.1158/0008-5472.can-03-2818.
- [17] A Bernard and AJ Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. In *Pacific Symposium on Biocomputing*, page 459, 2005.
- [18] Cosetta Bertoli, Jan M. Skotheim, and Robertus A. M. de Bruin. Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology*, 14(8):518 – 528, Jul 2013. doi:10.1038/nrm3629.
- [19] Adrian R. Black, Jennifer D. Black, and Jane Azizkhan-Clifford. Sp1 and krüppel-like factor family of transcription factors in cell growth regulation and cancer. *Journal of Cellular Physiology*, 188(2):143 – 160, Jun 2001. doi:10.1002/jcp.1111.
- [20] Alexandre Blais and Brian David Dynlacht. Constructing transcriptional regulatory networks. *Genes & Development*, 19(13):1499 – 1511, 2005.
- [21] Hamid Bolouri. *Computational modeling of gene regulatory networks: a primer*. World Scientific, 2008.
- [22] Hamid Bolouri. Modeling genomic regulatory networks with big data. *Trends in Genetics*, 30(5):182 – 191, 2014.
- [23] Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteynn Thorsson. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems biology data sets *de novo*. *Genome Biology*, 7(5):R36, 2006.
- [24] Luiz A Bovolenta, Marcio L Acencio, and Ney Lemke. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, 13(1):405, 2012. doi:10.1186/1471-2164-13-405.
- [25] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. GO::TermFinder — open source software

- for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710 – 3715, 2004.
- [26] Adrian P. Bracken, Marco Ciro, Andrea Cocito, and Kristian Helin. E2F target genes: unraveling the biology. *Trends in Biochemical Sciences*, 29(8):409 – 417, Aug 2004. doi:10.1016/j.tibs.2004.06.006.
- [27] Jacqueline Bromberg and James E Darnell. The role of STATs in transcriptional control and their impact on cellular function. *Oncogene*, 19(21):2468 – 2473, May 2000. doi:10.1038/sj.onc.1203476.
- [28] Youquan Bu, Yusuke Suenaga, Sayaka Ono, Tadayuki Koda, Fangzhou Song, Akira Nakagawara, and Toshinori Ozaki. Sp1-mediated transcriptional regulation of NFBD1/MDC1 plays a critical role in DNA damage response pathway. *Genes to Cells*, 13(1):53 – 66, Dec 2007. doi:10.1111/j.1365-2443.2007.01144.x.
- [29] Carol J Bult, Judith A Blake, Cynthia L Smith, James A Kadin, Joel E Richardson, A Anagnostopoulos, R Asabor, R M Baldarelli, J S Beal, S M Bello, and et al. Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801 – D806, Nov 2018. doi:10.1093/nar/gky1056.
- [30] Martha L Bulyk. Computational prediction of transcription factor binding site locations. *Genome Biology*, 5(1):1–11, 2003.
- [31] Christophe Bureau, Naima Hanoun, Jerome Torrisani, Jean-Pierre Vinel, Louis Buscail, and Pierre Cordelier. Expression and Function of Kruppel Like-Factors (KLF) in Carcinogenesis. *Current Genomics*, 10(5):353 – 360, Aug 2009. doi:10.2174/138920209788921010.
- [32] Valentina Calò, Manuela Migliavacca, Viviana Bazan, Marcella Macaluso, Maria Buscemi, Nicola Gebbia, and Antonio Russo. STAT proteins: From normal control of cellular events to tumorigenesis. *Journal of Cellular Physiology*, 197(2):157 – 168, Jul 2003. doi:10.1002/jcp.10364.

- [33] Irene Cantone, Lucia Marucci, Francesco Iorio, Maria Aurelia Ricci, Vincenzo Belcastro, Mukesh Bansal, Stefania Santini, Mario Di Bernardo, Diego Di Bernardo, and Maria Pia Cosma. A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172 – 181, 2009.
- [34] Angelo Canty and B. D. Ripley. boot: Bootstrap R (S-Plus) Functions. <https://cran.r-project.org/web/packages/boot/>, 2017. R package version 1.3-20.
- [35] Marc Carlson. org.Hs.eg.db: Genome wide annotation for Human. <https://bioconductor.org/packages/org.Hs.eg.db/>, 2018. R package version 3.7.0.
- [36] T Chen, HL He, and GM Church. Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing*, pages 29 – 40, 1999.
- [37] Xi Chen, Gerd A Müller, Marianne Quaas, Martin Fischer, Namshik Han, Benjamin Stutchbury, Andrew D Sharrocks, and Kurt Engeland. The forkhead transcription factor FOXM1 controls cell cycle-dependent gene expression through an atypical chromatin binding mechanism. *Molecular and Cellular Biology*, 33(2):227 – 236, 2013.
- [38] David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5(Oct):1287 – 1330, 2004.
- [39] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1):103 – 112, 2005.
- [40] Robert Clarke, Habtom Resson, Antai Wang, Jianhua Xuan, Minetta Liu, Edmund Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37 – 49, 2008.

- [41] Thomas Cokelaer, Mukesh Bansal, Christopher Bare, Erhan Bilal, Brian M Bot, Elias Chaibub Neto, Federica Eduati, Alberto de la Fuente, Mehmet Gönen, Steven M Hill, et al. DREAMTools: a Python package for scoring collaborative challenges. *F1000Research*, 4, 2015.
- [42] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(90001):258D – 261, Jan 2004. doi:10.1093/nar/gkh036.
- [43] UniProt Consortium et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158 – D169, 2017.
- [44] Athel Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021, 1985.
- [45] Robert H. Costa. FoxM1 dances with mitosis. *Nature Cell Biology*, 7(2):108 – 110, Feb 2005. doi:10.1038/ncb0205-108.
- [46] Modan K Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. In *BMC Bioinformatics*, volume 8, page S21. Springer, 2007.
- [47] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233 – 240, 2006.
- [48] W. Davis. Reciprocal regulation of expression of the human adenosine 5'-triphosphate binding cassette, sub-family A, transporter 2 (ABCA2) promoter by the early growth response-1 (EGR-1) and Sp-family transcription factors. *Nucleic Acids Research*, 31(3):1097 – 1107, Feb 2003. doi:10.1093/nar/gkg192.
- [49] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717 – 729, 2010.
- [50] James DeGregori. The genetics of the E2F family of transcription factors: shared functions and unique roles. *Biochimica et Biophysica Acta (BBA)* -

- Reviews on Cancer*, 1602(2):131 – 150, 2002. doi:[https://doi.org/10.1016/S0304-419X\(02\)00051-3](https://doi.org/10.1016/S0304-419X(02)00051-3).
- [51] James DeGregori, Timothy Kowalik, and Joseph R Nevins. Cellular targets for activation by the E2F1 transcription factor include DNA synthesis - and G1/S - regulatory genes. *Molecular and Cellular Biology*, 15(8):4215 – 4224, 1995.
- [52] Chu-Xia Deng. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Research*, 34(5):1416 – 1426, 2006.
- [53] Emmanuelle Deniaud, Joël Baguet, Roxane Chalard, Bariza Blanquier, Lilia Brinza, Julien Meunier, Marie-Cécile Michallet, Aurélie Laugraud, Claudette Ah-Soon, Anne Wierinckx, and et al. Overexpression of Transcription Factor Sp1 Leads to Gene Expression Perturbations and Cell Cycle Inhibition. *PLoS ONE*, 4(9):e7035, Sep 2009. doi:10.1371/journal.pone.0007035.
- [54] Edward R Dougherty and Ilya Shmulevich. On the limitations of biological knowledge. *Current Genomics*, 13(7):574, 2012.
- [55] Kevin T Duffy, Mary Frances McAleer, William R Davidson, Laszlo Kari, Csaba Kari, Chang-Gong Liu, Steven A Farber, Keith C Cheng, Jason R Mest, Eric Wickstrom, et al. Coordinate control of cell cycle regulatory genes in zebrafish development tested by cyclin D1 knockdown with morpholino phosphorodi-amidates and hydroxypropyl-phosphono peptide nucleic acids. *Nucleic Acids Research*, 33(15):4914 – 4921, 2005.
- [56] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439 – 3440, 2005.
- [57] Steffen Durinck, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4:1184 – 1191, 2009. URL: <https://bioconductor.org/packages/biomaRt/>.

- [58] Stéphanie Dutertre, Martine Cazales, Muriel Quaranta, Carine Froment, Valerie Trabut, Christine Dozier, Gladys Mirey, Jean-Pierre Bouché, Nathalie Theis-Febvre, Estelle Schmitt, et al. Phosphorylation of CDC25B by Aurora-A at the centrosome contributes to the G2 – M transition. *Journal of Cell Science*, 117(12):2523 – 2531, 2004.
- [59] Ron Edgar, Michael Domrachev, and Alex Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207 – 210, 2002.
- [60] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pages 569 – 593. Springer, 1992.
- [61] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004.
- [62] R. Elkon. Genome-Wide *In Silico* Identification of Transcriptional Regulators Controlling the Cell Cycle in Human Cells. *Genome Research*, 13(5):773 – 780, May 2003. doi:10.1101/gr.947203.
- [63] Frank Emmert-Streib, Galina Glazko, Ricardo De Matos Simoes, et al. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in Genetics*, 3:8, 2012.
- [64] Douglas H Erwin and Eric H Davidson. The evolution of hierarchical gene regulatory networks. *Nature Reviews Genetics*, 10(2):141 – 148, 2009.
- [65] Ahmed Essaghir and Jean-Baptiste Demoulin. A minimal connected network of transcription factors regulated in human tumors and its application to the quest for universal cancer biomarkers. *PloS One*, 7(6), 2012.
- [66] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8, 2007.

- [67] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348 – 1360, Dec 2001. doi:10.1198/016214501753382273.
- [68] Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10(Sept):2013 – 2038, 2009.
- [69] Donald Farrar and Robert Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92 – 107, 1967.
- [70] Chenchen Feng, Chao Song, Yuejuan Liu, Fengcui Qian, Yu Gao, Ziyu Ning, Qiuyu Wang, Yong Jiang, Yanyu Li, Meng Li, et al. KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Research*, 48(D1):D93 – D100, 2020.
- [71] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and Elastic Net regularized generalized linear models. *e version*, 1(4), 2009. URL: <https://glmnet.stanford.edu>.
- [72] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- [73] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799 – 805, 2004.
- [74] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe’er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601 – 620, 2000.
- [75] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. *ArXiv Preprint ArXiv:1301.7374*, 2013.
- [76] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150 – 3152, 2012.

- [77] André Fujita, Joao R Sato, Humberto M Garay-Malpartida, Rui Yamaguchi, Satoru Miyano, Mari C Sogayar, and Carlos E Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(1):1 – 11, 2007.
- [78] Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363 – 1375, Jul 2019. doi:10.1101/gr.240663.118.
- [79] Pascale Gaudet, Michael S Livstone, Suzanna E Lewis, and Paul D Thomas. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, 12(5):449 – 462, 2011.
- [80] Florian Geier, Jens Timmer, and Christian Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Systems Biology*, 1(1):11, 2007.
- [81] Pierre Geurts et al. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Scientific Reports*, 8(1):1 – 12, 2018.
- [82] Raluca Gordân, Alexander J Hartemink, and Martha L Bulyk. Distinguishing direct versus indirect transcription factor – DNA interactions. *Genome Research*, 19(11):2090 – 2100, 2009.
- [83] S Gordon, G Akopyan, H Garban, and B Bonavida. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene*, 25(8):1125 – 1142, Nov 2005. doi:10.1038/sj.onc.1209080.
- [84] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329 – 352, 1980.
- [85] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017 – 1018, Feb 2011. doi:10.1093/bioinformatics/btr064.
- [86] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky,

- Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569 – 576, 2015.
- [87] Edgar Grinstein, Franziska Jundt, Inge Weinert, Peter Wernet, and Hans-Dieter Royer. Sp1 as G1 cell cycle phase specific transcription factor in epithelial cells. *Oncogene*, 21(10):1485 – 1492, Feb 2002. doi:10.1038/sj.onc.1205211.
- [88] A. Gurtner, P. Fuschi, F. Martelli, I. Manni, S. Artuso, G. Simonte, V. Ambrosino, A. Antonini, V. Folgiero, R. Falcioni, and et al. Transcription Factor NF-Y Induces Apoptosis in Cells Expressing Wild-Type p53 through E2F1 Up-regulation and p53 Activation. *Cancer Research*, 70(23):9711 – 9720, Oct 2010. doi:10.1158/0008-5472.can-10-0721.
- [89] Aymone Gurtner, Paola Fuschi, Fabio Martelli, Isabella Manni, Simona Artuso, Giacomina Simonte, Valeria Ambrosino, Annalisa Antonini, Valentina Folgiero, Rita Falcioni, et al. Transcription factor NF-Y induces apoptosis in cells expressing wild-type p53 through E2F1 upregulation and p53 activation. *Cancer Research*, 70(23):9711 – 9720, 2010.
- [90] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1/3):389 – 422, 2002. doi:10.1023/a:1012487302797.
- [91] Peter Hall, James Stephen Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427 – 444, 2005.
- [92] Timothy C Hallstrom, Seiichi Mori, and Joseph R Nevins. An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. *Cancer Cell*, 13(1):11 – 22, 2008.
- [93] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasmom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, and et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380 – D386, Oct 2017. doi:10.1093/nar/gkx1013.

- [94] Peter Harris, Sue Nagy, Nicholas Vardaxis, and Nick Vardaxis. *Mosby's dictionary of medicine, nursing and health professions*. Elsevier Australia, 2009.
- [95] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402:C47 – C52, 1999.
- [96] Fatma A Hashim, Mai S Mabrouk, and Walid Al-Atabany. Review of different sequence motif finding algorithms. *Avicenna Journal of Medical Biotechnology*, 11(2):130, 2019.
- [97] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. TIGRESS: trustful inference of gene regulation using stability selection. *BMC Systems Biology*, 6(1):145, 2012.
- [98] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models — a review. *Biosystems*, 96(1):86 – 103, 2009.
- [99] J Michael Herry, Caroline Adler, Catherine Ball, Stephen A Chervitz, Selina S Dwight, Erich T Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, et al. SGD: *Saccharomyces* genome database. *Nucleic Acids Research*, 26(1):73 – 79, 1998.
- [100] Veronica F Hinman, Kristen A Yankura, and Brenna S McCauley. Evolution of gene regulatory network architectures: examples of subcircuit conservation and plasticity between classes of echinoderms. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1789(4):326 – 332, 2009.
- [101] Arthur Hoerl and Robert Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55 – 67, 1970.
- [102] GS Hongyi Li and Maddala. Bootstrapping time series models. *Econometric Reviews*, 15(2):115 – 158, 1996.
- [103] Sean D Hooper, Stephanie Boué, Roland Krause, Lars J Jensen, Christopher E Mason, Murad Ghanim, Kevin P White, Eileen EM Furlong, and Peer Bork. Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis. *Molecular Systems Biology*, 3(1):72, 2007.

- [104] Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research*, 47(D1):D33 – D38, Sep 2018. doi:10.1093/nar/gky822.
- [105] Jianjun Hu, Bin Li, and Daisuke Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33(15):4899 – 4913, 2005.
- [106] Jianjun Hu, Yifeng D Yang, and Daisuke Kihara. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC Bioinformatics*, 7(1):342, 2006.
- [107] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1 – 13, 2008.
- [108] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44 – 57, 2009.
- [109] Xiangwei Huang, Xia Li, and Bin Guo. KLF6 Induces Apoptosis in Prostate Cancer Cells through Up-regulation of ATF3. *Journal of Biological Chemistry*, 283(44):29795 – 29801, Aug 2008. doi:10.1074/jbc.m802515200.
- [110] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian Von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, 34(8):2115 – 2122, 2017.
- [111] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309 – D314, 2019.

- [112] Jason Hughes, Preston Estep, Saeed Tavazoie, and George Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296(5):1205 – 1214, 2000.
- [113] Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, and et al. Ensembl variation resources. *Database*, 2018, Jan 2018. doi:10.1093/database/bay119.
- [114] C. Huttenhower, E. M. Haley, M. A. Hibbs, V. Dumeaux, D. R. Barrett, H. A. Collier, and O. G. Troyanskaya. Exploring the human genome with functional maps. *Genome Research*, 19(6):1093 – 1106, Feb 2009. doi:10.1101/gr.082214.108.
- [115] Vân Anh Huynh-Thu and Guido Sanguinetti. Gene Regulatory Network Inference: An Introductory Survey. *Gene Regulatory Networks*, pages 1 – 23, Dec 2018. doi:10.1007/978-1-4939-8882-2_1.
- [116] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS One*, 5(9):e12776, 2010.
- [117] Ian B Jeffery, Desmond G Higgins, and Aedín C Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7(1), Jul 2006. doi:10.1186/1471-2105-7-359.
- [118] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, et al. STRING 8 — a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl_1):D412 – D416, 2008.
- [119] David G Johnson. Role of E2F in cell cycle control and cancer. *Frontiers in Bioscience*, 3(4):d447 – 458, 1998. doi:10.2741/a291.

- [120] Szymon Jozefczuk, Sebastian Klie, Gareth Catchpole, Jędrzej Szymanski, Alvaro Cuadros-Inostroza, Dirk Steinhauser, Joachim Selbig, and Lothar Willmitzer. Metabolomic and transcriptomic stress response of *Escherichia coli*. *Molecular Systems Biology*, 6(1):364, 2010.
- [121] LM Julian, Y Liu, CA Pakenham, D Dugal-Tessier, V Ruzhynsky, S Bae, SY Tsai, G Leone, RS Slack, and A Blais. Tissue-specific targeting of cell fate regulatory genes by E2F factors. *Cell Death & Differentiation*, 23(4):565 – 575, 2016.
- [122] Mun-Su Jung, Jeanho Yun, Hee-Don Chae, Jeong-Min Kim, Sun-Chang Kim, Tae-Saeng Choi, and Deug Y Shin. p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of NF-Y transcription factor. *Oncogene*, 20(41):5818 – 5825, Sep 2001. doi:10.1038/sj.onc.1204748.
- [123] Stephanie Kamgnia and Gregory Butler. BENIN: Combining Knockout Data with Time Series Gene Expression Data for the Gene Regulatory Network Inference. In *Proceedings of the Tenth International Conference on Computational Systems Biology and Bioinformatics*, CSBio '19. ACM, 2019. doi:10.1145/3365953.3365955.
- [124] M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27 – 30, Jan 2000. doi:10.1093/nar/28.1.27.
- [125] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947 – 1951, Sep 2019. doi:10.1002/pro.3715.
- [126] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590 – D595, Oct 2018. doi:10.1093/nar/gky962.
- [127] Ari Kassardjian, Raed Rizkallah, Sarah Riman, Samuel H. Renfro, Karen E. Alexander, and Myra M. Hurt. The Transcription Factor YY1 Is a Novel Substrate for Aurora B Kinase at G2/M Transition of the Cell Cycle. *PLoS ONE*, 7(11):e50645, Nov 2012. doi:10.1371/journal.pone.0050645.

- [128] Michael B. Kastan, Christine E. Canman, and Christopher J. Leonard. P53, cell cycle control and apoptosis: Implications for cancer. *Cancer and Metastasis Reviews*, 14(1):3 – 15, Mar 1995. doi:10.1007/bf00690207.
- [129] Mary M. Kavurma, Fernando S. Santiago, Emanuela Bonfoco, and Levon M. Khachigian. Sp1 Phosphorylation Regulates Apoptosis via Extracellular FasL-Fas Engagement. *Journal of Biological Chemistry*, 276(7):4964 – 4971, Oct 2000. doi:10.1074/jbc.m009251200.
- [130] Alexei E Kazakov, Michael J Cipriano, Pavel S Novichkov, Simon Minovitsky, Dmitry V Vinogradov, Adam Arkin, Andrey A Mironov, Mikhail S Gelfand, and Inna Dubchak. RegTransBase — a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Research*, 35(suppl 1):D407 – D412, 2007.
- [131] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996 – 1006, May 2002. doi:10.1101/gr.229102.
- [132] Hun Sik Kim and Myung-Shik Lee. STAT1 as a key modulator of cell death. *Cellular Signalling*, 19(3):454 – 465, 2007.
- [133] Nikolay A. Kolchanov, Olga A. Podkolodnaya, Elena A. Ananko, Elena V. Ignatieva, Irina L. Stepanenko, Olga V. Kel-Margoulis, Alexander E. Kel, Tatyana I. Merkulova, TN Goryachkovskaya, TV Busygina, et al. Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Research*, 28(1):298 – 301, 2000.
- [134] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [135] Arjun Krishnan, Ran Zhang, Victoria Yao, Chandra L Theesfeld, Aaron K Wong, Alicja Tadych, Natalia Volfovsky, Alan Packer, Alex Lash, and Olga G Troyanskaya. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 19(11):1454 – 1462, 2016.

- [136] Robert Küffner, Tobias Petri, Lukas Windhager, and Ralf Zimmer. Petri nets with fuzzy logic (PNFL): reverse engineering and parametrization. *PLoS One*, 5(9):e12807, 2010.
- [137] Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2013.
- [138] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008. URL: <http://bioconductor.org/biocLite.R>.
- [139] Jamila Laoukili, Matthijs RH Kooistra, Alexandra Brás, Jos Kauw, Ron M Kerkhoven, Ashby Morrison, Hans Clevers, and René H Medema. FOXM1 is required for execution of the mitotic programme and chromosome stability. *Nature Cell Biology*, 7(2):126 – 136, 2005.
- [140] Ritwik K Layek, Aniruddha Datta, and Edward R Dougherty. From biological pathways to regulatory networks. *Molecular BioSystems*, 7(3):843 – 851, 2011.
- [141] Sophie Lèbre. Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.
- [142] B. Lemon. Orchestrated response: a symphony of transcription factors for gene control. *Genes and Development*, 14(20):2551 – 2569, Oct 2000. doi:10.1101/gad.831000.
- [143] Robert Lesurf, Kelsy C Cotto, Grace Wang, Malachi Griffith, Katayoon Kasarian, Steven JM Jones, Stephen B Montgomery, Obi L Griffith, and Open Regulatory Annotation Consortium. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Research*, 44(D1):D126–D132, 2016.
- [144] Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147 – 151, Jul 2003. doi:10.1038/nature01763.
- [145] Weizhong Li and Adam Godzik. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658 – 1659, 2006.

- [146] Shoudan Liang, Stefanie Fuhrman, Roland Somogyi, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, volume 3, pages 18 – 29, 1998.
- [147] Néhémy Lim, Yasin Şenbabaoğlu, George Michailidis, and Florence d’Alché Buc. OKVAR-Boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks. *Bioinformatics*, 29(11):1416 – 1423, 2013.
- [148] Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, and Luonan Chen. Inference of gene regulatory network based on local bayesian networks. *PLoS Computational Biology*, 12(8):e1005024, 2016.
- [149] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015, 2015.
- [150] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), Dec 2014. doi:10.1186/s13059-014-0550-8.
- [151] Luong Linh Ly, Hideki Yoshida, and Masamitsu Yamaguchi. Nuclear transcription factor Y and its roles in cellular processes related to human disease. *American Journal of Cancer Research*, 3(4):339, 2013.
- [152] David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*, 89(428):1535 – 1546, 1994.
- [153] Patricia A. Madureira, Rana Varshochi, Demetra Constantinidou, Richard E. Francis, R. Charles Coombes, Kwok-Ming Yao, and Eric W.-F. Lam. The Forkhead Box M1 Protein Regulates the Transcription of the Estrogen Receptor α in Breast Cancer Cells. *Journal of Biological Chemistry*, 281(35):25167 – 25176, Jun 2006. doi:10.1074/jbc.m603906200.

- [154] Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448 – 3449, 2005.
- [155] Isabella Manni, Giuseppina Mazzaro, Aymone Gurtner, Roberto Mantovani, Ulrike Haugwitz, Karen Krause, Kurt Engeland, Ada Sacchi, Silvia Soddu, and Giulia Piaggio. NF-Y Mediates the Transcriptional Inhibition of the cyclin B1, cyclin B2, and cdc25C Promoters upon Induced G2 Arrest. *Journal of Biological Chemistry*, 276(8):5570 – 5576, Nov 2000. doi:10.1074/jbc.m006052200.
- [156] Daniel Marbach. *Evolutionary reverse engineering of gene networks*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2009.
- [157] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796 – 804, 2012.
- [158] Daniel Marbach, Robert Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286 – 6291, 2010.
- [159] Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. ARACNE : an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.
- [160] Donald W Marquardt and Ronald D Snee. Ridge regression in practice. *The American Statistician*, 29(1):3 – 20, 1975.
- [161] John R Masters. HeLa cells 50 years on: the good, the bad and the ugly. *Nature Reviews Cancer*, 2(4):315 – 319, 2002.
- [162] Veia Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E. Kel, Olga V. Kel-Margoulis,

- et al. TRANSFAC : transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374 – 378, 2003.
- [163] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62 – 69, 2012.
- [164] Pedro Mendes, Wei Sha, and Keying Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(suppl_2):ii122 – ii129, 2003.
- [165] Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020.
- [166] Pedro T Monteiro, Jorge Oliveira, Pedro Pais, Miguel Antunes, Margarida Palma, Mafalda Cavalheiro, Mónica Galocha, Cláudia P Godinho, Luís C Martins, Nuno Bourbon, and et al. YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Research*, 48(D1):D642 – D649, Oct 2019. doi:10.1093/nar/gkz859.
- [167] Kevin Patrick Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.
- [168] Isabel Nepomuceno-Chamorro, Jesus Aguilar-Ruiz, and Jose Riquelme. Inferring gene regression networks with model trees. *BMC Bioinformatics*, 11(1):517, 2010.
- [169] Joseph O Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6(S2), May 2012. doi:10.1186/1753-6561-6-s2-s10.

- [170] Carlota Oleaga, Sabine Welten, Audrey Belloc, Anna Solé, Laura Rodriguez, Núria Mencia, Elisabet Selga, Alicia Tapias, Veronique Noé, and Carlos J Ciudad. Identification of novel Sp1 targets involved in proliferation and cancer by functional genomics. *Biochemical Pharmacology*, 84(12):1581 – 1591, 2012.
- [171] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The MIntAct project — IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358 – D363, 2014.
- [172] David A Orlando, Charles Y Lin, Allister Bernard, Jean Y Wang, Joshua ES Socolar, Edwin S Iversen, Alexander J Hartemink, and Steven B Haase. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, 453(7197):944 – 947, 2008.
- [173] Bernhard Ø Palsson. *Systems Biology: simulation of dynamic network states*. Cambridge University Press, 2011.
- [174] Ranjit Kumar Paul. Multicollinearity: Causes, Effects and Remedies. *IASRI, New Delhi*, 2006.
- [175] Giulio Pavesi, Paolo Mereghetti, Giancarlo Mauri, and Graziano Pesole. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, 32(suppl 2):W199 – W203, 2004.
- [176] William R. Pearson. An Introduction to Sequence Similarity (“Homology”) Searching. *Current Protocols in Bioinformatics*, 42(1):3.1.1 – 3.1.8, Jun 2013. doi:10.1002/0471250953.bi0301s42.
- [177] Baikang Pei and Dong-Guk Shin. Reconstruction of biological networks by incorporating prior knowledge into Bayesian network models. *Journal of Computational Biology*, 19(12):1324 – 1334, 2012.
- [178] Francesca Petralia, Pei Wang, Jialiang Yang, and Zhidong Tu. Integrative random forest for gene regulatory network inference. *Bioinformatics*, 31(12):i197 – i205, 2015.

- [179] Andrea Pinna, Nicola Soranzo, and Alberto De La Fuente. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PloS One*, 5(10):e12912, 2010.
- [180] Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303 – 1313, 1994.
- [181] Elodie Portales-Casamar, David Arenillas, Jonathan Lim, Magdalena I Swanson, Steven Jiang, Anthony McCallum, Stefan Kirov, and Wyeth W Wasserman. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research*, 37(suppl_1):D54 – D60, 2009.
- [182] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS One*, 5(2):e9202, 2010.
- [183] Jonathan M Raser and Erin K O’Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010 – 2013, 2005.
- [184] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191 – W198, May 2019. doi:10.1093/nar/gkz369.
- [185] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551 – 1555, 2002.
- [186] John E Reid and Lorenz Wernisch. STEME: efficient EM to find motifs in large data sets. *Nucleic Acids Research*, 39(18):e126 – e126, 2011.
- [187] Bing Ren, Hieu Cam, Yasuhiko Takahashi, Thomas Volkert, Jolyon Terragni, Richard A Young, and Brian David Dynlacht. E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints. *Genes & Development*, 16(2):245 – 256, 2002.

- [188] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47 – e47, Jan 2015. doi:10.1093/nar/gkv007.
- [189] Karl J. Roberts. Index of kroberts Lecture. <http://academic.pgcc.edu/kroberts/Lecture/Chapter%207/regulation.html>, Sept 2006.
- [190] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651 – 657, 2007.
- [191] Dmitry A Rodionov. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chemical Reviews*, 107(8):3467 – 3497, 2007.
- [192] Michal Ronen, Revital Rosenberg, Boris I Shraiman, and Uri Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences*, 99(16):10555 – 10560, 2002.
- [193] Delphine Ropers, Hidde De Jong, and Hans Geiselmann. Mathematical modeling of genetic regulatory networks: Stress responses in it *Escherichia coli*. In P. Fu, M. Letterich, and S. Panke, editors, *Systems Biology and Synthetic Biology*, pages 235–271. John Wiley & Sons, 2009.
- [194] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507 – 2517, 2007.
- [195] Fikret Sahin and Todd L. Sladek. E2F-1 has dual roles depending on the cell cycle. *International Journal of Biological Sciences*, pages 116 – 128, 2010. doi:10.7150/ijbs.6.116.
- [196] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432, 2015.

- [197] Francesco Sambo, Barbara Di Camillo, and Gianna Toffolo. CNET: an algorithm for reverse engineering of causal gene networks. *NETTAB2008, Varenna, Italy*, 2008.
- [198] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91D – 94, Jan 2004. doi:10.1093/nar/gkh012.
- [199] Jun-ichi Satoh and Hiroko Tabunoki. A comprehensive profile of CHIP-Seq-based STAT1 target genes suggests the complexity of STAT1-mediated gene regulatory mechanisms. *Gene Regulation and Systems Biology*, 7:GRSB – S11433, 2013.
- [200] Eric Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218 – 223, 2009.
- [201] Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754 – 764, 2005.
- [202] Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: *in silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263 – 2270, 2011.
- [203] William F Scherer, Jerome T Syverton, and George O Gey. Studies on the propagation *in vitro* of poliomyelitis viruses: IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. *The Journal of Experimental Medicine*, 97(5):695 – 710, 1953.
- [204] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1), Jun 2006. doi:10.1186/1471-2105-7-302.

- [205] Thomas D Schneider and R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097 – 6100, 1990.
- [206] Eran Segal, Yoseph Barash, Itamar Simon, Nir Friedman, and Daphne Koller. From promoter sequence to expression: a probabilistic framework. In *Proceedings of the Sixth Annual International Conference on Computational Biology*, pages 263 – 272. ACM, 2002.
- [207] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe’er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166, 2003.
- [208] Eran Segal, R Yelensky, and Daphne Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(suppl_1):i273 – i282, 2003.
- [209] Yang Shi, Jeng-Shin Lee, and Katherine M. Galvin. Everything you have ever wanted to know about Yin Yang 1..... *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1332(2):F49 – F66, Apr 1997. doi:10.1016/s0304-419x(96)00044-3.
- [210] Teppei Shimamura, Seiya Imoto, Rui Yamaguchi, André Fujita, Masao Nagasaki, and Satoru Miyano. Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3(1):41, 2009.
- [211] Ali Shojaie and George Michailidis. Discovering graphical Granger causality using the truncating LASSO penalty. *Bioinformatics*, 26(18):i517 – i523, 2010.
- [212] Christopher A Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1 – 48, 1980.
- [213] Saurabh Sinha and Martin Tompa. YMF : a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 31(13):3586 – 3588, 2003.

- [214] Steven M Smith, Daniel C Fulton, Tansy Chia, David Thorneycroft, Andrew Chapple, Hannah Dunstan, Christopher Hylton, Samuel C Zeeman, and Alison M Smith. Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in *Arabidopsis* leaves. *Plant Physiology*, 136(1):2687 – 2699, 2004.
- [215] Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1 – 25, Jan 2004. doi:10.2202/1544-6115.1027.
- [216] Artem Sokolov, Daniel E Carlin, Evan O Paull, Robert Baertsch, and Joshua M Stuart. Pathway-based genomics prediction using generalized elastic net. *PLoS Computational Biology*, 12(3), 2016.
- [217] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle – regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273 – 3297, 1998.
- [218] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613 – 626, 2012.
- [219] J Stanelle, H Tu-Rapp, and B M Pützer. A novel mitochondrial protein DIP mediates E2F1-induced apoptosis independently of p53. *Cell Death and Differentiation*, 12(4):347 – 357, Nov 2004. doi:10.1038/sj.cdd.4401532.
- [220] Craig Stevens, Linda Smith, and Nicholas B. La Thangue. Chk2 activates E2F-1 in response to DNA damage. *Nature Cell Biology*, 5(5):401 – 409, Apr 2003. doi:10.1038/ncb974.
- [221] James H Stock and Mark W Watson. Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101 – 115, 2001.

- [222] Gary D Stormo, Thomas D Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the “Perceptron” algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9):2997 – 3011, 1982.
- [223] Matthew E Studham, Andreas Tjärnberg, Torbjörn EM Nordling, Sven Nelander, and Erik LL Sonnhammer. Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, 30(12):i130 – i138, 2014.
- [224] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. STRING v11: protein – protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607 – D613, 2019.
- [225] Matthew J. Thomas and Edward Seto. Unlocking the mechanisms of transcription factor YY1: are chromatin modifying enzymes the key? *Gene*, 236(2):197 – 208, Aug 1999. doi:10.1016/s0378-1119(99)00261-9.
- [226] Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, and Jacques van Helden. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, 40(4):e31 – e31, 2012.
- [227] Dawn Thompson, Aviv Regev, and Sushmita Roy. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annual Review of Cell and Developmental Biology*, 31:399 – 428, 2015.
- [228] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 267 – 288, 1996.
- [229] Hannah Tipney and Lawrence Hunter. An introduction to effective use of enrichment analysis software. *Human Genomics*, 4(3):202, 2010.
- [230] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137 – 144, 2005.

- [231] Geradrd J Tortora and SR Grabowski. *Principles of anatomy and physiology, 2000*. Wiley; 9th edition (2000), 2011.
- [232] Benjamin Jean-Marie Tremblay. *universalmotif: Import, Modify, and Export Motifs with R*, 2019. R package version 1.0.22. URL: <https://github.com/bjmt/universalmotif>.
- [233] Jeffrey M Trimarchi and Jacqueline A Lees. Sibling rivalry in the E2F family. *Nature Reviews Molecular Cell Biology*, 3(1):11 – 20, 2002.
- [234] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475 – 494, 2001.
- [235] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116 – 5121, Apr 2001. doi:10.1073/pnas.091062498.
- [236] Chiharu Uchida, Toshiaki Oda, Tsuyoshi Sugiyama, Sunao Otani, Masatoshi Kitagawa, and Arata Ichiyama. The Role of Sp1 and AP-2 in Basal and Protein Kinase A-induced Expression of Mitochondrial Serine:Pyruvate Aminotransferase in Hepatocytes. *Journal of Biological Chemistry*, 277(42):39082 – 39092, Aug 2002. doi:10.1074/jbc.m201380200.
- [237] UniProt Consortium et al. The universal protein resource (UniProt). *Nucleic Acids Research*, 36(suppl 1):D190 – D195, 2008.
- [238] Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.
- [239] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 2013.
- [240] Eberhard O Voit. *A first course in Systems Biology*. Garland Science, 2012.

- [241] Aijin Wang, Robin Schneider-Broussard, Addanki P. Kumar, Michael C. MacLeod, and David G. Johnson. Regulation of BRCA1 Expression by the Rb-E2F Pathway. *Journal of Biological Chemistry*, 275(6):4532 – 4536, Feb 2000. doi:10.1074/jbc.275.6.4532.
- [242] I-C. Wang, Y.-J. Chen, D. Hughes, V. Petrovic, M. L. Major, H. J. Park, Y. Tan, T. Ackerson, and R. H. Costa. Forkhead Box M1 Regulates the Transcriptional Network of Genes Essential for Mitotic Progression and Genes Encoding the SCF Skp2-Cks1 Ubiquitin Ligase. *Molecular and Cellular Biology*, 25(24):10875 – 10894, Nov 2005. doi:10.1128/mcb.25.24.10875-10894.2005.
- [243] Su Wang, Hanfei Sun, Jian Ma, Chongzhi Zang, Chenfei Wang, Juan Wang, Qianzi Tang, Clifford A Meyer, Yong Zhang, and X Shirley Liu. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nature Protocols*, 8(12):2502 – 2515, Nov 2013. doi:10.1038/nprot.2013.150.
- [244] Y Wang, O Deng, Z Feng, Z Du, X Xiong, J Lai, X Yang, M Xu, H Wang, D Taylor, and et al. RNF126 promotes homologous recombination via regulation of E2F1-mediated BRCA1 expression. *Oncogene*, 35(11):1363 – 1372, Aug 2015. doi:10.1038/onc.2015.198.
- [245] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, and et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431 – 1443, Sep 2014. doi:10.1016/j.cell.2014.08.009.
- [246] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462 – 11467, 2001.
- [247] Michael L. Whitfield, Gavin Sherlock, Alok J. Saldanha, John I. Murray, Catherine A. Ball, Karen E. Alexander, John C. Matese, Charles M. Perou,

- Myra M. Hurt, Patrick O. Brown, and et al. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Molecular Biology of the Cell*, 13(6):1977 – 2000, Jun 2002. doi:10.1091/mbc.02-02-0030.
- [248] Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, 5(11):R92, 2004.
- [249] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics*, 6(1):227, 2005.
- [250] Stephanie Kamgnia Wonkap and Gregory Butler. BENIN: Biologically enhanced network inference. *Journal of Bioinformatics and Computational Biology*, 18(03):2040007, 2020.
- [251] Diane R. Wonsey and Maximillian T. Follettie. Loss of the Forkhead Transcription Factor FoxM1 Causes Centrosome Amplification and Mitotic Catastrophe. *Cancer Research*, 65(12):5181 – 5189, Jun 2005. doi:10.1158/0008-5472.can-04-4059.
- [252] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3 – 34, 2015.
- [253] Hua Xiong, Wen-Yu Su, Qin-Chuan Liang, Zhi-Gang Zhang, Hui-Min Chen, Wan Du, Ying-Xuan Chen, and Jing-Yuan Fang. Inhibition of STAT5 induces G1 cell cycle arrest and reduces tumor cell invasion in human colorectal cancer cells. *Laboratory Investigation*, 89(6):717 – 725, Mar 2009. doi:10.1038/labinvest.2009.11.
- [254] Zheng Xu, Shao-Bo Xiao, Peng Xu, Qian Xie, Lu Cao, Dang Wang, Rui Luo, Yao Zhong, Huan-Chun Chen, and Liu-Rong Fang. miR-365, a Novel Negative Regulator of Interleukin-6 Gene Expression, Is Cooperatively Regulated by Sp1 and NF-KB. *Journal of Biological Chemistry*, 286(24):21401 – 21412, Apr 2011. doi:10.1074/jbc.m110.198630.

- [255] Rul Yamaguchi, Ryo Yoshida, Seiya Imoto, Tomoyuki Higuchi, and Satoru Miyano. Finding module-based gene networks with state-space models-Mining high-dimensional and short time-course gene expression data. *IEEE Signal Processing Magazine*, 24(1):37 – 46, 2007.
- [256] Ka Yee Yeung, Kenneth M Dombek, Kenneth Lo, John E Mittler, Jun Zhu, Eric E Schadt, Roger E Bumgarner, and Adrian E Raftery. Construction of regulatory networks using expression time-series data of a genotyped population. *Proceedings of the National Academy of Sciences*, 108(48):19436 – 19441, 2011.
- [257] Kiyotsugu Yoshida and Yoshio Miki. Role of BRCA1 and BRCA2 as regulators of dna repair, transcription, and cell cycle in response to dna damage. *Cancer Science*, 95(11):866–8711, 2004.
- [258] William Chad Young, Adrian E Raftery, and Ka Yee Yeung. Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Systems Biology*, 8(1):47, 2014.
- [259] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976 – 978, Feb 2010. URL: <http://dx.doi.org/10.1093/bioinformatics/btq064>, doi:10.1093/bioinformatics/btq064.
- [260] Jing Yu, Anne Smith, Paul Wang, Alexander Hartemink, and Erich Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594 – 3603, 2004.
- [261] Bowen Yuan, Youhong Liu, Xiaohui Yu, Linglong Yin, Yuchong Peng, Yingxue Gao, Qianling Zhu, Tuoyu Cao, Yinke Yang, Xuegong Fan, et al. FOXM1 contributes to taxane resistance by regulating UHRF1-controlled cancer cell stemness. *Cell Death & Disease*, 9(5):1 – 11, 2018.
- [262] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49 – 67, 2006.

- [263] Jeanho Yun, Hee-Don Chae, Hyon E. Choy, Jongkyeong Chung, Hyang-Sook Yoo, Moon-Hi Han, and Deug Y. Shin. p53 Negatively Regulatescdc2Transcription via the CCAAT-binding NF-Y Transcription Factor. *Journal of Biological Chemistry*, 274(42):29677 – 29682, Oct 1999. doi: 10.1074/jbc.274.42.29677.
- [264] Dimas Yusuf, Stefanie L Butland, Magdalena I Swanson, Eugene Bolotin, Amy Ticoll, Warren A Cheung, Xiao Yu Cindy Zhang, Christopher TD Dickman, Debra L Fulton, Jonathan S Lim, et al. The transcription factor encyclopedia. *Genome Biology*, 13(3):R24, 2012.
- [265] Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*, 1986.
- [266] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.
- [267] Yuji Zhang, Jianhua Xuan, G Benildo, Robert Clarke, and Habtom W Ressom. Reconstruction of gene regulatory modules in cancer cell cycle by multi-source data integration. *PloS One*, 5(4):e10268, 2010.
- [268] Fang Zhao, Zhenyu Xuan, Lihua Liu, and Michael Q Zhang. TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Research*, 33(suppl_1):D103 – D107, 2005.
- [269] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based *ab initio* prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171 – 1179, 2018.
- [270] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning – based sequence model. *Nature Methods*, 12(10):931 – 934, 2015.
- [271] Jun Zhu, Bin Zhang, Erin N Smith, Becky Drees, Rachel B Brem, Leonid Kruglyak, Roger E Bumgarner, and Eric E Schadt. Integrating large-scale

- functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7):854, 2008.
- [272] Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, 11(1):154, 2010.
- [273] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418 – 1429, Dec 2006. doi:10.1198/016214506000000735.
- [274] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301 – 320, 2005.
- [275] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4):1733, 2009.

Appendix A

Background

A.1 IUPAC degenerate base symbols

The table shows the list of degenerate symbols used in biochemistry to represent position in the DNA sequence where there is variation.

Table 22: List of Degenerate IUPAC base symbols

Description	Symbol	Bases represented				Complementary bases	
		#	A	C	G		T
Adenine	A		A				T
Cytosine	C			C			G
Guanine	G	1			G		C
Thymine	T					T	A
Uracil	U					U	A
Weak	W		A			T	W
Strong	S			C	G		S
Amino	M	2	A	C			K
Keto	K				G	T	M
Purine	R		A		G		Y
Pyrimidine	Y			C		T	R
Not A	B			C	G	T	V
Not C	D	3	A		G	T	H

Table 22 continued from previous page

Description	Symbol	Bases represented				Complementary bases	
		#	A	C	G		T
Not G	H		A	C		T	D
Not T	V		A	C	G		B
Any one base	N	4	A	C	G	T	N
Zero	Z	0					Z

The table reports the list of IUPAC base symbols used to report positional variation in DNA sequence. # stands for number of. The first column gives the description of the symbol. The second column gives the actual degenerate symbol. The 4th column gives the the number of nucleotides it represents. 5th – 8th columns gives the actual nucleotides it replaces. 9th column the complementary base.

Appendix B

BENIN

B.1 BENIN parameters setting

This section gives more details on the parameters setting used for running BENIN on the DREAM4 dataset. We divided the parameters into two sets. First, the general parameters: that we set to same values for both sub-challenge size. Second, the main parameters, which are the parameters that influence BENIN performance.

- Table 23 gives the values of BENIN general parameters when inferring size 10 and size 100 DREAM4 subchallenges.
- Table 24 (respectively Table 26) gives BENIN main parameters setting for reconstructing networks in the size 100 (respectively size 10) DREAM4 subchallenge, combining time-series and KO expression data.
- Table 25 (respectively Table 27) gives BENIN main parameters setting for reconstructing networks in the size 100 (respectively size 10) DREAM4 subchallenge, combining location data with time-series expression data.

B.2 BENIN results

In this section, we report the distribution of BENIN results when combining time series with the 11 generated location data for size 10 (Table 33) and size 100 (Table 32) DREAM4 subchallenge.

Table 23: BENIN General Parameter setting

Parameters	Values
λ (exponential distribution)	20
λ_{min} (integral lower limit)	1
λ_{max} (integral upper limit)	1000
β	0.5
λ_{Enet}	lambda.min
nbfolds (CV)	15
R (number of bootstrap)	1000
l (mean block length)	10
τ	0.5

The table summarizes the values assigned to each of BENIN general parameter when applying BENIN DREAM4 challenge.

Table 24: BENIN +KO parameters on size 100 subchallenge

Parameters	Net 1	Net 2	Net 3	Net 4	Net 5
γ	1.6	1.6	1.4	1.5	1.4
R	3000	4000	3000	3000	3000
α	0.9	0.99	0.9	0.9	0.9

BENIN parameters values when combining time series expression data and KO expression data for the inference of the five networks in the DREAM4 size 100 subchallenge.

Table 25: BENIN +Location parameters setting on size 100 subchallenge

Parameters	Net 1	Net 2	Net 3	Net 4	Net 5
γ	1	1	1	1	1
R	10000	10000	10000	10000	10000
α	0.7	0.7	0.7	0.8	0.9

BENIN parameter values when combining time-series expression data and location data for the inference of the five networks in the DREAM4 size 100 subchallenge.

Table 26: BENIN +KO parameters on size 10 subchallenge

Parameters	Net 1	Net 2	Net 3	Net 4	Net 5
γ	0.7	1.5	1.3	1.1	1.5
R	1000	2000	1000	1000	1000
α	0.9	0.99	0.9	0.9	0.9

BENIN parameters values when combining time series expression data and KO expression data for the inference of the five networks in the DREAM4 size 10 subchallenge.

Table 27: BENIN +Location data parameters on size 10 subchallenge

Parameters	Net 1	Net 2	Net 3	Net 4	Net 5
γ	1	1	1	1	1
R	1000	1000	1000	1000	1000
α	0.9	0.9	0.9	0.3	0.7

BENIN parameter values when combining time-series expression data and location data for the inference of the five networks in the DREAM4 size 10 subchallenge.

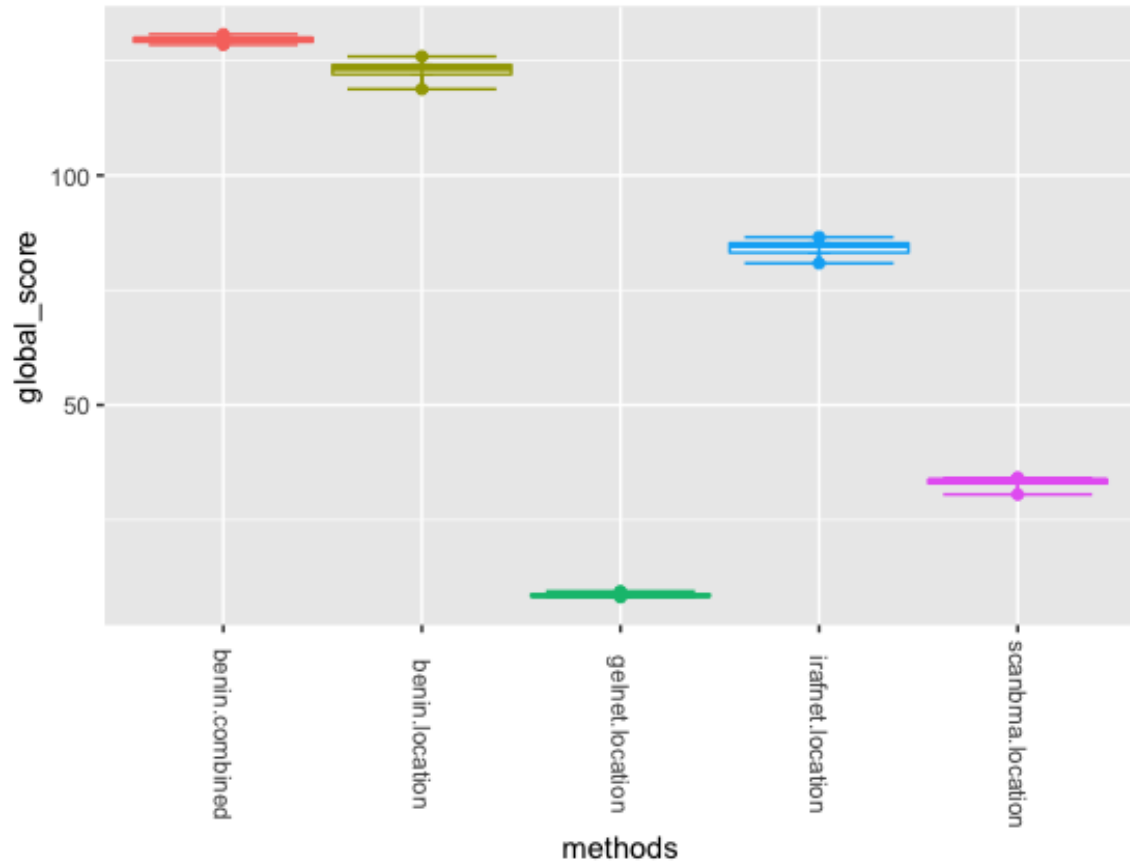


Figure 32: Global score Distribution for the DREAM4 size 100 subchallenge. Distribution of the global scores for the methods that combine the 11 generated location datasets with time series gene expression data to infer the five networks in size 100 DREAM4 subchallenge.

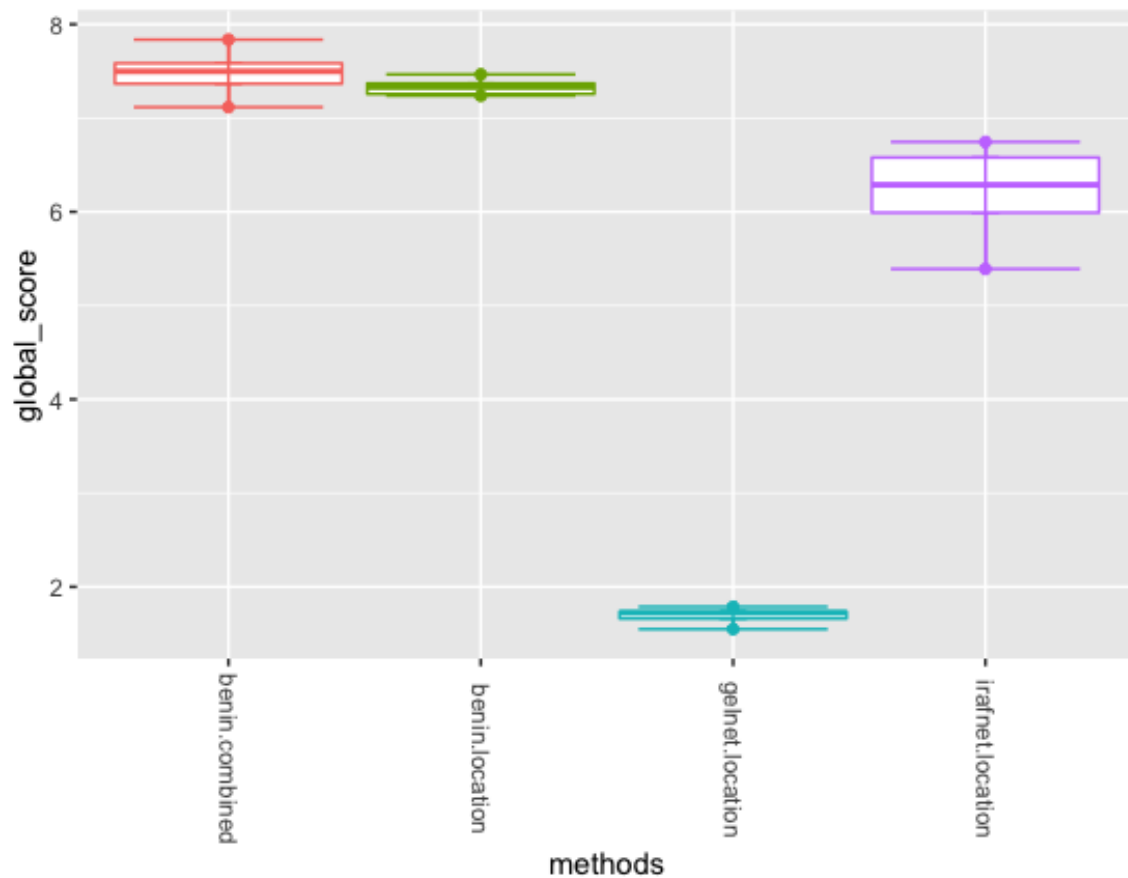


Figure 33: Global score Distribution for the DREAM4 size 10 subchallenge. Distribution of the global scores for the methods that combine the 11 generated location datasets with time series gene expression data to infer the five networks in size 10 DREAM4 subchallenge.

Appendix C

BENIN: Application to Human HeLa Cell Cycle GRN

C.1 Data

This section gives more details on the data used for the inference the HeLa cell cycle GRN.

- Table 28 presents list of the peak files downloaded from the UCSC web page that contains the peak regions of the TFBS. It gives details on the cell line considered, the TFs considered in each experiment, the lab that performed the experiment, and finally, the URLs for downloading the file.
- Table 29 gives the list of KD gene expression datasets. It provides details about the list of considered TFs. For each TF, the table gives datasets source as well as the list of their IDs. Our datasets are from GEO (Gene Expression Omnibus) and the ENCODE project.
- Table 30 gives details about the motifs downloaded from CisBP. It associates each motif to the TF information: the official name, its ID from external databases. It further provides information on the experiment that produces the motif.
- Table 31 provides the list of all the 628 genes that are considered from Whitfield time-series expression dataset [247].

- Table 32 gives the list the TFs considered in our experiments and their gene ontology annotation.
- Table 33 lists the datasets that we manually downloaded from the GEO database. If the dataset is already included in dataTable 29 from KnockTF, it is re-analyzed.
- To build our “gold standard” network, we downloaded a list of 132 gold standard networks from the HumanBase database <https://hb.flatironinstitute.org/download>. Then we concatenated all the 132 networks into a single network. We used a row concatenation. We then restricted the network to genes that are expressed in the cell cycle. Table 34 gives for each edge, how many time it is repeated in the concatenated network.
- To build our “gold standard” network, we collected two networks from Garcia Alonso *et.al* work [78]. The authors generated one network for cancer cells and one for normal cells. We combine the two networks row per row. In Table 35, we give the number of repetitions for each edge after combining the two networks.
- Table 36 and Table 37 give the list of edges in our gold standard network. Table 36 gives the list of positive links and Table 37 the list of negative links. In each table, the 1st column represents the TF. The 2nd column the TG. The 3rd column informs for each edge if it is present in the network (value of 1) or if it is absent (value of 0). The present edges are the positive links, and the absent edges are the negative links. For each edge, the number in the 4th column provides the number of times it was repeated before removing the duplicate edges from the network obtained by combining Alonso networks and HumanBase networks.
- To perform the ortholog information transfer, we first need to build our model organism regulatory network. We first collected regulatory interactions from TRRUST and RegNetwork databases. Table 38 gives the list of repeated regulatory interactions after merging the regulatory network downloaded from TRRUST and RegNetwork databases.

- Table 39 gives the list of repeated edges the mouse regulatory network obtained after merging networks from TRRUST, RegNetwork and STRINGDB databases.

Table 28: List of HeLa Peak Files

Cell	TFs	Lab	URL
HeLa-S3	BRCA1	Stanford	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3Brca1a300IggrabUniPk.narrowPeak.gz
HeLa-S3	CTCF	Broad	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsBroadHelas3CtcfUniPk.narrowPeak.gz
HeLa-S3	E2F1	USC	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3E2f1UniPk.narrowPeak.gz
HeLa-S3	NFYA	Harvard	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3NfyaIggrabUniPk.narrowPeak.gz
HeLa-S3	NFYB	Harvard	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3NfybIggrabUniPk.narrowPeak.gz
HeLa-S3	STAT1	Yale	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3Stat1fng30UniPk.narrowPeak.gz
HeLa-S3	TFAP2A	USC	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3Ap2alphaUniPk.narrowPeak.gz
HeLa-S3	ZNF143	Stanford	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsSydhHelas3Znf143IggrabUniPk.narrowPeak.gz

The table gives details about the list of peak files downloaded from the UCSC webpage. The 1st column gives the cell line used for the Chip-seq experiment. The 2nd column gives the TF concerned in the experiment. The 3rd column provides the lab names that generate the dataset, and finally, 4th column provides the URLs to access the file used.

Table 29: List of knockdown datasets

TF	Source	Profile ID
STAT1	GEO	GSE35551
SRF	GEO	GSE22606
SP1	GEO	GSE37935
NFYA	GEO	GSE40215
NFE2L2	GEO	GSE38332
MITF	GEO	GSE16249
ZNF521	GEO	GSE79110
MBD4	GEO	GSE52567
BRCA1	GEO	GSE54265
YY1	GEO	GSE14964
RUNX1	GEO	GSE94835, GSE79598, GSE62140, GSE45743, GSE34594, GSE24778, GSE16238, GSE16238
FOXMI	GEO	GSE55204, GSE40051, GSE31534
	ENCODE	ENCSR701TVL
HSF2	GEO	GSE48672, GSE31534
HF1A	GEO	GSE76581, GSE56989, GSE55212, GSE54360, GSE44943, GSE3188, GSE3188
NR3C1	GEO	GSE42538
KLF9	GEO	GSE54699
NFYB	ENCODE	ENCSR171KMM
	GEO	GSE61272
ZNF143	ENCODE	ENCSR781XJD
HOXB4	ENCODE	ENCSR359VJC
CTCF	GEO	GSE108869

The table gives the list of KD gene expression datasets. The 1st column provides the list of considered TFs. The 2nd column gives the original source of the datasets. The 3rd column provides the list of dataset IDs.

Table 30: Information Motif and Transcription Factor

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T010824.2.00	M02762.2.00	MS33.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Jolma2013	23332764
T010824.2.00	M02763.2.00	MS33.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Jolma2013	23332764
T010824.2.00	M02764.2.00	MS33.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Jolma2013	23332764
T010824.2.00	M02765.2.00	MS33.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Jolma2013	23332764
T010824.2.00	M02766.2.00	MS33.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Jolma2013	23332764
T010824.2.00	M02767.2.00	MS33.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Jolma2013	23332764
T010824.2.00	M04054.2.00	MS62.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Yin2017	28473536
T010824.2.00	M04055.2.00	MS62.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Yin2017	28473536
T010824.2.00	M07784.2.00	MS18.2.00	ENSG00000137203	TFAP2A	TF_AP-2	ENCODE	22955619
T010824.2.00	M08703.2.00	MS27.2.00	ENSG00000137203	TFAP2A	TF_AP-2	HocoMoco	23175603
T010824.2.00	M09755.2.00	MS59.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Transfac	16381825
T010824.2.00	M09756.2.00	MS59.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Transfac	16381825
T010824.2.00	M09757.2.00	MS59.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Transfac	16381825
T010824.2.00	M09758.2.00	MS59.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Transfac	16381825
T010824.2.00	M09759.2.00	MS59.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Transfac	16381825
T010824.2.00	M09760.2.00	MS59.2.00	ENSG00000137203	TFAP2A	TF_AP-2	Transfac	16381825
T034249.2.00	M02774.2.00	MS33.2.00	ENSG00000070444	MNT	HLH	Jolma2013	23332764
T034254.2.00	M08049.2.00	MS31.2.00	ENSG00000100644	HIF1A	HLH	JASPAR	24194598
T034254.2.00	M08713.2.00	MS27.2.00	ENSG00000100644	HIF1A	HLH	HocoMoco	23175603
T034254.2.00	M09454.2.00	MS28.2.00	ENSG00000100644	HIF1A	HLH	HOMER	20513432
T034254.2.00	M09807.2.00	MS59.2.00	ENSG00000100644	HIF1A	HLH	Transfac	16381825
T034254.2.00	M09808.2.00	MS59.2.00	ENSG00000100644	HIF1A	HLH	Transfac	16381825
T034254.2.00	M09809.2.00	MS59.2.00	ENSG00000100644	HIF1A	HLH	Transfac	16381825
T034254.2.00	M09810.2.00	MS59.2.00	ENSG00000100644	HIF1A	HLH	Transfac	16381825
T034254.2.00	M09811.2.00	MS59.2.00	ENSG00000100644	HIF1A	HLH	Transfac	16381825
T034335.2.00	M08058.2.00	MS31.2.00	ENSG00000187098	MITF	HLH	JASPAR	24194598
T034335.2.00	M08740.2.00	MS27.2.00	ENSG00000187098	MITF	HLH	HocoMoco	23175603
T034335.2.00	M09880.2.00	MS59.2.00	ENSG00000187098	MITF	HLH	Transfac	16381825
T034335.2.00	M09881.2.00	MS59.2.00	ENSG00000187098	MITF	HLH	Transfac	16381825
T059732.2.00	M08789.2.00	MS27.2.00	ENSG00000116044	NFE2L2	bZIP_1	HocoMoco	23175603
T059732.2.00	M09943.2.00	MS59.2.00	ENSG00000116044	NFE2L2	bZIP_1	Transfac	16381825
T059732.2.00	M09944.2.00	MS59.2.00	ENSG00000116044	NFE2L2	bZIP_1	Transfac	16381825
T059732.2.00	M09945.2.00	MS59.2.00	ENSG00000116044	NFE2L2	bZIP_1	Transfac	16381825
T059732.2.00	M09946.2.00	MS59.2.00	ENSG00000116044	NFE2L2	bZIP_1	Transfac	16381825
T059742.2.00	M01813.2.00	MS64.2.00	ENSG00000137504	CREBZF	bZIP_1	Zoo.01	25215497
T094796.2.00	M04400.2.00	MS62.2.00	ENSG00000067082	KLF6	zf-C2H2	Yin2017	28473536
T094796.2.00	M04401.2.00	MS62.2.00	ENSG00000067082	KLF6	zf-C2H2	Yin2017	28473536
T094796.2.00	M08857.2.00	MS27.2.00	ENSG00000067082	KLF6	zf-C2H2	HocoMoco	23175603
T094796.2.00	M10113.2.00	MS59.2.00	ENSG00000067082	KLF6	zf-C2H2	Transfac	16381825
T094821.2.00	M02663.2.00	MS31.2.00	ENSG00000099326	MZF1	zf-C2H2	JASPAR	24194598
T094821.2.00	M02664.2.00	MS31.2.00	ENSG00000099326	MZF1	zf-C2H2	JASPAR	24194598

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T094821.2.00	M08236.2.00	MS43.2.00	ENSG00000099326	MZF1	zf-C2H2	Najafabadi2015b	25690854
T094821.2.00	M08286.2.00	MS52.2.00	ENSG00000099326	MZF1	zf-C2H2	Schmitges2016	27852650
T094821.2.00	M08863.2.00	MS27.2.00	ENSG00000099326	MZF1	zf-C2H2	HocoMoco	23175603
T094821.2.00	M10133.2.00	MS59.2.00	ENSG00000099326	MZF1	zf-C2H2	Transfac	16381825
T094821.2.00	M10134.2.00	MS59.2.00	ENSG00000099326	MZF1	zf-C2H2	Transfac	16381825
T094821.2.00	M10135.2.00	MS59.2.00	ENSG00000099326	MZF1	zf-C2H2	Transfac	16381825
T094821.2.00	M10136.2.00	MS59.2.00	ENSG00000099326	MZF1	zf-C2H2	Transfac	16381825
T094823.2.00	M02877.2.00	MS33.2.00	ENSG00000100811	YY1	zf-C2H2	Jolma2013	23332764
T094823.2.00	M04406.2.00	MS62.2.00	ENSG00000100811	YY1	zf-C2H2	Yin2017	28473536
T094823.2.00	M04407.2.00	MS62.2.00	ENSG00000100811	YY1	zf-C2H2	Yin2017	28473536
T094823.2.00	M04408.2.00	MS62.2.00	ENSG00000100811	YY1	zf-C2H2	Yin2017	28473536
T094823.2.00	M04409.2.00	MS62.2.00	ENSG00000100811	YY1	zf-C2H2	Yin2017	28473536
T094823.2.00	M05845.2.00	MS30.2.00	ENSG00000100811	YY1	zf-C2H2	Isakova2017	28092692
T094823.2.00	M07855.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M07856.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M07857.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M07858.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M07859.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M07860.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M07861.2.00	MS18.2.00	ENSG00000100811	YY1	zf-C2H2	ENCODE	22955619
T094823.2.00	M08085.2.00	MS31.2.00	ENSG00000100811	YY1	zf-C2H2	JASPAR	24194598
T094823.2.00	M08237.2.00	MS43.2.00	ENSG00000100811	YY1	zf-C2H2	Najafabadi2015b	25690854
T094823.2.00	M08288.2.00	MS52.2.00	ENSG00000100811	YY1	zf-C2H2	Schmitges2016	27852650
T094823.2.00	M08865.2.00	MS27.2.00	ENSG00000100811	YY1	zf-C2H2	HocoMoco	23175603
T094823.2.00	M10138.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10139.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10140.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10141.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10142.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10143.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10144.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10145.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094823.2.00	M10146.2.00	MS59.2.00	ENSG00000100811	YY1	zf-C2H2	Transfac	16381825
T094831.2.00	M02878.2.00	MS33.2.00	ENSG00000102974	CTCF	zf-C2H2	Jolma2013	23332764
T094831.2.00	M05846.2.00	MS30.2.00	ENSG00000102974	CTCF	zf-C2H2	Isakova2017	28092692
T094831.2.00	M07550.2.00	MS49.2.00	ENSG00000102974	CTCF	zf-C2H2	Rhee2011	22153082
T094831.2.00	M07551.2.00	MS49.2.00	ENSG00000102974	CTCF	zf-C2H2	Rhee2011	22153082
T094831.2.00	M07552.2.00	MS49.2.00	ENSG00000102974	CTCF	zf-C2H2	Rhee2011	22153082
T094831.2.00	M07862.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07863.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07864.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07865.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T094831.2.00	M07908.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07909.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07910.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07911.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07912.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07913.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07914.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07915.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07916.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07917.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07918.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07919.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07920.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07921.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07922.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07923.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07924.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M07925.2.00	MS18.2.00	ENSG00000102974	CTCF	zf-C2H2	ENCODE	22955619
T094831.2.00	M08087.2.00	MS31.2.00	ENSG00000102974	CTCF	zf-C2H2	JASPAR	24194598
T094831.2.00	M08238.2.00	MS43.2.00	ENSG00000102974	CTCF	zf-C2H2	Najafabadi2015b	25690854
T094831.2.00	M08289.2.00	MS52.2.00	ENSG00000102974	CTCF	zf-C2H2	Schmitges2016	27852650
T094831.2.00	M08869.2.00	MS27.2.00	ENSG00000102974	CTCF	zf-C2H2	HocoMoco	23175603
T094831.2.00	M09503.2.00	MS28.2.00	ENSG00000102974	CTCF	zf-C2H2	HOMER	20513432
T094831.2.00	M09504.2.00	MS28.2.00	ENSG00000102974	CTCF	zf-C2H2	HOMER	20513432
T094831.2.00	M10152.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10153.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10154.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10155.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10156.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10157.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10158.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094831.2.00	M10159.2.00	MS59.2.00	ENSG00000102974	CTCF	zf-C2H2	Transfac	16381825
T094868.2.00	M08088.2.00	MS31.2.00	ENSG00000119138	KLF9	zf-C2H2	JASPAR	24194598
T094868.2.00	M08880.2.00	MS27.2.00	ENSG00000119138	KLF9	zf-C2H2	HocoMoco	23175603
T095017.2.00	M04509.2.00	MS62.2.00	ENSG00000162702	ZNF281	zf-C2H2	Yin2017	28473536
T095017.2.00	M04510.2.00	MS62.2.00	ENSG00000162702	ZNF281	zf-C2H2	Yin2017	28473536
T095017.2.00	M08321.2.00	MS52.2.00	ENSG00000162702	ZNF281	zf-C2H2	Schmitges2016	27852650
T095017.2.00	M08906.2.00	MS27.2.00	ENSG00000162702	ZNF281	zf-C2H2	HocoMoco	23175603
T095017.2.00	M10237.2.00	MS59.2.00	ENSG00000162702	ZNF281	zf-C2H2	Transfac	16381825
T095041.2.00	M02899.2.00	MS33.2.00	ENSG00000166478	ZNF143	zf-C2H2	Jolma2013	23332764
T095041.2.00	M07931.2.00	MS18.2.00	ENSG00000166478	ZNF143	zf-C2H2	ENCODE	22955619
T095041.2.00	M08910.2.00	MS27.2.00	ENSG00000166478	ZNF143	zf-C2H2	HocoMoco	23175603

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T095041.2.00	M09510.2.00	MS28.2.00	ENSG00000166478	ZNF143	zf-C2H2	HOMER	20513432
T095041.2.00	M10247.2.00	MS59.2.00	ENSG00000166478	ZNF143	zf-C2H2	Transfac	16381825
T095041.2.00	M10248.2.00	MS59.2.00	ENSG00000166478	ZNF143	zf-C2H2	Transfac	16381825
T095041.2.00	M10249.2.00	MS59.2.00	ENSG00000166478	ZNF143	zf-C2H2	Transfac	16381825
T095112.2.00	M10260.2.00	MS59.2.00	ENSG00000171940	ZNF217	zf-C2H2	Transfac	16381825
T095129.2.00	M08445.2.00	MS31.2.00	ENSG00000173404	INSM1	zf-C2H2	JASPAR	24194598
T095129.2.00	M08923.2.00	MS27.2.00	ENSG00000173404	INSM1	zf-C2H2	HocoMoco	23175603
T095173.2.00	M02914.2.00	MS33.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Jolma2013	23332764
T095173.2.00	M04579.2.00	MS62.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Yin2017	28473536
T095173.2.00	M04580.2.00	MS62.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Yin2017	28473536
T095173.2.00	M07937.2.00	MS18.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	ENCODE	22955619
T095173.2.00	M08095.2.00	MS31.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	JASPAR	24194598
T095173.2.00	M08926.2.00	MS27.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	HocoMoco	23175603
T095173.2.00	M10273.2.00	MS59.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Transfac	16381825
T095173.2.00	M10274.2.00	MS59.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Transfac	16381825
T095173.2.00	M10275.2.00	MS59.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Transfac	16381825
T095173.2.00	M10276.2.00	MS59.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Transfac	16381825
T095173.2.00	M10277.2.00	MS59.2.00	ENSG00000178951	ZBTB7A	zf-C2H2	Transfac	16381825
T095233.2.00	M02921.2.00	MS33.2.00	ENSG00000185591	SP1	zf-C2H2	Jolma2013	23332764
T095233.2.00	M04605.2.00	MS62.2.00	ENSG00000185591	SP1	zf-C2H2	Yin2017	28473536
T095233.2.00	M04606.2.00	MS62.2.00	ENSG00000185591	SP1	zf-C2H2	Yin2017	28473536
T095233.2.00	M08096.2.00	MS31.2.00	ENSG00000185591	SP1	zf-C2H2	JASPAR	24194598
T095233.2.00	M08363.2.00	MS52.2.00	ENSG00000185591	SP1	zf-C2H2	Schmitges2016	27852650
T095233.2.00	M08938.2.00	MS27.2.00	ENSG00000185591	SP1	zf-C2H2	HocoMoco	23175603
T095233.2.00	M10294.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095233.2.00	M10295.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095233.2.00	M10296.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095233.2.00	M10297.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095233.2.00	M10298.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095233.2.00	M10299.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095233.2.00	M10300.2.00	MS59.2.00	ENSG00000185591	SP1	zf-C2H2	Transfac	16381825
T095392.2.00	M07746.2.00	MS03.2.00	ENSG00000198466	ZNF587	zf-C2H2	Barazandeh2018	29146583
T095403.2.00	M10318.2.00	MS59.2.00	ENSG00000198795	ZNF521	zf-C2H2	Transfac	16381825
T159918.2.00	M08104.2.00	MS31.2.00	ENSG00000001167	NFYA	CBFB_NFYA	JASPAR	24194598
T159918.2.00	M09018.2.00	MS27.2.00	ENSG00000001167	NFYA	CBFB_NFYA	HocoMoco	23175603
T159918.2.00	M10403.2.00	MS59.2.00	ENSG00000001167	NFYA	CBFB_NFYA	Transfac	16381825
T159918.2.00	M10404.2.00	MS59.2.00	ENSG00000001167	NFYA	CBFB_NFYA	Transfac	16381825
T159918.2.00	M10405.2.00	MS59.2.00	ENSG00000001167	NFYA	CBFB_NFYA	Transfac	16381825
T172616.2.00	M02952.2.00	MS33.2.00	ENSG00000101412	E2F1	E2F_TDP	Jolma2013	23332764
T172616.2.00	M02953.2.00	MS33.2.00	ENSG00000101412	E2F1	E2F_TDP	Jolma2013	23332764
T172616.2.00	M02954.2.00	MS33.2.00	ENSG00000101412	E2F1	E2F_TDP	Jolma2013	23332764
T172616.2.00	M02955.2.00	MS33.2.00	ENSG00000101412	E2F1	E2F_TDP	Jolma2013	23332764

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T172616.2.00	M04700.2.00	MS62.2.00	ENSG00000101412	E2F1	E2F_TDP	Yin2017	28473536
T172616.2.00	M07938.2.00	MS18.2.00	ENSG00000101412	E2F1	E2F_TDP	ENCODE	22955619
T172616.2.00	M09030.2.00	MS27.2.00	ENSG00000101412	E2F1	E2F_TDP	HocoMoco	23175603
T172616.2.00	M09521.2.00	MS28.2.00	ENSG00000101412	E2F1	E2F_TDP	HOMER	20513432
T172616.2.00	M10444.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10445.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10446.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10447.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10448.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10449.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10450.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10451.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10452.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172616.2.00	M10453.2.00	MS59.2.00	ENSG00000101412	E2F1	E2F_TDP	Transfac	16381825
T172619.2.00	M02959.2.00	MS33.2.00	ENSG00000129173	E2F8	E2F_TDP	Jolma2013	23332764
T172619.2.00	M04703.2.00	MS62.2.00	ENSG00000129173	E2F8	E2F_TDP	Yin2017	28473536
T172619.2.00	M04704.2.00	MS62.2.00	ENSG00000129173	E2F8	E2F_TDP	Yin2017	28473536
T172620.2.00	M09032.2.00	MS27.2.00	ENSG00000133740	E2F5	E2F_TDP	HocoMoco	23175603
T185765.2.00	M09085.2.00	MS27.2.00	ENSG00000111206	FOXMI	Forkhead	HocoMoco	23175603
T185765.2.00	M10532.2.00	MS59.2.00	ENSG00000111206	FOXMI	Forkhead	Transfac	16381825
T185765.2.00	M10533.2.00	MS59.2.00	ENSG00000111206	FOXMI	Forkhead	Transfac	16381825
T185765.2.00	M10534.2.00	MS59.2.00	ENSG00000111206	FOXMI	Forkhead	Transfac	16381825
T185765.2.00	M10535.2.00	MS59.2.00	ENSG00000111206	FOXMI	Forkhead	Transfac	16381825
T209837.2.00	M03143.2.00	MS33.2.00	ENSG00000130675	MNX1	Homeobox	Jolma2013	23332764
T209837.2.00	M05126.2.00	MS62.2.00	ENSG00000130675	MNX1	Homeobox	Yin2017	28473536
T209837.2.00	M05127.2.00	MS62.2.00	ENSG00000130675	MNX1	Homeobox	Yin2017	28473536
T209837.2.00	M05128.2.00	MS62.2.00	ENSG00000130675	MNX1	Homeobox	Yin2017	28473536
T209869.2.00	M03178.2.00	MS33.2.00	ENSG00000160199	PKNOX1	Homeobox	Jolma2013	23332764
T209869.2.00	M05206.2.00	MS62.2.00	ENSG00000160199	PKNOX1	Homeobox	Yin2017	28473536
T209869.2.00	M05207.2.00	MS62.2.00	ENSG00000160199	PKNOX1	Homeobox	Yin2017	28473536
T209869.2.00	M05208.2.00	MS62.2.00	ENSG00000160199	PKNOX1	Homeobox	Yin2017	28473536
T209869.2.00	M05209.2.00	MS62.2.00	ENSG00000160199	PKNOX1	Homeobox	Yin2017	28473536
T209869.2.00	M09150.2.00	MS27.2.00	ENSG00000160199	PKNOX1	Homeobox	HocoMoco	23175603
T209869.2.00	M10712.2.00	MS59.2.00	ENSG00000160199	PKNOX1	Homeobox	Transfac	16381825
T209914.2.00	M03223.2.00	MS33.2.00	ENSG00000177426	TGIF1	Homeobox	Jolma2013	23332764
T209914.2.00	M05317.2.00	MS62.2.00	ENSG00000177426	TGIF1	Homeobox	Yin2017	28473536
T209914.2.00	M05318.2.00	MS62.2.00	ENSG00000177426	TGIF1	Homeobox	Yin2017	28473536
T209914.2.00	M09159.2.00	MS27.2.00	ENSG00000177426	TGIF1	Homeobox	HocoMoco	23175603
T209914.2.00	M10740.2.00	MS59.2.00	ENSG00000177426	TGIF1	Homeobox	Transfac	16381825
T209924.2.00	M05343.2.00	MS62.2.00	ENSG00000182742	HOXB4	Homeobox	Yin2017	28473536
T209924.2.00	M05344.2.00	MS62.2.00	ENSG00000182742	HOXB4	Homeobox	Yin2017	28473536
T209924.2.00	M05345.2.00	MS62.2.00	ENSG00000182742	HOXB4	Homeobox	Yin2017	28473536

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T209924.2.00	M09162.2.00	MS27.2.00	ENSG00000182742	HOXB4	Homeobox	HocoMoco	23175603
T240222.2.00	M02744.2.00	MS32.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Jolma2010	20378718
T240222.2.00	M03324.2.00	MS33.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Jolma2013	23332764
T240222.2.00	M05505.2.00	MS62.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Yin2017	28473536
T240222.2.00	M05506.2.00	MS62.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Yin2017	28473536
T240222.2.00	M05507.2.00	MS62.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Yin2017	28473536
T240222.2.00	M05508.2.00	MS62.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Yin2017	28473536
T240222.2.00	M09229.2.00	MS27.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	HocoMoco	23175603
T240222.2.00	M10859.2.00	MS59.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Transfac	16381825
T240222.2.00	M10860.2.00	MS59.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Transfac	16381825
T240222.2.00	M10861.2.00	MS59.2.00	ENSG00000025156	HSF2	HSF_DNA-bind	Transfac	16381825
T253657.2.00	M03341.2.00	MS33.2.00	ENSG00000112658	SRF	SRF-TF	Jolma2013	23332764
T253657.2.00	M03342.2.00	MS33.2.00	ENSG00000112658	SRF	SRF-TF	Jolma2013	23332764
T253657.2.00	M05553.2.00	MS62.2.00	ENSG00000112658	SRF	SRF-TF	Yin2017	28473536
T253657.2.00	M05554.2.00	MS62.2.00	ENSG00000112658	SRF	SRF-TF	Yin2017	28473536
T253657.2.00	M07981.2.00	MS18.2.00	ENSG00000112658	SRF	SRF-TF	ENCODE	22955619
T253657.2.00	M07982.2.00	MS18.2.00	ENSG00000112658	SRF	SRF-TF	ENCODE	22955619
T253657.2.00	M07983.2.00	MS18.2.00	ENSG00000112658	SRF	SRF-TF	ENCODE	22955619
T253657.2.00	M07984.2.00	MS18.2.00	ENSG00000112658	SRF	SRF-TF	ENCODE	22955619
T253657.2.00	M09249.2.00	MS27.2.00	ENSG00000112658	SRF	SRF-TF	HocoMoco	23175603
T253657.2.00	M10947.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10948.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10949.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10950.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10951.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10952.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10953.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10954.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10955.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T253657.2.00	M10956.2.00	MS59.2.00	ENSG00000112658	SRF	SRF-TF	Transfac	16381825
T260164.2.00	M09256.2.00	MS27.2.00	ENSG00000134046	MBD2	MBD	HocoMoco	23175603
T303216.2.00	M03366.2.00	MS33.2.00	ENSG00000113580	NR3C1	zf-C4	Jolma2013	23332764

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T303216.2.00	M05587.2.00	MS62.2.00	ENSG00000113580	NR3C1	zf-C4	Yin2017	28473536
T303216.2.00	M05588.2.00	MS62.2.00	ENSG00000113580	NR3C1	zf-C4	Yin2017	28473536
T303216.2.00	M07986.2.00	MS18.2.00	ENSG00000113580	NR3C1	zf-C4	ENCODE	22955619
T303216.2.00	M07987.2.00	MS18.2.00	ENSG00000113580	NR3C1	zf-C4	ENCODE	22955619
T303216.2.00	M09270.2.00	MS27.2.00	ENSG00000113580	NR3C1	zf-C4	HocoMoco	23175603
T303216.2.00	M09607.2.00	MS28.2.00	ENSG00000113580	NR3C1	zf-C4	HOMER	20513432
T303216.2.00	M11119.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T303216.2.00	M11120.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T303216.2.00	M11121.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T303216.2.00	M11122.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T303216.2.00	M11123.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T303216.2.00	M11124.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T303216.2.00	M11125.2.00	MS59.2.00	ENSG00000113580	NR3C1	zf-C4	Transfac	16381825
T319384.2.00	M09371.2.00	MS27.2.00	ENSG00000159216	RUNX1	Runt	HocoMoco	23175603
T319384.2.00	M09631.2.00	MS28.2.00	ENSG00000159216	RUNX1	Runt	HOMER	20513432
T319384.2.00	M09632.2.00	MS28.2.00	ENSG00000159216	RUNX1	Runt	HOMER	20513432
T319384.2.00	M11258.2.00	MS59.2.00	ENSG00000159216	RUNX1	Runt	Transfac	16381825
T319384.2.00	M11259.2.00	MS59.2.00	ENSG00000159216	RUNX1	Runt	Transfac	16381825
T319384.2.00	M11260.2.00	MS59.2.00	ENSG00000159216	RUNX1	Runt	Transfac	16381825
T319384.2.00	M11261.2.00	MS59.2.00	ENSG00000159216	RUNX1	Runt	Transfac	16381825
T319384.2.00	M11262.2.00	MS59.2.00	ENSG00000159216	RUNX1	Runt	Transfac	16381825
T319384.2.00	M11263.2.00	MS59.2.00	ENSG00000159216	RUNX1	Runt	Transfac	16381825
T324626.2.00	M05745.2.00	MS62.2.00	ENSG00000141905	NFIC	MH1	Yin2017	28473536
T324626.2.00	M05746.2.00	MS62.2.00	ENSG00000141905	NFIC	MH1	Yin2017	28473536
T324626.2.00	M08164.2.00	MS31.2.00	ENSG00000141905	NFIC	MH1	JASPAR	24194598
T324626.2.00	M09378.2.00	MS27.2.00	ENSG00000141905	NFIC	MH1	HocoMoco	23175603
T324626.2.00	M09635.2.00	MS28.2.00	ENSG00000141905	NFIC	MH1	HOMER	20513432
T324626.2.00	M09636.2.00	MS28.2.00	ENSG00000141905	NFIC	MH1	HOMER	20513432
T324626.2.00	M11278.2.00	MS59.2.00	ENSG00000141905	NFIC	MH1	Transfac	16381825
T324626.2.00	M11279.2.00	MS59.2.00	ENSG00000141905	NFIC	MH1	Transfac	16381825
T324628.2.00	M03480.2.00	MS33.2.00	ENSG00000162599	NFIA	MH1	Jolma2013	23332764
T324628.2.00	M03481.2.00	MS33.2.00	ENSG00000162599	NFIA	MH1	Jolma2013	23332764
T324628.2.00	M09379.2.00	MS27.2.00	ENSG00000162599	NFIA	MH1	HocoMoco	23175603
T324628.2.00	M11282.2.00	MS59.2.00	ENSG00000162599	NFIA	MH1	Transfac	16381825
T328056.2.00	M02487.2.00	MS64.2.00	ENSG00000064961	HMG20B	HMG_box	Zoo_01	25215497
T328057.2.00	M00195.2.00	MS02.2.00	ENSG00000079432	CIC	HMG_box	Badis09	19443739
T337444.2.00	M08013.2.00	MS18.2.00	ENSG00000115415	STAT1	STAT_bind	ENCODE	22955619
T337444.2.00	M08014.2.00	MS18.2.00	ENSG00000115415	STAT1	STAT_bind	ENCODE	22955619
T337444.2.00	M08171.2.00	MS31.2.00	ENSG00000115415	STAT1	STAT_bind	JASPAR	24194598
T337444.2.00	M08229.2.00	MS42.2.00	ENSG00000115415	STAT1	STAT_bind	modENCODE	22080565
T337444.2.00	M08230.2.00	MS42.2.00	ENSG00000115415	STAT1	STAT_bind	modENCODE	22080565
T337444.2.00	M09411.2.00	MS27.2.00	ENSG00000115415	STAT1	STAT_bind	HocoMoco	23175603

Table 30 continued from previous page

TF ID	Motif ID	MSource ID	DBID	TF Name	DBDs	MSource Identifier	PMID
T337444.2.00	M09642.2.00	MS28.2.00	ENSG00000115415	STAT1	STAT_bind	HOMER	20513432
T337444.2.00	M11348.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337444.2.00	M11349.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337444.2.00	M11350.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337444.2.00	M11351.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337444.2.00	M11352.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337444.2.00	M11353.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337444.2.00	M11354.2.00	MS59.2.00	ENSG00000115415	STAT1	STAT_bind	Transfac	16381825
T337450.2.00	M09417.2.00	MS27.2.00	ENSG00000173757	STAT5B	STAT_bind	HocoMoco	23175603
T337450.2.00	M11371.2.00	MS59.2.00	ENSG00000173757	STAT5B	STAT_bind	Transfac	16381825
T350252.2.00	M08178.2.00	MS31.2.00	ENSG00000120837	NFYB	UNKNOWN	JASPAR	24194598
T350252.2.00	M09444.2.00	MS27.2.00	ENSG00000120837	NFYB	UNKNOWN	HocoMoco	23175603
T350264.2.00	M11430.2.00	MS59.2.00	ENSG00000137947	GTF2B	UNKNOWN	Transfac	16381825
T350264.2.00	M11431.2.00	MS59.2.00	ENSG00000137947	GTF2B	UNKNOWN	Transfac	16381825

The table gives details about the motifs downloaded from CisBP and used for inferring the HeLa cell cycle GRN. The 1st column is the internal unique CisBP ID for the TF. The 2nd column is the internal CisBP ID for the associated motif. The 3rd column is the internal CisBP ID for the database or study the motif originates. The 4th column is the external ID of the TF. The 5th column gives the name of the TF. The 6th column gives the unique set of DBDs present in the TF. The 7th column is the ID for the source project of the motif. The 8th column is the Pubmed ID of the motif.

Table 31: Cell cycle genes

List of Considered Cell Cycle Genes					
GOLGA8A	PTP4A1	GINS3	GCSH	SHTN1	TTF2
FBXL20	INSR	DLGAP5	TMEM138	CEP55	GSE1
ZNF587	INADL	CKS2	CDKN2D	ODF2	POLQ
ZNHIT2	ADH4	MBD4	ITPR3	CDK20	HIST1H2AM
DCAF7	COQ6	MLLT4	DIS3	NCS1	PRR11
AHI1	IL18BP	KLF6	HIST1H4C	HSPA1L	TTC31
BARD1	PCNA	GCLM	ZBED5	SRD5A1	ARL4A
KATNA1	CENPF	HMGB2	UBE2C	DKC1	CBX3
KIAA1586	PDGFA	CCNB1	TUBA1A	MKI67	BCLAF1
HELLS	ANKRD10	CTNND1	ZNF414	ERN2	MDM2
FZR1	EXO1	PRIM1	TAF2N	GTF2B	CSGALNACT1
MASTL	HLA-DOA	GAS1	CAPS	ARL6IP1	NDE1
CDC27	HIST1H4H	CDR2	CLSPN	GOT1	ABCA7
HP1BP3	TOMM70A	MATN2	ROCK1	RBBP8	RAD51AP1
KNSTRN	BIRC2	KDM4A	RAD51	SDC1	ASIP
PPP1R10	CHAF1A	KIAA1147	ZNF217	MCM5	HIST1H4B
ZC3HC1	VEGFC	TOP2A	AMD1	IFIT1	HIST1H4E
DONSON	POLA1	CSH2	TUBB4B	STAT1	KIAA1524
RCCD1	ZWINT	PLCXD1	KATNBL1	DNAJB4	CCDC14
FAM105A	STIL	SHC1	KIAA0586	TIPIN	CCDC90B
SPAG5	CDC45	SV2B	MELK	KPNB1	ZSCAN5A
RPS25	CCNF	NEK2	NFYB	FOXM1	PCED1A
FAM110A	MAP2K6	CDC25C	FAM60A	NUP160	BAG3
DHFR	NEIL3	GPSM2	HLA-DRA	PSMD11	ARHGAP19
ZBTB7A	NSUN3	ACYP1	HOXB4	CTR9	HIST3H2A
DTL	OLR1	THRAP3	CDH24	RNF113A	FKBP1A
CDC25A	FABP1	PNN	NCOA5	RUNX1	BUB3
TNPO2	POC1A	RAD18	USB1	ADCY6	C5orf42
PTMS	STAG3L1	BAIAP2	TPX2	MAPK13	GMNN
ANLN	ZNF593	PRIM2	HJURP	GADD45A	CASP8AP2
HMMR	HN1	RRM2	NUCKS1	MRPS18B	TGIF1
KIFC1	ITPR1	ASPHD2	USP13	CREBZF	ORC1
MCM4	ADAMTS1	HMGB3	UHRF1	SLF2	NUP37
CDK7	FAM189B	TSG101	HSF2	TRAIP	PKNOX1
MCAM	UBR7	B2M	CXCL14	PSMG3	CHML
MAN1A2	CDC25B	SPDL1	HDAC3	FANCI	MYCBP2
DNAJC3	AOC3	MORF4L2	TAB2	SYNCRIP	RPL13A
ARMC1	KMO	NFE2L2	ZNF143	EIF4E	NUF2
SLC38A2	INSIG2	RCAN1	MZF1	CFD	ABCC2
WSB1	DR1	SLBP	USP1	NPM1	HERPUD2
G2E3	DNAJB9	ME3	MND1	CCNE1	NCOA3
SGK1	TRIM45	HMGCR	MCM6	RSRC2	TOPBP1
BIVM	KIF5B	LMNA	LMNB1	CDKN3	TRMT2A
DNA2	HMG20B	SLC4A1AP	UBE2T	NPAT	FAM83D
TMEM132A	LPP	RNPS1	GINS2	NKTR	RGS3
STAG1	LBR	COL7A1	OSGIN2	TACC3	CENPA

Table 31 continued from previous page

List of Considered Cell Cycle Genes					
TFAP2A	DMTF1	HIF1A	HORMAD1	USP6NL	RRP1
LRIF1	NUP98	MID1	RANGAP1	PIK3CD	PTTG1
MCM8	VCAM1	CKAP5	LMO4	STAT5B	GTSE1
DSCC1	UNG	CDC20	PBK	UACA	CEP44
RNPC3	SAP30	KAT7	RAB3A	SMC4	GRPEL1
PRR16	VPS37C	SRSF5	POLD3	CNN2	CDKN2AIP
ESPL1	FAN1	SRSF3	MBD3	CKAP2	CDC6
AFAP1	RAB23	PPP2CA	ANP32E	ARHGEF39	ATAD2
GRK6	ANTXR1	PAK1IP1	CNIH4	BRD8	CDKL5
MIS18BP1	RAD54L	OXR1	DMXL2	DEXI	CASP3
SERPINB3	HSPB8	HRAS	SH3GL2	USP16	DET1
PSEN1	H1FO	INPP5K	OSER1	AKIRIN2	KLF9
CENPE	TTK	SMTN	LINC00339	YY1	KRAS
DNAJB1	SLC25A27	PRC1	NFIC	BORA	VTA1
SETD8P1	SLC17A2	CDC42EP4	TOP3A	KPNA2	ZCCHC10
PLIN3	HSPA8	KIF23	RPA2	ANP32B	RFC4
NFYA	MDC1	STAG3	FYN	KIF11	CFLAR
QRICH1	HIST1H2BC	G3BP1	PRKAR1A	CCNE2	ARHGAP8
PKMYT1	CDC42EP1	DCTN6	RAD51C	MRPL19	TUBD1
KIF14	MNX1	GDF15	CENPL	SAP30BP	USP53
IDO1	TOP1	SEC62	TSKU	KDM5B	PANK2
DDX11	ATL2	TXNRD1	NR3C1	CADM1	IDI2
ZPBP	MSH2	HCP5	SP1	CRYBA1	SEPN1
KIF2C	HSPA13	UBL3	FANCD2	RERE	ZMYM1
PDXP	FRZB	SSR3	FAM214A	MGAT2	HAUS5
CDCA3	INSM1	SLC22A3	ARHGAP11A	KANK2	CWC15
PLK1	UBQLN2	CENPU	ENOSF1	GNB1	PRPSAP1
CYTH3	CRK	PLK2	MITF	SRF	MTCL1
TUBA3C	NMB	TSN	NIPBL	MUC1	PPP3CA
C14orf142	TUBB2A	GAS6	CAPN7	SEPHS1	KIAA0101
FLAD1	CHEK2	CENPQ	OGT	TOB2	MEGF9
SLC25A36	ADGRE5	BIRC5	ZNF281	VCL	CIT
SLC39A10	APEX2	ZNF521	AOC2	BBS2	PCF11
MET	NUSAP1	TMPO	CKS1B	IVNS1ABP	ZNFX1
CDC16	NFIA	MCM2	PPP1R2	ARHGDIB	NDC80
RHEB	TULP4	SLC44A2	HIST1H2AC	CDCA7L	KCTD2
DUSP4	ACD	MNT	CDC7	PYM1	RHOBTB3
CHAF1B	MEPCE	JADE2	AP3M2	BUB1B	CCNA2
VPS72	POM121	ELP3	SS18	PWP1	RECQL4
SHCBP1	CTCF	E2F5	FXR1	HAUS8	FEN1
LRRC17	TIMP1	DZIP3	LARP7	BRD7	HMG1
LYAR	DSP	CYTH2	CIC	NAB1	RAD21
SRSF7	AURKB	TOMM34	CDKN2C	NBPF14	NCAPD2
SFPQ	BRCA1	ARGLU1	RHNO1	MED31	TYMS
E2F1	SMARCD1	MRI1	CNOT10	AGFG1	DYNLL1
EBI3	TROAP	UBE2S	RMI1	BMP2	CDKN1B
NUDT4	ILF2	RNF126	INTS7	NASP	CCNB2

Table 31 continued from previous page

List of Considered Cell Cycle Genes					
KIF22	KBTBD2	MRPS2	CDCA8	FANCA	TRIP13
NCAPD3	KDELC1	TTC38	KIF20B	KAT2B	AP3D1
FANCG	CDCA7	CTSD	YWHAH	DCAF16	RRM1
PASK	NCAPH	BTBD3	ABHD10	E2F8	NNMT
PTPN9	C6	RAN	CENPM	TFF3	ITGB3
CYB5R2	NLRP2	HIST2H2BE	PPP6R3	HSD17B11	CDC42
FAM64A	ZNF207	DNAJC6	ZRANB2	TXNDC9	RFC2
FADD	FEM1B	SNUPN	ASF1B	DHX8	TSC22D1
SMARCB1	KCTD9	ATF7IP	HRSP12	SUCLG2	C4B
BUB1	VPS25	DEPDC1B	REEP1	AP4B1	UBE2D3
RBM8A	H2AFX	ADGRG6	MBD2	MAD2L1	
DNAJB6	CEP70	ORC3	MAP3K2	TRA2A	

The table gives the gene names of the 628 HeLa cell cycle genes from the Whitfield [247] HeLa dataset, which were considered in our analysis.

Table 32: Cell Cycle Transcription factor

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
CENPA	ENSG00000115163	GO:0051382 :kinetochore assembly GO:0071459 :protein localization to chromosome, centromeric region GO:0016032 :viral process GO:0000281 :mitotic cytokinesis GO:0000132 :establishment of mitotic spindle orientation GO:0034080 :CENP-A containing nucleosome assembly GO:0070345 :negative regulation of fat cell proliferation GO:0060252 :positive regulation of glial cell proliferation GO:0072332 :intrinsic apoptotic signaling pathway by p53 class mediator GO:1900740 :positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway GO:0045892 :negative regulation of transcription, DNA-templated GO:0045893 :positive regulation of transcription, DNA-templated GO:0071466 :cellular response to xenobiotic stimulus
E2F1	ENSG00000101412	GO:0006977 :DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest GO:0048255 :mRNA stabilization GO:0030900 :forebrain development GO:0006351 :transcription, DNA-templated GO:0010628 :positive regulation of gene expression GO:0071398 :cellular response to fatty acid GO:0043276 :anoikis GO:0048146 :positive regulation of fibroblast proliferation GO:2000045 :regulation of G1/S transition of mitotic cell cycle GO:0000077 :DNA damage checkpoint GO:0000122 :negative regulation of transcription from RNA polymerase II promoter GO:0016032 :viral process

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
FOXM1	ENSG0000011206	GO:0043392 :negative regulation of DNA binding
		GO:0008630 :intrinsic apoptotic signaling pathway in response to DNA damage
		GO:1990086 :lens fiber cell apoptotic process
		GO:0006355 :regulation of transcription, DNA-templated
		GO:1990090 :cellular response to nerve growth factor stimulus
		GO:0000083 :regulation of transcription involved in G1/S transition of mitotic cell cycle
		GO:0071456 :cellular response to hypoxia
		GO:0070317 :negative regulation of G0 to G1 transition
		GO:0045599 :negative regulation of fat cell differentiation
		GO:0051726 :regulation of cell cycle
		GO:0043065 :positive regulation of apoptotic process
		GO:0007283 :spermatogenesis
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0071930 :negative regulation of transcription involved in G1/S transition of mitotic cell cycle
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:0071156 :regulation of cell cycle arrest
		GO:2000377 :regulation of reactive oxygen species metabolic process
		GO:0006281 :DNA repair
		GO:0008284 :positive regulation of cell proliferation
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0006978 :DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator
		GO:0001570 :vasculogenesis
		GO:0000086 :G2/M transition of mitotic cell cycle
GO:0000122 :negative regulation of transcription from RNA polymerase II promoter		
GO:0045944 :positive regulation of transcription from RNA polymerase II promoter		
GO:2000781 :positive regulation of double-strand break repair		
GO:0032873 :negative regulation of stress-activated MAPK cascade		
GO:0090344 :negative regulation of cell aging		
GO:0042127 :regulation of cell proliferation		
GO:0051726 :regulation of cell cycle		
GO:0046578 :regulation of Ras protein signal transduction		
MBD4	ENSG00000129071	GO:0032355 :response to estradiol
		GO:0045008 :depyrimidination
		GO:0006281 :DNA repair
CTCF	ENSG00000102974	GO:0035065 :regulation of histone acetylation
		GO:0010628 :positive regulation of gene expression
		GO:0031060 :regulation of histone methylation
		GO:0070602 :regulation of centromeric sister chromatid cohesion
		GO:0006306 :DNA methylation
		GO:0006349 :regulation of gene expression by genetic imprinting
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0040029 :regulation of gene expression, epigenetic
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0016584 :nucleosome positioning
		GO:0071459 :protein localization to chromosome, centromeric region
		GO:0010216 :maintenance of DNA methylation
		GO:0008285 :negative regulation of cell proliferation
		GO:0007059 :chromosome segregation
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0040030 :regulation of molecular function, epigenetic

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
E2F8	ENSG00000129173	GO:0070365 :hepatocyte differentiation
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0060718 :chorionic trophoblast cell differentiation
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0008283 :cell proliferation
		GO:0002040 :sprouting angiogenesis
		GO:0006977 :DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest
		GO:0032466 :negative regulation of cytokinesis
		GO:0032877 :positive regulation of DNA endoreduplication
		GO:0060707 :trophoblast giant cell differentiation
		GO:0001890 :placenta development
		GO:0033301 :cell cycle comprising mitosis without cytokinesis
		GO:0051726 :regulation of cell cycle
		MZF1
GO:0045944 :positive regulation of transcription from RNA polymerase II promoter		
GO:0000122 :negative regulation of transcription from RNA polymerase II promoter		
RUNX1	ENSG00000159216	GO:0001503 :ossification
		GO:0032743 :positive regulation of interleukin-2 production
		GO:0045637 :regulation of myeloid cell differentiation
		GO:0045652 :regulation of megakaryocyte differentiation
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0048935 :peripheral nervous system neuron development
		GO:0045589 :regulation of regulatory T cell differentiation
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0001959 :regulation of cytokine-mediated signaling pathway
		GO:0030182 :neuron differentiation
		GO:0043378 :positive regulation of CD8-positive, alpha-beta T cell differentiation
		GO:0030111 :regulation of Wnt signaling pathway
		GO:0050855 :regulation of B cell receptor signaling pathway
		GO:0043371 :negative regulation of CD4-positive, alpha-beta T cell differentiation
		GO:0045766 :positive regulation of angiogenesis
		GO:1902036 :regulation of hematopoietic stem cell differentiation
		GO:2000810 :regulation of bicellular tight junction assembly
		GO:0002062 :chondrocyte differentiation
		GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0010629 :negative regulation of gene expression
		GO:0071425 :hematopoietic stem cell proliferation
		GO:0030097 :hemopoiesis
GO:0030854 :positive regulation of granulocyte differentiation		
GO:0045595 :regulation of cell differentiation		
GO:0033146 :regulation of intracellular estrogen receptor signaling pathway		
GO:0045616 :regulation of keratinocyte differentiation		
MNT	ENSG00000070444	GO:0006366 :transcription from RNA polymerase II promoter
		GO:0007275 :multicellular organism development
		GO:0051726 :regulation of cell cycle
		GO:2001234 :negative regulation of apoptotic signaling pathway
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:1903508 :positive regulation of nucleic acid-templated transcription
		GO:0007569 :cell aging
GO:0008285 :negative regulation of cell proliferation		

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
HSF2	ENSG00000025156	GO:0061408 :positive regulation of transcription from RNA polymerase II promoter in response to heat stress GO:0006366 :transcription from RNA polymerase II promoter GO:0034605 :cellular response to heat GO:0045944 :positive regulation of transcription from RNA polymerase II promoter GO:0007283 :spermatogenesis GO:0043618 :regulation of transcription from RNA polymerase II promoter in response to stress
MNX1	ENSG00000130675	GO:0007417 :central nervous system development GO:0031018 :endocrine pancreas development GO:0048812 :neuron projection morphogenesis GO:0009653 :anatomical structure morphogenesis GO:0006357 :regulation of transcription from RNA polymerase II promoter GO:0006959 :humoral immune response GO:0021520 :spinal cord motor neuron cell fate specification
DMTF1	ENSG00000135164	GO:0007049 :cell cycle GO:0006355 :regulation of transcription, DNA-templated GO:0006357 :regulation of transcription from RNA polymerase II promoter
HOXB4	ENSG00000182742	GO:0045944 :positive regulation of transcription from RNA polymerase II promoter GO:0048704 :embryonic skeletal system morphogenesis GO:0009952 :anterior/posterior pattern specification GO:0008283 :cell proliferation GO:0048539 :bone marrow development GO:0000122 :negative regulation of transcription from RNA polymerase II promoter GO:0060216 :definitive hemopoiesis GO:0060218 :hematopoietic stem cell differentiation GO:0048103 :somatic stem cell division GO:0002011 :morphogenesis of an epithelial sheet GO:2000738 :positive regulation of stem cell differentiation GO:0001501 :skeletal system development GO:0048536 :spleen development
CIC	ENSG00000079432	GO:0007420 :brain development GO:0048286 :lung alveolus development GO:0007613 :memory GO:0007612 :learning GO:0000122 :negative regulation of transcription from RNA polymerase II promoter GO:0045892 :negative regulation of transcription, DNA-templated GO:0035176 :social behavior
ZNF414	ENSG00000133250	GO:0006357 :regulation of transcription from RNA polymerase II promoter
CREBZF	ENSG00000137504	GO:0006357 :regulation of transcription from RNA polymerase II promoter GO:0009615 :response to virus GO:0051090 :regulation of sequence-specific DNA binding transcription factor activity GO:0006351 :transcription, DNA-templated GO:0045814 :negative regulation of gene expression, epigenetic GO:0045892 :negative regulation of transcription, DNA-templated

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
		GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0007259 :JAK-STAT cascade
		GO:0032870 :cellular response to hormone stimulus
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0033077 :T cell differentiation in thymus
		GO:0040014 :regulation of multicellular organism growth
		GO:0042127 :regulation of cell proliferation
		GO:0045647 :negative regulation of erythrocyte differentiation
		GO:0045648 :positive regulation of erythrocyte differentiation
		GO:0038110 :interleukin-2-mediated signaling pathway
		GO:0070670 :response to interleukin-4
		GO:0007595 :lactation
		GO:0019915 :lipid storage
		GO:0030856 :regulation of epithelial cell differentiation
		GO:0045588 :positive regulation of gamma-delta T cell differentiation
		GO:0045954 :positive regulation of natural killer cell mediated cytotoxicity
		GO:0048541 :Peyer's patch development
STAT5B	ENSG00000173757	GO:0071363 :cellular response to growth factor stimulus
		GO:0050729 :positive regulation of inflammatory response
		GO:0001779 :natural killer cell differentiation
		GO:0006952 :defense response
		GO:0007565 :female pregnancy
		GO:0042448 :progesterone metabolic process
		GO:0043029 :T cell homeostasis
		GO:0097531 :mast cell migration
		GO:0032355 :response to estradiol
		GO:0042104 :positive regulation of activated T cell proliferation
		GO:0071364 :cellular response to epidermal growth factor stimulus
		GO:0019218 :regulation of steroid metabolic process
		GO:0038111 :interleukin-7-mediated signaling pathway
		GO:0001553 :luteinization
		GO:0032819 :positive regulation of natural killer cell proliferation
		GO:0040018 :positive regulation of multicellular organism growth
		GO:0043434 :response to peptide hormone
		GO:0045579 :positive regulation of B cell differentiation
		GO:0045931 :positive regulation of mitotic cell cycle
		GO:0019221 :cytokine-mediated signaling pathway
		GO:0038113 :interleukin-9-mediated signaling pathway
		GO:0035723 :interleukin-15-mediated signaling pathway
		GO:0019530 :taurine metabolic process
		GO:0032825 :positive regulation of natural killer cell differentiation
		GO:0043066 :negative regulation of apoptotic process
		GO:0045086 :positive regulation of interleukin-2 biosynthetic process
		GO:0060397 :JAK-STAT cascade involved in growth hormone signaling pathway
		GO:0046543 :development of secondary female sexual characteristics
		GO:0046544 :development of secondary male sexual characteristics
		GO:0006952 :defense response
		GO:0009612 :response to mechanical stimulus
		GO:0043124 :negative regulation of I-kappaB kinase/NF-kappaB signaling
		GO:0043434 :response to peptide hormone
		GO:0045648 :positive regulation of erythrocyte differentiation
		GO:0072136 :metanephric mesenchymal cell proliferation involved in metanephros development
		GO:0001937 :negative regulation of endothelial cell proliferation
		GO:0008015 :blood circulation
		GO:0060337 :type I interferon signaling pathway
		GO:0071407 :cellular response to organic cyclic compound

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
	ENSG00000115415	
STAT1		<p>GO:0019221 :cytokine-mediated signaling pathway</p> <p>GO:0042981 :regulation of apoptotic process</p> <p>GO:0048661 :positive regulation of smooth muscle cell proliferation</p> <p>GO:0060334 :regulation of interferon-gamma-mediated signaling pathway</p> <p>GO:0072162 :metanephric mesenchymal cell differentiation</p> <p>GO:0035456 :response to interferon-beta</p> <p>GO:0000122 :negative regulation of transcription from RNA polymerase II promoter</p> <p>GO:0007259 :JAK-STAT cascade</p> <p>GO:0016525 :negative regulation of angiogenesis</p> <p>GO:0042542 :response to hydrogen peroxide</p> <p>GO:0046725 :negative regulation by virus of viral protein levels in host cell</p> <p>GO:0060333 :interferon-gamma-mediated signaling pathway</p> <p>GO:0038113 :interleukin-9-mediated signaling pathway</p> <p>GO:0035458 :cellular response to interferon-beta</p> <p>GO:0010742 :macrophage derived foam cell differentiation</p> <p>GO:0072308 :negative regulation of metanephric nephron tubule epithelial cell differentiation</p> <p>GO:0070757 :interleukin-35-mediated signaling pathway</p> <p>GO:0032727 :positive regulation of interferon-alpha production</p> <p>GO:0032869 :cellular response to insulin stimulus</p> <p>GO:0034097 :response to cytokine</p> <p>GO:0042127 :regulation of cell proliferation</p> <p>GO:0045944 :positive regulation of transcription from RNA polymerase II promoter</p> <p>GO:0070102 :interleukin-6-mediated signaling pathway</p> <p>GO:0071346 :cellular response to interferon-gamma</p> <p>GO:0061326 :renal tubule development</p> <p>GO:0007221 :positive regulation of transcription of Notch receptor target</p> <p>GO:0016032 :viral process</p> <p>GO:0033209 :tumor necrosis factor-mediated signaling pathway</p> <p>GO:0043542 :endothelial cell migration</p> <p>GO:0038114 :interleukin-21-mediated signaling pathway</p> <p>GO:0051770 :positive regulation of nitric-oxide synthase biosynthetic process</p> <p>GO:0003340 :negative regulation of mesenchymal to epithelial transition involved in metanephros morphogenesis</p> <p>GO:0007584 :response to nutrient</p> <p>GO:0045893 :positive regulation of transcription, DNA-templated</p> <p>GO:0051591 :response to cAMP</p> <p>GO:0051607 :defense response to virus</p> <p>GO:0070106 :interleukin-27-mediated signaling pathway</p> <p>GO:0002053 :positive regulation of mesenchymal cell proliferation</p> <p>GO:0002230 :positive regulation of defense response to virus by host</p>

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
NR3C1	ENSG00000113580	GO:0006355 :regulation of transcription, DNA-templated
		GO:0071385 :cellular response to glucocorticoid stimulus
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:1902895 :positive regulation of pri-miRNA transcription from RNA polymerase II promoter
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0006366 :transcription from RNA polymerase II promoter
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0042921 :glucocorticoid receptor signaling pathway
		GO:0006325 :chromatin organization
		GO:0071383 :cellular response to steroid hormone stimulus
		GO:0071549 :cellular response to dexamethasone stimulus
		GO:0007165 :signal transduction
		GO:0071560 :cellular response to transforming growth factor beta stimulus
		NR3C1
GO:0006367 :transcription initiation from RNA polymerase II promoter		
GO:0007049 :cell cycle		
GO:0007059 :chromosome segregation		
GO:0006351 :transcription, DNA-templated		
GO:0006915 :apoptotic process		
GO:0051301 :cell division		
GO:0003404 :optic vesicle morphogenesis		
GO:0010628 :positive regulation of gene expression		
GO:0042127 :regulation of cell proliferation		
GO:0045944 :positive regulation of transcription from RNA polymerase II promoter		
GO:0003409 :optic cup structural organization		
GO:0006357 :regulation of transcription from RNA polymerase II promoter		
GO:0035115 :embryonic forelimb morphogenesis		
GO:0070172 :positive regulation of tooth mineralization		
GO:2000378 :negative regulation of reactive oxygen species metabolic process		
GO:0010842 :retina layer formation		
GO:0010944 :negative regulation of transcription by competitive promoter binding		
GO:0060021 :palate development		
TFAP2A		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0021559 :trigeminal nerve development
		GO:0001822 :kidney development
		GO:0008285 :negative regulation of cell proliferation
		GO:0043525 :positive regulation of neuron apoptotic process
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0060349 :bone morphogenesis
		GO:0042472 :inner ear morphogenesis
		GO:0045595 :regulation of cell differentiation
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:0048701 :embryonic cranial skeleton morphogenesis
		GO:0048856 :anatomical structure development
		GO:0061029 :eyelid development in camera-type eye
		GO:0071281 :cellular response to iron ion
GO:0021623 :oculomotor nerve formation		
GO:0007605 :sensory perception of sound		
GO:0043066 :negative regulation of apoptotic process		
GO:0030501 :positive regulation of bone mineralization		
KLF6	ENSG00000067082	GO:0030183 :B cell differentiation
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0045893 :positive regulation of transcription, DNA-templated

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function	
PKNOX1	ENSG00000160199	GO:0001525 :angiogenesis	
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter	
		GO:0030218 :erythrocyte differentiation	
		GO:0043010 :camera-type eye development	
		GO:0006366 :transcription from RNA polymerase II promoter	
ZNF587	ENSG00000198466	GO:0030217 :T cell differentiation	
		GO:0006355 :regulation of transcription, DNA-templated	
E2F5	ENSG00000133740	GO:0030030 :cell projection organization	
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter	
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter	
		GO:0009887 :animal organ morphogenesis	
KLF9	ENSG00000119138	GO:0051726 :regulation of cell cycle	
		GO:0006357 :regulation of transcription from RNA polymerase II promoter	
		GO:0071387 :cellular response to cortisol stimulus	
		GO:0097067 :cellular response to thyroid hormone stimulus	
		GO:0007623 :circadian rhythm	
DR1	ENSG00000117505	GO:0045944 :positive regulation of transcription from RNA polymerase II promoter	
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter	
		GO:0006338 :chromatin remodeling	
		GO:0045898 :regulation of RNA polymerase II transcriptional preinitiation complex assembly	
		GO:0043966 :histone H3 acetylation	
		ENSG00000112658	GO:0030220 :platelet formation
			GO:0046716 :muscle cell cellular homeostasis
	GO:0051150 :regulation of smooth muscle cell differentiation		
	GO:0060055 :angiogenesis involved in wound healing		
	GO:1902894 :negative regulation of pri-miRNA transcription from RNA polymerase II promoter		
	GO:0030168 :platelet activation		
	GO:0031175 :neuron projection development		
	GO:0060947 :cardiac vascular smooth muscle cell differentiation		
	GO:0001829 :trophodermal cell differentiation		
	GO:0007616 :long-term memory		
	GO:0008285 :negative regulation of cell proliferation		
	GO:0035855 :megakaryocyte development		
GO:0035912 :dorsal aorta morphogenesis			
GO:0045597 :positive regulation of cell differentiation			
GO:0045944 :positive regulation of transcription from RNA polymerase II promoter			
GO:0046016 :positive regulation of transcription by glucose			
GO:0048538 :thymus development			
GO:0070830 :bicellular tight junction assembly			
GO:0071333 :cellular response to glucose stimulus			
GO:0090398 :cellular senescence			
GO:1902895 :positive regulation of pri-miRNA transcription from RNA polymerase II promoter			
GO:0021766 :hippocampus development			
GO:0033561 :regulation of water loss via skin			
GO:0060425 :lung morphogenesis			

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
SRF		GO:0001569 :branching involved in blood vessel morphogenesis
		GO:0007507 :heart development
		GO:0008306 :associative learning
		GO:0034097 :response to cytokine
		GO:0045987 :positive regulation of smooth muscle contraction
		GO:0030155 :regulation of cell adhesion
		GO:0051491 :positive regulation of filopodium assembly
		GO:0001707 :mesoderm formation
		GO:0001764 :neuron migration
		GO:0002011 :morphogenesis of an epithelial sheet
		GO:0048666 :neuron development
		GO:0060532 :bronchus cartilage development
		GO:0060534 :trachea cartilage development
		GO:0009725 :response to hormone
		GO:0001666 :response to hypoxia
		GO:0010669 :epithelial structure maintenance
		GO:0010735 :positive regulation of transcription via serum response element binding
		GO:0030878 :thyroid gland development
		GO:0043149 :stress fiber assembly
		GO:0045773 :positive regulation of axon extension
		GO:0060347 :heart trabecula formation
		GO:0061029 :eyelid development in camera-type eye
		GO:0045059 :positive thymic T cell selection
		GO:0090009 :primitive streak formation
		GO:0030036 :actin cytoskeleton organization
		GO:0060292 :long term synaptic depression
		GO:0061145 :lung smooth muscle development
		GO:0030336 :negative regulation of cell migration
		GO:0007160 :cell-matrix adhesion
		GO:1900222 :negative regulation of beta-amyloid clearance
		GO:0002042 :cell migration involved in sprouting angiogenesis
		GO:0051091 :positive regulation of sequence-specific DNA binding transcription factor activity
		GO:0060218 :hematopoietic stem cell differentiation
		GO:0060324 :face development
		GO:0003257 :positive regulation of transcription from RNA polymerase II promoter involved in myocardial precursor cell differentiation
		GO:0042789 :mRNA transcription from RNA polymerase II promoter
GO:0045214 :sarcomere organization		
GO:0001947 :heart looping		
GO:0009636 :response to toxic substance		
GO:0043589 :skin morphogenesis		
GO:0048589 :developmental growth		
GO:0048821 :erythrocyte development		
GO:0060261 :positive regulation of transcription initiation from RNA polymerase II promoter		
GO:0022028 :tangential migration from the subventricular zone to the olfactory bulb		
GO:0090136 :epithelial cell-cell adhesion		
GO:0055003 :cardiac myofibril assembly		
HMG20B	ENSG00000064961	GO:0010468 :regulation of gene expression
		GO:0045666 :positive regulation of neuron differentiation
		GO:0035914 :skeletal muscle cell differentiation
		GO:0006325 :chromatin organization
		GO:0007049 :cell cycle
		GO:0033234 :negative regulation of protein sumoylation
TGIF1	ENSG00000177426	GO:0007596 :blood coagulation
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0042493 :response to drug
		GO:0071363 :cellular response to growth factor stimulus
		GO:0007275 :multicellular organism development

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
INSM1	ENSG00000173404	GO:0060290 :transdifferentiation
		GO:0061549 :sympathetic ganglion development
		GO:0003323 :type B pancreatic cell development
		GO:2000179 :positive regulation of neural precursor cell proliferation
		GO:0008285 :negative regulation of cell proliferation
		GO:0030182 :neuron differentiation
		GO:0045597 :positive regulation of cell differentiation
		GO:0043254 :regulation of protein complex assembly
		GO:0071158 :positive regulation of cell cycle arrest
		GO:0003358 :noradrenergic neuron development
		GO:0031018 :endocrine pancreas development
		GO:0042421 :norepinephrine biosynthetic process
		GO:0007049 :cell cycle
		GO:0010468 :regulation of gene expression
		GO:0061104 :adrenal chromaffin cell differentiation
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0003309 :type B pancreatic cell differentiation
		GO:0001933 :negative regulation of protein phosphorylation
		GO:0030335 :positive regulation of cell migration
		GO:0010564 :regulation of cell cycle process
ZNF521	ENSG00000198795	GO:0003310 :pancreatic A cell differentiation
		GO:0048663 :neuron fate commitment
ZNF207	ENSG00000010244	GO:0006355 :regulation of transcription, DNA-templated
		GO:0050821 :protein stabilization
KDM5B	ENSG00000117139	GO:0006355 :regulation of transcription, DNA-templated
		GO:0007094 :mitotic spindle assembly checkpoint
		GO:0051301 :cell division
		GO:0008608 :attachment of spindle microtubules to kinetochore
		GO:0000070 :mitotic sister chromatid segregation
		GO:0090307 :mitotic spindle assembly
		GO:0001578 :microtubule bundle formation
		GO:0046785 :microtubule polymerization
		GO:0051983 :regulation of chromosome segregation
		GO:0006338 :chromatin remodeling
		GO:0009791 :post-embryonic development
		GO:0060444 :branching involved in mammary gland duct morphogenesis
		GO:0060763 :mammary duct terminal end bud growth
		GO:0060992 :response to fungicide
		GO:0010628 :positive regulation of gene expression
		GO:0034720 :histone H3-K4 demethylation
		GO:0055114 :oxidation-reduction process
		GO:2000864 :regulation of estradiol secretion
		GO:1990830 :cellular response to leukemia inhibitory factor
		GO:0007338 :single fertilization
GO:0048511 :rhythmic process		
GO:0045892 :negative regulation of transcription, DNA-templated		
GO:0033601 :positive regulation of mammary gland epithelial cell proliferation		
GO:0044344 :cellular response to fibroblast growth factor stimulus		
GO:0061038 :uterus morphogenesis		
GO:0070306 :lens fiber cell differentiation		
GO:0006357 :regulation of transcription from RNA polymerase II promoter		
NFIC	ENSG00000141905	GO:0034721 :histone H3-K4 demethylation, trimethyl-H3-K4-specific
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0006260 :DNA replication
		GO:0006366 :transcription from RNA polymerase II promoter
		GO:0042475 :odontogenesis of dentin-containing tooth
GO:0045944 :positive regulation of transcription from RNA polymerase II promoter		

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
TSC22D1	ENSG00000102804	GO:0006357 :regulation of transcription from RNA polymerase II promoter
ZNF281	ENSG00000162702	GO:0006366 :transcription from RNA polymerase II promoter GO:0000122 :negative regulation of transcription from RNA polymerase II promoter GO:0045892 :negative regulation of transcription, DNA-templated GO:0006355 :regulation of transcription, DNA-templated GO:0048863 :stem cell differentiation GO:0010172 :embryonic body morphogenesis GO:0010629 :negative regulation of gene expression GO:0045893 :positive regulation of transcription, DNA-templated
HIF1A	ENSG00000100644	GO:0006357 :regulation of transcription from RNA polymerase II promoter GO:0006366 :transcription from RNA polymerase II promoter GO:0035162 :embryonic hemopoiesis GO:0045926 :negative regulation of growth GO:0071456 :cellular response to hypoxia GO:2000378 :negative regulation of reactive oxygen species metabolic process GO:0035774 :positive regulation of insulin secretion involved in cellular response to glucose stimulus GO:0043687 :post-translational protein modification GO:0070244 :negative regulation of thymocyte apoptotic process GO:0032364 :oxygen homeostasis GO:0001755 :neural crest cell migration GO:0006089 :lactate metabolic process GO:1902895 :positive regulation of pri-miRNA transcription from RNA polymerase II promoter GO:0014850 :response to muscle activity GO:0032963 :collagen metabolic process GO:0021987 :cerebral cortex development GO:0006355 :regulation of transcription, DNA-templated GO:0010629 :negative regulation of gene expression GO:0010575 :positive regulation of vascular endothelial growth factor production GO:0060574 :intestinal epithelial cell maturation GO:1903377 :negative regulation of oxidative stress-induced neuron intrinsic apoptotic signaling pathway GO:0019221 :cytokine-mediated signaling pathway GO:0001666 :response to hypoxia GO:0007165 :signal transduction GO:0016567 :protein ubiquitination GO:0010039 :response to iron ion GO:0051541 :elastin metabolic process GO:0070101 :positive regulation of chemokine-mediated signaling pathway GO:0010870 :positive regulation of receptor biosynthetic process GO:0021502 :neural fold elevation formation GO:0051000 :positive regulation of nitric-oxide synthase activity GO:0061418 :regulation of transcription from RNA polymerase II promoter in response to hypoxia GO:2001054 :negative regulation of mesenchymal cell apoptotic process GO:0016239 :positive regulation of macroautophagy GO:0032722 :positive regulation of chemokine production GO:0043536 :positive regulation of blood vessel endothelial cell migration GO:0071347 :cellular response to interleukin-1 GO:0071542 :dopaminergic neuron differentiation GO:0001837 :epithelial to mesenchymal transition GO:0001922 :B-1 B cell homeostasis GO:0003208 :cardiac ventricle morphogenesis

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
		GO:0097411 :hypoxia-inducible factor-1alpha signaling pathway
		GO:1903715 :regulation of aerobic respiration
		GO:0010468 :regulation of gene expression
		GO:0010634 :positive regulation of epithelial cell migration
		GO:0016579 :protein deubiquitination
		GO:0002248 :connective tissue replacement involved in inflammatory response wound healing
		GO:0042593 :glucose homeostasis
		GO:0043619 :regulation of transcription from RNA polymerase II promoter in response to oxidative stress
		GO:0046716 :muscle cell cellular homeostasis
		GO:1903599 :positive regulation of mitophagy
		GO:0001892 :embryonic placenta development
		GO:0001947 :heart looping
		GO:0002052 :positive regulation of neuroblast proliferation
		GO:0003151 :outflow tract morphogenesis
		GO:0061419 :positive regulation of transcription from RNA polymerase II promoter in response to hypoxia
		GO:0007595 :lactation
		GO:0010628 :positive regulation of gene expression
		GO:0045648 :positive regulation of erythrocyte differentiation
		GO:0045821 :positive regulation of glycolytic process
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0046886 :positive regulation of hormone biosynthetic process
		GO:0051216 :cartilage development
		GO:0061072 :iris morphogenesis
		GO:0001938 :positive regulation of endothelial cell proliferation
		GO:0030502 :negative regulation of bone mineralization
		GO:0001525 :angiogenesis
		GO:0045766 :positive regulation of angiogenesis
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0061030 :epithelial cell differentiation involved in mammary gland alveolus development
		GO:0019896 :axonal transport of mitochondrion
		GO:0010573 :vascular endothelial growth factor production
		GO:0008542 :visual learning
		GO:0030949 :positive regulation of vascular endothelial growth factor receptor signaling pathway
		GO:0032909 :regulation of transforming growth factor beta2 production
		GO:0042541 :hemoglobin biosynthetic process
		GO:0048546 :digestive tract morphogenesis
		GO:0006879 :cellular iron ion homeostasis
		GO:0032007 :negative regulation of TOR signaling
		GO:0061298 :retina vasculature development in camera-type eye
NCOA3	ENSG00000124151	GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0030521 :androgen receptor signaling pathway
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0035624 :receptor transactivation
		GO:0045618 :positive regulation of keratinocyte differentiation
		GO:1902459 :positive regulation of stem cell population maintenance
		GO:0071392 :cellular response to estradiol stimulus
		GO:0010628 :positive regulation of gene expression
		GO:0032870 :cellular response to hormone stimulus
		GO:2000035 :regulation of stem cell division
		GO:0016573 :histone acetylation
		GO:2001141 :regulation of RNA biosynthetic process
		GO:0043697 :cell dedifferentiation

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
ZBTB7A	ENSG00000178951	GO:2000677 :regulation of transcription regulatory region DNA binding GO:0097680 :double-strand break repair via classical nonhomologous end joining GO:0045444 :fat cell differentiation GO:0006355 :regulation of transcription, DNA-templated GO:0045892 :negative regulation of transcription, DNA-templated GO:0006110 :regulation of glycolytic process GO:0006338 :chromatin remodeling GO:0006974 :cellular response to DNA damage stimulus GO:0060766 :negative regulation of androgen receptor signaling pathway GO:0034504 :protein localization to nucleus GO:0043249 :erythrocyte maturation GO:0000381 :regulation of alternative mRNA splicing, via spliceosome GO:0006325 :chromatin organization GO:0006351 :transcription, DNA-templated GO:0045746 :negative regulation of Notch signaling pathway GO:0051090 :regulation of sequence-specific DNA binding transcription factor activity GO:0051092 :positive regulation of NF-kappaB transcription factor activity GO:0030183 :B cell differentiation GO:0042981 :regulation of apoptotic process GO:0000122 :negative regulation of transcription from RNA polymerase II promoter GO:0030512 :negative regulation of transforming growth factor beta receptor signaling pathway
NFIA	ENSG00000162599	GO:0006260 :DNA replication GO:0045944 :positive regulation of transcription from RNA polymerase II promoter GO:0060074 :synapse maturation GO:0000122 :negative regulation of transcription from RNA polymerase II promoter GO:0006355 :regulation of transcription, DNA-templated GO:0072189 :ureter development GO:0019079 :viral genome replication
NFE2L2	ENSG00000116044	GO:0070301 :cellular response to hydrogen peroxide GO:0071498 :cellular response to fluid shear stress GO:1902176 :negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway GO:0046326 :positive regulation of glucose import GO:0034599 :cellular response to oxidative stress GO:0045995 :regulation of embryonic development GO:2000352 :negative regulation of endothelial cell apoptotic process GO:0007568 :aging GO:0010499 :proteasomal ubiquitin-independent protein catabolic process GO:0071499 :cellular response to laminar fluid shear stress GO:2000379 :positive regulation of reactive oxygen species metabolic process GO:0030968 :endoplasmic reticulum unfolded protein response GO:0045454 :cell redox homeostasis GO:0010976 :positive regulation of neuron projection development GO:0061419 :positive regulation of transcription from RNA polymerase II promoter in response to hypoxia

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
		GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0016032 :viral process
		GO:0043536 :positive regulation of blood vessel endothelial cell migration
		GO:0045766 :positive regulation of angiogenesis
		GO:0042149 :cellular response to glucose starvation
		GO:0006366 :transcription from RNA polymerase II promoter
		GO:0010628 :positive regulation of gene expression
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:1904753 :negative regulation of vascular associated smooth muscle cell migration
		GO:1903206 :negative regulation of hydrogen peroxide-induced cell death
		GO:2000121 :regulation of removal of superoxide radicals
		GO:0046223 :aflatoxin catabolic process
		GO:1903071 :positive regulation of ER-associated ubiquitin-dependent protein catabolic process
		GO:1904385 :cellular response to angiotensin
		GO:0006954 :inflammatory response
		GO:0030194 :positive regulation of blood coagulation
		GO:0071356 :cellular response to tumor necrosis factor
		GO:1902037 :negative regulation of hematopoietic stem cell differentiation
		GO:0036003 :positive regulation of transcription from RNA polymerase II promoter in response to stress
		GO:0036091 :positive regulation of transcription from RNA polymerase II promoter in response to oxidative stress
		GO:0010667 :negative regulation of cardiac muscle cell apoptotic process
		GO:0016567 :protein ubiquitination
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0036499 :PERK-mediated unfolded protein response
		GO:0043161 :proteasome-mediated ubiquitin-dependent protein catabolic process
		GO:0071280 :cellular response to copper ion
ZNF217	ENSG00000171940	GO:1903788 :positive regulation of glutathione biosynthetic process
		GO:0006351 :transcription, DNA-templated
		GO:0006355 :regulation of transcription, DNA-templated
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
ZBED5	ENSG00000236287	GO:0006357 :regulation of transcription from RNA polymerase II promoter
MITF	ENSG00000187098	GO:0006351 :transcription, DNA-templated
		GO:0044336 :canonical Wnt signaling pathway involved in negative regulation of apoptotic process
		GO:0042127 :regulation of cell proliferation
		GO:0030316 :osteoclast differentiation
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0006355 :regulation of transcription, DNA-templated
		GO:0010628 :positive regulation of gene expression
		GO:0045670 :regulation of osteoclast differentiation
		GO:2000144 :positive regulation of DNA-templated transcription, initiation
		GO:0030336 :negative regulation of cell migration
		GO:0045165 :cell fate commitment

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
GTF2B	ENSG00000137947	GO:0046849 :bone remodeling
		GO:0030318 :melanocyte differentiation
		GO:2001141 :regulation of RNA biosynthetic process
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0043010 :camera-type eye development
		GO:0065003 :macromolecular complex assembly
		GO:0001174 :transcriptional start site selection at RNA polymerase II promoter
		GO:0042795 :snRNA transcription from RNA polymerase II promoter
		GO:0043923 :positive regulation by host of viral transcription
		GO:0006367 :transcription initiation from RNA polymerase II promoter
		GO:0006473 :protein acetylation
		GO:0050434 :positive regulation of viral transcription
		GO:0006352 :DNA-templated transcription, initiation
		GO:0016573 :histone acetylation
		GO:0060261 :positive regulation of transcription initiation from RNA polymerase II promoter
		GO:1904798 :positive regulation of core promoter binding
GO:0006366 :transcription from RNA polymerase II promoter		
GO:0016032 :viral process		
GO:0051123 :RNA polymerase II transcriptional preinitiation complex assembly		
YY1	ENSG00000100811	GO:1990114 :RNA Polymerase II core complex assembly
		GO:0000724 :double-strand break repair via homologous recombination
		GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0034644 :cellular response to UV
		GO:0034696 :response to prostaglandin F
		GO:0051276 :chromosome organization
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0006403 :RNA localization
		GO:0010629 :negative regulation of gene expression
		GO:0061052 :negative regulation of cell growth involved in cardiac muscle cell development
		GO:0007283 :spermatogenesis
		GO:0010225 :response to UV-C
		GO:0006974 :cellular response to DNA damage stimulus
		GO:0010467 :gene expression
		GO:0016579 :protein deubiquitination
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0030154 :cell differentiation
GO:0032688 :negative regulation of interferon-beta production		
GO:0009952 :anterior/posterior pattern specification		
GO:0048593 :camera-type eye morphogenesis		
GO:1902894 :negative regulation of pri-miRNA transcription from RNA polymerase II promoter		
GO:0071347 :cellular response to interleukin-1		
KAT7	ENSG00000136504	GO:0006281 :DNA repair
		GO:0006260 :DNA replication
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0072708 :response to sorbitol
		GO:0072739 :response to anisomycin
		GO:0090240 :positive regulation of histone H4 acetylation
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:0043966 :histone H3 acetylation
		GO:1902035 :positive regulation of hematopoietic stem cell proliferation
		GO:0072720 :response to dithiothreitol
		GO:2000819 :regulation of nucleotide-excision repair

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
		GO:0031098 :stress-activated protein kinase signaling cascade
		GO:0045740 :positive regulation of DNA replication
		GO:0001779 :natural killer cell differentiation
		GO:0018393 :internal peptidyl-lysine acetylation
		GO:0043982 :histone H4-K8 acetylation
		GO:0043984 :histone H4-K16 acetylation
		GO:1900182 :positive regulation of protein localization to nucleus
		GO:0043967 :histone H4 acetylation
		GO:0072710 :response to hydroxyurea
		GO:0072716 :response to actinomycin D
		GO:0006355 :regulation of transcription, DNA-templated
		GO:0044154 :histone H3-K14 acetylation
		GO:0030174 :regulation of DNA-dependent DNA replication initiation
		GO:0043981 :histone H4-K5 acetylation
		GO:0045648 :positive regulation of erythrocyte differentiation
		GO:0032786 :positive regulation of DNA-templated transcription, elongation
		GO:0043983 :histone H4-K12 acetylation
MBD3	ENSG00000071655	GO:0031667 :response to nutrient levels
		GO:0048568 :embryonic organ development
		GO:0007568 :aging
		GO:0044030 :regulation of DNA methylation
		GO:0001701 :in utero embryonic development
		GO:0016573 :histone acetylation
		GO:0007420 :brain development
		GO:0043044 :ATP-dependent chromatin remodeling
		GO:0007507 :heart development
		GO:0009888 :tissue development
		GO:1901796 :regulation of signal transduction by p53 class mediator
		GO:0032355 :response to estradiol
SP1	ENSG00000185591	GO:0010628 :positive regulation of gene expression
		GO:0043923 :positive regulation by host of viral transcription
		GO:0048511 :rhythmic process
		GO:1904828 :positive regulation of hydrogen sulfide biosynthetic process
		GO:0016032 :viral process
		GO:0006355 :regulation of transcription, DNA-templated
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0033194 :response to hydroperoxide
		GO:1905564 :positive regulation of vascular endothelial cell proliferation
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0043536 :positive regulation of blood vessel endothelial cell migration
		GO:0042795 :snRNA transcription from RNA polymerase II promoter
		GO:0045766 :positive regulation of angiogenesis
		GO:0045540 :regulation of cholesterol biosynthetic process
		GO:1902004 :positive regulation of beta-amyloid formation
		GO:0032869 :cellular response to insulin stimulus
NFYB	ENSG00000120837	GO:0006355 :regulation of transcription, DNA-templated
		GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0045540 :regulation of cholesterol biosynthetic process
		GO:1990830 :cellular response to leukemia inhibitory factor
		GO:0045893 :positive regulation of transcription, DNA-templated
NFYA	ENSG00000001167	GO:0006355 :regulation of transcription, DNA-templated
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0010723 :positive regulation of transcription from RNA polymerase II promoter in response to iron
		GO:0045540 :regulation of cholesterol biosynthetic process
		GO:0006366 :transcription from RNA polymerase II promoter
		GO:0048511 :rhythmic process

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
MBD2	ENSG00000134046	GO:0042711 :maternal behavior
		GO:0000122 :negative regulation of transcription from RNA polymerase II promoter
		GO:0000183 :chromatin silencing at rDNA
		GO:0031667 :response to nutrient levels
		GO:0043044 :ATP-dependent chromatin remodeling
		GO:0042127 :regulation of cell proliferation
		GO:0009612 :response to mechanical stimulus
		GO:0034622 :cellular macromolecular complex assembly
		GO:0048568 :embryonic organ development
		GO:0007507 :heart development
		GO:0071407 :cellular response to organic cyclic compound
		GO:0035563 :positive regulation of chromatin binding
		GO:0006346 :methylation-dependent chromatin silencing
		GO:0007568 :aging
		GO:0030177 :positive regulation of Wnt signaling pathway
		GO:0044030 :regulation of DNA methylation
		GO:0032355 :response to estradiol
ZNF143	ENSG00000166478	GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0042795 :snRNA transcription from RNA polymerase II promoter
		GO:0006355 :regulation of transcription, DNA-templated
		GO:0006383 :transcription from RNA polymerase III promoter
		GO:0006366 :transcription from RNA polymerase II promoter
		GO:0006359 :regulation of transcription from RNA polymerase III promoter
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
BRCA1	ENSG00000012048	GO:0006974 :cellular response to DNA damage stimulus
		GO:0006978 :DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator
		GO:0042127 :regulation of cell proliferation
		GO:0045892 :negative regulation of transcription, DNA-templated
		GO:0006633 :fatty acid biosynthetic process
		GO:0051865 :protein autoubiquitination
		GO:0006301 :postreplication repair
		GO:0000729 :DNA double-strand break processing
		GO:0035066 :positive regulation of histone acetylation
		GO:0042981 :regulation of apoptotic process
		GO:0070317 :negative regulation of G0 to G1 transition
		GO:0006260 :DNA replication
		GO:0006349 :regulation of gene expression by genetic imprinting
		GO:0010212 :response to ionizing radiation
		GO:0016579 :protein deubiquitination
		GO:0045893 :positive regulation of transcription, DNA-templated
		GO:0046600 :negative regulation of centriole replication
		GO:0071158 :positive regulation of cell cycle arrest
		GO:0006357 :regulation of transcription from RNA polymerase II promoter
		GO:0007098 :centrosome cycle
		GO:0010575 :positive regulation of vascular endothelial growth factor production
		GO:0000724 :double-strand break repair via homologous recombination
GO:0051571 :positive regulation of histone H3-K4 methylation		
GO:0051572 :negative regulation of histone H3-K4 methylation		
GO:0071356 :cellular response to tumor necrosis factor		
GO:1902042 :negative regulation of extrinsic apoptotic signaling pathway via death domain receptors		

Table 32 continued from previous page

Gene Symbol	Ensembl Gene ID	GO - Molecular Function
		GO:2000617 :positive regulation of histone H3-K9 acetylation
		GO:0006359 :regulation of transcription from RNA polymerase III promoter
		GO:0016567 :protein ubiquitination
		GO:0008630 :intrinsic apoptotic signaling pathway in response to DNA damage
		GO:0031398 :positive regulation of protein ubiquitination
		GO:0033147 :negative regulation of intracellular estrogen receptor signaling pathway
		GO:0043009 :chordate embryonic development
		GO:0051573 :negative regulation of histone H3-K9 methylation
		GO:0051574 :positive regulation of histone H3-K9 methylation
		GO:0071681 :cellular response to indole-3-methanol
		GO:0085020 :protein K6-linked ubiquitination
		GO:0006302 :double-strand break repair
		GO:0006303 :double-strand break repair via nonhomologous end joining
		GO:0006915 :apoptotic process
		GO:0010628 :positive regulation of gene expression
		GO:0045944 :positive regulation of transcription from RNA polymerase II promoter
		GO:0070512 :positive regulation of histone H4-K20 methylation
		GO:0009048 :dosage compensation by inactivation of X chromosome
		GO:0044818 :mitotic G2/M transition checkpoint
		GO:0035067 :negative regulation of histone acetylation
		GO:0044030 :regulation of DNA methylation
		GO:0045717 :negative regulation of fatty acid biosynthetic process
		GO:0045739 :positive regulation of DNA repair
		GO:0045766 :positive regulation of angiogenesis
		GO:2000378 :negative regulation of reactive oxygen species metabolic process
		GO:0007059 :chromosome segregation
		GO:0043627 :response to estrogen
		GO:0030521 :androgen receptor signaling pathway
		GO:1901796 :regulation of signal transduction by p53 class mediator
		GO:0072425 :signal transduction involved in G2 DNA damage checkpoint
		GO:2000620 :positive regulation of histone H4-K16 acetylation

The table reports the list of TFs that are considered in our analysis. The 1st column reports the TF official name. The 2nd column report the TF Ensemble ID and, finally, the 3rd column gives the TF gene ontology annotation. More specifically, we report the biological process.

Table 33: Knockdown Data from Gene Expression Omnibus

Transcription Factor	GEO dataset ID
NFYA	GSE40215
NFE2L2	GSE38332
MITF	GSE16249
KAT7	GSE33220
ZNF521	GSE79110
MBD4	GSE52567
BRCA1	GSE54265
CTCF	GSE108869

Table 33 continued from previous page

Transcription Factor	GEO dataset ID
SP1	GSE37935
ZNF217	GSE35511
SRF	GSE22606

The table provides the list of KD expression datasets manually downloaded from **Gene Expression Omnibus (GEO)** database. The 1st column gives the name of the TF targeted by the KD experiment. The 2nd column provides the ID of the dataset.

Table 34: Edges repetition in networks from HumanBase

TF	TG	# dup
3091	55655	77
672	63967	77
672	51514	76
6777	3111	76
861	4233	76
2305	890	75
3091	729	75
5316	729	75
861	729	75
2305	23779	74
2305	4233	74
6722	729	74
861	10234	74
3091	7424	73
6667	4233	73
672	10595	73
672	580	73
5316	3111	71
672	4751	71
861	4286	71
3091	6491	70
672	3070	70
6777	10595	70
8202	3070	70
5316	2619	69
672	11065	69
672	5347	69
7050	1869	69
7528	9319	68
3091	4233	67
672	22909	67
672	55632	67
672	9319	67
6722	6491	67
6777	55655	67
8930	10595	67
7528	2619	66
8932	6382	66
6667	4286	65

Table 34 continued

TF	TG	# dup
672	5888	64
6722	3214	64
6777	332	64
6777	3690	64
7050	4286	64
10664	55723	63
58487	3434	63
7050	6491	63
7528	22909	63
7528	5888	63
2908	332	62
5316	4286	62
6777	3070	62
7050	2619	62
7528	993	62
8932	4286	62
6772	3434	61
7050	3110	61
6722	3690	60
7528	4286	60
3091	2487	59
6667	2619	59
6667	29899	59
672	2177	59
672	23354	59
6777	7412	59
6667	7020	58
672	9666	57
672	9682	57
3091	7412	56
7528	51514	56
8932	898	56
7528	650	55
10765	3070	54
3091	650	54
6667	3214	54
672	11200	54
1869	10492	53
6667	5864	53
1810	3070	52
8202	9682	52
10664	3070	51
6667	7412	51
861	2487	51
8930	11200	51

Table 34 continued

TF	TG	# dup
11143	3070	50
6667	2487	50
6722	650	49
2305	2487	48
53615	3070	48
7528	3214	48
2305	4281	46
6777	9510	46
7020	3214	45
8932	650	45
6777	11200	42
4286	2487	41
10765	9682	40
1810	9682	40
10664	9682	37
53615	9682	37
7528	56852	37
8932	2487	37
6667	55632	36
11143	9682	34
7020	3710	34
6722	55632	33
7020	2619	33
7020	2487	31
3214	7020	25
7020	650	23
4286	8932	20
4286	10766	18
5316	55790	18
7050	55790	17
10664	55632	16
2908	51170	16
6777	5293	16
7020	51141	16
1810	55632	15
3091	51170	15
6777	51170	15
6777	55790	15
7528	51170	15
8932	55790	15
10765	55632	14
5316	51170	14
8202	55632	14
3110	51715	13
3214	650	13

Table 34 continued

TF	TG	# dup
4286	10036	13
5316	5293	13
7020	7251	13
7050	51170	13
861	51170	13
1869	1977	12
1869	5422	12
3110	54820	12
4286	2782	12
4286	3337	12
4286	51343	12
4286	7050	12
5316	4863	12
6722	51170	12
6772	51170	12
1810	55790	11
2908	55790	11
3091	5293	11
3214	6667	11
4286	10347	11
4286	6667	11
4286	7296	11
4286	8473	11
5316	7412	11
7020	1832	11
7020	3337	11
7528	55632	11
861	5293	11
8930	51170	11
2908	699	10
3091	127	10
3110	1058	10
4286	2189	10
4286	687	10
4286	7528	10
4286	994	10
5316	3434	10
53615	55632	10
6667	23397	10
7020	1509	10
8932	51170	10
8932	65055	10
11143	55632	9
2305	3110	9
2305	3111	9

Table 34 continued

TF	TG	# dup
2908	898	9
3091	65055	9
3110	11104	9
3214	6118	9
3214	7528	9
4286	10051	9
4286	10905	9
4286	1500	9
4286	1810	9
4286	53615	9
4286	6944	9
4286	7374	9
4286	9134	9
4286	9531	9
4286	9646	9
4286	9793	9
5316	10068	9
5316	5347	9
53615	51514	9
6667	701	9
672	701	9
6722	5293	9
6722	65055	9
6772	55723	9
7050	55655	9
7528	701	9
8202	29899	9
8202	51170	9
8202	9212	9
8932	5293	9
8932	699	9
10664	6382	8
1810	10234	8
1869	10036	8
1869	29117	8
1869	3337	8
2305	51170	8
2305	5293	8
2908	65055	8
3091	3070	8
3110	55729	8
3110	7050	8
3214	10492	8
4286	10460	8
4286	11073	8

Table 34 continued

TF	TG	# dup
4286	1509	8
4286	1832	8
4286	23310	8
4286	3156	8
4286	3832	8
4286	3837	8
4286	4664	8
4286	4678	8
4286	473	8
4286	51715	8
4286	51763	8
4286	5316	8
4286	54407	8
4286	54820	8
4286	5530	8
4286	5573	8
4286	5608	8
4286	5663	8
4286	5780	8
4286	6241	8
4286	7251	8
4286	7298	8
4286	861	8
4286	996	8
4286	998	8
4780	10595	8
53615	55790	8
53615	7424	8
58487	9212	8
6667	10068	8
6667	51170	8
672	10234	8
6722	10234	8
6772	3070	8
6777	4233	8
6777	6491	8
6777	898	8
7020	1058	8
7020	4247	8
7050	11200	8
7050	23118	8
7050	23354	8
7050	23779	8
7050	4863	8
7528	6491	8

Table 34 continued

TF	TG	# dup
8202	55247	8
861	3642	8
8930	9212	8
8932	10234	8
8932	10873	8
10664	10234	7
10664	9212	7
10765	10234	7
10765	127	7
10765	29899	7
10765	9212	7
11143	10234	7
11143	6382	7
1810	2487	7
1810	51170	7
1810	890	7
1810	9319	7
1869	1031	7
1869	2177	7
1869	993	7
2305	10873	7
2305	65055	7
2908	23397	7
2908	5293	7
2908	55247	7
3091	9212	7
3110	23077	7
3110	3337	7
3110	6421	7
3110	9939	7
3214	10347	7
3214	10765	7
3214	11004	7
3214	1509	7
3214	22894	7
3214	23279	7
3214	25836	7
3214	5515	7
3214	5573	7
3214	5889	7
3214	6760	7
3214	6944	7
3214	7296	7
3214	84305	7
3642	55729	7

Table 34 continued

TF	TG	# dup
4286	10902	7
4286	23077	7
4286	23279	7
4286	23338	7
4286	2908	7
4286	3708	7
4286	3838	7
4286	4247	7
4286	54806	7
4286	55075	7
4286	6491	7
4286	6792	7
4286	729	7
4286	7323	7
4286	7414	7
4286	8930	7
4286	9133	7
4286	9918	7
4286	993	7
4780	51170	7
5316	1063	7
5316	23312	7
5316	3110	7
5316	332	7
5316	4281	7
5316	699	7
5316	701	7
5316	993	7
53615	23397	7
53615	23779	7
53615	3434	7
53615	3642	7
53615	51170	7
53615	701	7
6667	5293	7
672	5293	7
672	6581	7
6722	23397	7
6722	4233	7
6777	127	7
6777	3110	7
687	9212	7
7020	10036	7
7020	10116	7
7020	10347	7

Table 34 continued

TF	TG	# dup
7020	127	7
7020	1500	7
7020	8932	7
7050	1058	7
7050	4281	7
7050	7020	7
7050	9585	7
7528	23354	7
7528	23397	7
7528	29899	7
7528	3070	7
7528	3110	7
7528	5293	7
7528	6382	7
8202	10234	7
8202	3110	7
8202	5293	7
8930	11065	7
8930	3110	7
8930	580	7
8930	7153	7
8930	9319	7
8932	10595	7
10664	55247	6
10664	65055	6
11143	10068	6
11143	2487	6
11143	29899	6
11143	3642	6
11143	5293	6
11143	55723	6
11143	65055	6
11143	9212	6
1810	23354	6
1810	3214	6
1810	55247	6
1810	6491	6
1810	7424	6
1869	11104	6
1869	1509	6
1869	23279	6
1869	5573	6
1869	56852	6
1869	6456	6
1869	7050	6

Table 34 continued			Table 34 continued			Table 34 continued		
TF	TG	# dup	TF	TG	# dup	TF	TG	# dup
1869	9133	6	3214	6722	6	687	4281	6
2305	898	6	3214	6792	6	687	6456	6
2305	9212	6	3214	699	6	687	890	6
2908	10234	6	3214	7374	6	7020	1827	6
2908	1058	6	3642	9585	6	7020	23077	6
2908	11065	6	4286	1827	6	7020	23279	6
2908	4863	6	4286	4189	6	7020	3111	6
2908	6456	6	4286	4780	6	7020	3837	6
2908	701	6	4286	5293	6	7020	51715	6
2908	729	6	4286	567	6	7020	65055	6
3091	10234	6	4286	6093	6	7020	898	6
3091	3110	6	4286	6240	6	7020	9793	6
3091	55247	6	4286	9994	6	7050	10595	6
3091	898	6	4780	23354	6	7050	127	6
3110	10036	6	4780	4286	6	7050	29899	6
3110	10234	6	4780	6491	6	7050	3710	6
3110	10492	6	4780	9212	6	7050	7412	6
3110	11200	6	4780	9319	6	7050	7424	6
3110	23354	6	5316	11065	6	7050	9212	6
3110	5422	6	5316	127	6	7050	9401	6
3110	5573	6	5316	580	6	7050	993	6
3110	699	6	5316	6456	6	7528	10234	6
3110	9212	6	5316	7424	6	7528	10873	6
3214	10068	6	5316	79866	6	7528	4233	6
3214	10460	6	5316	9212	6	861	3070	6
3214	10921	6	53615	1063	6	861	55247	6
3214	2189	6	53615	127	6	861	6382	6
3214	22836	6	53615	4281	6	861	65055	6
3214	23338	6	53615	63967	6	861	701	6
3214	29896	6	58487	701	6	861	9212	6
3214	3337	6	58487	9319	6	8930	10234	6
3214	3832	6	6667	64403	6	8930	3642	6
3214	3837	6	672	55247	6	8930	4751	6
3214	3838	6	6722	10873	6	8930	51514	6
3214	3845	6	6722	127	6	8930	5347	6
3214	4171	6	6722	3110	6	8930	8564	6
3214	4173	6	6722	6382	6	8932	23118	6
3214	4175	6	6722	64403	6	8932	5864	6
3214	4247	6	6772	1058	6	8932	6491	6
3214	473	6	6772	23354	6	8932	701	6
3214	54407	6	6772	9212	6	8932	729	6
3214	5780	6	6777	10234	6	8932	9212	6
3214	58487	6	6777	65055	6	8932	993	6
3214	6421	6	6777	701	6	1058	3337	5
3214	65055	6	6777	9585	6	1058	5889	5

Table 34 continued

TF	TG	# dup
10664	10068	5
10664	10873	5
10664	127	5
10664	29899	5
10664	51170	5
10664	64403	5
10765	10068	5
10765	2177	5
10765	51170	5
10765	5293	5
10765	55723	5
11143	890	5
1810	10068	5
1810	2619	5
1810	29899	5
1810	3642	5
1810	3710	5
1810	5293	5
1810	6382	5
1810	7412	5
1810	8564	5
1869	1032	5
1869	10460	5
1869	2189	5
1869	55723	5
1869	5932	5
1869	687	5
1869	9510	5
2305	10234	5
2305	3070	5
2305	6382	5
2305	701	5
2908	1063	5
2908	1869	5
2908	2619	5
2908	3642	5
2908	4233	5
2908	55723	5
2908	580	5
2908	6432	5
2908	79866	5
2908	9212	5
2908	9319	5
3091	10873	5
3091	6456	5

Table 34 continued

TF	TG	# dup
3091	9510	5
3091	9666	5
3110	1063	5
3110	10873	5
3110	10921	5
3110	22894	5
3110	2534	5
3110	25896	5
3110	3643	5
3110	4175	5
3110	473	5
3110	4869	5
3110	54806	5
3110	6009	5
3110	65055	5
3110	687	5
3110	701	5
3110	7323	5
3110	7424	5
3110	79866	5
3110	8365	5
3110	8564	5
3110	993	5
3110	9967	5
3214	10036	5
3214	10051	5
3214	1031	5
3214	1032	5
3214	10458	5
3214	10600	5
3214	10714	5
3214	10769	5
3214	11104	5
3214	11135	5
3214	1647	5
3214	1827	5
3214	1977	5
3214	25896	5
3214	2730	5
3214	29117	5
3214	3091	5
3214	3146	5
3214	3156	5
3214	3312	5
3214	3642	5

Table 34 continued

TF	TG	# dup
3214	3643	5
3214	4664	5
3214	4780	5
3214	5111	5
3214	51141	5
3214	51763	5
3214	5316	5
3214	5422	5
3214	54806	5
3214	54962	5
3214	55075	5
3214	55729	5
3214	5663	5
3214	57153	5
3214	5932	5
3214	6009	5
3214	6093	5
3214	6240	5
3214	6241	5
3214	6464	5
3214	6491	5
3214	7050	5
3214	7251	5
3214	7298	5
3214	7884	5
3214	8364	5
3214	8365	5
3214	8366	5
3214	8473	5
3214	861	5
3214	8837	5
3214	891	5
3214	8932	5
3214	9793	5
3214	9939	5
3214	994	5
3214	996	5
3214	9967	5
3642	10036	5
3642	10492	5
3642	23279	5
3642	5573	5
4286	10873	5
4286	23354	5
4286	3111	5

Table 34 continued

TF	TG	# dup
4286	3267	5
4286	3845	5
4286	55723	5
4286	65055	5
4286	9212	5
4780	1063	5
4780	11065	5
4780	29899	5
4780	55247	5
4780	55632	5
4780	580	5
4780	6432	5
4780	64403	5
4780	65055	5
4780	7153	5
4780	79866	5
4780	9682	5
5316	10234	5
5316	10873	5
5316	23118	5
5316	51514	5
5316	55247	5
5316	55655	5
5316	7153	5
53615	10234	5
53615	10721	5
53615	23354	5
53615	5293	5
53615	9319	5
58487	1058	5
58487	23397	5
58487	3070	5
58487	4233	5
58487	55247	5
58487	56852	5
58487	729	5
6667	1058	5
6667	55247	5
6667	55723	5
6667	6491	5
6667	729	5
6667	9212	5
6667	9510	5
672	29899	5
672	3111	5

Table 34 continued

TF	TG	# dup
672	3642	5
672	56852	5
672	64403	5
6722	23354	5
6722	3111	5
6722	56852	5
6722	9212	5
6772	10234	5
6772	127	5
6772	3111	5
6772	55247	5
6772	898	5
6777	23354	5
6777	64403	5
6777	9319	5
687	1869	5
687	6432	5
687	898	5
7020	10766	5
7020	10905	5
7020	2189	5
7020	23397	5
7020	3838	5
7020	51170	5
7020	5154	5
7020	6240	5
7020	6792	5
7020	687	5
7020	7296	5
7050	1663	5
7050	23397	5
7050	3111	5
7050	3642	5
7050	3690	5
7050	5347	5
7050	55247	5
7050	580	5
7050	6382	5
7050	63967	5
7050	650	5
7050	699	5
7050	701	5
7050	7153	5
7528	1058	5
7528	3642	5

Table 34 continued

TF	TG	# dup
7528	4751	5
7528	6456	5
7528	9212	5
7528	9510	5
7528	9666	5
8202	10873	5
8202	127	5
8202	2177	5
8202	3111	5
8202	3214	5
8202	6382	5
8202	701	5
861	29899	5
861	3110	5
861	4751	5
861	64403	5
8930	1058	5
8930	1063	5
8930	1663	5
8930	1869	5
8930	23312	5
8930	23397	5
8930	3070	5
8930	3111	5
8930	332	5
8930	3710	5
8930	4281	5
8930	55247	5
8930	55632	5
8930	55655	5
8930	6432	5
8930	650	5
8930	9401	5
8930	9666	5
8932	10068	5
8932	1058	5
8932	23397	5
8932	332	5
8932	3434	5
8932	55247	5
8932	580	5
8932	5888	5
8932	63967	5
8932	7412	5
8932	9682	5

Table 34 continued

TF	TG	# dup
1058	127	4
1058	51170	4
10664	23354	4
10664	3642	4
10664	5293	4
10664	701	4
10664	898	4
10664	9666	4
10765	3642	4
10765	4233	4
10765	56852	4
10765	65055	4
10765	9510	4
11143	127	4
11143	2619	4
11143	332	4
11143	3710	4
11143	4233	4
11143	4281	4
11143	55247	4
11143	701	4
1810	3111	4
1810	332	4
1810	4233	4
1810	4281	4
1810	55723	4
1810	5888	4
1810	6581	4
1810	9401	4
1810	9585	4
1869	10902	4
1869	11073	4
1869	127	4
1869	1810	4
1869	3642	4
1869	4664	4
1869	4678	4
1869	4780	4
1869	51763	4
1869	53615	4
1869	54820	4
1869	64403	4
1869	7296	4
1869	7298	4
1869	7374	4

Table 34 continued

TF	TG	# dup
1869	9134	4
1869	9319	4
1869	9531	4
1869	994	4
2305	3214	4
2305	6491	4
2908	10595	4
2908	10873	4
2908	3070	4
2908	3111	4
2908	3690	4
2908	5864	4
2908	6491	4
3091	23397	4
3091	3111	4
3091	3642	4
3091	55723	4
3091	6382	4
3091	6581	4
3110	10068	4
3110	10116	4
3110	10595	4
3110	10902	4
3110	11004	4
3110	11073	4
3110	11130	4
3110	11135	4
3110	127	4
3110	1509	4
3110	2237	4
3110	2730	4
3110	3146	4
3110	3708	4
3110	3838	4
3110	4085	4
3110	5111	4
3110	51170	4
3110	5293	4
3110	54407	4
3110	55655	4
3110	6432	4
3110	64403	4
3110	6456	4
3110	6491	4
3110	6602	4

Table 34 continued

TF	TG	# dup
3110	7153	4
3110	7296	4
3110	7298	4
3110	8366	4
3110	8367	4
3110	84305	4
3110	890	4
3110	9319	4
3110	9510	4
3110	9646	4
3110	994	4
3214	10049	4
3214	10116	4
3214	1027	4
3214	1062	4
3214	10664	4
3214	10902	4
3214	10905	4
3214	11073	4
3214	11130	4
3214	11143	4
3214	1500	4
3214	1663	4
3214	1810	4
3214	1832	4
3214	2177	4
3214	2305	4
3214	23077	4
3214	23310	4
3214	23580	4
3214	23705	4
3214	2621	4
3214	27338	4
3214	2782	4
3214	2805	4
3214	2908	4
3214	29115	4
3214	3014	4
3214	3110	4
3214	3111	4
3214	3265	4
3214	3267	4
3214	329	4
3214	3708	4
3214	4286	4

Table 34 continued			Table 34 continued			Table 34 continued		
TF	TG	# dup	TF	TG	# dup	TF	TG	# dup
3214	4869	4	4780	1058	4	672	9510	4
3214	51715	4	4780	127	4	6722	3642	4
3214	5293	4	4780	23312	4	6722	4751	4
3214	53615	4	4780	23397	4	6772	3642	4
3214	55140	4	4780	3434	4	6772	4751	4
3214	55247	4	4780	51514	4	6772	6456	4
3214	55723	4	4780	5888	4	6772	729	4
3214	5608	4	4780	650	4	6777	1058	4
3214	57026	4	4780	699	4	6777	23397	4
3214	65057	4	4780	7020	4	6777	3642	4
3214	6598	4	4780	8564	4	6777	4751	4
3214	6772	4	5316	1869	4	6777	55632	4
3214	6777	4	5316	23397	4	6777	5864	4
3214	7323	4	5316	29899	4	6777	729	4
3214	7398	4	5316	6432	4	687	23354	4
3214	7533	4	5316	650	4	687	3111	4
3214	8202	4	5316	9666	4	687	699	4
3214	836	4	53615	10873	4	687	9682	4
3214	8367	4	53615	11200	4	687	993	4
3214	8930	4	53615	2619	4	7020	1027	4
3214	8943	4	53615	29899	4	7020	10460	4
3214	898	4	53615	5347	4	7020	1810	4
3214	9133	4	53615	55655	4	7020	29899	4
3214	9134	4	53615	580	4	7020	3070	4
3214	9156	4	53615	6382	4	7020	4189	4
3214	9184	4	58487	10234	4	7020	4664	4
3214	9646	4	58487	3110	4	7020	51763	4
3214	9918	4	58487	55723	4	7020	54806	4
3214	998	4	58487	6456	4	7020	54820	4
3214	9994	4	58487	898	4	7020	55632	4
3642	10347	4	6667	10234	4	7020	5608	4
3642	10873	4	6667	127	4	7020	567	4
3642	127	4	6667	3110	4	7020	6241	4
3642	4751	4	6667	3642	4	7020	6491	4
3642	55140	4	6667	65055	4	7020	6667	4
4286	127	4	6667	9319	4	7020	729	4
4286	2177	4	6667	993	4	7020	836	4
4286	23397	4	672	3110	4	7020	9133	4
4286	3642	4	672	3214	4	7050	10068	4
4286	4751	4	672	55723	4	7050	1063	4
4286	51170	4	672	6382	4	7050	10873	4
4286	6456	4	672	6456	4	7050	11065	4
4286	8881	4	672	65055	4	7050	3070	4
4780	10068	4	672	898	4	7050	3434	4
4780	10234	4	672	9212	4	7050	51514	4

Table 34 continued

TF	TG	# dup
7050	6432	4
7050	64403	4
7050	8564	4
7528	3111	4
7528	55247	4
7528	55723	4
8202	10068	4
8202	65055	4
8202	729	4
8202	9319	4
8202	993	4
861	10068	4
861	23354	4
861	6456	4
861	6491	4
861	898	4
861	9666	4
8930	2487	4
8930	3690	4
8930	4286	4
8930	5888	4
8930	6382	4
8930	6456	4
8930	65055	4
8930	7020	4
8930	729	4
8930	79866	4
8930	9682	4
8932	11200	4
8932	2177	4
8932	3642	4
8932	3710	4
8932	4233	4
8932	51514	4
8932	5347	4
8932	55723	4
8932	7153	4
8932	7424	4
8932	79866	4
8932	8564	4
8932	890	4
8932	9585	4
1058	10873	3
1058	10921	3
1058	11073	3

Table 34 continued

TF	TG	# dup
1058	7298	3
1058	9133	3
10664	3110	3
10664	3111	3
10664	4233	3
10664	4751	3
10664	6456	3
10664	6491	3
10765	3110	3
10765	3111	3
10765	55247	3
10765	64403	3
10765	6491	3
10765	729	3
10765	993	3
11143	11200	3
11143	3434	3
11143	6581	3
11143	729	3
11143	9510	3
11143	993	3
1810	10721	3
1810	1869	3
1810	55655	3
1810	65055	3
1810	701	3
1810	7020	3
1810	9212	3
1810	9510	3
1869	10068	3
1869	1027	3
1869	10347	3
1869	1827	3
1869	23077	3
1869	2534	3
1869	2621	3
1869	3111	3
1869	3156	3
1869	3265	3
1869	3267	3
1869	3708	3
1869	3837	3
1869	3845	3
1869	4189	3
1869	4193	3

Table 34 continued

TF	TG	# dup
1869	4247	3
1869	473	3
1869	5154	3
1869	51715	3
1869	5293	3
1869	54407	3
1869	54806	3
1869	55075	3
1869	5530	3
1869	5608	3
1869	5663	3
1869	5780	3
1869	6240	3
1869	65055	3
1869	6792	3
1869	6944	3
1869	8881	3
1869	8930	3
1869	8932	3
1869	898	3
1869	9212	3
1869	9994	3
2305	3642	3
2305	55247	3
2305	55723	3
2305	64403	3
2305	6456	3
2305	9319	3
2908	10721	3
2908	127	3
2908	1663	3
2908	23312	3
2908	29899	3
2908	4286	3
2908	55632	3
2908	56852	3
2908	64403	3
2908	9510	3
2908	9666	3
3091	1058	3
3091	3214	3
3091	4751	3
3091	55632	3
3091	64403	3
3091	9319	3

Table 34 continued

TF	TG	# dup
3110	10051	3
3110	1022	3
3110	1027	3
3110	1031	3
3110	1032	3
3110	10600	3
3110	1062	3
3110	10714	3
3110	11143	3
3110	1647	3
3110	1810	3
3110	2177	3
3110	2189	3
3110	22836	3
3110	23279	3
3110	23338	3
3110	27338	3
3110	29896	3
3110	3111	3
3110	3214	3
3110	3832	3
3110	3837	3
3110	4171	3
3110	4189	3
3110	4233	3
3110	4247	3
3110	4780	3
3110	51141	3
3110	51763	3
3110	5316	3
3110	53615	3
3110	5515	3
3110	55632	3
3110	5603	3
3110	5780	3
3110	5889	3
3110	6093	3
3110	6118	3
3110	6240	3
3110	6241	3
3110	65057	3
3110	6760	3
3110	6772	3
3110	6944	3
3110	7374	3

Table 34 continued

TF	TG	# dup
3110	7528	3
3110	7884	3
3110	8202	3
3110	836	3
3110	8364	3
3110	8837	3
3110	8850	3
3110	8932	3
3110	899	3
3110	9133	3
3110	9156	3
3110	9531	3
3110	9585	3
3110	9682	3
3110	9793	3
3110	9918	3
3110	996	3
3110	9994	3
3214	1022	3
3214	10234	3
3214	10595	3
3214	10766	3
3214	2237	3
3214	2280	3
3214	23397	3
3214	2534	3
3214	3434	3
3214	3690	3
3214	4000	3
3214	4189	3
3214	4436	3
3214	4678	3
3214	5154	3
3214	54820	3
3214	5530	3
3214	5603	3
3214	567	3
3214	580	3
3214	6382	3
3214	6432	3
3214	6602	3
3214	672	3
3214	687	3
3214	729	3
3214	7414	3

Table 34 continued

TF	TG	# dup
3214	83695	3
3214	84168	3
3214	8772	3
3214	8841	3
3214	8850	3
3214	8881	3
3214	899	3
3214	9212	3
3214	9531	3
3214	9585	3
3642	10068	3
3642	10714	3
3642	2177	3
3642	4171	3
3642	51170	3
3642	56852	3
3642	5889	3
3642	6432	3
3642	64403	3
3642	6491	3
3642	699	3
3642	994	3
4286	10068	3
4286	3110	3
4286	64403	3
4286	9319	3
4286	9510	3
4780	23779	3
4780	2619	3
4780	3070	3
4780	3111	3
4780	4233	3
4780	5347	3
4780	56852	3
4780	5864	3
4780	63967	3
4780	6581	3
4780	701	3
4780	890	3
4780	898	3
4780	9585	3
5316	1058	3
5316	10595	3
5316	1663	3
5316	2177	3

Table 34 continued

TF	TG	# dup
5316	3214	3
5316	3642	3
5316	3710	3
5316	4233	3
5316	5888	3
5316	6382	3
5316	65055	3
5316	6581	3
5316	890	3
5316	898	3
5316	9510	3
5316	9585	3
53615	10068	3
53615	1058	3
53615	11065	3
53615	1663	3
53615	1869	3
53615	2177	3
53615	23118	3
53615	3111	3
53615	3690	3
53615	3710	3
53615	4286	3
53615	55247	3
53615	55723	3
53615	5888	3
53615	64403	3
53615	650	3
53615	65055	3
53615	6581	3
53615	699	3
53615	729	3
53615	8564	3
53615	890	3
53615	898	3
53615	9401	3
53615	9510	3
53615	9666	3
58487	29899	3
58487	3111	3
58487	51170	3
58487	55632	3
58487	6491	3
58487	65055	3
6667	10873	3

Table 34 continued

TF	TG	# dup
6667	2177	3
6667	3070	3
6667	3111	3
6667	56852	3
6667	6581	3
672	4233	3
672	6491	3
6722	29899	3
6722	55247	3
6722	55723	3
6722	6456	3
6772	10873	3
6772	23397	3
6772	29899	3
6772	3110	3
6772	6491	3
6772	9319	3
6777	10873	3
6777	3214	3
6777	6382	3
6777	9401	3
6777	9666	3
687	10068	3
687	10721	3
687	11065	3
687	11200	3
687	127	3
687	1663	3
687	332	3
687	3642	3
687	3690	3
687	51514	3
687	5347	3
687	55655	3
687	580	3
687	6382	3
687	63967	3
687	6491	3
687	650	3
687	65055	3
687	7020	3
687	729	3
687	7412	3
687	79866	3
687	8564	3

Table 34 continued

TF	TG	# dup
687	9510	3
687	9666	3
7020	10873	3
7020	23310	3
7020	25836	3
7020	3156	3
7020	3267	3
7020	4780	3
7020	5316	3
7020	55075	3
7020	55247	3
7020	55723	3
7020	5573	3
7020	5780	3
7020	6093	3
7020	64403	3
7020	6456	3
7020	7414	3
7020	8930	3
7020	9319	3
7020	9918	3
7020	994	3
7020	996	3
7050	10234	3
7050	22909	3
7050	23312	3
7050	3214	3
7050	332	3
7050	4751	3
7050	5888	3
7050	65055	3
7050	79866	3
7050	9319	3
7050	9682	3
7528	127	3
7528	64403	3
7528	65055	3
7528	6581	3
7528	898	3
8202	56852	3
8202	64403	3
8202	6456	3
8202	6491	3
861	1058	3
861	10873	3

Table 34 continued

TF	TG	# dup
861	127	3
861	2177	3
861	23397	3
861	55632	3
861	9319	3
861	9510	3
8930	10068	3
8930	2177	3
8930	23118	3
8930	23354	3
8930	23779	3
8930	4863	3
8930	55723	3
8930	6491	3
8930	701	3
8930	890	3
8930	993	3
8932	10721	3
8932	127	3
8932	2619	3
8932	3070	3
8932	3110	3
8932	3690	3
8932	6432	3
8932	64403	3
8932	7020	3
8932	9319	3
1058	10068	2
1058	2237	2
1058	23279	2
1058	3642	2
1058	4247	2
1058	55140	2
1058	55790	2
1058	5780	2
1058	6421	2
1058	6456	2
1058	729	2
1058	7374	2
1058	84305	2
1058	8564	2
1058	9666	2
1058	9939	2
1058	994	2
10664	2177	2

Table 34 continued

TF	TG	# dup
10664	56852	2
10664	729	2
10664	9319	2
10664	9510	2
10765	3214	2
10765	898	2
11143	10873	2
11143	2177	2
11143	3111	2
11143	3214	2
11143	51170	2
11143	55655	2
11143	56852	2
11143	64403	2
11143	6456	2
11143	6491	2
11143	898	2
11143	9319	2
1810	1063	2
1810	11200	2
1810	127	2
1810	22909	2
1810	23312	2
1810	23779	2
1810	5864	2
1810	6432	2
1810	64403	2
1810	6456	2
1810	79866	2
1810	993	2
1869	10234	2
1869	10766	2
1869	10873	2
1869	10905	2
1869	1832	2
1869	23338	2
1869	2908	2
1869	3214	2
1869	3838	2
1869	51170	2
1869	5316	2
1869	567	2
1869	6093	2
1869	6241	2
1869	6382	2

Table 34 continued

TF	TG	# dup
1869	7251	2
1869	7323	2
1869	7414	2
1869	9646	2
2305	29899	2
2305	55632	2
2305	56852	2
2305	6581	2
2305	729	2
2305	9666	2
2908	23354	2
2908	3110	2
2908	4751	2
2908	55655	2
2908	6382	2
3091	23354	2
3091	29899	2
3091	701	2
3110	10460	2
3110	10664	2
3110	10765	2
3110	10766	2
3110	10769	2
3110	11065	2
3110	1832	2
3110	1977	2
3110	23312	2
3110	23580	2
3110	23705	2
3110	23779	2
3110	2621	2
3110	2782	2
3110	2805	2
3110	2908	2
3110	29117	2
3110	3014	2
3110	3148	2
3110	3156	2
3110	3265	2
3110	3267	2
3110	3312	2
3110	332	2
3110	3434	2
3110	3845	2
3110	4436	2

Table 34 continued			Table 34 continued			Table 34 continued		
TF	TG	# dup	TF	TG	# dup	TF	TG	# dup
3110	51514	2	3642	11004	2	4780	2487	2
3110	5347	2	3642	11065	2	4780	3214	2
3110	54962	2	3642	1509	2	4780	4281	2
3110	55723	2	3642	2237	2	4780	5293	2
3110	55790	2	3642	3070	2	4780	55655	2
3110	5608	2	3642	3214	2	4780	55723	2
3110	5888	2	3642	3337	2	4780	6382	2
3110	650	2	3642	3690	2	4780	6456	2
3110	6792	2	3642	4247	2	4780	729	2
3110	7533	2	3642	5111	2	4780	993	2
3110	8473	2	3642	5293	2	5316	10721	2
3110	8930	2	3642	54806	2	5316	22909	2
3110	8943	2	3642	5515	2	5316	23354	2
3110	898	2	3642	55247	2	5316	23779	2
3110	9184	2	3642	55790	2	5316	2487	2
3110	9666	2	3642	580	2	5316	3690	2
3214	1063	2	3642	6382	2	5316	5864	2
3214	10873	2	3642	63967	2	5316	64403	2
3214	127	2	3642	6760	2	5316	6491	2
3214	22909	2	3642	687	2	5316	7020	2
3214	23312	2	3642	7153	2	5316	9401	2
3214	2487	2	3642	729	2	5316	9682	2
3214	3070	2	3642	7296	2	53615	10595	2
3214	3148	2	3642	7374	2	53615	22909	2
3214	4085	2	3642	7424	2	53615	23312	2
3214	4193	2	3642	7884	2	53615	2487	2
3214	4281	2	3642	8366	2	53615	3110	2
3214	4751	2	3642	84305	2	53615	3214	2
3214	4863	2	3642	9134	2	53615	4751	2
3214	51170	2	3642	9212	2	53615	56852	2
3214	51343	2	3642	9319	2	53615	6432	2
3214	55655	2	3642	9666	2	53615	6491	2
3214	63967	2	3642	9682	2	53615	7153	2
3214	6456	2	3642	9793	2	53615	7412	2
3214	701	2	3642	9967	2	53615	9585	2
3214	7412	2	4286	3070	2	58487	127	2
3214	7424	2	4286	4233	2	58487	3214	2
3214	79866	2	4286	55247	2	58487	3642	2
3214	890	2	4286	56852	2	58487	4751	2
3214	9319	2	4286	701	2	58487	5293	2
3214	9510	2	4780	10721	2	6667	23354	2
3642	10234	2	4780	11200	2	6667	4751	2
3642	1063	2	4780	2177	2	6667	6382	2
3642	10721	2	4780	22909	2	6667	9666	2
3642	10921	2	4780	23118	2	6722	3070	2

Table 34 continued

TF	TG	# dup
6722	701	2
6722	898	2
6722	9319	2
6722	9666	2
6772	5293	2
6772	64403	2
6772	65055	2
6772	9510	2
6777	10721	2
6777	1663	2
6777	55247	2
6777	55723	2
6777	6581	2
6777	9212	2
687	10234	2
687	10595	2
687	23118	2
687	23312	2
687	29899	2
687	3434	2
687	3710	2
687	4233	2
687	4863	2
687	55723	2
687	5864	2
687	64403	2
687	701	2
687	7153	2
687	9401	2
687	9585	2
7020	10051	2
7020	10902	2
7020	23338	2
7020	23354	2
7020	3110	2
7020	3642	2
7020	3832	2
7020	3845	2
7020	4233	2
7020	4678	2
7020	473	2
7020	5293	2
7020	53615	2
7020	54407	2
7020	5530	2

Table 34 continued

TF	TG	# dup
7020	5663	2
7020	6382	2
7020	6581	2
7020	6944	2
7020	701	2
7020	7050	2
7020	7323	2
7020	7374	2
7020	9134	2
7020	9212	2
7020	9510	2
7020	9646	2
7020	998	2
7050	10721	2
7050	5293	2
7050	55632	2
7050	55723	2
7050	5864	2
7050	6456	2
7050	898	2
7050	9510	2
7528	729	2
8202	23354	2
8202	4233	2
8202	898	2
861	3111	2
861	55723	2
8930	10721	2
8930	10873	2
8930	2619	2
8930	29899	2
8930	3214	2
8930	4233	2
8930	5293	2
8930	5864	2
8930	63967	2
8930	64403	2
8930	6581	2
8930	7424	2
8930	898	2
8930	9510	2
8930	9585	2
8932	1063	2
8932	11065	2
8932	1869	2

Table 34 continued

TF	TG	# dup
8932	23312	2
8932	23354	2
8932	23779	2
8932	29899	2
8932	3214	2
8932	4281	2
8932	4863	2
8932	55632	2
8932	6581	2
8932	9510	2
8932	9666	2
1058	1022	1
1058	10492	1
1058	1062	1
1058	10714	1
1058	1509	1
1058	2189	1
1058	22909	1
1058	23312	1
1058	2487	1
1058	2805	1
1058	29896	1
1058	3111	1
1058	3710	1
1058	4171	1
1058	4175	1
1058	473	1
1058	5111	1
1058	5293	1
1058	54820	1
1058	55247	1
1058	55632	1
1058	55723	1
1058	5573	1
1058	56852	1
1058	6009	1
1058	64403	1
1058	6581	1
1058	6760	1
1058	7412	1
1058	7424	1
1058	8365	1
1058	8366	1
1058	8367	1
1058	8930	1

Table 34 continued

TF	TG	# dup
1058	9134	1
1058	9510	1
1058	9585	1
10664	3214	1
10664	6581	1
10664	993	1
10765	10873	1
10765	23354	1
10765	6382	1
10765	6581	1
10765	9319	1
11143	23354	1
11143	7412	1
11143	7424	1
11143	79866	1
1810	10873	1
1810	2177	1
1810	3110	1
1810	3434	1
1810	4286	1
1810	4863	1
1810	729	1
1810	898	1
1869	29899	1
1869	3070	1
1869	3110	1
1869	4233	1
1869	55247	1
1869	55632	1
1869	6581	1
1869	9666	1
2305	127	1
2908	3214	1
2908	9682	1
3091	56852	1
3110	10049	1
3110	10347	1
3110	10458	1
3110	10721	1
3110	10905	1
3110	1500	1
3110	1663	1
3110	2280	1
3110	22909	1
3110	23118	1

Table 34 continued

TF	TG	# dup
3110	23310	1
3110	23397	1
3110	2487	1
3110	29115	1
3110	29899	1
3110	329	1
3110	3690	1
3110	3710	1
3110	4173	1
3110	4193	1
3110	4281	1
3110	4664	1
3110	4678	1
3110	4751	1
3110	4863	1
3110	51343	1
3110	55075	1
3110	55140	1
3110	55247	1
3110	5663	1
3110	567	1
3110	56852	1
3110	57026	1
3110	57153	1
3110	580	1
3110	58487	1
3110	63967	1
3110	6581	1
3110	6667	1
3110	6777	1
3110	7020	1
3110	7398	1
3110	7414	1
3110	83695	1
3110	84168	1
3110	8772	1
3110	8841	1
3110	891	1
3110	9134	1
3110	9401	1
3110	998	1
3214	1058	1
3214	11065	1
3214	1869	1
3214	23118	1

Table 34 continued

TF	TG	# dup
3214	29899	1
3214	332	1
3214	3710	1
3214	51514	1
3214	5347	1
3214	55632	1
3214	55790	1
3214	56852	1
3214	5864	1
3214	64403	1
3214	7153	1
3214	8564	1
3214	9401	1
3214	9666	1
3214	993	1
3642	1058	1
3642	11130	1
3642	1977	1
3642	22909	1
3642	23118	1
3642	23338	1
3642	2487	1
3642	25896	1
3642	2730	1
3642	29115	1
3642	29899	1
3642	3434	1
3642	3832	1
3642	3838	1
3642	4000	1
3642	4175	1
3642	4281	1
3642	51141	1
3642	5347	1
3642	55632	1
3642	55655	1
3642	55723	1
3642	5932	1
3642	6241	1
3642	6421	1
3642	6456	1
3642	65057	1
3642	6667	1
3642	701	1
3642	7020	1

Table 34 continued

TF	TG	# dup
3642	7298	1
3642	7398	1
3642	79866	1
3642	8364	1
3642	8365	1
3642	8367	1
3642	83695	1
3642	8850	1
3642	890	1
3642	891	1
3642	898	1
3642	9133	1
3642	9531	1
3642	993	1
3642	9939	1
4286	29899	1
4286	3214	1
4286	6581	1
4286	9666	1
4780	10873	1
4780	1869	1
4780	3110	1
4780	332	1
4780	3642	1
4780	3690	1
4780	3710	1
4780	4863	1
4780	7412	1
4780	7424	1
4780	9401	1
4780	9510	1
4780	9666	1
5316	11200	1
5316	4751	1
5316	55632	1
5316	56852	1
5316	63967	1
53615	332	1
53615	4863	1
53615	5864	1
53615	6456	1
53615	7020	1
53615	79866	1
53615	9212	1
53615	993	1

Table 34 continued

TF	TG	# dup
58487	10873	1
58487	6382	1
58487	64403	1
58487	9510	1
672	729	1
6772	3214	1
6772	4233	1
6772	55632	1
6772	56852	1
6772	6382	1
6772	6581	1
6772	701	1
6772	9666	1
6777	29899	1
6777	56852	1
6777	6456	1
687	1058	1
687	10873	1
687	2177	1
687	23397	1
687	23779	1
687	3070	1
687	3214	1
687	5293	1
687	55790	1
687	56852	1
687	5888	1
687	6581	1
687	7424	1
687	9319	1
7020	10234	1
7020	56852	1
7020	8881	1
7020	9531	1
7050	2177	1
7050	2487	1
7050	56852	1
7050	6581	1
7050	729	1
7050	890	1
7050	9666	1
8202	6581	1
8202	9510	1
861	56852	1
861	6581	1

Table 34 continued

TF	TG	# dup
861	993	1
8930	3434	1
8930	55790	1
8930	56852	1
8930	699	1
8930	7412	1
8932	1663	1
8932	3111	1
8932	4751	1
8932	55655	1
8932	56852	1
8932	6456	1
8932	9401	1

The table gives for each edge the number of time it is repeated after concatenating all the 132 cell line networks collected from the **HumanBase** database <https://hb.flatironinstitute.org/download>. The 1^{st} column represents the TF entrez ID. The 2^{nd} column the TG entrez ID. The 3^{rd} column represent the of times the link is repeated in the final network.

Table 35: Edges repetition in Garcia networks

TF	TG	# dup
CTCF	ABCA7	2
CTCF	ABCC2	2
CTCF	ABHD10	2
CTCF	ADCY6	2
CTCF	ADH4	2
CTCF	AHI1	2
CTCF	AMD1	2
CTCF	ANLN	2
CTCF	ANP32B	2
CTCF	ANP32E	2
CTCF	AOC2	2
CTCF	AOC3	2
CTCF	AP3D1	2
CTCF	AP3M2	2
CTCF	AP4B1	2
CTCF	ARHGAP11A	2
CTCF	ARHGAP19	2
CTCF	ARHGAP8	2
CTCF	ARHGEF39	2
CTCF	ARL4A	2
CTCF	ARL6IP1	2
CTCF	ASF1B	2
CTCF	ASIP	2
CTCF	ASPHD2	2
CTCF	ATAD2	2
CTCF	ATF7IP	2
CTCF	ATL2	2
CTCF	AURKB	2
CTCF	B2M	2
CTCF	BAG3	2
CTCF	BAIAP2	2
CTCF	BBS2	2
CTCF	BCLAF1	2
CTCF	BIRC2	2
CTCF	BIVM	2
CTCF	BMP2	2
CTCF	BRCA1	2
CTCF	BRD7	2
CTCF	BTBD3	2
CTCF	BUB3	2

Table 35 continued

TF	TG	# dup
CTCF	C6	2
CTCF	CADM1	2
CTCF	CAPN7	2
CTCF	CASP3	2
CTCF	CCDC90B	2
CTCF	CCNE1	2
CTCF	CCNF	2
CTCF	CDC16	2
CTCF	CDC20	2
CTCF	CDC25A	2
CTCF	CDC25B	2
CTCF	CDC25C	2
CTCF	CDC42	2
CTCF	CDC42EP1	2
CTCF	CDC42EP4	2
CTCF	CDC45	2
CTCF	CDC6	2
CTCF	CDC7	2
CTCF	CDCA7	2
CTCF	CDCA7L	2
CTCF	CDK20	2
CTCF	CDK7	2
CTCF	CDKN1B	2
CTCF	CDKN2AIP	2
CTCF	CDKN2C	2
CTCF	CDKN3	2
CTCF	CENPA	2
CTCF	CENPE	2
CTCF	CENPF	2
CTCF	CENPM	2
CTCF	CEP44	2
CTCF	CEP55	2
CTCF	CEP70	2
CTCF	CFD	2
CTCF	CFLAR	2
CTCF	CHAF1B	2
CTCF	CHEK2	2
CTCF	CIC	2
CTCF	CIT	2
CTCF	CKAP5	2
CTCF	CKS2	2
CTCF	CLSPN	2
CTCF	CNN2	2
CTCF	CNOT10	2
CTCF	COQ6	2

Table 35 continued

TF	TG	# dup
CTCF	CREBZF	2
CTCF	CRK	2
CTCF	CRYBA1	2
CTCF	CSH2	2
CTCF	CTCF	2
CTCF	CTNND1	2
CTCF	CTR9	2
CTCF	CTSD	2
CTCF	CWC15	2
CTCF	CXCL14	2
CTCF	CYB5R2	2
CTCF	CYTH3	2
CTCF	DCAF16	2
CTCF	DCAF7	2
CTCF	DCTN6	2
CTCF	DDX11	2
CTCF	DEPDC1B	2
CTCF	DET1	2
CTCF	DHX8	2
CTCF	DLGAP5	2
CTCF	DMTF1	2
CTCF	DMXL2	2
CTCF	DNAJB1	2
CTCF	DNAJB4	2
CTCF	DNAJB6	2
CTCF	DNAJB9	2
CTCF	DNAJC3	2
CTCF	DNAJC6	2
CTCF	DTL	2
CTCF	DUSP4	2
CTCF	DYNLL1	2
CTCF	DZIP3	2
CTCF	E2F1	2
CTCF	E2F5	2
CTCF	E2F8	2
CTCF	EBI3	2
CTCF	EIF4E	2
CTCF	ELP3	2
CTCF	ENOSF1	2
CTCF	ERN2	2
CTCF	ESPL1	2
CTCF	EXO1	2
CTCF	FABP1	2
CTCF	FADD	2
CTCF	FAM105A	2

Table 35 continued

TF	TG	# dup
CTCF	FAM110A	2
CTCF	FAM189B	2
CTCF	FAM214A	2
CTCF	FAM60A	2
CTCF	FANCA	2
CTCF	FANCI	2
CTCF	FBXL20	2
CTCF	FEM1B	2
CTCF	FEN1	2
CTCF	FKBP1A	2
CTCF	FLAD1	2
CTCF	FXR1	2
CTCF	G2E3	2
CTCF	G3BP1	2
CTCF	GAS1	2
CTCF	GAS6	2
CTCF	GDF15	2
CTCF	GINS2	2
CTCF	GINS3	2
CTCF	GMNN	2
CTCF	GNB1	2
CTCF	GOLGA8A	2
CTCF	GOT1	2
CTCF	GPSM2	2
CTCF	GRK6	2
CTCF	GRPEL1	2
CTCF	GTF2B	2
CTCF	GTSE1	2
CTCF	H2AFX	2
CTCF	HAUS5	2
CTCF	HAUS8	2
CTCF	HCP5	2
CTCF	HELLS	2
CTCF	HERPUD2	2
CTCF	HIF1A	2
CTCF	HIST1H4C	2
CTCF	HIST1H4E	2
CTCF	HIST1H4H	2
CTCF	HJURP	2
CTCF	HLA-DOA	2
CTCF	HLA-DRA	2
CTCF	HMG20B	2
CTCF	HMGCR	2
CTCF	HMMR	2
CTCF	HRAS	2

Table 35 continued

TF	TG	# dup
CTCF	HSD17B11	2
CTCF	HSF2	2
CTCF	HSPA13	2
CTCF	HSPB8	2
CTCF	IDO1	2
CTCF	ILF2	2
CTCF	INADL	2
CTCF	INPP5K	2
CTCF	INSIG2	2
CTCF	INSM1	2
CTCF	INSR	2
CTCF	INTS7	2
CTCF	ITPR3	2
CTCF	IVNS1ABP	2
CTCF	KANK2	2
CTCF	KAT2B	2
CTCF	KCTD2	2
CTCF	KDM4A	2
CTCF	KDM5B	2
CTCF	KIAA0586	2
CTCF	KIAA1147	2
CTCF	KIAA1524	2
CTCF	KIF11	2
CTCF	KIF14	2
CTCF	KIF20B	2
CTCF	KIF22	2
CTCF	KIF5B	2
CTCF	KIFC1	2
CTCF	KLF6	2
CTCF	KLF9	2
CTCF	KMO	2
CTCF	KPNA2	2
CTCF	KPNB1	2
CTCF	KRAS	2
CTCF	LARP7	2
CTCF	LMNB1	2
CTCF	LMO4	2
CTCF	LPP	2
CTCF	LRIF1	2
CTCF	LYAR	2
CTCF	MAD2L1	2
CTCF	MAN1A2	2
CTCF	MAP2K6	2
CTCF	MAP3K2	2
CTCF	MAPK13	2

Table 35 continued

TF	TG	# dup
CTCF	MATN2	2
CTCF	MBD2	2
CTCF	MBD3	2
CTCF	MCAM	2
CTCF	MCM5	2
CTCF	MCM8	2
CTCF	MDC1	2
CTCF	MDM2	2
CTCF	ME3	2
CTCF	MED31	2
CTCF	MEGF9	2
CTCF	MELK	2
CTCF	MET	2
CTCF	MGAT2	2
CTCF	MID1	2
CTCF	MIS18BP1	2
CTCF	MITF	2
CTCF	MKI67	2
CTCF	MLLT4	2
CTCF	MND1	2
CTCF	MNT	2
CTCF	MXN1	2
CTCF	MORF4L2	2
CTCF	MRPL19	2
CTCF	MRPS2	2
CTCF	MSH2	2
CTCF	MTCL1	2
CTCF	MYCBP2	2
CTCF	MZF1	2
CTCF	NAB1	2
CTCF	NCAPD2	2
CTCF	NCAPD3	2
CTCF	NCAPH	2
CTCF	NCOA3	2
CTCF	NCOA5	2
CTCF	NCS1	2
CTCF	NDE1	2
CTCF	NEIL3	2
CTCF	NEK2	2
CTCF	NFIC	2
CTCF	NFYA	2
CTCF	NFYB	2
CTCF	NIPBL	2
CTCF	NKTR	2
CTCF	NMB	2

Table 35 continued

TF	TG	# dup
CTCF	NNMT	2
CTCF	NPAT	2
CTCF	NPM1	2
CTCF	NR3C1	2
CTCF	NSUN3	2
CTCF	NUCKS1	2
CTCF	NUDT4	2
CTCF	NUF2	2
CTCF	NUP160	2
CTCF	NUP37	2
CTCF	ODF2	2
CTCF	OGT	2
CTCF	OLR1	2
CTCF	ORC3	2
CTCF	OSER1	2
CTCF	PANK2	2
CTCF	PCNA	2
CTCF	PDGFA	2
CTCF	PDXP	2
CTCF	PIK3CD	2
CTCF	PKMYT1	2
CTCF	PLIN3	2
CTCF	PLK1	2
CTCF	PLK2	2
CTCF	POC1A	2
CTCF	POLA1	2
CTCF	POLD3	2
CTCF	POLQ	2
CTCF	POM121	2
CTCF	PPP1R2	2
CTCF	PPP3CA	1
CTCF	PPP6R3	1
CTCF	PRIM1	1
CTCF	PRIM2	1
CTCF	PRKAR1A	1
CTCF	PRPSAP1	1
CTCF	PRR11	1
CTCF	PRR16	1
CTCF	PSEN1	1
CTCF	PSMD11	1
CTCF	PSMG3	1
CTCF	PTMS	1
CTCF	PTP4A1	1
CTCF	PTPN9	1
CTCF	PTTG1	1

Table 35 continued

TF	TG	# dup
CTCF	PWP1	1
CTCF	QRICH1	1
CTCF	RAB23	1
CTCF	RAB3A	1
CTCF	RAD18	1
CTCF	RAD21	1
CTCF	RAD51	1
CTCF	RAD51C	1
CTCF	RAD54L	1
CTCF	RAN	1
CTCF	RANGAP1	1
CTCF	RBBP8	1
CTCF	RBM8A	1
CTCF	RCAN1	1
CTCF	REEP1	1
CTCF	RFC4	1
CTCF	RGS3	1
CTCF	RHEB	1
CTCF	RHOBTB3	1
CTCF	RNF126	1
CTCF	ROCK1	1
CTCF	RPL13A	1
CTCF	RRM1	1
CTCF	RRM2	1
CTCF	RRP1	1
CTCF	SAP30	1
CTCF	SAP30BP	1
CTCF	SDC1	1
CTCF	SEC62	1
CTCF	SEPHS1	1
CTCF	SEPN1	1
CTCF	SGK1	1
CTCF	SH3GL2	1
CTCF	SHCBP1	1
CTCF	SLBP	1
CTCF	SLC17A2	1
CTCF	SLC22A3	1
CTCF	SLC25A27	1
CTCF	SLC25A36	1
CTCF	SLC38A2	1
CTCF	SLC39A10	1
CTCF	SLC44A2	1
CTCF	SLC4A1AP	1
CTCF	SMARCB1	1
CTCF	SMARCD1	1

Table 35 continued

TF	TG	# dup
CTCF	SMC4	1
CTCF	SMTN	1
CTCF	SNUPN	1
CTCF	SP1	1
CTCF	SPDL1	1
CTCF	SRF	1
CTCF	SS18	1
CTCF	SSR3	1
CTCF	STAG3	1
CTCF	STAT1	1
CTCF	STAT5B	1
CTCF	STIL	1
CTCF	SUCLG2	1
CTCF	TAB2	1
CTCF	TFAP2A	1
CTCF	TGIF1	1
CTCF	THRAP3	1
CTCF	TMPO	1
CTCF	TNPO2	1
CTCF	TOMM34	1
CTCF	TOP1	1
CTCF	TOP2A	1
CTCF	TPX2	1
CTCF	TRA2A	1
CTCF	TRAIP	1
CTCF	TRIM45	1
CTCF	TRIP13	1
CTCF	TROAP	1
CTCF	TSC22D1	1
CTCF	TSKU	1
CTCF	TSN	1
CTCF	TTC31	1
CTCF	TTF2	1
CTCF	TTK	1
CTCF	TUBB2A	1
CTCF	TUBB4B	1
CTCF	TUBD1	1
CTCF	TULP4	1
CTCF	TXNRD1	1
CTCF	TYMS	1
CTCF	UACA	1
CTCF	UBE2D3	1
CTCF	UBE2S	1
CTCF	UBL3	1
CTCF	UBR7	1

Table 35 continued

TF	TG	# dup
CTCF	UHRF1	1
CTCF	UNG	1
CTCF	USP1	1
CTCF	USP13	1
CTCF	USP53	1
CTCF	USP6NL	1
CTCF	VCAM1	1
CTCF	VCL	1
CTCF	VEGFC	1
CTCF	VPS37C	1
CTCF	VPS72	1
CTCF	VTA1	1
CTCF	WSB1	1
CTCF	YWHAH	1
CTCF	YY1	1
CTCF	ZBED5	1
CTCF	ZBTB7A	1
CTCF	ZC3HC1	1
CTCF	ZMYM1	1
CTCF	ZNF143	1
CTCF	ZNF217	1
CTCF	ZNF281	1
CTCF	ZNF414	1
CTCF	ZNF521	1
CTCF	ZNF593	1
CTCF	ZNFX1	1
CTCF	ZNHIT2	1
CTCF	ZPBP	1
CTCF	ZRANB2	1
CTCF	ZSCAN5A	1
E2F1	ABCC2	1
E2F1	ADAMTS1	1
E2F1	ADH4	1
E2F1	AHI1	1
E2F1	AMD1	1
E2F1	ANTXR1	1
E2F1	AP3D1	1
E2F1	AP3M2	1
E2F1	ARHGAP19	1
E2F1	ARHGAP8	1
E2F1	ASF1B	1
E2F1	ATF7IP	1
E2F1	ATL2	1
E2F1	AURKB	1
E2F1	BAG3	1

Table 35 continued

TF	TG	# dup
E2F1	BIRC5	1
E2F1	BORA	1
E2F1	CADM1	1
E2F1	CAPS	1
E2F1	CCNA2	1
E2F1	CCNB1	1
E2F1	CCNE1	1
E2F1	CCNF	1
E2F1	CDC27	1
E2F1	CDC45	1
E2F1	CDC6	1
E2F1	CDCA3	1
E2F1	CDCA7	1
E2F1	CDK7	1
E2F1	CDKL5	1
E2F1	CDKN3	1
E2F1	CENPE	1
E2F1	CENPF	1
E2F1	CHAF1A	1
E2F1	CHEK2	1
E2F1	CIT	1
E2F1	CKAP5	1
E2F1	CNIH4	1
E2F1	CNOT10	1
E2F1	COL7A1	1
E2F1	COQ6	1
E2F1	CTSD	1
E2F1	CYTH2	1
E2F1	DET1	1
E2F1	DHFR	1
E2F1	DTL	1
E2F1	E2F1	1
E2F1	E2F8	1
E2F1	FABP1	1
E2F1	FAM60A	1
E2F1	FANCA	1
E2F1	FANCD2	1
E2F1	FEN1	1
E2F1	FLAD1	1
E2F1	FOXM1	1
E2F1	FXR1	1
E2F1	FYN	1
E2F1	G2E3	1
E2F1	GCLM	1
E2F1	GDF15	1

Table 35 continued

TF	TG	# dup
E2F1	GINS3	1
E2F1	GMNN	1
E2F1	GOT1	1
E2F1	GPSM2	1
E2F1	HELLS	1
E2F1	HERPUD2	1
E2F1	HIST1H2AC	1
E2F1	HIST1H4E	1
E2F1	HLA-DOA	1
E2F1	HRAS	1
E2F1	HRSP12	1
E2F1	HSPB8	1
E2F1	INSR	1
E2F1	ITPR1	1
E2F1	KDM5B	1
E2F1	KIAA0586	1
E2F1	KIF14	1
E2F1	KIF20B	1
E2F1	KIF23	1
E2F1	KIF2C	1
E2F1	KIFC1	1
E2F1	KRAS	1
E2F1	LBR	1
E2F1	LPP	1
E2F1	LRIF1	1
E2F1	LRRC17	1
E2F1	MAD2L1	1
E2F1	MAN1A2	1
E2F1	MAP2K6	1
E2F1	MAPK13	1
E2F1	MCM8	1
E2F1	ME3	1
E2F1	MEGF9	1
E2F1	MELK	1
E2F1	MET	1
E2F1	MKI67	1
E2F1	MND1	1
E2F1	MNX1	1
E2F1	MRI1	1
E2F1	MRPS18B	1
E2F1	MSH2	1
E2F1	MZF1	1
E2F1	NCOA3	1
E2F1	NCS1	1
E2F1	NPAT	1

Table 35 continued

TF	TG	# dup
E2F1	NUDT4	1
E2F1	NUP160	1
E2F1	NUP37	1
E2F1	ODF2	1
E2F1	ORC3	1
E2F1	OSER1	1
E2F1	PBK	1
E2F1	PKNOX1	1
E2F1	PLIN3	1
E2F1	PLK1	1
E2F1	POC1A	1
E2F1	POM121	1
E2F1	PPP1R2	1
E2F1	PRIM2	1
E2F1	PRKAR1A	1
E2F1	PSEN1	1
E2F1	PTTG1	1
E2F1	PWP1	1
E2F1	QRICH1	1
E2F1	RAD18	1
E2F1	RAD51	1
E2F1	RAD54L	1
E2F1	REEP1	1
E2F1	RFC2	1
E2F1	RFC4	1
E2F1	RGS3	1
E2F1	RPA2	1
E2F1	RRM1	1
E2F1	RRM2	1
E2F1	RUNX1	1
E2F1	SAP30BP	1
E2F1	SEPHS1	1
E2F1	SGK1	1
E2F1	SLBP	1
E2F1	SLC44A2	1
E2F1	SP1	1
E2F1	SRD5A1	1
E2F1	SRSF5	1
E2F1	STAT5B	1
E2F1	STIL	1
E2F1	SUCLG2	1
E2F1	THRAP3	1
E2F1	TIMP1	1
E2F1	TMEM132A	1
E2F1	TOMM70A	1

Table 35 continued

TF	TG	# dup
E2F1	TOP1	1
E2F1	TOP2A	1
E2F1	TOP3A	1
E2F1	TOPBP1	1
E2F1	TRA2A	1
E2F1	TRIM45	1
E2F1	TRIP13	1
E2F1	TROAP	1
E2F1	TSG101	1
E2F1	TUBB2A	1
E2F1	TULP4	1
E2F1	TYMS	1
E2F1	UACA	1
E2F1	UBE2S	1
E2F1	UBE2T	1
E2F1	UBL3	1
E2F1	UBQLN2	1
E2F1	UHRF1	1
E2F1	USP1	1
E2F1	VCAM1	1
E2F1	VEGFC	1
E2F1	VPS37C	1
E2F1	WSB1	1
E2F1	YY1	1
E2F1	ZBED5	1
E2F1	ZBTB7A	1
E2F1	ZC3HC1	1
E2F1	ZMYM1	1
E2F1	ZNF143	1
E2F1	ZNF521	1
E2F1	ZSCAN5A	1
E2F1	ZWINT	1
E2F5	ASF1B	1
E2F5	BRCA1	1
E2F8	E2F1	1
FOXM1	AURKB	1
FOXM1	BIRC5	1
FOXM1	CCNB1	1
FOXM1	CDC25A	1
FOXM1	CDC6	1
FOXM1	CDKN1B	1
FOXM1	CKS1B	1
FOXM1	PDGFA	1
FOXM1	PLK1	1
HIF1A	ADAMTS1	1

Table 35 continued

TF	TG	# dup
HIF1A	ARL4A	1
HIF1A	CDK7	1
HIF1A	CDKN1B	1
HIF1A	DNAJB9	1
HIF1A	DYNLL1	1
HIF1A	FANCD2	1
HIF1A	FOXM1	1
HIF1A	GRPEL1	1
HIF1A	HERPUD2	1
HIF1A	HIF1A	1
HIF1A	HMMR	1
HIF1A	INSIG2	1
HIF1A	KDM5B	1
HIF1A	MET	1
HIF1A	MUC1	1
HIF1A	NR3C1	1
HIF1A	PCF11	1
HIF1A	PDXP	1
HIF1A	PLIN3	1
HIF1A	POM121	1
HIF1A	PPP6R3	1
HIF1A	PRPSAP1	1
HIF1A	RBM8A	1
HIF1A	RHOBTB3	1
HIF1A	RRM2	1
HIF1A	SAP30	1
HIF1A	TFF3	1
HIF1A	TIMP1	1
HIF1A	TOMM34	1
HIF1A	TOP3A	1
HIF1A	TYMS	1
HIF1A	VEGFC	1
HIF1A	WSB1	1
HIF1A	ZNF217	1
HSF2	HIF1A	1
INSM1	INSM1	1
KDM5B	BRCA1	1
KLF6	PTTG1	1
KLF9	TFAP2A	1
MITF	ABCC2	1
MITF	ACD	1
MITF	AFAP1	1
MITF	AHI1	1
MITF	AMD1	1
MITF	ANKRD10	1

Table 35 continued

TF	TG	# dup
MITF	ANP32B	1
MITF	ANTXR1	1
MITF	AP3D1	1
MITF	AP3M2	1
MITF	ARHGEF39	1
MITF	ARL4A	1
MITF	ASF1B	1
MITF	ASIP	1
MITF	ATF7IP	1
MITF	ATL2	1
MITF	BAG3	1
MITF	BMP2	1
MITF	BRCA1	1
MITF	BTBD3	1
MITF	BUB3	1
MITF	C6	1
MITF	CADM1	1
MITF	CBX3	1
MITF	CCNB1	1
MITF	CCNE1	1
MITF	CDC16	1
MITF	CDC25B	1
MITF	CDC7	1
MITF	CDKN1B	1
MITF	CDKN2AIP	1
MITF	CDKN2C	1
MITF	CENPA	1
MITF	CENPM	1
MITF	CFLAR	1
MITF	CHEK2	1
MITF	CIC	1
MITF	CIT	1
MITF	CKS2	1
MITF	CNOT10	1
MITF	CSGALNACT1	1
MITF	CTNND1	1
MITF	CYB5R2	1
MITF	DDX11	1
MITF	DEXI	1
MITF	DKC1	1
MITF	DMXL2	1
MITF	DNAJB1	1
MITF	DNAJB4	1
MITF	DNAJB6	1
MITF	DNAJB9	1

Table 35 continued

TF	TG	# dup
MITF	DR1	1
MITF	DSP	1
MITF	DUSP4	1
MITF	DYNLL1	1
MITF	E2F5	1
MITF	E2F8	1
MITF	FADD	1
MITF	FAM189B	1
MITF	FAM60A	1
MITF	FAM64A	1
MITF	FANCA	1
MITF	FEM1B	1
MITF	FEN1	1
MITF	FKBP1A	1
MITF	FZR1	1
MITF	GAS1	1
MITF	GAS6	1
MITF	GNB1	1
MITF	GTF2B	1
MITF	HAUS5	1
MITF	HAUS8	1
MITF	HERPUD2	1
MITF	HIF1A	1
MITF	HIST2H2BE	1
MITF	HOXB4	1
MITF	HP1BP3	1
MITF	HRAS	1
MITF	HSF2	1
MITF	HSPA8	1
MITF	IDI2	1
MITF	INADL	1
MITF	ITPR3	1
MITF	IVNS1ABP	1
MITF	JADE2	1
MITF	KANK2	1
MITF	KAT2B	1
MITF	KBTBD2	1
MITF	KDELC1	1
MITF	KDM4A	1
MITF	KDM5B	1
MITF	KIFC1	1
MITF	KLF6	1
MITF	KLF9	1
MITF	KPNA2	1
MITF	LBR	1

Table 35 continued

TF	TG	# dup
MITF	MAN1A2	1
MITF	MAPK13	1
MITF	MBD3	1
MITF	MCAM	1
MITF	MCM8	1
MITF	ME3	1
MITF	MIS18BP1	1
MITF	MNT	1
MITF	MNX1	1
MITF	MORF4L2	1
MITF	MSH2	1
MITF	MTCL1	1
MITF	MZF1	1
MITF	NAB1	1
MITF	NCAPH	1
MITF	NCOA3	1
MITF	NCOA5	1
MITF	NDE1	1
MITF	NFE2L2	1
MITF	NFIC	1
MITF	NPM1	1
MITF	NSUN3	1
MITF	OSER1	1
MITF	PAK1IP1	1
MITF	PANK2	1
MITF	PCF11	1
MITF	PDGFA	1
MITF	PDXP	1
MITF	PIK3CD	1
MITF	PLIN3	1
MITF	PLK1	1
MITF	POC1A	1
MITF	POLA1	1
MITF	PPP1R10	1
MITF	PRIM2	1
MITF	PRKAR1A	1
MITF	PRR16	1
MITF	PTP4A1	1
MITF	PTTG1	1
MITF	PWP1	1
MITF	QRICH1	1
MITF	RAB3A	1
MITF	RAN	1
MITF	RCCD1	1
MITF	RHEB	1

Table 35 continued

TF	TG	# dup
MITF	RMI1	1
MITF	RRM2	1
MITF	RRP1	1
MITF	SAP30	1
MITF	SAP30BP	1
MITF	SGK1	1
MITF	SLC25A36	1
MITF	SLC38A2	1
MITF	SMARCB1	1
MITF	SMTN	1
MITF	SP1	1
MITF	SRSF3	1
MITF	SS18	1
MITF	SSR3	1
MITF	STAG1	1
MITF	STAT1	1
MITF	SV2B	1
MITF	SYNCRIP	1
MITF	TAB2	1
MITF	TACC3	1
MITF	TFAP2A	1
MITF	TGIF1	1
MITF	TOB2	1
MITF	TOMM34	1
MITF	TOP1	1
MITF	TOP3A	1
MITF	TRAIP	1
MITF	TRIP13	1
MITF	TSC22D1	1
MITF	TSG101	1
MITF	TSKU	1
MITF	TSN	1
MITF	TTC38	1
MITF	TUBB2A	1
MITF	TUBB4B	1
MITF	TULP4	1
MITF	TXNRD1	1
MITF	UACA	1
MITF	UBE2D3	1
MITF	UBL3	1
MITF	UHRF1	1
MITF	UNG	1
MITF	USP1	1
MITF	USP13	1
MITF	VEGFC	1

Table 35 continued

TF	TG	# dup
MITF	VPS37C	1
MITF	WSB1	1
MITF	YWHAH	1
MITF	YY1	1
MITF	ZBED5	1
MITF	ZC3HC1	1
MITF	ZCCHC10	1
MITF	ZNF217	1
MITF	ZNFX1	1
MITF	ZNHIT2	1
NCOA3	BRCA1	1
NFE2L2	BRCA1	1
NFIA	NR3C1	1
NFIC	HRAS	1
NFIC	INSR	1
NFIC	NR3C1	1
NFIC	TFAP2A	1
NFYA	CDC25A	1
NFYA	CDCA8	1
NFYA	CDKN1B	1
NFYA	E2F1	1
NFYA	GADD45A	1
NFYA	HOXB4	1
NFYA	MCM8	1
NFYA	PTTG1	1
NFYB	CDKN1B	1
NFYB	HLA-DOA	1
NFYB	HLA-DRA	1
NFYB	HSPA13	1
NR3C1	BRCA1	1
NR3C1	NR3C1	1
NR3C1	SRF	1
NR3C1	STAT1	1
RUNX1	ADAMTS1	1
RUNX1	BBS2	1
RUNX1	BCLAF1	1
RUNX1	BIRC2	1
RUNX1	C5orf42	1
RUNX1	CDC25B	1
RUNX1	CENPF	1
RUNX1	CENPL	1
RUNX1	CEP70	1
RUNX1	CKAP2	1
RUNX1	CKS2	1
RUNX1	CTR9	1

Table 35 continued

TF	TG	# dup
RUNX1	CXCL14	1
RUNX1	DEPDC1B	1
RUNX1	DNA2	1
RUNX1	DNAJC3	1
RUNX1	EIF4E	1
RUNX1	FAM105A	1
RUNX1	FXR1	1
RUNX1	GPSM2	1
RUNX1	HIST1H2BC	1
RUNX1	HSF2	1
RUNX1	INADL	1
RUNX1	IVNS1ABP	1
RUNX1	KLF6	1
RUNX1	KPNA2	1
RUNX1	LARP7	1
RUNX1	MAD2L1	1
RUNX1	MAN1A2	1
RUNX1	MAP2K6	1
RUNX1	ME3	1
RUNX1	MKI67	1
RUNX1	MND1	1
RUNX1	MTCL1	1
RUNX1	NCOA3	1
RUNX1	NEIL3	1
RUNX1	NSUN3	1
RUNX1	NUP98	1
RUNX1	ORC3	1
RUNX1	PIK3CD	1
RUNX1	PPP6R3	1
RUNX1	PRIM2	1
RUNX1	PTP4A1	1
RUNX1	ROCK1	1
RUNX1	SGK1	1
RUNX1	SLC25A27	1
RUNX1	SLC25A36	1
RUNX1	SLC38A2	1
RUNX1	SLC39A10	1
RUNX1	SPAG5	1
RUNX1	STAG1	1
RUNX1	SUCLG2	1
RUNX1	TRIP13	1
RUNX1	TSKU	1
RUNX1	UACA	1
RUNX1	VCL	1
RUNX1	WSB1	1

Table 35 continued

TF	TG	# dup
RUNX1	ZBPB	1
RUNX1	ZRANB2	1
SP1	BIRC5	1
SP1	BRCA1	1
SP1	BUB1B	1
SP1	C4B	1
SP1	CASP3	1
SP1	CCNA2	1
SP1	CCNB1	1
SP1	CDC25A	1
SP1	CDC25C	1
SP1	CDKN1B	1
SP1	CDKN2C	1
SP1	CDKN2D	1
SP1	COL7A1	1
SP1	CTSD	1
SP1	CXCL14	1
SP1	DHFR	1
SP1	DKC1	1
SP1	E2F1	1
SP1	EXO1	1
SP1	FOXM1	1
SP1	HIF1A	1
SP1	HSD17B11	1
SP1	HSPA8	1
SP1	ITGB3	1
SP1	KIF2C	1
SP1	LMO4	1
SP1	MCAM	1
SP1	MDM2	1
SP1	NR3C1	1
SP1	PDGFA	1
SP1	POLA1	1
SP1	PSEN1	1
SP1	PTTG1	1
SP1	RECQL4	1
SP1	SP1	1
SP1	TIMP1	1
SP1	TMPO	1
SP1	TYMS	1
SP1	UNG	1
SRF	KPNB1	1
SRF	UBE2S	1
STAT1	ABCA7	1
STAT1	ABCC2	1

Table 35 continued

TF	TG	# dup
STAT1	ADAMTS1	1
STAT1	ADCY6	1
STAT1	AFAP1	1
STAT1	AGFG1	1
STAT1	AHI1	1
STAT1	AKIRIN2	1
STAT1	ANKRD10	1
STAT1	ANP32B	1
STAT1	ANP32E	1
STAT1	ANTXR1	1
STAT1	AP3M2	1
STAT1	ARHGAP11A	1
STAT1	ARHGAP19	1
STAT1	ARHGDIB	1
STAT1	ARHGEF39	1
STAT1	ARL6IP1	1
STAT1	ARMC1	1
STAT1	ASF1B	1
STAT1	ATAD2	1
STAT1	ATF7IP	1
STAT1	B2M	1
STAT1	BAG3	1
STAT1	BARD1	1
STAT1	BCLAF1	1
STAT1	BIRC2	1
STAT1	BMP2	1
STAT1	BRC A1	1
STAT1	BRD7	1
STAT1	BTBD3	1
STAT1	BUB3	1
STAT1	C5orf42	1
STAT1	C6	1
STAT1	CADM1	1
STAT1	CASP3	1
STAT1	CBX3	1
STAT1	CCDC90B	1
STAT1	CCNA2	1
STAT1	CCNE1	1
STAT1	CDC16	1
STAT1	CDC20	1
STAT1	CDC25B	1
STAT1	CDC25C	1
STAT1	CDC27	1
STAT1	CDC42EP1	1
STAT1	CDC42EP4	1

Table 35 continued

TF	TG	# dup
STAT1	CDC45	1
STAT1	CDCA7	1
STAT1	CDCA7L	1
STAT1	CDKN1B	1
STAT1	CDKN2AIP	1
STAT1	CDKN2C	1
STAT1	CDR2	1
STAT1	CENPA	1
STAT1	CENPE	1
STAT1	CENPM	1
STAT1	CEP44	1
STAT1	CFD	1
STAT1	CHAF1A	1
STAT1	CHEK2	1
STAT1	CIC	1
STAT1	CIT	1
STAT1	CKS2	1
STAT1	CLSPN	1
STAT1	CNIH4	1
STAT1	CNOT10	1
STAT1	CREBZF	1
STAT1	CRK	1
STAT1	CRYBA1	1
STAT1	CSGALNACT1	1
STAT1	CSH2	1
STAT1	CTCF	1
STAT1	CTNND1	1
STAT1	CTR9	1
STAT1	CTSD	1
STAT1	CYTH2	1
STAT1	CYTH3	1
STAT1	DCTN6	1
STAT1	DEPDC1B	1
STAT1	DHFR	1
STAT1	DHX8	1
STAT1	DIS3	1
STAT1	DLGAP5	1
STAT1	DNAJB1	1
STAT1	DNAJB6	1
STAT1	DNAJB9	1
STAT1	DNAJC3	1
STAT1	DNAJC6	1
STAT1	DR1	1
STAT1	DSCC1	1
STAT1	DTL	1

Table 35 continued

TF	TG	# dup
STAT1	DUSP4	1
STAT1	DYNLL1	1
STAT1	DZIP3	1
STAT1	E2F1	1
STAT1	E2F8	1
STAT1	ELP3	1
STAT1	ERN2	1
STAT1	ESPL1	1
STAT1	FADD	1
STAT1	FAM105A	1
STAT1	FAM110A	1
STAT1	FAM214A	1
STAT1	FAM60A	1
STAT1	FAM83D	1
STAT1	FANCA	1
STAT1	FANCG	1
STAT1	FANCI	1
STAT1	FEM1B	1
STAT1	FEN1	1
STAT1	FKBP1A	1
STAT1	FLAD1	1
STAT1	G2E3	1
STAT1	G3BP1	1
STAT1	GADD45A	1
STAT1	GCSH	1
STAT1	GINS3	1
STAT1	GMNN	1
STAT1	GOT1	1
STAT1	GRK6	1
STAT1	GTF2B	1
STAT1	H1F0	1
STAT1	HCP5	1
STAT1	HERPUD2	1
STAT1	HIF1A	1
STAT1	HIST1H2AC	1
STAT1	HIST1H4H	1
STAT1	HIST2H2BE	1
STAT1	HMGCR	1
STAT1	HMMR	1
STAT1	HN1	1
STAT1	HP1BP3	1
STAT1	HRAS	1
STAT1	HSD17B11	1
STAT1	HSF2	1
STAT1	HSPA1L	1

Table 35 continued

TF	TG	# dup
STAT1	HSPA8	1
STAT1	HSPB8	1
STAT1	IDO1	1
STAT1	IFIT1	1
STAT1	IL18BP	1
STAT1	ILF2	1
STAT1	INADL	1
STAT1	INPP5K	1
STAT1	INSIG2	1
STAT1	INSR	1
STAT1	INTS7	1
STAT1	ITPR3	1
STAT1	IVNS1ABP	1
STAT1	KAT2B	1
STAT1	KATNBL1	1
STAT1	KBTBD2	1
STAT1	KCTD2	1
STAT1	KDM4A	1
STAT1	KDM5B	1
STAT1	KIAA0101	1
STAT1	KIAA1524	1
STAT1	KIF11	1
STAT1	KIF14	1
STAT1	KIF20B	1
STAT1	KIF22	1
STAT1	KIF5B	1
STAT1	KPNA2	1
STAT1	KRAS	1
STAT1	LBR	1
STAT1	LMNB1	1
STAT1	LMO4	1
STAT1	LRIF1	1
STAT1	LRRC17	1
STAT1	MAN1A2	1
STAT1	MAP2K6	1
STAT1	MATN2	1
STAT1	MBD2	1
STAT1	MCM2	1
STAT1	MCM4	1
STAT1	MDC1	1
STAT1	MDM2	1
STAT1	ME3	1
STAT1	MED31	1
STAT1	MEGF9	1
STAT1	MELK	1

Table 35 continued

TF	TG	# dup
STAT1	MET	1
STAT1	MID1	1
STAT1	MKI67	1
STAT1	MND1	1
STAT1	MNX1	1
STAT1	MORF4L2	1
STAT1	MTCL1	1
STAT1	MZF1	1
STAT1	NAB1	1
STAT1	NASP	1
STAT1	NCAPH	1
STAT1	NCOA3	1
STAT1	NCOA5	1
STAT1	NCS1	1
STAT1	NDC80	1
STAT1	NDE1	1
STAT1	NEIL3	1
STAT1	NEK2	1
STAT1	NFIC	1
STAT1	NFYB	1
STAT1	NIPBL	1
STAT1	NKTR	1
STAT1	NNMT	1
STAT1	NSUN3	1
STAT1	NUCKS1	1
STAT1	NUF2	1
STAT1	NUP160	1
STAT1	NUP98	1
STAT1	OGT	1
STAT1	OLR1	1
STAT1	OSER1	1
STAT1	OSGIN2	1
STAT1	OXR1	1
STAT1	PAK1IP1	1
STAT1	PBK	1
STAT1	PDGFA	1
STAT1	PIK3CD	1
STAT1	PKNOX1	1
STAT1	PLIN3	1
STAT1	PLK2	1
STAT1	POC1A	1
STAT1	POLD3	1
STAT1	POLQ	1
STAT1	POM121	1
STAT1	PPP1R2	1

Table 35 continued

TF	TG	# dup
STAT1	PPP3CA	1
STAT1	PPP6R3	1
STAT1	PRC1	1
STAT1	PRIM2	1
STAT1	PRPSAP1	1
STAT1	PRR11	1
STAT1	PRR16	1
STAT1	PSEN1	1
STAT1	PSMG3	1
STAT1	PTP4A1	1
STAT1	PTTG1	1
STAT1	RAB23	1
STAT1	RAD18	1
STAT1	RAD21	1
STAT1	RAD51	1
STAT1	RAD51AP1	1
STAT1	RAD51C	1
STAT1	RAD54L	1
STAT1	RANGAP1	1
STAT1	RBBP8	1
STAT1	RCAN1	1
STAT1	RCCD1	1
STAT1	REEP1	1
STAT1	RFC2	1
STAT1	RFC4	1
STAT1	RGS3	1
STAT1	RHEB	1
STAT1	RHNO1	1
STAT1	RHOBTB3	1
STAT1	RMI1	1
STAT1	RNPC3	1
STAT1	RNPS1	1
STAT1	RRM2	1
STAT1	RRP1	1
STAT1	RSRC2	1
STAT1	SAP30	1
STAT1	SAP30BP	1
STAT1	SDC1	1
STAT1	SEPHS1	1
STAT1	SERPINB3	1
STAT1	SFPQ	1
STAT1	SH3GL2	1
STAT1	SHC1	1
STAT1	SLC22A3	1
STAT1	SLC25A36	1

Table 35 continued

TF	TG	# dup
STAT1	SLC38A2	1
STAT1	SLC39A10	1
STAT1	SLC4A1AP	1
STAT1	SMARCB1	1
STAT1	SMARCD1	1
STAT1	SNUPN	1
STAT1	SP1	1
STAT1	SPAG5	1
STAT1	SRD5A1	1
STAT1	SRF	1
STAT1	SRSF3	1
STAT1	SS18	1
STAT1	STAG3	1
STAT1	STAT5B	1
STAT1	STIL	1
STAT1	SUCLG2	1
STAT1	SV2B	1
STAT1	SYNCRIP	1
STAT1	TAB2	1
STAT1	TACC3	1
STAT1	TFAP2A	1
STAT1	TFF3	1
STAT1	TGIF1	1
STAT1	THRAP3	1
STAT1	TIMP1	1
STAT1	TIPIN	1
STAT1	TMPO	1
STAT1	TOB2	1
STAT1	TOMM34	1
STAT1	TOP1	1
STAT1	TOP2A	1
STAT1	TOP3A	1
STAT1	TPX2	1
STAT1	TRA2A	1
STAT1	TRIP13	1
STAT1	TSG101	1
STAT1	TSKU	1
STAT1	TSN	1
STAT1	TTC31	1
STAT1	TTF2	1
STAT1	TUBA1A	1
STAT1	TUBB2A	1
STAT1	TULP4	1
STAT1	TXNRD1	1
STAT1	UACA	1

Table 35 continued

TF	TG	# dup
STAT1	UBE2D3	1
STAT1	UBL3	1
STAT1	UBQLN2	1
STAT1	UHRF1	1
STAT1	USP1	1
STAT1	USP6NL	1
STAT1	VCAM1	1
STAT1	VEGFC	1
STAT1	VPS25	1
STAT1	VPS72	1
STAT1	VTA1	1
STAT1	WSB1	1
STAT1	YY1	1
STAT1	ZBTB7A	1
STAT1	ZC3HC1	1
STAT1	ZCCHC10	1
STAT1	ZNF143	1
STAT1	ZNF207	1
STAT1	ZNF217	1
STAT1	ZNF281	1
STAT1	ZNF414	1
STAT1	ZNF521	1
STAT1	ZNFX1	1
STAT1	ZNHIT2	1
STAT1	ZPBP	1
STAT1	ZRANB2	1
STAT1	ZSCAN5A	1
STAT1	ZWINT	1
STAT5B	MET	1
STAT5B	MUC1	1
STAT5B	RAD51	1
TFAP2A	CCNB1	1
TFAP2A	CTSD	1
TFAP2A	DHX8	1
TFAP2A	HSPA8	1
TFAP2A	MCAM	1
TFAP2A	NR3C1	1
TFAP2A	RECQL4	1
TFAP2A	TFAP2A	1
TFAP2A	TIMP1	1
TFAP2A	TOP1	1
TFAP2A	VEGFC	1
YY1	BRCA1	1
YY1	CDC6	1
YY1	DKC1	1

Table 35 continued

TF	TG	# dup
YY1	DNAJB4	1
YY1	HDAC3	1
YY1	HIF1A	1
YY1	HLA-DRA	1
YY1	MCM5	1
YY1	NUP160	1
YY1	PCNA	1
YY1	SAP30	1
YY1	TFAP2A	1
ZNF143	BUB1B	1

The table presents for each edge the number of time it appears in the network obtained after concatenating the two networks collected from the work of Alonso *et.al* [78]. The authors generated one gold standard network for cancer cells and one for normal cells. The 1st column represents the TF official name. The 2nd column the TG official name. The 3rd column represents the of times the link is repeated after a per row concatenation of the two networks.

Table 36: HeLa “gold-standard” network

- Positive links

TF	TG	W	# dup
BRCA1	BARD1	1	1
BRCA1	CHEK2	1	1
BRCA1	DTL	1	1
BRCA1	DZIP3	1	1
BRCA1	ERN2	1	1
BRCA1	FAN1	1	1
BRCA1	FANCD2	1	1
BRCA1	G2E3	1	1
BRCA1	HAUS5	1	1
BRCA1	HELLS	1	1
BRCA1	KDM4A	1	1
BRCA1	NEK2	1	1
BRCA1	PLK1	1	1
BRCA1	RAD51	1	1
BRCA1	TRIP13	1	1
CENPA	NDE1	1	1
CTCF	ABCA7	1	2
CTCF	ABCC2	1	2
CTCF	ABHD10	1	2
CTCF	ADCY6	1	2
CTCF	ADH4	1	3
CTCF	AHI1	1	2
CTCF	AMD1	1	2
CTCF	ANLN	1	2
CTCF	ANP32B	1	2
CTCF	ANP32E	1	2
CTCF	AOC2	1	2
CTCF	AOC3	1	2
CTCF	AP3D1	1	2
CTCF	AP3M2	1	2
CTCF	AP4B1	1	2
CTCF	ARHGAP11A	1	2
CTCF	ARHGAP19	1	2
CTCF	ARHGAP8	1	2
CTCF	ARHGEF39	1	2
CTCF	ARL4A	1	2
CTCF	ARL6IP1	1	2
CTCF	ASF1B	1	3
CTCF	ASIP	1	2
CTCF	ASPHD2	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	ATAD2	1	2
CTCF	ATF7IP	1	2
CTCF	ATL2	1	2
CTCF	AURKB	1	3
CTCF	B2M	1	2
CTCF	BAG3	1	2
CTCF	BAIAP2	1	2
CTCF	BBS2	1	2
CTCF	BCLAF1	1	2
CTCF	BIRC2	1	2
CTCF	BIVM	1	2
CTCF	BMP2	1	2
CTCF	BRCA1	1	2
CTCF	BRD7	1	2
CTCF	BTBD3	1	2
CTCF	BUB3	1	2
CTCF	C6	1	3
CTCF	CADM1	1	2
CTCF	CAPN7	1	2
CTCF	CASP3	1	2
CTCF	CCDC90B	1	2
CTCF	CCNE1	1	3
CTCF	CCNF	1	2
CTCF	CDC16	1	2
CTCF	CDC20	1	2
CTCF	CDC25A	1	3
CTCF	CDC25B	1	2
CTCF	CDC25C	1	2
CTCF	CDC42	1	2
CTCF	CDC42EP1	1	2
CTCF	CDC42EP4	1	2
CTCF	CDC45	1	2
CTCF	CDC6	1	2
CTCF	CDC7	1	2
CTCF	CDCA7	1	2
CTCF	CDCA7L	1	2
CTCF	CDK20	1	2
CTCF	CDK7	1	2
CTCF	CDKN1B	1	2
CTCF	CDKN2AIP	1	2
CTCF	CDKN2C	1	2
CTCF	CDKN3	1	2
CTCF	CENPA	1	2
CTCF	CENPE	1	2
CTCF	CENPF	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	CENPM	1	2
CTCF	CEP44	1	2
CTCF	CEP55	1	2
CTCF	CEP70	1	2
CTCF	CFD	1	2
CTCF	CFLAR	1	2
CTCF	CHAF1B	1	2
CTCF	CHEK2	1	2
CTCF	CIC	1	2
CTCF	CIT	1	2
CTCF	CKAP5	1	2
CTCF	CKS2	1	2
CTCF	CLSPN	1	2
CTCF	CNN2	1	2
CTCF	CNOT10	1	2
CTCF	COQ6	1	2
CTCF	CREBZF	1	2
CTCF	CRK	1	2
CTCF	CRYBA1	1	2
CTCF	CSH2	1	2
CTCF	CTCF	1	2
CTCF	CTNND1	1	2
CTCF	CTR9	1	2
CTCF	CTSD	1	2
CTCF	CWC15	1	2
CTCF	CXCL14	1	2
CTCF	CYB5R2	1	2
CTCF	CYTH3	1	2
CTCF	DCAF16	1	2
CTCF	DCAF7	1	2
CTCF	DCTN6	1	2
CTCF	DDX11	1	2
CTCF	DEPDC1B	1	2
CTCF	DET1	1	2
CTCF	DHX8	1	2
CTCF	DLGAP5	1	2
CTCF	DMTF1	1	2
CTCF	DMXL2	1	2
CTCF	DNAJB1	1	2
CTCF	DNAJB4	1	2
CTCF	DNAJB6	1	2
CTCF	DNAJB9	1	2
CTCF	DNAJC3	1	2
CTCF	DNAJC6	1	2
CTCF	DTL	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	DUSP4	1	2
CTCF	DYNLL1	1	2
CTCF	DZIP3	1	3
CTCF	E2F1	1	2
CTCF	E2F5	1	2
CTCF	E2F8	1	2
CTCF	EBI3	1	2
CTCF	EIF4E	1	2
CTCF	ELP3	1	2
CTCF	ENOSF1	1	2
CTCF	ERN2	1	2
CTCF	ESPL1	1	2
CTCF	EXO1	1	2
CTCF	FABP1	1	2
CTCF	FADD	1	2
CTCF	FAM105A	1	2
CTCF	FAM110A	1	2
CTCF	FAM189B	1	2
CTCF	FAM214A	1	2
CTCF	FAM60A	1	2
CTCF	FANCA	1	2
CTCF	FANCI	1	2
CTCF	FBXL20	1	2
CTCF	FEM1B	1	2
CTCF	FEN1	1	2
CTCF	FKBP1A	1	2
CTCF	FLAD1	1	2
CTCF	FXR1	1	2
CTCF	G2E3	1	3
CTCF	G3BP1	1	2
CTCF	GAS1	1	2
CTCF	GAS6	1	2
CTCF	GDF15	1	2
CTCF	GINS2	1	2
CTCF	GINS3	1	2
CTCF	GMNN	1	2
CTCF	GNB1	1	2
CTCF	GOLGA8A	1	2
CTCF	GOT1	1	2
CTCF	GPSM2	1	3
CTCF	GRK6	1	2
CTCF	GRPEL1	1	2
CTCF	GTF2B	1	2
CTCF	GTSE1	1	2
CTCF	H2AFX	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	HAUS5	1	3
CTCF	HAUS8	1	2
CTCF	HCP5	1	2
CTCF	HELLS	1	3
CTCF	HERPUD2	1	2
CTCF	HIF1A	1	2
CTCF	HIST1H4C	1	2
CTCF	HIST1H4E	1	2
CTCF	HIST1H4H	1	2
CTCF	HJURP	1	2
CTCF	HLA-DOA	1	3
CTCF	HLA-DRA	1	2
CTCF	HMG20B	1	2
CTCF	HMGCR	1	2
CTCF	HMMR	1	2
CTCF	HRAS	1	2
CTCF	HSD17B11	1	3
CTCF	HSF2	1	2
CTCF	HSPA13	1	2
CTCF	HSPB8	1	2
CTCF	IDO1	1	2
CTCF	ILF2	1	2
CTCF	INADL	1	2
CTCF	INPP5K	1	2
CTCF	INSIG2	1	2
CTCF	INSM1	1	3
CTCF	INSR	1	2
CTCF	INTS7	1	2
CTCF	ITPR3	1	2
CTCF	IVNS1ABP	1	2
CTCF	KANK2	1	2
CTCF	KAT2B	1	2
CTCF	KCTD2	1	2
CTCF	KDM4A	1	3
CTCF	KDM5B	1	2
CTCF	KIAA0586	1	2
CTCF	KIAA1147	1	2
CTCF	KIAA1524	1	2
CTCF	KIF11	1	2
CTCF	KIF14	1	2
CTCF	KIF20B	1	2
CTCF	KIF22	1	2
CTCF	KIF5B	1	2
CTCF	KIFC1	1	2
CTCF	KLF6	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	KLF9	1	2
CTCF	KMO	1	2
CTCF	KPNA2	1	2
CTCF	KPNB1	1	2
CTCF	KRAS	1	2
CTCF	LARP7	1	2
CTCF	LMNB1	1	2
CTCF	LMO4	1	2
CTCF	LPP	1	2
CTCF	LRIF1	1	2
CTCF	LYAR	1	2
CTCF	MAD2L1	1	2
CTCF	MAN1A2	1	2
CTCF	MAP2K6	1	2
CTCF	MAP3K2	1	2
CTCF	MAPK13	1	2
CTCF	MATN2	1	2
CTCF	MBD2	1	2
CTCF	MBD3	1	2
CTCF	MCAM	1	2
CTCF	MCM5	1	2
CTCF	MCM8	1	2
CTCF	MDC1	1	2
CTCF	MDM2	1	2
CTCF	ME3	1	3
CTCF	MED31	1	2
CTCF	MEGF9	1	2
CTCF	MELK	1	2
CTCF	MET	1	3
CTCF	MGAT2	1	2
CTCF	MID1	1	2
CTCF	MIS18BP1	1	2
CTCF	MITF	1	2
CTCF	MKI67	1	2
CTCF	MLLT4	1	2
CTCF	MND1	1	2
CTCF	MNT	1	2
CTCF	MNX1	1	3
CTCF	MORF4L2	1	2
CTCF	MRPL19	1	2
CTCF	MRPS2	1	2
CTCF	MSH2	1	2
CTCF	MTCL1	1	2
CTCF	MYCBP2	1	2
CTCF	MZF1	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	NAB1	1	2
CTCF	NCAPD2	1	2
CTCF	NCAPD3	1	2
CTCF	NCAPH	1	2
CTCF	NCOA3	1	2
CTCF	NCOA5	1	2
CTCF	NCS1	1	2
CTCF	NDE1	1	2
CTCF	NEIL3	1	3
CTCF	NEK2	1	3
CTCF	NFIC	1	2
CTCF	NFYA	1	2
CTCF	NFYB	1	2
CTCF	NIPBL	1	2
CTCF	NKTR	1	2
CTCF	NMB	1	2
CTCF	NNMT	1	2
CTCF	NPAT	1	2
CTCF	NPM1	1	2
CTCF	NR3C1	1	2
CTCF	NSUN3	1	2
CTCF	NUCKS1	1	2
CTCF	NUDT4	1	2
CTCF	NUF2	1	2
CTCF	NUP160	1	2
CTCF	NUP37	1	2
CTCF	ODF2	1	2
CTCF	OGT	1	2
CTCF	OLR1	1	2
CTCF	ORC3	1	2
CTCF	OSER1	1	2
CTCF	PANK2	1	2
CTCF	PCNA	1	2
CTCF	PDGFA	1	2
CTCF	PDXP	1	2
CTCF	PIK3CD	1	3
CTCF	PKMYT1	1	2
CTCF	PLIN3	1	2
CTCF	PLK1	1	2
CTCF	PLK2	1	2
CTCF	POC1A	1	2
CTCF	POLA1	1	2
CTCF	POLD3	1	2
CTCF	POLQ	1	2
CTCF	POM121	1	2

Table 36 continued

TF	TG	W	# dup
CTCF	PPP1R2	1	2
CTCF	PPP3CA	1	1
CTCF	PPP6R3	1	1
CTCF	PRIM1	1	1
CTCF	PRIM2	1	1
CTCF	PRKAR1A	1	1
CTCF	PRPSAP1	1	1
CTCF	PRR11	1	1
CTCF	PRR16	1	1
CTCF	PSEN1	1	1
CTCF	PSMD11	1	1
CTCF	PSMG3	1	1
CTCF	PTMS	1	1
CTCF	PTP4A1	1	1
CTCF	PTPN9	1	1
CTCF	PTTG1	1	1
CTCF	PWP1	1	1
CTCF	QRICH1	1	1
CTCF	RAB23	1	1
CTCF	RAB3A	1	1
CTCF	RAD18	1	2
CTCF	RAD21	1	1
CTCF	RAD51	1	1
CTCF	RAD51C	1	1
CTCF	RAD54L	1	1
CTCF	RAN	1	1
CTCF	RANGAP1	1	1
CTCF	RBBP8	1	1
CTCF	RBM8A	1	1
CTCF	RCAN1	1	1
CTCF	REEP1	1	2
CTCF	RFC4	1	1
CTCF	RGS3	1	1
CTCF	RHEB	1	1
CTCF	RHOBTB3	1	1
CTCF	RNF126	1	1
CTCF	ROCK1	1	1
CTCF	RPL13A	1	1
CTCF	RRM1	1	1
CTCF	RRM2	1	1
CTCF	RRP1	1	1
CTCF	SAP30	1	1
CTCF	SAP30BP	1	1
CTCF	SDC1	1	2
CTCF	SEC62	1	1

Table 36 continued

TF	TG	W	# dup
CTCF	SEPHS1	1	1
CTCF	SEPN1	1	1
CTCF	SGK1	1	1
CTCF	SH3GL2	1	2
CTCF	SHCBP1	1	1
CTCF	SLBP	1	1
CTCF	SLC17A2	1	1
CTCF	SLC22A3	1	2
CTCF	SLC25A27	1	1
CTCF	SLC25A36	1	1
CTCF	SLC38A2	1	1
CTCF	SLC39A10	1	1
CTCF	SLC44A2	1	1
CTCF	SLC4A1AP	1	1
CTCF	SMARCB1	1	1
CTCF	SMARCD1	1	1
CTCF	SMC4	1	1
CTCF	SMTN	1	1
CTCF	SNUPN	1	1
CTCF	SP1	1	1
CTCF	SPDL1	1	1
CTCF	SRF	1	1
CTCF	SS18	1	1
CTCF	SSR3	1	1
CTCF	STAG3	1	1
CTCF	STAT1	1	1
CTCF	STAT5B	1	1
CTCF	STIL	1	2
CTCF	SUCLG2	1	1
CTCF	TAB2	1	1
CTCF	TFAP2A	1	1
CTCF	TGIF1	1	1
CTCF	THRAP3	1	1
CTCF	TMPO	1	1
CTCF	TNPO2	1	1
CTCF	TOMM34	1	1
CTCF	TOP1	1	1
CTCF	TOP2A	1	1
CTCF	TPX2	1	1
CTCF	TRA2A	1	1
CTCF	TRAIP	1	1
CTCF	TRIM45	1	1
CTCF	TRIP13	1	2
CTCF	TROAP	1	1
CTCF	TSC22D1	1	1

Table 36 continued

TF	TG	W	# dup
CTCF	TSKU	1	1
CTCF	TSN	1	1
CTCF	TTC31	1	1
CTCF	TTF2	1	1
CTCF	TTK	1	1
CTCF	TUBB2A	1	1
CTCF	TUBB4B	1	1
CTCF	TUBD1	1	1
CTCF	TULP4	1	1
CTCF	TXNRD1	1	1
CTCF	TYMS	1	1
CTCF	UACA	1	1
CTCF	UBE2D3	1	1
CTCF	UBE2S	1	1
CTCF	UBL3	1	1
CTCF	UBR7	1	1
CTCF	UHRF1	1	1
CTCF	UNG	1	1
CTCF	USP1	1	1
CTCF	USP13	1	1
CTCF	USP53	1	1
CTCF	USP6NL	1	1
CTCF	VCAM1	1	1
CTCF	VCL	1	1
CTCF	VEGFC	1	1
CTCF	VPS37C	1	1
CTCF	VPS72	1	1
CTCF	VTA1	1	1
CTCF	WSB1	1	1
CTCF	YWHAH	1	1
CTCF	YY1	1	1
CTCF	ZBED5	1	1
CTCF	ZBTB7A	1	1
CTCF	ZC3HC1	1	1
CTCF	ZMYM1	1	1
CTCF	ZNF143	1	1
CTCF	ZNF217	1	1
CTCF	ZNF281	1	1
CTCF	ZNF414	1	1
CTCF	ZNF521	1	1
CTCF	ZNF593	1	1
CTCF	ZNFX1	1	1
CTCF	ZNHIT2	1	1
CTCF	ZPBP	1	1
CTCF	ZRANB2	1	1

Table 36 continued

TF	TG	W	# dup
CTCF	ZSCAN5A	1	1
E2F1	ABCC2	1	1
E2F1	ADAMTS1	1	2
E2F1	ADH4	1	2
E2F1	AHI1	1	2
E2F1	AMD1	1	1
E2F1	ANTXR1	1	1
E2F1	AP3D1	1	1
E2F1	AP3M2	1	1
E2F1	ARHGAP19	1	1
E2F1	ARHGAP8	1	1
E2F1	ASF1B	1	2
E2F1	ATF7IP	1	1
E2F1	ATL2	1	1
E2F1	AURKB	1	2
E2F1	BAG3	1	2
E2F1	BIRC5	1	1
E2F1	BORA	1	1
E2F1	BRD7	1	1
E2F1	CADM1	1	1
E2F1	CAPS	1	1
E2F1	CCNA2	1	1
E2F1	CCNB1	1	1
E2F1	CCNE1	1	2
E2F1	CCNF	1	1
E2F1	CDC27	1	1
E2F1	CDC45	1	1
E2F1	CDC6	1	1
E2F1	CDCA3	1	1
E2F1	CDCA7	1	1
E2F1	CDK7	1	1
E2F1	CDKL5	1	2
E2F1	CDKN1B	1	1
E2F1	CDKN2C	1	1
E2F1	CDKN2D	1	1
E2F1	CDKN3	1	1
E2F1	CENPE	1	1
E2F1	CENPF	1	1
E2F1	CHAF1A	1	2
E2F1	CHEK2	1	1
E2F1	CIT	1	1
E2F1	CKAP5	1	1
E2F1	CNIH4	1	1
E2F1	CNOT10	1	1
E2F1	COL7A1	1	1

Table 36 continued

TF	TG	W	# dup
E2F1	COQ6	1	1
E2F1	CTSD	1	2
E2F1	CYTH2	1	1
E2F1	DET1	1	1
E2F1	DHFR	1	1
E2F1	DTL	1	1
E2F1	E2F1	1	1
E2F1	E2F8	1	1
E2F1	EIF4E	1	1
E2F1	FABP1	1	1
E2F1	FAM60A	1	1
E2F1	FANCA	1	1
E2F1	FANCD2	1	2
E2F1	FEN1	1	1
E2F1	FLAD1	1	1
E2F1	FOXM1	1	1
E2F1	FXR1	1	1
E2F1	FYN	1	2
E2F1	G2E3	1	2
E2F1	GAS6	1	1
E2F1	GCLM	1	1
E2F1	GDF15	1	1
E2F1	GINS3	1	1
E2F1	GMNN	1	1
E2F1	GOT1	1	1
E2F1	GPSM2	1	2
E2F1	HELLS	1	2
E2F1	HERPUD2	1	1
E2F1	HIST1H2AC	1	1
E2F1	HIST1H4E	1	1
E2F1	HLA-DOA	1	2
E2F1	HRAS	1	2
E2F1	HRSP12	1	1
E2F1	HSPB8	1	1
E2F1	INSR	1	1
E2F1	ITPR1	1	2
E2F1	KATNA1	1	1
E2F1	KDM5B	1	1
E2F1	KIAA0586	1	1
E2F1	KIF14	1	1
E2F1	KIF20B	1	1
E2F1	KIF23	1	1
E2F1	KIF2C	1	1
E2F1	KIFC1	1	1
E2F1	KRAS	1	2

Table 36 continued

TF	TG	W	# dup
E2F1	LBR	1	1
E2F1	LPP	1	1
E2F1	LRIF1	1	1
E2F1	LRRC17	1	2
E2F1	MAD2L1	1	1
E2F1	MAN1A2	1	2
E2F1	MAP2K6	1	2
E2F1	MAPK13	1	1
E2F1	MCM8	1	1
E2F1	MDM2	1	1
E2F1	ME3	1	2
E2F1	MEGF9	1	1
E2F1	MELK	1	1
E2F1	MET	1	2
E2F1	MKI67	1	1
E2F1	MND1	1	1
E2F1	MNX1	1	2
E2F1	MRI1	1	1
E2F1	MRPS18B	1	1
E2F1	MSH2	1	1
E2F1	MZF1	1	1
E2F1	NCOA3	1	1
E2F1	NCS1	1	1
E2F1	NDE1	1	1
E2F1	NPAT	1	1
E2F1	NUDT4	1	1
E2F1	NUP160	1	2
E2F1	NUP37	1	1
E2F1	ODF2	1	1
E2F1	ORC3	1	1
E2F1	OSER1	1	1
E2F1	PBK	1	1
E2F1	PDGFA	1	1
E2F1	PKNOX1	1	2
E2F1	PLIN3	1	1
E2F1	PLK1	1	1
E2F1	POC1A	1	1
E2F1	POLA1	1	1
E2F1	POM121	1	1
E2F1	PPP1R2	1	1
E2F1	PPP3CA	1	1
E2F1	PRIM2	1	1
E2F1	PRKAR1A	1	2
E2F1	PSEN1	1	2
E2F1	PTTG1	1	1

Table 36 continued

TF	TG	W	# dup
E2F1	PWP1	1	1
E2F1	QRICH1	1	1
E2F1	RAD18	1	2
E2F1	RAD51	1	1
E2F1	RAD54L	1	1
E2F1	RBBP8	1	1
E2F1	REEP1	1	2
E2F1	RFC2	1	1
E2F1	RFC4	1	1
E2F1	RGS3	1	1
E2F1	RPA2	1	1
E2F1	RRM1	1	2
E2F1	RRM2	1	2
E2F1	RUNX1	1	1
E2F1	SAP30BP	1	1
E2F1	SEPHS1	1	1
E2F1	SGK1	1	1
E2F1	SLBP	1	1
E2F1	SLC44A2	1	1
E2F1	SP1	1	1
E2F1	SRD5A1	1	1
E2F1	SRSF5	1	1
E2F1	STAT5B	1	1
E2F1	STIL	1	1
E2F1	SUCLG2	1	1
E2F1	SYNCRIP	1	1
E2F1	TACC3	1	1
E2F1	TGIF1	1	1
E2F1	THRAP3	1	1
E2F1	TIMP1	1	1
E2F1	TMEM132A	1	1
E2F1	TOMM70A	1	1
E2F1	TOP1	1	1
E2F1	TOP2A	1	1
E2F1	TOP3A	1	1
E2F1	TOPBP1	1	2
E2F1	TRA2A	1	1
E2F1	TRIM45	1	1
E2F1	TRIP13	1	2
E2F1	TROAP	1	1
E2F1	TSG101	1	2
E2F1	TUBB2A	1	1
E2F1	TULP4	1	1
E2F1	TYMS	1	2
E2F1	UACA	1	2

Table 36 continued

TF	TG	W	# dup
E2F1	UBE2S	1	1
E2F1	UBE2T	1	1
E2F1	UBL3	1	1
E2F1	UBQLN2	1	1
E2F1	UHRF1	1	1
E2F1	USP1	1	1
E2F1	VCAM1	1	1
E2F1	VEGFC	1	1
E2F1	VPS37C	1	1
E2F1	WSB1	1	1
E2F1	YY1	1	1
E2F1	ZBED5	1	1
E2F1	ZBTB7A	1	1
E2F1	ZC3HC1	1	1
E2F1	ZMYM1	1	1
E2F1	ZNF143	1	1
E2F1	ZNF521	1	1
E2F1	ZSCAN5A	1	1
E2F1	ZWINT	1	1
E2F5	ASF1B	1	1
E2F5	BRCA1	1	1
E2F8	E2F1	1	1
FOXM1	ARHGAP8	1	1
FOXM1	AURKB	1	2
FOXM1	BIRC5	1	1
FOXM1	CCNA2	1	1
FOXM1	CCNB1	1	1
FOXM1	CDC25A	1	1
FOXM1	CDC6	1	1
FOXM1	CDKN1B	1	1
FOXM1	CKS1B	1	1
FOXM1	MID1	1	1
FOXM1	PDGFA	1	1
FOXM1	PLK1	1	1
HIF1A	ADAMTS1	1	2
HIF1A	ARL4A	1	1
HIF1A	C6	1	1
HIF1A	CDK7	1	1
HIF1A	CDKN1B	1	1
HIF1A	DNAJB9	1	1
HIF1A	DYNLL1	1	1
HIF1A	FANCD2	1	1
HIF1A	FOXM1	1	1
HIF1A	FRZB	1	1
HIF1A	GRPEL1	1	1

Table 36 continued

TF	TG	W	# dup
HIF1A	HERPUD2	1	1
HIF1A	HIF1A	1	1
HIF1A	HMMR	1	1
HIF1A	INSIG2	1	1
HIF1A	KDM5B	1	1
HIF1A	MET	1	2
HIF1A	MUC1	1	1
HIF1A	NLRP2	1	1
HIF1A	NR3C1	1	1
HIF1A	PCF11	1	1
HIF1A	PDXP	1	1
HIF1A	PLIN3	1	1
HIF1A	POM121	1	1
HIF1A	PPP6R3	1	1
HIF1A	PRPSAP1	1	1
HIF1A	RBM8A	1	1
HIF1A	RHOBTB3	1	1
HIF1A	RRM2	1	1
HIF1A	SAP30	1	1
HIF1A	STIL	1	1
HIF1A	TFF3	1	1
HIF1A	TIMP1	1	1
HIF1A	TOMM34	1	1
HIF1A	TOP3A	1	1
HIF1A	TYMS	1	1
HIF1A	VCAM1	1	1
HIF1A	VEGFC	1	2
HIF1A	WSB1	1	1
HIF1A	ZNF217	1	1
HOXB4	NIPBL	1	1
HOXB4	PSEN1	1	1
HOXB4	SP1	1	1
HOXB4	SRF	1	1
HOXB4	TFAP2A	1	1
HOXB4	YY1	1	1
HSF2	HIF1A	1	1
INSM1	INSM1	1	1
KDM5B	BRCA1	1	1
KLF6	PTTG1	1	1
KLF9	TFAP2A	1	2
MITF	ABCC2	1	1
MITF	ACD	1	1
MITF	AFAP1	1	1
MITF	AHI1	1	2
MITF	AMD1	1	1

Table 36 continued

TF	TG	W	# dup
MITF	ANKRD10	1	1
MITF	ANP32B	1	1
MITF	ANTXR1	1	1
MITF	AP3D1	1	1
MITF	AP3M2	1	1
MITF	ARHGEF39	1	1
MITF	ARL4A	1	1
MITF	ASF1B	1	2
MITF	ASIP	1	1
MITF	ATF7IP	1	1
MITF	ATL2	1	1
MITF	BAG3	1	2
MITF	BMP2	1	1
MITF	BRCA1	1	1
MITF	BTBD3	1	1
MITF	BUB3	1	1
MITF	C6	1	2
MITF	CADM1	1	1
MITF	CBX3	1	1
MITF	CCNB1	1	1
MITF	CCNE1	1	1
MITF	CDC16	1	2
MITF	CDC25B	1	2
MITF	CDC42	1	1
MITF	CDC7	1	1
MITF	CDKN1B	1	1
MITF	CDKN2AIP	1	1
MITF	CDKN2C	1	1
MITF	CENPA	1	1
MITF	CENPM	1	1
MITF	CFLAR	1	1
MITF	CHEK2	1	1
MITF	CIC	1	1
MITF	CIT	1	1
MITF	CKS2	1	1
MITF	CNOT10	1	1
MITF	CSGALNACT1	1	1
MITF	CTNND1	1	2
MITF	CYB5R2	1	1
MITF	DDX11	1	1
MITF	DEXI	1	1
MITF	DKC1	1	1
MITF	DMXL2	1	1
MITF	DNAJB1	1	2
MITF	DNAJB4	1	1

Table 36 continued

TF	TG	W	# dup
MITF	DNAJB6	1	1
MITF	DNAJB9	1	2
MITF	DR1	1	2
MITF	DSP	1	2
MITF	DUSP4	1	1
MITF	DYNLL1	1	1
MITF	E2F5	1	1
MITF	E2F8	1	1
MITF	FADD	1	1
MITF	FAM189B	1	1
MITF	FAM60A	1	1
MITF	FAM64A	1	1
MITF	FANCA	1	1
MITF	FEM1B	1	1
MITF	FEN1	1	1
MITF	FKBP1A	1	1
MITF	FRZB	1	1
MITF	FZR1	1	2
MITF	GAS1	1	1
MITF	GAS6	1	1
MITF	GNB1	1	2
MITF	GTF2B	1	1
MITF	HAUS5	1	2
MITF	HAUS8	1	1
MITF	HERPUD2	1	1
MITF	HIF1A	1	1
MITF	HIST2H2BE	1	1
MITF	HOXB4	1	2
MITF	HP1BP3	1	1
MITF	HRAS	1	1
MITF	HSF2	1	1
MITF	HSPA8	1	1
MITF	IDI2	1	1
MITF	INADL	1	1
MITF	ITPR3	1	1
MITF	IVNS1ABP	1	1
MITF	JADE2	1	2
MITF	KANK2	1	1
MITF	KAT2B	1	1
MITF	KBTBD2	1	1
MITF	KDELC1	1	1
MITF	KDM4A	1	1
MITF	KDM5B	1	1
MITF	KIFC1	1	1
MITF	KLF6	1	1

Table 36 continued

TF	TG	W	# dup
MITF	KLF9	1	2
MITF	KPNA2	1	2
MITF	LBR	1	1
MITF	MAN1A2	1	2
MITF	MAPK13	1	1
MITF	MBD2	1	1
MITF	MBD3	1	2
MITF	MCAM	1	1
MITF	MCM8	1	1
MITF	ME3	1	2
MITF	MIS18BP1	1	1
MITF	MNT	1	1
MITF	MNX1	1	2
MITF	MORF4L2	1	1
MITF	MSH2	1	1
MITF	MTCL1	1	1
MITF	MZF1	1	1
MITF	NAB1	1	2
MITF	NCAPH	1	2
MITF	NCOA3	1	1
MITF	NCOA5	1	1
MITF	NDE1	1	2
MITF	NFE2L2	1	2
MITF	NFIC	1	1
MITF	NPM1	1	1
MITF	NSUN3	1	1
MITF	OGT	1	1
MITF	OSER1	1	1
MITF	PAK11P1	1	1
MITF	PANK2	1	1
MITF	PCF11	1	1
MITF	PDGFA	1	1
MITF	PDXP	1	1
MITF	PIK3CD	1	2
MITF	PKNOX1	1	1
MITF	PLIN3	1	1
MITF	PLK1	1	1
MITF	POC1A	1	1
MITF	POLA1	1	1
MITF	PPP1R10	1	1
MITF	PRIM2	1	1
MITF	PRKAR1A	1	2
MITF	PRR16	1	1
MITF	PSEN1	1	1
MITF	PTP4A1	1	1

Table 36 continued

TF	TG	W	# dup
MITF	PTTG1	1	1
MITF	PWP1	1	1
MITF	QRICH1	1	1
MITF	RAB3A	1	1
MITF	RAN	1	1
MITF	RCCD1	1	1
MITF	RHEB	1	1
MITF	RMI1	1	1
MITF	RRM2	1	2
MITF	RRP1	1	1
MITF	RUNX1	1	1
MITF	SAP30	1	1
MITF	SAP30BP	1	1
MITF	SGK1	1	1
MITF	SLC25A36	1	1
MITF	SLC38A2	1	2
MITF	SMARCB1	1	1
MITF	SMTN	1	1
MITF	SP1	1	2
MITF	SRSF3	1	1
MITF	SS18	1	1
MITF	SSR3	1	1
MITF	STAG1	1	1
MITF	STAT1	1	1
MITF	SV2B	1	1
MITF	SYNCRIP	1	1
MITF	TAB2	1	1
MITF	TACC3	1	2
MITF	TFAP2A	1	1
MITF	TGIF1	1	2
MITF	TOB2	1	2
MITF	TOMM34	1	1
MITF	TOP1	1	1
MITF	TOP3A	1	1
MITF	TRAIP	1	1
MITF	TRIP13	1	2
MITF	TSC22D1	1	1
MITF	TSG101	1	2
MITF	TSKU	1	1
MITF	TSN	1	1
MITF	TTC38	1	1
MITF	TUBB2A	1	1
MITF	TUBB4B	1	1
MITF	TULP4	1	1
MITF	TXNRD1	1	2

Table 36 continued

TF	TG	W	# dup
MITF	UACA	1	2
MITF	UBE2D3	1	2
MITF	UBL3	1	1
MITF	UHRF1	1	1
MITF	UNG	1	2
MITF	USP1	1	1
MITF	USP13	1	1
MITF	VEGFC	1	1
MITF	VPS37C	1	1
MITF	WSB1	1	1
MITF	YWHAH	1	1
MITF	YY1	1	2
MITF	ZBED5	1	1
MITF	ZC3HC1	1	1
MITF	ZCCHC10	1	1
MITF	ZNF217	1	1
MITF	ZNFX1	1	1
MITF	ZNHIT2	1	1
MNX1	CDC42	1	1
MNX1	FYN	1	1
MNX1	GAS6	1	1
MNX1	INSR	1	1
MNX1	KATNA1	1	1
MNX1	MYCBP2	1	1
MNX1	NDE1	1	1
MNX1	PSEN1	1	1
MNX1	RAB23	1	1
MNX1	TGIF1	1	1
NCOA3	BRCA1	1	1
NFE2L2	BRCA1	1	1
NFIA	NR3C1	1	1
NFIC	HRAS	1	1
NFIC	INSR	1	1
NFIC	NR3C1	1	1
NFIC	TFAP2A	1	1
NFYA	CDC25A	1	1
NFYA	CDCA8	1	1
NFYA	CDKN1B	1	1
NFYA	E2F1	1	1
NFYA	GADD45A	1	1
NFYA	HOXB4	1	1
NFYA	MCM8	1	1
NFYA	PTTG1	1	1
NFYB	CDKN1B	1	1
NFYB	HLA-DOA	1	1

Table 36 continued

TF	TG	W	# dup
NFYB	HLA-DRA	1	1
NFYB	HSPA13	1	1
NR3C1	BRCA1	1	1
NR3C1	NR3C1	1	1
NR3C1	SRF	1	1
NR3C1	STAT1	1	1
PKNOX1	C6	1	1
PKNOX1	GAS1	1	1
PKNOX1	HLA-DOA	1	1
PKNOX1	MITF	1	1
RUNX1	ADAMTS1	1	2
RUNX1	BBS2	1	1
RUNX1	BCLAF1	1	1
RUNX1	BIRC2	1	1
RUNX1	C5orf42	1	1
RUNX1	CDC25B	1	1
RUNX1	CENPF	1	1
RUNX1	CENPL	1	1
RUNX1	CEP70	1	1
RUNX1	CKAP2	1	1
RUNX1	CKS2	1	1
RUNX1	CTR9	1	1
RUNX1	CXCL14	1	1
RUNX1	DEPDC1B	1	1
RUNX1	DNA2	1	1
RUNX1	DNAJC3	1	1
RUNX1	EIF4E	1	1
RUNX1	FAM105A	1	1
RUNX1	FRZB	1	1
RUNX1	FXR1	1	1
RUNX1	GPSM2	1	2
RUNX1	HIST1H2BC	1	1
RUNX1	HSF2	1	1
RUNX1	INADL	1	1
RUNX1	IVNS1ABP	1	1
RUNX1	KLF6	1	1
RUNX1	KPNA2	1	1
RUNX1	LARP7	1	1
RUNX1	LRRC17	1	1
RUNX1	MAD2L1	1	1
RUNX1	MAN1A2	1	1
RUNX1	MAP2K6	1	1
RUNX1	ME3	1	2
RUNX1	MET	1	1
RUNX1	MITF	1	1

Table 36 continued

TF	TG	W	# dup
RUNX1	MKI67	1	1
RUNX1	MND1	1	1
RUNX1	MTCL1	1	1
RUNX1	NCOA3	1	1
RUNX1	NEIL3	1	2
RUNX1	NSUN3	1	1
RUNX1	NUP98	1	1
RUNX1	ORC3	1	1
RUNX1	PIK3CD	1	2
RUNX1	PPP6R3	1	1
RUNX1	PRIM2	1	1
RUNX1	PTP4A1	1	1
RUNX1	ROCK1	1	1
RUNX1	SGK1	1	1
RUNX1	SLC25A27	1	1
RUNX1	SLC38A2	1	1
RUNX1	SLC39A10	1	1
RUNX1	SPAG5	1	1
RUNX1	STAG1	1	1
RUNX1	SUCLG2	1	1
RUNX1	TRIP13	1	2
RUNX1	TSKU	1	1
RUNX1	UACA	1	1
RUNX1	VCL	1	1
RUNX1	WSB1	1	1
RUNX1	ZPBP	1	1
RUNX1	ZRANB2	1	1
SP1	BIRC5	1	1
SP1	BRCA1	1	1
SP1	BUB1B	1	2
SP1	C4B	1	1
SP1	CASP3	1	1
SP1	CCNA2	1	1
SP1	CCNB1	1	1
SP1	CDC25A	1	2
SP1	CDC25C	1	1
SP1	CDKN1B	1	1
SP1	CDKN2C	1	1
SP1	CDKN2D	1	1
SP1	COL7A1	1	1
SP1	CTSD	1	1
SP1	CXCL14	1	1
SP1	DHFR	1	1
SP1	DKC1	1	1
SP1	E2F1	1	1

Table 36 continued

TF	TG	W	# dup
SP1	EXO1	1	1
SP1	FOXM1	1	1
SP1	FRZB	1	1
SP1	G2E3	1	1
SP1	GAS1	1	1
SP1	GPSM2	1	1
SP1	HIF1A	1	1
SP1	HSD17B11	1	2
SP1	HSPA8	1	1
SP1	ITGB3	1	1
SP1	KIF2C	1	1
SP1	LMO4	1	1
SP1	MCAM	1	1
SP1	MDM2	1	1
SP1	MET	1	1
SP1	NR3C1	1	1
SP1	PDGFA	1	1
SP1	POLA1	1	1
SP1	PSEN1	1	1
SP1	PTTG1	1	1
SP1	RECQL4	1	1
SP1	SP1	1	1
SP1	TFAP2A	1	1
SP1	TIMP1	1	1
SP1	TMPO	1	1
SP1	TYMS	1	1
SP1	UNG	1	1
SP1	VCAM1	1	1
SRF	HOXB4	1	1
SRF	KPNB1	1	1
SRF	STIL	1	1
SRF	UBE2S	1	1
STAT1	ABCA7	1	1
STAT1	ABCC2	1	1
STAT1	ADAMTS1	1	2
STAT1	ADCY6	1	1
STAT1	AFAP1	1	1
STAT1	AGFG1	1	1
STAT1	AHI1	1	1
STAT1	AKIRIN2	1	1
STAT1	ANKRD10	1	1
STAT1	ANP32B	1	1
STAT1	ANP32E	1	1
STAT1	ANTXR1	1	1
STAT1	AP3M2	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	ARHGAP11A	1	1
STAT1	ARHGAP19	1	1
STAT1	ARHGDIB	1	1
STAT1	ARHGEF39	1	1
STAT1	ARL4A	1	1
STAT1	ARL6IP1	1	1
STAT1	ARMC1	1	1
STAT1	ASF1B	1	2
STAT1	ATAD2	1	1
STAT1	ATF7IP	1	1
STAT1	B2M	1	1
STAT1	BAG3	1	1
STAT1	BARD1	1	1
STAT1	BCLAF1	1	1
STAT1	BIRC2	1	1
STAT1	BMP2	1	1
STAT1	BRCA1	1	1
STAT1	BRD7	1	1
STAT1	BTBD3	1	1
STAT1	BUB3	1	1
STAT1	C5orf42	1	1
STAT1	C6	1	2
STAT1	CADM1	1	1
STAT1	CASP3	1	1
STAT1	CBX3	1	1
STAT1	CCDC90B	1	1
STAT1	CCNA2	1	1
STAT1	CCNE1	1	2
STAT1	CDC16	1	1
STAT1	CDC20	1	1
STAT1	CDC25B	1	1
STAT1	CDC25C	1	1
STAT1	CDC27	1	1
STAT1	CDC42EP1	1	1
STAT1	CDC42EP4	1	1
STAT1	CDC45	1	1
STAT1	CDCA7	1	1
STAT1	CDCA7L	1	1
STAT1	CDKN1B	1	1
STAT1	CDKN2AIP	1	1
STAT1	CDKN2C	1	1
STAT1	CDR2	1	1
STAT1	CENPA	1	2
STAT1	CENPE	1	1
STAT1	CENPM	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	CEP44	1	1
STAT1	CFD	1	1
STAT1	CHAF1A	1	1
STAT1	CHEK2	1	1
STAT1	CIC	1	1
STAT1	CIT	1	1
STAT1	CKS2	1	1
STAT1	CLSPN	1	1
STAT1	CNIH4	1	1
STAT1	CNOT10	1	1
STAT1	CREBZF	1	1
STAT1	CRK	1	1
STAT1	CRYBA1	1	1
STAT1	CSGALNACT1	1	1
STAT1	CSH2	1	1
STAT1	CTCF	1	1
STAT1	CTNND1	1	1
STAT1	CTR9	1	1
STAT1	CTSD	1	1
STAT1	CYTH2	1	1
STAT1	CYTH3	1	1
STAT1	DCTN6	1	1
STAT1	DEPDC1B	1	1
STAT1	DHFR	1	1
STAT1	DHX8	1	1
STAT1	DIS3	1	1
STAT1	DLGAP5	1	1
STAT1	DNAJB1	1	1
STAT1	DNAJB6	1	1
STAT1	DNAJB9	1	1
STAT1	DNAJC3	1	1
STAT1	DNAJC6	1	1
STAT1	DR1	1	1
STAT1	DSCC1	1	1
STAT1	DTL	1	1
STAT1	DUSP4	1	1
STAT1	DYNLL1	1	1
STAT1	DZIP3	1	2
STAT1	E2F1	1	1
STAT1	E2F8	1	1
STAT1	ELP3	1	1
STAT1	ERN2	1	1
STAT1	ESPL1	1	1
STAT1	FADD	1	1
STAT1	FAM105A	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	FAM110A	1	1
STAT1	FAM214A	1	1
STAT1	FAM60A	1	1
STAT1	FAM83D	1	1
STAT1	FANCA	1	1
STAT1	FANCG	1	1
STAT1	FANCI	1	1
STAT1	FEM1B	1	1
STAT1	FEN1	1	1
STAT1	FKBP1A	1	1
STAT1	FLAD1	1	1
STAT1	G2E3	1	2
STAT1	G3BP1	1	1
STAT1	GADD45A	1	1
STAT1	GCSH	1	1
STAT1	GINS3	1	1
STAT1	GMNN	1	1
STAT1	GOT1	1	1
STAT1	GRK6	1	1
STAT1	GTF2B	1	1
STAT1	H1F0	1	1
STAT1	HCP5	1	1
STAT1	HERPUD2	1	1
STAT1	HIF1A	1	1
STAT1	HIST1H2AC	1	1
STAT1	HIST1H4H	1	1
STAT1	HIST2H2BE	1	1
STAT1	HMGCR	1	1
STAT1	HMMR	1	1
STAT1	HN1	1	1
STAT1	HP1BP3	1	1
STAT1	HRAS	1	1
STAT1	HSD17B11	1	2
STAT1	HSF2	1	1
STAT1	HSPA1L	1	1
STAT1	HSPA8	1	1
STAT1	HSPB8	1	1
STAT1	IDO1	1	1
STAT1	IFIT1	1	2
STAT1	IL18BP	1	1
STAT1	ILF2	1	1
STAT1	INADL	1	1
STAT1	INPP5K	1	1
STAT1	INSIG2	1	1
STAT1	INSR	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	INTS7	1	1
STAT1	ITPR3	1	1
STAT1	IVNS1ABP	1	1
STAT1	KAT2B	1	1
STAT1	KATNBL1	1	1
STAT1	KBTBD2	1	1
STAT1	KCTD2	1	1
STAT1	KDM4A	1	1
STAT1	KDM5B	1	1
STAT1	KIAA0101	1	1
STAT1	KIAA1524	1	1
STAT1	KIF11	1	1
STAT1	KIF14	1	1
STAT1	KIF20B	1	1
STAT1	KIF22	1	1
STAT1	KIF5B	1	1
STAT1	KPNA2	1	1
STAT1	KRAS	1	1
STAT1	LBR	1	1
STAT1	LMNB1	1	1
STAT1	LMO4	1	1
STAT1	LRIF1	1	1
STAT1	LRRC17	1	2
STAT1	MAN1A2	1	1
STAT1	MAP2K6	1	1
STAT1	MATN2	1	1
STAT1	MBD2	1	1
STAT1	MCM2	1	1
STAT1	MCM4	1	1
STAT1	MDC1	1	1
STAT1	MDM2	1	1
STAT1	ME3	1	2
STAT1	MED31	1	1
STAT1	MEGF9	1	1
STAT1	MELK	1	1
STAT1	MET	1	2
STAT1	MID1	1	1
STAT1	MKI67	1	1
STAT1	MND1	1	1
STAT1	MNX1	1	2
STAT1	MORF4L2	1	1
STAT1	MTCL1	1	1
STAT1	MZF1	1	1
STAT1	NAB1	1	1
STAT1	NASP	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	NCAPH	1	2
STAT1	NCOA3	1	1
STAT1	NCOA5	1	1
STAT1	NCS1	1	1
STAT1	NDC80	1	1
STAT1	NDE1	1	1
STAT1	NEIL3	1	2
STAT1	NEK2	1	2
STAT1	NFIC	1	1
STAT1	NFYB	1	1
STAT1	NIPBL	1	1
STAT1	NKTR	1	1
STAT1	NNMT	1	1
STAT1	NSUN3	1	1
STAT1	NUCKS1	1	1
STAT1	NUF2	1	1
STAT1	NUP160	1	1
STAT1	NUP98	1	1
STAT1	OGT	1	1
STAT1	OLR1	1	1
STAT1	OSER1	1	1
STAT1	OSGIN2	1	1
STAT1	OXR1	1	1
STAT1	PAK1IP1	1	1
STAT1	PBK	1	1
STAT1	PDGFA	1	1
STAT1	PIK3CD	1	2
STAT1	PKNOX1	1	1
STAT1	PLIN3	1	1
STAT1	PLK2	1	1
STAT1	POC1A	1	1
STAT1	POLD3	1	1
STAT1	POLQ	1	1
STAT1	POM121	1	1
STAT1	PPP1R2	1	1
STAT1	PPP3CA	1	1
STAT1	PPP6R3	1	1
STAT1	PRC1	1	1
STAT1	PRIM2	1	1
STAT1	PRPSAP1	1	1
STAT1	PRR11	1	1
STAT1	PRR16	1	1
STAT1	PSEN1	1	1
STAT1	PSMG3	1	1
STAT1	PTP4A1	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	PTTG1	1	1
STAT1	RAB23	1	1
STAT1	RAD18	1	2
STAT1	RAD21	1	1
STAT1	RAD51	1	1
STAT1	RAD51AP1	1	1
STAT1	RAD51C	1	1
STAT1	RAD54L	1	1
STAT1	RANGAP1	1	1
STAT1	RBBP8	1	1
STAT1	RCAN1	1	1
STAT1	RCCD1	1	1
STAT1	REEP1	1	2
STAT1	RFC2	1	1
STAT1	RFC4	1	1
STAT1	RGS3	1	1
STAT1	RHEB	1	1
STAT1	RHNO1	1	1
STAT1	RHOBTB3	1	1
STAT1	RMI1	1	1
STAT1	RNPC3	1	1
STAT1	RNPS1	1	1
STAT1	RRM2	1	1
STAT1	RRP1	1	1
STAT1	RSRC2	1	1
STAT1	SAP30	1	1
STAT1	SAP30BP	1	1
STAT1	SDC1	1	2
STAT1	SEPHS1	1	1
STAT1	SERPINB3	1	1
STAT1	SFPQ	1	1
STAT1	SH3GL2	1	2
STAT1	SHC1	1	1
STAT1	SLC22A3	1	2
STAT1	SLC25A36	1	1
STAT1	SLC38A2	1	1
STAT1	SLC39A10	1	1
STAT1	SLC4A1AP	1	1
STAT1	SMARCB1	1	1
STAT1	SMARCD1	1	1
STAT1	SNUPN	1	1
STAT1	SP1	1	1
STAT1	SPAG5	1	1
STAT1	SRD5A1	1	1
STAT1	SRF	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	SRSF3	1	1
STAT1	SS18	1	1
STAT1	STAG3	1	1
STAT1	STAT5B	1	1
STAT1	STIL	1	2
STAT1	SUCLG2	1	1
STAT1	SV2B	1	1
STAT1	SYNCRIP	1	1
STAT1	TAB2	1	1
STAT1	TACC3	1	1
STAT1	TFAP2A	1	1
STAT1	TFF3	1	1
STAT1	TGIF1	1	1
STAT1	THRAP3	1	1
STAT1	TIMP1	1	1
STAT1	TIPIN	1	1
STAT1	TMPO	1	1
STAT1	TOB2	1	1
STAT1	TOMM34	1	1
STAT1	TOP1	1	1
STAT1	TOP2A	1	1
STAT1	TOP3A	1	1
STAT1	TPX2	1	1
STAT1	TRA2A	1	1
STAT1	TRIP13	1	2
STAT1	TSG101	1	1
STAT1	TSKU	1	1
STAT1	TSN	1	1
STAT1	TTC31	1	1
STAT1	TTF2	1	1
STAT1	TUBA1A	1	1
STAT1	TUBB2A	1	1
STAT1	TULP4	1	1
STAT1	TXNRD1	1	1
STAT1	UACA	1	1
STAT1	UBE2D3	1	1
STAT1	UBL3	1	1
STAT1	UBQLN2	1	1
STAT1	UHRF1	1	1
STAT1	USP1	1	1
STAT1	USP6NL	1	1
STAT1	VCAM1	1	1
STAT1	VEGFC	1	1
STAT1	VPS25	1	1
STAT1	VPS72	1	1

Table 36 continued

TF	TG	W	# dup
STAT1	VTA1	1	1
STAT1	WSB1	1	1
STAT1	YY1	1	1
STAT1	ZBTB7A	1	1
STAT1	ZC3HC1	1	1
STAT1	ZCCHC10	1	1
STAT1	ZNF143	1	1
STAT1	ZNF207	1	1
STAT1	ZNF217	1	1
STAT1	ZNF281	1	1
STAT1	ZNF414	1	1
STAT1	ZNF521	1	1
STAT1	ZNFX1	1	1
STAT1	ZNHIT2	1	1
STAT1	ZPBP	1	1
STAT1	ZRANB2	1	1
STAT1	ZSCAN5A	1	1
STAT1	ZWINT	1	1
STAT5B	ADAMTS1	1	1
STAT5B	ERN2	1	1
STAT5B	HLA-DOA	1	1
STAT5B	ITGB3	1	1
STAT5B	MET	1	2
STAT5B	MUC1	1	1
STAT5B	NLRP2	1	1
STAT5B	RAD51	1	1
STAT5B	VCAM1	1	1
TFAP2A	CASP3	1	1
TFAP2A	CCNB1	1	1
TFAP2A	CDC42	1	1
TFAP2A	CDKN1B	1	1
TFAP2A	CTNND1	1	1
TFAP2A	CTSD	1	2
TFAP2A	DHX8	1	1
TFAP2A	DSP	1	1
TFAP2A	FEM1B	1	1
TFAP2A	FRZB	1	1
TFAP2A	HSPA8	1	1
TFAP2A	INSIG2	1	1
TFAP2A	MCAM	1	1
TFAP2A	NIPBL	1	1
TFAP2A	NR3C1	1	1
TFAP2A	PDGFA	1	1
TFAP2A	PSEN1	1	1
TFAP2A	RECQL4	1	1

Table 36 continued

TF	TG	W	# dup
TFAP2A	SP1	1	1
TFAP2A	TFAP2A	1	1
TFAP2A	TIMP1	1	1
TFAP2A	TOP1	1	1
TFAP2A	TSG101	1	1
TFAP2A	VEGFC	1	1
TGIF1	MNX1	1	1
YY1	BMP2	1	1
YY1	BRCA1	1	1
YY1	CDC25A	1	1
YY1	CDC6	1	1
YY1	DKC1	1	1
YY1	DNAJB4	1	1
YY1	DTL	1	1
YY1	FAN1	1	1
YY1	GAS1	1	1
YY1	HDAC3	1	1
YY1	HIF1A	1	1
YY1	HLA-DRA	1	1
YY1	HOXB4	1	1
YY1	MCM5	1	1
YY1	NUP160	1	1
YY1	PCNA	1	1
YY1	RAD51	1	1
YY1	SAP30	1	1
YY1	TFAP2A	1	1
ZNF143	BUB1B	1	1

The table gives the list of positive edges in our gold-standard network. The 1st column represents the TF. The 2nd column the TG. The 3rd column informs for each edge if it is present in the network (value of 1) or if it is absent (value of 0). The present edges are the positive links, and the absent edges are the negative links. For each edge, the number in the 4th column provides the number of times it was repeated before removing the duplicate edges from the network obtained by combining Alonso networks and **HumanBase** networks.

Table 37: HeLa “gold-standard” network-
Negative links

TF	TG	W	# dup
BRCA1	ADAMTS1	0	1
BRCA1	ASF1B	0	1
BRCA1	AURKB	0	1
BRCA1	BUB1B	0	1
BRCA1	C6	0	1
BRCA1	CCNE1	0	1
BRCA1	CDH24	0	1
BRCA1	CLSPN	0	1
BRCA1	GPSM2	0	1
BRCA1	HLA-DOA	0	1
BRCA1	HOXB4	0	1
BRCA1	INSM1	0	1
BRCA1	LRRC17	0	1
BRCA1	MET	0	1
BRCA1	MNX1	0	1
BRCA1	NEIL3	0	1
BRCA1	PIK3CD	0	1
BRCA1	RAD18	0	1
BRCA1	REEP1	0	1
BRCA1	SDC1	0	1
BRCA1	SH3GL2	0	1
BRCA1	SLC22A3	0	1
BRCA1	STIL	0	1
BRCA1	UBE2C	0	1
CENPA	ADAMTS1	0	1
CENPA	ADH4	0	1
CENPA	ASF1B	0	1
CENPA	C6	0	1
CENPA	CCNB2	0	1
CENPA	CCNE2	0	1
CENPA	CDC25B	0	1
CENPA	CDH24	0	1
CENPA	CDK7	0	1
CENPA	CENPE	0	1
CENPA	CSGALNACT1	0	1
CENPA	CTSD	0	1
CENPA	DMXL2	0	1
CENPA	DNAJB1	0	1
CENPA	DZIP3	0	1
CENPA	ELP3	0	1

Table 37 continued

TF	TG	W	# dup
CENPA	FAN1	0	1
CENPA	FANCG	0	1
CENPA	FEN1	0	1
CENPA	FRZB	0	1
CENPA	G2E3	0	1
CENPA	GOT1	0	1
CENPA	HIST1H4B	0	1
CENPA	HIST1H4E	0	1
CENPA	HIST1H4H	0	1
CENPA	HLA-DOA	0	1
CENPA	HSD17B11	0	1
CENPA	IL18BP	0	1
CENPA	INSM1	0	1
CENPA	ITPR3	0	1
CENPA	KIF20B	0	1
CENPA	KMO	0	1
CENPA	MBD4	0	1
CENPA	MCM2	0	1
CENPA	MCM6	0	1
CENPA	ME3	0	1
CENPA	MGAT2	0	1
CENPA	NEIL3	0	1
CENPA	NUP160	0	1
CENPA	PCNA	0	1
CENPA	PIK3CD	0	1
CENPA	POLD3	0	1
CENPA	PRKAR1A	0	1
CENPA	PTPN9	0	1
CENPA	PYM1	0	1
CENPA	RAD18	0	1
CENPA	RAD51C	0	1
CENPA	RBM8A	0	1
CENPA	RERE	0	1
CENPA	RHEB	0	1
CENPA	RNPS1	0	1
CENPA	SFPQ	0	1
CENPA	SH3GL2	0	1
CENPA	SLC22A3	0	1
CENPA	SS18	0	1
CENPA	SYNCRIP	0	1
CENPA	TOPBP1	0	1
CENPA	TRA2A	0	1
CENPA	TYMS	0	1
CENPA	UNG	0	1
CENPA	VCAM1	0	1

Table 37 continued

TF	TG	W	# dup
CREBZF	VEGFC	0	1
CREBZF	ADAMTS1	0	1
CREBZF	ADH4	0	1
CREBZF	ASF1B	0	1
CREBZF	AURKB	0	1
CREBZF	BUB1B	0	1
CREBZF	C6	0	1
CREBZF	CCNE1	0	1
CREBZF	CDH24	0	1
CREBZF	CENPA	0	1
CREBZF	G2E3	0	1
CREBZF	GPSM2	0	1
CREBZF	HELLS	0	1
CREBZF	HLA-DOA	0	1
CREBZF	HOXB4	0	1
CREBZF	HSD17B11	0	1
CREBZF	IFIT1	0	1
CREBZF	INSM1	0	1
CREBZF	LRRC17	0	1
CREBZF	ME3	0	1
CREBZF	MET	0	1
CREBZF	MNX1	0	1
CREBZF	NCAPH	0	1
CREBZF	NEIL3	0	1
CREBZF	NEK2	0	1
CREBZF	PIK3CD	0	1
CREBZF	RAD18	0	1
CREBZF	REEP1	0	1
CREBZF	SDC1	0	1
CREBZF	SH3GL2	0	1
CREBZF	STIL	0	1
CREBZF	TRIP13	0	1
CTCF	ADAMTS1	0	1
CTCF	BUB1B	0	1
CTCF	CDH24	0	1
CTCF	FANCD2	0	1
CTCF	HOXB4	0	1
CTCF	IL18BP	0	1
CTCF	LRRC17	0	1
DR1	ADAMTS1	0	1
DR1	ADH4	0	1
DR1	ARHGAP8	0	1
DR1	ASF1B	0	1
DR1	AURKB	0	1
DR1	BIRC5	0	1

Table 37 continued

TF	TG	W	# dup
DR1	BORA	0	1
DR1	BUB1B	0	1
DR1	C6	0	1
DR1	CCNA2	0	1
DR1	CCNE1	0	1
DR1	CDC25A	0	1
DR1	CDH24	0	1
DR1	CENPF	0	1
DR1	CHEK2	0	1
DR1	CSGALNACT1	0	1
DR1	DMXL2	0	1
DR1	E2F1	0	1
DR1	FAN1	0	1
DR1	FANCD2	0	1
DR1	FRZB	0	1
DR1	G2E3	0	1
DR1	GAS1	0	1
DR1	GPSM2	0	1
DR1	HAUS5	0	1
DR1	HELLS	0	1
DR1	HLA-DOA	0	1
DR1	HOXB4	0	1
DR1	HSD17B11	0	1
DR1	IFIT1	0	1
DR1	IL18BP	0	1
DR1	INSM1	0	1
DR1	ITPR3	0	1
DR1	KDM4A	0	1
DR1	KIF20B	0	1
DR1	KMO	0	1
DR1	LRRC17	0	1
DR1	ME3	0	1
DR1	MET	0	1
DR1	MID1	0	1
DR1	MITF	0	1
DR1	MNX1	0	1
DR1	NEIL3	0	1
DR1	NLRP2	0	1
DR1	NPAT	0	1
DR1	PIK3CD	0	1
DR1	POLQ	0	1
DR1	RAB3A	0	1
DR1	RAD51	0	1
DR1	RECQL4	0	1
DR1	REEP1	0	1

Table 37 continued

TF	TG	W	# dup
DR1	SDC1	0	1
DR1	SH3GL2	0	1
DR1	SLC22A3	0	1
DR1	SRSF7	0	1
DR1	STIL	0	1
DR1	TFAP2A	0	1
DR1	TRIP13	0	1
DR1	VCAM1	0	1
DR1	VEGFC	0	1
E2F1	ABCA7	0	1
E2F1	AGFG1	0	1
E2F1	B2M	0	1
E2F1	BRD8	0	1
E2F1	CASP8AP2	0	1
E2F1	CCNB2	0	1
E2F1	CCNE2	0	1
E2F1	CDC16	0	1
E2F1	CDC25A	0	1
E2F1	CDC25B	0	1
E2F1	CDH24	0	1
E2F1	CTR9	0	1
E2F1	DNAJB1	0	1
E2F1	DNAJB9	0	1
E2F1	DR1	0	1
E2F1	DSP	0	1
E2F1	DZIP3	0	1
E2F1	FANCG	0	1
E2F1	HMGCR	0	1
E2F1	HOXB4	0	1
E2F1	HSD17B11	0	1
E2F1	IL18BP	0	1
E2F1	INPP5K	0	1
E2F1	INSM1	0	1
E2F1	JADE2	0	1
E2F1	KLF9	0	1
E2F1	KPNA2	0	1
E2F1	KPNB1	0	1
E2F1	MBD2	0	1
E2F1	MBD3	0	1
E2F1	MBD4	0	1
E2F1	MGAT2	0	1
E2F1	MYCBP2	0	1
E2F1	NAB1	0	1
E2F1	NASP	0	1
E2F1	NEIL3	0	1

Table 37 continued

TF	TG	W	# dup
E2F1	NFE2L2	0	1
E2F1	NR3C1	0	1
E2F1	PIK3CD	0	1
E2F1	PTPN9	0	1
E2F1	RAB23	0	1
E2F1	RCAN1	0	1
E2F1	RERE	0	1
E2F1	ROCK1	0	1
E2F1	SDC1	0	1
E2F1	SH3GL2	0	1
E2F1	SLC22A3	0	1
E2F1	SLC38A2	0	1
E2F1	TOB2	0	1
E2F1	TXNRD1	0	1
E2F1	UBE2D3	0	1
E2F1	UNG	0	1
E2F1	VCL	0	1
E2F1	VPS72	0	1
FOXM1	ADH4	0	1
FOXM1	ASF1B	0	1
FOXM1	BUB1B	0	1
FOXM1	C6	0	1
FOXM1	CCNE1	0	1
FOXM1	CDH24	0	1
FOXM1	DZIP3	0	1
FOXM1	FRZB	0	1
FOXM1	G2E3	0	1
FOXM1	GPSM2	0	1
FOXM1	HELLS	0	1
FOXM1	HLA-DOA	0	1
FOXM1	HOXB4	0	1
FOXM1	HSD17B11	0	1
FOXM1	INSM1	0	1
FOXM1	LRRC17	0	1
FOXM1	ME3	0	1
FOXM1	MET	0	1
FOXM1	MNX1	0	1
FOXM1	NEIL3	0	1
FOXM1	PIK3CD	0	1
FOXM1	RAD18	0	1
FOXM1	REEP1	0	1
FOXM1	SDC1	0	1
FOXM1	SH3GL2	0	1
FOXM1	SLC22A3	0	1
FOXM1	STIL	0	1

Table 37 continued

TF	TG	W	# dup
FOXM1	TRIP13	0	1
HIF1A	ADH4	0	1
HIF1A	ASF1B	0	1
HIF1A	AURKB	0	1
HIF1A	BMP2	0	1
HIF1A	BUB1B	0	1
HIF1A	CCNE1	0	1
HIF1A	CDH24	0	1
HIF1A	CENPA	0	1
HIF1A	DZIP3	0	1
HIF1A	G2E3	0	1
HIF1A	GPSM2	0	1
HIF1A	HAUS5	0	1
HIF1A	HELLS	0	1
HIF1A	HLA-DOA	0	1
HIF1A	HOXB4	0	1
HIF1A	HSD17B11	0	1
HIF1A	INSM1	0	1
HIF1A	LRRC17	0	1
HIF1A	ME3	0	1
HIF1A	MNX1	0	1
HIF1A	NCAPH	0	1
HIF1A	NEIL3	0	1
HIF1A	NEK2	0	1
HIF1A	PIK3CD	0	1
HIF1A	RAD18	0	1
HIF1A	REEP1	0	1
HIF1A	SDC1	0	1
HIF1A	SH3GL2	0	1
HIF1A	SLC22A3	0	1
HIF1A	TRIP13	0	1
HOXB4	ABCA7	0	1
HOXB4	ACD	0	1
HOXB4	ADAMTS1	0	1
HOXB4	ADH4	0	1
HOXB4	AGFG1	0	1
HOXB4	AHI1	0	1
HOXB4	ANTXR1	0	1
HOXB4	AP3D1	0	1
HOXB4	ASF1B	0	1
HOXB4	ATF7IP	0	1
HOXB4	AURKB	0	1
HOXB4	B2M	0	1
HOXB4	BAG3	0	1
HOXB4	BAIAP2	0	1

Table 37 continued

TF	TG	W	# dup
HOXB4	BARD1	0	1
HOXB4	BIRC2	0	1
HOXB4	BIRC5	0	1
HOXB4	BMP2	0	1
HOXB4	BORA	0	1
HOXB4	BRCA1	0	1
HOXB4	BRD7	0	1
HOXB4	BRD8	0	1
HOXB4	BUB1	0	1
HOXB4	BUB1B	0	1
HOXB4	BUB3	0	1
HOXB4	C6	0	1
HOXB4	CADM1	0	1
HOXB4	CASP3	0	1
HOXB4	CASP8AP2	0	1
HOXB4	CCNA2	0	1
HOXB4	CCNB1	0	1
HOXB4	CCNB2	0	1
HOXB4	CCNE1	0	1
HOXB4	CCNE2	0	1
HOXB4	CCNF	0	1
HOXB4	CDC16	0	1
HOXB4	CDC25A	0	1
HOXB4	CDC25B	0	1
HOXB4	CDC27	0	1
HOXB4	CDC42	0	1
HOXB4	CDC42EP1	0	1
HOXB4	CDC42EP4	0	1
HOXB4	CDH24	0	1
HOXB4	CDK7	0	1
HOXB4	CDKL5	0	1
HOXB4	CDKN1B	0	1
HOXB4	CDKN2C	0	1
HOXB4	CDKN2D	0	1
HOXB4	CENPA	0	1
HOXB4	CENPE	0	1
HOXB4	CENPF	0	1
HOXB4	CFLAR	0	1
HOXB4	CHAF1A	0	1
HOXB4	CKAP5	0	1
HOXB4	CLSPN	0	1
HOXB4	CREBZF	0	1
HOXB4	CSGALNACT1	0	1
HOXB4	CTCF	0	1
HOXB4	CTNND1	0	1

Table 37 continued

TF	TG	W	# dup
HOXB4	CTR9	0	1
HOXB4	CTSD	0	1
HOXB4	DDX11	0	1
HOXB4	DIS3	0	1
HOXB4	DMXL2	0	1
HOXB4	DNAJB1	0	1
HOXB4	DNAJB6	0	1
HOXB4	DNAJB9	0	1
HOXB4	DR1	0	1
HOXB4	DSP	0	1
HOXB4	DTL	0	1
HOXB4	DZIP3	0	1
HOXB4	E2F1	0	1
HOXB4	EIF4E	0	1
HOXB4	ELP3	0	1
HOXB4	ERN2	0	1
HOXB4	EXO1	0	1
HOXB4	FADD	0	1
HOXB4	FAN1	0	1
HOXB4	FANCD2	0	1
HOXB4	FANCG	0	1
HOXB4	FEM1B	0	1
HOXB4	FEN1	0	1
HOXB4	FKBP1A	0	1
HOXB4	FOXM1	0	1
HOXB4	FRZB	0	1
HOXB4	FYN	0	1
HOXB4	FZR1	0	1
HOXB4	G2E3	0	1
HOXB4	GADD45A	0	1
HOXB4	GAS6	0	1
HOXB4	GCLM	0	1
HOXB4	GNB1	0	1
HOXB4	GOT1	0	1
HOXB4	GPSM2	0	1
HOXB4	H2AFX	0	1
HOXB4	HDAC3	0	1
HOXB4	HELLS	0	1
HOXB4	HIF1A	0	1
HOXB4	HIST1H4B	0	1
HOXB4	HIST1H4C	0	1
HOXB4	HIST1H4E	0	1
HOXB4	HIST1H4H	0	1
HOXB4	HLA-DOA	0	1
HOXB4	HMG1	0	1

Table 37 continued

TF	TG	W	# dup
HOXB4	HMGB2	0	1
HOXB4	HMGCR	0	1
HOXB4	HRAS	0	1
HOXB4	HSD17B11	0	1
HOXB4	HSPA8	0	1
HOXB4	IFIT1	0	1
HOXB4	IL18BP	0	1
HOXB4	INPP5K	0	1
HOXB4	INSIG2	0	1
HOXB4	INSM1	0	1
HOXB4	INSR	0	1
HOXB4	INTS7	0	1
HOXB4	ITGB3	0	1
HOXB4	ITPR1	0	1
HOXB4	ITPR3	0	1
HOXB4	JADE2	0	1
HOXB4	KAT2B	0	1
HOXB4	KAT7	0	1
HOXB4	KATNA1	0	1
HOXB4	KDM5B	0	1
HOXB4	KIF11	0	1
HOXB4	KIF20B	0	1
HOXB4	KIF2C	0	1
HOXB4	KLF9	0	1
HOXB4	KMO	0	1
HOXB4	KPNA2	0	1
HOXB4	KPNB1	0	1
HOXB4	KRAS	0	1
HOXB4	LMNA	0	1
HOXB4	LRRC17	0	1
HOXB4	MAD2L1	0	1
HOXB4	MAN1A2	0	1
HOXB4	MAP2K6	0	1
HOXB4	MAPK13	0	1
HOXB4	MBD2	0	1
HOXB4	MBD3	0	1
HOXB4	MBD4	0	1
HOXB4	MCM2	0	1
HOXB4	MCM4	0	1
HOXB4	MCM6	0	1
HOXB4	MDM2	0	1
HOXB4	ME3	0	1
HOXB4	MGAT2	0	1
HOXB4	MID1	0	1
HOXB4	MITF	0	1

Table 37 continued

TF	TG	W	# dup
HOXB4	MNX1	0	1
HOXB4	MSH2	0	1
HOXB4	MYCBP2	0	1
HOXB4	NAB1	0	1
HOXB4	NASP	0	1
HOXB4	NCAPD2	0	1
HOXB4	NCAPD3	0	1
HOXB4	NCAPH	0	1
HOXB4	NCOA3	0	1
HOXB4	NDE1	0	1
HOXB4	NEIL3	0	1
HOXB4	NEK2	0	1
HOXB4	NFE2L2	0	1
HOXB4	NLRP2	0	1
HOXB4	NPAT	0	1
HOXB4	NPM1	0	1
HOXB4	NR3C1	0	1
HOXB4	NUP160	0	1
HOXB4	OGT	0	1
HOXB4	PCNA	0	1
HOXB4	PDGFA	0	1
HOXB4	PDXP	0	1
HOXB4	PIK3CD	0	1
HOXB4	PKNOX1	0	1
HOXB4	PLK1	0	1
HOXB4	PLK2	0	1
HOXB4	POLA1	0	1
HOXB4	POLD3	0	1
HOXB4	PPP2CA	0	1
HOXB4	PPP3CA	0	1
HOXB4	PRKAR1A	0	1
HOXB4	PTPN9	0	1
HOXB4	PYM1	0	1
HOXB4	RAB23	0	1
HOXB4	RAB3A	0	1
HOXB4	RAD18	0	1
HOXB4	RAD51C	0	1
HOXB4	RBBP8	0	1
HOXB4	RBM8A	0	1
HOXB4	RCAN1	0	1
HOXB4	RECQL4	0	1
HOXB4	REEP1	0	1
HOXB4	RERE	0	1
HOXB4	RHEB	0	1
HOXB4	RHNO1	0	1

Table 37 continued

TF	TG	W	# dup
HOXB4	RHOBTB3	0	1
HOXB4	RNPS1	0	1
HOXB4	ROCK1	0	1
HOXB4	RPA2	0	1
HOXB4	RRM1	0	1
HOXB4	RRM2	0	1
HOXB4	RUNX1	0	1
HOXB4	SAP30BP	0	1
HOXB4	SDC1	0	1
HOXB4	SFPQ	0	1
HOXB4	SH3GL2	0	1
HOXB4	SHC1	0	1
HOXB4	SLBP	0	1
HOXB4	SLC38A2	0	1
HOXB4	SLC44A2	0	1
HOXB4	SMARCB1	0	1
HOXB4	SMARCD1	0	1
HOXB4	SMC4	0	1
HOXB4	SRSF7	0	1
HOXB4	SS18	0	1
HOXB4	STAT1	0	1
HOXB4	STAT5B	0	1
HOXB4	STIL	0	1
HOXB4	SYNCRIP	0	1
HOXB4	TAB2	0	1
HOXB4	TACC3	0	1
HOXB4	TGIF1	0	1
HOXB4	THRAP3	0	1
HOXB4	TIPIN	0	1
HOXB4	TOB2	0	1
HOXB4	TOP2A	0	1
HOXB4	TOPBP1	0	1
HOXB4	TRA2A	0	1
HOXB4	TRIP13	0	1
HOXB4	TSG101	0	1
HOXB4	TXNRD1	0	1
HOXB4	TYMS	0	1
HOXB4	UACA	0	1
HOXB4	UBE2C	0	1
HOXB4	UBE2D3	0	1
HOXB4	UBE2S	0	1
HOXB4	UNG	0	1
HOXB4	USP1	0	1
HOXB4	USP16	0	1
HOXB4	VCAM1	0	1

Table 37 continued

TF	TG	W	# dup
HOXB4	VCL	0	1
HOXB4	VPS72	0	1
HOXB4	YWHAH	0	1
HOXB4	ZWINT	0	1
INSM1	ABCA7	0	1
INSM1	ACD	0	1
INSM1	ADH4	0	1
INSM1	AGFG1	0	1
INSM1	AHI1	0	1
INSM1	ASF1B	0	1
INSM1	ATF7IP	0	1
INSM1	AURKB	0	1
INSM1	BAG3	0	1
INSM1	BARD1	0	1
INSM1	BORA	0	1
INSM1	BUB1	0	1
INSM1	BUB1B	0	1
INSM1	C6	0	1
INSM1	CCNA2	0	1
INSM1	CCNB1	0	1
INSM1	CCNB2	0	1
INSM1	CCNE1	0	1
INSM1	CCNE2	0	1
INSM1	CDC25A	0	1
INSM1	CDC25B	0	1
INSM1	CDH24	0	1
INSM1	CENPA	0	1
INSM1	CENPF	0	1
INSM1	CHAF1A	0	1
INSM1	CKAP5	0	1
INSM1	CLSPN	0	1
INSM1	CSGALNACT1	0	1
INSM1	CTSD	0	1
INSM1	DNAJB1	0	1
INSM1	DZIP3	0	1
INSM1	EIF4E	0	1
INSM1	ELP3	0	1
INSM1	FAN1	0	1
INSM1	FANCD2	0	1
INSM1	FEN1	0	1
INSM1	FRZB	0	1
INSM1	G2E3	0	1
INSM1	GCLM	0	1
INSM1	GPSM2	0	1
INSM1	HELLS	0	1

Table 37 continued

TF	TG	W	# dup
INSM1	HIST1H4B	0	1
INSM1	HIST1H4C	0	1
INSM1	HIST1H4E	0	1
INSM1	HIST1H4H	0	1
INSM1	HOXB4	0	1
INSM1	HSD17B11	0	1
INSM1	IFIT1	0	1
INSM1	IL18BP	0	1
INSM1	INSIG2	0	1
INSM1	INTS7	0	1
INSM1	ITGB3	0	1
INSM1	JADE2	0	1
INSM1	KAT2B	0	1
INSM1	KDM4A	0	1
INSM1	KIF11	0	1
INSM1	KIF20B	0	1
INSM1	KIF2C	0	1
INSM1	KLF9	0	1
INSM1	KPNA2	0	1
INSM1	LMNA	0	1
INSM1	LRRC17	0	1
INSM1	MCM2	0	1
INSM1	MCM6	0	1
INSM1	ME3	0	1
INSM1	MGAT2	0	1
INSM1	MID1	0	1
INSM1	NEIL3	0	1
INSM1	NEK2	0	1
INSM1	NLRP2	0	1
INSM1	NUP160	0	1
INSM1	PCNA	0	1
INSM1	PIK3CD	0	1
INSM1	PLK1	0	1
INSM1	POLD3	0	1
INSM1	POLQ	0	1
INSM1	PPP2CA	0	1
INSM1	PRKAR1A	0	1
INSM1	PYM1	0	1
INSM1	RAD18	0	1
INSM1	RAD51C	0	1
INSM1	RBBP8	0	1
INSM1	RBM8A	0	1
INSM1	RHNO1	0	1
INSM1	RNPS1	0	1
INSM1	RRM2	0	1

Table 37 continued

TF	TG	W	# dup
INSM1	SAP30BP	0	1
INSM1	SDC1	0	1
INSM1	SFPQ	0	1
INSM1	SH3GL2	0	1
INSM1	SLBP	0	1
INSM1	SP1	0	1
INSM1	SRSF7	0	1
INSM1	SS18	0	1
INSM1	STIL	0	1
INSM1	SYNCRIP	0	1
INSM1	TAB2	0	1
INSM1	TFAP2A	0	1
INSM1	THRAP3	0	1
INSM1	TOP2A	0	1
INSM1	TRIP13	0	1
INSM1	TXNRD1	0	1
INSM1	TYMS	0	1
INSM1	UBE2C	0	1
INSM1	UNG	0	1
INSM1	USP1	0	1
INSM1	VEGFC	0	1
INSM1	ZWINT	0	1
KAT7	ADAMTS1	0	1
KAT7	ADH4	0	1
KAT7	ASF1B	0	1
KAT7	AURKB	0	1
KAT7	BIRC5	0	1
KAT7	BORA	0	1
KAT7	BUB1B	0	1
KAT7	C6	0	1
KAT7	CCNA2	0	1
KAT7	CCNE1	0	1
KAT7	CDC25A	0	1
KAT7	CDH24	0	1
KAT7	CHEK2	0	1
KAT7	FANCD2	0	1
KAT7	FRZB	0	1
KAT7	G2E3	0	1
KAT7	GAS1	0	1
KAT7	GPSM2	0	1
KAT7	HAUS5	0	1
KAT7	HELLS	0	1
KAT7	HLA-DOA	0	1
KAT7	HOXB4	0	1
KAT7	HSD17B11	0	1

Table 37 continued

TF	TG	W	# dup
KAT7	IFIT1	0	1
KAT7	IL18BP	0	1
KAT7	INSM1	0	1
KAT7	ITPR3	0	1
KAT7	KDM4A	0	1
KAT7	LRRC17	0	1
KAT7	ME3	0	1
KAT7	MET	0	1
KAT7	MID1	0	1
KAT7	NEIL3	0	1
KAT7	NLRP2	0	1
KAT7	PIK3CD	0	1
KAT7	RAD18	0	1
KAT7	REEP1	0	1
KAT7	SDC1	0	1
KAT7	SH3GL2	0	1
KAT7	SLC22A3	0	1
KAT7	STIL	0	1
KAT7	TRIP13	0	1
KAT7	VCAM1	0	1
KAT7	VEGFC	0	1
KDM5B	ADAMTS1	0	1
KDM5B	ADH4	0	1
KDM5B	ASF1B	0	1
KDM5B	AURKB	0	1
KDM5B	C6	0	1
KDM5B	CCNE1	0	1
KDM5B	CDC25A	0	1
KDM5B	CDH24	0	1
KDM5B	FANCD2	0	1
KDM5B	G2E3	0	1
KDM5B	GPSM2	0	1
KDM5B	HAUS5	0	1
KDM5B	HELLS	0	1
KDM5B	HLA-DOA	0	1
KDM5B	HOXB4	0	1
KDM5B	HSD17B11	0	1
KDM5B	IL18BP	0	1
KDM5B	INSM1	0	1
KDM5B	KDM4A	0	1
KDM5B	LRRC17	0	1
KDM5B	ME3	0	1
KDM5B	MET	0	1
KDM5B	MNX1	0	1
KDM5B	NEIL3	0	1

Table 37 continued

TF	TG	W	# dup
KDM5B	PIK3CD	0	1
KDM5B	RAD18	0	1
KDM5B	REEP1	0	1
KDM5B	SDC1	0	1
KDM5B	SLC22A3	0	1
KDM5B	STIL	0	1
KDM5B	TRIP13	0	1
KLF9	ADAMTS1	0	1
KLF9	ADH4	0	1
KLF9	ARHGAP8	0	1
KLF9	ASF1B	0	1
KLF9	AURKB	0	1
KLF9	BARD1	0	1
KLF9	BIRC5	0	1
KLF9	BMP2	0	1
KLF9	BORA	0	1
KLF9	BUB1	0	1
KLF9	BUB1B	0	1
KLF9	C6	0	1
KLF9	CCNA2	0	1
KLF9	CCNE1	0	1
KLF9	CDC25A	0	1
KLF9	CDH24	0	1
KLF9	CENPA	0	1
KLF9	CHEK2	0	1
KLF9	CLSPN	0	1
KLF9	CSGALNACT1	0	1
KLF9	DDX11	0	1
KLF9	DMXL2	0	1
KLF9	DTL	0	1
KLF9	DZIP3	0	1
KLF9	E2F1	0	1
KLF9	ERN2	0	1
KLF9	FANCD2	0	1
KLF9	GPSM2	0	1
KLF9	HAUS5	0	1
KLF9	HELLS	0	1
KLF9	HLA-DOA	0	1
KLF9	HOXB4	0	1
KLF9	IFIT1	0	1
KLF9	IL18BP	0	1
KLF9	INSM1	0	1
KLF9	ITGB3	0	1
KLF9	ITPR3	0	1
KLF9	KDM4A	0	1

Table 37 continued

TF	TG	W	# dup
KLF9	KIF20B	0	1
KLF9	KMO	0	1
KLF9	LRRC17	0	1
KLF9	ME3	0	1
KLF9	MET	0	1
KLF9	MID1	0	1
KLF9	NCAPH	0	1
KLF9	NLRP2	0	1
KLF9	NPAT	0	1
KLF9	PIK3CD	0	1
KLF9	PLK1	0	1
KLF9	POLQ	0	1
KLF9	RAB3A	0	1
KLF9	RAD18	0	1
KLF9	RAD51	0	1
KLF9	RECQL4	0	1
KLF9	REEP1	0	1
KLF9	SDC1	0	1
KLF9	SH3GL2	0	1
KLF9	SLC22A3	0	1
KLF9	SRSF7	0	1
KLF9	STIL	0	1
KLF9	TAB2	0	1
KLF9	TOP2A	0	1
KLF9	TRIP13	0	1
KLF9	UBE2C	0	1
KLF9	VCAM1	0	1
KLF9	VEGFC	0	1
MBD2	ADAMTS1	0	1
MBD2	ADH4	0	1
MBD2	ARHGAP8	0	1
MBD2	ASF1B	0	1
MBD2	AURKB	0	1
MBD2	BARD1	0	1
MBD2	BIRC5	0	1
MBD2	BMP2	0	1
MBD2	BORA	0	1
MBD2	BUB1	0	1
MBD2	BUB1B	0	1
MBD2	C6	0	1
MBD2	CCNA2	0	1
MBD2	CCNE1	0	1
MBD2	CDC25A	0	1
MBD2	CDH24	0	1
MBD2	CENPA	0	1

Table 37 continued

TF	TG	W	# dup
MBD2	CENPF	0	1
MBD2	CHEK2	0	1
MBD2	CLSPN	0	1
MBD2	CSGALNACT1	0	1
MBD2	DDX11	0	1
MBD2	DMXL2	0	1
MBD2	DTL	0	1
MBD2	DZIP3	0	1
MBD2	E2F1	0	1
MBD2	ERN2	0	1
MBD2	FANCD2	0	1
MBD2	FRZB	0	1
MBD2	G2E3	0	1
MBD2	GAS1	0	1
MBD2	GPSM2	0	1
MBD2	HAUS5	0	1
MBD2	HELLS	0	1
MBD2	HLA-DOA	0	1
MBD2	HOXB4	0	1
MBD2	HSD17B11	0	1
MBD2	IFIT1	0	1
MBD2	IL18BP	0	1
MBD2	INSM1	0	1
MBD2	ITGB3	0	1
MBD2	ITPR3	0	1
MBD2	KDM4A	0	1
MBD2	KIF20B	0	1
MBD2	KMO	0	1
MBD2	LRRC17	0	1
MBD2	ME3	0	1
MBD2	MET	0	1
MBD2	MID1	0	1
MBD2	MITF	0	1
MBD2	MNX1	0	1
MBD2	NCAPH	0	1
MBD2	NEIL3	0	1
MBD2	NEK2	0	1
MBD2	NLRP2	0	1
MBD2	NPAT	0	1
MBD2	PIK3CD	0	1
MBD2	PLK1	0	1
MBD2	POLQ	0	1
MBD2	RAB3A	0	1
MBD2	RAD18	0	1
MBD2	RAD51	0	1

Table 37 continued

TF	TG	W	# dup
MBD2	RECQL4	0	1
MBD2	REEP1	0	1
MBD2	SDC1	0	1
MBD2	SH3GL2	0	1
MBD2	SLC22A3	0	1
MBD2	SRSF7	0	1
MBD2	STIL	0	1
MBD2	TAB2	0	1
MBD2	TFAP2A	0	1
MBD2	TOP2A	0	1
MBD2	TRIP13	0	1
MBD2	UBE2C	0	1
MBD2	VCAM1	0	1
MBD2	VEGFC	0	1
MBD3	ADAMTS1	0	1
MBD3	ADH4	0	1
MBD3	ARHGAP8	0	1
MBD3	ASF1B	0	1
MBD3	AURKB	0	1
MBD3	BARD1	0	1
MBD3	BIRC5	0	1
MBD3	BMP2	0	1
MBD3	BORA	0	1
MBD3	BUB1	0	1
MBD3	BUB1B	0	1
MBD3	C6	0	1
MBD3	CCNA2	0	1
MBD3	CCNE1	0	1
MBD3	CDC25A	0	1
MBD3	CDH24	0	1
MBD3	CENPA	0	1
MBD3	CENPF	0	1
MBD3	CHEK2	0	1
MBD3	CLSPN	0	1
MBD3	CSGALNACT1	0	1
MBD3	DDX11	0	1
MBD3	DMXL2	0	1
MBD3	DTL	0	1
MBD3	DZIP3	0	1
MBD3	E2F1	0	1
MBD3	ERN2	0	1
MBD3	FAN1	0	1
MBD3	FANCD2	0	1
MBD3	FRZB	0	1
MBD3	G2E3	0	1

Table 37 continued

TF	TG	W	# dup
MBD3	GAS1	0	1
MBD3	GPSM2	0	1
MBD3	HAUS5	0	1
MBD3	HELLS	0	1
MBD3	HLA-DOA	0	1
MBD3	HOXB4	0	1
MBD3	HSD17B11	0	1
MBD3	IFIT1	0	1
MBD3	IL18BP	0	1
MBD3	INSM1	0	1
MBD3	ITGB3	0	1
MBD3	ITPR3	0	1
MBD3	KDM4A	0	1
MBD3	KIF20B	0	1
MBD3	KMO	0	1
MBD3	LRRC17	0	1
MBD3	ME3	0	1
MBD3	MID1	0	1
MBD3	MITF	0	1
MBD3	MNX1	0	1
MBD3	NCAPH	0	1
MBD3	NEIL3	0	1
MBD3	NEK2	0	1
MBD3	NLRP2	0	1
MBD3	NPAT	0	1
MBD3	PIK3CD	0	1
MBD3	PLK1	0	1
MBD3	POLQ	0	1
MBD3	RAB3A	0	1
MBD3	RAD18	0	1
MBD3	RAD51	0	1
MBD3	RECQL4	0	1
MBD3	REEP1	0	1
MBD3	SDC1	0	1
MBD3	SH3GL2	0	1
MBD3	SLC22A3	0	1
MBD3	SRSF7	0	1
MBD3	STIL	0	1
MBD3	TAB2	0	1
MBD3	TFAP2A	0	1
MBD3	TOP2A	0	1
MBD3	TRIP13	0	1
MBD3	UBE2C	0	1
MBD3	VCAM1	0	1
MBD3	VEGFC	0	1

Table 37 continued

TF	TG	W	# dup
MBD4	ADAMTS1	0	1
MBD4	ARHGAP8	0	1
MBD4	ASF1B	0	1
MBD4	AURKB	0	1
MBD4	BARD1	0	1
MBD4	BIRC5	0	1
MBD4	BMP2	0	1
MBD4	BORA	0	1
MBD4	BUB1	0	1
MBD4	BUB1B	0	1
MBD4	C6	0	1
MBD4	CCNA2	0	1
MBD4	CCNE1	0	1
MBD4	CDC25A	0	1
MBD4	CDH24	0	1
MBD4	CENPA	0	1
MBD4	CENPF	0	1
MBD4	CHEK2	0	1
MBD4	CLSPN	0	1
MBD4	CSGALNACT1	0	1
MBD4	DDX11	0	1
MBD4	DMXL2	0	1
MBD4	DTL	0	1
MBD4	DZIP3	0	1
MBD4	E2F1	0	1
MBD4	ERN2	0	1
MBD4	FANCD2	0	1
MBD4	FRZB	0	1
MBD4	G2E3	0	1
MBD4	GAS1	0	1
MBD4	GPSM2	0	1
MBD4	HAUS5	0	1
MBD4	HELLS	0	1
MBD4	HLA-DOA	0	1
MBD4	HOXB4	0	1
MBD4	HSD17B11	0	1
MBD4	IFIT1	0	1
MBD4	IL18BP	0	1
MBD4	INSM1	0	1
MBD4	ITGB3	0	1
MBD4	ITPR3	0	1
MBD4	KDM4A	0	1
MBD4	KIF20B	0	1
MBD4	KMO	0	1
MBD4	LRRC17	0	1

Table 37 continued

TF	TG	W	# dup
MBD4	ME3	0	1
MBD4	MET	0	1
MBD4	MID1	0	1
MBD4	MITF	0	1
MBD4	MNX1	0	1
MBD4	NCAPH	0	1
MBD4	NEIL3	0	1
MBD4	NEK2	0	1
MBD4	NLRP2	0	1
MBD4	NPAT	0	1
MBD4	PIK3CD	0	1
MBD4	PLK1	0	1
MBD4	POLQ	0	1
MBD4	RAB3A	0	1
MBD4	RAD18	0	1
MBD4	RAD51	0	1
MBD4	RECQL4	0	1
MBD4	REEP1	0	1
MBD4	SDC1	0	1
MBD4	SH3GL2	0	1
MBD4	SLC22A3	0	1
MBD4	SRSF7	0	1
MBD4	STIL	0	1
MBD4	TAB2	0	1
MBD4	TFAP2A	0	1
MBD4	TOP2A	0	1
MBD4	TRIP13	0	1
MBD4	UBE2C	0	1
MBD4	VCAM1	0	1
MBD4	VEGFC	0	1
MITF	ABCA7	0	1
MITF	ADAMTS1	0	1
MITF	ADH4	0	1
MITF	AGFG1	0	1
MITF	AURKB	0	1
MITF	B2M	0	1
MITF	BRD8	0	1
MITF	BUB1B	0	1
MITF	CASP8AP2	0	1
MITF	CCNB2	0	1
MITF	CCNE2	0	1
MITF	CDC25A	0	1
MITF	CDC27	0	1
MITF	CDH24	0	1
MITF	CDKL5	0	1

Table 37 continued

TF	TG	W	# dup
MITF	CHAF1A	0	1
MITF	CKAP5	0	1
MITF	CTR9	0	1
MITF	CTSD	0	1
MITF	DZIP3	0	1
MITF	FANCD2	0	1
MITF	FANCG	0	1
MITF	GPSM2	0	1
MITF	HELLS	0	1
MITF	HLA-DOA	0	1
MITF	HMGCR	0	1
MITF	HSD17B11	0	1
MITF	IL18BP	0	1
MITF	INPP5K	0	1
MITF	INSM1	0	1
MITF	ITPR1	0	1
MITF	KIF11	0	1
MITF	KPNB1	0	1
MITF	KRAS	0	1
MITF	MAP2K6	0	1
MITF	MBD4	0	1
MITF	MET	0	1
MITF	MGAT2	0	1
MITF	MYCBP2	0	1
MITF	NASP	0	1
MITF	NCAPD2	0	1
MITF	NCAPD3	0	1
MITF	NEIL3	0	1
MITF	NEK2	0	1
MITF	NR3C1	0	1
MITF	NUP160	0	1
MITF	PPP3CA	0	1
MITF	PTPN9	0	1
MITF	RAB23	0	1
MITF	RAD18	0	1
MITF	RCAN1	0	1
MITF	REEP1	0	1
MITF	RERE	0	1
MITF	ROCK1	0	1
MITF	RRM1	0	1
MITF	SH3GL2	0	1
MITF	SLC22A3	0	1
MITF	SMC4	0	1
MITF	STIL	0	1
MITF	TOPBP1	0	1

Table 37 continued

TF	TG	W	# dup
MITF	TYMS	0	1
MITF	VCL	0	1
MITF	VPS72	0	1
MNX1	ABCA7	0	1
MNX1	ACD	0	1
MNX1	ADAMTS1	0	1
MNX1	ADH4	0	1
MNX1	AGFG1	0	1
MNX1	AHI1	0	1
MNX1	ANTXR1	0	1
MNX1	AP3D1	0	1
MNX1	ARHGAP8	0	1
MNX1	ASF1B	0	1
MNX1	ATF7IP	0	1
MNX1	AURKB	0	1
MNX1	B2M	0	1
MNX1	BAG3	0	1
MNX1	BAIAP2	0	1
MNX1	BARD1	0	1
MNX1	BIRC2	0	1
MNX1	BIRC5	0	1
MNX1	BMP2	0	1
MNX1	BORA	0	1
MNX1	BRD7	0	1
MNX1	BRD8	0	1
MNX1	BUB1	0	1
MNX1	BUB1B	0	1
MNX1	BUB3	0	1
MNX1	CADM1	0	1
MNX1	CASP3	0	1
MNX1	CASP8AP2	0	1
MNX1	CCNA2	0	1
MNX1	CCNB1	0	1
MNX1	CCNB2	0	1
MNX1	CCNE1	0	1
MNX1	CCNE2	0	1
MNX1	CCNF	0	1
MNX1	CDC25A	0	1
MNX1	CDC25B	0	1
MNX1	CDC27	0	1
MNX1	CDC42EP1	0	1
MNX1	CDC42EP4	0	1
MNX1	CDH24	0	1
MNX1	CDK7	0	1
MNX1	CDKL5	0	1

Table 37 continued

TF	TG	W	# dup
MNX1	CDKN1B	0	1
MNX1	CDKN2C	0	1
MNX1	CDKN2D	0	1
MNX1	CENPA	0	1
MNX1	CENPE	0	1
MNX1	CENPF	0	1
MNX1	CFLAR	0	1
MNX1	CHAF1A	0	1
MNX1	CHEK2	0	1
MNX1	CKAP5	0	1
MNX1	CLSPN	0	1
MNX1	CREBZF	0	1
MNX1	CSGALNACT1	0	1
MNX1	CTCF	0	1
MNX1	CTNND1	0	1
MNX1	CTR9	0	1
MNX1	CTSD	0	1
MNX1	DDX11	0	1
MNX1	DIS3	0	1
MNX1	DMXL2	0	1
MNX1	DNAJB1	0	1
MNX1	DNAJB6	0	1
MNX1	DNAJB9	0	1
MNX1	DR1	0	1
MNX1	DSP	0	1
MNX1	DTL	0	1
MNX1	DZIP3	0	1
MNX1	EIF4E	0	1
MNX1	ELP3	0	1
MNX1	ERN2	0	1
MNX1	EXO1	0	1
MNX1	FADD	0	1
MNX1	FAN1	0	1
MNX1	FANCD2	0	1
MNX1	FANCG	0	1
MNX1	FEM1B	0	1
MNX1	FEN1	0	1
MNX1	FKBP1A	0	1
MNX1	FRZB	0	1
MNX1	FZR1	0	1
MNX1	G2E3	0	1
MNX1	GADD45A	0	1
MNX1	GCLM	0	1
MNX1	GNB1	0	1
MNX1	GOT1	0	1

Table 37 continued

TF	TG	W	# dup
MNX1	GPSM2	0	1
MNX1	H2AFX	0	1
MNX1	HAUS5	0	1
MNX1	HDAC3	0	1
MNX1	HIST1H4B	0	1
MNX1	HIST1H4C	0	1
MNX1	HIST1H4E	0	1
MNX1	HIST1H4H	0	1
MNX1	HLA-DOA	0	1
MNX1	HMG1	0	1
MNX1	HMGB2	0	1
MNX1	HMGCR	0	1
MNX1	HOXB4	0	1
MNX1	HRAS	0	1
MNX1	HSD17B11	0	1
MNX1	HSPA8	0	1
MNX1	IFIT1	0	1
MNX1	IL18BP	0	1
MNX1	INPP5K	0	1
MNX1	INSIG2	0	1
MNX1	INTS7	0	1
MNX1	ITGB3	0	1
MNX1	ITPR1	0	1
MNX1	ITPR3	0	1
MNX1	JADE2	0	1
MNX1	KAT2B	0	1
MNX1	KAT7	0	1
MNX1	KDM4A	0	1
MNX1	KDM5B	0	1
MNX1	KIF11	0	1
MNX1	KIF20B	0	1
MNX1	KIF2C	0	1
MNX1	KLF9	0	1
MNX1	KMO	0	1
MNX1	KPNA2	0	1
MNX1	KPNB1	0	1
MNX1	KRAS	0	1
MNX1	LRRC17	0	1
MNX1	MAD2L1	0	1
MNX1	MAN1A2	0	1
MNX1	MAP2K6	0	1
MNX1	MAPK13	0	1
MNX1	MBD2	0	1
MNX1	MBD3	0	1
MNX1	MBD4	0	1

Table 37 continued

TF	TG	W	# dup
MNX1	MCM2	0	1
MNX1	MCM4	0	1
MNX1	MCM6	0	1
MNX1	MDM2	0	1
MNX1	ME3	0	1
MNX1	MET	0	1
MNX1	MGAT2	0	1
MNX1	MID1	0	1
MNX1	MSH2	0	1
MNX1	NAB1	0	1
MNX1	NASP	0	1
MNX1	NCAPD2	0	1
MNX1	NCAPD3	0	1
MNX1	NCAPH	0	1
MNX1	NCOA3	0	1
MNX1	NEIL3	0	1
MNX1	NEK2	0	1
MNX1	NFE2L2	0	1
MNX1	NLRP2	0	1
MNX1	NPAT	0	1
MNX1	NPM1	0	1
MNX1	NR3C1	0	1
MNX1	NUP160	0	1
MNX1	OGT	0	1
MNX1	PCNA	0	1
MNX1	PDXP	0	1
MNX1	PIK3CD	0	1
MNX1	PKNOX1	0	1
MNX1	PLK1	0	1
MNX1	PLK2	0	1
MNX1	POLA1	0	1
MNX1	POLD3	0	1
MNX1	POLQ	0	1
MNX1	PPP2CA	0	1
MNX1	PRKAR1A	0	1
MNX1	PTPN9	0	1
MNX1	PYM1	0	1
MNX1	RAD18	0	1
MNX1	RAD51	0	1
MNX1	RAD51C	0	1
MNX1	RBM8A	0	1
MNX1	RECQL4	0	1
MNX1	REEP1	0	1
MNX1	RERE	0	1
MNX1	RHEB	0	1

Table 37 continued

TF	TG	W	# dup
MNX1	RHNO1	0	1
MNX1	RHOBTB3	0	1
MNX1	RNPS1	0	1
MNX1	ROCK1	0	1
MNX1	RPA2	0	1
MNX1	RRM1	0	1
MNX1	RRM2	0	1
MNX1	SAP30BP	0	1
MNX1	SFPQ	0	1
MNX1	SH3GL2	0	1
MNX1	SLBP	0	1
MNX1	SLC22A3	0	1
MNX1	SLC38A2	0	1
MNX1	SLC44A2	0	1
MNX1	SMARCD1	0	1
MNX1	SMC4	0	1
MNX1	SP1	0	1
MNX1	SRSF7	0	1
MNX1	SS18	0	1
MNX1	STAT1	0	1
MNX1	STAT5B	0	1
MNX1	STIL	0	1
MNX1	SYNCRIP	0	1
MNX1	TAB2	0	1
MNX1	TACC3	0	1
MNX1	TFAP2A	0	1
MNX1	THRAP3	0	1
MNX1	TIPIN	0	1
MNX1	TOB2	0	1
MNX1	TOP2A	0	1
MNX1	TOPBP1	0	1
MNX1	TRA2A	0	1
MNX1	TRIP13	0	1
MNX1	TXNRD1	0	1
MNX1	TYMS	0	1
MNX1	UACA	0	1
MNX1	UBE2C	0	1
MNX1	UBE2D3	0	1
MNX1	UBE2S	0	1
MNX1	UNG	0	1
MNX1	USP1	0	1
MNX1	USP16	0	1
MNX1	VCL	0	1
MNX1	VEGFC	0	1
MNX1	VPS72	0	1

Table 37 continued

TF	TG	W	# dup
MNX1	YWHAH	0	1
MNX1	YY1	0	1
MNX1	ZWINT	0	1
NCOA3	ADAMTS1	0	1
NCOA3	ADH4	0	1
NCOA3	AURKB	0	1
NCOA3	BUB1B	0	1
NCOA3	C6	0	1
NCOA3	CCNE1	0	1
NCOA3	CDC25A	0	1
NCOA3	CDH24	0	1
NCOA3	FANCD2	0	1
NCOA3	G2E3	0	1
NCOA3	GPSM2	0	1
NCOA3	HAUS5	0	1
NCOA3	HELLS	0	1
NCOA3	HLA-DOA	0	1
NCOA3	HOXB4	0	1
NCOA3	HSD17B11	0	1
NCOA3	IL18BP	0	1
NCOA3	KDM4A	0	1
NCOA3	LRRC17	0	1
NCOA3	ME3	0	1
NCOA3	MET	0	1
NCOA3	MNX1	0	1
NCOA3	NEIL3	0	1
NCOA3	PIK3CD	0	1
NCOA3	RAD18	0	1
NCOA3	REEP1	0	1
NCOA3	SDC1	0	1
NCOA3	SH3GL2	0	1
NCOA3	SLC22A3	0	1
NCOA3	STIL	0	1
NCOA3	TRIP13	0	1
NFE2L2	ADAMTS1	0	1
NFE2L2	ADH4	0	1
NFE2L2	ARHGAP8	0	1
NFE2L2	ASF1B	0	1
NFE2L2	AURKB	0	1
NFE2L2	BARD1	0	1
NFE2L2	BIRC5	0	1
NFE2L2	BMP2	0	1
NFE2L2	BORA	0	1
NFE2L2	BUB1	0	1
NFE2L2	BUB1B	0	1

Table 37 continued

TF	TG	W	# dup
NFE2L2	C6	0	1
NFE2L2	CCNA2	0	1
NFE2L2	CCNE1	0	1
NFE2L2	CDC25A	0	1
NFE2L2	CDH24	0	1
NFE2L2	CENPA	0	1
NFE2L2	CENPF	0	1
NFE2L2	CHEK2	0	1
NFE2L2	CLSPN	0	1
NFE2L2	DMXL2	0	1
NFE2L2	DTL	0	1
NFE2L2	DZIP3	0	1
NFE2L2	E2F1	0	1
NFE2L2	ERN2	0	1
NFE2L2	FAN1	0	1
NFE2L2	FANCD2	0	1
NFE2L2	FRZB	0	1
NFE2L2	G2E3	0	1
NFE2L2	GAS1	0	1
NFE2L2	GPSM2	0	1
NFE2L2	HAUS5	0	1
NFE2L2	HELLS	0	1
NFE2L2	HLA-DOA	0	1
NFE2L2	HOXB4	0	1
NFE2L2	HSD17B11	0	1
NFE2L2	IFIT1	0	1
NFE2L2	IL18BP	0	1
NFE2L2	INSM1	0	1
NFE2L2	ITGB3	0	1
NFE2L2	ITPR3	0	1
NFE2L2	KDM4A	0	1
NFE2L2	KIF20B	0	1
NFE2L2	KMO	0	1
NFE2L2	LRRRC17	0	1
NFE2L2	ME3	0	1
NFE2L2	MET	0	1
NFE2L2	MID1	0	1
NFE2L2	MITF	0	1
NFE2L2	MNX1	0	1
NFE2L2	NCAPH	0	1
NFE2L2	NEIL3	0	1
NFE2L2	NLRP2	0	1
NFE2L2	NPAT	0	1
NFE2L2	PIK3CD	0	1
NFE2L2	PLK1	0	1

Table 37 continued

TF	TG	W	# dup
NFE2L2	POLQ	0	1
NFE2L2	RAB3A	0	1
NFE2L2	RAD18	0	1
NFE2L2	RAD51	0	1
NFE2L2	RECQL4	0	1
NFE2L2	REEP1	0	1
NFE2L2	SDC1	0	1
NFE2L2	SH3GL2	0	1
NFE2L2	SLC22A3	0	1
NFE2L2	SRSF7	0	1
NFE2L2	STIL	0	1
NFE2L2	TAB2	0	1
NFE2L2	TFAP2A	0	1
NFE2L2	TOP2A	0	1
NFE2L2	TRIP13	0	1
NFE2L2	UBE2C	0	1
NFE2L2	VCAM1	0	1
NFE2L2	VEGFC	0	1
NR3C1	ADAMTS1	0	1
NR3C1	ADH4	0	1
NR3C1	ASF1B	0	1
NR3C1	AURKB	0	1
NR3C1	BARD1	0	1
NR3C1	BIRC5	0	1
NR3C1	BORA	0	1
NR3C1	BUB1	0	1
NR3C1	BUB1B	0	1
NR3C1	C6	0	1
NR3C1	CCNE1	0	1
NR3C1	CDH24	0	1
NR3C1	CENPA	0	1
NR3C1	CENPF	0	1
NR3C1	CSGALNACT1	0	1
NR3C1	DDX11	0	1
NR3C1	DMXL2	0	1
NR3C1	DZIP3	0	1
NR3C1	E2F1	0	1
NR3C1	ERN2	0	1
NR3C1	G2E3	0	1
NR3C1	GAS1	0	1
NR3C1	GPSM2	0	1
NR3C1	HAUS5	0	1
NR3C1	HELLS	0	1
NR3C1	HLA-DOA	0	1
NR3C1	HOXB4	0	1

Table 37 continued

TF	TG	W	# dup
NR3C1	HSD17B11	0	1
NR3C1	INSM1	0	1
NR3C1	ITGB3	0	1
NR3C1	KDM4A	0	1
NR3C1	LRRC17	0	1
NR3C1	ME3	0	1
NR3C1	MET	0	1
NR3C1	MITF	0	1
NR3C1	MNX1	0	1
NR3C1	NCAPH	0	1
NR3C1	NEIL3	0	1
NR3C1	NEK2	0	1
NR3C1	NLRP2	0	1
NR3C1	NPAT	0	1
NR3C1	PIK3CD	0	1
NR3C1	POLQ	0	1
NR3C1	RAB3A	0	1
NR3C1	RAD18	0	1
NR3C1	REEP1	0	1
NR3C1	SDC1	0	1
NR3C1	SH3GL2	0	1
NR3C1	SRSF7	0	1
NR3C1	STIL	0	1
NR3C1	TRIP13	0	1
NR3C1	UBE2C	0	1
PKNOX1	ADAMTS1	0	1
PKNOX1	ADH4	0	1
PKNOX1	ARHGAP8	0	1
PKNOX1	AURKB	0	1
PKNOX1	BARD1	0	1
PKNOX1	BIRC5	0	1
PKNOX1	BMP2	0	1
PKNOX1	BORA	0	1
PKNOX1	BUB1	0	1
PKNOX1	BUB1B	0	1
PKNOX1	CCNA2	0	1
PKNOX1	CCNE1	0	1
PKNOX1	CDC25A	0	1
PKNOX1	CDH24	0	1
PKNOX1	CENPA	0	1
PKNOX1	CENPF	0	1
PKNOX1	CHEK2	0	1
PKNOX1	CLSPN	0	1
PKNOX1	CSGALNACT1	0	1
PKNOX1	DDX11	0	1

Table 37 continued

TF	TG	W	# dup
PKNOX1	DMXL2	0	1
PKNOX1	DTL	0	1
PKNOX1	DZIP3	0	1
PKNOX1	E2F1	0	1
PKNOX1	ERN2	0	1
PKNOX1	FAN1	0	1
PKNOX1	FANCD2	0	1
PKNOX1	FRZB	0	1
PKNOX1	G2E3	0	1
PKNOX1	GPSM2	0	1
PKNOX1	HAUS5	0	1
PKNOX1	HOXB4	0	1
PKNOX1	HSD17B11	0	1
PKNOX1	IFIT1	0	1
PKNOX1	IL18BP	0	1
PKNOX1	INSM1	0	1
PKNOX1	ITGB3	0	1
PKNOX1	ITPR3	0	1
PKNOX1	KDM4A	0	1
PKNOX1	KIF20B	0	1
PKNOX1	LRRC17	0	1
PKNOX1	ME3	0	1
PKNOX1	MET	0	1
PKNOX1	MID1	0	1
PKNOX1	MNX1	0	1
PKNOX1	NCAPH	0	1
PKNOX1	NEIL3	0	1
PKNOX1	NEK2	0	1
PKNOX1	NLRP2	0	1
PKNOX1	NPAT	0	1
PKNOX1	PIK3CD	0	1
PKNOX1	PLK1	0	1
PKNOX1	POLQ	0	1
PKNOX1	RAB3A	0	1
PKNOX1	RAD18	0	1
PKNOX1	RAD51	0	1
PKNOX1	RECQL4	0	1
PKNOX1	REEP1	0	1
PKNOX1	SDC1	0	1
PKNOX1	SH3GL2	0	1
PKNOX1	SLC22A3	0	1
PKNOX1	SRSF7	0	1
PKNOX1	STIL	0	1
PKNOX1	TAB2	0	1
PKNOX1	TFAP2A	0	1

Table 37 continued

TF	TG	W	# dup
PKNOX1	TOP2A	0	1
PKNOX1	TRIP13	0	1
PKNOX1	UBE2C	0	1
PKNOX1	VCAM1	0	1
PKNOX1	VEGFC	0	1
RUNX1	ADH4	0	1
RUNX1	ASF1B	0	1
RUNX1	AURKB	0	1
RUNX1	BUB1B	0	1
RUNX1	C6	0	1
RUNX1	CCNE1	0	1
RUNX1	CDC25A	0	1
RUNX1	CDH24	0	1
RUNX1	CENPA	0	1
RUNX1	DZIP3	0	1
RUNX1	FANCD2	0	1
RUNX1	G2E3	0	1
RUNX1	HAUS5	0	1
RUNX1	HELLS	0	1
RUNX1	HLA-DOA	0	1
RUNX1	HSD17B11	0	1
RUNX1	IL18BP	0	1
RUNX1	INSM1	0	1
RUNX1	MNX1	0	1
RUNX1	NCAPH	0	1
RUNX1	NEK2	0	1
RUNX1	RAD18	0	1
RUNX1	REEP1	0	1
RUNX1	SDC1	0	1
RUNX1	SH3GL2	0	1
RUNX1	SLC22A3	0	1
RUNX1	STIL	0	1
SP1	ADAMTS1	0	1
SP1	ADH4	0	1
SP1	ASF1B	0	1
SP1	AURKB	0	1
SP1	C6	0	1
SP1	CDH24	0	1
SP1	CENPA	0	1
SP1	DZIP3	0	1
SP1	FANCD2	0	1
SP1	HAUS5	0	1
SP1	HELLS	0	1
SP1	HLA-DOA	0	1
SP1	HOXB4	0	1

Table 37 continued

TF	TG	W	# dup
SP1	IL18BP	0	1
SP1	INSM1	0	1
SP1	LRRC17	0	1
SP1	ME3	0	1
SP1	MITF	0	1
SP1	MNX1	0	1
SP1	NCAPH	0	1
SP1	NEIL3	0	1
SP1	NEK2	0	1
SP1	PIK3CD	0	1
SP1	RAB3A	0	1
SP1	RAD18	0	1
SP1	REEP1	0	1
SP1	SDC1	0	1
SP1	SLC22A3	0	1
SP1	STIL	0	1
SP1	TRIP13	0	1
SRF	ADH4	0	1
SRF	ASF1B	0	1
SRF	AURKB	0	1
SRF	BMP2	0	1
SRF	BUB1B	0	1
SRF	C6	0	1
SRF	CCNE1	0	1
SRF	CDH24	0	1
SRF	DZIP3	0	1
SRF	G2E3	0	1
SRF	GPSM2	0	1
SRF	HAUS5	0	1
SRF	HELLS	0	1
SRF	HLA-DOA	0	1
SRF	HSD17B11	0	1
SRF	INSM1	0	1
SRF	ITGB3	0	1
SRF	LRRC17	0	1
SRF	ME3	0	1
SRF	MET	0	1
SRF	MNX1	0	1
SRF	NCAPH	0	1
SRF	NEIL3	0	1
SRF	NEK2	0	1
SRF	PIK3CD	0	1
SRF	RAD18	0	1
SRF	REEP1	0	1
SRF	SDC1	0	1

Table 37 continued

TF	TG	W	# dup
SRF	SH3GL2	0	1
SRF	TRIP13	0	1
STAT1	ADH4	0	1
STAT1	AURKB	0	1
STAT1	BUB1B	0	1
STAT1	CDH24	0	1
STAT1	GPSM2	0	1
STAT1	HAUS5	0	1
STAT1	HELLS	0	1
STAT1	HLA-DOA	0	1
STAT1	HOXB4	0	1
STAT1	INSM1	0	1
STAT5B	ADH4	0	1
STAT5B	ASF1B	0	1
STAT5B	AURKB	0	1
STAT5B	BIRC5	0	1
STAT5B	BUB1B	0	1
STAT5B	C6	0	1
STAT5B	CCNE1	0	1
STAT5B	CDH24	0	1
STAT5B	CENPA	0	1
STAT5B	CHEK2	0	1
STAT5B	CSGALNACT1	0	1
STAT5B	DDX11	0	1
STAT5B	DZIP3	0	1
STAT5B	G2E3	0	1
STAT5B	GPSM2	0	1
STAT5B	HAUS5	0	1
STAT5B	HELLS	0	1
STAT5B	HOXB4	0	1
STAT5B	HSD17B11	0	1
STAT5B	INSM1	0	1
STAT5B	KIF20B	0	1
STAT5B	LRRC17	0	1
STAT5B	ME3	0	1
STAT5B	MNX1	0	1
STAT5B	NCAPH	0	1
STAT5B	NEIL3	0	1
STAT5B	NEK2	0	1
STAT5B	PIK3CD	0	1
STAT5B	POLQ	0	1
STAT5B	RAB3A	0	1
STAT5B	RAD18	0	1
STAT5B	RECQL4	0	1
STAT5B	REEP1	0	1

Table 37 continued

TF	TG	W	# dup
STAT5B	SDC1	0	1
STAT5B	SH3GL2	0	1
STAT5B	SLC22A3	0	1
STAT5B	STIL	0	1
STAT5B	TRIP13	0	1
TFAP2A	ABCA7	0	1
TFAP2A	ADAMTS1	0	1
TFAP2A	ADH4	0	1
TFAP2A	AGFG1	0	1
TFAP2A	AHI1	0	1
TFAP2A	ASF1B	0	1
TFAP2A	AURKB	0	1
TFAP2A	B2M	0	1
TFAP2A	BAG3	0	1
TFAP2A	BMP2	0	1
TFAP2A	BRD8	0	1
TFAP2A	BUB1B	0	1
TFAP2A	C6	0	1
TFAP2A	CCNB2	0	1
TFAP2A	CCNE1	0	1
TFAP2A	CCNE2	0	1
TFAP2A	CDC16	0	1
TFAP2A	CDC25B	0	1
TFAP2A	CDC27	0	1
TFAP2A	CDH24	0	1
TFAP2A	CDKL5	0	1
TFAP2A	CENPA	0	1
TFAP2A	CHAF1A	0	1
TFAP2A	CKAP5	0	1
TFAP2A	CTR9	0	1
TFAP2A	DNAJB1	0	1
TFAP2A	DNAJB9	0	1
TFAP2A	DR1	0	1
TFAP2A	FANCG	0	1
TFAP2A	G2E3	0	1
TFAP2A	GAS1	0	1
TFAP2A	GPSM2	0	1
TFAP2A	HAUS5	0	1
TFAP2A	HELLS	0	1
TFAP2A	HLA-DOA	0	1
TFAP2A	HMGCR	0	1
TFAP2A	HOXB4	0	1
TFAP2A	HSD17B11	0	1
TFAP2A	INPP5K	0	1
TFAP2A	INSM1	0	1

Table 37 continued

TF	TG	W	# dup
TFAP2A	ITPR3	0	1
TFAP2A	JADE2	0	1
TFAP2A	KIF11	0	1
TFAP2A	KLF9	0	1
TFAP2A	KPNA2	0	1
TFAP2A	KPNB1	0	1
TFAP2A	KRAS	0	1
TFAP2A	LRRC17	0	1
TFAP2A	MAN1A2	0	1
TFAP2A	MAP2K6	0	1
TFAP2A	MBD2	0	1
TFAP2A	MBD3	0	1
TFAP2A	MBD4	0	1
TFAP2A	ME3	0	1
TFAP2A	MET	0	1
TFAP2A	MGAT2	0	1
TFAP2A	MNX1	0	1
TFAP2A	MYCBP2	0	1
TFAP2A	NAB1	0	1
TFAP2A	NASP	0	1
TFAP2A	NCAPD2	0	1
TFAP2A	NCAPD3	0	1
TFAP2A	NCAPH	0	1
TFAP2A	NDE1	0	1
TFAP2A	NEIL3	0	1
TFAP2A	NFE2L2	0	1
TFAP2A	NUP160	0	1
TFAP2A	PIK3CD	0	1
TFAP2A	PKNOX1	0	1
TFAP2A	PPP3CA	0	1
TFAP2A	PRKAR1A	0	1
TFAP2A	PTPN9	0	1
TFAP2A	RAB23	0	1
TFAP2A	RAD18	0	1
TFAP2A	RCAN1	0	1
TFAP2A	REEP1	0	1
TFAP2A	RERE	0	1
TFAP2A	ROCK1	0	1
TFAP2A	RRM1	0	1
TFAP2A	RRM2	0	1
TFAP2A	SDC1	0	1
TFAP2A	SH3GL2	0	1
TFAP2A	SLC22A3	0	1
TFAP2A	SLC38A2	0	1
TFAP2A	SMC4	0	1

Table 37 continued

TF	TG	W	# dup
TFAP2A	STIL	0	1
TFAP2A	TACC3	0	1
TFAP2A	TGIF1	0	1
TFAP2A	TOB2	0	1
TFAP2A	TRIP13	0	1
TFAP2A	TXNRD1	0	1
TFAP2A	UACA	0	1
TFAP2A	UBE2D3	0	1
TFAP2A	UNG	0	1
TFAP2A	VCL	0	1
TFAP2A	VPS72	0	1
TGIF1	ADAMTS1	0	1
TGIF1	ADH4	0	1
TGIF1	ARHGAP8	0	1
TGIF1	ASF1B	0	1
TGIF1	AURKB	0	1
TGIF1	BARD1	0	1
TGIF1	BIRC5	0	1
TGIF1	BMP2	0	1
TGIF1	BORA	0	1
TGIF1	BUB1	0	1
TGIF1	BUB1B	0	1
TGIF1	C6	0	1
TGIF1	CCNA2	0	1
TGIF1	CCNE1	0	1
TGIF1	CDC25A	0	1
TGIF1	CDH24	0	1
TGIF1	CENPA	0	1
TGIF1	CENPF	0	1
TGIF1	CHEK2	0	1
TGIF1	CLSPN	0	1
TGIF1	CSGALNACT1	0	1
TGIF1	DDX11	0	1
TGIF1	DMXL2	0	1
TGIF1	DTL	0	1
TGIF1	DZIP3	0	1
TGIF1	E2F1	0	1
TGIF1	ERN2	0	1
TGIF1	FAN1	0	1
TGIF1	FANCD2	0	1
TGIF1	FRZB	0	1
TGIF1	G2E3	0	1
TGIF1	GAS1	0	1
TGIF1	GPSM2	0	1
TGIF1	HAUS5	0	1

Table 37 continued

TF	TG	W	# dup
TGIF1	HELLS	0	1
TGIF1	HLA-DOA	0	1
TGIF1	HOXB4	0	1
TGIF1	HSD17B11	0	1
TGIF1	IFIT1	0	1
TGIF1	IL18BP	0	1
TGIF1	INSM1	0	1
TGIF1	ITGB3	0	1
TGIF1	ITPR3	0	1
TGIF1	KDM4A	0	1
TGIF1	KIF20B	0	1
TGIF1	KMO	0	1
TGIF1	LRRC17	0	1
TGIF1	ME3	0	1
TGIF1	MID1	0	1
TGIF1	MITF	0	1
TGIF1	NCAPH	0	1
TGIF1	NEIL3	0	1
TGIF1	NEK2	0	1
TGIF1	NLRP2	0	1
TGIF1	NPAT	0	1
TGIF1	PIK3CD	0	1
TGIF1	PLK1	0	1
TGIF1	POLQ	0	1
TGIF1	RAB3A	0	1
TGIF1	RAD18	0	1
TGIF1	RAD51	0	1
TGIF1	RECQL4	0	1
TGIF1	REEP1	0	1
TGIF1	SDC1	0	1
TGIF1	SH3GL2	0	1
TGIF1	SLC22A3	0	1
TGIF1	SRSF7	0	1
TGIF1	STIL	0	1
TGIF1	TAB2	0	1
TGIF1	TFAP2A	0	1
TGIF1	TOP2A	0	1
TGIF1	TRIP13	0	1
TGIF1	UBE2C	0	1
TGIF1	VCAM1	0	1
TGIF1	VEGFC	0	1
YY1	ADAMTS1	0	1
YY1	ADH4	0	1
YY1	ASF1B	0	1
YY1	AURKB	0	1

Table 37 continued

TF	TG	W	# dup
YY1	BUB1B	0	1
YY1	C6	0	1
YY1	CCNE1	0	1
YY1	CDH24	0	1
YY1	CENPA	0	1
YY1	DZIP3	0	1
YY1	G2E3	0	1
YY1	GPSM2	0	1
YY1	HAUS5	0	1
YY1	HELLS	0	1
YY1	HLA-DOA	0	1
YY1	HSD17B11	0	1
YY1	INSM1	0	1
YY1	LRRC17	0	1
YY1	ME3	0	1
YY1	MET	0	1
YY1	MITF	0	1
YY1	MNX1	0	1
YY1	NCAPH	0	1
YY1	NEIL3	0	1
YY1	NEK2	0	1
YY1	PIK3CD	0	1
YY1	RAD18	0	1
YY1	REEP1	0	1
YY1	SDC1	0	1
YY1	SH3GL2	0	1
YY1	SLC22A3	0	1
YY1	STIL	0	1
YY1	TRIP13	0	1

The table gives the list of negative edges in our gold standard network. The 1st column represents the TF. The 2nd column the TG. The 3rd column informs for each edge, if it is present in the network (value of 1) or if it is absent (value of 0). The present edges are the positive links and the absent edges are the negative links. For each edge, the number in the 4th column provides the number of times it was repeated before removing the duplicate edges from the network obtained by combining Alonso networks and HumanBase networks.

Table 38: Duplicate regulatory interaction from TRRUST and RegNetwork

TF	TG	Number duplicate
SP1	MET	4
E2F1	CCNA2	3
E2F1	CCNE1	2
E2F1	CCNE2	2
E2F1	DHFR	2
E2F1	NPAT	2
E2F1	POLA1	2
HIF1A	CDKN1B	2
HIF1A	PDGFA	2
HIF1A	TIMP1	2
HIF1A	VEGFC	2
NFE2L2	TXNRD1	2
SP1	TYMS	2
STAT1	VEGFC	2
TFAP2A	ITPR1	2
YY1	TOP3A	2

The table gives the edges that are repeated in the network obtained after merging the network from TRRUST and RegNetwork databases. The 1st column represents the TF. The 2nd column the TG. The 3rd column informs for each edge, its number of repetitions.

Table 39: Edges duplicated in our mouse regulatory network

TF	TG	# dup
ENSMUSP0000001326	ENSMUSP00000079324	4
ENSMUSP0000001326	ENSMUSP00000111102	4
ENSMUSP0000001326	ENSMUSP00000111103	4
ENSMUSP0000001326	ENSMUSP00000117856	4
ENSMUSP0000001326	ENSMUSP00000118755	4
ENSMUSP0000001326	ENSMUSP00000121923	4
ENSMUSP00000126143	ENSMUSP00000079324	4
ENSMUSP00000126143	ENSMUSP00000111102	4
ENSMUSP00000126143	ENSMUSP00000111103	4
ENSMUSP00000126143	ENSMUSP00000117856	4
ENSMUSP00000126143	ENSMUSP00000118755	4
ENSMUSP00000126143	ENSMUSP00000121923	4
ENSMUSP00000127445	ENSMUSP00000079324	4
ENSMUSP00000127445	ENSMUSP00000111102	4
ENSMUSP00000127445	ENSMUSP00000111103	4
ENSMUSP00000127445	ENSMUSP00000117856	4
ENSMUSP00000127445	ENSMUSP00000118755	4
ENSMUSP00000127445	ENSMUSP00000121923	4
ENSMUSP00000127714	ENSMUSP00000079324	4
ENSMUSP00000127714	ENSMUSP00000111102	4

Table 39 continued from previous page

TF	TG	# dup
ENSMUSP00000127714	ENSMUSP00000111103	4
ENSMUSP00000127714	ENSMUSP00000117856	4
ENSMUSP00000127714	ENSMUSP00000118755	4
ENSMUSP00000127714	ENSMUSP00000121923	4
ENSMUSP00000129638	ENSMUSP00000079324	4
ENSMUSP00000129638	ENSMUSP00000111102	4
ENSMUSP00000129638	ENSMUSP00000111103	4
ENSMUSP00000129638	ENSMUSP00000117856	4
ENSMUSP00000129638	ENSMUSP00000118755	4
ENSMUSP00000129638	ENSMUSP00000121923	4
ENSMUSP00000130747	ENSMUSP00000079324	4
ENSMUSP00000130747	ENSMUSP00000111102	4
ENSMUSP00000130747	ENSMUSP00000111103	4
ENSMUSP00000130747	ENSMUSP00000117856	4
ENSMUSP00000130747	ENSMUSP00000118755	4
ENSMUSP00000130747	ENSMUSP00000121923	4
ENSMUSP00000000894	ENSMUSP00000029270	3
ENSMUSP00000000894	ENSMUSP00000118239	3
ENSMUSP00000000894	ENSMUSP00000142946	3
ENSMUSP00000099434	ENSMUSP00000029270	3
ENSMUSP00000099434	ENSMUSP00000118239	3
ENSMUSP00000099434	ENSMUSP00000142946	3
ENSMUSP00000000894	ENSMUSP00000006856	2
ENSMUSP00000000894	ENSMUSP00000022218	2
ENSMUSP00000000894	ENSMUSP00000029866	2
ENSMUSP00000000894	ENSMUSP00000048709	2
ENSMUSP00000000894	ENSMUSP00000103658	2
ENSMUSP00000000894	ENSMUSP00000103960	2
ENSMUSP00000000894	ENSMUSP00000117662	2
ENSMUSP00000000894	ENSMUSP00000130693	2
ENSMUSP00000000894	ENSMUSP00000145532	2
ENSMUSP00000001326	ENSMUSP00000026846	2
ENSMUSP00000001326	ENSMUSP00000123377	2
ENSMUSP00000001326	ENSMUSP00000142970	2
ENSMUSP00000001326	ENSMUSP00000143001	2
ENSMUSP00000001326	ENSMUSP00000143540	2
ENSMUSP00000001326	ENSMUSP00000143552	2
ENSMUSP00000021530	ENSMUSP00000003115	2
ENSMUSP00000021530	ENSMUSP00000009530	2
ENSMUSP00000021530	ENSMUSP00000033919	2
ENSMUSP00000021530	ENSMUSP00000038870	2
ENSMUSP00000021530	ENSMUSP00000065832	2
ENSMUSP00000021530	ENSMUSP00000075463	2
ENSMUSP00000021530	ENSMUSP00000106521	2
ENSMUSP00000021530	ENSMUSP00000106522	2
ENSMUSP00000021530	ENSMUSP00000110999	2
ENSMUSP00000021530	ENSMUSP00000145056	2

Table 39 continued from previous page

TF	TG	# dup
ENSMUSP00000021530	ENSMUSP00000148210	2
ENSMUSP00000021692	ENSMUSP00000002891	2
ENSMUSP00000021692	ENSMUSP00000099729	2
ENSMUSP00000021692	ENSMUSP00000113057	2
ENSMUSP00000021692	ENSMUSP00000113653	2
ENSMUSP00000021692	ENSMUSP00000115727	2
ENSMUSP00000021787	ENSMUSP00000032192	2
ENSMUSP00000021787	ENSMUSP00000144880	2
ENSMUSP00000021787	ENSMUSP00000145177	2
ENSMUSP00000021787	ENSMUSP00000145339	2
ENSMUSP00000021787	ENSMUSP00000145522	2
ENSMUSP00000021787	ENSMUSP00000145526	2
ENSMUSP00000021787	ENSMUSP00000148284	2
ENSMUSP00000066743	ENSMUSP00000033919	2
ENSMUSP00000066743	ENSMUSP00000148210	2
ENSMUSP00000073041	ENSMUSP00000085581	2
ENSMUSP00000099434	ENSMUSP00000006856	2
ENSMUSP00000099434	ENSMUSP00000022218	2
ENSMUSP00000099434	ENSMUSP00000029866	2
ENSMUSP00000099434	ENSMUSP00000048709	2
ENSMUSP00000099434	ENSMUSP00000103658	2
ENSMUSP00000099434	ENSMUSP00000103960	2
ENSMUSP00000099434	ENSMUSP00000117662	2
ENSMUSP00000099434	ENSMUSP00000130693	2
ENSMUSP00000099434	ENSMUSP00000145532	2
ENSMUSP00000099733	ENSMUSP00000020484	2
ENSMUSP00000099733	ENSMUSP00000151409	2
ENSMUSP00000099733	ENSMUSP00000151629	2
ENSMUSP00000099733	ENSMUSP00000151825	2
ENSMUSP00000099733	ENSMUSP00000152046	2
ENSMUSP00000105822	ENSMUSP00000032192	2
ENSMUSP00000105822	ENSMUSP00000144880	2
ENSMUSP00000105822	ENSMUSP00000145177	2
ENSMUSP00000105822	ENSMUSP00000145339	2
ENSMUSP00000105822	ENSMUSP00000145522	2
ENSMUSP00000105822	ENSMUSP00000145526	2
ENSMUSP00000105822	ENSMUSP00000148284	2
ENSMUSP00000106088	ENSMUSP00000003115	2
ENSMUSP00000106088	ENSMUSP00000009530	2
ENSMUSP00000106088	ENSMUSP00000033919	2
ENSMUSP00000106088	ENSMUSP00000038870	2
ENSMUSP00000106088	ENSMUSP00000065832	2
ENSMUSP00000106088	ENSMUSP00000075463	2
ENSMUSP00000106088	ENSMUSP00000106521	2
ENSMUSP00000106088	ENSMUSP00000106522	2
ENSMUSP00000106088	ENSMUSP00000110999	2
ENSMUSP00000106088	ENSMUSP00000145056	2

Table 39 continued from previous page

TF	TG	# dup
ENSMUSP00000106088	ENSMUSP00000148210	2
ENSMUSP00000106091	ENSMUSP00000003115	2
ENSMUSP00000106091	ENSMUSP00000009530	2
ENSMUSP00000106091	ENSMUSP00000033919	2
ENSMUSP00000106091	ENSMUSP00000038870	2
ENSMUSP00000106091	ENSMUSP00000065832	2
ENSMUSP00000106091	ENSMUSP00000075463	2
ENSMUSP00000106091	ENSMUSP00000106521	2
ENSMUSP00000106091	ENSMUSP00000106522	2
ENSMUSP00000106091	ENSMUSP00000110999	2
ENSMUSP00000106091	ENSMUSP00000145056	2
ENSMUSP00000106091	ENSMUSP00000148210	2
ENSMUSP00000122403	ENSMUSP00000043909	2
ENSMUSP00000126143	ENSMUSP00000026846	2
ENSMUSP00000126143	ENSMUSP00000123377	2
ENSMUSP00000126143	ENSMUSP00000142970	2
ENSMUSP00000126143	ENSMUSP00000143001	2
ENSMUSP00000126143	ENSMUSP00000143540	2
ENSMUSP00000126143	ENSMUSP00000143552	2
ENSMUSP00000127445	ENSMUSP00000026846	2
ENSMUSP00000127445	ENSMUSP00000123377	2
ENSMUSP00000127445	ENSMUSP00000142970	2
ENSMUSP00000127445	ENSMUSP00000143001	2
ENSMUSP00000127445	ENSMUSP00000143540	2
ENSMUSP00000127445	ENSMUSP00000143552	2
ENSMUSP00000127714	ENSMUSP00000026846	2
ENSMUSP00000127714	ENSMUSP00000123377	2
ENSMUSP00000127714	ENSMUSP00000142970	2
ENSMUSP00000127714	ENSMUSP00000143001	2
ENSMUSP00000127714	ENSMUSP00000143540	2
ENSMUSP00000127714	ENSMUSP00000143552	2
ENSMUSP00000129638	ENSMUSP00000026846	2
ENSMUSP00000129638	ENSMUSP00000123377	2
ENSMUSP00000129638	ENSMUSP00000142970	2
ENSMUSP00000129638	ENSMUSP00000143001	2
ENSMUSP00000129638	ENSMUSP00000143540	2
ENSMUSP00000129638	ENSMUSP00000143552	2
ENSMUSP00000130747	ENSMUSP00000026846	2
ENSMUSP00000130747	ENSMUSP00000123377	2
ENSMUSP00000130747	ENSMUSP00000142970	2
ENSMUSP00000130747	ENSMUSP00000143001	2
ENSMUSP00000130747	ENSMUSP00000143540	2
ENSMUSP00000130747	ENSMUSP00000143552	2
ENSMUSP00000139746	ENSMUSP00000033919	2
ENSMUSP00000139746	ENSMUSP00000148210	2
ENSMUSP00000140482	ENSMUSP00000033919	2
ENSMUSP00000140482	ENSMUSP00000148210	2

Table 39 continued from previous page

TF	TG	# dup
ENSMUSP00000140518	ENSMUSP00000033919	2
ENSMUSP00000140518	ENSMUSP00000148210	2
ENSMUSP00000140643	ENSMUSP00000033919	2
ENSMUSP00000140643	ENSMUSP00000148210	2
ENSMUSP00000140875	ENSMUSP00000033919	2
ENSMUSP00000140875	ENSMUSP00000148210	2
ENSMUSP00000141125	ENSMUSP00000033919	2
ENSMUSP00000141125	ENSMUSP00000148210	2
ENSMUSP00000141132	ENSMUSP00000033919	2
ENSMUSP00000141132	ENSMUSP00000148210	2
ENSMUSP00000141144	ENSMUSP00000033919	2
ENSMUSP00000141144	ENSMUSP00000148210	2
ENSMUSP00000153149	ENSMUSP00000032192	2
ENSMUSP00000153149	ENSMUSP00000144880	2
ENSMUSP00000153149	ENSMUSP00000145177	2
ENSMUSP00000153149	ENSMUSP00000145339	2
ENSMUSP00000153149	ENSMUSP00000145522	2
ENSMUSP00000153149	ENSMUSP00000145526	2
ENSMUSP00000153149	ENSMUSP00000148284	2
ENSMUSP00000153271	ENSMUSP00000032192	2
ENSMUSP00000153271	ENSMUSP00000144880	2
ENSMUSP00000153271	ENSMUSP00000145177	2
ENSMUSP00000153271	ENSMUSP00000145339	2
ENSMUSP00000153271	ENSMUSP00000145522	2
ENSMUSP00000153271	ENSMUSP00000145526	2
ENSMUSP00000153271	ENSMUSP00000148284	2
ENSMUSP00000153522	ENSMUSP00000032192	2
ENSMUSP00000153522	ENSMUSP00000144880	2
ENSMUSP00000153522	ENSMUSP00000145177	2
ENSMUSP00000153522	ENSMUSP00000145339	2
ENSMUSP00000153522	ENSMUSP00000145522	2
ENSMUSP00000153522	ENSMUSP00000145526	2
ENSMUSP00000153522	ENSMUSP00000148284	2
ENSMUSP00000153667	ENSMUSP00000032192	2
ENSMUSP00000153667	ENSMUSP00000144880	2
ENSMUSP00000153667	ENSMUSP00000145177	2
ENSMUSP00000153667	ENSMUSP00000145339	2
ENSMUSP00000153667	ENSMUSP00000145522	2
ENSMUSP00000153667	ENSMUSP00000145526	2
ENSMUSP00000153667	ENSMUSP00000148284	2
ENSMUSP00000153667	ENSMUSP00000145522	2
ENSMUSP00000153667	ENSMUSP00000145526	2
ENSMUSP00000153667	ENSMUSP00000148284	2

The table gives the edges that are repeated in the Mouse regulatory network after merging networks from TRRUST, RegNetwork and STRINGDB databases. The 1st column represents the TF. The 2nd column the TG. The 3rd column informs for each edge, its number of repetitions.

C.2 Other

This section gives more details on **BENIN**'s execution time on the ENCS speed cluster. Table 40 gives **BENIN**'s execution time on different network sizes, and without integrating any prior knowledge data. We performed all the computations on the ENCS speed cluster. It has sixteen, 32-core nodes, each with 512 GB of memory and approximately 1 TB of volatile-scratch disk space. We requested 25 cores for all computations. Note that size 10 and 100 networks are obtained from DREAM4 challenge data, and the size 628 network is the human network data.

Table 40: **BENIN** execution time on different network sizes

# Genes	# TFs	# Time points	Execution time
10	8	105	125s
100	41	210	931s
628	54	48	5335s

The table reports **BENIN**'s execution time on a ENCS speed cluster, for different network sizes and different expression datasets. The speed cluster has sixteen, 32-core nodes, each with 512 GB of memory and approximately 1 TB of volatile-scratch disk space. We requested 25 cores for all computations. Note that size 10 and 100 networks are obtained from DREAM4 challenge data, and the size 628 network is the human network data. The 1st column gives the number of genes in the network/dataset. The 2nd column gives the number of TFs in the network/dataset. The 3rd column gives the number of time points in the time-series expression dataset. The 4th column gives **BENIN** execution time without considering any prior knowledge data.