UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXX CICLO - 2017

A DATA FUSION APPROACH FOR LEARNING TRANSCRIPTIONAL BAYESIAN NETWORKS IN CHRONIC LEUKEMIA

PhD Thesis by
ELISABETTA SAUTA

Advisor:
Prof. Riccardo Bellazzi

PhD Program Chair:
Prof. Riccardo Bellazzi

Abstract (Italiano)

La crescente disponibilità di dati omici ha determinato un importante cambiamento nel paradigma della ricerca scientifica, passando da uno studio “contesto specifico” focalizzato su un singolo aspetto biologico, ad un studio su larga scala guidato dai dati. L’analisi simultanea di diversi livelli omici potrebbe aiutare a chiarire la relazione tra caratteristiche o perturbazioni del sistema molecolare non rilevate in precedenza con un fenotipo specifico, specialmente nel caso di malattie complesse, come il cancro. A tal fine, un approccio computazionale integrativo in grado di gestire l'eterogeneità dei dati e la complessità biologica può consentire un'indagine approfondita di programmi di espressione genica disregolati responsabili dei meccanismi di insorgenza e di progressione della malattia. La ricostruzione dei pattern regolatori dei fattori determinanti della trascrizione (fattori di trascrizione, TF), che presiedono allo schema di espressione genica, potrebbe anche aiutare a ottenere informazioni sulle firme molecolari che guidano i fenotipi della malattia, offrendo così nuove ipotesi di ricerca.

In questa tesi è stato sviluppato un approccio di “data fusion”, incentrato sull'integrazione a più livelli di dati omici per la modellizzazione di background trascrizionali su larga scala. La sua strategia di ricerca combina efficacemente un approccio network-centrico per ricostruire l'interattoma trascrizionale con la modellizzazione offerta dalla teoria Bayesiana, ed è in grado di indagare probabilisticamente, su scala genomica, le regolazioni trascrizionali e le sottostanti firme molecolari.

Questo lavoro di ricerca fa parte del progetto "Rete Ematologica Lombarda (REL) cluster biotecnologico per l'implementazione dell'analisi genomica e lo sviluppo di trattamenti innovativi nelle neoplasie ematologiche", che mira a stabilire un centro di riferimento per lo studio delle neoplasie ematologiche, con particolare attenzione alle neoplasie mieloidi.

La metodologia proposta è stata infatti applicata ad un tipo di patologia mieloide, la leucemia mieloide cronica (LMC), di cui è noto l’evento genetico causale, ma l’alterato ruolo trascrizionale alla base della progressione della malattia non è stato ancora approfondito a livello genomico.

Abstract (English)

The increasing availability of omics data has caused an important paradigmatic shift in scientific research from case-based studies towards large scale data-driven research. The simultaneous interrogation of different omics levels, could help to elucidate the interrelation of previously-undetected system features or perturbations with a specific phenotype, especially in complex diseases, such as cancer. To this aim, an integrative computational approach able to deal with data heterogeneity and biological complexity may allow a deep investigation of dysregulated gene expression programs responsible of disease onset and progression mechanisms. The reconstruction of transcriptional determinants (transcription factors, TFs) regulatory patterns, which preside over the gene expression scheme could also help to gain insights into molecular signatures driving disease phenotypes, offering new research hypotheses.

In this thesis, I have developed a data fusion approach focused on "multi-layered" omics data integration for modeling large-scale transcriptional background. Its framework efficiently combines a network-centric approach to reconstruct the transcriptional interactome to probabilistically inspect, on a genome-wide scale, the transcriptional regulations and the underlying regulative signatures.

This work is part of the project "*Rete Ematologica Lombarda (REL) biotechnology cluster for the implementation of genomic analysis and the development of innovative treatments in hematological malignancies*", which aims at establishing a reference center for the study of hematological malignancies, with focus on myeloid neoplasms.

The proposed methodology has been applied to the case of a myeloid disorder, the Chronic Myeloid Leukemia (CML), whose causative genetic event is known but its emerging transcriptional altered role in disease progression has not yet been deeply investigated at a genomic level.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This first introductory chapter focuses on the importance of data integration in the biomedical research, which has been revolutionized in the last years by the advance of high-throughput technologies and the resulting increase of omics data volume. The following sections go back over the main steps of this phenomenon, describing the challenging characteristics of this type of data. Moreover, the need of exploiting computational integrative models for data organization and interpretation, which may help to translate novel biological knowledge into improved diseases understanding, will be also discussed.

The final aim of this chapter is to provide an overview of the key concepts on which the methodology proposed in this PhD thesis is focused.

1.1 The Omics revolution

Over the past decade, the development of next generation sequencing (NGS) technologies has considerably expanded the biological knowledge at molecular level, providing the possibility to study the underlying mechanisms involved in human disease or human health processes on genome scale. This so-called "O*mics revolution*" enabled the investigation of biological systems through massively parallel sequence acquisition or simultaneous molecular measurements, providing a holistic description of the considered cellular phenomenon [1–3].

Omics is a term used to indicate all data gathered from high-throughput techniques, and each omics research field (or domain) specifies the category of experimental data to which it refers [4]. The most important domains are briefly described below.

- *Genomics*, which concerns the study of organisms' whole genome and its genetic variations.

- *Transcriptomics*, which enables the genome-wide assessment of gene expression patterns in cells and tissues.
- *Proteomics,* which deals with the study of proteins and their molecular modifications, trying to assess the cellular levels of each protein encoded in the genome.
- *Epigenomics*, which focuses on genome-wide characterization of reversible modifications of DNA (which does not change its sequence) or DNA-associated proteins, such as histones and transcription factors, with the aim to understand the regulations of the gene expression.
- *Metabolomics,* which simultaneously quantifies multiple small molecule types (metabolites), produced by cellular metabolic functions.

The continued progression of new sequencing technologies has encouraged to develop large-scale sequencing projects, such as 1000 Genomes Project [5], The Cancer Genome Atlas (TCGA) [6], the Encyclopedia of DNA Elements (ENCODE) [7], and other big genomics projects reported in Fig. 1.1.

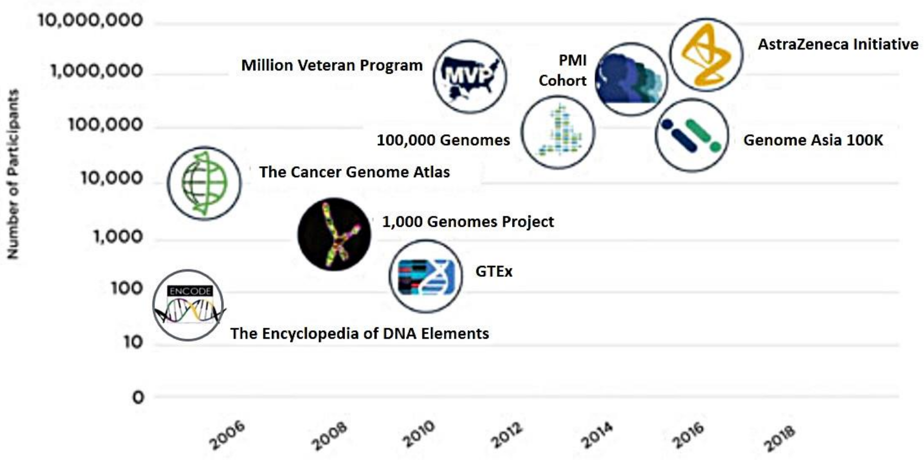


Figure 1.1: Big genomics projects diffusion over the last decade.
Source: adapted from Brandi D. [8]

The international 1000 Genomes Project is a government backed initiative launched in 2008 that aims to sequence the entire genome of thousands of people from around the world and it is continuing to grow as the largest worldwide data set on human genetic variation.

The US-funded TCGA instead contains cancer genomic data from 33 different tumor types and clinical data from more than 11,000 patients. Another ongoing project is ENCODE, a consortium found in 2003 by National Human Genome Research Institute (NHGRI), whose main objective is to map and characterize all functional elements within the genome, using different omics approaches applied not only to the human genome but also to genomes of several model organisms.

The spread of big genomics projects determined an increasing volume of sequencing data (see Fig. 1.2), which is anticipated to exceed 2 exabytes (2 million terabytes) by 2025 [8,9].

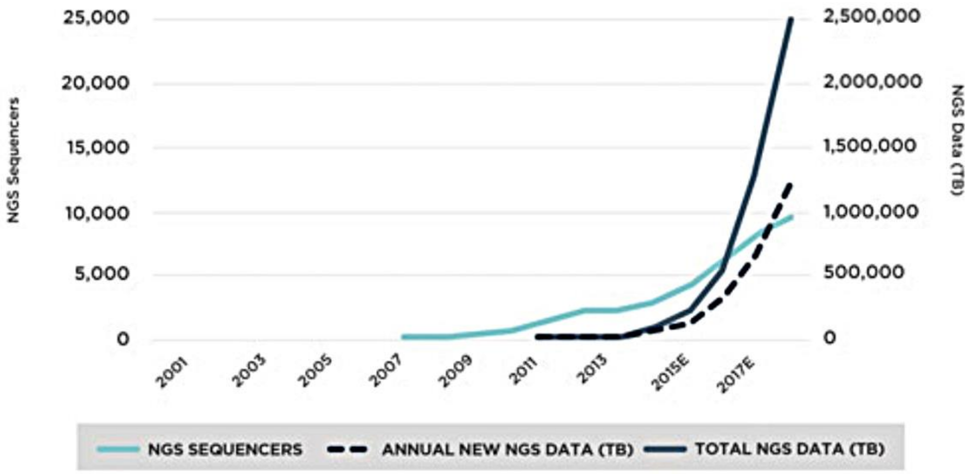


Figure 1.2: The growth of NGS data in the last decade. The chart represents the exponential increase of sequencing data for genomics research.
Source: Brandi D. [8]

The resulting increased availability of biological data and clinical data, generated at unprecedented speed and scale, led the biomedical research to the realm of Big Data [10]. This widely accepted definition encloses four important features, commonly known as the 4 Vs: *Volume* of produced data, *Velocity*, the measure of how fast the data is generated, *Variety* of data sources and time scales from which data are collected, and *Veracity*, the uncertainty characterizing the data quality. Within the Omics era of life sciences, all of biological data produced from high-throughput techniques can be defined as Big "Omics" Data [11].

1.2 Big Omics data integration challenge

The increasing availability of big omics data determined an important paradigm shift in scientific research from case-based studies toward large scale data-driven research. Whilst this phenomenon has revolutionized biomedical studies, the intrinsic omics data structure determined by various biological principles and experiment designs raises challenging characteristics in addition to the aforementioned 4Vs [11], reported in Table 1.1 below.

Table 1.1: Big Omics data characteristics

Big Omics data characteristics	Description
Hierarchical	Data is generated at different biological levels ranging from molecules, cell tissues to systems
(Highly) Heterogeneous	Data is produced using different omics methods and the resulting datasets differ in size, format and dimensionality
Complex	Data can be recorded as multi-level information obtained simultaneously from over thousands of molecules
Dynamic	Data provides only a snapshot of biological processes or states that change with conditions and over time

This table summarizes big biological data inherent characteristics.
Source: Adapted from Li Y, Chen L [11]

Given these properties, the main challenge is translating the driving force or causal relationship among biological elements depicted by omics data, into meaningful knowledge of clinical relevance, helping to decipher the mechanisms of biological processes and complex diseases, such as cancer.

From this perspective, interrogating more different omics levels instead a single layer could help to elucidate the genotype and phenotype interrelation and the combined influence on disease onset and progression [12]. For example, the integration of genomics and proteomics data from brain tumor tissues has allowed the identification of biomarker signatures, resulting in a better diagnosis accuracy, as demonstrated by Petrik et al. [13]. Other studies, like the one presented by Sohal et al., demonstrated the viability of integrating genomic data collected from different laboratories and public databases, such as the NCBI's Gene Expression Omnibus (GEO) [14], discovering common gene expression signature characteristics of cells involved in leukemia processes.

It clearly emerges that the integration process of heterogeneous omics data, or so-called *omics data fusion*, becomes a key point for biomedical research to capture previously-undetected system features or perturbations within a pathological scenario [15]. Moreover, in the context of precision medicine, combining genome-scale molecular data with patients-specific clinical information can shed a light on diseases molecular process, and on novel biomarkers discovery, improving molecular-targeted diagnosis and personalized therapeutics (see Fig. 1.3) [16].

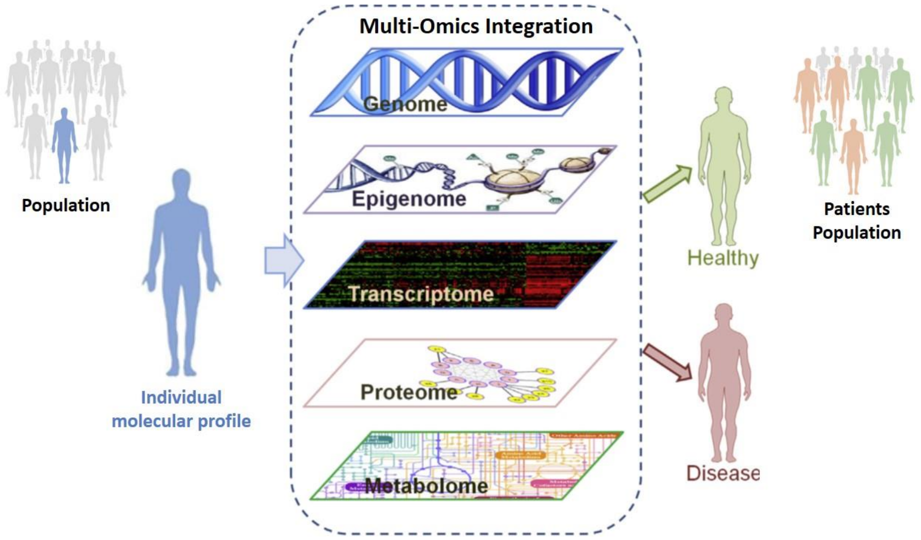


Figure 1.3: Conceptual model of multi-omics data integration for precision medicine. Source: adapted from Sun Y. et al [12]

In this challenging scenario, Systems Biology provides a new way for system-wide study exploiting heterogeneous data integration, to discover coherent biological signatures underlying data and to predict phenotypic outcomes.

1.3 Systems Biology for Omics data integration

Biological regulation is the results of a structured multi-dimensional circuits of relations among biological entities (i.e. genes, proteins, or metabolites), whose functions redundancy, driving cooperation or competition are the hallmarks of system complexity and robustness to external environment. Describing such composite system through its inner components provides a representation of the global entity, whose important properties can be missed by analyzing its elements separately.

The research field of Systems Biology (SB) revolves around this main concept, considering that the phenotype of any individual organism is the reflection of the simultaneous multitude of molecular interactions combined in a holistic manner to produce such a phenotype [17]. Its final objective is to mathematically model biological systems to describe their structure, dynamics and changes after perturbations, trying to simulate the outcome responses for a given input stimulus.

An emerging SB branch is **network biology**, whose approach consists to emphasize intracellular molecular interactions, translating them into mathematically well-defined networks. A key characteristic is the possibility to integrate data from heterogeneous sources, mapping omics data onto biological networks, enhancing the data with the connectivity information encoded within the network architecture.

The use of network biology as an integrative approach is considerably grown over the last ten years, following also the big omics data spread [18,19], as demonstrated by the number of related publications, reported in Fig. 1.4.

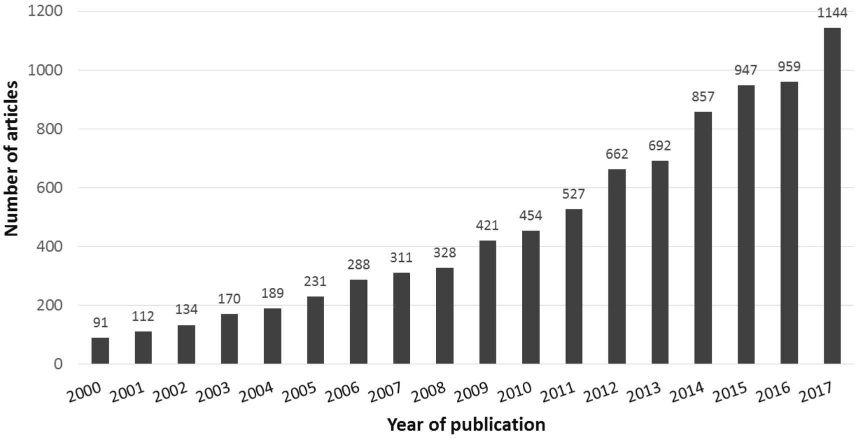


Figure 1.4: Network Biology as integrative approach related publications. Data are extracted from PubMed using the query ‘network’ AND ‘integration’ for title and abstract word of published publications.

Whilst network biology offers a natural scaffold upon which omics data can be integrated, the development of bioinformatics pipelines to support this integration, applying appropriate standards and quality controls-metrics on this noisy data is fundamental. The reconstruction of hundreds to thousands of molecular interrelating relationships, which globally constitute an interactome, requires robust computational methods to scale and investigate such complexity [18], in order to prioritize novel biological hypotheses generated from data for experimental validation.

Combining a network-based integrative approach with computational modeling could indeed reveal crucial mechanisms of regulation presiding over physiological functions and their dysregulated counterpart in disease. Within this perspective, a deeply investigation of transcriptional interactions which serve as convergence points of oncogenic and pathogenic signaling, could be a useful strategy, since under these regulations relies the first level signature of gene activities, whose expression patterns are altered by the disease.

Following these considerations, in this dissertation we present a multi-layered data fusion approach and the results of our research applied to the Chronic Myeloid Leukemia (CML) case-study. This PhD thesis is part of the project “Rete Ematologica Lombarda (REL) biotechnology cluster for the implementation of genomic analysis and the development of innovative treatments in hematological malignancies”, aimed to define the molecular basis of this type of cancers, with a specific focus on myeloid neoplasms.

Chronic myeloid leukemia is a myeloid disorder that originates in the bone marrow, and is caused by a specific mutation. Despite this important discovery allowed to develop a targeted therapy, the resistance to the approved treatment is a recurring phenomenon in an increasing proportion of patients. The treatment indeed does not eradicate cancer cells, which continue to progress within a landscape of unclear molecular mechanisms of the disease. To overcome these barriers, innovative integrative approaches are definitely needed, to investigate the multiplicity and complexity of genetic and epigenetic changes underlying the molecular cross-talk of signaling pathways, which may be altered by leukemia.

The developed approach exploits the complementarity of the information gained from omics experimental sources, and applied to the case-study, reconstructs, with a network-based approach, its transcriptional genomic interactome. In order to scale and infer this complexity, the data fusion method provides a Bayesian modeling and a simulating framework to investigate transcriptional signatures on a genome-wide scale.

1.4 Thesis Outline

The content of the thesis is organized as follows:

Chapter 2 gives a theory overview of the crucial points faced with the proposed data fusion approach, to provide all conceptual bases for a better comprehension of the developed method. The chapter gives a description of omics data sources useful for our aim, some background on biological networks and on the related properties, with a main focus on transcriptional network.

Since the approach is based on a Bayesian formalism, an explanation of the underlying Bayesian theory, which allowed the modeling and inference of the considered transcriptional context, is provided. Moreover, an outline of the Chronic Myeloid Leukemia disease background, to which the method has been applied, is depicted.

Chapter 3 all the steps of the proposed approach will be described in details, starting from a computational integrative analysis for reconstructing the transcriptional interactome, passing through its Bayesian modeling within a hybrid structure learning scheme, to obtain a probabilistic model which can be exploited to investigate transcriptional signatures.

Chapter 4 the robustness of the proposed method compared to other regulatory network reconstruction strategies, and the related results will be illustrated.

Chapter 5 the obtained results applying the methodology to the case of Chronic Myeloid Leukemia will be described.

Chapter 6 some concluding remarks, a final discussion on the results obtained and future directions of the work will be presented.

An **Appendix** will follow containing supplementary materials useful to better understand the results achieved by applying our approach.

Chapter 2

Background

The aim of this chapter is to give all the theoretical fundamentals for understanding the proposed data fusion approach applied to the transcriptional context of the case study.

The first paragraph presents the omics data source that can be used for extract useful information to reconstruct transcriptional landscapes. Since the implemented transcriptional modeling is based on a network-centric approach, the following section describes its biological principles and the topological measures derived from graph theory to investigate specific molecular patterns.

In the third part of the chapter, a brief overview of the transcriptional regulatory network reconstruction methods is provided, with a particular focus on the Bayesian formalism exploited to probabilistically inspect the reconstructed transcriptional model.

2.1 Omics sources for transcriptional background reconstruction

The vast amount of generated biological data in recent years has determined the necessity to store and organize experimental data within public repositories, and make them available for scientific community. Structuring the data-driven information with the existing knowledge, available on online databases, is fundamental to accurately reconstruct the molecular interactions behind the considered biological processes. To this aim, the translation of the experimental-derived information into functional meaningful associations inside specific pathways becomes a natural part of the integration process [20], and can expand the knowledge-based networks in the context of human disease.

Given the considerable number of existing databases publicly accessible online retrieving biological data and knowledge, as reported by Fernández-

Suárez X. et al [21] and by Pathguide resource [22], in the following sections, we distinguish between data-driven repositories, which store high-throughput data, and literature-derived repositories, covering the information on cellular signaling cascades and processes. The next databases description is focused on those which store biological information useful for supporting transcriptional background reconstruction tackled in this dissertation.

Within the context of data-driven databases, a brief description of the exploited omics data will be provided.

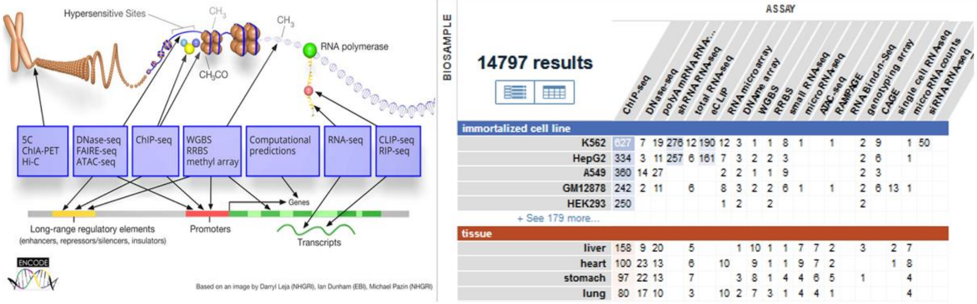
2.1.1 High-throughput data repositories

In the growing field of omics science, the storage and organization of high-throughput quality data became a very important goal to promote the reuse of archived data. Following specific submission pipelines and quality control metrics, every scientist or laboratory that, for example, belongs to a research project or consortium, can share and deposit its experimental data in these databases.

ENCODE

The international Encyclopedia of DNA Elements (ENCODE) consortium started in 2003 with the aim to identify all functional elements in the human genome, and then extended to model organisms. It is currently implemented in four phases, from the initial pilot phase focused on 1% of the genome, to the actual fourth one on more than 80% of the genome. To store and organize the vast amount of data produced by several research groups during these years, ENCODE has created a big project portal to freely access to the data, ensuring that specific quality standards are met before data releasing [23].

The database offers a wide-range of data from different high-throughput sequencing (seq, in short-form) techniques, as shown in Fig. 2.1, depending on the targeted biological process to explore. These included DNase-seq (Dnase I hypersensitive sites seq) for studying the DNA accessibility through its conformation; RNA-seq to investigate the expression of genes and in terms of transcripts abundance; RRBS (Reduced representation bisulfite seq), WGBS (whole-genome bisulfite seq) for studying the genome methylation state, or ChIP-seq to map physical binding events along the genome.


Hypersensitive Sites
CH3
CH3
CH3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
ENCODE
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
BIOSAMPLE
ASSAY

14797 results

	ChIP-seq	DNase-seq	polyA mRNA RNA-...	shRNA RNA-seq	total RNA-seq	eCLIP	RNA microarray	DNAme array	WGBS	RRBS	small RNA-seq	microRNA-seq	ATAC-seq	RAMPAGE	RNA Bind-n-Seq	genotyping array	CAGE	single cell RNA-seq	microRNA counts	siRNA RNA-seq
immortalized cell line																				
K562	627	7	19	276	12	190	12	3	1	1	8	1		1		2	9		1	50
HepG2	334	3	11	257	6	161	7	3	2	2	3					2	6		1	
A549	380	14	27				2	2	1	1	9					2	3			
GM12878	242	2	11		6		8	3	2	2	6	1		1		2	6	13	1	
HEK293	250						1	2		2						2				
+ See 179 more...																				
tissue																				
liver	158	9	20		5			1	10	1	1	7	7	2		3		2	7	
heart	100	23	13		6		10		9	1	1	9	7	2				1	8	
stomach	97	22	13		7			3	8	1	4	4	6	5		1			4	
lung	80	17	10		3		10	2	7	3	1	4	4	1					4	

Figure 2.1: ENCODE assays and data matrix snapshots from the project portal.

Data are obtainable in raw and processed format, correlated with its experimental metadata (i.e. the assay protocol, sample replicates details, the obtained files list), and quality tags (i.e. sequencing read length and depth, sample replicate concordance, inconsistencies in the analysis pipeline), for different types of tissues and cell lines, classified by availability of data e.g. cell line defined as *Tier 1* means that a prodigious volume of data are accessible with a remarkable potential for combinatorial and integrative analyses.

As highlighted by the numbers of available biosamples reported in Fig. 2.1, the most prominent and applied technique is Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq).

ChIP-seq is a mainstream method in genomics and epigenomics, which sequences genomic DNA fragments that co- precipitate with a DNA-binding protein that is under study (see Fig. 2.2), typically a molecule that, acting on a specific DNA portion, performs its function as regulator of the transcriptional process (a transcription factor, TF), or as chromatin-modifying enzyme (that can be both a TF or a histone), or components of the basal transcriptional machinery (RNA polymerase) [24]. The target molecule is identified through a specific antibody that allows to isolate the DNA portion bound to it. Each experiment is paired to a control sample, ("*input DNA sample*") in which no specific antibody is used, and typically consists of genomic DNA. This sample is necessary to establish the noisy background and estimate the fragments distribution in the absence of specific binding.

DNA fragments are then purified and sequenced as *reads*, which are then mapped onto the reference genome. The regions that are significantly enriched for ChIP-reads, can be detected using ad-hoc algorithms as *peaks*, through the comparison with input reads. This peaks along the genome represent candidate binding regions of the studied molecule.

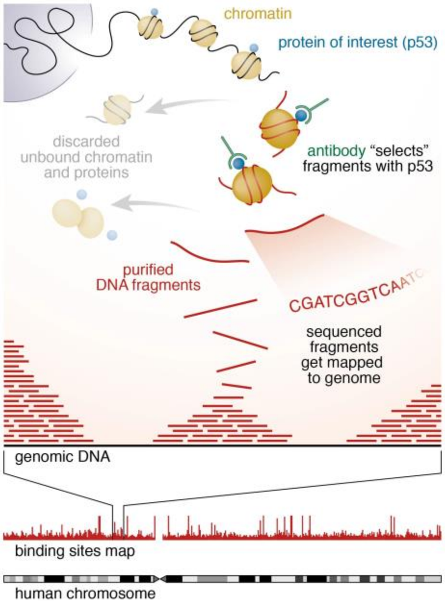

chromatin
protein of interest (p53)
discarded unbound chromatin and proteins
antibody "selects" fragments with p53
purified DNA fragments
CGATCGGTCAATC
sequenced fragments get mapped to genome
genomic DNA
binding sites map
human chromosome

Figure 2.2: Experimental overview of the ChIP-sequencing technology.
Source: Botcheva K., et al. [25]

This technology has the potential to identify all DNA segments in the genome physically associated with one of the aforementioned molecular targets, offering a genome-wide perspective of its biological binding events.

All the binding information gained from ChIP-Seq experiments targeting transcription factors has also been organized in another ENCODE-related database, Factorbook, providing a complete set of sequences features and structural DNA information around the genomic regions bound by all the considered TFs.

Furthermore, ChIP-seq results can also be integrated with other types of genomic assays, including gene expression, DNA methylation or chromatin conformation, to understand mechanisms of genomic functions from multiple aspects, leading to important discoveries related to disease-associated transcriptional regulation, tissue-specificity of epigenetic regulation, and chromatin organization [26].

NCBI – GEO

The Gene Expression Omnibus (GEO) is an international public repository built and managed by the National Center for Biotechnology Information (NCBI). It hosts and freely distributes more than 32,000 data series, comprising raw data, processed data and metadata which are indexed and cross-linked. The related submissions are deposited by individual laboratories, Data Coordinating Centers, or by microarray facilities on behalf of their clients [27]. As depicted in Fig. 2.3., GEO provides a wide range of archived data, comprising microarray technology, next-generation sequencing data, which exhibits a rapid increase since 2008, and other forms of high-throughput functional genomic data. Methods like ChIP-seq, included under 'genome binding/occupancy profiling by NGS' definition are increasing at higher rate than other NGS assays, highlighting their important impact on research community.

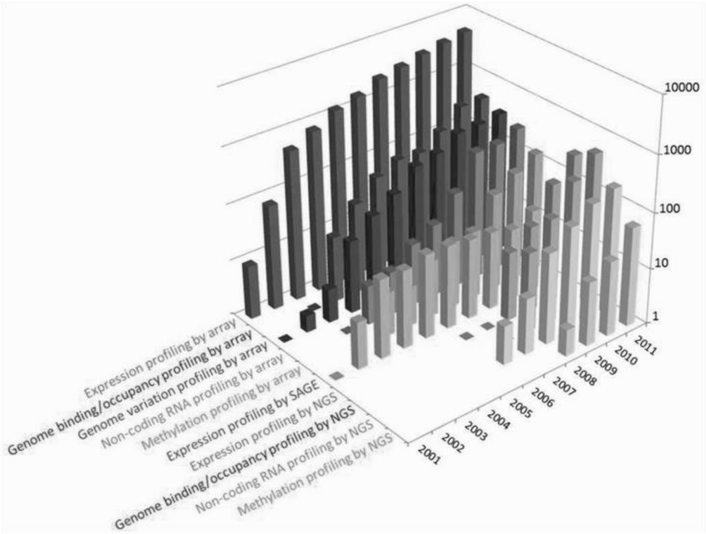


Figure 2.3: Distribution of the number and types of selected studies released by GEO each year since inception. Source: Barrett T., et al. [27]

Microarray assay described by 'Expression profiling by array' term represents the most common study type submitted to the database by an order of magnitude compared to the recent NGS techniques, although its growth rate is slowing. It is the first approach that made the transcriptomics analysis possible, and despite the increasingly turning to RNA-seq technology, remains a well-established approach for measuring gene expression levels, both in its static or dynamic profiling.

The first type of experimental design is a static sampling experiment where samples are collected from distinct biological groups without respect to time. The gene expression profiling in its dynamic scheme is instead a temporal experiment, where samples are collected over a time window to characterize temporal dynamic spectrum and underlying developmental or progressive biological mechanisms.

In both forms, the aim is profiling the entire repertoire of RNA transcripts of a cell or a tissue, globally defined as transcriptome.

The experimental starting point for a **microarray** is a set of short genomic DNA probes complementary to the cellular transcripts. As represented in Fig. 2.4., transcripts are extracted from samples of the biological site to be investigated, labeled with fluorescent dyes (either one color or two), washed to remove unbound sample, hybridized to the array, and scanned with a laser. Probes that match with transcribed RNA hybridize to their complementary target and emanate a fluorescent signal, which is detected by a scanner. Raw data consists in light signals whose intensity is then used as a measure of gene expression. This raw information can be retrieved in CEL format, which stores results of the intensity calculations.

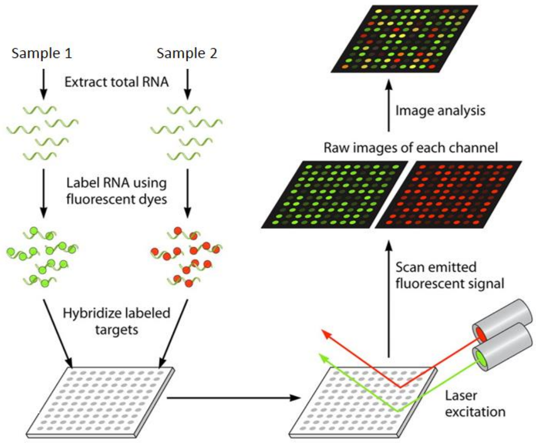

Sample 1
Sample 2
Extract total RNA
Label RNA using fluorescent dyes
Hybridize labeled targets
Laser excitation
Scan emitted fluorescent signal
Raw images of each channel
Image analysis

Figure 2.4: Workflow summary of microarray experimental procedure.
Source: Miller M.B., et al. [28]

During these steps, some artifacts can be generated, due to the experimental variability that affects the procedure. The advance in the identification of such biases has led to the development of both quality control standards and computational methods to deal with systematic variation, ensuring well-performed microarray experiments.

Instead of using molecular hybridization to capture transcript molecules of interest, RNA-seq samples transcripts present in the starting biological material by direct sequencing. Transcript sequences are then mapped back to a reference genome and counted to assess the level of each expressed gene in the genome. Despite the depth of the information gained with this technique, its relative high cost, the inherent heterogeneity, and the lacking of appropriate and defined standards, which are currently being established, microarray continues to remain a widely used tool to reconstruct transcriptomics profiles [29].

EMBL-EBI – ArrayExpress

The ArrayExpress Archive of Functional Genomics Data is an international functional genomics database hosted by the European Bioinformatics Institute (EMBL-EBI). It archives data from over 7,000 public sequencing and 42,000 array-based studies comprising over 1.5 million assays in total [30]. Alongside GEO, ArrayExpress is regarded by major journals as an important extended high-throughput data repository.

Data available for download is represented in a structured and standardized MAGE-TAB (MicroArray Gene Expression Tabular) format, where investigation design, array descriptions, and processed data are described. This format also facilitates linking to open source analysis environments such as Bioconductor [31]. For sequencing data, ArrayExpress stores raw data to the European Nucleotide Archive (ENA) [32], processed data (e.g. gene expression levels) and its metadata, describing the sample properties and the experimental design, are instead available on the ArrayExpress portal.

To facilitate reproducible research, data compliance is promoted using Minimum Information About a Microarray Experiment (MIAME) or Minimum Information about Sequencing Experiment (MINSEQE) guidelines. In this way, each submission is automatically scored by these criteria allowing users to quickly recognize high-quality data sets.

2.1.2 Literature-curated repositories

An important goal of the research once the network interactions are achieved, is that the reconstruction gives rise to a signaling pathway in a biologically consistent and meaningful manner so as to allow the mathematical analysis of the emerging properties of the network. This expansion of the knowledge-based network toward a better comprehension of leading complex diseases mechanisms can be supported by current understanding on cellular signaling systems.

A variety of repositories containing information on cell signaling pathways have been developed in conjunction with methodologies to access and analyze the data for getting insights on biological dynamics. The main pathway annotation databases are Reactome, Kyoto Encyclopedia of Genes and Genomes (KEGG), WikiPathways, Nature Pathway Interaction Database (PID), Pathway Commons and Gene Ontology [33,34].

In the following section, a description of Reactome, indicated as one of the most complete and curated pathway databases, and of the Gene Ontology Consortium, a comprehensive resource regarding the functions of genes and gene products are provided.

Reactome

The Reactome Pathway Knowledgebase provides reactions for any type of biological process, ranging from signal transduction, cellular transport, DNA replication and metabolism, organizing them as an ordered hierarchical network of molecular transformations. It has entries for 10,719 human genes, and supports the annotation of 24,704 specific forms of proteins, providing a coverage for 22 non-human species. Pathways are presented as chains of chemical reactions and the data model is based on classes, such as event or physical entity, with given properties (e.g. type of molecular interaction). Physical entities comprise proteins, RNA, DNA, small molecules and molecular complexes. An event, instead, can be either a *ReactionLikeEvent*, representing all reactions that convert an input into an output, or a *PathwayLikeEvent*, which groups together several related events. Cross references to several external databases are provided, and every two years all the information is reviewed, to keep it updated.

Reactome can be directly browsed or queried by text, or using ad-hoc tools through a web interface, or programmatically accessed, allowing data download for its visualization and analysis.

Gene Ontology (GO)

The Gene Ontology (GO) Consortium represents the most complete resource currently available for computable knowledge concerning the functions of genes and their products. It was established in 2000 to provide a controlled vocabulary for annotating homologous gene and protein sequences in different organisms.

Its knowledgebase is characterized by two main components. The first is the Gene Ontology (GO), which provides the logical structure of the biological functions (‘*GO terms*’), each one classified with a unique identifiers, and how these functions are related to each other (‘*relations*’), displayed as a directed acyclic graph.

GO relations are represented using a graph-based terminology, and *node* is used to refer to GO terms.

To define the relationships among nodes, a parent-child framework has been implemented, where a *parent* refers to the node closer to the root(s) of the graph, and *child* to that closer to the leaf nodes, as represented in Fig. 2.5 below.

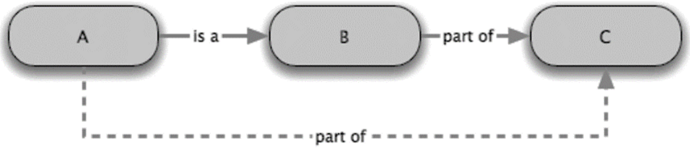


Figure 2.5: Graph-based representation example of GO relations. The arrowhead indicates the direction of the relationship; dotted line represents an inferred relationship, i.e. one that has not been expressly stated. The formal mathematical/logical representation of the inference made in the graph above would be “*is a part of→part of*”.
Source: http://www.geneontology.org/page/ontology-relations

The second component is the corpus of GO annotations, evidence-based statements relating a specific gene product (a protein, non-coding RNA, or macromolecular complex) to a specific ontology term. Each annotation is characterized by a unique identifier, and is linked to the evidence supporting that biological conclusion, typically a specific publication from the biomedical literature.

The GO describes functions considering three interrelating perspectives:

- *molecular function* (MF), which refers to the molecular-level of activities performed by gene products,
- *cellular component* (CC), describing the locations relative to cellular structures in which a gene product performs its function,
- *biological process* (BP), the larger processes, also defined as ‘biological program’ accomplished by multiple molecular activities

The number of terms and relationships categorized in these three aspects of the Gene Ontology are reported in Table 2.1 below.

Table 2.1. Gene Ontology annotated terms statistics

Aspects	Terms (classes)	Relationships
Molecular function	10,417	14,039
Cellular Component	4,022	7,854
Biological Process	29,146	71,372

Latest data as at October 2016; *Source:* The Gene Ontology Consortium [35]

Currently, the GO knowledgebase includes experimental findings from almost 140,000 published papers, represented as over 600,000 experimentally-supported GO annotations. These provide the core dataset for additional inference of over 6 million functional annotations for a diverse set of organisms spanning the tree of life [35].

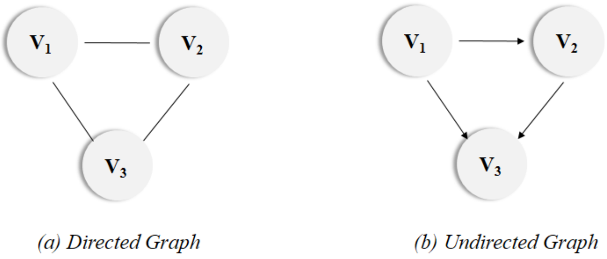
2.2 Basic concepts of Biological Networks

Availability of biomedical pathways and networks based on large-scale data gathering through diverse omics data sources offers new opportunities to explain the causality of relationships among biological entities, unraveling disease mechanisms [17,20,36]. Within this network-centric framework, the starting point is to use the mathematical concept of *graph* for representing omics layers as a *network*, and topological measures, belonging to the graph theory, to identify valuable biological properties.

In the first part of this section, the empirical and the mathematical description of graphs, that represent networks, are introduced with some of the basic definitions behind graph theory, useful to study network structure. Then, a brief description of the main categories of biological networks will be provided, focusing on the one that defines transcriptional regulations, the transcriptional regulatory network, and its challenging features to model.

2.2.1 Definitions and mathematical preliminaries

The basic mathematical concept used to model networks is a graph. This can be formally represented by a graphical structure G = (V, E), composed of a set of N nodes or vertices, $V = \{v_1, v_2, \dots, v_n\}$, and a set of edges or links, $E = \{(V_i, V_j): V_i, V_j \ \in V)$. The single edge e=$(V_i, V_j)$ represents the relation occurring between two nodes, and depending on the nature of the interactions in the graph, it can be directed (see Fig. 2.6, (a)) or undirected (see Fig. 2.6, (b)).



(a) Directed Graph

(b) Undirected Graph

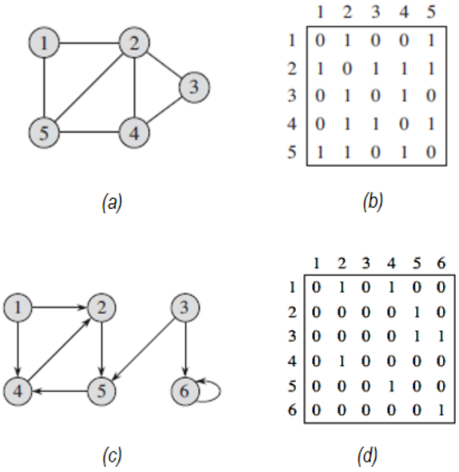
Figure 2.6: Graph representation. Two graphs with three nodes each one: (a) directed, (b) undirected

In a **directed graph**, an edge *e = (i, j)* ∈ *E* is an ordered pair, which represents the direction of the relation. The edge, in this case, is composed of a source node *s(e) = i* and a target node *t(e) = j*. Directed graphs are mostly suitable for the representation of schemas describing biological pathways or procedures which show the sequential interaction of elements at one or multiple time points and the flow of information throughout the network.

On the other hand, in an **undirected** graph, an edge is an unordered pair, since there is no direction associated with an edge. The two nodes joined by the edge *e* can be considered as source or target indifferently.

A biological network can be also described through a **weighted** graph in which each edge is associated to a weight function w: E→R, where R denotes the set of all real numbers. The weight w_{ij} of the edge between nodes *i* and *j* represents the relevance of the connection. Usually, a larger weight corresponds to higher reliability, or affinity of a connection.

A widely used way to represent the structural information stored in a network is through an **adjacency matrix**. The adjacency matrix A is defined as an NxN squared matrix in which each entry a_{ij} corresponds to the link between the nodes *i* and *j*. In particular, for an unweighted link a_{ij} will be 1 when there is a link between $(i, j) \in$ V and 0 otherwise. For a weighted graph, the values a_{ij} of the related adjacency matrix will correspond to the edge weights.



	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

	1	2	3	4	5	6
1	0	1	0	1	0	0
2	0	0	0	0	1	0
3	0	0	0	0	1	1
4	0	1	0	0	0	0
5	0	0	0	1	0	0
6	0	0	0	0	0	1

Figure 2.7: Graphs representation using adjacency matrixes. (a) An undirected graph with 5 vertices and 7 edges and its adjacency matrix (b). A directed graph with 5 vertices and 8 edges (c) and its adjacency matrix (d).

2.2.2 Structural properties

Looking at different network properties can provide valuable insight into the internal organization of a biological network. The topological measures give also insights into the evolution, stability, and dynamic responses of the system [37].

In the following, we provided an overview of the main properties that are commonly analyzed in networks and can be exploited in the context of transcriptional interactome modeling.

Degree Centrality

The most elementary characteristic of a node is its degree k, which shows the number of interactions of a given node, also indicating the relevance of a particular node to the large scale structure of a network. For a node i, the degree centrality is calculated as

$$C_d(i) = deg(i) \tag{2.1}$$

For directed graphs, each node is obviously characterized by two degree centrality measures, the *in-degree* (the number of edges ending in i), and *out-degree* (the number of edges from i to other nodes), respectively reported in Eq. 2.2.

$$\begin{gathered} C_{d\,in}(i) = deg_{in}(i) \\ C_{d\,out}(i) = deg_{out}(i) \end{gathered} \tag{2.2}$$

The average degree <*k*> is the average of all the vertex degree in the graph. Nodes with high degree are called **hubs** since they are connected to many adjacent nodes (neighbors) and tend to be essential for sustaining the integrity of the network [38]. Formally, the degree of a node i (k_i) is given as

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{2.3}$$

where A is the adjacency matrix and n is the number of network nodes.

The degree distribution P(*k*) of a network measures the proportion of selected nodes with degree *k*. Formally, if there are *n* nodes in total in a network, and n_k of them have degree *k*, the degree distribution is calculated as P(*k*) = n_k /*n*.

The degree of a node explains the general topological features of the network and can only capture the local structure of network nodes, since only the immediate neighborhood (nearest neighbors) of the vertex of interest is considered. To this aim, several global centrality measures are used in graph theory [39] to investigate patterns and rules hidden in the structural network domains. Most of these rely on the path concept.

A **path** from a vertex i to a vertex j is a sequence of edges which must be crossed to go from i to j with no edge traversed more than once. The graph distance $\delta(i, j)$ is the length of a path, and among all possible paths, the one with the smallest length is called **shortest path**. On the contrary, the **diameter** of a graph G is the longest shortest path taken over all pair of distinct nodes, $i, j \in V(G)$ which are connected by at least one path.

It follows that a graph is **connected** when there is a path between every pair of vertices, and there are no unreachable vertices, otherwise if exist two nodes in G such that no path in G has those nodes as endpoints, the graph is disconnected.

Betweenness Centrality

Betweenness centrality (BC) of a given node i is related to how frequently a node occurs on the shortest paths between all the pairs of nodes in the network. Formally, for distinct nodes $i, j, w \in V(G)$ let σ_{ij} be the total number of shortest paths between i and j and $\sigma_{ij}(w)$ be the number of shortest paths from i to j that pass through w. Moreover, for $w \in V(G)$, let $V(i)$ denote the set of all ordered pairs, (i, j) in $V(G) \times V(G)$ such that i, j, w are all distinct. Then the BC is calculated as

$$BC(w) = \sum_{(i,j)\in V(w)}^{n} \frac{\sigma_{ij}(w)}{\sigma_{ij}} \tag{2.4}$$

It is a widely applied measure in the context of regulatory networks, since nodes with high BC, termed as “bottlenecks”, exerting a key role in the essential functional and dynamic properties, and their disruption could greatly affect the network capacity of response and robustness [40,41].

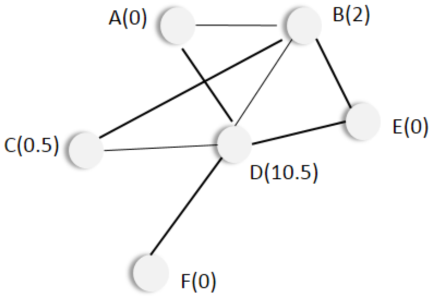


Figure 2.8: Graphical representation of betweenness centrality measure. Nodes A, B, C, D, E and F are well connected and maintain efficient network communication. Numbers in parentheses refer to each node’s BC, which indicates how many of the shortest paths between all other node pairs in the network pass through it. For example, to reach node C from node F, information flow is efficient and only passes through D.

Clustering Coefficient

The clustering coefficient C_i of a node n is a measure of the fraction of connected neighbors of the considered node. A node with k_i links can have at most $\binom{k_i}{2} = k_i(k_i - 2)/2$ pairs of its neighbors connected to each other. Denoting t_i as the number of links among the neighbors of node n, then the clustering coefficient is defined as

$$C_i = \frac{2t_i}{k_i(k_i - 1)} \tag{2.5}$$

It provides an idea of the level of interconnectivity in the neighborhood of a node, indicating also the modularity and connectivity patterns at a lower (more local) scale [42].

The average of C_i over all nodes in the network is

$$< C >= \frac{1}{N}\sum_{i=1}^{N} C_i \tag{2.6}$$

Higher values of the average clustering coefficient can be related to greater redundancy and robustness in biological networks.

For decades, molecular interaction networks were considered either completely regular or completely random. However, the obvious existence of molecules with a very high number of interactions is a fact that cannot be explained by either of the two models. In regular networks, all nodes have the same connectivity. In *random* graphs (see Fig. 2.9), links are placed randomly among nodes and their connectivity follows a Poisson distribution, which means that the existence of nodes with an extraordinarily high number of links is very improbable. Recent studies [37,43] have shown that in many biological networks, the degree distribution follows a power-law distribution that is $P(k)\sim k^{-\gamma}$ with parameter γ being often between 2 or 3. In such networks, most nodes have a small number of links, but a small fraction of nodes (hubs) have a very large number of edges. Because in such networks no 'typical node' (typical 'scale') exists, they are called **scale-free**, as shown in Fig. 2.9.

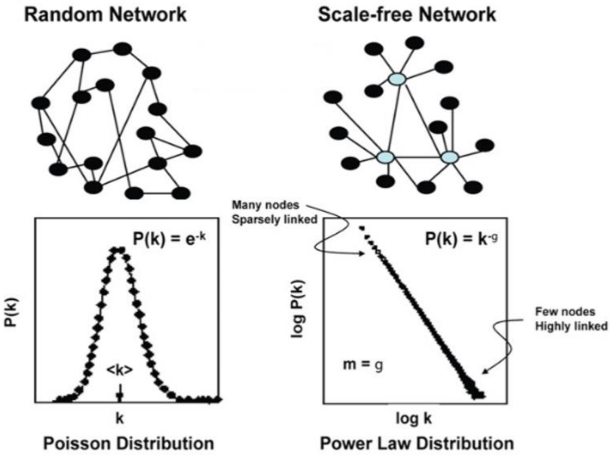


Figure 2.9: Random network and Scale-free network properties.
Source: Chan S.Y, Loscalzo J [44].

Scale-free networks are very robust, extraordinarily resilient to random component failures. Even after a high number of nodes are removed, the rest are still held together by the hubs so that the network often does not become disintegrated and can still fulfill its function. As the number of hubs is relatively very small compared to the number of nodes with few links, the chance that a randomly removed node is a hub is small. The intentional removal of hubs, on the other hand, is often critical for network's integrity and proper function, that is, scale-free networks have a high hub vulnerability.

Another important feature of biological networks is **modularity**, the tendency to contain nodes communities. Since genome-wide interaction networks are highly connected, modules should not be understood as disconnected components but rather as components that have dense intracomponent connectivity and sparse intercomponent connectivity.

2.2.3 Biological Networks models

In network biology, according to Barabasi et al. [45], we can distinguish different types of networks, in relation to the molecular interactions which they model. The five main categories are briefly described below.

- **Protein-protein interaction (PPI) networks.** Nodes of PPI networks are proteins and edges represent their physical interactions.
- **Metabolic Network**. Nodes of metabolic networks are metabolites that are linked if they participate in the same biochemical reactions.
- **Signaling networks** show how extracellular signals are propagated in the cells using multiple signal transduction pathways.
- **Co-expression networks** in which genes with similar co-expression patterns are linked.
- **Gene Regulatory network (GRN)**. In this network, a node can represent a gene, the transcribed mRNA, and the coded protein simultaneously. The links are directed and indicate a regulatory interaction which governs the cellular gene expression process.

Since the proposed approach focuses on a particular type of GRN, the transcriptional regulatory network, next section will provide a complete description of the underlying transcriptional biological mechanisms which these kind of network tries to model, and its challenging characteristics.

Transcriptional regulatory networks

Genes and gene products interact on several levels. At the genomic level, transcription factors (TFs) can activate or inhibit the transcription of genes finalized to the production of mRNA transcripts, which are transduced into proteins (see Fig. 2.10 (A)). This represents a major control point in gene expression processes operated by TFs, presiding over precise spatial and temporal control mechanisms responsible for the intricate cellular processes of developmental specification and adult tissue homeostasis.

TFs account for almost 7% of genes (~1,400) in the human genome, and to exert their role, TFs bind in a DNA sequence-specific manner the promoter region of a target gene, near its Transcription Start Site (TSS) to allow the initiation of the transcription process.

This event is also triggered by the interaction of TFs with other transcriptional machinery components, such as RNA polymerase and chromatin-remodeling complexes (i.e. transcriptional co-activators and co-repressors), and by the behavior of TFs as epigenetic regulators, acting on the conformation of DNA for its accessibility [46] (see Fig. 2.10 (B)).

Moreover, TFs have the ability to directly regulate their expression, through the control of their own gene transcription, and, interacting cooperatively with other TFs [47], giving rise to intricate TFs regulatory circuits (see Fig. 2.10 (C)). A target gene can be controlled by more than one TF, providing a flexible regulation in a combinatorial manner, that is very likely to confer a fitness advantage under different environmental conditions [48].

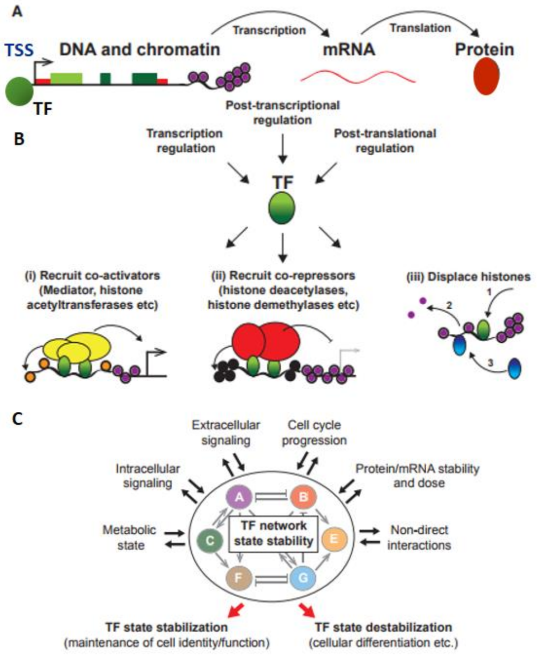

A
TSS
DNA and chromatin
Transcription
mRNA
Translation
Protein
TF
B
Post-transcriptional regulation
Transcription regulation
Post-translational regulation
TF
(i) Recruit co-activators (Mediator, histone acetyltransferases etc)
(ii) Recruit co-repressors (histone deacetylases, histone demethylases etc)
(iii) Displace histones
1
2
3
C
Extracellular signaling
Cell cycle progression
Intracellular signaling
Protein/mRNA stability and dose
Metabolic state
Non-direct interactions
A
B
C
E
F
G
TF network state stability
TF state stabilization (maintenance of cell identity/function)
TF state destabilization (cellular differentiation etc.)

Figure 2.10: Molecular functions of TFs. (A) The gene expression process, from genes, encoded within genomic DNA and packaged inside chromatin structure, to their transcribed products (mRNAs) and then proteins. (B) Regulation of TFs expression and activity. (C) Mechanisms of TFs network state stability, influenced by numerous extrinsic and intrinsic mechanisms.
Source: Adapted from Wilkinson A.C et al [47]

Given the determinant role of TFs in defining gene expression profiles in response to several cellular signaling cascades, not surprisingly then, mutations to transcription factors and molecules that comprise and modify the chromatin landscape, commonly underlie the altered gene expression profiles that are characteristic of cancer cells. Centrally to the realization of personalized medicine, investigating genetic mutations alone often fail to accurately predict disease progression [47]. Understanding the TF network states associated with a certain disease within a unique output (e.g., gene expression profile), may help to accurately predict clinical response and outcome [46,49].

To this aim, TFs interplay can be modeled with **transcriptional regulatory networks** (TRNs), whose nodes represent both transcription factors and their target genes (TGs), and directed edges define the regulatory interactions among TFs, and from TFs to their targets.

Given the synergic behavior of TFs, at local level TRNs are characterized by several regulation **motifs**, configurations of regulators and target genes that occur repeatedly within network structure, suggesting a modular network organization. Such motifs represent the simplest units of the network architecture required to create specific patterns of inter-regulation among TFs and TGs. They are conserved in diverse organisms from bacteria to human, and carry out specific dynamic cellular functions [50]. Examples of transcriptional motifs described below are reported in Fig. 2.11.

Negative autoregulation occurs when a transcription factor represses the transcription of its own gene; on the contrary, cascades of gene expression create consecutive activation of genes. The downstream gene is activated when its regulator reaches the relevant threshold of activation, and using also a negative regulation, genes can be sequentially stimulated and repressed. Feedback loops are made of two TFs that regulate each other. Feed-forward loops (FFL) consists of three layers of regulation in which, at the top, the *master regulator*, indicated with TF_1, in Fig. 2.11 (D), regulates the two underlying strata. The intermediate regulator, *middle manager* or *broker,* (TF_2) together with the master regulator control the TF at the bottom (TF_3), which is therefore identified as the regulated vertex or *workhorse*. Each of the three interactions in the FFL can be either activation or repression mechanisms [50].

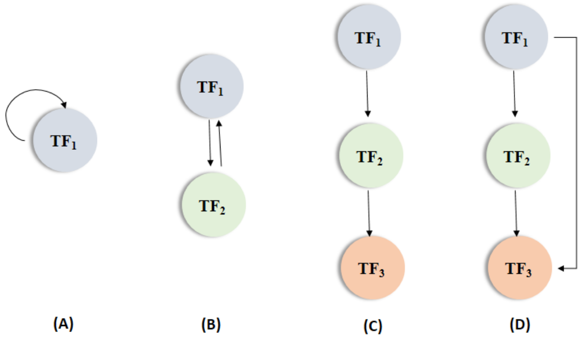


Figure 2.11: TRNs motifs. (A) TF autoregulation; (B) Feedback loop; (C) Transcriptional cascades; (D) Feedforward loop.
Arrowheads in this representation do not discriminate between activating and repressing transcriptional functions.

As can emerge from these considerations, the underlying transcriptional architecture in TRNs is hierarchical. The high molecular complexity due to the combinatorial nature of TFs interactions, enriched in **loops** (cycles) of regulation, impacts on network size, culminating in large "hairball" structure. It becomes clear that the aforementioned structural organization cannot be easily detectable, making it difficult to formulate simple conclusions regarding the logic or outputs of these networks, especially if the model reflects a genome-wide perspective [51].

Traditionally, biomedical research has applied a reductionist approach to study the transcriptional background, focusing on a specific known mutated TF [52,53] or on small fraction of crucial TFs to explore a specific cellular process [54,55], isolating them from other regulatory elements that collectively form the context in which TFs operate. To instead maintain a global point of view on transcriptional interactions and cooperation, its study on a genomic level could shed a light on understanding the molecular mechanisms of human biology and pathogenesis.

2.3 Transcriptional regulatory networks reconstruction: an overview

The availability of completely sequenced genomes and the wealth of literature on gene regulation have enabled researchers to model the transcriptional interactions system of some model organisms in the form of a network. The study and characterization of such interactomes started from simple model organisms, from the metazoan *Caenorhabditis elegans*, to the bacterium *Escherichia coli* and yeast *Saccaromices cerevisiae*, easier to investigate than human networks, since their genomes contain less genes than human genome, and can also be engineered through targeted experiments.

The increasing advance in experimental techniques as well as in computational methods make genome-scale regulatory network reconstruction a feasible task, at least for these well-studied organisms. The obtained knowledge led the resulting networks to be considered as gold standards, whose validated interactions are available in RegulonDB [56] for E. *coli*, and in Saccharomyces Genome Database (SGD) [57] and YEASTRACT [58] repositories for S. *cerevisiae*, respectively.

On the other hand, reconstruct such networks in non-model organisms, as in the human context, requires robust computational approaches to learn directly from data or from existing knowledge (i.e. curated databases or from published experimental research works) the interactions of a state-specific regulatory circuitry, which remains largely unknown.

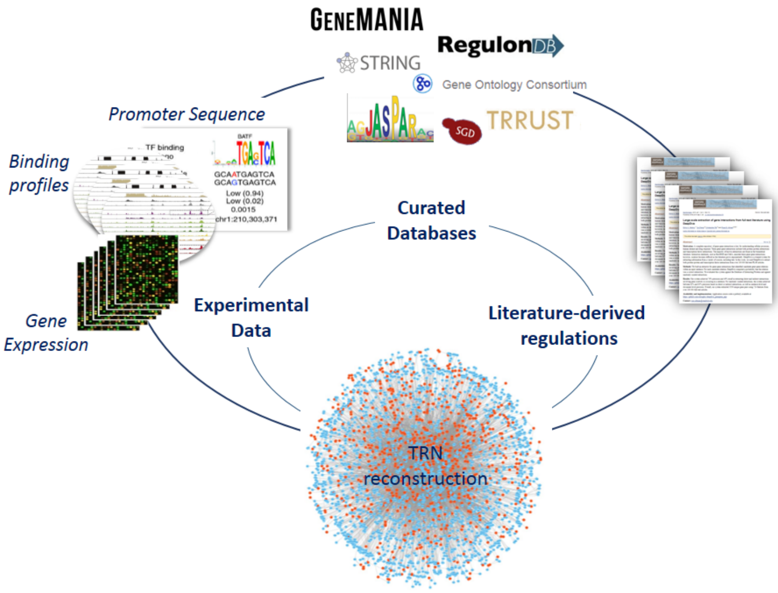


Figure 2.12: Transcriptional regulatory networks reconstruction approaches

Several approaches have been developed through recent years, trying to reconstruct TRNs and to make inference on TFs activities.

Template based methods transfer interactions between homologous components from a model organism to the organism of interest [59]. Starting with a known regulatory network (used as a template), the information about interactions can be transferred to genes that have been determined to be orthologous in a target genome of interest.

Reverse engineering approaches [60], aimed at determining the expression state of a genome, use microarray experiments to detect similar patterns in gene expression that stem from similar regulatory interactions.

Other "physical" methods are based on the principle that TFs recognize their targets through specific sequences (*binding motifs*). Genes that share common sequences in their regulatory regions are more likely to be under similar regulation. This logic has been extensively used to infer TF binding motifs. On the other hand, if the TF motif is known, a gene whose regulatory region contains one or more instance of this motif is more likely to be the regulatory target of this TF.

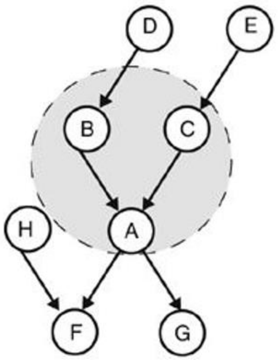
Most of these inferential strategies rely on exploiting a single source of data, providing a partial and potentially biased reconstruction. As He B. and Tan K. [61] pointed out in their recent review, among current computational approaches for constructing TRN models there is a lack of integrative genome-wide methods which, combining multiple, independently generated observations (such as gene expression, *in vivo* TF binding and chromatin modification states, protein abundance measure, etc.), can strengthen the resulting models and provide novel insights from the inferred network structure. A particular issue is to find a method able to deal the biological complexity of these systems, and sufficiently robust to scale their genomic dimension allowing multiple data integration. Among the mathematical formalisms used to model the transcriptional information, as linear regression, statistical correlation or Bayesian theory, this last one, through Bayesian networks, offers an ideal framework for heterogeneous data integration, using a combination of two mathematical areas: probability and graph theory [62].

For such reasons, the proposed data fusion approach exploits the Bayesian formalism to jointly analyze complementary transcriptional data under a single unified framework.

2.4 Modelling transcriptional regulations using Bayesian Networks

2.4.1 Bayesian Networks

A Bayesian network (BN) is a graphical representation of the joint probability distribution (JPD) of a set of random variables $X = \{X_1, \dots, X_n\}$. BN is described as $B =< S, \Theta >$, where the encoding of this probability distribution is defined by a network structure S and a set of model parameters Θ, which describes the probability distribution of model's variables [63]. Model structure S is represented as a directed acyclic graph (DAG), whose vertices (or nodes) are the random variables, and their conditional dependencies are described by directed edges. In particular, each variable is assumed to be independent of its non-descendants given its set of parents, denoted as $\boldsymbol{pa}(X_n)$.



P(A|B,C,D,E)= P (A | B,C)

Figure 2.13: Graphical representation of a Bayesian Network. Node A is conditionally independent of D and E given B and C. The BN relationships can be described through the factorization of the full JPD into component conditional distributions: P(A,B,C,D,E,F,G,H)= P(D) PI P(H) P(B|D) P(C|E) P(A|B,C) P(F|A,H) P(G|A)
Source: Adapted from Bansal M. et al [64]

Under this Markov assumption, the joint probability distribution of all nodes of the model is given as

$$P(X) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{pa}(X_i) = \prod_{i=1}^{n} \theta_{X_{i|\boldsymbol{pa}(X_i)}} \qquad (2.7)$$

where each variable X_i is described by a parameters' set (θ_i) which defines the variable distribution conditional on its parents.

Given a DAG, let $\{X_1, \ldots, X_n\}$ be a topological ordering of variables of S, where with *parents* nodes (*ancestors*) are ordered before *children* (*descendants*). The set $\{X_1, \ldots, X_{i-1}\}$ includes only parents and non-descendants of X*ᵢ*.

In the case of transcriptional models, a TF node (*parent node*) can regulates the expression of a target node (*child node*) which can be either a TF or a gene vertex. The underlying structure represents the causal relationships among variables that, in this context, are the regulatory interactions among transcription factors (TFs) and from TFs to genes.

In this way, relationships among BNs variables can be described at both qualitative and quantitative level. At a qualitative level, since the edges represent simply dependence and conditional independence relations. At quantitative level, with family of joint probability distributions, whose form depends on the type of network variable, which can be discrete or continuous.

In the case of *discrete* variables, each node takes finite values so that the JPD representation is given by a conditional probability table (CPT), specifying probabilities according to all possible joint configurations of parents. For *continuous* nodes, multivariate continuous distributions do not have a unique representation, and it is possible to use a linear Gaussian conditional distribution for each node and hence the multivariate Gaussian as joint distribution. In this case, the linear Gaussian density of *Xᵢ* given its $\boldsymbol{pa}(X_i) = \{U_1, \ldots, U_k\}$ implies that is normally distributed around a mean value that depends linearly on the values of $\boldsymbol{pa}(X_i)$. The variance of this Normal distribution is independent of the parents' values. In this representation $\theta_{X_i|\{u_1,..,u_k\}} = \langle a_0, \ldots, a_k, \sigma \rangle$.

$$P(X_i|u_1, \ldots, u_k) \sim N\left(a_0 + \sum_i a_i \cdot u_i, \sigma^2\right) \tag{2.8}$$

Continuous data allow the inference of network models without suffering from information loss due to discretization. Moreover, continuous models are more parsimonious than discrete models since they require fewer parameters to describe variable dependencies.

2.4.2 Learning Bayesian Networks

In order to perform an efficient inference and correct representations of transcriptional dependencies, the definition of BN model from a TRN can be implemented learning its structure from experimental data.

Given a dataset $D = \{D_1, \ldots, D_n\}$ where D is an instantiation of all the variables in *X*, learning BN structure from *D* corresponds to finding a model structure that best fits the observed data.

The model-learning algorithm usually assumes a specific form of the conditional probability function. Any function can be used, including Boolean and linear functions. But there will be a tradeoff between model realism and model simplicity. More realistic models will have more parameters, which will require more experimental data and greater computational effort to solve.

Finding the optimal BN represents indeed an NP-hard (nondetermistic polynomial-time) problem [65], both the running time and memory usage of exact learning are exponential in the number of variables. In order to face this limit, several empirical investigations have been carried out on developing approximation methods, which collectively have been classified in the literature as *constraints-based* and *score-based* structure learning algorithms [66].

The first algorithms class learns a BN structure as a constraint satisfaction problem. In this approach, properties of conditional independence among variables are estimated by a statistical hypothesis test, such as mutual information test or the exact Student's *t* test, to construct a partially oriented graph, retaining or rejecting candidate edges using the observed dependencies and independencies.

The second approach learns a Bayesian network as a heuristic optimization problem, exploiting a statistically motivated scoring function. To evaluate the goodness of fit of each candidate structure model (G*), the process assigns to it a score, which the algorithm tries to maximize.

$$G^* = argmax_{G \in G_n} Score(G|D) \tag{2.9}$$

where G_n is the set of all possible structures (DAGs), *Score(G|D)* is a given score function measuring the degree of fitness of any candidate DAG (G). The score typically approximates the probability of the structure given the data and represents a compromise between how well the network fits the data and how complex the network is.

An important property of the scoring function is its *decomposability*. A scoring function is *decomposable* if the value assigned to each structure can be expressed as a sum of local values that depend only on each node (X_i) and its parents, as denoted in the Eq. 2.10.

$$Score(G \mid D) = \sum_{i=1}^{n} FamScore\big(X_i \,\big|\, \boldsymbol{pa}^{\boldsymbol{G}}(X_i) \,\big|\, D\big) \tag{2.10}$$

The *Bayesian Information Criterion* (BIC) [67] is among the most popular scoring metric, which asymptotically approximates the posterior probability of the DAG. It is based on the maximization of criteria that combines a term describing the likelihood of the observations, and another one that penalizes the complexity of the model.

$$BIC(B, D) = logPr(D|G, \theta^{ML}) - \frac{1}{2} Dim(G) logN \qquad (2.11)$$

where θ^{ML} is an estimate of the model parameters, obtained by likelihood maximization, and *Dim*(G) is the network dimension, also defined in Eq. 2.10. It represents the number of model parameters, and N the size of the dataset. The second term of Eq. 2.11 is a penalty term which has the effect of discouraging overly complicated structures and acting to automatically protect from overfitting.

Constraint-based approaches have been shown to be sensitive to error propagation [68] and do not give an indication of the relative confidence in the model, which is instead provided, on the other hand, to the score-based methods. Both strategies scale to large networks poorly, because the number of possible graph structures or tests rises exponentially as the size of a network increases.

A third class of learning structure methodologies is represented by *hybrid algorithms*, that, exploiting the best of both worlds, have therefore proved successful in learning causal graphs from data [69,70]. Typically, they start with a constraint-based search to find the skeleton of the network and then employ a score-based scheme to identify a high-scoring network structure.

To circumvent the high-dimensional search space problem of all possible structures, and to reduce the inherent uncertainty of models retrieved by heuristic learning schema, is possible to incorporate *prior knowledge* into the algorithm framework.

2.4.3 Prior knowledge integration in Bayesian Network learning

In recent years, different studies have so far been carried out in order to develop BN learning methodologies for recovering transcriptional regulatory networks. Several methods [71,72] have focused their learning procedure only on gene expression profiles in their static or dynamic (or time-series) forms.

The early works in this area attempted to reconstruct networks from microarray data alone. Friedman et al. [73] and Murphy and Mian [74] were among the first to apply a Bayesian structure learning strategy on time-series data, trying to capture transcriptional dynamics in the temporal domain. The limited number of monitored time points, and, for the static gene expression profiles case, the relative limited number of experimental samples, are nevertheless statistically insufficient for reconstructing even a network with moderate size.

Moreover, data only gives a partial picture of regulatory mechanisms, which, combined to inherent noisy and sparse nature of experimental source, potentially affect the truthfulness of inferred results, leading to incorporate unreliable biological transcriptional regulations [75,76].

To improve the performance and the accuracy of network reconstruction process, is possible to introduce prior knowledge into the learning process.

Several types of prior information can be introduced during network learning such as known associations between genes and transcription factors (TFs), TF binding-sites or genomic context information. The biological prior integration can be effectively achieved within the framework of BNs as they offers a well-founded way to introduce prior knowledge, by exploiting the possibility to specify prior probabilities for network models. Moreover, BNs present the characteristic to decompose the global model in local ones and, for this reason, also the prior introduction over a network can be realized considering probabilities for each individual edge.

Several studies have so far been carried out in order to develop methodologies to integrate prior information in BN learning.

Le Phillip et al. [77] approached this problem clamping edges and non-edges which means that knowledge about interactions is transformed into hard constraints. The presence of a relationship leads to set the respective prior probability to 1 and conversely absence of an edge to a prior probability equal to 0; each simulation selects randomly clamped edges and non-edges.

Bernard and Hartemink [78] selected as model prior transcription factor binding location data, forcing the search procedure to add arcs in a specific position, and eliminating all graphs lacking these recommended edges. Data about TF binding location suggests the presence of a connection by means of a p-value that is inversely related to the network edge probability. Therefore, they derived a function to map the described evidence into probabilistic terms. Using the edge-wise decomposition that is the subdivision in local models, a factorable informative prior over networks is obtained.

Imoto et al. [79] and Werhli et al. [80] expressed biological priors in terms of energy function, measuring the degree of agreement between the explored network and the prior information. The total energy can be decomposed into the sum of local contributes, that is local energy defined by a gene and its parents. This formulation allows to evaluate the difference between prior structure and learned network as a unique quantitative probabilistic value without transforming it into edge probabilities. Even though this framework has been successfully used by several authors, it is limited in the application to small networks because of complexity and computational time [81].

Other work allows the integration of multiple types of prior knowledge into a Bayesian framework [79,82]. While it may be obvious that incorporating more data or prior knowledge into the network reconstruction process will give better learning results, there currently exists no quantitative analysis of the effects of data set size and amount of prior knowledge on learning performance for networks with realistic topology. In addition, these methods are always applied on networks or biological pathways with a small group of variables from model organisms, such as S. cerevisiae, lacking of an integrative method which could handle large-scale interactions.

All the above mentioned works highlighted an improved fidelity of network reconstruction using prior knowledge in their learning schema. They further confirms that exploiting a unique source of information is usually not sufficient for an accurate and robust regulatory mechanisms reconstruction, which can be overcome through data integration, as already discussed in the previous sections (see Sec. 1.3; Sec. 2.2, Sec. 2.3), and which is instead a central focus of the work done in this thesis.

2.4.4 Bayesian networks inference

BNs have been studied also as an instrument of inference, supporting reasoning about events in a domain with inherent uncertainty. A Bayesian network is a complete simulation system able to predict the value of an unobserved variable under particular conditions (the posterior probability of a variable given the observations gathered on any of the other variables) and, on the other hand, able to find the most probable set of initial conditions for an observed status.

In the case of transcriptional models, once the underlying structure have been reconstructed, their probabilistic inference through a Bayesian model allows to prioritize transcriptional interactions and as a consequence, the related gene expression regulators, with the aim to uncover the dynamics of the underlying regulatory programs.

Given the model structure and a set of input values, referred to commonly as the *findings* or *evidence*, BN derives the *posterior probabilities* for a target of interest. This is known as *probabilistic inference* on the target, and the value with the highest belief or probability is its prediction.

By inference, we mean computing

$$P(X_i|X_j) \propto \sum_{k \neq i,j} P(X_i, X_j, X_k) \tag{2.12}$$

where X$_j$ represents a set of observed variables, X$_i$ represents a set of hidden variables whose value we are interested in estimating, and X$_k$ are the irrelevant (nuisance) hidden variables.

For instance, given evidence e (i.e. the expression level) of a target node I in the regulatory network, inferences about the likely values of other nodes of the model or of a subset of them (Y) can be made as $P(Y|E = e)$.

More generally, inferences of the values of a set of variables may be made given the evidence of another set of variables, by marginalizing over unknown variables. This marginalization operation is equivalent to consider all possible values that the unknown variables may take, and averaging over them [83].

Conceptually, inference is straightforward, P(A|B) is calculated as a product of relevant conditional probability distributions, using Bayes' rule to calculate any posterior probabilities.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.13}$$

where P(A|B) and P(B|A) are conditional probabilities, and P(A) and P(B) are the probabilities of observing A and B independently of each other, called marginal probabilities.

Computationally, a number of methods have been developed, exploiting the structure of the BN model to derive efficient exact or approximate inference algorithms, collectively defined as *inference engines.* They allow to interrogate the BN using the evidence, for computing solutions to queries against the knowledge base [84].

Exact inference, that means having a closed form solution, is only possible in a very limited set of cases, meaning that given a model P_{Φ}, a variable X and a value $x \in Val(X)$, compute $P_{\Phi}(X = x)$ is NP-hard.

The "easier" situations for this type of probabilistic inference are when all hidden nodes are discrete, or when all nodes (hidden and observed) have linear Gaussian distributions, in which case the network is just a sparse parameterization of a joint multivariate Gaussian [85]. Exploiting the chain-rule decomposition of the joint $P(X) = P(X_1), P(X_2|X_1), P(X_3|X_1, X_2), \ldots$, the algorithm constructs the join distribution over all nodes and then marginalize it.

When the exact inference is not computationally feasible, or there is no closed-form solution, an *approximate inference* can be applied. Within the field of approximate Bayesian inference, variational and Monte Carlo methods, and Belief propagation are currently the mainstay techniques, which will be not used in the context of this work, and whose detailed description can be found in Murphy K.'s study [86].

2.4.5 Chronic Myeloid Leukemia: a case study

Chronic myeloid leukemia or chronic myelogenous leukemia (CML) is a myeloproliferative neoplasm that originates in the hematopoietic stem cell (HSC) of the bone marrow, and is caused by a specific mutation. The hallmark of CML is indeed the presence in this cells of a balanced translocation between the long arms of chromosomes 9 and 22, t(9;22)(q34;q11.2), which is known as the Philadelphia (Ph) chromosome. This translocation results in the formation of the BCR-ABL1 fusion gene, as illustrated in Fig. 2.14, which, in turn, is translated into a chimeric Bcr-Abl protein with deregulated tyrosine kinase activity.

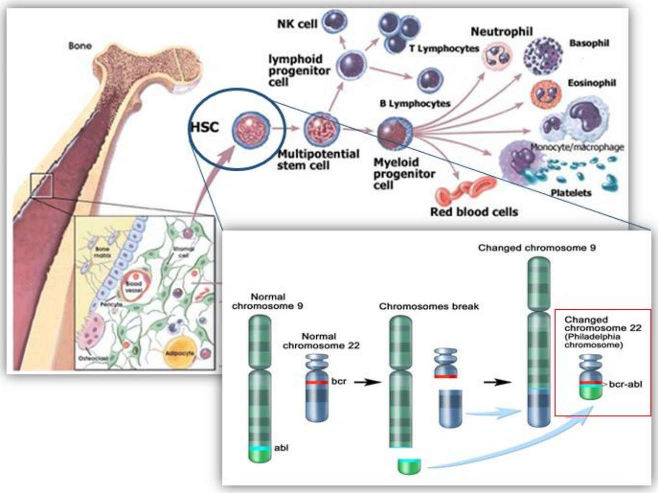


Figure 2.14: Molecular biology of CML disease.

The normal tyrosine kinase activity of the ABL protein is tightly regulated, but it changes into constitutive activity due to the traslocation. In this way, BCR/ABL is able to transduce signals in various cellular processes in an autonomous fashion, triggering multiple downstream pathways, which lead to enhanced cell proliferation and transformation, reduced growth factor dependence, resistance to apoptosis, and genetic instability [87].

This results in the expansion of the leukemic cell population, initially characterized by overproduction of mature myeloid cells with normal morphology (*chronic phase*). As the disease advances (*accelerated phase*, followed by *blast crisis*), leukemic stem cells acquire additional chromosomal aberrations and mutations, which involve transcription factors [88,89], contributing to disease progression. However, at present, little is known about the molecular mechanisms underlying disease progression, but, most likely, activation of oncogenic factors and/or mutations leading to loss of function of tumor suppressor genes in hematopoietic stem cells are involved [90]. Since TFs control expression of genes essential for the normal functioning of the hematopoietic system and regulate development of distinct blood cell types, in the event of genetic perturbations, their molecular roles can be altered, resulting in uncontrolled proliferation of immature blood cell lineages and sometimes depletion of one or more blood cell lineage as occurs in leukemia [89]. The lack of a deep understanding of the molecular mechanisms underlying the disease reflects on the problem of drug resistance, which is poorly understood.

The therapy of choice uses targeted inhibitors of the enzymatic activity of the BCR-ABL1 protein product. This treatment does not eradicate cancer cells, which continue to progress, and an ever increasing percentage of patients fail primary cure, and only 10–20% can discontinue therapy and achieve long-term treatment-free remission [91].

In order to improve the current CML knowledge and, consequently, the therapeutic strategies, there is still a significant clinical need to develop novel integrative approaches to investigate on a large-scale the pleiotropic effect of constitutive BCR-ABL1 activity. In support of this, expression studies revealed that BCR-ABL1 dramatically perturbs the CML transcriptome [92], resulting in altered expression of genes.

It became clear that such scenario is a suitable candidate for the proposed data fusion approach, aimed at omics data integration to reconstruct and investigate the transcriptional signatures on a genome-wide perspective.

Chapter 3

A Bayesian Data Fusion based approach for learning genome-wide TRNs

Since our data fusion method relies on omics data integration, the first step is represented by the TRN reconstruction on a genome-wide scale, which may help to define the global picture of the physiological or disease status at the molecular network level.

Once the backbone of the transcriptional system is defined, in order to scale its complexity and infer the underlying transcriptional signatures, it is then converted into a Bayesian model and integrated with transcriptomics data for its probabilistic investigation.

3.1. Genome-wide TRN construction

Formally, a TRN can be defined as a directed graph $TRN = \langle V, E \rangle$, where V is the set of TFs and genes vertices, and E is a ordered pairs set of genomic edges composed in turn by two subsets, describing the regulatory interactions among TFs (E_1) and from TFs to genes (E_2).

$$V = \{TF_1, \dots, TF_i\, ; G_1, \dots, G_k\}$$

$$E = \begin{cases} E_1 = \{(TF_1, TF_2), \dots, (TF_i, TF_j)\} & \forall_i \forall_j, i \neq j \\ E_2 = \{(TF_1, G_1), \dots, (TF_i, G_k)\} \end{cases}$$

To accurately reconstruct the genomic transcriptional regulations which constitute a TRN, ChIP-seq data is the suitable source of information to achieve this goal, as described in Sec. 2.1.1 and Sec. 2.3. Their potential lies in revealing the high-dimensional interrelationship level of TFs binding sites across the entire genome. If on one side it is useful for understand the cooperation and the interactions on a genome-scale, on the other hand, as all experimental data source, has a noisy nature which may lead to false positive associations. This feature can be controlled starting the study from raw data and adopting some appropriate expedients at different steps of the computational analysis.

For the first stage of omics data integration process within the proposed data fusion approach, a bioinformatics pipeline has been developed in order to handle the data volume and its inherent experimental heterogeneity, with the final aim of building up the genomics transcriptional profiles.

3.1.1 Computational integrative analysis for TRN design

Combining quality control metrics with stringent p-value cut-offs for binding signals (*peaks*) detection allows to discover, filter and evaluate TFs-specific profiles along the genome from sequence alignment data (BAM file) of each considered ChIP-seq experiment. Moreover, a scoring method to quantitatively weight the strength of the target-TF interaction has been also introduced to remove weak and potentially false relations.

The analysis is carried out on UNIX command line, and is integrated with some genomic tools to allow the set-up of the pipeline. Its main steps are reported in Fig. 3.1, and are described below.

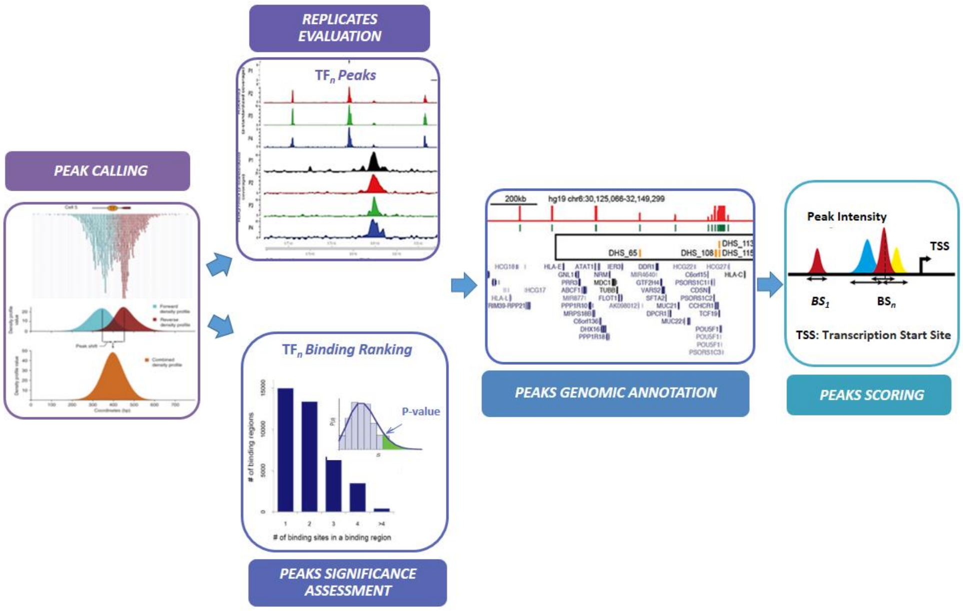


Figure 3.1: ChIP-seq bioinformatics analysis pipeline

Peak calling

Peaks detection of a ChIP sample involves the use of specific algorithms looking for the regions of significant tag enrichment that are typically assumed to reflect transcription factor binding to the sequence region. The Model-based analysis of ChIP-seq (MACS) method in its stable version (v.1.4.2) [93] was implemented as peak calling algorithm.

Starting from reads count data, it removes redundant reads that are repeatedly mapped to the same location, and calculates the reads distribution for each genomic position along the DNA double-strands, comparing and normalizing it to the background (reads from input/control sample). Peaks mapped to the two strands are treated separately to build two coverage density profiles. The distance between the modes of the two distributions represent the fragment size (d) bound by a certain TF (see Fig. 3.2 (A)), and will be used by MACS to detect regions significantly enriched in the ChIP sample.

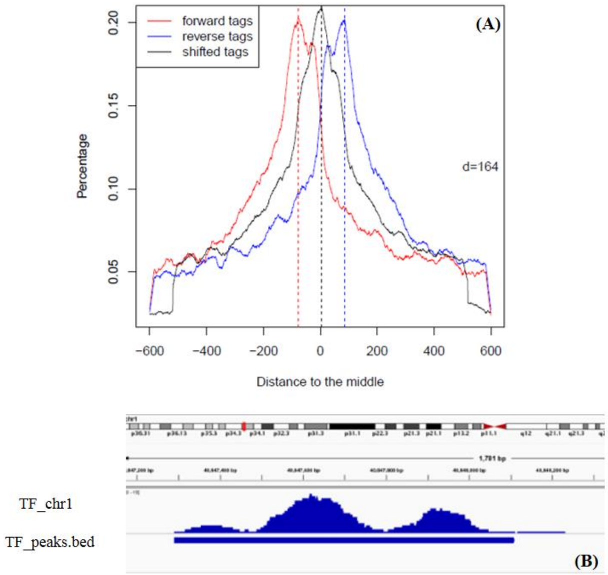


Figure 3.2: Peak model and binding profile built by MACS. (A) TF Peak model, where d=164 represents the estimated DNA fragment size. The red and blue curves model the percentage of reads (tags) at each base pair on the two DNA strands, respectively the forward and reverse strands. The black one represents the union of the two distributions. (B) From the final TF binding profile, peaks within a specific chromosome region is represented

In the end, an empirical false discovery rate is calculated for each peak, assessing its statistical significance. As result, MACS retains only peaks whose p-value is $< 1.00e^{-05}$.

The final output consists of a *BED* (Browser Extensible Data) file with genomic locations of the called peaks, that are the peaks' lengths, and a *summits* file, which reported the summit height of each identified peak, corresponding to its intensity.

The genome-wide binding profile of the considered TF is in this way reconstructed, and visualized as in Fig. 3.2 (B).

Since each ChIP-experiment can be conducted with more than one replicate to assess its biological variability (*biological replicates*), the aforementioned procedure has to be repeated for each replicate for each TF sample.

Replicates evaluation

In order to derive a consensus binding profile for each analyzed TF, replicates from the same experiment are sorted by genomic coordinates, concatenated and merged, combining the *overlapping peaks* of a genomic interval into a single peak, which spans all of the combined features. To this aim, BEDTools [94] is applied to finally retain those peaks observed in all the considered replicates.

Peaks significance assessment

All peaks of each TF consensus BED file, obtained from the previous step of analysis, are ranked by their p-values calculates by MACS. To further evaluate their statistical significance, with the aim of avoiding the inclusion of spurious interactions, a stringent p-value cut-off of 10^{-9} has been applied as additional constraint. Peaks with a p-pvalue less than this threshold are retained for the last steps of the analysis.

Peaks genomic annotation

To associate a genomic region to a specific gene bound by a specific TF, all the genomic coordinates of peaks, that passed the statistical significance filtering, are annotated to the human reference genome (GRCh37/ hg19 version). If the binding coordinates are from a previous genome assembly, such as hg18, the CrossMap (Convert Genome Coordinates Between Assemblies) tool [95], integrated to the pipeline, can be used to convert the coordinates ranges between genome assemblies before the annotation step.

Only those peaks which map the promoter of a gene, which is defined through its Transcription Starting Site (TSS), a region of DNA that initiates transcription of a particular gene, will be retained in the final output, and classified as *promoter-associate peaks*.

In this way, for each initially analyzed TF-ChIP sample, we will obtain a genome-wide TF binding profile.

This annotation step is computationally linked to the following phase of analysis since both depend to the same R package, TFTargetCaller [96].

Peaks scoring

Given the heterogeneity of the TF binding signal around the TSS region, that can be narrow for factors requiring binding close to the promoter, whereas it will be broader for factors that may bind further away and still affect the expression of their targets. Moreover, the co-occurrence of TF peaks in the proximity of the promoter, assumes that (1) genes with many peaks in proximity to their TSS are more likely to be targets and (2) peak proximity to the TSS increases the probability of the gene being a target. This determines several overlapping peaks with different width and intensity and in order to minimize artifacts and false positive interactions, within the applied pipeline they will be quantitatively weighted.

The scoring metric developed by Sikora-Wohlfeld et al. [96] gives a measure of the confidence of the TF binding and as a consequence, this numerical value is directly proportional to the strength of the interaction.

Using a *ClosestGene* approach which (1) assigns peaks to their closest gene, (2) scores peaks based on the distribution of all peaks around the TSSs, and (3) considers (summing) all peaks assigned to a particular gene, all the peaks along the genome for all the evaluated TFs will be weighted (see Fig. 3.3).

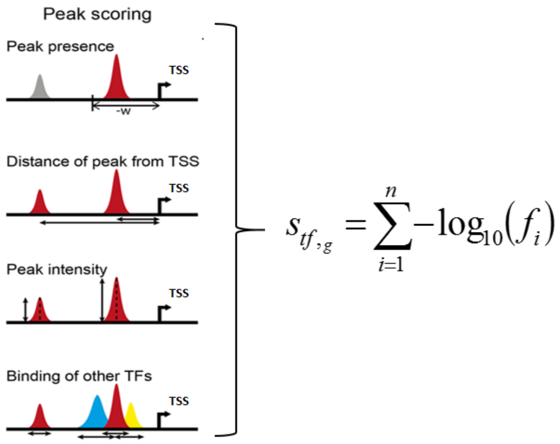


Figure 3.3: Overview of the ChIP-seq scoring procedure.
Source: Adapted from Sikora-Wohlfeld et al. [96]

The scoring function calculates, for each TF profile, the cumulative distribution of peaks distances to their closest genes, and computes the scores calculating the fraction of peaks observed at the given distance from the TSS (f_i), interpreting it as a probability. The interaction score ($s_{tf,g}$) between a

TF and a target gene is calculated as the sum of $-log_{10}$ transformed scores f_i for individual peaks assigned to a given gene (the formula is reported in Fig. 3.3). Higher is the score, stronger is the binding.

In this way, each TF binding profile is annotated to the reference genome assembly, each gene is then associated with its TSS, and the related mapped peaks are scored. All the interactions to which correspond a zero score will be discarded.

The omics TF binding profiles, obtained with the aforementioned bioinformatics analysis are then computationally integrated in order to reconstruct the transcriptional interactome maintaining the genome-wide perspective. Each profile can be view as a graph with a single regulator node and all its genomic regulated genes as target nodes.

Since TFs cooperatively regulates each other, as described in Sec. 2.2.3, all the resulting graphs (one for each of the considered ChIP-seq experiments) are integrated, exploiting these regulatory modules to build a genomic Transcriptional Regulatory Network. Moreover, since to each directed interactions is assigned a score representing the strength of the binding, the transcriptional relationships of the TRN will be weighted, with a relative measure of the interactions confidence.

3.2 TRN Bayesian modeling

The Bayesian framework conceived within this work exploits the multi-layered omics data integration to model large-scale transcriptional networks.

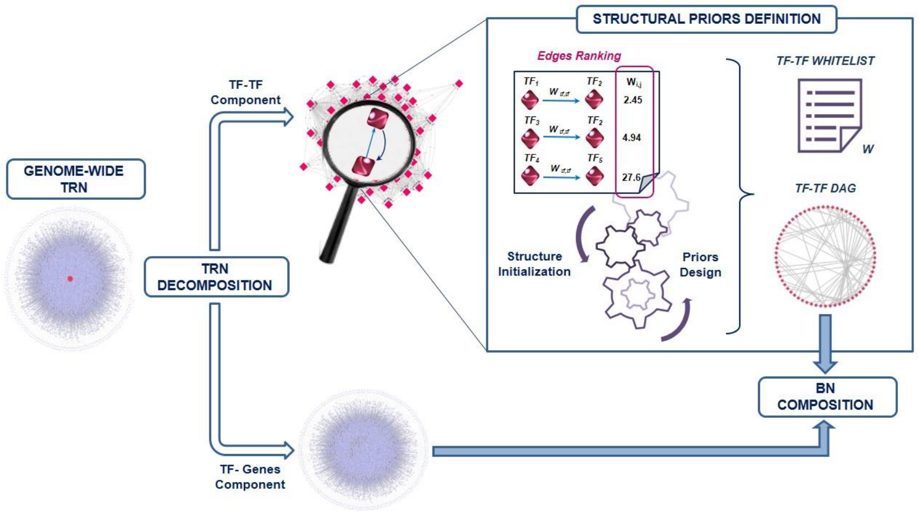
To this aim, a hybrid structure learning algorithm has been developed, able to use the data-driven transcriptional interactions as a prior knowledge. The algorithm also exploits integrated gene expression profiles for both assigning prior probabilities to each individual transcriptional relation, and for learning the model parameters during its search process.

3.2.1 Transcriptional Bayesian model definition

The Bayesian theory requires that the network which has to be modeled must be a directed acyclic graph (DAG), lacking of directed cycles or loops (see Sec. 2.4.1). Transcriptional networks are instead characterized by many loops of regulation, a classical property of the dynamic crosstalk among TFs, as already described in Sec. 2.2.3. To match this biological peculiarity with the acyclicity constraint, the proposed framework exploits the property of TRNs, whose regulations set can be divided in turn in two subsets, as described in the previous section, defining the interactions among TFs (which contains the regulatory loops), and among TFs and genes.

The approach for modeling a TRN into a Bayesian Network (BN) firstly decomposes such TRN into its fundamental parts, as shown in Fig. 3.4: a TF-TF Component, consisting of TF-TF edges, and a TF-Genes Component, consisting of edges from TFs to genes.

Since the loops issue is just included in the TF-TF component, it undergoes to an iterative process aimed not only at removing cycles, but also at initializing the model structure and defining the priors of the algorithm. Within this scheme, as first step, the procedure evaluates the type of edges among TFs, ranking and sorting them in decreasing order if they are weighted, otherwise it shuffles all the arcs and assigns an equal weight to them. In this case, all TF-TF edges are weighted and were ranked by their relative weight, the binding score, in a decreasing order. The process, then tries to remove one arc at a time, starting from edges with lower weight, and checking, at every iteration, if the TF-TF component is still connected. The procedure ends when a minimal connected DAG is found.



.

Figure 3.4: Transcriptional BN definition. Decomposition of a genomic Transcriptional Regulatory Network (TRN) allows to operate on the TF-TF component, characterized by many regulatory loops, i.e. feedback loops (as shown in the magnifying glass) to initialize the BN structure model and its structural constraints. The obtained DAG is then combined with the TF-Genes Component to define a genomic transcriptional BN.

All the TF-TF edges excluded from this structure initialization constituted an arcs whitelist (W), which will represent the search space of the possible structural models of the algorithm.

The resulting DAG is combined with the TF-Genes Component, to obtain again a genomic transcriptional network, but designed as a Bayesian model (TBN).

The second step for defining the initial BN, graphically depicted in Fig. 3.5, is represented by TBN integration with transcriptomics data that is gene expression (GE) profiles, in order to obtain a fully observable Bayesian network.

The underlying distribution is modeled as a joint multivariate Gaussian, where the conditional density of each variable (a TF, or a gene) given its parents, can be represented as a linear Gaussian model (see Sec. 2.4.1).

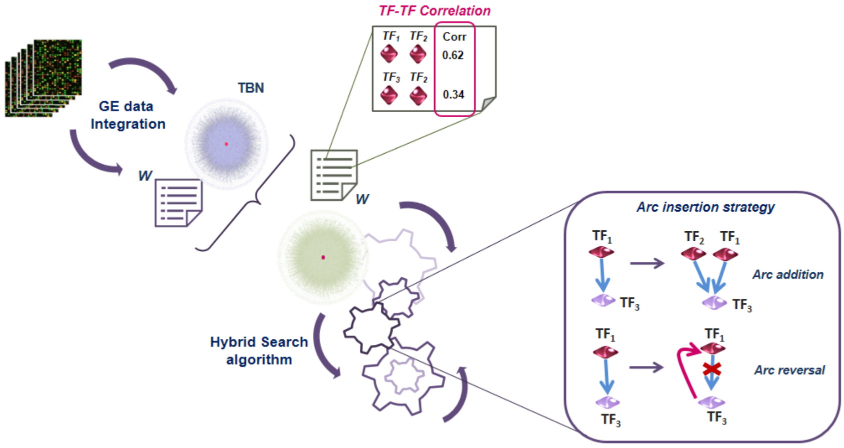


Figure 3.5: GE data integration step of the transcriptional BN. The GE data were integrated in the Transcriptional BN (TBN) and in the related arcs whitelist, defining them as inputs of the search algorithm. The box on the right highlights a peculiarity of the learning procedure, described in the Section 3.2.2.

Moreover, this omics data source is also used to calculate the correlation by the classic Pearson Correlation Coefficient among the expression values of TFs nodes included in the TBN. This measure, which defines the linear dependence of each TFs pair, is estimated as follows

$$\rho\left(TF_{1,}TF_{2}\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{TF_{1i}-\mu_{TF_1}}{\sigma_{TF_1}}\right)\left(\frac{TF_{2i}-\mu_{TF_2}}{\sigma_{TF_2}}\right) \qquad (3.1)$$

where $\mu_{TF_1}, \sigma_{TF_1}$ are the mean and standard deviation of TF1, respectively, and $\mu_{TF_2}, \sigma_{TF_2}$ are the mean and standard deviation of TF2, considering that each variable has N scalar observations.

This correlation is then assigned to each TF-TF interaction and is exploited as an extraction probability associated to each arc.

The algorithm scheme indeed plans on evaluating in the model each edge, belonging to the whitelist. The probability of an arc to be sampled is equivalent to the estimated correlation.

3.2.2 The hybrid structure learning algorithm

The developed algorithm follows the search and score paradigm, typical of the hybrid class of structure learning methodologies, as described in Sec. 2.4.2.

It proposes a heuristic search over the space of all possible structures derived from the whitelist, which encloses the informative structural priors concerning the TF-TF relations. Every extracted arc from the whitelist will be evaluated in the TBN, using the Bayesian Information Criterion (BIC) as scoring metric.

Its mathematical formulation is slightly different of that reported in the Eq. 2.11. Since the distribution of the TBN is assumed to be jointly multivariate Gaussian, in this context, the BIC score can be expressed in terms of the *residual sum of squares* (RSS)

$$BIC = n \log(RSS/n) + k \log(n) \tag{3.2}$$

where n is the number of observations (the GE dataset size), and k is the number of parameters in the model.

Moreover, supposing that each variable of the BN model is linearly dependent upon its continuous parents, we consider the TBN as the sum of all local models.

Thus, we modeled two BIC scores, a *local* one that is used to assess the local improvement in the network before and after a whitelisted arc addition, and a *global* one which represents the BN score computed as the sum of all BIC scores from local models, as shown in Eq. (3.3) and Eq. (3.4), respectively.

$$\begin{aligned} BIC_{local} = \Delta BIC = BIC_{old} - BIC_{new} \\ = n \log(RSS_{old}/RSS_{new}) - \Delta\, k * \log(n) \end{aligned} \tag{3.3}$$

$$BIC_{global} = \sum_{i=1}^{m} BIC \tag{3.4}$$

where m denotes the number of local models composing the transcriptional BN.

The second term in Eq. (3.3) is a penalty term that takes into account the edge changes; since many of the whitelisted arcs comes from TRN regulatory loops, the algorithm can add a new arc between two nodes ($\Delta k = 1$) or reverse the directionality of an existing BN arc ($\Delta k = 0$), as illustrated in the box of Figure 3.5.

All the steps of learning process are detailed below and presented in Fig. 3.6, in which is reported the pseudo code of the algorithm.

At each iteration, the algorithm randomly draws from the whitelist a group of arcs ($\underline{w}$) (i.e. one hundred) to test in the transcriptional BN (step 5). This sampling process is guided by correlation, which is exploited as an extraction probability associated to each whitelisted edge.

The algorithm adds every sampled arc, one by one, to the BN model, learns the model parameters from gene expression (GE) data, and evaluates the newly obtained BN using BIC$_{\text{local}}$ score. Since our learning schema is designed for parallel computing, all the arcs extracted from the whitelist are tested simultaneously.

Thus, the process evaluates all the computed $\mathrm{BIC}_{\mathrm{local}}$, selects as best model the solution that maximizes Eq. (3.3) (step 8), and then includes the corresponding arc into the model (step 9). The BN structure and its new score ($BICglobal_{new}$) are updated, and the process moves forward (steps 9-11) until the stop criterion (defined at the step 4) is met.

The algorithm ends its iterations when the new model score does not improve more than a fixed threshold compared to the score of the previous network ($BICglobal_{old}$).

Given the vulnerability of structure learning methods to getting trapped in a local optimal network during their search phase, the learning procedure provides also a strategy to prevent this problem (steps 13-14). When the stop condition is verified, the algorithm tries to move out of this potential local minimum for 10 consecutive times, combining an increased arcs sampling size (w_i) (i.e. the dimension of the whitelist is doubled) with a correspondingly augmented proportion of arcs to test. We considered the $\mathrm{BIC}_{\mathrm{global}}$ computed on the model before starting this procedure as $BICglobal_{old}$; if in any of these steps the new solution is not better than the old one, at the last iteration the algorithm stops, otherwise it accepts the new model structure and continues the search process. At the end of each algorithm run, the heuristic procedure returns as output a learned transcriptional BN.

Hybrid Structure Learning Algorithm

```
1: Procedure Hybrid Struct. Learning (TBN, D, W)
   Inputs:  TBN, transcriptional BN model containing the genomic variables $X_i$, $i=1, ..., n$
            D, Gene Expression Dataset representing the evidence for all $X_i$
            W, arcs whitelist
   Output:  TBN*, learned TBN
2: $BICglobal_{old} \leftarrow BIC_{global}$ estimation on TBN
3: cnt=0
4: while $(BICglobal_{new} - BICglobal_{old})/BICglobal_{new}$ > threshold
      %Phase I: Whitelisted arcs evaluation
5:    extract $\underline{w} \subset W$, where $(TF_i, TF_j) \in W$
6:    $B = \emptyset$
7:    for all arc $\in \underline{w}$ do
7.1**:    insert the arc in the TBN
7.2:      learn model parameters from D
7.3:      B(arc) $\leftarrow BIC_{local}$ estimation
7.4:  end for
      %Phase II: Update the TBN model
8:    $(TF_i, TF_j)_{best} = \max(B)$
9:    $tBN^* = tBN \cup (TF_i, TF_j)_{best}$
10:   delete $(TF_i, TF_j)_{best}$ from W
11:   $BICglobal_{new} \leftarrow BIC_{global}$ update on TBN*
      %Phase III: Escape from Local Minimum
13:   if ($BICglobal_{new} > BICglobal_{old}$) then
13.1:     if cnt < 10 then
13.2:         extract $\underline{w_i} \subset W$ where size($\underline{w_i}$)=2*size($\underline{w}$)
13.3:         cnt= cnt+1
13.4:         $\underline{w_i} = \underline{w}$
13.5:         continue   %go to step 6
13.6:     elseif cnt == 10
13.7:         end procedure
13.8:     end if
13.9: else
14:       cnt =0
14.1:     $BICglobal_{old} = BICglobal_{new}$
14.2:     continue   %go to step 5
14.3: end if
15: end while

**Possible arc operations ($\forall_i \forall_j, i \neq j$):
   1. arc addition: $TBN' \leftarrow TBN \cup (TF_i, TF_j)$
   2. arc reversal: $TBN' \leftarrow TBN \setminus (TF_i, TF_j) \cup (TF_j, TF_i)$
```

Figure 3.6: Pseudo code of the hybrid algorithm for learning a transcriptional BN structure

When learning BN structures from experimental data, the uncertainty about individual network structures has to be taken into account, especially in the absence of any gold standard network as for the human transcriptional context. For this reason, we delineated a "*consensus approach*" for the identification of structural consistencies across all the learned models.

3.2.3 Consensus Transcriptional BN definition

To assess the variability and the unavoidable uncertainty about the correct network structure, it is necessary to evaluate all added edges from learned TBNs in order to find a single consensus BN structure.

To this aim, a confidence threshold has been defined, considering it as the minimum degree of confidence for an edge to be significantly accepted in a final Consensus Bayesian Network.

For each learned TF-TF edge (*e_{ij}),* we compute its strength (w_{ij}) considering the BN models (m), in which this transcriptional relationship appeared, and their related scores (BIC$_{global}$).

$$w_{ij} = \sum_{m=1}^{n} \left(BIC_{global}(m)\right) \tag{3.5}$$

Edges with high confidence (significant edges present in more than half of the learned network structures, and in the best scenario, present in all the network structures) are strongly weighted and more likely to be included in the final consensus model.

The percentile distribution of the edge weights combined with the edge frequencies were used to rank all the considered arcs and to assess a confidence threshold, ensuring that the obtained transcriptional consensus BN is acyclic and fully connected.

3.3 TRN Inference

The systematic perturbation of transcriptional networks enables the elucidation of gene functions and regulatory relations that underlie biological processes. Current experimental methods of modulating transcriptional networks mainly rely on targeted single-gene overexpression (inducing the gene to increase its protein product, knockout (deleting its functionality), and knockdown (reducing of a certain threshold its activity).

Although these technologies provide powerful strategies for perturbing individual genes, they may not be suitable for global or combinatorial perturbation of transcriptional networks. Many complex diseases, as well as treatments required to counteract those conditions, may involve simultaneous or dynamic changes in the expression levels of many genes, which are not accessible by screens that target genes one at a time.

Moreover, without a specific hypothesis, the target of these in-vitro experiments potentially is each molecule of the system, and this may be unfeasible to do in practice.

To this aim, simulating *in silico* a perturbation of the system and propagate the effect of such intervention on the entire network, could help to get insights in the underlying regulatory modules.

We defined a *perturbation model* for each TF included in the consensus transcriptional BN (TBN), implementing a knockout effect on the transcriptional expression, to allow the investigation of TFs influence.

Since the distribution of the modeled TBN is assumed to be a multivariate Gaussian, the conditional probabilities have the form of a Gaussian model, as described in Sec. 2.4.1, and can also be view as a set of regression equations.

If J is a set of nodes, then denotes the vector of variables indexed by J. In the following notation, $X_{C(J)}$ are the conditioning variables of X_J.

Each conditional variable $\{X_J|X_{C(J)}\}$ has a normally distributed mean, defined as $(\mu_J + \sum_{k\epsilon C(J)} b_{kJ}(X_k - \mu_k))$, variance v_J (fixed for a given set of conditioning variables), and linear coefficients b_{kJ}; the resulting conditional model for *j=1,…,n* is given by

$$X_J = \mu_J + \sum_{k\epsilon C(J)} b_{kJ}(X_k - \mu_k) + (v_J)^{1/2}Z_J) \qquad (3.6)$$

in which $Z_1,\ldots,Z_n$ are independent standard normal random variables. The matrix **B**=$[b_{kJ}]$ can be thought as regression coefficients [85].

In the model, when the mean of a variable changes (i.e. if the node is perturbed), and if this node has a successor, the mean of this last one changes consequently.

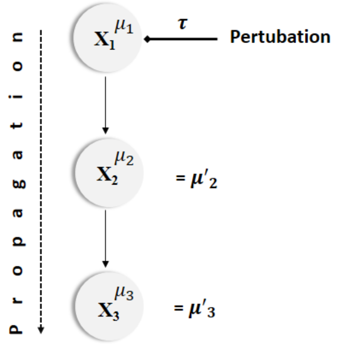


Figure 3.7: Perturbation propagation model

If the mean of the perturbed node X_1 changes from μ_1 to μ'_1, the new value for μ_2 is

$$\begin{aligned} \mu'_2 = \mathrm{E}[X_2] = E[E[\,X_2|X_1] = E[\mu_2 + b_{12}(X_1 - \mu_1) \\ = \mu_2 + b_{12}(\mu'_1 - \mu_1) \end{aligned} \qquad (3.7)$$

The mean μ_3 of the X_3 vertex, which is the successor of X_2, also changes as a result of the μ_2 variation.

More generally, if the mean of a node i is perturbed, the effect of this change (τ) is propagated and can be summed along every directed path which emanates from node i [85].

In this way, instantiating a knockout effect on the expression/function of each TF included in the consensus TBN that consists in setting to zero the expression value of a TF, the perturbation is then propagated along each its transcriptional regulation, resulting in an expression change of its genomic targets (both TFs and genes).

Through the calculation of the marginal probability (see Sec. 2.4.4) on the network nodes, the effect of such perturbation is then estimated, comparing the mean before and post the "stimulus" induction.

$$\mu_{pert} = \mu_{post} - \mu_{pre} = \Delta\mu \tag{3.8}$$

For each node, we can estimate the perturbation effect (τ) in terms of expression value changes ($\Delta\mu$).

Considering the scale of the perturbation impact on μ_{post} and comparing it to μ_{pre} in the context of this probabilistic model, it is possible to evaluate the type of regulation exerted by a certain TF on the considered target (i.e. potential repression or activation). If $\mu_{post} > \mu_{pre}$, we could infer that the TF originally has an inhibitory effect, otherwise ($\mu_{post} < \mu_{pre}$,) it acts as an activator.

Moreover, for each perturbation model, we also obtain the distribution of τ from $\Delta\mu_n$ calculated for all nodes (n) of the network. Using a variation threshold applied on the model standard deviation (σ_{pert}), the perturbed targets (PTs) can be ranked and filtered in order to identify the signature of the knocked TF.

3.4 Method Implementation

The implementation of the described methodology for parallel execution has been done in Matlab, tested both on a standard PC (P7 CPU 4.0 GHz, 32 GB RAM), and on a high performance computing environment, the HiPerGator 2.0 cluster (30,000 Intel cores, with 4 GB RAM per core).

For perturbation model simulations some code from Bayesian Network Toolbox (BNT, Murphy K. [86]) has been exploited.

Data preprocessing, file management, and integration aimed at creation of the molecular networks used in the following were performed with Python scripts that integrate several functions of several libraries, in particular Bioinformatics extension of the Orange Data Mining Suite [97] and NetworkX [98] packages. This last one allows to create network in Cytoscape compatible format.

All the networks represented and topologically analyzed in this work has been visualized with Cytoscape (v.3.3.0) [99], an open source bioinformatics platform for the visualization, integration and analysis of molecular interaction networks.

Chapter 4

Validation of the Methodology

The peculiarities of our novel approach optimized for learning large-scale transcriptional BNs make finding other similar methods difficult, especially in the class of hybrid BN learning algorithms, which exploit prior knowledge, GE data and directed regulations but without forcing the search process with severe constraints.

To evaluate the performance of our method, we selected SAGA algorithm [100], the only approach with some common grounds with our strategy, and ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks), which is the most widely used technique for regulatory network reconstruction from gene expression data [101].

For the aforementioned reason, despite some common features, we excluded a recent Bayesian structure learning tool, Bnlearn [102] due to its strict way to specify the structural priors, inconsistent with our methodology.

The validation has been accomplished using data from yeast *S. cerevisiae*, since only a few experimentally verified eukaryotic transcriptional networks are available as gold standards, like yeast and *E. coli*. This last one has a transcriptional network not sufficiently large and complex to apply our hybrid learning strategy, and is poorly enriched of TFs coregulations.

In the following sections, a briefly introduction of the regulatory network reconstruction approaches considered in this work is provided.

Given the inherent noisy nature of omics data sources, which potentially contains incorrect information, the robustness of the proposed Bayesian data fusion based approach has been tested to increasing percentages of false priors and compared to the other selected strategies. The obtained results from such comparison are then reported.

4.1 Competing methods

A brief overview of the regulatory network reconstruction methods selected for the comparison is provided.

SAGA - Banjo

SAGA is a hybrid Bayesian learning algorithm, implemented in the Banjo (Bayesian Network Inference with Java Objects) software [103], which combines Simulated Annealing with a greedy search, using Bayesian Dirichlet equivalence as a scoring metric to evaluate the generated network.

It allows arc addition and reversal, and the possibility to specify a structural prior as well as a list of forbidden arcs that must not be added (blacklist) to the model. This method does not exploit an arcs whitelist strategy, but it infers the network structure from discretized gene expression data. Banjo ends its search when one of the termination criteria are met (i.e. fixed number of explored networks, search threshold time, maximum number of restarts reached), and returns as output the learned network with the best score.

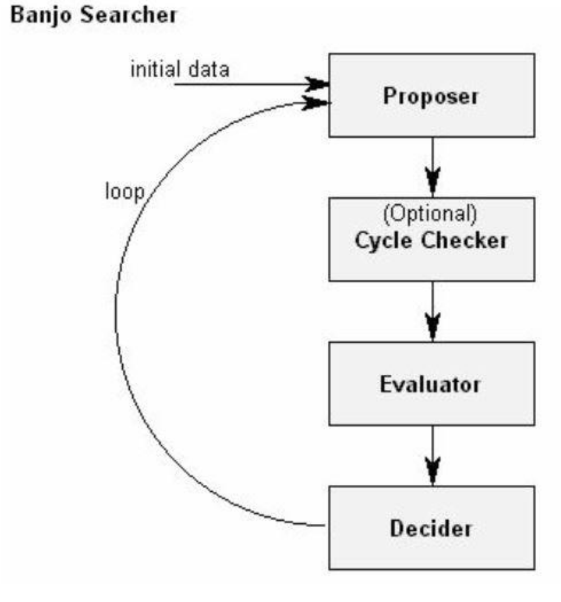


Figure 4.1: BANJO components

Within the search loop (**Searcher** is the core of the Banjo algorithm), Banjo allows various combinations of Proposer, CycleChecker, Evaluator, and Decider components to handle the aforementioned aspects of each iteration step.

The first step, **Proposer**, implements the SAGA algorithm which searches a graph structure (G_{rough}) to be evaluated according to the data. After a change in the graph, G_{rough} is proposed, then it is scanned for cycles, through the **Cycle Checker**. If it contains a cycle, G_{rough} is discarded, and the search goes back to the Proposer to request another network change; if not, G_{rough} goes to the next step. The acyclic graph *G* is then evaluated, **Evaluator**, according to the scoring function described above. The **Decider** considers, possibly stochastically, whether to accept the proposed network (as the new current network) and best scored networks are then reported.

ARACNe

ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) is an information-theoretic based approach that implements a Gaussian Kernel estimator [104] to estimate the joint probability distributions and then computes the mutual-information (MI) between all pair of variables, using the calculated pairwise MIs to build up the gene network. Considering two variables X and Y, the MI is calculated as follow

$$MI(X;Y) = \frac{1}{M}\sum_{i=1}^{M} log\left(\frac{\hat{f}(x_i y_i)}{\widehat{f(x_\imath)}\widehat{f(y_\imath)}}\right) \tag{4.1}$$

MI(X;Y) is equal to 0 if and only if X and Y are statistically independent. In experimental setting, the estimated MI never equals zero. Under this scenario, the recovered gene network would be full connected (each gene is connected to all the other genes of the network).

To remove redundant hence false predicted connections among genes, ARACNe implements a bootstrap strategy that allows to compute a random MI given the number of observations. This approach allows to set a threshold that discriminates between statistically dependent and independent pairs of genes, given the data. The threshold over the MIs consents to remove most of the false positives predicted interactions. To this aim, the *Data Processing Inequality* (DPI) has also been implemented.

The data processing inequality in information theory states that given three random variables X, Y and Z, then the mutual information between X and Y is greater than or equal to the mutual information between X and Z. That is *MI(X;Y) > MI(X;Z).*

In this context, it is used as a pruning strategy, in order to remove indirect interactions, scanning all the full connected triplets of genes in the network and removing the recovered connection with lowest MI.

For our test, we used the last version of this algorithm, ARACNe-AP [105], that works on reconstructing transcriptional networks taking as inputs a GE dataset and a predefined list of regulators (TFs). Its strategy consists of computing MI for every TF/target pair, without estimating it for all pair of network variables, and reconstructing MI networks from bootstrapped GE samples. A consensus network is then generated from the significant edges detected across all bootstrap runs.

Bnlearn

Another tool for learning BN structures and estimating their parameters is the R package Bnlearn, which however cannot be used for our purposes. It implements a hybrid algorithm, the Max-Min Hill-Climbing, to reconstruct the network from GE data, combing network reconstruction with a Bayesian-scoring greedy hill-climbing search to orient the edges. It allows to specify a structural prior in the form of a DAG, but it forces all the arcs designed as priors to be included in the final model, preventing the addition of any other extra arc. This constraint makes this approach not appropriate to handle transcriptional network problems, as regulatory loops, and for this reason, it has been discarded from our comparison.

4.2 Yeast transcriptional benchmark network reconstruction

Due to the scarcity of validated Eukaryotic transcriptional networks, regarded as gold standards, among the existing model organisms, yeast *S. cerevisiae* regulatory interactions have been deeply investigated in recent years. This information is retrieved in yeast databases, such as Saccharomyces Genome Database (SGD) [57] and YEASTRACT [58].

We retrieved all available transcriptional regulations in yeast among known TFs and target genes, which map only verified ORFs (Open Reading Frames, which identify the codifying portion of the DNA sequence) from both repositories. The resulting regulatory interactions were then integrated to obtain a complete transcriptional relationships set.

As transcriptomics source, we used the normalized GE dataset from the study of Spellman et al [106], considering only those genes identified by the authors as cell-cycle regulated. Given the presence of missing values, we, selected only those genes with a missing values rate less than 10%. Then, a k-nearest-neighbor imputation has been performed, obtaining a final complete dataset of 473 cell-cycle related genes expressed in 77 samples.

Combing the validated transcriptional binding information with GE data, we defined as ground truth a yeast regulatory network (yTRN) composed of 33 TFs and 437 target genes, and 3,299 transcriptional regulations, as illustrated in Fig. 4.2.

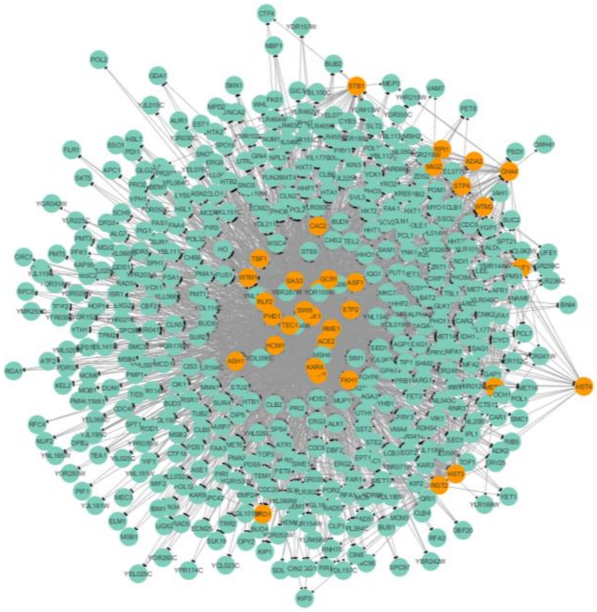


Figure 4.2: The reconstructed yeast transcriptional regulatory network (yTRN)

4.3 False prior information setting

Our Bayesian data fusion based approach exploits a data-driven prior knowledge, and as previously described, the inherent data nature is noisy and potentially contains incorrect information.

To test the robustness of our method to false priors, we randomly added an increasing number of false edges to the yTRN, from 10% to 60% of the total number of TF-TF regulations. We considered each known interaction as true positive (TP), and every additional incorrect arc as false positive (FP). The performance of our method was evaluated for each FPs percentage and then compared to Banjo and ARACNe-AP.

The yTRN underwent the BN definition procedure, as illustrated in Sec. 3.2.1, and is decomposed in its fundamental parts: a TF-TF Component, characterized by 33 TFs and 249 interactions, and a TF-Genes Component with 470 nodes and 3,050 edges. Using the option for unweighted transcriptional data, the TF-TF Component underwent to the iterative process in order to obtain an initial DAG with 33 nodes and 32 TF-TF arcs. Combining it with the other Component, the structure of the initialized model of 470 nodes and 3,082 edges is defined.

This starting TRN and the arcs whitelist, whose dimension varied according to the considered FPs rate, as described in Table 4.1, were used for testing the proposed approach in all of the six incorrect prior conditions.

Table 4.1: Incorrect priors rates tested for the method validation

	Edges whitelist	
Tested FPs rate	**#of FPs**	**Total #of edges (FPs+TPs)**
10%	25	242
20%	50	267
30%	75	292
40%	100	317
50%	125	342
60%	150	366

This table summarizes the number of FPs and TPs edges included in the whitelist for each considered FPs rate

4.4 Robustness evaluation to false prior results

We collected 100 learned transcriptional BNs for each tested FPs percentage, for which we evaluated the computational performance of our learning schema, considering the time used by our algorithm to learn all the obtained networks, as illustrated in Fig. 4.3. The average computational time estimated on the total number of transcriptional BN models for all the FPs levels varied from 1.61 minutes to 2.00 minutes.

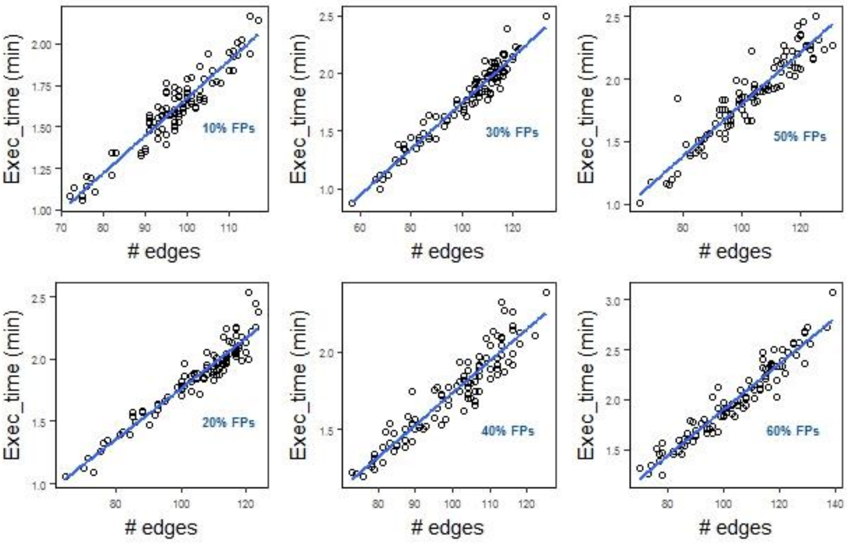


Figure 4.3: Performance evaluation of the hybrid structure learning algorithm on yeast network. For all the considered FPs levels, we analyzed the execution time of the search strategy to learn the BN model structures, comparing it to the number of edges of each learned network.

We then applied the “consensus” approach, described in Sec. 3.2.3, on each set of learned networks. For all the analyzed FPs percentages, we selected the 25th percentile of the arcs weights distribution to find the confidence threshold which guarantees a full connectivity of the related yeast Consensus BNs (yTBNs).

To examine the performance of our algorithm throughout all FPs tested levels, we considered the percentage of FPs added in the final consensus yTBN comparing it to the total rate of FPs enclosed in each whitelist, reported in Table 4.1. The number of FPs edges added in the every final model is low and it remains so despite the increasing false priors available in the whitelist, as shown in Fig. 4.4, and then in Fig. 4.5 and in Table 4.2.

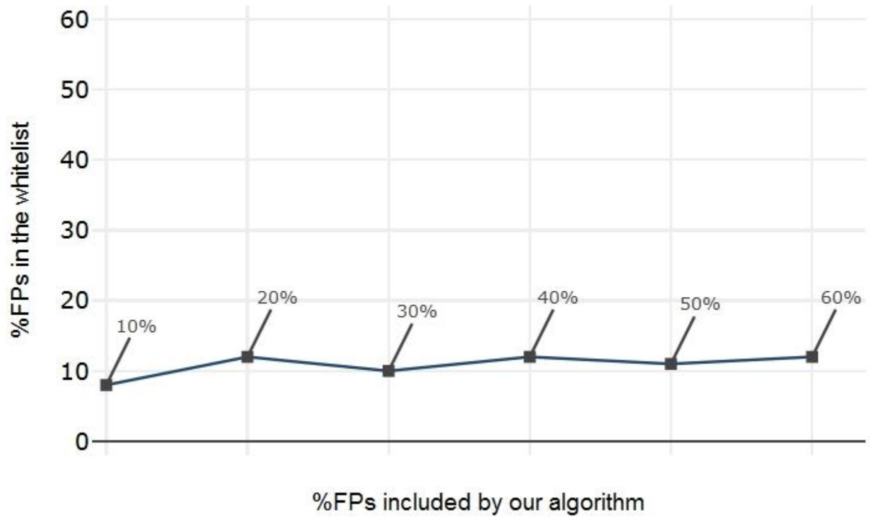


Figure 4.4: Robustness of our method to erroneous priors. For each considered FP rate (reported as percentage on the performance line), we reported the percentage of FPs added in the final consensus BN compared to the total rate of FPs enclosed in each whitelist. We can conclude that the method is quite robust to wrong prior information.

Application of Banjo algorithm

In order to estimate the joint probability distribution of all the variables in the network, Banjo first discretizes GE data using a quantile discretization procedure. The Proposer/Searcher strategies were set to random local move and simulated annealing, respectively. The amount of time Banjo uses to explore the yTBN space was set to five hours (designated as stop criterion), since this time window has been indicated by the author as optimal to reach the highest sensitivity [103]. All the other parameters such as reannealingTemperature, coolingFactor, and so on, were left with their default values. The parameters setting can be found in the Appendix Sec. A.1.

Banjo was evaluated in each incorrect prior scenario taking as input data the discretized GE yeast dataset, the same initial structures (i.e. DAG structures characterized by TF-TF interactions) exploited by our approach, and a blacklist, to avoid gene-gene interactions and unrealistic regulations from genes to TFs. All results are summarized in Table 4.2 and in Fig. 4.5, where the comparison of Banjo and our developed method is represented.

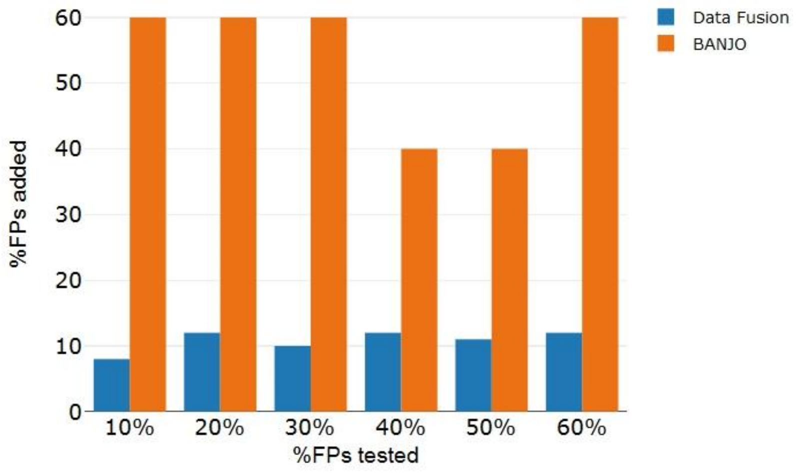


Figure 4.5: Performance comparison of the proposed Data Fusion approach and Banjo. We analyzed the FPs edges added in the Consensus BNs for each FPs rate tested with our hybrid structure learning strategy and Banjo.

Application of ARACNe-AP algorithm

ARACNe-AP cannot be evaluated under these incorrect prior conditions since it infers the network structure using the GE data and a list of regulators (the considered 33 yeast TFs).

Once specified the input data, the algorithm calculates the MI threshold using the input GE matrix, computes 100 reproducible bootstraps from gene expression samples, and then consolidates these results using an implemented edges significance test to return a final consensus network.

A MI threshold of 0.2989 has been calculated on data and the obtained yeast consensus network after the bootstrap step is composed of 348 nodes and 1,003 interactions, whose representation is shown in Fig. 4.6

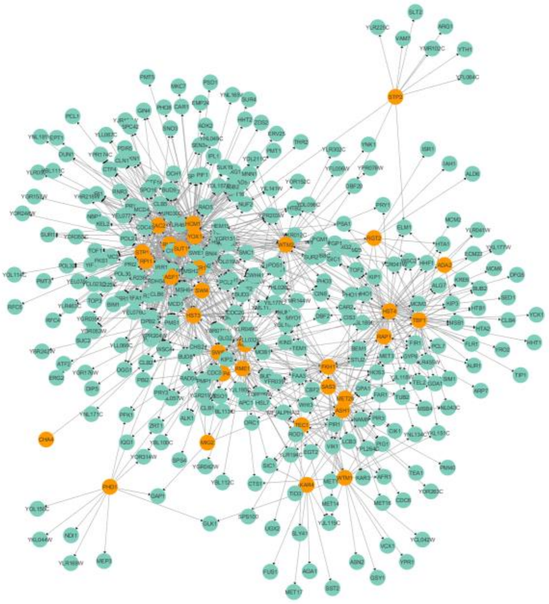


Figure 4.6: The consensus yeast TRN reconstructed by ARACNe-AP

Due to its implementation, ARACNe-AP does not allow to specify a structural prior. In this way, to determine its performance on our input data, we compared the interactions of this network with those included in the reconstructed yTRN, as described in Sec. 4.2, to estimate the number of TPs and FPs, whose percentage is reported in Table 4.2.

Table 4.2: Methods comparison results

	DF approach		**BANJO**		**ARACNe-AP**	
FPs rate	#of Consensus edges	Added FPs	#of Consensus edges	Added FPs	#of Consensus edges	Added FPs
10%	50	8%	69	60%	1003	70%
20%	58	12%	69	60%		
30%	56	10%	69	60%		
40%	60	12%	69	40%		
50%	69	11%	69	40%		
60%	76	12%	69	60%		

As shown in Table 4.2 and in the figures above, our Bayesian data fusion based approach is robust to the increasing amount of false positive prior information. From the comparison with Banjo and ARACNe-AP, it can be pointed out that our algorithm outperforms both methods, producing a significant improvement in structural accuracy, even with a progressively higher FPs rate.

ARACNe-AP bases its structural reconstruction only on a single source of data (GE data), and this penalizes the correctness of the inferred transcriptional relations, 70% of which are false positives predicted interactions.

On the other hand, Banjo allows to specify a structural prior, but its implemented constraints and parameters, whose setting is not trivial (i.e. initialTemperature, coolingFactor, reannealingTemperature, etc.), does not enable to acquire an accurate learning. Moreover, it has a limitation on the maximum number of parents allowed for each network node. As the author advised in Banjo documentation, this criterion must be less than 7 for memory requirements needed for the learning. Banjo requires also a list of forbidden arcs, to avoid the insertion of interactions from genes to TFs, whose definition for large-scale transcriptional networks is equivalent to 2^n (where n is the number of genes) interactions to exclude.

The search schema of our approach does not impose a constraint on the number of interacting variables or on the number of parents for each variable, and it is fast and scale well as illustrated in Fig. 4.3, thanks to its learning schema based on local learning executable in parallel.

Chapter 5

Results from the Data Fusion approach applied to the CML case

In this chapter, we present the results obtained by applying the proposed approach to the case of Chronic Myeloid Leukemia, starting from the CML transcriptional regulatory network reconstruction, and passing through its Bayesian network modeling to probabilistically assess the underlying structure with the hybrid learning algorithm. Finally, after a consensus transcriptional interactome has been defined, it can be exploited as a predictive perturbation model, aimed at investigating the hidden transcriptional signatures.

5.1 CML genome-wide transcriptional regulatory network

The CML transcriptional regulatory network is represented by all transcriptional interactions identified along the genome, inferred from the analysis of ChIP-seq experiments, available on the ENCODE repository for the K562 leukemia Tier 1 cell line, specific for the considered disease.

Despite the quality criteria established by the ENCODE Consortium for data publication, not all the required standards has been respected. To better control the noise and the experimental variability among replicates of the same sample, further quality criteria, described below, have been applied for selecting the data to analyze.

- *Sample accessibility*: raw ChIP-seq experimental data in BAM format.
- *Control sample availability:* each ChIP-seq experiment must have a corresponding control experiment.
- *Biological replicate availability:* each sample must have a minimum of two biological replicates to assess its experimental variability.
- *Sample treatment:* no pharmacological treatment has been administered.
- *Samples sequencing depth (for both "case" and control samples)*: 20 million usable fragments.
- *Priority on the laboratory that produces data*: if a ChIP-seq experiment for a certain TF is available from two different labs of the consortium, the lab with more complete metadata and with data matching the aforementioned filters is preferred.

Applying such criteria, 65 TF ChIP-seq experiments has been considered, each one isolating binding data of a specific TF, to retrieve the related raw data, controls and biological replicates.

TFs experiments were then analyzed with the integrative bioinformatics pipeline, described in detail in Sec. 3.1.1, whose steps performed peak calling, replicates evaluation, peak significance analysis, annotation of promoter-overlapping peaks, and finally a quantitative weighting of TF-target interactions. The final aim of this procedure is to filter and integrate the transcriptional information underlying each ChIP-experiment for reconstructing the binding profile of every TF along the genome. Moreover, the statistical constraints set across the analysis steps allows to discard potentially false interactions, as represented by the numbers in Table A1 of the Appendix Sec. A.2. This Table also reports the TFs list analyzed in this study.

Every obtained omics binding profile represents the *regulon* of each considered TF, in other words, the group of genomic targets regulated by the analyzed transcription factor. All TFs regulon size is illustrated in Fig. 5.1 below. On average, the number of target nodes regulated per TF is 7,363.

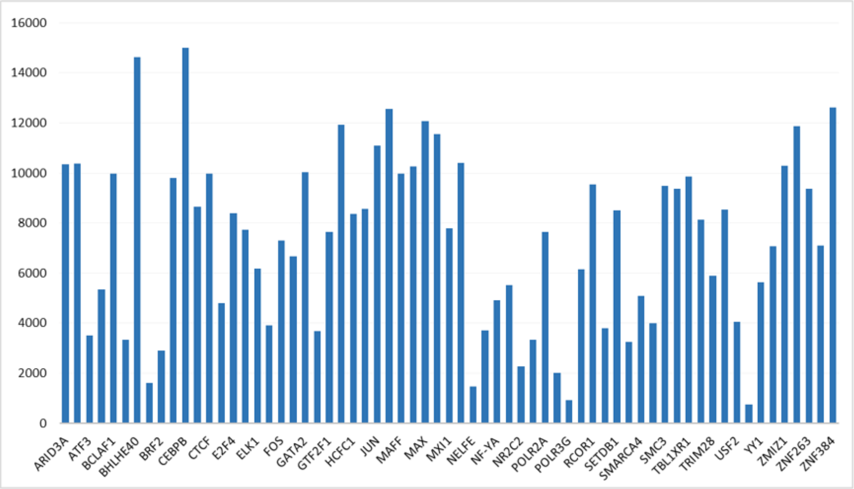


Figure 5.1: Regulon size of each analyzed TF.

A functional annotation of the considered TFs has been performed using the Epifactors database [107] and GO molecular functions terms, whose results are annotated in Table A2 of the Appendix Sec. A.3. In summary, among the analyzed 65 TFs, 27 of them have a sequence-specific binding (TFSS) with a particular molecular motif, 12 TFs exert also an epigenetic function, 11 TFs have a role combining the TFSS and epigenetic functions, and 13 TFs have a general function, as cofactors of the transcriptional machinery.

Since each transcriptional interaction of all regulons is weighted with the binding score (*bs*), introduced during the integrative analysis, the score distribution along these relationships is represented as follows

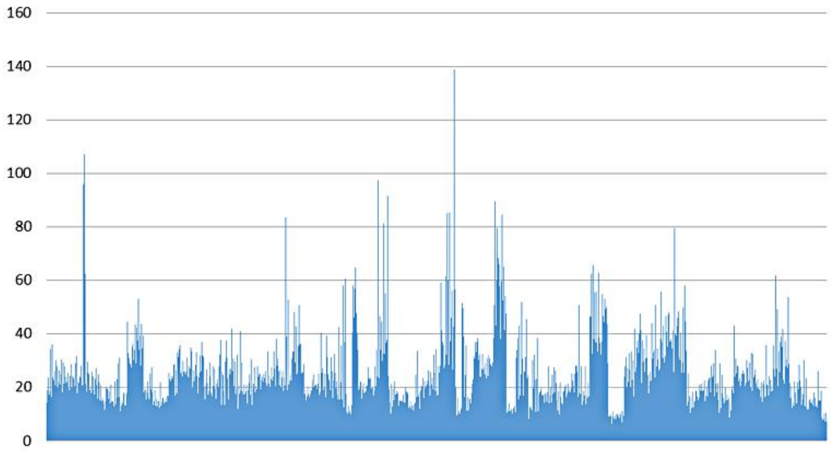


Figure 5.2: Binding scores distribution.

Each regulon can be described through a graph, with the considered TF as a single regulator node and all its genomic targets as regulated nodes, as depicted in Fig. 5.3. The underlying transcriptional interactions are symbolized by directed and weighted edges from the regulator TF to its targets (which can be both genes and TFs), and the weight (*w*) is the *bs*, directly proportional to the strength of the transcriptional relation.

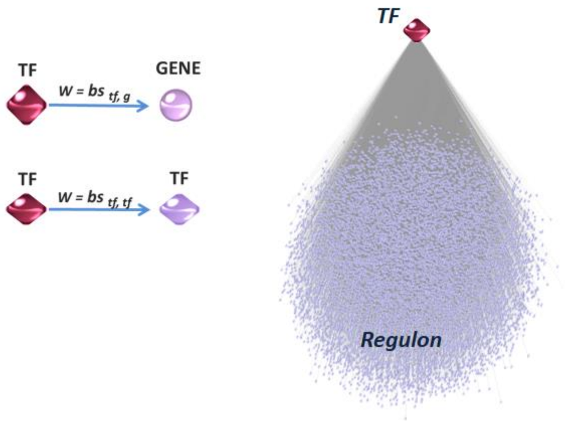


Figure 5.3: Graph representation of an omics TF binding profile.

We obtained a graph for each TF profile, and exploiting the property of transcriptional networks, in which TFs have a synergic behavior, regulating each other as described in Sec. 2.2.3, all the resulting graphs were computationally integrated in order to build a genome-wide Transcriptional Regulatory Network for the considered disease context. The reconstructed CML genome-wide TRN is illustrated in Fig. 5.4.

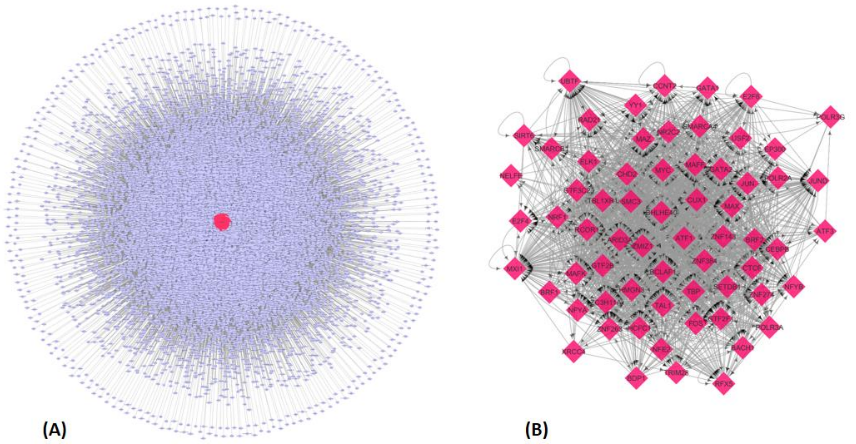


Figure 5.4: The reconstructed genome-wide TRN for CML. (A) the genomic TRN; (B) TFs subnetwork, representing the core of the TRN.

This network reflects the genomic landscape of transcriptional targets, since it is composed of 20,876 nodes (65 TFs and 20,811 target genes), and 478,558 directed and weighted edges. The amaranth colored nucleus represents the TFs core, whose subnetwork is illustrated in Fig. 5.4 (B). This core is characterized by 1,857 coregulatory interactions, and 30 of which are autoregulation, self-controlled regulation operated by the TF itself.

Through the genomic annotation step of the analysis pipeline, some preliminary functional considerations can be done on the represented interactome, as reported in Table 5.1.

Table 5.1. Functional nodes of the TRN

Nodes type	#of Nodes
TFs regulator	65
Protein coding genes	19,427
Other TFs target	177
miRNA coding genes	1,207

Among the 20,876 nodes, there are 19,427 protein coding genes, 177 nodes are genes which codify for TFs, for which the transcriptional binding profile is not available, and 1,207 of the total number of nodes are miRNA coding genes, which are regulated by all the considered 65 TFs.

Given the high dimensional space of the transcriptional regulations in such interactome, it is not possible to make inference, since each TF from a topological point of view, is equally important to another TF of the network, as shown in Fig. 5.5 with the in–coming connectivity (In-degree) distribution that can be approximated by an exponential fit.

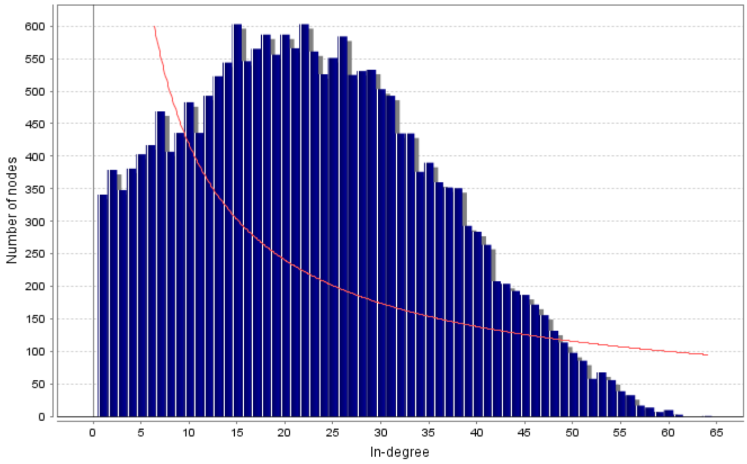


Figure 5.5: In-Degree distribution of the TRN.

The out-coming connectivity (Out-degree) distribution instead is slightly different from zero, since the TRN aim is to include only regulations that start from TFs, excluding the interactions among genes.

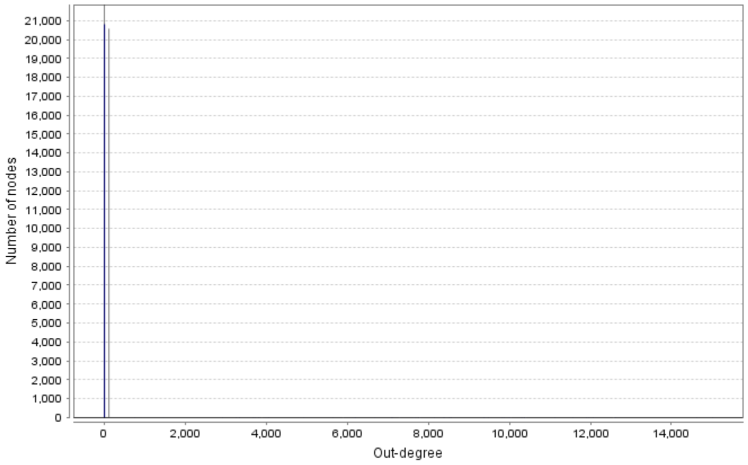


Figure 5.6: Out-Degree distribution of the TRN.

The compactness of the network is also demonstrated through Table 5.2, reporting some topological metrics calculated on the entire network, and only on the TF-TF component, where the core of the transcriptional relationships is embedded.

Table 5.2: TRN and TF-TF subnetwork topological metrics

Network metrics	On TRN	On TF-TF subnt.
Connected components	1	1
Clustering coefficient	0.562	0.546
Network diameter	5	4
Avg. number of neighbors	45,803	41,877

The diameter is relatively small if compared to the huge number of existing nodes in the TRN, highlighting a graph compactness, as shown by the connected components measure, and by the intermediate value of the clustering coefficient, whose values range from 0 to 1.

Analyzing the Betweeness Centrality distribution of the TF-TF subnetwork, we have to take into account that the higher the value, the higher the relevance of the TF as organizing regulatory molecule in the network. As illustrated in Fig. 5.7, there is only one TF with a high BC, which however is a RNA polymerase, and not functionally informative for the considered disease scenario. Most of TFs relies instead on the same BC range that is equivalent to the same transcriptional importance.

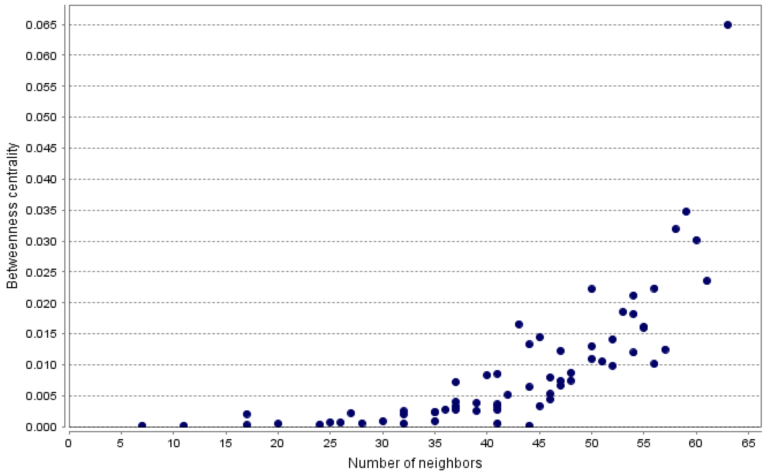


Figure 5.7: Betweenness Centrality distribution among TFs nodes.

5.2 From a genomic TRN to a Bayesian Network model

As described in the section above, in this "hairball" regulatory network no inference can be done to investigate the transcriptional impact of each TF.

To this aim, the TRN is transformed into a probabilistic model, and underwent to the Bayesian network (BN) definition process.

As first step, we dissected this genomic network into a TF-TF Component, characterized by 1,827 edges among TFs, excluding the loops of autoregulation, and a TF-Genes Component, which included the remaining network edges.

Applying the BN design process (see Sec. 3.2.1) to the TF-TF Component, all the weights of the arcs were sorted in decreasing order and ranked according to their binding score values. The procedure tried to remove one arc at a time, starting from edges with lower weight, to find a minimal connected DAG.

At the end, we obtained a whitelist of 1,763 transcriptional relations and a minimal connected DAG, defined by 64 interactions. This DAG combined with the TF-Genes Component constituted the initial BN model, as depicted in Fig. 5.8.

Transcriptomics integration

Given the integrative scheme of our method, a second omics source of data is necessary to integrate the obtained BN.

A gene expression (GE) compendium from microarray data of 122 CML patients is generated, deriving it through the integration of five GE datasets, retrieved from GEO and ArrayExpress databases (GEO accessions

GSE13159 [108], GSE47927 [109], GSE24739 [110]) (ArrayExpress accessions E-MTAB-2581 [111], E-MEXP-480 [112]) in CEL format. All transcriptional raw data were RMA normalized [113] and expressed on $\log_2$ scale. All transcript probes were then annotated with the relative Gene Symbol. Probes mapping the same gene were median averaged, and all of them lacking of functional annotations (i.e. control probes, probes mapping uncharacterized loci) were discarded. In order to obtain a unique gene expression panel, only those genes expressed in all the evaluated profiles were retained in the final dataset. These steps of analysis were performed by *limma* package in R environment. [114].

The obtained transcriptomics compendium was integrated in the transcriptional BN, with the purpose of achieving a fully observable network, whose underlying probability distribution is modeled as (conditionally) Gaussian (see Sec. 3.2.1).

Moreover, this omics source was also used to calculate the correlation among the expression values of TFs included in the network. This measure is assigned to each whitelisted arc, and will be exploited as a sampling probability by the learning structure algorithm, as detailed in Sec. 3.2.1 and in Sec. 3.2.2. The whitelist became a *correlation whitelist* (c_w), and with the initialized GE integrated BN (TBN), composed by 11,986 (60 TFs) nodes and 282,533 edges, represented the input of the hybrid structure learning algorithm.

These steps are summarized in the Fig. 5.8.

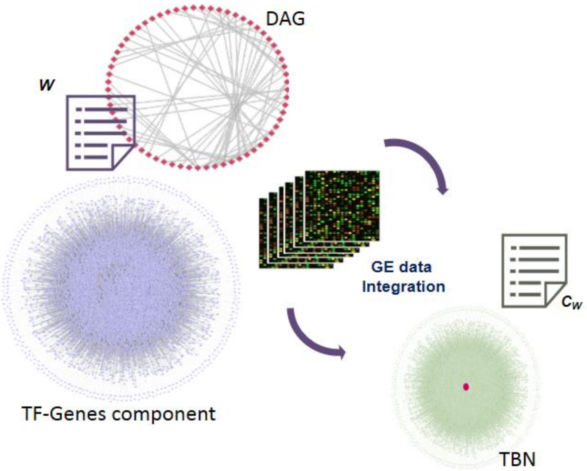


Figure 5.8: TRN conversion into a Bayesian model

5.3 Structural assessment on CML transcriptional BN

The obtained TBN and the related correlation whitelist underwent to the hybrid structure learning process of the proposed algorithm.

After 100 runs, we collected 100 transcriptional BN models. The computational time required for learning all the obtained genomic networks is shown in Fig. 5.9, with an average learning time of 2.5 hours to obtain a final genome-wide TBN within a complete run of the algorithm.

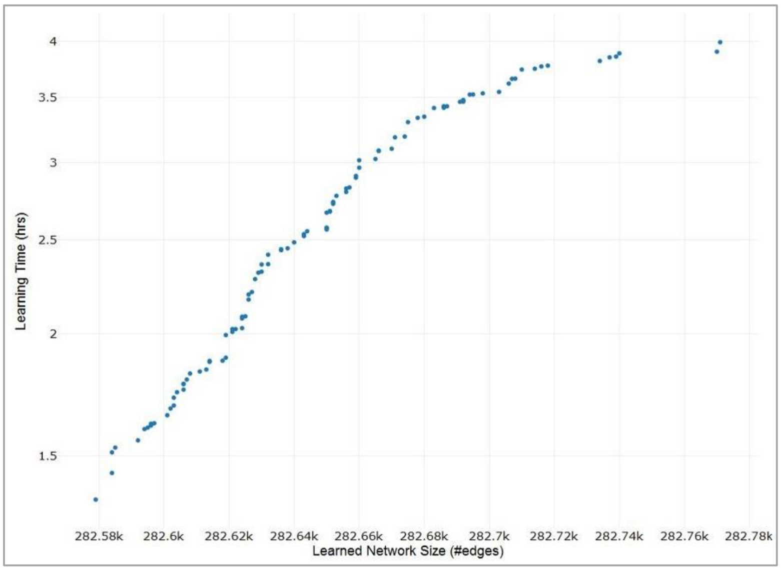


Fig. 5.9. Performance evaluation of the hybrid structure learning algorithm on the CML network. The execution time of the search strategy to learn the BN model structures is compared to the number of edges of each learned genomic network.

In order to obtain a consensus transcriptional BN, we ranked and weighted all the TF-TF relations from the learned network structures, following the approach described in Sec. 3.2.3. We chose as confidence threshold the weight value corresponding to the 5^{th} percentile of the arcs weights distribution, to avoid the inclusion of edges with low confidence.

The resulting genome-wide consensus TBN is composed 11,986 nodes and 282,544 transcriptional interactions. The TF-TF core is defined by 70 TF-TF edges, 30 of which had been reversed by the algorithm, as an effect of TRN regulatory loops. The topological metrics estimated on the TBN are reported in Table 5.3.

Table 5.3: Consensus TBN and TF-TF subnetwork topological metrics.

Network metrics	On TBN	On TF-TF subnt.
Connected components	1	1
Clustering coefficient	0.025	0.0
Network diameter	7	6
Avg. number of neighbors	47,146	2,333

Analyzing the incoming and the outcoming connectivity of each consensus TF node, we compare the difference between out- and in- degree (O-I), which measures the direction of the transcriptional information flow within a metric called *hierarchy height (h)*, introduced in the study of Gerstein M.K. et al [115]. With possible values ranging from -1 to 1, this metric provides a normalized measure of the disparity between a given TF's roles as a regulating factor and a regulated target.

Specifically, it is calculated by normalizing the difference between the out- and the in-degrees by the sum of the out- and in-degrees.

$$h = \frac{(O - I)}{(O + I)}$$

Lower h values indicate that a TF is heavily regulated and without many targets of its own (i.e., it is lower within a regulatory hierarchy), whereas higher h values indicate that a TF is a regulator of many other elements, and without many other elements responsible for its regulation (i.e., it is higher within a regulatory hierarchy).

In other words, TFs with no in-degree (i.e., those regulated by no other TFs) have a range values of (0.5, 1], using this metric. TFs with no out-degree (i.e., those regulating no other TFs) fall in this values interval [-1, -0.5). TFs with balanced regulation (i.e., those regulated by the same number of TFs that they themselves regulate) have a range of [-0.5, 0.5] using this statistic.

The resulting distribution of h, within our consensus TF-TF subnetwork, illustrated in Fig. 5.10 (B) allows to classify the regulator activity of different TFs classes that are master regulators, middle managers, and workhorses TFs. As the hierarchy is constructed by maximizing the number of edges from top to bottom, out-degree hubs are more likely to be found in the upper levels, while in-degree hubs are more likely to be found in the lower levels.

These categories was not distinctly detectable in the initial TF-TF component where all nodes are interconnected to each other without an hierarchical order, as illustrated in Fig. 5.4, and with its calculated h distribution in Fig. 5.10 (A).

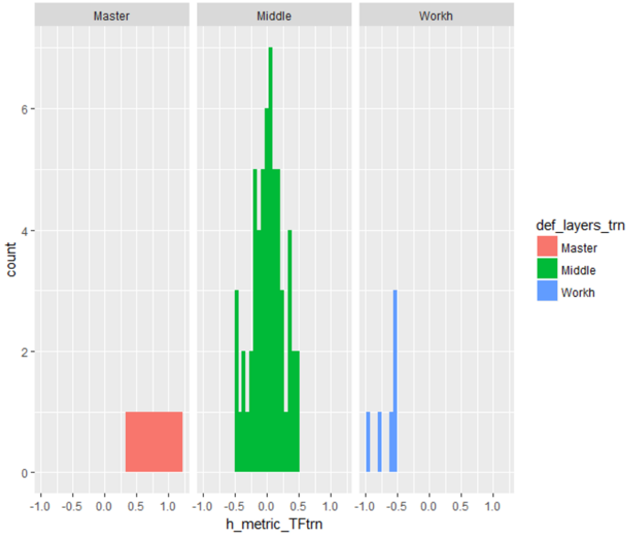


Figure 5.10 (A): *h* distribution on the TF-TF Component in the initial TRN.

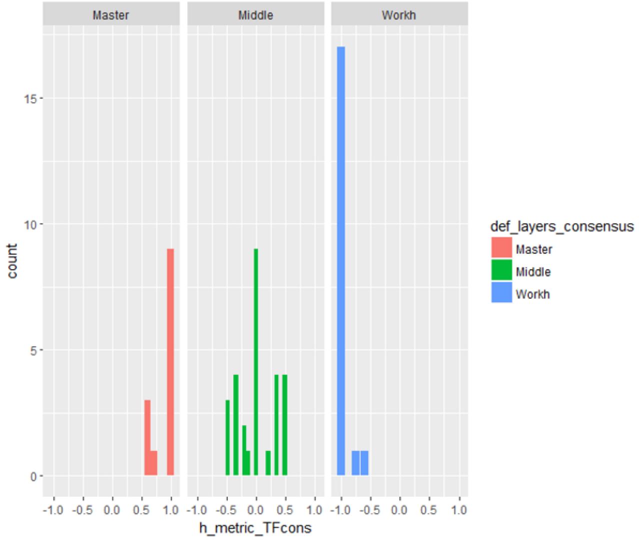


Figure 5.10 (B): *h* distribution on the TF-TF Component in the learned consensus.

The transcriptional flow detected in the consensus TF-TF Component instead allowed to organize its underlying interactions into a three-layered hierarchy, represented in the figure below.

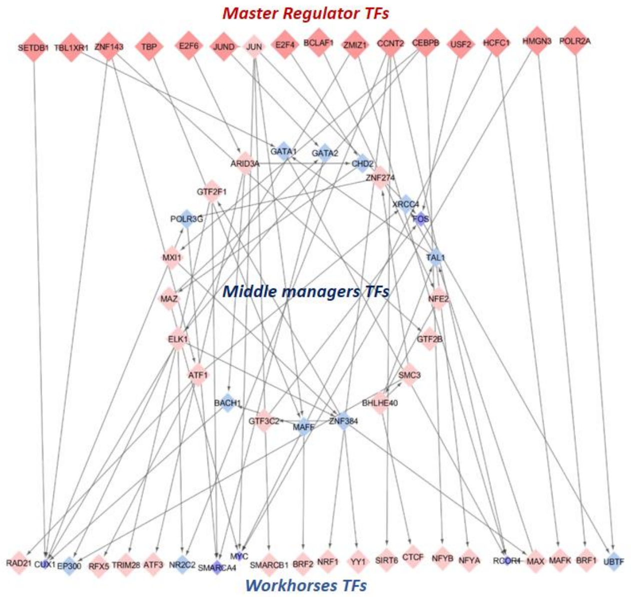


Figure 5.11: Transcriptional hierarchy for the CML. The color intensity and the size of TFs nodes are proportional to the incoming connectivity (e.g. small size combined with a darker color for a high in-degree).

This hierarchy is composed of 16 master regulator TFs, at the top, 21 brokers or middle managers, and the remaining 23 workhorses TFs, at the bottom. This topological organization can be view in a transcriptional functionality perspective, using the TFs annotations reported in Table A3 of the Appendix Sec. A.3., to correlate the impact on each TF on gene expression with its molecular function, as shown in Fig. 5.12.

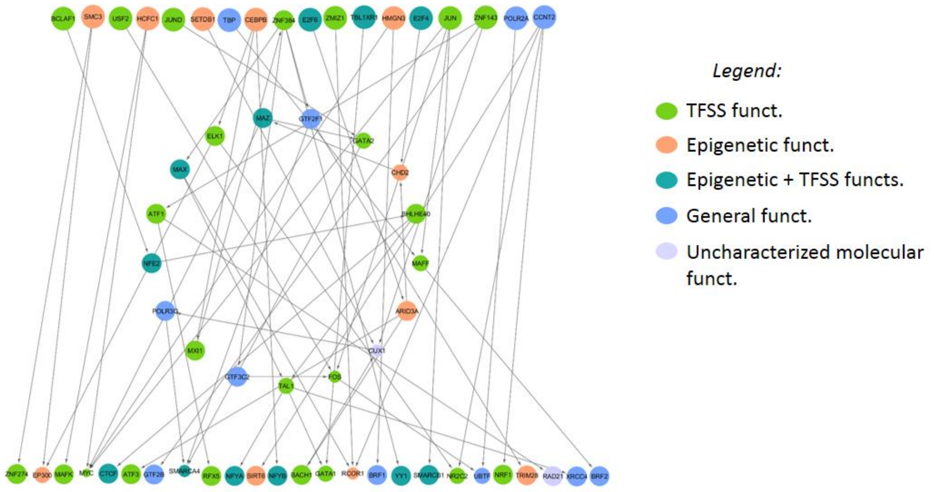


Figure 5.12: TFs functional hierarchical organization underlying the TBN.

5.4 Simulation of transcriptional perturbations

Once the transcriptional structure has been assessed, the resulting TBN can be exploited to investigate the signatures of TFs, which are rising from the topological analysis performed at the previous step.

Considering one TF at time, a knockout effect is implemented, setting its expression values to zero. From this process, we excluded the RNA polymerase (i.e. POLR3G, POLR2A), given its general function of catalyzing the DNA transcription for mRNA production.

This *in-silico* perturbation is then propagated to the each network nodes, and quantified in terms of average variation (μ_{pert} or Δ_{pert}), considering the average before and after the propagation of the perturbation. Clearly, if the variation is equal or slightly different from zero, we can conclude that the considered node is not influenced by the knocked TF or simply, it is not its target.

The reconstructed distribution of this genomic variation allowed to rank the perturbed targets (PTs), considering the standard deviation of the estimated Δ_{pert} , to identify those PTs that are more influenced after knocking a certain TF, as depicted in Fig. 5.13.

Given the considered high dimensionality of the transcriptional perturbations, we chose two thresholds to sort Δ_{pert} . The PTs whose variability is more than ±3σ or lies within the range ±3σ to ±2σ were finally retrieved. For each perturbation model, in this way, we obtained two different lists of ranked PTs.

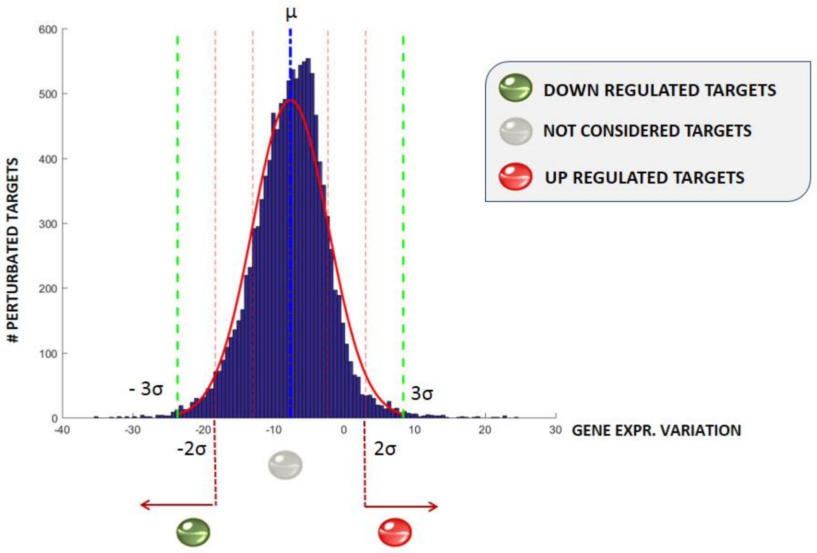


Figure 5.13: Transcriptional perturbation model distribution.

The variation detected in the average expression value of the PTs can be correlated to the function exerted by the knocked TF on them. In other words, if $\mu_{post} > \mu_{pre}$, we could infer that the considered TF negatively modulated the expression of its targets, reducing it; if instead $\mu_{post} < \mu_{pre}$, the TF originally acted as a transcriptional activator.

All knocked TF are investigated following this concept, counting the number on genomic PTs inhibited or activated after the effect propagation, as shown in Figs. 5.14-5.17.

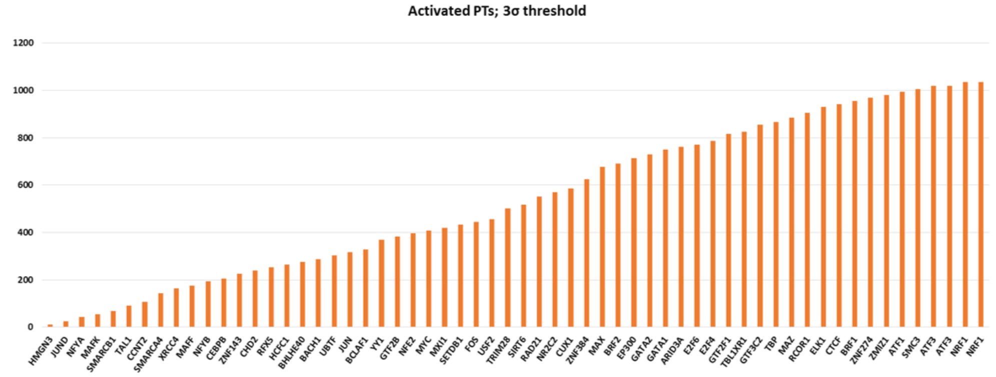


Fig. 5.14: PTs ranking for the ±3σ variation threshold – inferring transcriptional activation function.

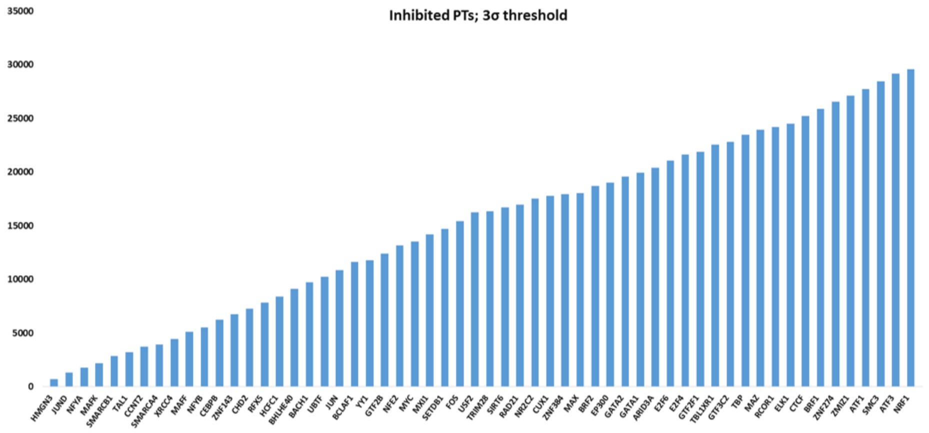


Fig. 5.15: PTs ranking for the ±3σ variation threshold – inferring transcriptional inhibition function.

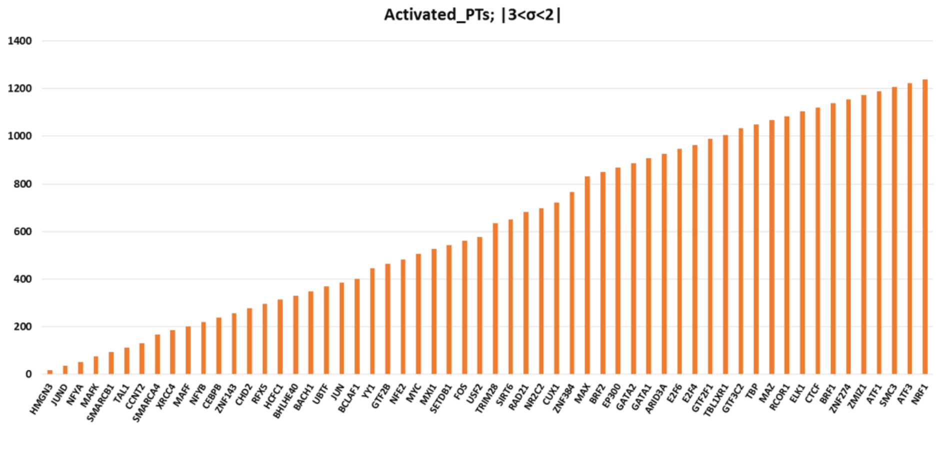


Fig. 5.16: PTs ranking for the [±3σ to ±2σ] variation threshold – inferring transcriptional activation function.

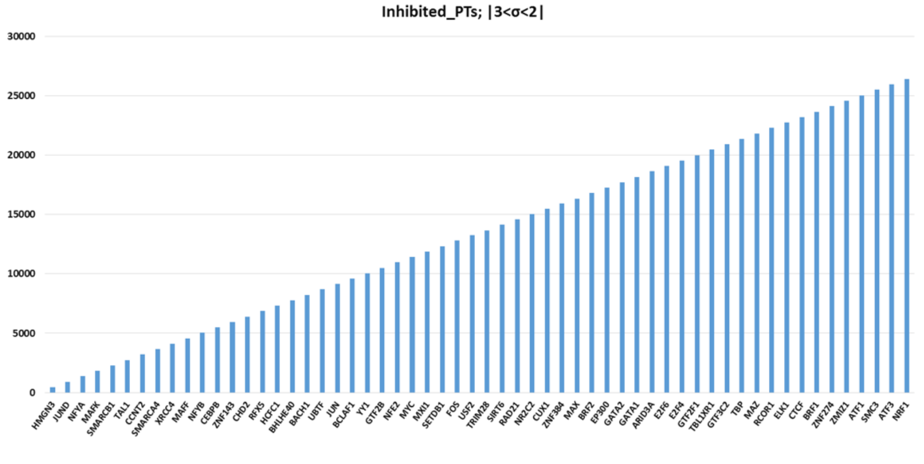


Fig. 5.17: PTs ranking for the [±3σ to ±2σ] variation threshold – inferring transcriptional inhibition function.

Moreover, to test the significance of the two PTs groups, one for each variation threshold, an enrichment analysis has been performed, using the Reactome biological pathways database.

All significant pathways (p-value <0.05, using the Benjamini Hochberg correction for multiple testing) were then aggregated for the TFs belonging to the same hierarchical layer, in order to reconstruct the common signatures.

This classification is graphically reported in Figs. 5.18-5.20 below for the ±3σ variation threshold. Each pathway is identified through a Reactome code, e.g. *R-HSA-69306*, where *R* refers to the database, *HSA*, the considered organism, "Homo sapiens", and the numerical ID is the pathway identifier. From the list of enriched pathways, we did not consider the gene expression and transcription related processes that clearly are significant for all the evaluated layers.

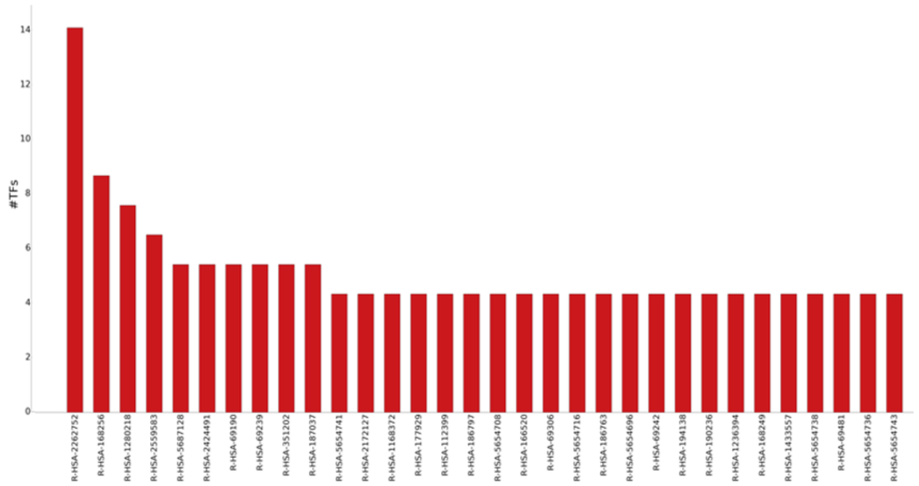


Figure 5.18: Perturbed targets of master regulator TFs pathways enrichment.

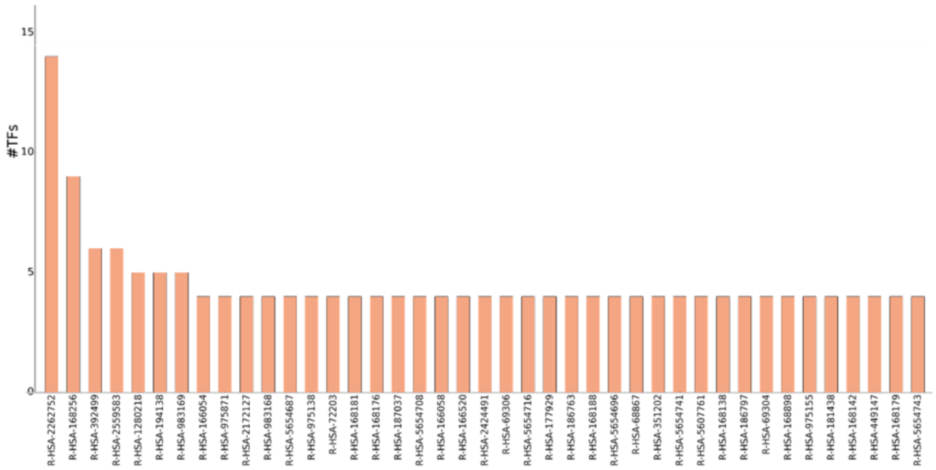


Figure 5.19: Perturbed targets of middle manager TFs pathways enrichment.

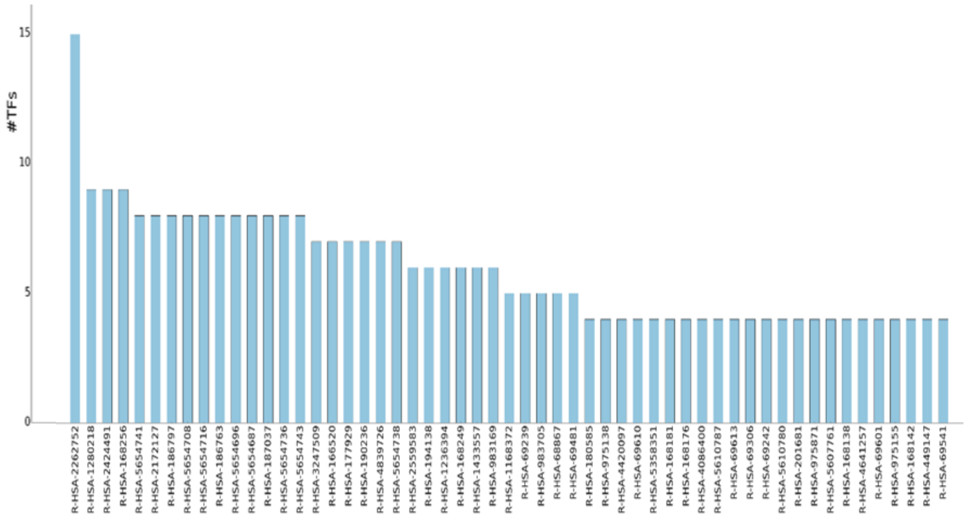


Figure 5.20: Perturbed targets of workhorses TFs pathways enrichment.

For the other threshold, given the high number of PTs and the related pathways analyzed, we reported the first ten most significant biological processes for each considered TF in Table A3, Appendix Sec. A.4.

Since the investigated disease is characterized by a molecular hallmark (see Sec. 2.4.5) involving the BCR and ABL1 genes, we explored the expression perturbations involving these two targets across all the knockout models. The μ_{pert} distribution is represented in Fig. 5.21.

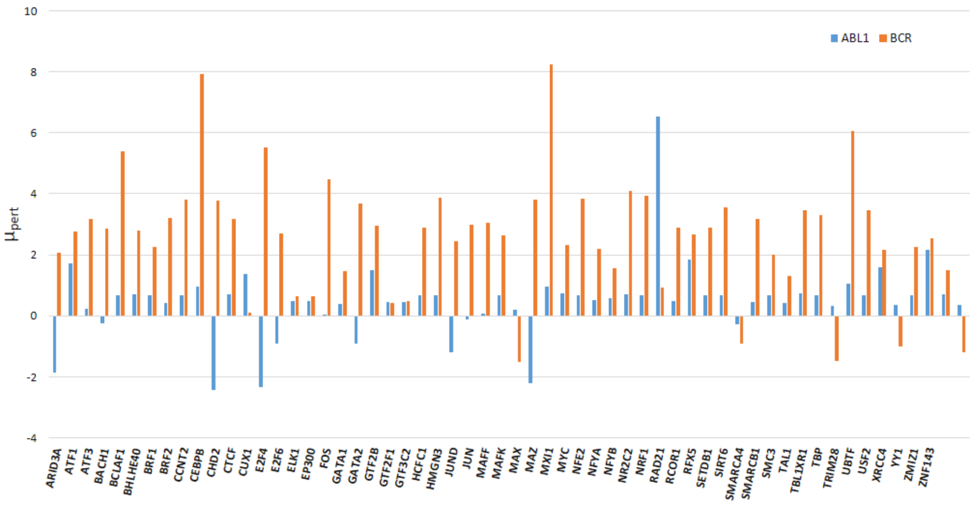


Figure 5.21: Quantification of the expression perturbation for ABL1 and BCR genes.

Chapter 6

Discussion and Conclusions

In the era of 'Omics', data integration represents a challenging tool to deliver more comprehensive insights into the biological systems under study, helping to translate novel molecular knowledge into improved diseases understanding.

In cancer, the context, in which dysregulated gene expression programs take place, has a profound impact on patients' disease mechanisms and on preventive and curative therapies responses. The investigation of such context, aimed at reconstructing the transcriptional determinants of the underlying altered expression patterns, may allow to gain insights into molecular signatures driving disease phenotypes. To this aim, the development of robust computational approaches able to deal with omics data heterogeneity and the biological complexity is necessary.

Given this challenging picture, within this three-year project, a data fusion approach has been developed, focused on multi-layered omics data integration for modeling large-scale transcriptional background. Its framework threads on three main aspects (i) the reconstruction of a transcriptional interactome using a network-centric approach, whose principles are inherited from graph theory and can be exploited to study the considered system; (ii) its mathematical modeling through a Bayesian formalism and consequently the probabilistic inspection on a genome-wide scale of the underlying transcriptional regulations with a hybrid structure learning; (iii) the investigation of the intrinsic transcriptional signatures, which characterize the resulting Bayesian model, simulating perturbations on system regulators at molecular level and propagating this effect following the transcriptional flow.

In this work, we have investigated the application of the proposed methodology to the case of CML, a subtype of blood cancers whose causative genetic event is known but its transcriptional architecture has not been deeply investigated yet on a genomic level. This becomes of particular interest, since novel hypotheses regarding altered transcriptomics and epigenomics patterns are emerging [116].

For CML, the transcriptional regulatory network (TRN) was constructed starting from raw ChIP-seq data, applying a specific bioinformatics pipeline to control and correct the biological variability inherent of this kind of NGS data, using both quality metrics and statistical constraints to prune the set of genomic transcriptional interactions of potential false positive relationships, as highlighted in Table A1. The scoring method, introduced in the analysis pipeline, exploited as a filter to detect relevant bindings, has allowed also to weight the strength of the considered interactions. This first part enabled the regulon reconstruction of each considered CML regulator (TF), modeled as a graph, whose properties has been computationally integrated leading to the TRN definition on a genome-wide scale.

The topological characterization of the TRN showed a high compactness and a complex connectivity among TF-TF interactions due to the synergic and cooperatively behavior of TFs, without leading back to meaningful biological conclusions.

In order to assess the role of each regulator, assigning it to a level within a "chain-of-command" hierarchy, the obtained regulatory system was modeled as a Bayesian network (BN), inserting it in a hybrid structure learning framework. Its scheme, able to computational scale the genomic size of the modeled network (see Chapter 4 and Sec. 5.3), exploits the reconstructed transcriptional backbone as prior knowledge, integrating it with a further complementary omics data source.

Within this joint learning framework, the network is integrated with a gene expression panel, composed by 122 transcriptomics profiles of CML patients, since our final aim is to identify transcriptional interactions which play a crucial role in the dynamic regulation of the gene expression program underlying the disease phenotype.

Following this Bayesian learning strategy, the probabilistic structure is then assessed and the resulting model allowed to get topological insights of the binding patterns, organized into a stratified hierarchy representing the overall system-level regulatory wiring, which was not observable in the initial TRN (see Sec. 5.1).

A three-tiered pyramidal structure was identified analyzing the ratio between the incoming and outcoming connectivity of all TFs presiding over all interactions of the network. This hierarchy clearly shows that the regulatory information is passed from the top to the bottom. A path within this topological organization represents a specific regulation of a downstream TF by an upstream one. Considering each path as a unique flow of transcriptional information, the number of paths through each node quantifies the amount of flow it controls. The specified TFs levels collectively regulate the non-regulator targets, lying in a lowest fourth layer that, due to its large dimension, has not been possible to graphically show in Fig. 5.1. Overall, the identified classes of TFs (master regulators, middle managers and workhorses TFs) can be interpret as the effect of their different regulatory impacts on gene expression cellular programs, since the learning phase that allowed such reconstruction is driven by the transcriptome expression. Moreover, from a functional point of view, TFs which share an epigenetic function are located at opposite ends of the hierarchy, and the central layer is instead characterized by TFs with sequence specific binding. The correlation between the topological and functional aspects for TFs, established within this hierarchy, represents an interesting novelty for the considered disease. This emerging perspective for many of the analyzed TFs could be further experimentally investigated. On the contrary, for some TFs the transcriptional role outlined in one of the aforementioned classes is supported by the scientific literature, since its importance for the hematopoietic system has already been examined, as described below.

CEBPB TF, that in this context is characterized as a master regulator (MR), within the hematopoietic system is effectively indicated as MR of steady-state granulopoiesis (i.e. process production of a sub-type of white-blood cells called granulocytes), expressed at high levels to regulate genes involved in immune and inflammatory responses. Under stress conditions, such as the cancer microenvironments, CEBPB is involved in BCR–ABL-mediated myeloid expansion and leukemic stem cell exhaustion in CML chronic-phase [117].

Members of the Jun family (JUN and JUND), that are key subunits of the transcription factor AP-1, are designated as MRs in healthy and cancer cells [118], given their crucial role in cell cycle progression, differentiation and programmed cell death. Not surprisingly, they are frequently overexpressed in leukemia, and their leukemogenesis actions are BCR-ABL1-induced [119].

Despite RAD21 and SMC3 TFs belong to the same cohesin complex involved in DNA damage repair and whose composing genes are frequently mutated in myeloid neoplasms [120], these regulators are located at the opposite network layers as a result of their different effects in their regulating modules.

GATA1 and GATA2 are two fundamental TFs which play a crucial role in gene regulation during development and differentiation of hematopoietic cells. They belong to the same layer, and their molecular recruitment is sequential; it is indeed know that GATA2 binds the promoter region of GATA1 whose expression can be repressed in the hematopoietic stem and progenitor cells [121].

To further study the underlying transcriptional signatures, a knockout effect was simulated on each considered TF, and then it was propagated following the genomic transcriptional flow of the Bayesian consensus structure. Given the high number of genes targets, to filter the perturbation effect on their expression (μ_{pert}), two variability thresholds has been taken into account: $\mu_{pert} < \pm 3\sigma$ and $\pm 3\sigma < \mu_{pert} < \pm 2\sigma$. The resulting two groups of perturbed genes (PTs) were firstly evaluated to investigate the transcriptional action of each knocked TF (activation or repression), considering the shift of the average expression before and after the perturbation propagation. Due to the broad number of the considered PTs (2,208 and 447 unique targets for the two variation cut-offs, respectively) in this research context, we point out some of the obtained results with a literature correspondence, whose underlying hypotheses could be further experimentally investigated.

In a recent study conducted by Prasad P. et al [122], the role of a SNF2 family enzymes, which comprises several chromatin remodeling genes, has been investigated in blood cells. SNF2 family enzymes are crucial for the execution of normal blood cell developmental program, and defects in chromatin remodeling, caused by mutations or aberrant expression of these proteins, may contribute to leukemogenesis.

Among them, they highlighted CHD2 TF and its interacting molecules as abundantly expressed in the blood cells related to their importance for the hematopoietic system physiology. These genes can be found as significantly perturbed for the $\pm3\sigma < \mu_{pert} < \pm2\sigma$ threshold in our work and are reported in Table 6.1 below.

Table 6.1. Some of the PTs after CHD2 TF knockout simulation.

PTs	μ_{pre}	μ_{post}
SMARCA4	2.679669	5.880250
MYC	-0.209379	9.564277
MORF4L2	-6.198733	4.487805
SMARCA5	-16.425829	6.160511
CHD4	2.974422	4.691987
HELLS	-7.770306	2.288726
BTAF1	7.724028	6.941390
BAZ1A	0.108133	5.662284
SMARCA1	-0.573362	3.621152

Moreover, within this signaling cascade, we emphasize the presence of MYC and SMARCA4 genes, since this last one is required for enhancer activation of MYC to stimulate its oncogenic transcription in leukemia [122].

Another useful comparison is represented by the knocked CTCF and MYC TFs. It is known that MYC is a target for the transcriptional repression exerted by CTCF, and Torrano V. et al [123] have experimentally demonstrated that the inhibition of CTCF expression in K562 cell line correlates with MYC overexpression, and conversely, the inhibition of MYC determines an expression increase for CTCF, as we can find in Table 6.2.

Table 6.2. CTCF and MYC expressions variation comparison.

TF KO	μ_{pre}	μ_{post}
	MYC	
CTCF	0.930602	9.309476
	CTCF	
MYC	6.172792	8.395217

As further example, we can consider the components of cohesin complex, recurrently mutated in myeloid malignancies, representing one of just nine categories of genetic alterations thought to actively contribute to leukemogenesis [120]. The major four subunits are SMC1A, SMC3, RAD21 and STAG1/2, which frequently co-locate on chromosome with CTCF transcription factor. Thus, we looked into their expression variations, having the RAD21, SMC3 and CTCF knockout simulations.

Moreover, since co-occuring mutations involve also NPM1, DNMT3A and FLT3 genes, we also considered these targets in our comparative analysis, despite FLT3 not shown a significant perturbation in RAD21 knockout. All evaluations are reported in following Tables 6.3-6.5.

Table 6.3. The considered PTs after RAD21 TF knockout simulation.

PTs	μ_{pre}	μ_{post}
STAG1	3.288470	5.473105
STAG2	7.976062	5.576123
SMC3	2.113087	7.194185
SMC1A	**0.994701**	**7.769760**
CTCF	6.172792	8.395217
FLT3	15.501185	15.016965
NPM1	**-0.978991**	**8.456430**
DNMT3A	**0.757536**	**6.175466**

Table 6.4. The considered PTs after SMC3 TF knockout simulation.

PTs	μ_{pre}	μ_{post}
STAG1	2.041964	4.594141
STAG2	12.244961	8.877212
RAD21	5.917111	8.947476
SMC1A	1.465333	6.636560
CTCF	6.676750	6.676750
FLT3	2.005014	7.027773
NPM1	**-0.698319**	**11.668327**
DNMT3A	**1.188773**	**6.538875**

Table 6.5. The considered PTs after CTCF TF knockout simulation.

PTs	μ_{pre}	μ_{post}
STAG1	1.300532	4.117925
STAG2	13.226377	9.004134
SMC3	2.113087	0.813818
SMC1A	1.462021	6.636560
RAD21	5.917111	8.947476
FLT3	2.558819	9.253860
NPM1	**0.251016**	**12.491815**
DNMT3A	**0.507685**	**6.013849**

As emerge from the bold highlighted genes, NPM1 and DNMT3 are the most influenced targets of the aforementioned simulations. In the last years, these two epigenetic modifiers have been regarded as powerful follow-up markers for another type of myeloid disorder, the acute myeloid leukemia [124], suggesting a common molecular link to impede myeloid differentiation to the benefit of disease progression [125], that can be also investigated within this pathological context.

From the experimental suppression of the epigenetic regulator EP300 in K562 cell line in the study of Giotopoulos G. et al. [126], a significant enrichment for genes involved in DNA replication, DNA repair, the control of mitosis and of the cell cycle has come out. In particular, among them, they found as not downregulated the multiple members of the minichromosome maintenance pre-replication complex (MCM3, MCM4 and MCM5), as well as its interacting proteins (MCM10) and loading/regulatory factors for replication origin licensing (CDT1 and GMNN). The related perturbations, in accord to the aforementioned results, are described in Table 6.6 for our EP300 *in-silico* knockout experiment.

Table 6.6. The considered PTs after EP300 TF knockout simulation.

PTs	μ_{pre}	μ_{post}
MCM3	-3.477105	6.909959
MCM4	-0.016194	6.909951
MCM5	3.148037	6.909641
MCM10	-2.868991	3.130641
CDT1	9.594106	10.274231
GMNN	1.841851	4.575728

In order to explore the impact of the transcriptional perturbed effects on the hallmark genes of CML, we quantified the perturbation on BCR and ABL1 targets, as shown in Fig. 5.21. Among the knocked TFs, BCR and ABL1 show a greater influence for the following regulators: ARID3A, important for B lineage commitment for human hematopoiesis [127], MAX, whose binding together with MYC is required for BCR upregulation [128], E2F4-E2F6 TFs, which play a crucial role in cell growth control [129], CEBPB, as already mentioned, essential in leukemogenesis-induced granulopoiesis, YY1 and MXI1 involved in hematopoietic stem cell differentiation [130].

The PTs groups were then assessed for a functional association with the disease phenotype, through a pathways enrichment analysis. To resume the common transcriptional signatures among the investigated biological processes, all significant pathways has been clustered by hierarchical layer to which each TF belong and ranked by significance level.

The prominent signature commonly shared among layers is the Immune system response (R-HSA-168256, R-HSA-1280218, R-HSA-5663205) combined to cellular response to stress (R-HSA-2262752). Inspecting each TFs class enrichment, we can highlight:

- For *master regulator TFs*, the most representative pathways concern cellular survival processes, as metabolism (R-HSA-1430728, R-HSA-392499), cell proliferation (R-HSA-1640170, R-HSA-69278, R-HSA-69275, R-HSA-453274), DNA replication (R-HSA-69306) and apoptosis (R-HSA-109581).
- For *middle manager TFs*, cell signaling pathways are predominant (R-HSA-194315, R-HSA-162582, R-HSA-5607761, R-HSA-4086400, R-HSA-5687128, R-HSA-1169091), cell cycle related checkpoints processes (R-HSA-453274, R-HSA-453279, R-HSA-68886) and pathways involving DNA repair mechanisms (R-HSA-73894).
- For *workhorses TFs* the signatures converge on two type of cellular signaling: external stimulus transduction through the plasma membrane and its receptors (R-HSA-187037, R-HSA-143357, R-HSA-5621487, R-HSA-1944138, R-HSA-74752, R-HSA-376176, R-HSA-190236) and internal cellular signaling exploiting the vesicle trafficking (R-HSA-1660514, R-HSA-199992, R-HSA-421837).

The obtained findings show the potential of the proposed methodology which, focusing on omics data integration, provides a data-driven platform for transcriptional regulatory network inference on a genome-wide perspective. Thanks to the probabilistic framework, it is possible to test biological hypotheses and extract meaningful information in order to better understand the considered context, that can be a pathological landscape, as investigated within this project, or a pharmacological scenario, where treatment signatures on a genomic scale may be explored for assessing the molecular effects of the administered drug in terms of genes target expression perturbations, or a personalized medicine context, combining the topological changes, which reflect altered regulatory interactions with the disease status of a certain subtype of patients (i.e. known genetic mutations that modify the expression of specific genes).

To our knowledge, the transcriptional CML background have not been investigated with an omics data fusion approach on a genomic scale. Of course, to better understand the emerging regulative signatures from the perturbation models, an experimental validation step is needed, planning targeted *in-vitro* experiments.

Anyway, under this perspective, the developed method may be considered as a reliable data driven strategy for the definition of new research hypothesis, allowing to *in-silico* test them on a large scale of potential targets, to then narrow the search field planning focused experimental procedur

Appendix

A.1 Parameters setting for the tested regulatory network reconstruction methods

ARACNe-AP

- **MI threshold calculation step**:
  p-value threshold= $1e^{-8}$
  seed= 1

  MI calculated threshold: 0.2989
- **Bootstrap step on GE input matrix:**
  100 reproducible bootstraps are obtained using the seed and p-value thresholds set at the previous step
- **Consolidation step:**
  A consensus BN is built using a p-value threshold of 0.05 for the edge significance test

----

Banjo

SearcherChoice:	SimAnneal
initialTemperature:	1000
coolingFactor:	0.7
reannealingTemperature	500
maxAcceptedNetworkBeforeCooling	1000
maxProposedNetworkBeforeCooling	10000
minAcceptedNetworkBeforeReannealing	200
ProposerChoice:	RandomLocalMove
EvaluatorChoice:	default
DeciderChoice:	default
DiscretizationPolicy:	Q5
minMarkovLag (for static data):	0
maxMarkovLag (for static data):	0
equivalentSampleSize:	1.0
maxParentCount:	6
maxTime:	5 hrs
minNetworkBeforeChecking:	1000

A.2 ChIP-seq data filtering through the pipeline

Table A1: The aim of this table is to highlight the filtering potential of the developed pipeline. Given the high number of the total analyzed replicates (more than two for each TF sample), here are reported only the identified peaks for the first replicate (Rep1). The "merged peaks" column refers to the combined peaks from each TF ChIP-replicate peak; in the last one, the final number of annotated peaks (target genes) and filtered by the binding score for each TF is reported.

TF	**Rep1 peaks**	**Filtered by p-value**	**Merged peaks**	**Targets filtered by Score**
ARID3A	26,858	14,167	45,608	10,365
ATF1	18,410	10,738	42,722	10,371
ATF3	5,206	4,103	7,953	3,515
BACH1	9,887	5,659	14,349	3,515
BCLAF1	16,236	10,339	29,093	9,981
BDP1	2,710	2,108	8,763	3,349
BHLHE40	38,943	15,294	79,902	27,058
BRF1	1,024	711	3,721	1,623
BRF2	1,848	1,760	5,957	2,912
CCNT2	26,797	9,415	48,941	9,817
CEBPB	95,501	40,258	69,412	15,008
CHD2	25,009	10,437	20,937	8,660
CTCF	62,883	13,440	26,004	9,986
CUX1	5,275	3,359	10,366	4,797
E2F4	24,848	9,739	28,300	8,392
E2F6	34,584	10,375	23,662	7,748
ELK1	7,115	5,022	11,999	6,191
EP300	5,628	3,575	7,972	3,898
FOS	22,057	9,255	23,697	7,293
GATA1	13,320	7,749	16,964	6,658
GATA2	37,026	17,886	33,206	10,043
GTF2B	6,872	4,189	7,739	3,667
GTF2F1	11,246	7,533	17,100	7,643
GTF3C2	10,681	8,120	52,264	11,947
HCFC1	26,207	10,035	21,064	8,356
HMGN3	19,355	8,680	20,713	8,558
JUN	32,362	16,935	32,874	11,111
JUND	62,351	23,234	45,930	12,570
MAFF	25,106	10,548	29,491	9,976
MAFK	23,028	10,548	27,600	10,271
MAX	74,411	25,757	42,156	12,072
MAZ	33,091	10,848	34,643	11,558
MXI1	15,994	8,398	16,309	7,788
MYC	50,678	17,636	31,347	10,404
NELFE	3,365	2,280	3,503	1,456
NFE2	8,020	4,306	8,022	3,715
NFYA	10,888	5,691	9,164	4,918
NFYB	14,002	4,151	10,421	5,518
NR2C2	2,338	1,646	5,367	2,287
NRF1	7,297	3,178	5,150	3,325
POLR2A	44,845	17,839	34,959	7,652

POLR3A	6,236	2,857	4,221	2,014
POLR3G	5,633	2,203	3,740	926
RAD21	23,660	7,706	15,224	6151
RCOR1	35,577	17,440	24,357	9,541
RFX5	4,994	3,458	7,108	3,802
SETDB1	22,897	16,025	29,854	8,505
SIRT6	6,170	4,646	7,381	3,243
SMARCA4	11,019	8,075	14,928	5,099
SMARCB1	6,838	6,343	9,309	3,985
SMC3	32,077	11,038	22,349	9,476
TAL1	31,127	9,944	23,383	9,361
TBL1XR1	11,867	7,777	27,856	9,860
TBP	33,202	12,455	18,873	8,140
TRIM28	14,716	11,389	19,428	5,860
UBTF	36,322	15,646	31,212	8,544
USF2	4,877	2,917	7,256	4,047
XRCC4	1,349	1,107	1,923	743
YY1	12,527	6,031	14,346	5,640
ZC3H11A	5,579	4,197	14,854	7,082
ZMIZ1	22,175	9,711	27,105	10,307
ZNF143	37,202	17,941	34,204	11,883
ZNF263	18,289	11,905	35,552	9,382
ZNF274	6,425	4,792	20,573	7,110
ZNF384	44,592	21,689	45,847	12,620

A.3 Functional annotation of the considered TFs

Table A2: For each TF is reported the Gene Family to which it belongs and the annotated molecular function, if it is known: Epigenetic function, and/or binding specificity (Sequence-Specific TF, TFSS), general function, if it exerts a generic action within the transcriptional machinery, as RNA polymerase.

TF	**Gene Family**	**Molecular Function**
ARID3A	AT-rich interaction domain containing	Epigenetic funct
ATF1	Basic leucine zipper proteins	TFSS
ATF3	Basic leucine zipper proteins	TFSS
BACH1	Basic leucine zipper proteins	TFSS
BCLAF1	no_info_gene_family	TFSS
BDP1	Myb/SANT domain containing	general
BHLHE40	Basic helix-loop-helix proteins	TFSS
BRF1	General transcription factors	general
BRF2	no_info_gene_family	general
CCNT2	Cyclins	general
CEBPB	Basic leucine zipper proteins	Epigenetic funct, TFSS
CHD2	DNA helicases	Epigenetic funct
CTCF	Zinc fingers C2H2-type	Epigenetic funct, TFSS
CUX1	CUT class homeoboxes and pseudogenes	-
E2F4	E2F transcription factors	Epigenetic funct, TFSS
E2F6	E2F transcription factors	Epigenetic funct, TFSS
ELK1	ETS transcription factor family	TFSS
EP300	Zinc fingers ZZ-type	Epigenetic funct
FOS	Basic leucine zipper proteins	TFSS
GATA1	GATA zinc finger domain containing	TFSS
GATA2	GATA zinc finger domain containing	TFSS
GTF2B	General transcription factors	general
GTF2F1	General transcription factors	general
GTF3C2	WD repeat domain containing	general

HCFC1	X-linked mental retardation	Epigenetic funct
HMGN3	Canonical high mobility group	Epigenetic funct
JUN	Basic leucine zipper proteins	TFSS
JUND	Basic leucine zipper proteins	TFSS
MAFF	Basic leucine zipper proteins	TFSS
MAFK	Basic leucine zipper proteins	TFSS
MAX	Basic helix-loop-helix proteins	Epigenetic funct, TFSS
MAZ	Zinc fingers C2H2-type	Epigenetic funct
MXI1	Basic helix-loop-helix proteins	TFSS
MYC	Basic helix-loop-helix proteins	TFSS
NELFE	RNA binding motif containing	general
NFE2	Basic leucine zipper proteins	TFSS
NFYA	no_info_gene_family	Epigenetic funct, TFSS
NFYB	no_info_gene_family	Epigenetic funct, TFSS
NR2C2	Nuclear hormone receptors	TFSS
NRF1	no_info_gene_family	TFSS
POLR2A	RNA polymerase subunits	general
POLR3A	RNA polymerase subunits	general
POLR3G	RNA polymerase subunits	general
RAD21	Cohesin complex	-
RCOR1	Myb/SANT domain containing	Epigenetic funct
RFX5	Regulatory factor X family	TFSS
SETDB1	Lysine methyltransferases	Epigenetic funct
SIRT6	Sirtuins	Epigenetic funct
SMARCA4	no_info_gene_family	Epigenetic funct, TFSS
SMARCB1	Protein phosphatase 1 regulatory subunits	Epigenetic funct
SMC3	Proteoglycans	Epigenetic funct
TAL1	Basic helix-loop-helix proteins	TFSS
TBL1XR1	WD repeat domain containing	Epigenetic funct, TFSS
TBP	General transcription factors	general
TRIM28	Ring finger proteins	Epigenetic funct

UBTF	no_info_gene_family	general
USF2	Basic helix-loop-helix proteins	TFSS
XRCC4	no_info_gene_family	general
YY1	Zinc fingers C2H2-type	Epigenetic funct, TFSS
ZC3H11A	Zinc fingers CCCH-type	TFSS
ZMIZ1	Zinc fingers MIZ-type	TFSS
ZNF143	Zinc fingers C2H2-type	TFSS
ZNF263	Zinc fingers C2H2-type	TFSS
ZNF274	Zinc fingers C2H2-type	TFSS
ZNF384	Zinc fingers C2H2-type	TFSS

A.4 Reactome Pathways Enrichment for each TF perturbation model

Table A3: For each TF the first ten most significant enriched pathways are reported, categorizing the TFs per hierarchical level. The enrichment has been performed for the perturbed targets (PTs) extracted applying μ_{pert} range ±3σ to ±2σ as variability threshold.

MASTER TFs

HMGN3	
Metabolism_R-HSA-1430728	9.15E-10
Cell Cycle_R-HSA-1640170	3.28E-08
Cell Cycle, Mitotic_R-HSA-69278	3.28E-08
Immune System_R-HSA-168256	3.54E-06
Processing of Capped Intron-Containing Pre-mRNA_R-HSA-72203	4.07E-06
Viral Messenger RNA Synthesis_R-HSA-168325	4.98E-06

JUND	
Metabolism_R-HSA-1430728	1.42E-08
Cell Cycle, Mitotic_R-HSA-69278	1.28E-06
Immune System_R-HSA-168256	1.48E-06
Cell Cycle_R-HSA-1640170	1.59E-06
Mitotic G2-G2/M phases_R-HSA-453274	4.51E-06
Metabolism of proteins_R-HSA-392499	4.76E-06

CCNT2	
Cell Cycle_R-HSA-1640170	3.91E-06
Cell Cycle, Mitotic_R-HSA-69278	1.29E-05
Immune System_R-HSA-168256	1.29E-05
Metabolism_R-HSA-1430728	1.29E-05
Cellular responses to stress_R-HSA-2262752	0.000156
Viral Messenger RNA Synthesis_R-HSA-168325	0.000198

SUMOylation of DNA replication proteins_R-HSA-4615885	0.000335
CEBPB	
Metabolism_R-HSA-1430728	3.03E-11
Cell Cycle_R-HSA-1640170	0.000173
Immune System_R-HSA-168256	0.000173
Metabolism of proteins_R-HSA-392499	0.000229
DNA Repair_R-HSA-73894	0.000617
Cell Cycle, Mitotic_R-HSA-69278	1.09E-03
Signaling by Rho GTPases_R-HSA-194315	1.12E-03
ZNF143	
Cell Cycle_R-HSA-1640170	4.76E-08
Cell Cycle, Mitotic_R-HSA-69278	4.76E-08
DNA Replication_R-HSA-69306	7.15E-05
G2/M Transition_R-HSA-69275	7.28E-05
Immune System_R-HSA-168256	7.61E-05
Mitotic G2-G2/M phases_R-HSA-453274	7.61E-05
Synthesis of DNA_R-HSA-69239	8.58E-05
HCFC1	
Immune System_R-HSA-168256	3.78E-08
Metabolism_R-HSA-1430728	3.87E-08
Cell Cycle_R-HSA-1640170	9.11E-07
Cell Cycle, Mitotic_R-HSA-69278	2.74E-06
Signaling by NGF_R-HSA-166520	3.16E-06
Innate Immune System_R-HSA-168249	7.32E-06
Infectious disease_R-HSA-5663205	1.12E-05
JUN	
Cell Cycle, Mitotic_R-HSA-69278	3.84E-08
Cell Cycle_R-HSA-1640170	1.70E-07
Metabolism_R-HSA-1430728	1.70E-07
Metabolism of proteins_R-HSA-392499	3.07E-07
Viral Messenger RNA Synthesis_R-HSA-168325	3.09E-06
Infectious disease_R-HSA-5663205	3.94E-06
BCLAF1	
Metabolism_R-HSA-1430728	5.13E-10
Cell Cycle_R-HSA-1640170	2.44E-07
Cell Cycle, Mitotic_R-HSA-69278	6.78E-07
Metabolism of proteins_R-HSA-392499	3.44E-06
Immune System_R-HSA-168256	7.80E-06
Cytokine Signaling in Immune system_R-HSA-1280215	2.10E-05
Infectious disease_R-HSA-5663205	3.16E-05

Cellular responses to stress_R-HSA-2262752	6.71E-05

SETDB1

Cell Cycle_R-HSA-1640170	5.18E-08
Metabolism_R-HSA-1430728	1.04E-07
Cell Cycle, Mitotic_R-HSA-69278	1.12E-07
SUMOylation of DNA replication proteins_R-HSA-4615885	1.87E-05
Immune System_R-HSA-168256	1.93E-05
Viral Messenger RNA Synthesis_R-HSA-168325	3.86E-05
SUMOylation_R-HSA-2990846	6.89E-05
SUMO E3 ligases SUMOylate target proteins_R-HSA-3108232	7.22E-05

USF2

Metabolism_R-HSA-1430728	6.32E-12
Cell Cycle, Mitotic_R-HSA-69278	7.32E-11
Cell Cycle_R-HSA-1640170	8.49E-11
Metabolism of proteins_R-HSA-392499	9.12E-09
Mitotic G2-G2/M phases_R-HSA-453274	4.00E-06
M Phase_R-HSA-68886	5.16E-06
Infectious disease_R-HSA-5663205	5.16E-06
G2/M Transition_R-HSA-69275	5.16E-06

E2F6

Cell Cycle_R-HSA-1640170	8.60E-08
Immune System_R-HSA-168256	8.60E-08
Metabolism_R-HSA-1430728	1.52E-07
Cell Cycle, Mitotic_R-HSA-69278	5.06E-07
Metabolism of proteins_R-HSA-392499	9.71E-06
Class I MHC mediated antigen processing & presentation_R-HSA-983169	1.15E-05

E2F4

Cell Cycle_R-HSA-1640170	8.38E-08
Cell Cycle, Mitotic_R-HSA-69278	6.37E-07
G2/M Checkpoints_R-HSA-69481	3.47E-06
Programmed Cell Death_R-HSA-5357801	2.05E-05
Cell Cycle Checkpoints_R-HSA-69620	2.73E-05
Metabolism_R-HSA-1430728	2.73E-05
Apoptosis_R-HSA-109581	2.73E-05

TBL1XR1

Metabolism_R-HSA-1430728	1.05E-08
Cytokine Signaling in Immune system_R-HSA-1280215	4.89E-06
Cell Cycle_R-HSA-1640170	2.22E-05
Cell Cycle, Mitotic_R-HSA-69278	2.22E-05
Infectious disease_R-HSA-5663205	4.20E-05

Immune System_R-HSA-168256	5.10E-05
EPHB-mediated forward signaling_R-HSA-3928662	5.10E-05

TBP

Cell Cycle, Mitotic_R-HSA-69278	3.01E-08
Metabolism_R-HSA-1430728	6.80E-08
Cell Cycle_R-HSA-1640170	8.12E-08
Metabolism of proteins_R-HSA-392499	1.37E-05
Infectious disease_R-HSA-5663205	8.16E-05
Mitotic G2-G2/M phases_R-HSA-453274	0.000253
G2/M Transition_R-HSA-69275	0.000414
S Phase_R-HSA-69242	0.00046

ZMIZ1

Cell Cycle, Mitotic_R-HSA-69278	1.13E-09
Cell Cycle_R-HSA-1640170	1.13E-09
Mitotic G2-G2/M phases_R-HSA-453274	3.50E-07
Metabolism_R-HSA-1430728	4.89E-07
G2/M Transition_R-HSA-69275	4.89E-07
Immune System_R-HSA-168256	5.90E-06
MAPK6/MAPK4 signaling_R-HSA-5687128	8.29E-06

MIDDLE MANAGERS TFs

MAZ

Immune System_R-HSA-168256	2.00E-06
Cellular responses to stress_R-HSA-2262752	1.64E-05
Metabolism_R-HSA-1430728	1.64E-05
Cell Cycle_R-HSA-1640170	7.24E-05
Cell Cycle, Mitotic_R-HSA-69278	0.000401
Innate Immune System_R-HSA-168249	0.000401

TAL1

Cellular responses to stress_R-HSA-2262752	1.12E-03
G1/S Transition_R-HSA-69206	1.13E-03
S Phase_R-HSA-69242	1.13E-03
C-type lectin receptors (CLRs)_R-HSA-5621481	0.00113
Cell Cycle, Mitotic_R-HSA-69278	0.001776
Activation of NF-kappaB in B cells_R-HSA-1169091	0.002636
MAPK6/MAPK4 signaling_R-HSA-5687128	0.002636
Mitotic G2-G2/M phases_R-HSA-453274	0.002636
Mitotic G1-G1/S phases_R-HSA-453279	0.002636
Cyclin E associated events during G1/S transition_R-HSA-69202	2.64E-03

MAFF

Metabolism_R-HSA-1430728	3.17E-09
Cell Cycle, Mitotic_R-HSA-69278	2.28E-07
Metabolism of proteins_R-HSA-392499	2.28E-07
Cell Cycle_R-HSA-1640170	4.51E-07
Viral Messenger RNA Synthesis_R-HSA-168325	3.65E-06
Infectious disease_R-HSA-5663205	2.89E-05
SUMOylation of DNA replication proteins_R-HSA-4615885	4.49E-05

ZNF384

Cellular responses to stress_R-HSA-2262752	2.40E-02

BHLHE40

Metabolism_R-HSA-1430728	1.22E-07
Cell Cycle_R-HSA-1640170	9.46E-07
Cell Cycle, Mitotic_R-HSA-69278	1.85E-06
Infectious disease_R-HSA-5663205	1.60E-05
Metabolism of lipids and lipoproteins_R-HSA-556833	5.37E-05

BACH1

Immune System_R-HSA-168256	7.06E-08
Cell Cycle, Mitotic_R-HSA-69278	2.94E-07
Cell Cycle_R-HSA-1640170	2.99E-07
Metabolism_R-HSA-1430728	2.99E-07
Processing of Capped Intron-Containing Pre-mRNA_R-HSA-72203	2.44E-06
Metabolism of proteins_R-HSA-392499	1.20E-05

NFE2

Metabolism_R-HSA-1430728	3.98E-11
Cell Cycle_R-HSA-1640170	2.18E-08
Cell Cycle, Mitotic_R-HSA-69278	2.56E-08
Cellular responses to stress_R-HSA-2262752	6.55E-06
Infectious disease_R-HSA-5663205	7.95E-06
Metabolism of proteins_R-HSA-392499	1.52E-05
Innate Immune System_R-HSA-168249	1.69E-05

FOS

Metabolism_R-HSA-1430728	7.87E-08
Metabolism of proteins_R-HSA-392499	1.21E-06
Cell Cycle, Mitotic_R-HSA-69278	3.38E-06
Cell Cycle_R-HSA-1640170	5.78E-06
Immune System_R-HSA-168256	0.000153
Asparagine N-linked glycosylation_R-HSA-446203	0.000201
Transport to the Golgi and subsequent modification_R-HSA-948021	2.67E-04
Innate Immune System_R-HSA-168249	6.81E-04
DNA Repair_R-HSA-73894	6.81E-04

ZNF274	
Infectious disease_R-HSA-5663205	2.44E-06
Metabolism_R-HSA-1430728	7.37E-06
Disease_R-HSA-1643685	2.44E-05
Cellular responses to stress_R-HSA-2262752	2.44E-05
Cell Cycle, Mitotic_R-HSA-69278	8.20E-05
Cell Cycle_R-HSA-1640170	0.000112

CHD2	
Immune System_R-HSA-168256	5.99E-09
Cell Cycle_R-HSA-1640170	6.71E-07
Cell Cycle, Mitotic_R-HSA-69278	8.80E-07
Cellular responses to stress_R-HSA-2262752	8.80E-07
Metabolism_R-HSA-1430728	9.43E-07
Programmed Cell Death_R-HSA-5357801	1.82E-05
S Phase_R-HSA-69242	2.17E-05
Apoptosis_R-HSA-109581	2.30E-05

MAX	
Signaling by Robo receptor_R-HSA-376176	3.05E-02
Beta-catenin independent WNT signaling_R-HSA-3858494	3.05E-02
TP53 Regulates Transcription of Cell Cycle Genes_R-HSA-6791312	3.05E-02
Cellular responses to stress_R-HSA-2262752	0.048477
DNA Replication_R-HSA-69306	0.048477
S Phase_R-HSA-69242	0.048477
PCP/CE pathway_R-HSA-4086400	0.048477

GATA2	
Cell Cycle_R-HSA-1640170	4.75E-09
Metabolism_R-HSA-1430728	3.62E-08
Cell Cycle, Mitotic_R-HSA-69278	3.72E-08
Immune System_R-HSA-168256	3.72E-08
Processing of Capped Intron-Containing Pre-mRNA_R-HSA-72203	4.01E-06
S Phase_R-HSA-69242	9.84E-06

GATA1	
S Phase_R-HSA-69242	1.49E-05
Cell Cycle, Mitotic_R-HSA-69278	3.79E-05
G1/S Transition_R-HSA-69206	3.79E-05
Cell Cycle_R-HSA-1640170	4.83E-05
Cellular responses to stress_R-HSA-2262752	5.67E-05
DNA Repair_R-HSA-73894	5.67E-05
Synthesis of DNA_R-HSA-69239	5.81E-05
DNA Replication_R-HSA-69306	5.81E-05

Immune System_R-HSA-168256	5.81E-05
ARID3A	
Immune System_R-HSA-168256	3.84E-08
Class I MHC mediated antigen processing & presentation_R-HSA-983169	4.34E-07
Cell Cycle_R-HSA-1640170	1.10E-05
Antigen processing: Ubiquitination & Proteasome degradation_R-HSA-983168	1.84E-05
Cell Cycle, Mitotic_R-HSA-69278	3.86E-05
Metabolism_R-HSA-1430728	3.86E-05
Adaptive Immune System_R-HSA-1280218	8.10E-05
Programmed Cell Death_R-HSA-5357801	1.02E-04
GTF2F1	
Cellular responses to stress_R-HSA-2262752	0.002259
Clathrin derived vesicle budding_R-HSA-421837	0.008771
trans-Golgi Network Vesicle Budding_R-HSA-199992	0.008771
C-type lectin receptors (CLRs)_R-HSA-5621481	0.021314
G1/S Transition_R-HSA-69206	0.021314
MAPK6/MAPK4 signaling_R-HSA-5687128	0.024419
Class I MHC mediated antigen processing & presentation_R-HSA-983169	0.026602
Assembly of the pre-replicative complex_R-HSA-68867	2.66E-02
S Phase_R-HSA-69242	2.66E-02
GTF3C2	
Cellular responses to stress_R-HSA-2262752	0.032419
Class I MHC mediated antigen processing & presentation_R-HSA-983169	0.032419
MAPK6/MAPK4 signaling_R-HSA-5687128	0.032825
G1/S Transition_R-HSA-69206	0.032825
Clathrin derived vesicle budding_R-HSA-421837	3.28E-02
trans-Golgi Network Vesicle Budding_R-HSA-199992	3.28E-02
Assembly of the pre-replicative complex_R-HSA-68867	4.93E-02
S Phase_R-HSA-69242	4.93E-02
ER-Phagosome pathway_R-HSA-1236974	4.93E-02
C-type lectin receptors (CLRs)_R-HSA-5621481	4.93E-02
ELK1	
Cellular responses to stress_R-HSA-2262752	9.17E-06
Dectin-1 mediated noncanonical NF-kB signaling_R-HSA-5607761	0.012584
S Phase_R-HSA-69242	0.012584
Signal Transduction_R-HSA-162582	0.013404
G1/S Transition_R-HSA-69206	0.013404
Class I MHC mediated antigen processing & presentation_R-HSA-983169	0.013404
C-type lectin receptors (CLRs)_R-HSA-5621481	0.013404
NIK-->noncanonical NF-kB signaling_R-HSA-5676590	0.013404

MXI1

Metabolism_R-HSA-1430728	6.00E-11
Immune System_R-HSA-168256	3.39E-05
Metabolism of proteins_R-HSA-392499	9.44E-05
DNA Repair_R-HSA-73894	3.11E-04
Signaling by Rho GTPases_R-HSA-194315	3.48E-04
Innate Immune System_R-HSA-168249	4.56E-04
Transcriptional Regulation by TP53_R-HSA-3700989	9.35E-04

ATF1

Cell Cycle, Mitotic_R-HSA-69278	1.75E-08
Cell Cycle_R-HSA-1640170	1.75E-08
Metabolism_R-HSA-1430728	2.24E-07
G2/M Transition_R-HSA-69275	8.01E-06
Mitotic G2-G2/M phases_R-HSA-453274	9.69E-06
Immune System_R-HSA-168256	5.55E-04
M Phase_R-HSA-68886	5.71E-04
Programmed Cell Death_R-HSA-5357801	5.71E-04

SMC3

Metabolism_R-HSA-1430728	5.04E-08
Infectious disease_R-HSA-5663205	8.35E-07
Cell Cycle_R-HSA-1640170	1.35E-06
Cell Cycle, Mitotic_R-HSA-69278	1.58E-06
M Phase_R-HSA-68886	3.57E-05

WORKHORSES TFs

RCOR1

Cellular responses to stress_R-HSA-2262752	2.63E-06
Chromatin modifying enzymes_R-HSA-3247509	0.000485
Chromatin organization_R-HSA-4839726	0.000485
Immune System_R-HSA-168256	0.002509
Signalling by NGF_R-HSA-166520	0.00275
NGF signalling via TRKA from the plasma membrane_R-HSA-187037	0.003395
Innate Immune System_R-HSA-168249	0.009326
Signaling by SCF-KIT_R-HSA-1433557	0.017065
Signaling by VEGF_R-HSA-194138	0.017065

MAFK

Immune System_R-HSA-168256	2.81E-07
Metabolism_R-HSA-1430728	1.01E-05
Signaling by VEGF_R-HSA-194138	1.06E-05
Class I MHC mediated antigen processing & presentation_R-HSA-983169	1.14E-05

Infectious disease_R-HSA-5663205	2.61E-05
MAPK family signaling cascades_R-HSA-5683057	3.03E-05
Signaling by Insulin receptor_R-HSA-74752	3.03E-05

YY1

Cellular responses to stress_R-HSA-2262752	0.033672
Immune System_R-HSA-168256	0.033672
Clathrin derived vesicle budding_R-HSA-421837	0.033672
trans-Golgi Network Vesicle Budding_R-HSA-199992	0.033672
Class I MHC mediated antigen processing & presentation_R-HSA-983169	0.038476
Synthesis of PIPs at the Golgi membrane_R-HSA-1660514	4.60E-02

SMARCA4

Cellular responses to stress_R-HSA-2262752	0.004824

XRCC4

Cell Cycle, Mitotic_R-HSA-69278	3.04E-07
Cell Cycle_R-HSA-1640170	3.76E-07
Immune System_R-HSA-168256	2.74E-05
Viral Messenger RNA Synthesis_R-HSA-168325	2.74E-05
G2/M Transition_R-HSA-69275	2.74E-05
Metabolism_R-HSA-1430728	3.00E-05
Mitotic G2-G2/M phases_R-HSA-453274	3.04E-05

TRIM28

Cellular responses to stress_R-HSA-2262752	0.021067
Signaling by Robo receptor_R-HSA-376176	0.021067

NFYB

Signaling by Wnt_R-HSA-195721	9.53E-05
Signalling by NGF_R-HSA-166520	1.24E-04
Cell Cycle_R-HSA-1640170	1.24E-04
Immune System_R-HSA-168256	0.000124
Signaling by FGFR_R-HSA-190236	0.000213
Cell Cycle, Mitotic_R-HSA-69278	0.000221
Signaling by FGFR2_R-HSA-5654738	0.000229
C-type lectin receptors (CLRs)_R-HSA-5621481	0.000229

UBTF

Processing of Capped Intron-Containing Pre-mRNA_R-HSA-72203	2.49E-06
Tat-mediated HIV elongation arrest and recovery_R-HSA-167243	1.97E-05
HIV elongation arrest and recovery_R-HSA-167287	1.97E-05
Elongation arrest and recovery_R-HSA-112387	1.97E-05
mRNA Splicing - Major Pathway_R-HSA-72163	1.97E-05
Viral Messenger RNA Synthesis_R-HSA-168325	2.33E-05

GTF2B

Cell Cycle_R-HSA-1640170	1.06E-08
Cell Cycle, Mitotic_R-HSA-69278	1.06E-08
Metabolism_R-HSA-1430728	9.67E-06
G2/M Transition_R-HSA-69275	6.49E-05
Mitotic G2-G2/M phases_R-HSA-453274	7.83E-05
M Phase_R-HSA-68886	8.00E-05
Immune System_R-HSA-168256	2.57E-04
Regulation of PLK1 Activity at G2/M Transition_R-HSA-2565942	7.77E-04

MYC

Signalling by NGF_R-HSA-166520	3.40E-08
Immune System_R-HSA-168256	4.23E-08
Signaling by VEGF_R-HSA-194138	4.69E-08
MAPK family signaling cascades_R-HSA-5683057	1.24E-07
Interleukin-2 signaling_R-HSA-451927	1.54E-07
Signalling to ERKs_R-HSA-187687	1.54E-07
Signalling to RAS_R-HSA-167044	1.70E-07
VEGFA-VEGFR2 Pathway_R-HSA-4420097	2.20E-07

SMARCB1

Cell Cycle_R-HSA-1640170	1.51E-08
Cell Cycle, Mitotic_R-HSA-69278	1.51E-08
Metabolism_R-HSA-1430728	1.51E-08
Metabolism of proteins_R-HSA-392499	1.59E-05
Infectious disease_R-HSA-5663205	2.35E-05
Regulation of PLK1 Activity at G2/M Transition_R-HSA-2565942	3.70E-05

SIRT6

Cellular responses to stress_R-HSA-2262752	1.58E-06
Immune System_R-HSA-168256	7.07E-05
Transcriptional Regulation by TP53_R-HSA-3700989	0.000509
Innate Immune System_R-HSA-168249	0.004515
Activation of anterior HOX genes in hindbrain development during early embryogenesis_R-HSA-5617472	4.52E-03
Activation of HOX genes during differentiation_R-HSA-5619507	4.52E-03

NR2C2

Metabolism_R-HSA-1430728	4.39E-06
Immune System_R-HSA-168256	3.22E-05
Cell Cycle_R-HSA-1640170	7.86E-05
Cell Cycle, Mitotic_R-HSA-69278	0.000474
Transcriptional Regulation by TP53_R-HSA-3700989	0.000559
Innate Immune System_R-HSA-168249	1.39E-03

CUX1

Cell Cycle, Mitotic_R-HSA-69278	1.05E-05
Cell Cycle_R-HSA-1640170	1.05E-05
Immune System_R-HSA-168256	1.05E-05
Metabolism_R-HSA-1430728	1.05E-05
Cellular responses to stress_R-HSA-2262752	6.55E-05
MAPK6/MAPK4 signaling_R-HSA-5687128	7.43E-05
Viral Messenger RNA Synthesis_R-HSA-168325	9.54E-05
DNA Replication_R-HSA-69306	1.30E-04
Metabolism of polyamines_R-HSA-351202	2.43E-04

RAD21

Immune System_R-HSA-168256	6.30E-05
Growth hormone receptor signaling_R-HSA-982772	0.003541
Innate Immune System_R-HSA-168249	0.01155
Metabolism of lipids and lipoproteins_R-HSA-556833	1.86E-02
Metabolism_R-HSA-1430728	1.88E-02
Cytokine Signaling in Immune system_R-HSA-1280215	2.15E-02
Adaptive Immune System_R-HSA-1280218	2.15E-02
CD209 (DC-SIGN) signaling_R-HSA-5621575	2.61E-02

BRF2

Cell Cycle, Mitotic_R-HSA-69278	4.09E-09
Cell Cycle_R-HSA-1640170	6.95E-09
Metabolism_R-HSA-1430728	7.76E-09
Metabolism of proteins_R-HSA-392499	2.07E-07
Infectious disease_R-HSA-5663205	7.20E-07
Viral Messenger RNA Synthesis_R-HSA-168325	7.20E-07

EP300

Cellular responses to stress_R-HSA-2262752	9.17E-06
Dectin-1 mediated noncanonical NF-kB signaling_R-HSA-5607761	1.26E-02
S Phase_R-HSA-69242	0.012584
Signal Transduction_R-HSA-162582	0.013404
G1/S Transition_R-HSA-69206	0.013404
Class I MHC mediated antigen processing & presentation_R-HSA-983169	1.34E-02
C-type lectin receptors (CLRs)_R-HSA-5621481	1.34E-02
NIK-->noncanonical NF-kB signaling_R-HSA-5676590	1.34E-02

RFX5

Cell Cycle_R-HSA-1640170	8.25E-08
Cell Cycle, Mitotic_R-HSA-69278	8.25E-08
Metabolism_R-HSA-1430728	1.90E-05
Immune System_R-HSA-168256	7.44E-05
G2/M Transition_R-HSA-69275	0.00015

Mitotic G2-G2/M phases_R-HSA-453274	1.80E-04
M Phase_R-HSA-68886	1.07E-03
NFYA	
Metabolism_R-HSA-1430728	1.37E-06
Cell Cycle, Mitotic_R-HSA-69278	1.03E-05
DNA Replication_R-HSA-69306	1.75E-05
G1/S Transition_R-HSA-69206	1.75E-05
S Phase_R-HSA-69242	3.43E-05
Cyclin A:Cdk2-associated events at S phase entry_R-HSA-69656	3.43E-05
ATF3	
Cell Cycle_R-HSA-1640170	1.06E-08
Cell Cycle, Mitotic_R-HSA-69278	1.53E-08
Metabolism_R-HSA-1430728	1.53E-08
Processing of Capped Intron-Containing Pre-mRNA_R-HSA-72203	8.82E-07
Metabolism of proteins_R-HSA-392499	1.73E-06
mRNA Splicing - Major Pathway_R-HSA-72163	1.05E-05
Viral Messenger RNA Synthesis_R-HSA-168325	1.09E-05
BRF1	
Metabolism_R-HSA-1430728	4.88E-10
Cell Cycle, Mitotic_R-HSA-69278	2.45E-09
Cell Cycle_R-HSA-1640170	2.45E-09
Infectious disease_R-HSA-5663205	2.31E-06
Viral Messenger RNA Synthesis_R-HSA-168325	2.31E-06
Processing of Capped Intron-Containing Pre-mRNA_R-HSA-72203	4.77E-06
Metabolism of proteins_R-HSA-392499	4.92E-06
CTCF	
Metabolism_R-HSA-1430728	5.18E-06
Immune System_R-HSA-168256	1.00E-05
Cellular responses to stress_R-HSA-2262752	3.21E-05
Signalling by NGF_R-HSA-166520	0.000614
Metabolism of lipids and lipoproteins_R-HSA-556833	0.000614
Membrane Trafficking_R-HSA-199991	0.000959
NCAM signaling for neurite out-growth_R-HSA-375165	9.79E-04
Innate Immune System_R-HSA-168249	9.79E-04
NRF1	
Cellular responses to stress_R-HSA-2262752	9.53E-05
Immune System_R-HSA-168256	1.85E-04
Transcriptional Regulation by TP53_R-HSA-3700989	0.000221
Cell Cycle_R-HSA-1640170	0.001495
HDR through Homologous Recombination (HR) or Single Strand Annealing (SSA)_R-HSA-5693567	0.001495

Generic Transcription Pathway_R-HSA-212436	0.001624
Innate Immune System_R-HSA-168249	0.002729
Homology Directed Repair_R-HSA-5693538	0.002729

Bibliography

[1] Chen R, Snyder M. Promise of Personalized Omics to Precision Medicine. Wiley Interdiscip Rev Syst Biol Med 2013;5:73–82. doi:10.1002/wsbm.1198.

[2] Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 2015;8:33. doi:10.1186/s12920-015-0108-y.

[3] Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 2016. doi:10.1093/bib/bbw114.

[4] Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, et al. Data Standards for Omics Data: The Basis of Data Sharing and Reuse. Methods Mol Biol Clifton NJ 2011;719:31–69. doi:10.1007/978-1-61779-027-0_2.

[5] 1000 Genomes Project Promises Closer Look at Variation in Human Genome | Genetics and Genomics | JAMA | The JAMA Network n.d. https://jamanetwork.com/journals/jama/article-abstract/183079?redirect=true (accessed December 7, 2017).

[6] The Cancer Genome Atlas Home Page. Cancer Genome Atlas - Natl Cancer Inst n.d. https://cancergenome.nih.gov/ (accessed December 11, 2017).

[7] An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 2012;489:57–74. doi:10.1038/nature11247.

[8] Brandi Davis-Dusenbery. Precision Medicine Research in the Million-Genome Era. G E N n.d. https://www.genengnews.com/gen-articles/precision-medicine-research-in-the-million-genome-era/5944 (accessed December 4, 2017).

[9] Costa FF. Big data in biomedicine. Drug Discov Today 2014;19:433–40. doi:10.1016/j.drudis.2013.10.012.

[10] Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights 2016;8:1–10. doi:10.4137/BII.S31559.

[11] Li Y, Chen L. Big Biological Data: Challenges and Opportunities. Genomics Proteomics Bioinformatics 2014;12:187–9. doi:10.1016/j.gpb.2014.10.001.

[12] Sun YV, Hu Y-J. Chapter Three - Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. In: Friedmann T, Dunlap JC, Goodwin SF, editors. Adv. Genet.,

vol. 93, Academic Press; 2016, p. 147–90. doi:10.1016/bs.adgen.2015.11.004.

[13] OMICS and brain tumour biomarkers: British Journal of Neurosurgery: Vol 20, No 5 n.d. http://www.tandfonline.com/doi/abs/10.1080/02688690600999620?journalCode=ibjn20 (accessed December 14, 2017).

[14] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.

[15] Herr TM, Bielinski SJ, Bottinger E, Brautbar A, Brilliant M, Chute CG, et al. A conceptual model for translating omic data into clinical action. J Pathol Inform 2015;6. doi:10.4103/2153-3539.163985.

[16] Mathew JP, Taylor BS, Bader GD, Pyarajan S, Antoniotti M, Chinnaiyan AM, et al. From Bytes to Bedside: Data Integration and Computational Biology for Translational Cancer Research. PLoS Comput Biol 2007;3. doi:10.1371/journal.pcbi.0030012.

[17] Ayers D, Day PJ. Systems Medicine: The Application of Systems Biology Approaches for Modern Medical Research and Drug Development. Mol Biol Int 2015;2015. doi:10.1155/2015/698169.

[18] Lefebvre C, Rieckhof G, Califano A. Reverse-engineering human regulatory networks. Wiley Interdiscip Rev Syst Biol Med 2012;4:311–25. doi:10.1002/wsbm.1159.

[19] Likić VA, McConville MJ, Lithgow T, Bacic A. Systems Biology: The Next Frontier for Bioinformatics. Adv Bioinforma 2010;2010. doi:10.1155/2010/268925.

[20] Greene CS, Troyanskaya OG. Integrative Systems Biology for Data Driven Knowledge Discovery. Semin Nephrol 2010;30:443–54. doi:10.1016/j.semnephrol.2010.07.002.

[21] Fernández-Suárez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Res 2014;42:D1–6. doi:10.1093/nar/gkt1282.

[22] Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res 2006;34:D504-506. doi:10.1093/nar/gkj126.

[23] Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. Nucleic Acids Res 2016;44:D726–32. doi:10.1093/nar/gkv1160.

[24] Liu ET, Pott S, Huss M. Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biol 2010;8:56. doi:10.1186/1741-7007-8-56.

[25] Botcheva K, McCorkle SR, McCombie W, Dunn JJ, Anderson CW. Distinct p53 genomic binding patterns in normal and cancer-derived human cells. Cell Cycle 2011;10:4237–49. doi:10.4161/cc.10.24.18383.

[26] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform 2017;18:279–90. doi:10.1093/bib/bbw023.

[27] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data

sets—update. Nucleic Acids Res 2013;41:D991–5. doi:10.1093/nar/gks1193.

[28] Miller MB, Tang Y-W. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. Clin Microbiol Rev 2009;22:611–33. doi:10.1128/CMR.00019-09.

[29] Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 2011;9:34. doi:10.1186/1741-7007-9-34.

[30] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res 2015;43:D1113–6. doi:10.1093/nar/gku1057.

[31] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80–R80. doi:10.1186/gb-2004-5-10-r80.

[32] Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, et al. Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res 2009;37:D19-25. doi:10.1093/nar/gkn765.

[33] Bauer-Mehren A, Furlong LI, Sanz F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol Syst Biol 2009;5:290. doi:10.1038/msb.2009.47.

[34] Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet 2012;28:323–32. doi:10.1016/j.tig.2012.03.004.

[35] Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 2017;45:D331–8. doi:10.1093/nar/gkw1108.

[36] Wang J, Zuo Y, Man Y, Avital I, Stojadinovic A, Liu M, et al. Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. J Cancer 2015;6:54–65. doi:10.7150/jca.10631.

[37] Albert R. Scale-free networks in cell biology. J Cell Sci 2005;118:4947–57. doi:10.1242/jcs.02714.

[38] Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. Nature 2000;406:378. doi:10.1038/35019019.

[39] Koschützki D, Lehmann KA, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O. Centrality Indices. Netw. Anal., Springer, Berlin, Heidelberg; 2005, p. 16–61. doi:10.1007/978-3-540-31955-9_3.

[40] Potapov AP, Voss N, Sasse N, Wingender E. Topology of Mammalian Transcription Networks. Genome Inform 2005;16:270–8. doi:10.11234/gi1990.16.2_270.

[41] Koschützki D, Schreiber F. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. Gene Regul Syst Biol 2008;2:193–201.

[42] Hao D, Ren C, Li C. Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. BMC Syst Biol 2012;6:34. doi:10.1186/1752-0509-6-34.

[43] Kamburov A. Structure of biological networks 2007.

[44] Chan SY, Loscalzo J. The Emerging Paradigm of Network Medicine in the Study of Human Disease. Circ Res 2012;111:359–74. doi:10.1161/CIRCRESAHA.111.258541.

[45] Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101. doi:10.1038/nrg1272.

[46] Brettingham-Moore KH, Taberlay PC, Holloway AF. Interplay between Transcription Factors and the Epigenome: Insight from the Role of RUNX1 in Leukemia. Front Immunol 2015;6. doi:10.3389/fimmu.2015.00499.

[47] Wilkinson AC, Nakauchi H, Göttgens B. Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. Cell Syst 2017;5:319–31. doi:10.1016/j.cels.2017.07.004.

[48] Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 2004;14:283–91. doi:10.1016/j.sbi.2004.05.004.

[49] Ng SWK, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature 2016;540:433. doi:10.1038/nature20598.

[50] Shoval O, Alon U. SnapShot: Network Motifs. Cell 2010;143:326–326.e1. doi:10.1016/j.cell.2010.09.050.

[51] Sorrells TR, Johnson AD. Making Sense of Transcription Networks. Cell 2015;161:714–23. doi:10.1016/j.cell.2015.04.014.

[52] Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell 2012;150:12–27. doi:10.1016/j.cell.2012.06.013.

[53] Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell 2012;151:68–79. doi:10.1016/j.cell.2012.08.033.

[54] Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature 2010;467:430–5. doi:10.1038/nature09380.

[55] Ng H-H, Surani MA. The transcriptional and signalling networks of pluripotency. Nat Cell Biol 2011;13:490–6. doi:10.1038/ncb0511-490.

[56] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res 2016;44:D133-143. doi:10.1093/nar/gkv1156.

[57] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 2012;40:D700–5. doi:10.1093/nar/gkr1029.

[58] Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, Santos D, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res 2014;42:D161–6. doi:10.1093/nar/gkt1015.

[59] Thompson D, Regev A, Roy S. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. Annu Rev Cell Dev Biol 2015;31:399–428. doi:10.1146/annurev-cellbio-100913-012908.

[60] Liu Z-P. Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data. Curr Genomics 2015;16:3–22. doi:10.2174/1389202915666141110210634.

[61] He B, Tan K. Understanding transcriptional regulatory networks using computational models. Curr Opin Genet Dev 2016;37:101–8. doi:10.1016/j.gde.2016.02.002.

[62] Sun N, Zhao H. Reconstructing transcriptional regulatory networks through genomics data. Stat Methods Med Res 2009;18:595–617. doi:10.1177/0962280209351890.

[63] Jensen FV. Introduction to Bayesian Networks. 1st ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 1996.

[64] Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol 2007;3:78. doi:10.1038/msb4100120.

[65] Chickering DM. Learning Bayesian networks is NP-complete. Learn Data Artif Intell Stat V 1996;112:121–130.

[66] Neapolitan RE. Learning Bayesian Networks. Pearson Prentice Hall; 2004.

[67] Schwarz : Estimating the Dimension of a Model n.d. https://projecteuclid.org/euclid.aos/1176344136 (accessed December 24, 2017).

[68] Spirtes P. Introduction to causal inference. J Mach Learn Res 2010;11:1643–1662.

[69] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006;65:31–78. doi:10.1007/s10994-006-6889-7.

[70] Triantafillou S, Tsamardinos I. Score-based vs Constraint-based Causal Learning in the Presence of Confounders. CFA UAI, 2016, p. 59–67.

[71] Penfold CA, Wild DL. How to infer gene networks from expression profiles, revisited. Interface Focus 2011;1:857–70. doi:10.1098/rsfs.2011.0053.

[72] Emmert-Streib F, Glazko GV, Altay G, de Matos Simoes R. Statistical Inference and Reverse Engineering of Gene Regulatory Networks from Observational Expression Data. Front Genet 2012;3. doi:10.3389/fgene.2012.00008.

[73] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian Networks to Analyze Expression Data. J Comput Biol 2000;7:601–20. doi:10.1089/106652700750050961.

[74] Murphy K, Mian S, others. Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA; 1999.

[75] Ghanbari M, Lasserre J, Vingron M. Reconstruction of gene networks using prior knowledge. BMC Syst Biol 2015;9. doi:10.1186/s12918-015-0233-4.

[76] Isci S, Dogan H, Ozturk C, Otu HH. Bayesian network prior: network analysis of biological data using external knowledge. Bioinformatics 2014;30:860–7. doi:10.1093/bioinformatics/btt643.

[77] Le Phillip P, Bahl A, Ungar LH. Using prior knowledge to improve genetic network reconstruction from microarray data. In Silico Biol 2004;4:335–53.

[78] Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. Pac. Symp. Biocomput., vol. 7, 2002, p. 437–449.

[79] Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. J Bioinform Comput Biol 2004;2:77–98.

[80] Werhli AV, Husmeier D. Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. Stat Appl Genet Mol Biol 2007;6. doi:10.2202/1544-6115.1282.

[81] Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinforma Oxf Engl 2013;29:1060–7. doi:10.1093/bioinformatics/btt099.

[82] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci 2003;100:8348–53. doi:10.1073/pnas.0832373100.

[83] Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. A Primer on Learning in Bayesian Networks for Computational Biology. PLoS Comput Biol 2007;3. doi:10.1371/journal.pcbi.0030129.

[84] Conrath DW. Organizational decision making behavior under varying conditions of uncertainty. Manag Sci 1967;13:B–487.

[85] Shachter RD, Kenley CR. Gaussian Influence Diagrams. Manag Sci 1989;35:527–50. doi:10.1287/mnsc.35.5.527.

[86] Murphy KP. The Bayes Net Toolbox for MATLAB. Comput Sci Stat 2001;33:2001.

[87] Trela E, Glowacki S, Błasiak J. Therapy of Chronic Myeloid Leukemia: Twilight of the Imatinib Era? Int Sch Res Not 2014. doi:10.1155/2014/596483.

[88] Zhang S. The role of aberrant transcription factor in the progression of chronic myeloid leukemia. Leuk Lymphoma 2008;49:1463–9. doi:10.1080/10428190802163305.

[89] Prange KHM, Singh AA, Martens JHA. The genome-wide molecular signature of transcription factors in leukemia. Exp Hematol 2014;42:637–50. doi:10.1016/j.exphem.2014.04.012.

[90] Grossmann V, Kohlmann A, Zenger M, Schindela S, Eder C, Weissmann S, et al. A deep-sequencing study of chronic myeloid leukemia patients in blast crisis (BC-CML) detects mutations in 76.9% of cases. Leukemia 2011;25:557. doi:10.1038/leu.2010.298.

[91] Perrotti D, Jamieson C, Goldman J, Skorski T. Chronic myeloid leukemia: mechanisms of blastic transformation. J Clin Invest 2010;120:2254–64. doi:10.1172/JCI41246.

[92] Yong ASM, Melo JV. The impact of gene profiling in chronic myeloid leukaemia. Best Pract Res Clin Haematol 2009;22:181–90. doi:10.1016/j.beha.2009.04.002.

[93] Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc 2012;7:1728–40. doi:10.1038/nprot.2012.101.

[94] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. doi:10.1093/bioinformatics/btq033.

[95] Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinforma Oxf Engl 2014;30:1006–7. doi:10.1093/bioinformatics/btt730.

[96] Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data. PLOS Comput Biol 2013;9:e1003342. doi:10.1371/journal.pcbi.1003342.

[97] Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, et al. Orange: Data Mining Toolbox in Python. J Mach Learn Res 2013;14:2349–53.

[98] Hagberg A, Swart P, S Chult D. Exploring Network Structure, Dynamics, and Function Using NetworkX. Proc. 7th Python Sci. Conf., 2008.

[99] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504. doi:10.1101/gr.1239303.

[100] Adabor ES, Acquaah-Mensah GK, Oduro FT. SAGA: A hybrid search algorithm for Bayesian Network structure learning of transcriptional regulatory networks. J Biomed Inform 2015;53:27–35. doi:10.1016/j.jbi.2014.08.010.

[101] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 2006;7:S7. doi:10.1186/1471-2105-7-S1-S7.

[102] Learning Bayesian Networks with the bnlearn R Package | Scutari | Journal of Statistical Software n.d. doi:10.18637/jss.v035.i03.

[103] Banjo: Bayesian Network Inference with Java Objects n.d. https://users.cs.duke.edu/~amink/software/banjo/ (accessed October 16, 2017).

[104] Beirlant J, Dudewicz EJ, Györfi L, Van der Meulen EC. Nonparametric entropy estimation: An overview. Int J Math Stat Sci 1997;6:17–39.

[105] Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics 2016;32:2233–5. doi:10.1093/bioinformatics/btw216.

[106] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Mol Biol Cell 1998;9:3273–97.

[107] Medvedeva YA, Lennartsson A, Ehsani R, Kulakovskiy IV, Vorontsov IE, Panahandeh P, et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. Database J Biol Databases Curation 2015;2015. doi:10.1093/database/bav067.

[108] Kohlmann A, Kipps TJ, Rassenti LZ, Downing JR, Shurtleff SA, Mills KI, et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. Br J Haematol 2008;142:802–7. doi:10.1111/j.1365-2141.2008.07261.x.

[109] Personalized synthetic lethality induced by targeting RAD52 in leukemias identified by gene mutation and expression profile | Blood Journal n.d. http://www.bloodjournal.org/content/122/7/1293?sso-checked=true (accessed October 14, 2017).

[110] Affer M, Dao S, Liu C, Olshen AB, Mo Q, Viale A, et al. Gene Expression Differences between Enriched Normal and Chronic Myelogenous Leukemia Quiescent Stem/Progenitor Cells and Correlations with Biological Abnormalities. J Oncol 2011. doi:10.1155/2011/798592.

[111] Abraham SA, Hopcroft LE, Carrick E, Drotar ME, Dunn K, Williamson AJ, et al. Dual targeting of p53 and c-Myc selectively eliminates leukaemic stem cells. Nature 2016;534:341–6. doi:10.1038/nature18288.

[112] Zheng C, Li L, Haak M, Brors B, Frank O, Giehl M, et al. Gene expression profiling of CD34+ cells identifies a molecular signature of chronic myeloid leukemia blast crisis. Leukemia 2006;20:1028–34. doi:10.1038/sj.leu.2404227.

[113] Exploration, normalization, and summaries of high density oligonucleotide array probe level data | Biostatistics | Oxford Academic n.d. https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/4.2.249 (accessed October 9, 2017).

[114] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47. doi:10.1093/nar/gkv007.

[115] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature 2012;489:91–100. doi:10.1038/nature11245.

[116] Machova Polakova K, Koblihova J, Stopka T. Role of epigenetics in chronic myeloid leukemia. Curr Hematol Malig Rep 2013;8:28–36. doi:10.1007/s11899-012-0152-z.

[117] Hirai H, Yokota A, Tamura A, Sato A, Maekawa T. Non-steady-state hematopoiesis regulated by the C/EBPβ transcription factor. Cancer Sci 2015;106:797–802. doi:10.1111/cas.12690.

[118] Vaňhara P, Šmarda J. Jun: the master regulator in healthy and cancer cells. J Appl Biomed 2006;4:163–170.

[119] Zhou C, Martinez E, Di Marcantonio D, Solanki-Patel N, Aghayev T, Peri S, et al. JUN is a key transcriptional regulator of the unfolded protein response in acute myeloid leukemia. Leukemia 2017;31:1196–205. doi:10.1038/leu.2016.329.

[120] Leeke B, Marsman J, O'Sullivan JM, Horsfield JA. Cohesin mutations in myeloid malignancies: underlying mechanisms. Exp Hematol Oncol 2014;3:13. doi:10.1186/2162-3619-3-13.

[121] Guo Y, Fu X, Huo B, Wang Y, Sun J, Meng L, et al. GATA2 regulates GATA1 expression through LSD1-mediated histone modification. Am J Transl Res 2016;8:2265–74.

[122] The Roles of SNF2/SWI2 Nucleosome Remodeling Enzymes in Blood Cell Differentiation and Leukemia n.d. https://www.hindawi.com/journals/bmri/2015/347571/ (accessed January 6, 2018).

[123] Torrano V, Chernukhin I, Docquier F, D'Arcy V, León J, Klenova E, et al. CTCF Regulates Growth and Erythroid Differentiation of Human Myeloid Leukemia Cells. J Biol Chem 2005;280:28152–61. doi:10.1074/jbc.M501481200.

[124] Schnittger S, Haferlach C, Alpermann T, Nadarajah N, Meggendorfer M, Perglerová K, et al. DNMT3A is a Powerful Follow-up Marker in NPM1 mutated AML. Blood 2014;124:122–122.

[125] Mahfouz RZ, Enane F, Hu Z, Clemente MJ, Przychodzen BP, Sekeres MA, et al. A Specific Mechanism By Which NPM1 mutations Impede Myeloid Differentiation Also Explains The Link With DNMT3A Mutation. Blood 2013;122:1254–1254.

[126] Giotopoulos G, Chan W-I, Horton S, Ruau D, Gallipoli P, Fowler A, et al. The epigenetic regulators CBP and p300 facilitate leukemogenesis and represent therapeutic targets in acute myeloid leukemia. Oncogene 2016;35:279–89. doi:10.1038/onc.2015.92.

[127] Ratliff ML, Mishra M, Frank MB, Guthridge JM, Webb CF. The Transcription Factor ARID3a is Important for In Vitro Differentiation of Human Hematopoietic Progenitors. J Immunol Baltim Md 1950 2016;196:614–23. doi:10.4049/jimmunol.1500355.

[128] Sharma N, Magistroni V, Piazza R, Citterio S, Mezzatesta C, Khandelwal P, et al. BCR/ABL1 and BCR are under the transcriptional control of the MYC oncogene. Mol Cancer 2015;14. doi:10.1186/s12943-015-0407-0.

[129] Adams MR, Sears R, Nuckolls F, Leone G, Nevins JR. Complex Transcriptional Regulatory Mechanisms Control Expression of the E2F3 Locus. Mol Cell Biol 2000;20:3633–9. doi:10.1128/MCB.20.10.3633-3639.2000.

[130] Green MR, Monti S, Dalla-Favera R, Pasqualucci L, Walsh NC, Schmidt-Supprian M, et al. Signatures of murine B-cell development implicate Yy1 as a regulator of the germinal center-specific program. Proc Natl Acad Sci U S A 2011;108:2873–8. doi:10.1073/pnas.1019537108.