UNIVERSITÉ DU
LUXEMBOURG

PhD-FSTC-2018-16
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 09/02/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Sascha JUNG GEB. ZICKENROTT
Born on 12 May 1988 in Göttingen, (Germany)

COMPUTATIONAL INTEGRATIVE MODELS FOR CELLULAR CONVERSION: APPLICATION TO CELLULAR REPROGRAMMING AND DISEASE MODELING

Dissertation defence committee

Dr Antonio del Sol, dissertation supervisor
Professor, Université du Luxembourg

Dr Lasse Sinkkonen
Research Scientist, Université du Luxembourg

Dr Jorge Goncalves, Chairman
Professor, Université du Luxembourg

Dr Miguel Andrade
Professor, Johannes Gutenberg University Mainz
Institute of Molecular Biology GmbH

Dr Noel J. Buckley, Vice Chairman
Professor, University of Oxford

Affidavit

I hereby confirm that the PhD thesis entitled "Computational Integrative Models For Cellular Conversion: Application To Cellular Reprogramming and Disease Modeling" has been written independently and without any other sources than cited.

Sascha Jung
Luxemburg
January 9, 2018

Acknowledgements

During the course of my PhD, I met a number of people who supported my personal and academic developments and would like to use this opportunity for expressing my gratitude to them.

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Antonio del Sol, for the continuous support of my research and the encouragement to aim for bigger achievements and ideas. He created a truly amazing environment of critical scientific discussions that greatly helped in focusing my research and provided invaluable feedback to my work. His door was always open to discuss new ideas, so that my productivity was only limited by my own working capacity and ability to develop in new directions.

I would also like to thank the members my evaluation committee, Prof. Dr. Jorge Goncalves and Prof. Dr. Noel J. Buckley. Their critical assessment of my research and suggestions of new research directions were invaluable in defining the topics addressed during my study and, ultimately, my PhD thesis.

During the course of my PhD studies, I collaborated with several people, which greatly contributed to the research projects reflected in this thesis. Therefore, I would like to thank Prof. Dr. Miguel Andrade, Dr. Lasse Sinkkonen, Bimal Babu Upadhyaya and Julia Becker.

I would also like to thank the LCSB for hosting me and providing an amazing environment that stimulates collaborative, interdisciplinary research. It has been a privilege to work with the people here and I will never forget this experience. Especially my colleagues in the Computational Biology group provided invaluable personal and academic support whenever I needed it. Dr. Srikanth Ravichandran, Dr. Andras Hartmann, Dr. Vladimir Espinosa Angarica, Gaia Zaffaroni and Muhammad Ali provided great feedback when I was stuck with a problem and stimulated the investigation of new possibility, whenever I needed it.

Finally, I would like to thank my parents for their continuous support. I would also particularly like to thank my wife, Patrizia, for her unconditional love and support in each and every situation and my kids, Anna and Felix, for showing me that distraction sometimes helps to get a clear view when being stuck with a problem.

Figures & Tables

Figures

Tables

Summary

The groundbreaking identification of only four transcription factors that are able to induce pluripotency in any somatic cell upon perturbation stimulated the discovery of copious amounts of instructive factors triggering different cellular conversions. Such conversions are highly significant to regenerative medicine with its ultimate goal of replacing or regenerating damaged and lost cells. Precise directed conversion of damaged cells into healthy cells offers the tantalizing prospect of promoting regeneration *in situ*.

In the advent of high-throughput sequencing technologies, the distinct transcriptional and accessible chromatin landscapes of several cell types have been characterized. This characterization provided clear evidences for the existence of cell type specific gene regulatory networks determined by their distinct epigenetic landscapes that control cellular phenotypes. Further, these networks are known to dynamically change during the ectopic expression of genes initiating cellular conversions and stabilize again to represent the desired phenotype.

Over the years, several computational approaches have been developed to leverage the large amounts of high-throughput datasets for a systematic prediction of instructive factors that can potentially induce desired cellular conversions. To date, the most promising approaches rely on the reconstruction of gene regulatory networks for a panel of well-studied cell types relying predominantly on transcriptional data alone. Though useful, these methods are not designed for newly identified cell types as their frameworks are restricted only to the panel of cell types originally incorporated. More importantly, these approaches rely majorly on gene expression data and cannot account for the cell type specific regulations modulated by the interplay of the transcriptional and epigenetic landscape.

In this thesis, a computational method for reconstructing cell type specific gene regulatory networks is proposed that aims at addressing the aforementioned limitations of current approaches. This method integrates transcriptomics, chromatin accessibility assays and available prior knowledge about gene regulatory interactions for predicting instructive factors that can potentially induce desired cellular conversions. Its application to the prioritization of drugs for reverting pathologic phenotypes and the identification of instructive factors for inducing the cellular conversion of adipocytes into osteoblasts underlines the potential to assist in the discovery of novel therapeutic interventions.

CHAPTER 1 Introduction & Literature Review

Regenerative medicine, with its ultimate goal to replace or regenerate damaged and lost human cells, is striving to develop efficient protocols for (re)generating cells, tissues or organs impaired in existing pathologies (Mason and Dunnill, 2008). Historically, the term was first coined 25 years ago (Kaiser, 1992) and comprises a wide variety of techniques. There are two major lines of regenerative medicine that can be distinguished. The first approaches involved the transplantation from donor tissues and aimed at restoring the normal function of tissues and cells. In the late 1960s, the first successful cell transplantation of bone marrow was performed in living humans (Starzl, 2000) and laid the groundwork for future clinical applications. One of the earliest developments is the use of engineered tissue transplants in human. Pioneered by the engineering and transplantation of skin in 1981 (Burke *et al.*, 1981), additional tissues, like artificial bladders (Atala *et al.*, 2006) and re-engineered livers (Uygun *et al.*, 2010), were successfully transplanted in living humans.

Complementary to tissue transplantation, the second line of research in regenerative medicine can be described as cell-based therapy and follows the strategy of injecting novel and healthy cells in pathologic tissues (Sampogna *et al.*, 2015). Several approaches currently exist for obtaining healthy cells for transplantation. First, differentiated cells can be directly collected from the patient's respective tissue, expanded *in vitro* and implanted again, without alterations. However, expanding the collected cells typically bears problems due to the change in the microenvironment of the cells, i.e. the culturing conditions (Atala, 2012). A more promising approach constitutes the cellular conversion of donor cells into induced pluripotent stem cells (iPSCs) and their subsequent differentiation or, more generally, their direct trans-differentiation into the desired cell type.

With the groundbreaking finding that the cellular conversion of somatic cells to pluripotent stem cells (iPSCs) can be induced by up-regulating only four transcription factors, OCT4, SOX2, KLF4 and c-MYC (Takahashi and Yamanaka, 2006), the directed cellular conversion into particular cell types became widely applicable. Many protocols have been discovered for various cellular conversion such as the differentiation of iPSCs into, for example, cortical neurons (Shi *et al.*, 2012), astrocytes (Shaltouki *et al.*, 2013), skeletal muscle cells (Maffioletti *et al.*, 2015) or cardiomyocytes (Burridge *et al.*, 2014) and the trans-differentiation of fibroblasts into cardiomyocytes (Ieda *et al.*, 2010), hematopoetic progenitors (Szabo *et al.*, 2010) or motor neurons (Son *et al.*, 2011). However, the instructive factors of desired conversion are unknown in many cases, due to the limited understanding of the cellular gene regulatory networks modulating cellular identity.

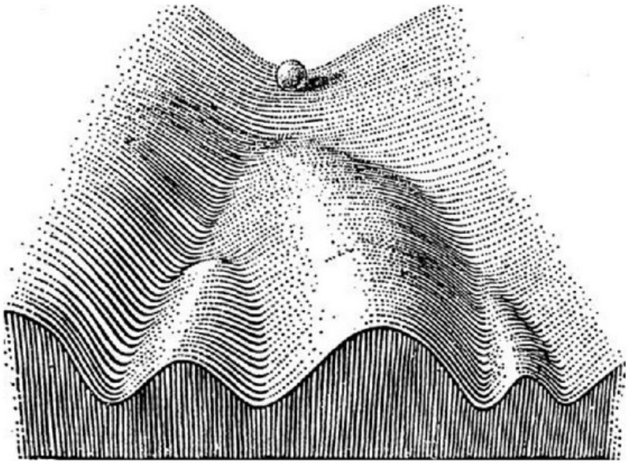
The discovery of instructive factors for cellular conversions has been previously accomplished by experimental testing of conjectures derived from literature. In the advent of high-throughput sequencing technologies, computational methods attempted the reconstruction of the complex gene regulatory networks underlying cellular phenotypes for enabling their systematic analysis (Liu, 2015). Based on these network models, strategies for identifying important transcriptional regulators or for examining the cellular responses to perturbations by model simulation have been proposed in order to facilitate the prediction of instructive factors (Crespo *et al.*, 2013; Cahan *et al.*, 2014; Rackham *et al.*, 2016). However, there exist several significant limitations that prevent these methods from being generally applicable.

The remainder of this chapter reviews the current knowledge about cellular gene regulatory networks and discusses the limitations of existing methods for identifying instructive factors of cellular conversions. The organization is as follows: Section 1.1 (Modulation of Cell Types through Gene Regulation) describes the modulation and stabilization of cellular phenotypes through the interplay of transcriptional and epigenetic regulation. Section 1.2 (Gene Regulatory Network Inference from Condition Specific Transcriptomics Data) provides a detailed review of modeling formalisms for representing gene regulatory networks with various levels of detail and the practical issues with their identification. Then, section 1.3 (Systematic Identification of Instructive Factors for Cellular Conversions) reviews current computational methods for the identification of instructive factors with and without the aid of computational network models. Finally, section 1.4 concludes with a summary including the main limitations of current computational methods.

1.1 Modulation of Cell Types through Gene Regulation

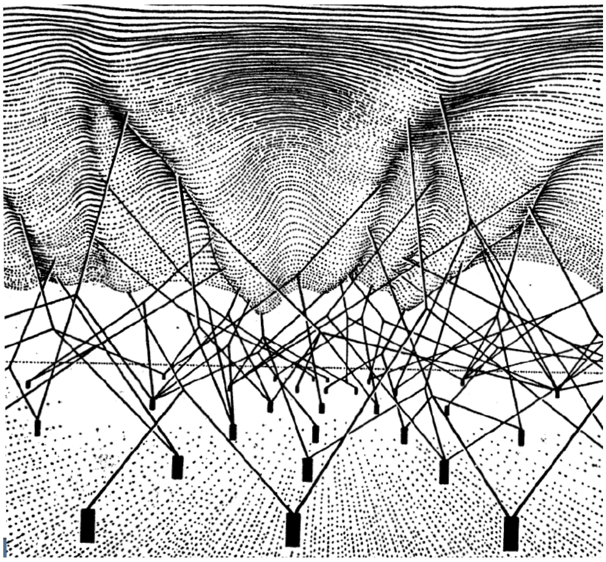
In 1957, Conrad Waddington introduced the concept of an "epigenetic landscape" representing the developmental process of cells (Figure 1.1) depicted by a ball rolling down a valley (Waddington, 1957). On its path, a number of forks, or decision points, occur that separate different valleys and represent the differentiation towards a more mature cell type. Each fork represents here the metastable state of stem and progenitor cells, ultimately giving rise to fully differentiated cells when no fork exists anymore. In most cases, forks represent a binary cell fate specification that forces cells to adopt one of two cell states (Hayward *et al.*, 2008) and are mediated by transcriptional changes. Thus, the "epigenetic landscape" is shaped by the coordinated expression of genes (Figure 1.2) requiring an interaction of transcription factors (TFs) that control the expression of genes and histone modifiers/chromatin remodelers for determining the accessibility of other transcription factors to the DNA and their binding stability (Moris *et al.*, 2016). Particularly, transcription factors

Figure 1.1. Waddington landscape



Representation of development as a ball rolling down a landscape with several, alternative paths. Distinct fates are separated by hills and inhibit a natural switch of committed cells. Similarly, not all fates are equally likely, as can be seen on the second decision of the left part. (Picture taken from (Waddington, 1957))

Figure 1.2. Shaping of the epigenetic landscape through gene regulation



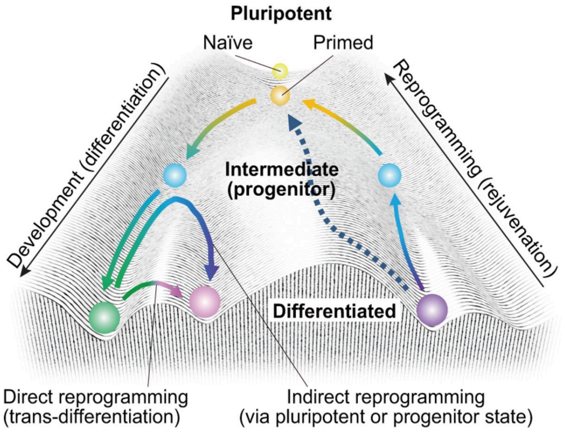
The developmental paths are modulated by the interplay of genes, depicted as bars. Combinatorial effects of gene expression shape the height of hills and wideness of valleys and as such determine cell fate decisions. (Picture taken from (Waddington, 1957))

have been shown to actively regulate and control cell fate specification during development in a variety of cases, including the retinal ganglion (Wu *et al.*, 2015), myeloid cells (Rosmarin *et al.*, 2005), cardiomyocytes (Bai *et al.*, 2015) and neurons (Zhang *et al.*, 2013).

The "epigenetic landscape" proposed by Conrad Waddington served for many years as an intuitive explanation of cellular transitions. However, the pioneering work of Davis et al. in 1987 proved the feasibility of direct cellular conversions of fibroblasts into myoblasts upon overexpression of a single transcription factor, MYOD (Davis *et al.*, 1987), which significantly changed the perception of cellular conversion. Subsequent studies confirmed the phenomenon of transcription factor induced cell fate conversions in several blood cell types (Kulessa *et al.*, 1995) and the pancreas (Shen *et al.*, 2000), which led to a major conceptual change of the epigenetic landscape. The concept was further shattered with the groundbreaking identification that somatic cells could be reprogrammed to a pluripotent state by ectopic expression of only four TFs (Takahashi and Yamanaka, 2006; Takahashi *et al.*, 2007). Instead of a unidirectional landscape of cellular developmental processes, these findings underline the multidirectionality of cellular conversions including trans-differentiation and rejuvenation (Figure 1.3, (Takahashi and Yamanaka, 2015)).

Even though Waddington's classic proposition of an "epigenetic landscape" underwent conceptual revisions, the experimental evidence for cellular conversions supports the existence of gene regulatory networks (GRNs) modulating different cell

Figure 1.3. Cellular conversions as displacements in the Waddington landscape



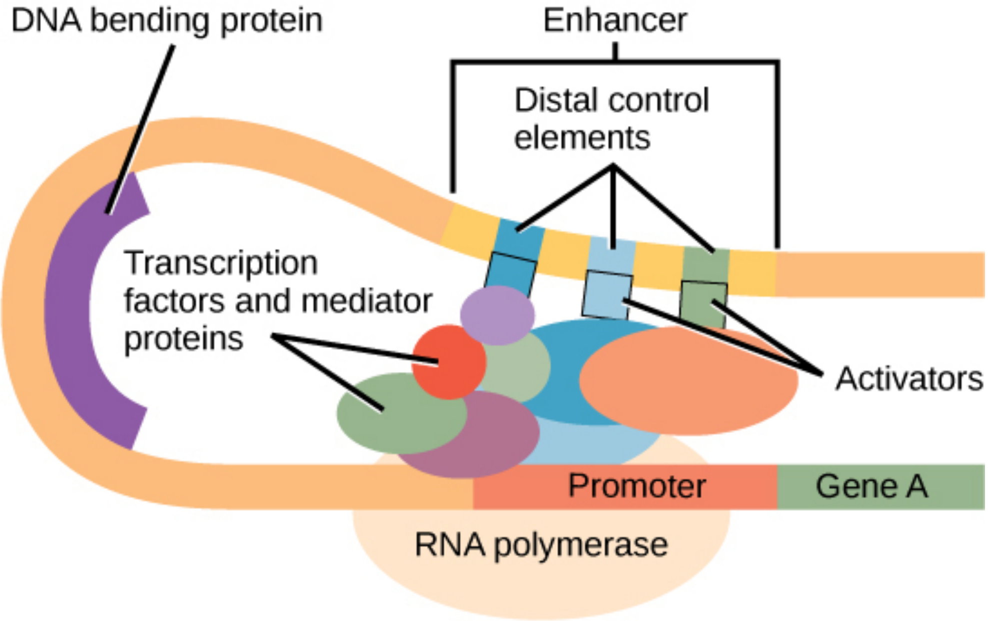
Developmental paths of embryonic to somatic cells can be altered through the perturbation of the underlying gene regulatory network. Three types of cellular conversions can be distinguished. Classical reprogramming involves the dedifferentiation of cells towards a pluripotent state not necessarily following the developmental path. Trans-differentiation constitutes the conversion of differentiated cells either by direct induction or indirectly by following developmental paths. (Picture taken from (Takahashi and Yamanaka, 2015))

types that are sensitive to perturbations. Understanding the composition and function of these cell type specific GRNs is therefore crucial to identify perturbations inducing desired cellular conversion.

1.1.1 Condition Specific Transcriptional Regulation

Transcription describes the process of generating an RNA copy of a particular segment of DNA and is orchestrated through the binding of transcription factors. In this process, TFs act alone or in conjunction with other proteins to promote or inhibit the initiation of transcription by recruitment of RNA polymerases. The regulatory elements transcription factors bind to can be divided into three main classes, i.e. promoters, enhancers and insulators (Figure 1.4). Promoters are proximal regulatory regions located close to the transcription start site (TSS) and are sufficient for attracting the transcriptional machinery. However, transcription is often weak in the absence of more distal regulatory elements (Shlyueva *et al.*, 2014). These distal elements, called enhancers, are potentially located hundreds of thousand base pairs away from the transcription start site (TSS) but remain spatially close through DNA looping in the nucleus (Amano *et al.*, 2009). Due to the three dimensional organization of DNA, multiple enhancers might contribute concurrently to the expression of their target

Figure 1.4. Transcriptional mediation through proximal and distal regulatory regions



Gene transcription is achieved through the regulation of proximal and distal regulatory elements. DNA bending spatially brings together enhancer and promoter regions allowing for the attraction of the transcriptional machinery by transcription factors. (Picture taken from https://archive.cnx.org/contents/53013107-747b-41b0-ad43-f4e97bd69ef1@2/gene-expression-eukaryotic-transcriptional-regulation-gpc)

genes in a mostly additive manner (Shlyueva *et al.*, 2014). In contrast to the active regulation in enhancer and promoter regions, insulators are passive regulatory regions separating enhancers and promoters to block their interactions (Gaszner and Felsenfeld, 2006).

The presence of multiple regulatory elements targeting the same gene supports a combinatorial binding of transcription factors ultimately allowing cells to give rise to multiple cell types and respond to environmental stimuli. Despite their individual contributions to the phenotype of cells, transcription factors form regulatory cores of mutual regulation to maintain a stable, cell type specific gene expression profile (Neph *et al.*, 2012). In particular, regulatory cores are typically composed of only a few factors and possess a strong maintenance capacity while suppressing transcription factors implicated in other lineages (Hikichi *et al.*, 2013). In this regard, the most studied example to date is certainly the regulatory core of pluripotent stem cells, composed of OCT4, SOX2 and NANOG (Ng and Surani, 2011; Young, 2011), which maintains pluripotency and, thus, suppresses the differentiation to trophectoderm, mesendoderm or neural ectoderm (Thomson *et al.*, 2011; Niwa *et al.*, 2000).

A deeper understanding of the transcriptional modules and regulatory interactions responsible for cellular stability requires the genome-wide dissection of transcription factor binding. In essence, binding of transcription factors to DNA is facilitated through the sequence-specific recognition of cis-regulatory elements, i.e. genomic sequences, ranging in size from 4 to 30 base pairs. For example, a widely studied cis-regulatory element is the TATA-Box, a DNA-sequence located in the core promoter of genes, at which the preinitiation complex forms after binding of the general transcription factor TFIID. The binding affinities of proteins, i.e. the strength of the chemical bond established between proteins and DNA, vary greatly from site to site depending on the actual sequence and the three dimensional shape of the DNA (Stormo and Zhao, 2010; Rohs *et al.*, 2009). In addition to these features, recent studies underpinned the crucial role of epigenetic modifications in establishing competent protein-DNA interactions (Liu *et al.*, 2015).

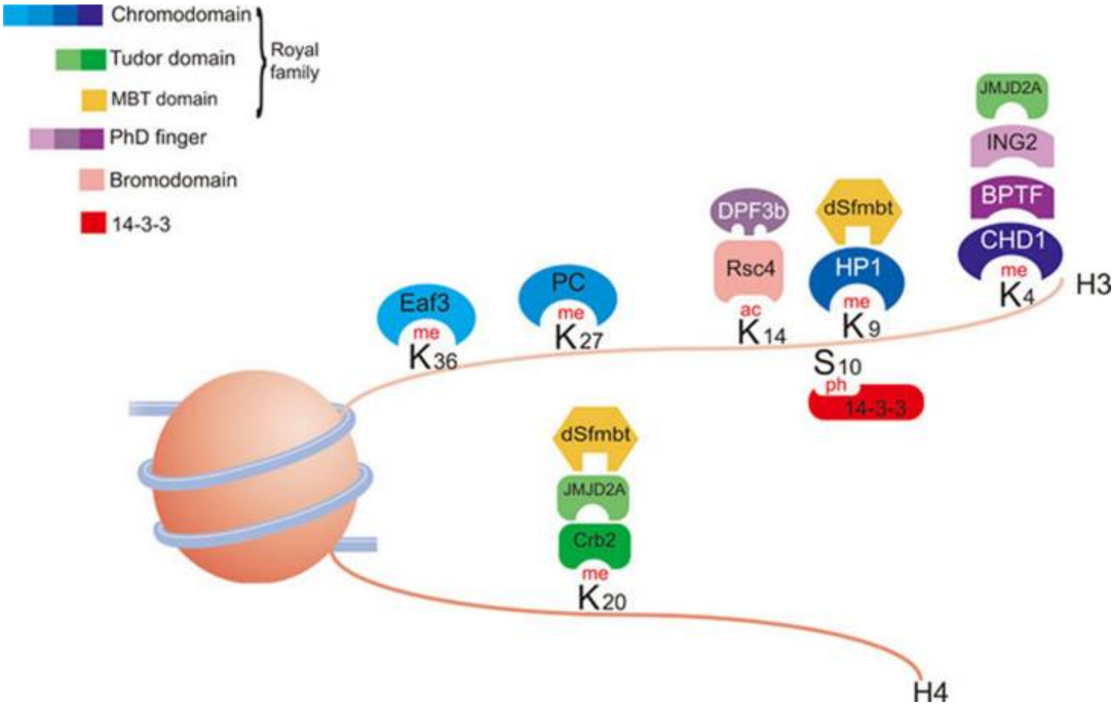
1.1.2 Condition Specific Epigenetic Landscapes Govern Transcription

The term “epigenetics” comprises various, partially heritable changes in gene function not related to changes in the DNA sequence (Dupont *et al.*, 2009). Among them, especially covalent histone modifications and chromatin accessibility are, to date, the most widely studied, though not necessarily most important, forms of epigenetic mechanisms that will be discussed in the remainder of this section.

Chromatin is typically packaged into nucleosomes consisting of histone octamers wrapped by approximately 147 base pairs of DNA. Positioning of the nucleosomes plays an important role in transcriptional regulation as it determines the availability of regulatory regions or genes to the transcriptional machinery (Radman-Livaja and Rando, 2010; A P Boyle *et al.*, 2008; Song *et al.*, 2011; Thurman *et al.*, 2012; Mercer *et al.*, 2013; Neph *et al.*, 2012). The accessibility of regulatory regions in turn allows for crosstalk between the epigenetic and transcriptional regulation in which the binding of specific TFs promotes dynamic alterations of the chromatin landscape (McVicker *et al.*, 2013; Kilpinen *et al.*, 2013; Kasowski *et al.*, 2013). Not surprisingly, variations in the chromatin landscape are an essential part of many phenotypic transitions during development or cellular conversions (Apostolou and Hochedlinger, 2013; Lara-Astiaso *et al.*, 2014; Gaspar-Maia *et al.*, 2011). Due to the importance of accessible chromatin regions, great efforts have been devoted to develop experimental and computational assays for profiling active regulatory regions in accessible chromatin. Particularly the identification of DNase hypersensitive sites by DNase-seq, formaldehyde-assisted isolation of regulatory elements by FAIRE-seq and accessible regulatory regions sensitive to transposon integration by ATAC-seq are frequently used experimental techniques (Song *et al.*, 2011; Buenrostro *et al.*, 2013; Hesselberth *et al.*, 2009), which identify accessible genomic regions. In this context, DNase-seq and ATAC-seq are of particular interest due to their pronounced ability of pinpointing genomic regions bound by proteins. However, covalent chromatin modifications play a critical role for mediating which transcription factors bind these regions and contribute to the stabilization of the gene expression profile.

The four histone proteins contained in nucleosomes can be modified by at least 80 covalent chromatin modification that define the so called “histone code” (Tsompana and Buck, 2014; Jenuwein and Allis, 2001). These post-translational modifications include the acetylation, methylation, phosphorylation, ubiquitylation and sumoylation of numerous occasions on the N-terminals of histones (Bannister and Kouzarides, 2011). While every histone mark serves its own, distinct function, two main classes of mechanisms can be distinguished. Histone acetylation and phosphorylation induce structural changes by reducing the positive charge of the histones and, as a consequence, reducing the interaction between histones and the negatively charged DNA leading to less compact chromatin structures (Bannister and Kouzarides, 2011). Examples include the acetylation of lysines four, nine and 27 on histone 3 that are enriched at active enhancer and promoter regions and facilitate transcriptional regulation (Wang *et al.*, 2008).

Figure 1.5. Interacting factors of histone modifications



Proteins interact with histone modifications through distinct domains to modify the chromatin structure or recruit additional factors. For example, ING proteins bind to methylated H3K4 and recruit either histone acetyltransferases or deacetylation complexes. On the other hand, HP1 recognizes H3K9me and is necessary for chromatin compaction. (Picture taken from (Bannister and Kouzarides, 2011))

However, in general terms, acetylated lysines on histone tails are not necessary to induce structural rearrangements and sometimes even prevent their induction (Bannister and Kouzarides, 2011). Complementary to their role in inducing structural changes, histone modifications mediate the binding of chromatin-modifying proteins (Figure 1.5). Proteins are able to interact with modified histones through different structural domains to dynamically regulate the epigenetic landscape of cells. Acetylated lysines are, for example, recognized by histone deacetyltransferases (HATs) through their bromodomain, which enables chromatin remodeling. However, different chromatin remodeling complexes or their co-factors contain the same domains and as such compete for the formation of accessible and inaccessible chromatin regions (Mujtaba *et al.*, 2007; Hassan *et al.*, 2002; Bannister and Kouzarides, 2011).

In contrast to the profiling of transcriptomics data, the experimental information about the epigenetic landscape is typically highly incomplete, which hinders the identification of causal relationships between the epigenome and the transcriptome. In fact, a previous study that jointly profiled chromatin accessibility, DNA methylation and the transcriptome in single cells and identified genes whose expression is not correlated with their methylation status (Clark *et al.*, 2018). This supports the hypothesis that multiple epigenetic states are associated to the same transcriptome. Conversely, epigenetic modifications have been related with transcriptional noise, i.e. variability in gene expression. For example, promoter regions enriched in H3K27me3, H3K4me1 or H3K9me3 are associated with increased noise while genes whose gene body is enriched in H3K36me3, H3K4me3 or H3K9ac are consistently associated to low noise (Faure *et al.*, 2017). Thus, multiple transcriptional states can arise from the same epigenetic landscape. However, these effects of the epigenome on the transcriptional state are not unidirectional. Rather the interplay of epigenetic and transcriptional regulatory elements regulates maintenance and differentiation of cells, as exemplified by P53 in mammalian stem cells (Levine and Berger, 2017). In particular, in human embryonic stem cells, deacetylating lysines at positions 120 and 373 of the P53 protein transcriptionally inactivates it. These reduced acetylation levels are maintained by the deacetylase SIRT1 whose transcription is activated by OCT4, a transcription factor responsible for maintenance of pluripotency (Ng and Surani, 2011; Young, 2011). Despite its ability to deacetylate lysines residues of P53, SIRT1 is a known histone deacetylase and leads to transcriptional repression (Zhang and Kraus, 2010). In case P53 becomes activated in embryonic stem cells, it activates transcription of two microRNAs, MIR349 and MIR145, which inhibit key transcription factors for maintaining pluripotency and induce differentiation (Levine and Berger, 2017).

Apart from the ability of some proteins to sense and interact with modified histone tails, another important class of proteins was discovered (Cirillo *et al.*, 2002). These proteins, or more precisely transcription factors, can directly bind to silent chromatin regions that are not marked by repressive chromatin modifications, such as H3K27me3 or H3K9me3 (Soufi *et al.*, 2012; van Oevelen *et al.*, 2015), and create an accessible chromatin conformation independent of other chromatin remodelers (Cirillo *et al.*, 2002). Despite the experimental confirmation of several pioneer factors (Table 1.1, (Iwafuchi-Doi and Zaret, 2014)), a common, determining structural property has not been identified, yet. FOXA1, for example, interacts with histones for opening the chromatin (Cirillo *et al.*, 2002) while OCT4, SOX2 and KLF4 recognize partial binding motifs (Soufi *et al.*, 2015). The importance of pioneer factors with respect to cellular conversions becomes evident when regarding the validated instances of induced cellular transitions in which they are involved (Table 1.1 and (Morris, 2016)). These include the direct conversion from fibroblast to cell types of all germ layers such as macrophages (Feng *et al.*, 2008), hepatocytes (Huang *et al.*, 2011) and motor neurons (Son *et al.*, 2011).

Table 1.1. Validated pioneer factors

Pioneer factors	Cellular context	Predicted/Validated pioneer activity	References
FoxA	Transdifferentiation from fibroblast to iHep	Induce transdifferentiation Binding to silent chromatin	(Sekiya and Suzuki, 2011; Huang *et al.*, 2011; Gualdi *et al.*, 1996)
Class V Pou (e.g. Oct3/4, Pou5f3)	Reprogramming from fibroblast to iPSCs	Binding to DNase-insensitive chromatin with absent histone modifications	(Soufi *et al.*, 2012)
Group B1 Sox (e.g. Sox2)	Reprogramming from fibroblast to iPSCs	Binding to DNase-insensitive chromatin with absent histone modifications	(Soufi *et al.*, 2012)
Klf4	Reprogramming from fibroblast to iPSCs	Induce reprogramming	(Takahashi and Yamanaka, 2006)
Ascl1	Transdifferentiation from fibroblast to induced neurons	Induce transdifferentiation Binding to DNase-insensitive chromatin in vivo	(Vierbuchen *et al.*, 2010; Wapinski *et al.*, 2013; Son *et al.*, 2011; Caiazzo *et al.*, 2011)
PU.1	Transdifferentiation from fibroblast to macrophages	Induce transdifferentiation Increase accessibility in DNase-insensitive chromatin	(Barozzi *et al.*, 2014; Heinz *et al.*, 2010; Ghisletti *et al.*, 2010; Feng *et al.*, 2008)
GATA4	Transdifferentiation from fibroblast to iHep Liver development	Induce transdifferentiation Binding to silent liver enhancer	(Bossard and Zaret, 1998; Huang *et al.*, 2011)
GATA1	Mitotic bookmarking	Binding to mitotic chromatin	(Kadauke *et al.*, 2012)

Validated pioneer factors and their implication in cellular conversions (modified from (Iwafuchi-Doi and Zaret, 2014))

1.1.3 Viewing Diseases as Cellular Conversions

The complex interconnection of transcriptional and epigenetic regulation enables cells to respond to environmental cues, such as temperature change or chemical exposure, for preserving cellular functions. These signals lead to changes in RNA and protein concentrations, protein-protein interactions or chromatin conformation, which eventually is reflected in modifications of transcriptional and epigenetic interactions. In addition to extrinsic fluctuations, inherent, internal fluctuations of transcription and translation arise naturally due to the stochastic nature of TF-DNA binding and transcriptional initiation by the formation of the pre-initiation complex. If cells are not able to restore proper functioning, i.e. the amount of RNA and proteins as well as the epigenetic configuration, changes manifest in modified cellular behavior, proliferative ability or the induction of differentiation of stem and progenitor cells. Since intrinsic and extrinsic fluctuations are an essential part of cellular regulation for proper functioning and development, the gene regulatory network has evolved to a robust yet flexible system responding only to specific perturbations (Macneil and Walhout, 2011). Some of these perturbations, thus, lead to the stabilization of molecular dysregulations causing the loss of normal cellular function, a disease state. Similar to cellular conversions, this conceptually results in the repositioning of cells on the Waddington landscape to a disease-related valley, which is maintained by the robust design of gene regulation and might be escaped upon cellular perturbation. Therefore, the formation of diseases can be interpreted as the induction of cellular conversions.

Systemic Lupus Erythematosus (SLE), for instance, is a complex autoimmune disease whose onset is caused by genetic and environmental factors. A key pathophysiological indication of SLE is the production of auto-antibodies of the immune system that are deposited in multiple organs and cause constant inflammation. As a consequence, multiple symptoms could occur, such as arthritis, fever, neurologic dysfunctions or butterfly rash, which hampers an appropriate diagnosis in the early stages of the disease (Kaul *et al.*, 2016). In addition to the difficulty of early diagnosis, the development of drugs for treating the disease pathology is significantly impeded by the incomplete understanding of the molecular mechanisms. Notable, rare mutations have been identified in the complement system proteins C2, C4 and C1Q and other genes responsible for clearance of cellular debris, nucleic acid degeneration and increased activation of the adaptive and innate immune system (Kaul *et al.*, 2016). Besides the occurrence of high-risk mutations, large-scale variations of the epigenetic landscape in T-cells have been observed that might be inducible by environmental exposure to chemicals. A previous study identified procainamide, a DNA methyltransferase inhibitor, to induce similar methylation patterns of disease-related genomic loci in CD4$^+$ cells of SLE patients (Lu *et al.*, 2005). Due to the association of DNA methylation with stable silencing of genes, which prevents their transcription, global hypomethylation suggests new potential transcriptional regulatory interactions. In addition, CD4$^+$ cells from SLE patients show global hypoacetylation of histones 3 and 4, which correlates with decreased mRNA levels of chromatin modifying genes EP300, CREBBP, HDAC2, HDAC7, SUV39H2 and EZH2, as well as global H3K9 hypomethylation (Hu *et al.*, 2008). In contrast to DNA methylation, acetylation of histones is associated to less compact, permissive chromatin structures that enable transcription. Thus, its global loss can be presumed to prevent transcription of various genes. Altogether, even though the concepts of cellular conversions and disease manifestation seem to be very different, the genetic, epigenetic and transcriptional dysregulations in Systemic Lupus Erythematosus together with its inducibility by compounds supports their perception as cellular conversions.

Another notable example of diseases, which can be perceived as cellular conversions, constitutes osteoporosis, a progressive bone pathology characterized by elevated bone marrow fat accumulation and reduced bone formation (Hu *et al.*, 2018). In particular, bone marrow fat is accumulated by increased differentiation of mesenchymal stem cells (MSCs) into fat cells, i.e. adipocytes, while differentiation of MSCs into osteoblasts is decreased, which reduces bone formation (Hu *et al.*, 2018). This dysregulation of the adipogenic and osteogenic differentiation potential can be partly attributed to alterations in the transcriptional regulatory network. Besides other transcription factors, RUNX2 and SP7 have been demonstrated to be an integral component of osteogenic differentiation. In particular, MSCs deficient in either of these genes are unable to give rise to osteoblasts (Hu *et al.*, 2018). On the other hand, adipogenic differentiation is directed by PPARG, while transcription factors such as GATA2, FOXA1 and HOXC8 inhibit adipocyte maturation (Hu *et al.*, 2018). These factors are incorporated into a bigger gene regulatory network including (post-)transcriptional regulation by microRNAs and signaling pathways that are susceptible to environmental factors. Therefore, osteoporosis can be regarded as the cellular conversion of healthy mesenchymal stem cells into a diseased state in which the adipogenic differentiation potential is significantly increased. However, other interpretations exist at the level of differentiated adipocytes and osteoblasts. In particular, the shift of healthy to pathologic cellular populations could be regarded as a cellular conversion of osteoblasts into adipocytes resulting in a pathologic proportion.

When viewing diseases as cellular conversions and, therefore, as displacements of cells in different wells of the Waddington landscape, an important medical questions is the search for suitable external stimuli or perturbations reverting the disease state. Classical pharmacology approaches relied on the identification of compounds able to revert the disease phenotype (Takenaka, 2001). Cells from human or mouse disease models were extracted and *in vitro* screened with multiple compounds that were expected to have the desired effects based on functional activity in the disease. After obtaining a suitable drug candidate, its targets were determined to understand the molecular mechanism leading to the phenotypic reversion and facilitate the discovery of compounds for treating other disease pathologies. However, in the era of profiling the transcriptional and epigenetic state of cells and the ever-growing knowledge about the regulatory interactions stabilizing disease phenotypes, this classical approach has become too inefficient. Instead, the focus has moved to prior bioinformatics and system biology analysis to preselect proteins that are effective targets (Wooller *et al.*, 2017; Xia, 2017).

Great efforts have been devoted to identify disease-related genes and their interactions among each other. More specifically, protein-protein interaction networks have been analyzed to obtain a systems level understanding of molecular mechanisms implicated in disease pathologies (Schadt *et al.*, 2009). From a network perspective, the identified genes show a number of interesting properties. First, cancer related genes have an increased node connectivity compared to non-cancer genes, i.e. they are more likely to be hub-genes (Jonsson and Bates, 2006). The importance of these highly connected proteins is supported by experiments in *S. cerevisiae* showing their criticality in phenotypic stabilization and organism survival (Jeong *et al.*, 2001) together with their ability to modulate toxicity (Said *et al.*, 2004). Second, a reconstructed disease-gene network revealed that genes of the same disorder class are organized in strongly connected modules suggesting distinct pathological mechanisms (Goh *et al.*, 2007). Indeed, further analysis confirmed that genes involved in the same disorder are more likely to interact through protein-protein interactions, show increased tissue expression homogeneity, are more highly correlated and are predominantly co-expressed (Goh *et al.*, 2007).

The information obtained from network analyses have been exploited for proposing predictive methodologies inferring drugs and drug-targets as potential therapeutics of disease pathologies. Franke et al. integrated gene ontology (Ashburner *et al.*, 2000), gene co-expression and protein-protein interactions in a single Bayesian network for prioritizing disease-related genes. In particular, this study found that disease genes should be close to each other in the network due to their related biological function and modularization (Franke *et al.*, 2006). Application of the method yielded, for example, the identification of important breast cancer genes like BRCA1 and TP53, but only 34% of known disease related genes were recapitulated in the top 10 predictions, overall. Other approaches follow the same rationale and predict candidate proteins that are close to already known disease genes with respect to random walks and flows in protein-protein interaction networks (Vanunu *et al.*, 2010; Köhler *et al.*, 2008).

Despite the amount of research conducted for detecting disease-related genes, there is a lack of methods for establishing disease-gene-drug relationships. Currently, information about chemical perturbations and drug response is organized in databases and linked to diseases based on gene signatures (Hu and Agarwal, 2009; Lamb, 2007; Lamb *et al.*, 2006). A notable example of these databases is the Connectivity Map (CMap) (Lamb, 2007), which has been successfully used for predicting effective drugs in different human diseases (Hieronymus *et al.*, 2006; Wei *et al.*, 2006; D'Arcy *et al.*, 2011). However, these methods disregard the underlying gene regulatory network that ultimately controls drug response. In view of diseases as cellular conversions, the development of a network-based approach for predicting disease-gene-drug relationships is a tantalizing prospect.

Overall, due to the complex interconnection of transcriptional and epigenetic landscapes stabilizing condition specific phenotypes, it is essential to understand the underlying gene regulatory network for the prediction of instructive factors inducing cellular transitions upon perturbation. Therefore, it is of utmost importance to gain a systems level understanding of the underlying mechanisms provided by the reconstruction of regulatory interaction networks. The next section discusses current approaches for inferring gene regulatory networks in more detail.

1.2 Gene Regulatory Network Inference from Condition Specific Transcriptomics Data

The increasing availability of condition specific gene expression profiles has offered the unique opportunity to reconstruct the gene regulatory networks stabilizing phenotypic identity of cell types or pathological conditions (Liu, 2015). In general, there are different approaches to model GRNs at different levels of detail (Liu, 2015). For instance, the following are the commonly employed levels of detail for modeling gene regulatory interactions:

- **(Is there a regulatory interaction?)** The minimal amount of information required to reconstruct a gene regulatory network is whether there exists a regulatory interaction between two genes. In practice, this notion is often further simplified to identify co-expressed genes sharing a similar gene expression pattern in several conditions, cell or tissue types.
- **(Who is the regulator?)** The second level of information is considering the directionality of the regulatory interactions. Here, the goal is to identify the regulator for every two genes exhibiting an interaction.
- **(What is the mode of action?)** When the directionality of the interactions is known, they can be further distinguished by their mode of action. In the context of transcriptional regulation, genes can act as activators or inhibitors resulting in increased or decreased transcription, respectively.
- **(What is the regulatory strength?)** The regulatory strength of an interaction is the most complex level of information due to the influence of various mechanisms, like the binding affinity of the TF to DNA or the strength of the protein-protein interaction with the transcriptional machinery.

In the remainder of this section, current methodologies, which model gene regulatory networks of different detail, are discussed, including their particular advantages and limitations.

Gene co-expression networks offer a naïve, yet widely used view on the coordinated control of genes in living organisms (Eisen *et al.*, 1998; Ben-Dor *et al.*). Based on the identification of patterns across different gene expression profiles, relationships between genes are established representing their synchronized actions. Thus, co-expression networks offer insights into coordinated expression indicating which genes are simultaneously active and are therefore assumed to belong to the same biological process. The most recent advancement includes the introduction of differential co-expression analysis that aims to identify genes significantly more co-expressed in one condition compared to another (Fiannaca *et al.*, 2015; Bhar *et al.*, 2013; Amar *et al.*, 2013). Several earlier studies reported differentially co-expressed TFs to be involved in cancer (Kostka and Spang, 2004; Lai *et al.*, 2004; Amar *et al.*, 2013) and found a significant decrease in co-expression compared to healthy samples (Kostka and Spang, 2004) highlighting the importance of differential analysis of regulation. However, gene co-expression networks solely offer insights about correlations, i.e. coordinated expression, rather than identifying causal regulatory relationships, which constitutes the major drawback of these methods. In the presence of co-expression, it is not determined what mechanism or set of genes causes the observed pattern and whether the same process regulates all strongly correlated genes at the same time. Also, from a methodological point of view, co-expression analysis assumes the observed correlation to be consistent across unobserved datasets, which does not hold in most of the cases. Even differential co-expression networks cannot fully address this issue as they impose the correlation to be consistent in unobserved samples of the same condition. As an example, cancer cell populations exhibit highly heterogeneous expression patterns, which invalidates this assumption (Mentzen *et al.*, 2009).

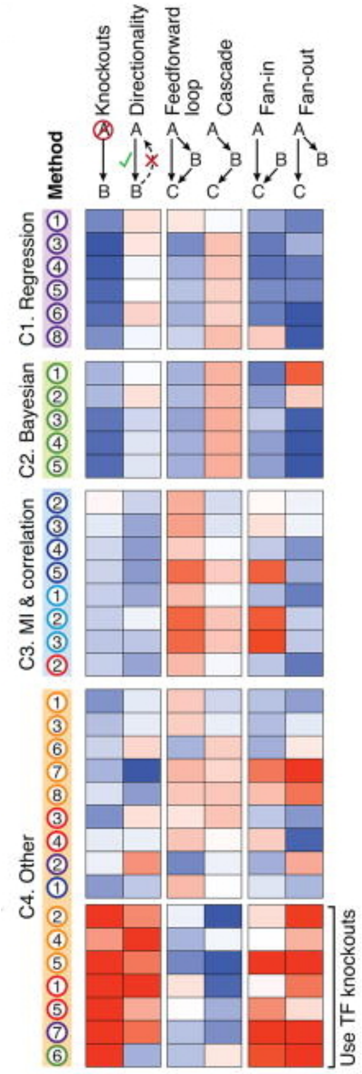
In order to address the existing limitations of undirected gene co-expression networks, great efforts are being devoted to the reconstruction of directed gene regulatory networks (Marbach *et al.*, 2012). The developed methods are based on a variety of mathematical techniques while most of them can be classified into three main categories, regression-based, mutual information-based and Bayesian networks. All of these methods share the goal of identifying networks that are directed but do not contain information about the mode of action of individual interactions or their strength.

Most available methods relying on regression analysis for identifying regulatory interactions, or gene regulatory networks in general, perform least absolute shrinkage and selection operator (LASSO) regression (Meinshausen and Bühlmann, 2010; Yuan and Lin, 2006; van Someren *et al.*, 2006; Lèbre *et al.*, 2010; Haury *et al.*, 2012). Since its introduction in 1996, LASSO regression got increasing attention due to its ability to perform feature selection and regularization at the same time. In the context of gene regulatory network inference, other regression techniques would retain small weights for every possible interaction while LASSO is able to efficiently select which interactions should not appear at all and pinpoint most influential regulatory mechanisms.

Methods based on mutual information aim to identify the interdependence of two genes. In particular, the mutual information obtained from expression sample of two genes across several conditions is a measure of how much information of one gene can be obtained by the other. Since this measure is symmetric, it does not readily provide information about the directionality of interactions justifying the need for combining it with other formalisms. Available methods additionally use Bayesian network approaches for inferring the causal relationship between two genes (Faith *et al.*, 2007; Mani and Cooper, 2004). The goal is to recover a network structure from the data representing the conditional dependencies between genes. However, these methods typically result in dense regulatory networks, since they cannot faithfully identify the absence of interactions between conditionally independent genes (Figure 1.6, “Fan-in” column).

A thorough comparison of available methods based on mutual information, regression and Bayesian networks revealed the elevated importance of choosing the appropriate mathematical framework compared to the actual implementation (Marbach *et al.*, 2012). The reconstructed networks of all methods are more similar among the same mathematical framework than among others, each having particular advantages and limitations for correctly predicting certain regulatory motifs (Figure 1.6). While mutual information based methods provide more confidence in the identification of feed-forward loops, they are mostly disadvantageous in the prediction of regulatory cascades. Complementary, regression-based and Bayesian network techniques provide more confidence in the predicted cascades compared to feed-forward loops. However, it is important to note that all methods perform similarly in identifying the directionality of interactions even though mutual information approaches show slightly reduced performance overall, which hinders the recapitulation of dynamic changes induces by transcription factor knockout experiments.

Figure 1.6. Comparison of available network reconstruction methods



Comparison of network reconstruction methods on the basis of known motifs. Rows represent individual methods and the color-coding depicts the prediction confidence in comparison to other methods from less (blue) to more confident (red). Methods from the same class show highly similar performance biases indicating that the choice of the modeling framework is more important than the actual implementation. Regression and Bayesian network based methods show stronger performance in resembling cascades while mutual Information and correlation based methods are more prompt to detect feed-forward loops. Other methods using transcription factor knockout experiments show clear advantages over all other methods in reproducing knockouts. However, they lack the ability of faithfully reconstructing network motifs. (Modified picture from (Marbach *et al.*, 2012)).

Regarding the accurate modeling of cellular gene regulatory networks and its subsequent utilization for identifying instructive factors for cellular conversions, these results highlight the inability of level two network models to provide faithful predictions and, therefore, the need for more sophisticated formalisms. Continuous modeling frameworks are the most desirable form of network representation as they allow the precise dissection of the transient network behavior through the interplay of regulatory interactions with different strengths. In this setting, an interaction is typically formulated in terms of hill functions extrapolated from experimental evidence in the context of effector concentration on the synthesis rate of enzymes (Polynikis *et al.*, 2009; Yagil and Yagil, 1971). Depending on the regulatory mode of action, these functions are increasing if the gene is an activator and decreasing if it is an inhibitor. The joint effect of multiple interactions on the same target gene is then modeled as an ordinary differential equation combining individual interactions additively. This modeling framework has got considerable attention for reconstructing gene regulatory networks of, for example, the cell cycle (Nachman *et al.*, 2004), circadian system in *A. thaliana* (Locke *et al.*, 2005) or carbon starvation response in *E. coli* (Ropers *et al.*, 2006).

Even though continuous models of regulation are desirable for modeling gene regulatory networks, due to their explicit inclusion of underlying chemical properties, considerable practical constraints limit their applicability. In particular, the kinetic parameters are mostly unknown in the context of transcriptional regulation (Le Novère, 2015) and are therefore subject to inference based on transcriptional (time-series) data. Albeit the dramatic increase of available experimental data in the advent of next generation sequencing techniques (e.g. RNA-seq), the reconstructed networks are typically restricted to a few genes whose interactions can be faithfully determined from gene expression data. More importantly, continuous models utilize gene expression measurements as a proxy for protein expression. However, there exists only a moderate relationship between these two quantities (Maier *et al.*, 2009). This significantly impedes the reliability of the observed transient evolution of the system upon perturbation, which is an integral part in the determination of instructive factors for cellular conversions. Finally, these models usually contain many parameters that are subject to estimation and pose substantial computational challenges to the identification of medium and large gene regulatory networks (Vijesh *et al.*, 2013).

In the early 1970s, Kauffman introduced Boolean networks, a particular class of logical models, as mathematical representations of gene regulatory networks (Glass and Kauffman, 1973; Kauffman, 1969). In contrast to the models discussed before, a gene can only take on two values, active and inactive, and gene regulation is represented by Boolean logic functions over a set of regulators determining the state of the regulated gene. Importantly, the Boolean representation of genes decreases the complexity of network inference considerable while the superimposition of Boolean functions connecting the regulatory interactions contribute to the complexity. In this context, cell-types are represented by discretized gene expression profiles, i.e. genes are deemed active or inactive based on their mRNA abundance, that are stable states of the networks. A variety of methods were developed over the past years aiming for more accurate network inference (Dorier *et al.*, 2016; Terfve *et al.*, 2012; Barman and Kwon, 2017; Crespo *et al.*, 2013; Melas *et al.*, 2013) and applied to study haematopoiesis (Bonzanni *et al.*, 2013), embryonic stem cells (Xu *et al.*, 2014) or cancer (Grieco *et al.*, 2013; Wittmann *et al.*, 2009; Calzone *et al.*, 2010). Due to their reduced complexity, Boolean networks address the main limitations of continuous models for gene regulation. First, the parameter search space is significantly smaller as an interaction can be only present or absent and the regulatory strength does not differ. Therefore, the amount of data needed for faithfully reconstructing gene regulatory networks is dramatically reduced. Second, even though gene expression is utilized as a proxy for protein expression, like in continuous modeling frameworks, discretization by differential expression analysis is assumed to yield higher correlations with protein abundance compared to raw measurements and, as such, increases the reliability of the model (Kosti *et al.*, 2016). Altogether, these advantages allow for the reconstruction of larger networks compared to continuous modeling frameworks, which provides a more complete view on the regulatory mechanisms responsible for phenotypic stability. Undoubtedly, Boolean networks also possess limitations. While, for example, the discretization of gene expression by differential expression analysis increases the correlation with protein abundance, quantitative differences influencing protein concentrations, translating to the number of concurrent regulatory events, are neglected. In addition, kinetic preferences of certain genes to proximal or distal regulatory regions are assumed to play a minor role in gene regulation and every regulator corresponds equally to the activation or inhibition of their target genes. Finally, Boolean networks possess only a discrete time representation that does not allow for modeling slow or fast processes.

The different mathematical modeling frameworks reviewed in this section have been used in various ways to uncover regulatory interactions of the underlying cellular network. In this regard, several methods have been developed in past years aiming at the exploitation of these models for predicting instructive factors of cellular conversions rather than solely describing static gene regulatory networks.

1.3 Systematic Identification of Instructive Factors for Cellular Conversions

Despite the interest in identifying cell type specific networks that closely resemble the real, partially known gene regulatory networks, a model's quality is typically assessed in terms of its predictive rather than descriptive power. The advantage of models describing the regulation of particular genes in the context of other genes has been exploited in many methods for identifying instructive factors triggering the transition from one condition or cell type to another. In this section, the computational models for predicting these instructive factors (see Table 1.2, modified from (Bian and Cahan, 2016)) are comprehensively reviewed.

Two of the presented methods are network-free relying solely on gene expression or histone modification data. *D'Alessio et al.* (D'Alessio *et al.*, 2015) developed a method for identifying the expression specificity of transcription factors in a set of 233 cell and tissue types. For each sample, a set of ten most specific core TFs has been computed that are most representative of the given cell or tissue type. These core TFs then serve as a starting point to induce cellular conversions and have been demonstrated in the conversion from fibroblasts to retinal pigment epithelial cells. One important limitation of this approach constitutes the biasedness of core TF identification towards high expression. Established collections of expression profiles already revealed that genes are expressed on different scales and typically lowly expressed TFs cannot be captured by this approach. From a practical point of view, the set of core TFs is too large to be efficiently used in experiments, but could be further reduced by identifying the underlying regulatory network governing the initial cell type to be conversed.

Table 1.2. Current methods for identifying instructive factors

Method	Input Data	Species	Cell/Tissue Types	Output
CellNet	Microarray gene expression	Human, mouse	20 cell/tissue types	Similarity to known cell/tissue types and predicted instructive factors
Mogrify	Initial/Final cell type	Human	Gene expression of 300 cell and tissue types	Predicted TFs for cellular conversion
D'Alesio	Target cell type	Human	TF expression of 233 cell and tissue types	Predicted TFs for cellular conversion
Crespo	Gene expression profile/Prior knowledge GRN	Human, mouse	Published networks of various cell types	Predicted set of core TFs for cellular conversion
Davis	Gene expression/Chromatin profiles	Human, mouse	65 datasets	Predicted TFs for cellular conversion

The other method, presented by *Davis and Eddy* (Davis and Eddy, 2013), uses gene expression profiles and polycomb repressed genomic regions marked by H3K27me3 for identifying potential instructive factors. Based on the analysis of 65 published sets of instructive factors, genes that are prompt to induce the trans-differentiation from an initial to a desired cell type are more likely to be polycomb repressed than genes not associated to the induction of the conversion. Therefore, a prioritization scheme is proposed based on the differential polycomb repression and expression of genes in the initial and final cell type. Nevertheless, the practical utility of this approach is significantly impeded by the absence of a threshold for selecting potential candidate genes based on the proposed analysis. In case the threshold is chosen inappropriately, the number of identified genes is intractably high or too small, respectively. It further neglects the causal relationship among genes and their ability to directly or indirectly modify the epigenetic landscape through the downstream expression of TFs and their co-factors.

The second class of methods relies on the identification of networks from various data sources and, to date, constitute the most promising approaches for systematically identifying the instructive factors for triggering desired cellular conversions. CellNet aims to reconstruct directed (level 2) networks from cell/tissue specific gene expression and transcription factor binding (ChIP-seq) datasets (Cahan *et al.*, 2014). For the initial study, every cell type was required to be represented by at least 60 gene expression profiles including perturbation experiments for inferring a single gene regulatory network from the complete dataset of all cell and tissue types. Condition-specific GRNs are subsequently obtained by identifying densely interconnected sub-networks. The resulting compendium of cell and tissue type specific GRNs can be used for confirmatory and exploratory analysis. For that, CellNet first estimates probabilities for a query expression sample to correspond to one of the included cell types. Second, CellNet ranks transcription factors based on a postulated 'Network Influence Score' measuring their expression in comparison to other cell types. This strategy has been applied to increase the fidelity of the B-cell to macrophage conversion by additionally knocking down the expression of Pou2af1and EBF1, which have been identified as essential B-cell genes (Morris *et al.*, 2014). However, even though CellNet proved its utility in certain cases, its predictive power is significantly hampered by the vast amount of data required for building the background GRN. More specifically, the background network is assembled from publicly available ChIP-seq datasets for determining potential regulatory interactions between transcription factors. For introducing new cell or tissue types in the approach, specific transcription factor binding site data must be collected. However, necessary amounts of data are usually solely available for widely studied cell types. To date, CellNet consists of only 16 human and 20 mouse cell types making it impractical as a general tool for studying cellular conversions.

Mogrify (Rackham *et al.*, 2016) follows a network-based approach, as well. It uses publicly available databases of regulatory interactions to estimate the individual effect of overexpressing a single TF in the starting cell type and subsequently identifies a minimal set of TFs that is potentially able to up-regulate all TFs associated to the target cell type. Following this approach, novel instructive factors have been identified for converting dermal fibroblasts to keratinocytes and keratinocytes to microvascular endothelial cells. The inherent disadvantage of the transcription factor prioritization performed by Mogrify is the use of networks composed of undirected interactions (level 1) from curated databases as well as predictions. The absence of directionality does therefore not guarantee the desired effect upon up-regulation of a TF while the predicted interactions introduce an unquantifiable level of noise in the network.

Finally, *Crespo et al.* (Crespo *et al.*, 2013) developed a method that reconstructs Boolean gene regulatory networks given a prior knowledge network and a differential expression profile of the initial and final cell type. The reconstructed network allows the identification of higher-level regulatory motifs (e.g. positive feedback loops) upon which a minimal set of transcription factors is selected corresponding to the ones with the highest impact on the network state. Thus far, it has been shown to reproduce existing trans-differentiation and reprogramming protocols like the conversion from myeloid to erythroid and fibroblasts to IPSCs. Nonetheless, the approach requires explaining the gene expression programs of highly unrelated cell types within the same gene regulatory network topology. However, as described in section 1.1, the gene regulatory interactions are modulated through epigenetic differences of the initial and final cell types. Another limitation of Boolean network-based approaches, like the one from *Crespo et al.,* is the requirement of discretized gene expression profiles. Typically, this discretization is obtained through differential expression analysis of the cell types under study, which creates an isolated view of the two conditions under study (Hudson *et al.*, 2012). This makes it impossible to compare the condition specific networks of multiple phenotypes.

1.4 Summary

There is a continuous development of new methodologies for identifying instructive factors of cellular conversions. GRN based methodologies, which rely on the analysis or reconstruction of gene regulatory networks representing the complex regulatory relationships between genes and transcription factors, are widely used and have shown their utility in predicting instructive factors of cellular conversions. Based on ever-growing information about the gene regulatory interactions, every cell type is determined by its own gene regulatory network stabilizing the resulting phenotype through the interplay of transcriptional activators, inhibitors and epigenetic modifiers. These networks dynamically change during the ectopic expression of transcription factors initiating cellular conversions and stabilize again to represent a desired phenotype.

Existing methods for predicting instructive factors suffer from at least one of three major limitations. First, network-based methodologies do not account for the differences in transcriptional regulation modulated by the epigenetic landscape of covalent histone modifications and chromatin accessibility and model multiple phenotypes within a single network topology. Second, the data requirements limit the general applicability of these methods to cell types or conditions not given broad attention. Especially for studying disease pathologies where transcriptomics and epigenetics datasets are typically highly limited, methodologies with low data requirements are essential for the identification of instructive factors inducing disease phenotype reversion. Finally, some of the approaches only offer a predefined catalogue of different cell or tissue types for computing instructive factors for cellular conversions and are not extendible with user-defined input data. Similar to high data requirements, the lack of extendibility of the computational approaches constitutes a severe limitation in the analysis of disease-related phenotypes.

Addressing these main limitations are the keynote aims of this thesis and the developed solutions will be presented in the remainder of this thesis.

CHAPTER 2 Scope & Aims of Thesis

The complexity of transcriptional regulation presents a major challenge to the systematic identification of instructive factors inducing desired cellular transitions upon perturbation. While next generation sequencing methodologies enabled a cost efficient and accurate screening of cellular phenotypes, chromatin accessibility and transcription factor binding sites, the understanding of the interplay of transcriptional and epigenetic regulation in the stabilization of cellular phenotypes is still limited. The virtual infeasibility of compiling complete gene regulatory networks for specific cell types or pathological conditions hinders the development of systematic approaches for deducing multitarget combinations of instructive factors that potentially induce cellular conversions. Nevertheless, experimental evidence has already provided an impression of the complexity and specificity of gene regulation in distinct cell types.

Therefore, the major issue in developing a systematic approach for studying cellular conversions is the understanding of the gene regulatory circuitry maintaining and destabilizing particular phenotypes. The analysis is further complicated by the limited amount of data for providing a window into the accessible chromatin landscape and the computational tools for analyzing transcriptomics and chromatin accessibility data that do not provide reliable information in the context of network reconstruction.

Due to these difficulties a wealth of methods has been developed for identifying instructive factors of desired cellular conversions. The most promising approaches reconstruct gene regulatory network models of various levels of detail, as described in the first chapter (see section 1.2), and deduce potential combinations of factors based on their analysis. All of these approaches have their own advantages and limitations while all network-based methodologies share the inability to reconstruct condition specific regulatory networks.

This thesis proposes to approach the modeling of gene regulatory networks by reconstructing condition specific network models and exploiting their topological characteristics combined with *in silico* perturbations to predict and score multitarget combinations of instructive factors that potentially induce desired cellular transitions.

2.1 Thesis Aims

Aim 1. Reconstruct condition specific networks by integrating available biological information on experimentally validated regulatory interactions in diverse cell types with condition specific transcriptomics and chromatin accessibility data. This involves the selection of a suitable modeling framework able to represent transcriptional regulation and the design of a strategy for reconstructing networks. Subsequently, the approach will be validated with condition specific gene regulatory interactions identified by ChIP-seq experiments.

Aim 2. Integrative analysis for identification of candidate instructive factors for cellular conversion by examining the topological features of the initial and final condition specific networks that were reconstructed from transcriptomics data and/or available chromatin accessibility information. This requires the choice of informative topological network motifs for obtaining dependable results. Identified candidate genes and multitarget combinations are then validated against previous studies implicating them in phenotypic stabilization and applied to the cellular conversion of adipocytes into osteoblasts as well as to the prioritization of drugs for inducing desired cellular transitions.

Aim 3. Overcome limitations of downstream computational tools for processing transcriptomics and chromatin accessibility data in order to obtain reliable datasets that can be readily used for reconstructing and analyzing gene regulatory networks. First, a machine-learning framework is presented that predicts gene-level accessibility for each gene in order to overcome the limited amount of available chromatin accessibility data. Validation of the predictions will be performed against experimental chromatin accessibility assays processed with current peak-calling methodologies on the basis of a compiled gold standard dataset. Second, a method for absolute discretization of gene expression data is developed that classifies genes into active and inactive states. With it, the consistent comparison of multiple cell types and conditions within the same network reconstruction methodology will be feasible. Validation of the method will consist of the qualitative and quantitative comparison to current state-of-the-art approaches.

2.2 Originality

The integrative gene regulatory network reconstruction approach combines multiple, available sources of biological information to better characterize condition or cell type specific regulatory networks. With it, the chromatin accessibility landscape can be projected onto the transcriptional regulatory network and utilized for studying differential regulatory mechanisms of cellular conditions. However, on one hand, current computational tools for processing experimental chromatin accessibility assays are not able to reliably detect accessible genomic regions. On the other hand, despite the ever-growing amount of experimental data, chromatin accessibility assays are not available for many cell types or cellular conditions. The proposed approach addresses these limitations by predicting whether genes are located in accessible or inaccessible chromatin domains in a condition specific manner. This way, it assists in overcoming the unavailability of experimental data and can be further utilized to optimize current tools for interpreting experimental data. Overall, the combination of the developed methodologies for gene regulatory network reconstruction and chromatin accessibility prediction allows the identification of candidate instructive factors that potentially induce desired cellular conversions. Previous approaches have used network based inference methods, but require tremendous amounts of data, limiting their applicability to other cell types or conditions, and do not take into account the distinct gene regulatory interactions of different cell types. However, these distinct regulatory programs, which are mediated by transcriptional and epigenetic landscapes, have been demonstrated by combined large scale probing of the gene expression and chromatin profiles. That these changes impact the cellular response to gene perturbations is undeniable and makes them valuable to model.

This thesis aims to address the main limitations of current methodologies for modeling gene regulatory networks and, thus, provides a systematic analysis approach integrating available biological knowledge about transcriptional gene regulation, transcriptomics and epigenetics data that could serve as a general strategy for selecting suitable combinations of genes that potentially induce desired cellular transitions upon perturbation.

CHAPTER 3 Methods

The prediction of instructive factors for cellular conversion, i.e. the identification of genes with the most pronounced effect on the phenotype upon perturbation, requires an understanding of the transcriptional regulatory processes occurring in the cell types under study. Current methodologies impose the strong assumption that the underlying gene regulatory networks do not change but only differ in the expression of their constituent genes. This implies that the set of regulated genes stays the same for each transcription factor, regardless of the cell type, and only the influence on transcription changes. With the availability of the latest experimental protocols to measure accessible genomic regulatory regions (e.g. DNase-seq, FAIRE-seq, ATAC-seq) this assumption has been invalidated.

To increase the reliability of identified combinations of instructive factors for cellular conversions a different strategy is needed that considers different transcriptional regulatory programs in different cell types. This thesis proposes one such method, which reconstructs condition specific transcriptional regulatory networks that are subsequently analyzed to identify candidate instructive factors.

For constructing cell-type specific transcriptional regulatory networks a suitable mathematical modeling formalism is needed. The Boolean network model has been chosen to represent genes and their effects on the transcription of other factors. The network is reconstructed by selecting a subset of experimentally validated interactions between genes from prior knowledge networks, in which one factor activates or inhibits the transcription of other factors such that the model represents the phenotype of a single cell type or condition. Candidate instructive factors are obtained by analyzing the reconstructed Boolean network models and scored through *in silico* simulation of their perturbation effects on the network.

The reconstruction of Boolean gene regulatory networks crucially depends on an accurate classification of gene expression measures (e.g. from RNA-seq or microarray experiments) into active or inactive states. Traditionally, a variety of statistical tests are carried out to obtain significantly up- and down-regulated genes in the initial and final cell types, which serves as the desired binary classification. However, in the presence of multiple cellular conditions of the same cell type, these tests are either not applicable or have low statistical power. Therefore, the method for reconstructing regulatory networks is complemented with a reference-based approach for accurately classifying gene expression data allowing for the eventual comparison of multiple cell types or conditions on the network level.

Following the approach of classifying genes into active or inactive states is necessary for reconstructing regulatory networks. However, cellular transcriptional regulation is governed by the epigenetic landscape of the cells and, in particular, accessible chromatin domains allow the transcription of genes. Conversely, if a gene is classified to be inactive or not expressed it can be either due to transcriptional repression or its location in inaccessible chromatin regions preventing transcription. In a network model, genes in inaccessible chromatin regions do not have to be explained by transcriptional repressors and should thus be disregarded. However, experimental accessibility measures are not available for all cell types and conditions making it necessary to predict the genes located in inaccessible chromatin domains.

This chapter provides the methodological background of the developed approaches for addressing these issues. Section 3.1 provides a formal introduction of Boolean networks and their analysis with respect to their application as a model for gene regulatory networks. Section 3.2 gives a detailed overview of the method for reconstructing cell-type specific gene regulatory networks and section 3.3 describes the analysis of these networks and the strategy for selecting combinations of candidate instructive factors of cellular conversions. Section 3.4 describes the machine learning approach for predicting the accessibility of genes from transcriptomics data and, finally, section 3.5 describes the method for classifying gene expression measures into active and inactive states by statistical comparison against a reference distribution for enabling the consistent comparison of multiple cellular conditions.

3.1 Boolean Networks as Models of Gene Regulation

A Boolean network can be intuitively envisioned as a graph consisting of nodes that are connected by directed edges. Each node can take on one of two values (0 or 1) and is governed by a logic function consisting of all parent nodes in the graph. Definition 1 formalizes this notion of a Boolean network.

Definition 1 (Boolean network). *A Boolean network B = (V,E,F,s) is a directed graph with nodes V := [1,n] for some $n \in \mathbb{N}$, edges $E \subseteq V \, x \, V$, a set of Boolean functions*

$$F := \{f_i: \{0,1\}^n \rightarrow \{0,1\} | i \in V\}$$

and sign-function

$$s : E \rightarrow \{-1,1\}$$

that assigns interaction types to the edges $e \in E$.

*Let $v \in V$ be a node of the network. Then $V_{-}(v)$*and *$V_{+}(v)$ denote the set of predecessors and successors, respectively. A predecessor $v' \in V_{-}(v)$ is then called a regulator of v.*

In the context of gene regulatory networks, the nodes describe genes and edges denote the transcriptional regulatory effect of a gene on another gene. For convenience, the terms nodes and genes as well as edges and interactions, respectively, will be used interchangeably in the following. In this setting, genes can either activate or inhibit their target genes, denoted by the sign-function that assigns *1* to activating and *-1* to inhibiting interactions.

The set of Boolean functions defining the network allows the specification of the dynamic behavior, i.e. the transition from one network state to another. In this context, a state is an assignment of 0 (inactive) or 1 (active) to every gene and the transition from one state to another corresponds to the application of the Boolean functions to this assignment. The syntax and semantics considered for Boolean functions in this thesis are defined in Definition 2 and 3.

Definition 2 (Syntax of Boolean Functions). *A Boolean function can be inductively defined, using the Backus Naur form* (Knuth, 1964)*, as*

$$\varphi := true \mid false \mid (\varphi \,\&\, \varphi) \mid (\varphi \mid \varphi) \mid !\phi \mid v$$

where $v \in V$ is a node.

Definition 3 (Semantics of Boolean Functions).

- *true is always true and false is never true.*
- *$(\varphi \,\&\, \varphi')$ is true, if φ is true and φ' is true.*
- *$(\varphi \mid \varphi')$ is true, if φ is true or φ' is true or both are true.*
- *$!\,\varphi$ is true, if φ is false.*
- *v is true, if $x_v = 1$ for a given state x*

Informally, genes contribute to the regulation of their targets only if they are active (expressed), which is encoded as *true* in the definition of the Boolean syntax and semantics, and can have *and* (&) and *or* (|) relationships. *And*-relationships require that all regulators are active, such as in the presence of protein complexes required for regulation, while *or*-relationships require the activity of a single factor. Each gene in the network is represented as a literal v while its discretized expression is denoted by x_v. Based on the previous definitions of Boolean functions, the one-step dynamics of the network is provided in Definition 4.

Definition 4 (One-step dynamics of Boolean networks). *Let B = (V,E,F,s) be a Boolean network and $x^t \in \{0,1\}^n$ a state of the network at a discrete time instance t. The state of node i at time t+1 can be defined as:*

$$x_i^{t+1} = f_i(x^t)$$

Then, the state of the entire network B can be defined as

$$x^{t+1} = F(x^t)$$

where the i-th function in F determines the i-th state at time t+1.

Boolean networks evolve in discrete time-steps and the state of each node of the network in the next step is solely controlled by the logic function for this state and the expression of its regulators. A successive application of these functions is called a *simulation run* – or simply simulation – of the network (Definition 5).

Definition 5 (k-Step Simulation). *Let B = (V,E,F,s) be a Boolean network and $x^t \in \{0,1\}^n$ a state of the network at time t. A k-step simulation of the network starting from x is a path*

$$x^t \to x^{t+1} \to x^{t+2} \to \cdots \to x^{t+k}$$

such that $x^{t+i} = F(x^{t+i-1})$ for $1 \leq i \leq k$.

While the next state of each node in the network is determined by the application of the corresponding logic function, there exist two different methods, called update schemes, for computing the network state. First, the asynchronous updating in which only one randomly selected node is updated and, second, synchronous updating in which all of the nodes in the network are updated simultaneously to constitute the next state. Throughout this thesis, the simulation of the Boolean networks follows the synchronous updating scheme exclusively.

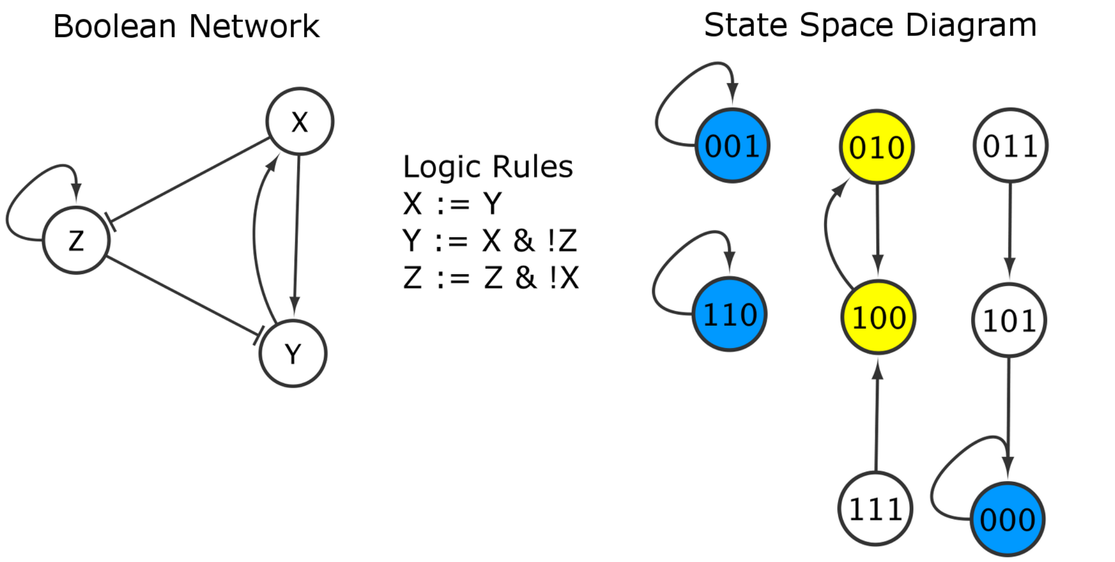
Suppose to simulate a finite Boolean network for an infinite number of steps, then – because of the finite state space – at least one network state must be visited multiple times. All the states having this property of being visited multiple times for a simulation starting from any state are called attractors. Two distinct types of attractors can be distinguished, fixed-point and cyclic attractors. Fixed-point attractors, sometimes referred to as steady states, are states whose individual node states do not change upon invocation of the Boolean functions whereas cyclic attractors are a series of network states that indefinitely repeats in a simulation. Definition 6 defines fixed-point and cyclic attractors more precisely.

Definition 6 (Attractors). *Let B = (V,E,F,s) be a Boolean network and* $x, x^t \in \{0,1\}^n$ *network states. Then*

x is called a fixed-point attractor if and only if $x = F(x)$.

A path is called a cyclic attractor if and only if the path is of the form $x \rightarrow x^t \rightarrow \cdots \rightarrow x^{t+i} \rightarrow x$ *such that* $x = F\left(x^{t+i}\right), x^t = F(x)$ and $x^{t+j} = F(x^{t+j-1})$

Figure 3.1. Boolean Network Example



Boolean networks are defined by their topology (left) and corresponding logic rules (middle). In the context of transcriptional regulatory networks, genes repressing the transcription of their target are represented by dashed lines and activators as arrows. The logic rules define which network configurations activates the gene. The networks state space shows all possible transitions of network configurations when simulated with a synchronous updating scheme. In this example, there exist three point attractors (blue) and one cyclic attractor (yellow).

Figure 3.1 shows an example network with its corresponding Boolean functions (logic rules) and state space diagram. It contains three genes that are activating (arrows) or inhibiting (dashes) other genes. The topology of the network in conjunction with the given logic rules entirely define the Boolean network. Here, inhibition is formalized as the requirement of the gene to be inactive or, more precisely, not active. The state space diagram contains all possible network states and subsumes all paths of a simulation with edges representing the one-step dynamics of the network. Steady states, shown in blue, are characterized by self-loops in the diagram corresponding to their definition of being a fixed-point of the Boolean functions whereas the only cyclic attractor, shown in yellow, corresponds to an alternation of states ‘100’ and ‘010’ as soon as one of these states is reached.

3.2 Reconstruction of cell-type specific gene regulatory network models

Gene regulatory networks and, more particularly, transcriptional regulatory networks represent the activatory or inhibitory effect of genes on each other. The reconstruction of condition specific networks thus requires determining the interactions between genes that take place in a particular condition while not including those specific to other conditions. The combination of these interactions gives then rise to a specific phenotype partly defining the condition under study.

The reconstruction of gene regulatory networks foremost requires a mathematical modeling formalism representing the condition under study. The formalism chosen here is that of Boolean networks introduced in the previous section of this thesis. Recall that nodes of the network represent genes while interactions specify a gene regulatory effect of one gene on another, i.e. activating or inhibiting another gene. Depending on the amount of mRNA of each gene in the network it is assigned a node state of 0 (inactive) or 1 (active). The state of the network consists of the states of the individual nodes.

It is important to note that employing Boolean networks makes a number of assumptions and simplifications that have to be recognized. First, by using the amount of mRNA to define the network state, the model assumes that the amount of proteins is high if the gene is labeled active and low if it is labeled inactive, and these proteins are able to actively regulate the expression of other genes. Second, Boolean networks are deterministic and, thus, do not account for the inherent stochasticity of the gene regulatory mechanisms within the cells. Finally, the strength of each interaction is the same and does not account for different affinities of proteins to the DNA.

The approach for reconstructing cell type specific networks followed in this thesis selects a set of interactions from a prior knowledge network such that the underlying Boolean network is compatible with a given binary gene expression profile. Specifically, the network composed of the selected interactions is required to have a point attractor corresponding to a discretized gene expression profile, which reflects the stability of the condition under study. The method presented here utilizes a genetic algorithm for reconstructing condition specific networks, which will be detailed in the remainder of this section.

3.2.1 Prior knowledge network

The network reconstruction process relies on the information of potential gene regulatory effects stabilizing the cell type under study. Given a set of genes that should be part of the network, experimentally validated interactions are compiled from scientific publications. In particular, MetaCore™ from Thomson Reuters, a highly curated database of molecular interactions reported in literature, is queried to obtain all experimentally validated direct interactions among these genes. The selected interactions are classified to be directly involved in the regulation of the target gene and their effect, i.e. activation or inhibition, is mostly experimentally supported. Those interactions whose effect is unspecified, i.e. unsigned interactions, can also be included in the prior knowledge network. The mode of action is then subject to inference during the reconstruction of the network. Importantly, even though literature-based interactions are used in the work presented in this thesis, the prior knowledge network could be generated using different approaches.

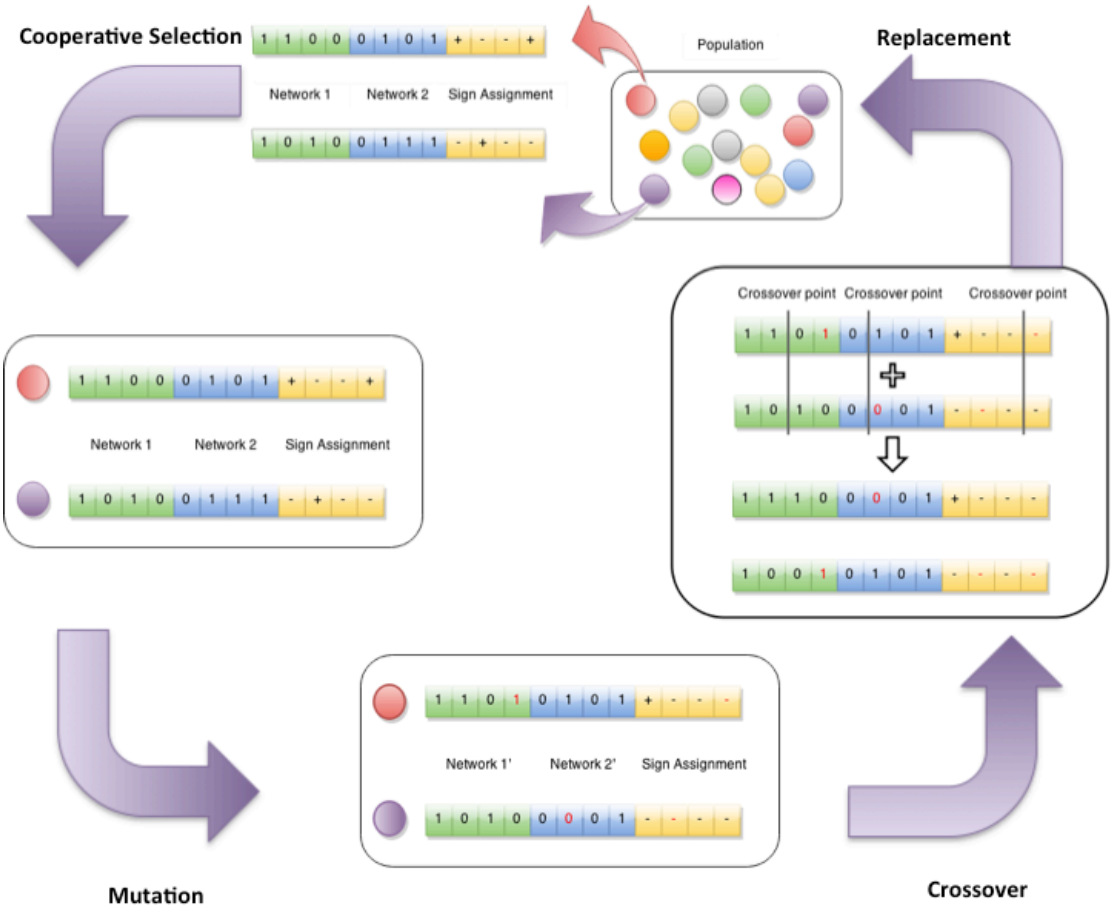
3.2.2 Genetic Algorithm for Boolean network reconstruction

The prior knowledge networks from MetaCore™ are rather heterogeneous due to their composition of interactions from different cell lines, cell types or tissues with different experimental conditions, which necessitates a procedure for selecting the interactions giving rise to the phenotype under study. For determining the network dynamics, the majority rule is selected as the logic scheme for each gene. With it, a gene is active if the number of active activators exceeds the number of active repressors. In the context of Boolean networks, the point attractors of the prior knowledge network do not necessarily match the binary gene expression profile of the studied biological condition. For obtaining a network that does have a corresponding attractor, the prior knowledge network must be contextualized to the phenotype. In order to accomplish this task, the proposed method utilizes a genetic algorithm for finding a sub-network by iteratively refining the result. With regard to its usage for identifying instructive factors of cellular conversions given an initial and final cell type or condition, the method reconstructs both corresponding networks at the same time to save computational resources.

As previously described, the prior knowledge network might contain unsigned interactions with unspecified mode of action. A sign for these interactions has to be determined during the contextualization process since the attractors of the network depend on them. Therefore, the mode of action is inferred such that it is consistent with both the initial and final network.

The method for reconstructing the cell type specific networks is based on a genetic algorithm comprised of a population of individuals. Each individual is implemented as a binary array representing the presence or absence of interactions from the prior knowledge network for both conditions and an inferred sign assignment (Figure 3.2).

Figure 3.2. Workflow of the genetic algorithm



The genetic algorithm for reconstructing networks contains an initial population of random individuals. Each individual contains three parts, the network for the first phenotype, the network for the second phenotype and a predicted mode of action for each unsigned interaction (activation: '+', inhibition: '-'). If an interaction from the prior knowledge network is present it is denoted by '1' and '0' if it is absent. The population is updated iteratively following a three-step approach for producing new, better individuals. First, two individuals are selected. These individuals are randomly changed and subsequently recombined giving rise to two new individuals. This procedure is repeated for a pre-defined number of times.

The presence of interactions is encoded in a binary scheme where '1' represents their presence and '0' their absence. Similarly, for the mode of action '1' represents activation and '0' inhibition (denoted as '+' and '-' in Figure 3.2). Thus, one individual in the genetic algorithm is divided into three sub-arrays, the network corresponding to the first phenotype (green), the network corresponding to the second phenotype (blue) and the inferred mode of action of unsigned interactions (yellow).

Besides having an efficient encoding of the network, the choice of the background algorithm plays a key role for a reasonable performance. For this reason, the jMetal (Durillo and Nebro, 2011) implementation of NSGA-II (Deb *et al.*, 2002), an elitist multiobjective genetic algorithm, has been chosen as it reduces the complexity compared to other implementations (Deb *et al.*, 2002). Moreover, NSGA-II is an elitist algorithm so that the best individuals are kept for the next iteration and are not subject to the three main steps of the algorithm. The fitness function, or score, for determining the best individuals is composed of three distinct parts. First, the manhattan distance of the network attractor and the required discrete phenotype it should represent for reducing the mismatch with the experimental data. Second, the closeness of the reached attractor upon simulation starting from the network state corresponding to the desired discretized phenotype. Finally, the number of removed interactions is minimized, which ensures that only inconsistent interactions are pruned. After scoring, the genetic algorithm iteratively evolves following a three-step approach until a user-defined number of iterations is exceeded.

3.2.2.1 Selection

The first step in each iteration of the genetic algorithm constitutes the selection of two individuals from the entire population. Cooperative selection (Nepomuceno *et al.*, 2007) is a scheme that preserves diversity while ensuring the selection of individuals whose attractors are closest to the phenotype with respect to the previously defined fitness. Briefly, this strategy selects the best individual and updates its score to the average score of the second- and third-best individuals, which implies the selection of a different individual in the next iteration. Importantly, multiple individuals may have the same score so that the selected individual is not necessarily the third network in the queue after the score was updated. Following this approach guarantees the selection of the best individuals in the population while preserving diversity in the next generation.

3.2.2.2 Mutation

The second step constitutes the mutation of the selected individuals. Each entry of the individual arrays is probabilistically altered, i.e. a present interaction is pruned while an absent interaction is re-introduced, or the inferred sign of an unsigned interaction is transposed with a certain probability. Here, the selection of an appropriate mutation rate is crucial for the performance of the genetic algorithm. A low mutation rate leads to faster convergence of the entire population towards a single solution while higher mutation rates lead to a greater exploration of the state space and slower convergence. Low mutation rates bear the risk of identifying local minima instead of the globally optimal solution.

3.2.2.3 Crossover

The last step of the genetic algorithm constitutes the crossover, i.e. a probabilistic recombination, of the two selected individuals. Given a probability p the two individuals are recombined and, consequently, left unchanged with probability *1-p*. Typically, one position in the array is chosen at which the two individuals are split with new individuals created from the first part of the first individual together with the second part of the second individual and vice versa. However, since the arrays in this implementation of the genetic algorithm are composed of three different components, the single-point crossover strategy has been adapted to a three-point crossover that selects a split position in each part of the array. Thus, the strategy for combining the split parts is the same as for the single-point crossover treating the subarrays individually.

3.3 Inference of candidate instructive factors for cellular conversions

Network motifs, such as positive and negative feedback loops, have an important role in maintaining network stability. While positive feedback loops are typically associated with the stabilization of point attractors (Plahte *et al.*, 1995; Snoussi, 1998; Gouzé, 1998; Soulè, 2003; Thomas, 1994), negative feedback is a necessary condition for a network to possess at least one cyclic attractor (Thomas, 1981) and a combination of both is implicated in maintaining a stable network state (Remy and Ruet, 2008). In order to detect these feedback loops, the proposed methodology implements Johnson's algorithm (Johnson, 1975) for identifying all elementary cycles of both reconstructed networks. Elementary cycles can be defined as a path starting and ending in the same network node while all nodes in the path only occur once. Definition 7 formalizes this notion in the context of Boolean Networks.

Definition 7 (Elementary cycle). *Let G = (V,E,F,s) be a Boolean network and* $x, x_1, \dots, x_n \in V$ *nodes in the network. An elementary cycle is either a self-loop, i.e.* $(x, x) \in E$ *or fulfills the condition*

$$(x, x_1) \in E \land (x_n, x) \in E \land \forall i. 1 \leq i \leq n - 1. (x_i, x_{i+1}) \in E$$

where $x, x_1, \dots, x_n$*are pairwise unequal. An elementary cycle is called positive if* $s(x, x_1) \cdot s(x_n, x) \cdot \prod_{i=1}^{n-1} s(x_i, x_{i+1}) = 1$ *and negative otherwise.*

Feedback loops common to both cell type specific networks are determined and all genes they are composed of are assumed to be candidate instructive factors. Further, the approach must take into account the topological differences of the two networks. Therefore, differentially regulated genes, i.e. genes being regulated by different sets of transcription factors, are identified as they are predestined targets that cannot be explained within the same network topology. It is important to note that the identified genes are not necessarily responsible for disease onset or maintenance, in the context of diseases as cellular conversions, but are rather expected to revert the disease phenotype upon perturbation.

A minimal multitarget combination of candidate instructive factors is obtained by randomly sampling from all possible combinations. For keeping the runtime tractable, the maximum number of simulated perturbations is limited to one million per combination size. Here, the size of a combination refers to the number of genes included in the perturbation. Networks are *in silico* perturbed with each combination individually and the propagated effect is simulated until a point-attractor is reached or for a maximum of 1000 steps. The difference between the last state of the simulation and the desired phenotype constitutes the effect of the combination and is penalized by the number of nodes in the network if no point-attractor has been reached. Of note, the perturbation effect is assumed to be dominant, i.e. an up-regulated gene is kept up-regulated regardless of its underlying logic rule. Therefore, larger perturbations are prompt to artificially increase their effect. In order to provide a comparable measure, the absolute perturbation effect is defined by the number of inflicted *in silico* changes to the phenotype subtracted by the number of perturbed genes. Importantly, the absolute perturbation effects can become negative if the number of induced changes on the phenotype is lower than the number of perturbed genes. Furthermore, the normalized perturbation effect is defined as the absolute perturbation effect of a multitarget combination divided by the highest observed absolute perturbation effect of all multitarget combinations, resulting in a scaling between 0 and 1.

3.3.1. Ranking of small compounds for inducing cellular conversions

The identification of potential drugs for inducing a desired cellular conversion utilizes the aforementioned multitarget perturbation effects to provide a qualitative measure of whether important genes are targeted. In summary, the enrichment of reported drug effects contained in simulated multitarget perturbations is computed by weighting the drug effects based on their individually induced change of the phenotype upon perturbation. Those drugs showing more influential effects in high scoring multitarget combinations and low enrichment in low scoring combinations are considered to have greater effects on the overall phenotype.

More formally, first, a weight w_g is assigned to every candidate instructive factors corresponding to the normalized single-gene perturbation effect. Given drug d and a set of reported drug effects D, the enrichment of drug effects in multitarget combination I can be defined as $m_I^d = \sum_{j \in I \wedge j \in D} w_j$, i.e. the sum of weights of genes contained in I having a reported drug effect. In order to obtain values between 0 and 1, these enrichment scores are normalized by the total sum of weights yielding $\widehat{m}_I^d = m_I^d / \sum_{j \in D} w_j$. From the normalized enrichment and perturbation effects for each multitarget combination, the enrichment distribution of drug d over the space of induced phenotypic changes can be computed. Informally, the probability mass function p_d of this distribution can be defined as the fraction of all drug effect enrichments in multitarget combinations having a particular score divided by the sum of all enrichments of all scores. Thus, $p_d(x) = M_d(x) / \sum_{y \in \mathbb{Z}} M_d(y)$ with $M_d(x) = \sum_{I \in \{i \mid gec(i) = x\}} \widehat{m}_I^d$ where *gec(i)* refers to the normalized perturbation effect of multitarget combination *i*. With it, the cumulative enrichment in multitarget combinations inducing less phenotypic changes than a given threshold can be directly derived as $c_d(x) = \sum_{j=0}^{\max(k \in \mathbb{Z} | k \leq x)} p_d(j)$.

The final quality score of a drug, with respect to its predicted potential of inducing desired phenotypic changes, can eventually be determined by considering the area under the cumulative enrichment function $c_d(x)$, defined as $AUC_d = \sum_{x = \min(y | c_d(y) \neq 0)}^{\min(y | c_d(y) = 1) - 1} c_d(x)$. Here, lower values of AUC_d correspond to more favorable drugs while higher values predominantly show enrichment in low scoring multitarget combinations. For example, a drug having an area under the cumulative enrichment curve of zero has reported effects solely in genes contained in the highest scoring multitarget combinations. Similarly, drugs with an AUC of one only have reported effects in genes inducing subtle changes on the phenotype *in silico*.

3.4 Prediction of binary accessibility measures

The identification of accessible and inaccessible genes is an inevitable step towards condition specific networks due to the heterogeneity of the epigenetic landscape in different cell types. State-of-the-art experimental methods, like DNase-seq and ATAC-seq, are based on cutting exposed regions of the DNA that are subsequently sequenced and aligned to the genome. Peaks corresponding to regions enriched in aligned reads can then be identified to pinpoint accessible chromatin regions throughout the genome. However, due to the employed experimental techniques, these peaks are typically not distributed over complete genes but rather identify active regulatory regions (Song *et al.*, 2011). When working with gene regulatory Boolean networks where nodes represent genes, the peaks throughout the genome must be related to genes and translated to gene-level accessibility. Another, more important, impediment constitutes the unavailability of experimental data for many cell types and conditions under study hindering the refinement of reconstructed regulatory networks.

The method presented in the remainder of this section addresses these limitations by defining a consistent gene-level accessibility measure that can be readily used in the context of Boolean networks and provides a tool for predicting gene-level accessibility from gene expression data. For this purpose, a stacked classification tree model was developed that learns the relationship between expression and chromatin accessibility data and can be applied to unseen transcriptomics datasets for predicting the chromatin accessibility of genes.

3.4.1 Dataset for model training

For training the stacked classification tree model, it is necessary to compile a dataset including both gene expression and chromatin accessibility data of a particular organism. The training data used for validating the approach are human RNA-seq and DNase-seq samples from ENCODE (Dunham *et al.*, 2012) aligned to Human Genome Version 19 (hg19). In total 18 samples have been processed including various cell lines and cell types ranging from alveolar basal epithelial cells over skeletal muscle cells up to embryonic stem cells. A complete list of obtained cell lines/types and corresponding Gene Expression Omnibus (GEO) accession numbers can be found in Table 3.1.

Table 3.1. Training dataset for predicting gene-level accessibility

Cell line/Cell type	RNA-Seq (GEO-Accession)	DNase-seq (GEO-Accession)
A549	GSM758564	GSM736580, GSM736506
AG04450	GSM758561	GSM736514, GSM736563
BJ	GSM758562	GSM736518, GSM736596
Monocytes CD14+	GSM984609	GSM1008582
GM12878	GSM758559	GSM736496, GSM736620
H1-hESC	GSM758566	GSM736582
HeLa-S3	GSM765402	GSM736564, GSM736510
HepG2	GSM758575	GSM736637, GSM736639
HMEC	GSM758571	GSM736634, GSM736552
HSMM	GSM758578	GSM736560, GSM736553
HUVEC	GSM758563	GSM736575, GSM736533
IMR90	GSM981249	GSM1008586
K562	GSM765405	GSM736629, GSM736566
MCF7	GSM765388	GSM736581, GSM736588
NHEK	GSM765401	GSM736545, GSM736556
NHLF	GSM765394	GSM736612, GSM736536
SK-N-SH	GSM981253	GSM1008585
SK-N-SH (Retinoid Acid)	GSM765395	GSM736559, GSM736578

DNase-seq sequences aligned to the hg19 genome were obtained from ENCODE (Dunham *et al.*, 2012; ENCODE Datasets) and further processed to identify enriched genomic regions, also known as hypersensitive regions, using the HotSpot algorithm (version 5.1) (John *et al.*, 2011; Sabo *et al.*, 2004) with default parameters imposing a false discovery rate of 1% on all identified sites. Subsequently, these identified regions were annotated with genomic information relating the position with, for example, gene names, introns, exons and gene types, using HOMER (Heinz *et al.*, 2010). A gene is then labeled accessible if the gene-coding region, including the 3' and 5' UTR exons or its promoter region, contains at least one DNase hypersensitive site, and inaccessible otherwise.

RNA-Seq samples were obtained from ENCODE (Dunham *et al.*, 2012) as aligned long PolyA+ sequences. Transcript abundance was estimated using HOMER in a two-step approach. First, a tag directory was created transforming the aligned reads into a data structure that can be interpreted by HOMER using the 'makeTagDirectory' command. Second, transcript abundance was quantified as fragments per kilobase of exon per million reads mapped (Mortazavi *et al.*, 2008) (FPKM) running the ‘analyzeRepeats’ program. Consequently, only reads mapping to exons on either strand of the DNA were counted and different transcripts of the same gene were condensed to the gene level.

3.4.2 Stacked classification tree model

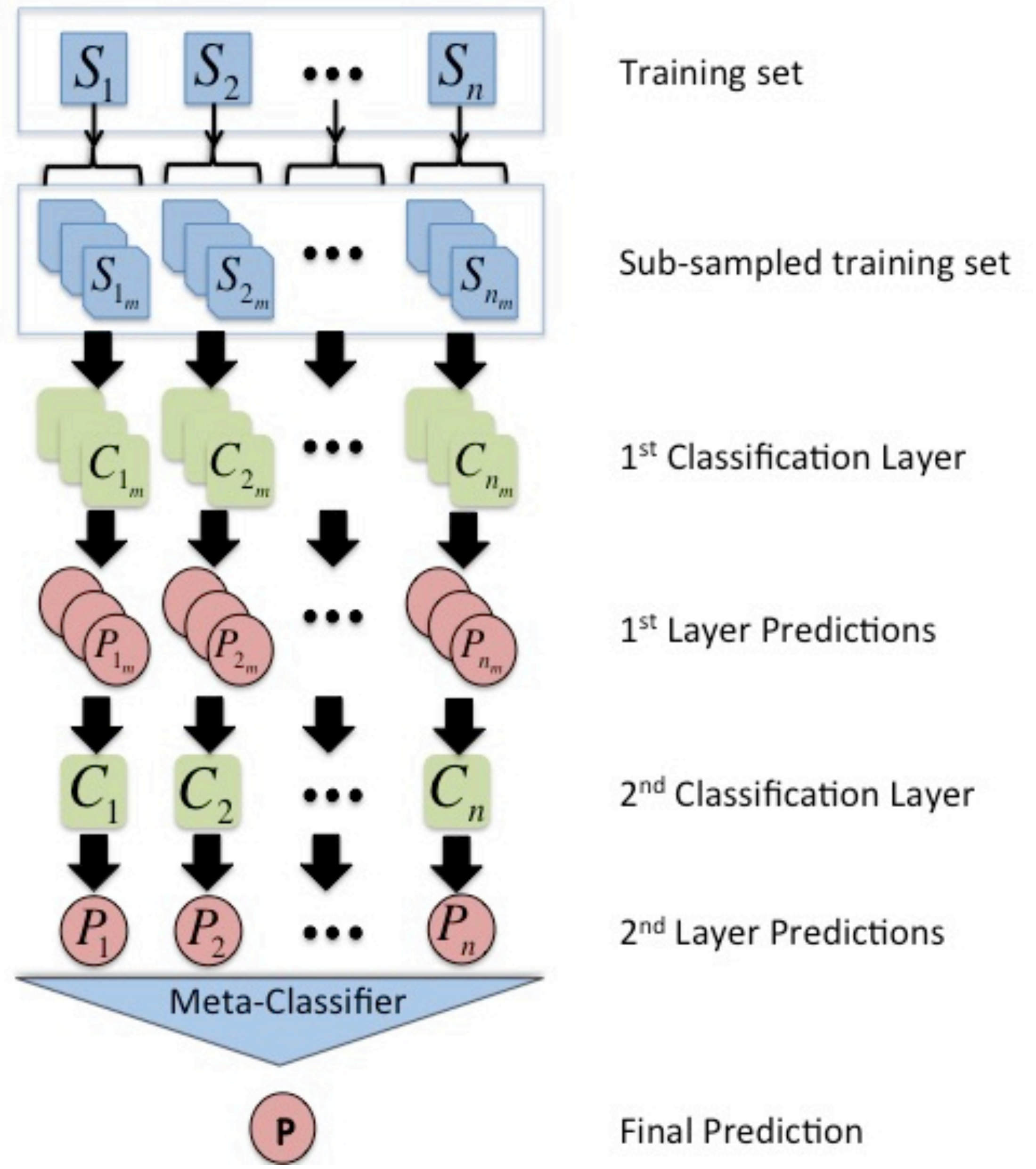
The model for predicting gene-level accessibility contains three classification layers that are stacked, i.e. hierarchically combined, in order to obtain the final predictions. For convenience, we denote the bottom, middle and upper layer classification trees as L_B, L_M and L_U, respectively. While L_B classifiers operate directly on the gene expression/chromatin accessibility training data, upstream classification layers relate the predictions of the lower layers to gene-level accessibility. Figure 3.3 illustrates the workflow of the stacked classification model.

For the first classification layer, each training sample is sub-sampled 1000 times where each sub-sample contains 1000 expression/accessibility pairs. The number of sub-samples drawn for each training sample was empirically estimated to guarantee that each gene is selected in at least one sample and multiplied by four to obtain well mixed training datasets. This step is necessary to ensure that the resulting classification trees do not contain too fine-grained conditions for classifying genes into being accessible or inaccessible and are, thus, less likely to overfit the data.

To address this issue, each sub-sample contains only 1000 expression/accessibility pairs for training to reduce the number of similar, low expression values while preserving the distribution of the whole sample.

Besides the raw gene expression values, another predictor is included in the training set, that is derived from both expression and accessibility. In particular, the empirical expression distributions of accessible and inaccessible genes are computed and the distance of each gene expression value to both distributions is determined by the Mahalanobis distance (Mahalanobis, 1936). For an expression value e the Mahalanobis distance to any empirical distribution X is defined as $d(e) = (e - \mu_X) \cdot S_X^{-1} \cdot (e - \mu_X)'$. Here, μ_X represents the mean of the empirical distribution X and S_X^{-1} is the inverse of the sample covariance matrix of X.

Figure 3.3. Illustration of stacked classification tree model



The stacked learning framework for predicting gene-level accessibility treats each sample in the training set separately. First, the sample is sub-sampled 1000 times for constructing a classification tree. The predictions of each tree for the whole sample are then used as an input for constructing a single classifier for the whole sample. The predictions of whole-sample classifiers are then combined in a meta-classifier to obtain the final predictions.

After sub-sampling the initial samples and training a classifier for each of the sub-sampled datasets, which constitute the first classification layer (L_B), the L_M classifiers are trained by relating the predicted probabilities for a gene to be accessible or inaccessible to the experimentally determined gene-level accessibility. As a result, L_M classifiers return the probability that the probabilities predicted by L_B trees are correct.

As depicted in Figure 3.3, the second classification layer contains as many trees as the number of samples in the original training set. The algorithm employed for training the second classification layer, RUSBoost (Seiffert *et al.*, 2010), addresses an important, inherent issue with the training samples. Based on the experimental data, there is an uneven ratio of accessible and inaccessible genes, which typically leads to a prediction bias. RUSBoost overcomes this limitation by randomly removing training data from the majority class for each of constructed classification tree.

As a final step, the predictions based on the individual samples are combined using a single meta-classifier L_U. In accordance with the first and second classification layer, a classification tree is constructed by employing the RUSBoost algorithm (Seiffert *et al.*, 2010), but this time on the predictions of the second layer (L_M) predictions.

3.5 Reference-based binarization of gene expression

The use of Boolean networks as models of gene regulation requires the classification of continuous gene expression measurements from RNA-seq or microarray experiments into discrete values reflecting whether the gene is active or inactive. For enabling the consistent comparison of more than two conditions a novel approach has been developed that utilizes a user-defined gene expression library serving as a reference for classifying active and inactive genes. Definition 8 provides a formal introduction of a reference library that is considered in this method.

Definition 8 (Reference library). *Let $\vec{g}$ be an expression profile of a single gene throughout different cell types/lines and conditions. Then*

$$G = \{gene : \vec{g} = (g_1, g_2, \dots, g_n) \mid g_1 = S_1(gene) \wedge g_2 = S_2(gene) \wedge \dots \wedge g_n = S_n(gene)\}$$

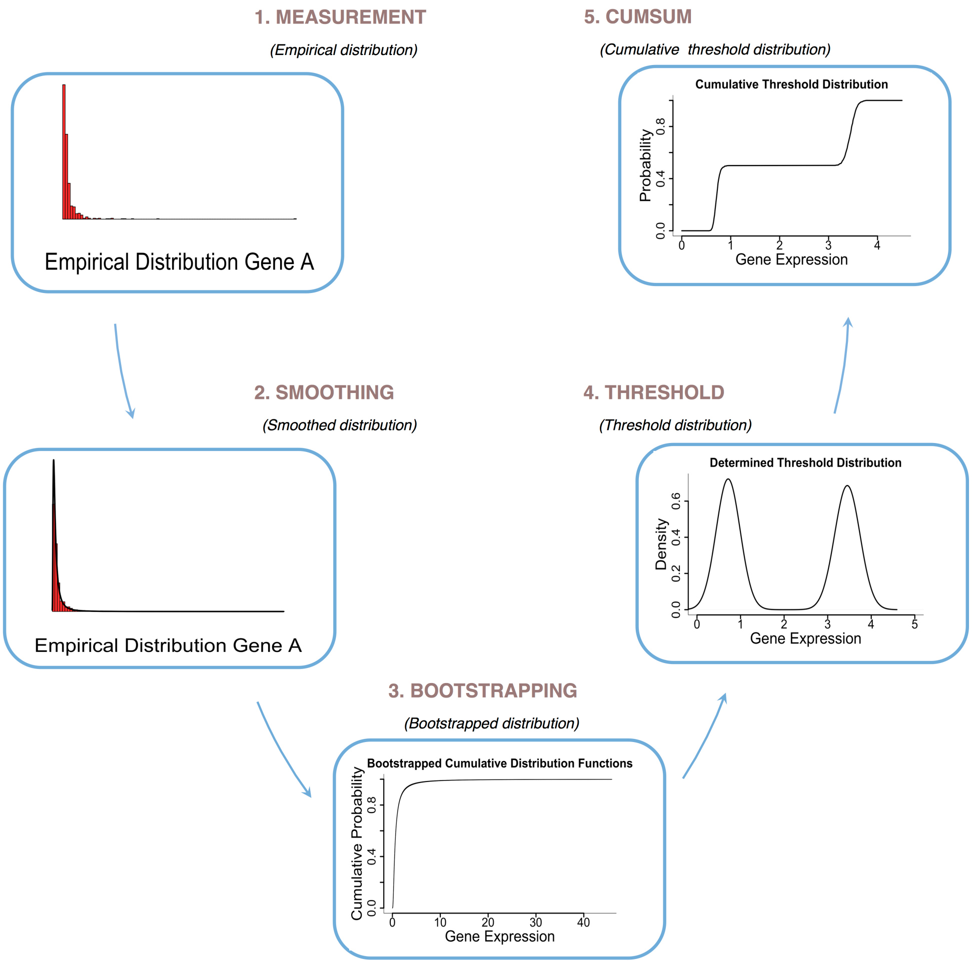
is a reference library where $S_1, \dots, S_n$ representing gene expression profiles of different cell types/lines or conditions.

The gene expression profile S_i in the definition of the reference libraries is a function returning the expression value of a gene for phenotype *i*. For convenience, each gene $\vec{g}$ is labeled with its corresponding gene name. Of note, typically all genes in the reference have the same number of expression values, i.e. all vectors are of the same length, since they were derived from the same transcriptomics datasets. Thus, G can be regarded as a matrix with rows corresponding to genes and columns to different expression samples.

The distribution given by the gene expression profiles contained in the reference library is assumed to be the same as the real, unknown distribution that would be obtained by performing gene expression experiments for every possible cell type. Given this assumption, the method performs three main steps to discretize gene expression values (Figure 3.4). (1) Due to the undeniable incompleteness of the reference, the empirical distribution for each gene in G is smoothed by selecting the best fitting parametric distribution to represent it. (see Figure3.4 part 1 and 2) (2) This process is repeated for a pre-defined number of bootstrap-samples in order to obtain a distribution of parameter values for the parametric distribution. (see Figure 3.4 part 3) (3) To identify thresholds for calling a gene active or inactive, the method approximates each bootstrapped distribution by two step-functions. The first step function minimizes the number of false positive assignments, i.e. classifying a gene as being active if it is inactive, while the second function minimizes the number of false negative assignments, i.e. classifying a gene inactive while it is actually active. The gene expression value at the jump discontinuity of each step function is then defined as the threshold for discretization. This procedure is conducted on all the bootstrapped distributions and gives rise to the threshold distribution used for classifying the individual genes. (see Figure 3.4 part 4 and 5)

The remainder of this section describes the main steps of the method and how statistical significance measures are assigned to query expression values. It concludes with a statistical assessment of how many expression profiles from different cell types are needed for approximating the real, unobservable distribution closely.

Figure 3.4. Workflow for discretizing gene expression data



The workflow for discretizing gene expression requires the computation of thresholds obtained within five steps. The gene expression distribution is approximated by a parameterized distribution family. After the best fitting distribution family has been determined, parameters are obtained for several bootstrap samples of the gene expression and lower and upper thresholds are computed for each fitted distribution. The resulting cumulative threshold distribution provides the statistical significance for a query expression value to be active or inactive.

3.5.1 Smoothing of empirical gene expression distributions

The chosen reference library G is usually not covering all cell types of an organism leading to an inherent uncertainty in the empirical gene expression distribution. Particularly those expression ranges that have not been profiled, appearing as no increase in the empirical cumulative distribution function (Definition 10), could be explained by two different assumptions.

Definition 9 (Indicator function). *Let X be a set and $A \subseteq X$. The indicator function $I_A: X \rightarrow \{0,1\}$ is defined as*

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A \end{cases}$$

Definition 10 (Empirical cumulative distribution function). *Let $\vec{g} \in G$ be a gene expression sample for a gene and I be the indicator function. Then*

$$ecdf_{\vec{g}}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{g_i \leq x}(g_i)$$

is the empirical distribution function of $\vec{g}$.

Either there exists any cell type or condition in which the gene under consideration takes on these expression values but G does not contain it or these expression values do not occur at all. Based on previous studies defining gene expression as a stochastic process (Blake *et al.*, 2003; Eldar and Elowitz, 2010; Elowitz, 2002; Syeed *et al.*, 2010; Kærn *et al.*, 2005; McCullagh *et al.*, 2009; Paulsson, 2005; Raj and van Oudenaarden, 2008), the basic assumption of the method is that these values have not been sampled and, thus, the distribution for these expression ranges should be approximated to account for this insufficient sampling. Therefore, the empirical distribution is replaced by the best fitting parametric distribution (Table 3.2) with respect to a maximum likelihood estimator (Wilks, 1938). In the remainder, we denote the best fitting probability distribution of gene g parameterized by $\vec{p}$ as $D_g(x|\vec{p})$.

Definition 11 (Negative Log-Likelihood). *The likelihood of a set of parameters $\vec{p}$ given (an expression value) x is the probability of observing x given $\vec{p}$, i.e.*

$$\mathcal{L}(\vec{p}|x) = P(x|\vec{p})$$

with $P(x|\vec{p}) = D_g(x|\vec{p})\frac{d}{dx}$. The negative log-likelihood is then defined as

$$-\log\left(\mathcal{L}(\vec{p}|x)\right)$$

The typical goodness-of-fit measure for assessing the suitability of a fitted distribution is the negative log-likelihood (Definition 11). However, this measure does not account for the different number of parameters of different distribution types. For example, the exponential distribution family constitutes a special case of generalized pareto distributions with shape and location parameters equal to zero. Therefore, a generalized pareto distribution never provides a worse fit because the other two parameters can be

Table 3.2. Candidate distributions fitted to empirical distributions

Distribution	1st parameter	2nd parameter	3rd parameter
Beta	1st shape – $a > 0$	2nd shape – $b > 0$	–
Birnbaum-Saunders	Scale – $\beta > 0$	Shape – $\gamma > 0$	–
Burr	Scale – $\alpha > 0$	1st shape – $c > 0$	2nd shape –$k > 0$
Exponential	Mean – $\mu > 0$	–	–
Extreme Value	Location – $-\infty < \mu < \infty$	Scale – $\sigma \geq 0$	–
Gamma	Shape – $a > 0$	Scale – $b \geq 0$	–
Generalized Extreme Value	Shape – $-\infty < k < \infty$	Scale – $\sigma \geq 0$	Location – $-\infty < \mu < \infty$
Generalized Pareto	Shape – $-\infty < k < \infty$	Scale – $\sigma \geq 0$	Location – $-\infty < \mu < \infty$
Inverse Gaussian	Scale – $\mu > 0$	Shape – $\lambda > 0$	–
Logistic	Mean – $-\infty < \mu < \infty$	Scale – $\sigma \geq 0$	–
Log-Logistic	Log mean – $\mu > 0$	Log scale – $\sigma > 0$	–
Lognormal	Log mean – $-\infty < \mu < \infty$	Log stdev – $\sigma \geq 0$	–
Nakagami	Shape – $\mu > 0$	Scale – $\omega > 0$	–
Normal	Mean – $-\infty < \mu < \infty$	Stdev – $\sigma \geq 0$	–
Rayleigh	Defining parameter – $b > 0$	–	–
Rician	Noncentrality – $s \geq 0$	Scale – $\sigma > 0$	–
T Locations-Scale	Location – $-\infty < \mu < \infty$	Scale – $\sigma > 0$	Shape – $\nu > 0$
Weibull	Scale – $a > 0$	Shape – $b > 0$	–
Gaussian Mixture Model Means – see Normal Distribution	Stdevs – see Normal Distribution	Weights	

The empirical distribution is fitted to the 19 different distribution types in this table. Each distribution is defined by at most three parameters with its corresponding parameter ranges shown. The exponential distribution constitutes a special case of generalized pareto distributions.

adjusted accordingly, even though the uncertainty in the parameter estimation increases. More generally, distributions with more parameters are more likely to fit the empirical distribution under negative log-likelihood scoring, as they possess more degrees of freedom for making them compatible with the observations. The approach taken in this thesis instead uses the Akaike Information Criterion (Akaike, 1974) (AIC) that incorporates the number of parameters in the calculation of the goodness-of-fit calculation (Definition 12). The distribution family of the distribution with the lowest AIC is then selected for further processing.

Definition 12 (Akaike Information Criterion). *Given a parameterization $\vec{p}$ that minimizes the negative log-likelihood function $-\log(\mathcal{L}(\vec{p}|x))$ and k the number of parameters, the Akaike Information Criterion is defined as*

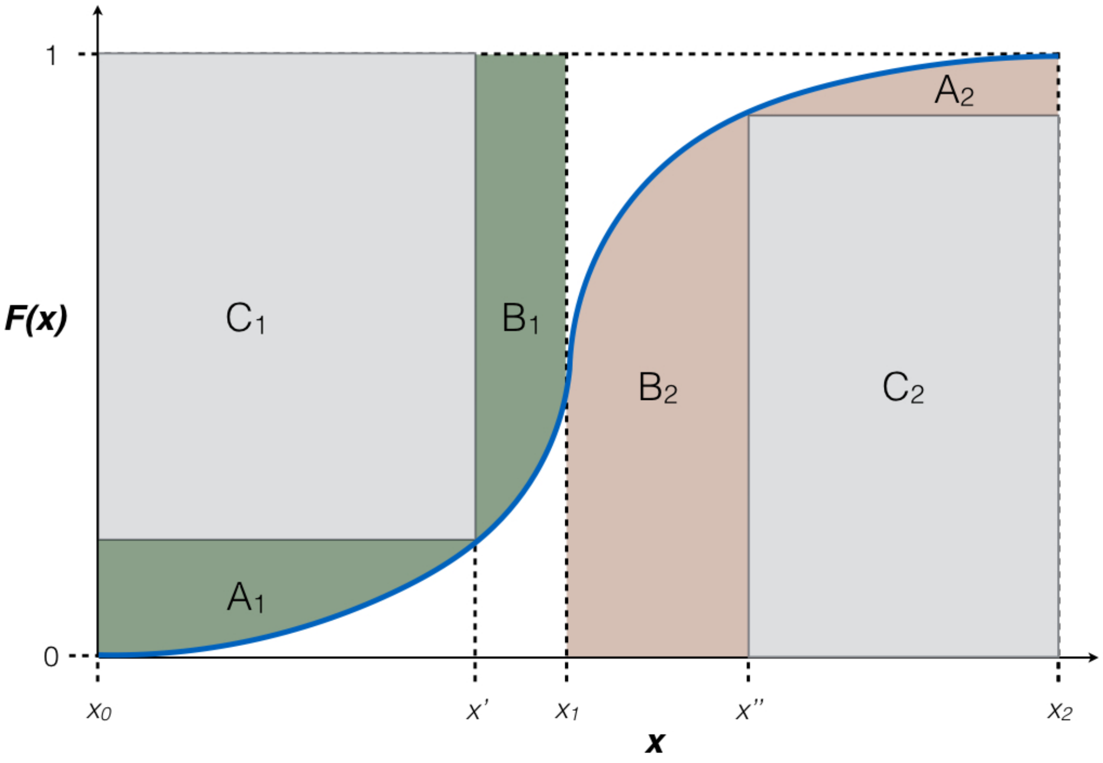
$$AIC = 2(k - \log(\mathcal{L}(\vec{p}|x)))$$

3.2.2 Parameter Distribution Approximation

After the best fitting distribution family is derived for each gene, the empirical parameter distributions are computed by bootstrapping the gene expression samples $\vec{g} \in G$. This procedure allows the estimation of the parameter variability that intuitively corresponds to how well the previously identified distribution family is suited to represent the given expression sample. In particular, the lower the individual parameter variance the more defined the real distribution is. Besides these theoretical considerations, this approach enables a statistical significance analysis of the discretized values after applying all subsequent steps.

Bootstrapping essentially is a resampling technique in which elements of a sample are randomly selected with replacement. With respect to a gene expression sample $\vec{g} = (g_1, g_2, \dots, g_n) \in G$ the set of bootstrap samples can be defined as $B_g = \{\overrightarrow{g_b} \mid \overrightarrow{g_b} = (g_{1_b}, g_{2_b}, \dots, g_{n_b})\}$ where $|B_g|$ constitutes the number of samples drawn. All analysis in this thesis are carried out using 1000 bootstrap samples, which is sufficient for approximating the parameter distribution. For each of the bootstrap samples $\overrightarrow{g_b} \in B_g$ a parameterization of the previously identified distribution family $\overrightarrow{p_{g_b}}$ is obtained. The parameter distribution sample is then defined as $P_{B_g} = \{\overrightarrow{p_{g_b}} \mid \overrightarrow{g_b} \in B_g\}$.

Figure 3.5. Idealized representation of threshold computation



Lower and upper thresholds are computed by minimizing the areas A and B, which is equivalent to maximizing the areas C. These thresholds constitute the optimal tradeoffs between false positive and false negative assignments. In case of lower thresholds minimizing B leads to a minimization of false negatives while a minimization of A renders minimal false positives.

3.2.3 Derivation of Thresholds

The parameter distribution samples serve as the basis for the derivation of thresholds for active and inactive gene expression by approximating each distribution from the sample with two step-functions. The first step function minimizes the error with the left tail of the distribution by computing an optimal trade-off between minimizing false negatives and maximizing the expression value at the jump discontinuity (see x' in Figure 3.5). In contrast, the second step function finds the optimal trade-off between minimizing the expression value at the jump discontinuity and minimizing the number of false positives (see x'' in Figure 3.5). It is worth noting that the terms false positive/negative do not refer to a measureable quantity due to the absence of a gold standard dataset. Instead, it is based on the rationale that increasing/decreasing the threshold for a gene to be active/inactive increases the number of false positive/negative assignments.

Computing the lower and upper thresholds x' and x'' (Figure 3.5) described before corresponds to the minimization of the area of $A_1 + B_1$ and $A_2 + B_2$, respectively. However, this problem formulation can be simplified to the maximization of the areas C_1 and C_2. Using the notation in Figure 3.5 and defining $a_i(x)$, $b_i(x)$ and $c_i(x)$ as the areas of A_i, B_i and C_i the equivalency between both problem formulations can be formalized in Lemma 1 and Lemma 2 (Jung *et al.*, 2017).

Lemma 1. *The minimization of* $A_1 + B_1$ *is equivalent with the maximization of* C_1*. In particular*

$$x' = \underset{x}{\operatorname{argmin}}\, a_1(x) + b_1(x) = \underset{x}{\operatorname{argmax}}\, c_1(x)$$

Proof. *The area functions* $a_1(x)$ and $b_1(x)$ *can be expressed in terms of integrals as*

$$a_1(x) = x - x_0 - \int_{x_0}^{x} F(y)\,dy - c_1(x)$$

$$b_1(x) = x_2 - x - \int_{x}^{x_1} F(y)\,dy$$

Applying the interval addition property of definite integrals, the sum of the two functions can be written as

$$a_1(x) + b_1(x) = x_2 - x_0 - \int_{x_0}^{x_1} F(y)\,dy - c_1(x)$$

in which all terms are constant except $c_1(x)$*. It follows that the maximum of* $c_1(x)$ *constitutes the minimum of* $a_1(x) + a_2(x)$.

Lemma 2. *The minimization of* $A_2 + B_2$ *is equivalent with the maximization of* C_2*. In particular*

$$x'' = \underset{x}{\operatorname{argmin}}\, a_2(x) + b_2(x) = \underset{x}{\operatorname{argmax}}\, c_2(x)$$

Proof. *Like in the proof of Lemma 1, the area functions* $a_2(x)$ and $b_2(x)$ *can be expressed in terms of integrals as*

$$a_2(x) = \int_{x}^{x_2} F(y)\,dy - c_2(x)$$

$$b_2(x) = \int_{x_1}^{x} F(y)\,dy$$

Applying the interval addition property of definite integrals to the sum of $a_2(x)$ and $b_2(x)$ *yields*

$$a_2(x) + b_2(x) = \int_{x_1}^{x_2} F(y)\,dy - c_2(x)$$

which resolves to the area of C_2 subtracted from a constant. Thus, $a_2(x) + b_2(x)$ takes on its minimal value when $c_2(x)$ is maximal.

While Lemma 1 and 2 proof the equivalency of the problem formulations, they omit a formal definition. Given a parameterization $\overrightarrow{p_{g_b}} \in P_{B_g}$ and the cumulative distribution function, $cdf\left(D_g\left(x\middle|\overrightarrow{p_{g_b}}\right)\right)$, of the parameterized distribution, the optimal jump discontinuities x' and x'' serving as lower and upper thresholds for a gene being active or inactive, are selected by solving the following expressions:

$$x' = \underset{0 \le x \le \max(\vec{g})}{\operatorname{argmax}} \underbrace{\left(x \cdot \left(1 - cdf\left(D_g\left(x\middle|\overrightarrow{p_{g_b}}\right)\right)\right)\right)}_{C_1}$$

$$x'' = \underset{0 \le x \le \max(\vec{g})}{\operatorname{argmax}} \underbrace{\left(cdf\left(D_g\left(x\middle|\overrightarrow{p_{g_b}}\right)\right) \cdot (\max(\vec{g}) - x)\right)}_{C_2}$$

It is important to note that both expressions are varying in the presence of outliers because the maximum is constrained by the maximum observed expression value in $\vec{g}$. Consequently, the method prioritizes the minimization of false assignments over the minimization of the expression value at the jump discontinuity. Especially the upper threshold x'' is constant or decreases upon removal of outliers. The results in this thesis are based on the complete reference library without removing outliers and can, thus, be considered to be more conservative thresholds.

Two of the distribution families described in Table 3.2, generalized Pareto distributions and Gaussian Mixture Models, require distinct treatments that will be discussed in the following.

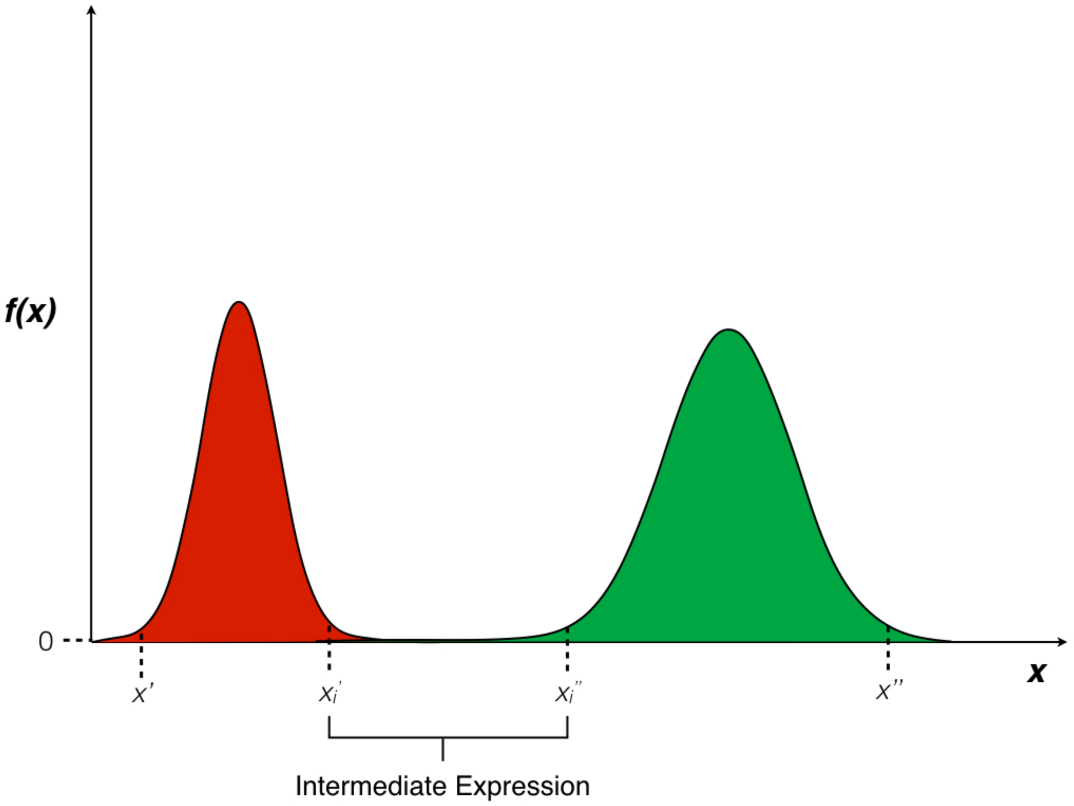
The unimodality of distributions typically guarantees that the maxima of C_1 and C_2 correspond to extreme values of the distributions (proof omitted here). However, this assertion is not valid for generalized Pareto distributions including the exponential distribution. Due to the mode being located at the boundary of the expression range, i.e. it is located at zero expression, the local maximum of C_1 might be located at negative expression values. Instead of taking the negative expression value as a threshold, the method sets it to zero since a gene not undergoing active transcription cannot be translated into proteins and, thus, must be inactive.

Gaussian Mixture Models differ from all other distribution families in that they possess multiple modes requiring adapting the strategy for lower and upper thresholds. In particular, a Gaussian Mixture Model is defined as the weighted sum of normal distributions possibly possessing different means and standard deviations. In accordance with the approach for unimodal distributions, multimodal distributions are decomposed into their individual components of which lower and upper thresholds are derived.

When solving the expressions of x' and x'' for all parameterizations obtained from the bootstrap samples, the empirical threshold distribution can be defined by the sets of lower and upper thresholds as $T_g^l = \left\{x' | x' = \underset{0 \le x \le \max(\vec{g})}{\operatorname{argmax}} \left(x \cdot \left(1 - cdf\left(D_g\left(x|\overrightarrow{p_{g_b}}\right)\right)\right)\right) \wedge \overrightarrow{p_{g_b}} \in P_{B_g}\right\}$ and $T_g^u = \left\{x'' | x'' = \underset{0 \le x \le \max(\vec{g})}{\operatorname{argmax}} \left(cdf\left(D_g\left(x|\overrightarrow{p_{g_b}}\right)\right) \cdot (\max(\vec{g}) - x)\right) \wedge \overrightarrow{p_{g_b}} \in P_{B_g}\right\}$, respectively, and are used to quantify the statistical significance of the classification.

3.5.4. Statistical Significance Computation

Figure 3.6. Threshold distribution used for classification



The derived threshold distributions define three discrete states. A gene is (1) inactive if its expression is below x', (2) active if its expression is greater than x'' and (3) intermediately expressed if its expression is between x_i' and x_i''. Expression values close to the mean of the lower (red) and upper thresholds (green) cannot be reliably determined. The pre-defined significance values determine the location of these thresholds.

The derivation of lower and upper thresholds leads to a threshold distribution consisting of three parts: (i) the lower threshold mode with its associated distribution, (ii) the upper threshold distribution mode with its associated distribution and (iii) the area between these distributions containing a low proportion of the probability mass. Figure 3.6 shows a threshold distribution illustrating the aforementioned three parts.

The lower threshold distribution (red) is defined as the empirical cumulative distribution function of T_g^l, $ecdf_{T_g^l}(x)$, and is utilized to define the statistical significance of assigning a gene g with gene expression $y_g$to be inactive. In particular, a p-value for gene g being inactive is defined as

$$p_{y_g}^i = ecdf_{T_g^l}(y_g)$$

and can be interpreted as a one-sided test against the null hypothesis that g is not inactive. Likewise, the empirical cumulative distribution function of upper threshold distribution (green) T_g^u, $ecdf_{T_g^u}(x)$, defines a p-value for gene g being active as

$$p_{y_g}^u = 1 - ecdf_{T_g^u}(y_g)$$

corresponding to a one-sided test against the null hypothesis that g is not active.

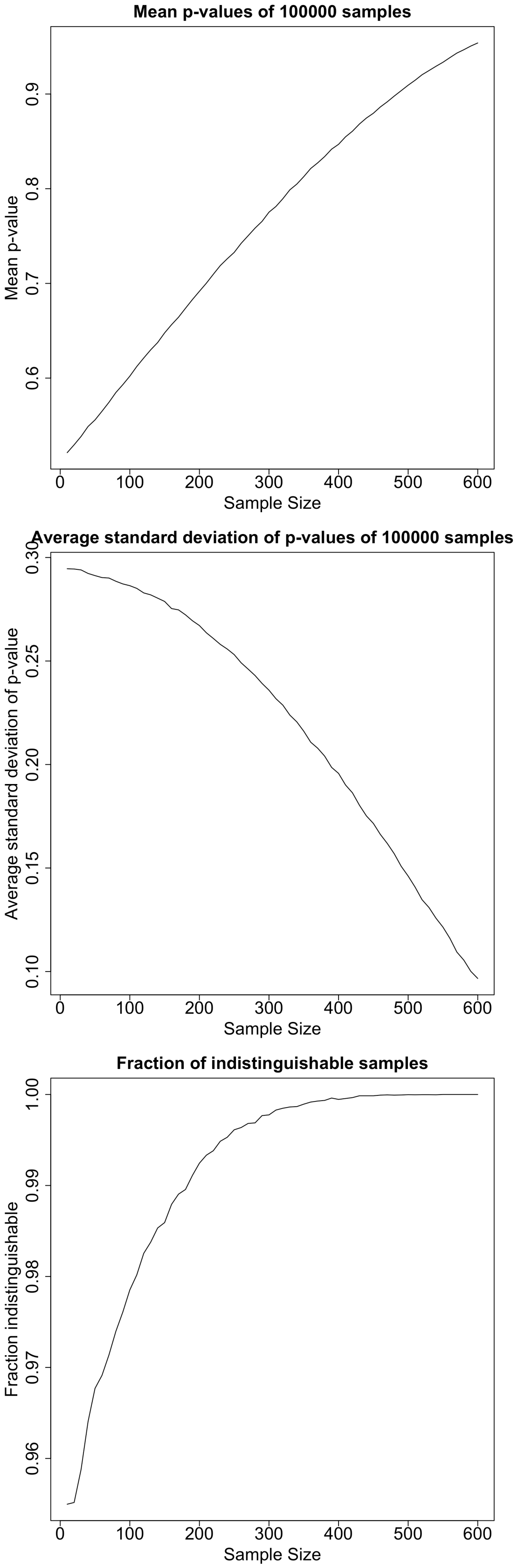
According to the p-value definitions, significantly active and inactive genes are solely defined based on the tails of the distribution. However, the intermediate expression range, or more precisely the area between both distributions only containing a small amount of the probability mass, should be treated differently than the area close to means of the distributions. While the expression range close to the means of the lower and upper threshold distributions corresponds to the uncertainty in the assignment, the intermediate expression range is defined as having certainty that the gene is neither active nor inactive. Thus, it requires its own classification based on the empirical cumulative distribution functions of T_g^u and T_g^l. Given a significance cutoff α, for any expression value y_g satisfying $1 - ecdf_{T_g^l}(y_g) < \alpha$ and $ecdf_{T_g^u}(y_g) < \alpha$ g is considered to be intermediately expressed. Notably, this definition allows the intermediate expression range to be empty if the two distributions are overlapping or the significance cutoff is too stringent.

3.5.5. Statistical assessment of the reference-library size

The choice of the reference distribution is crucial for the successful application of the previously described methodology for classifying active and inactive genes. Therefore, two questions have to be addressed. First, how many samples are sufficient and, second, can they be randomly sampled? For this purpose a collection of 27887 microarray samples (Torrente *et al.*, 2016) was collected and analyzed with respect to a sufficient sample size that is statistically indistinguishable from the whole set of samples. In particular, the set of samples was randomly subsampled to sizes between 10 to 600 and compared to the whole set of samples by means of a two-sample Kolmogorov-Smirnoff test that assesses the statistical equivalence of the distributions. High p-values mean that the distributions are indistinguishable while low p-values reject this hypothesis at the obtained significance level. This step was conducted 100,000 times for each sub-sample size to obtain conclusive and statistically sound results.

As expected, the average p-values increase with increasing sample size almost linearly while the average standard deviation declines exponentially (Figure 3.7 upper and middle panel). Complementary to this

Figure 3.7. Minimal reference distribution sizes



The minimal amount of samples in the reference distribution was assessed by subsampling 27887 microarray samples and comparing the subsampled to the complete reference using a two-sample Kolmogorov-Smirnoff test.
With increasing subsampling size, p-values increase (upper panel) while their standard deviation decreases (middle panel) suggesting the homogenization of the distribution. High p-values correspond to more indistinguishable samples (lower panel).

assessment, the fraction of samples that are indistinguishable from the complete distribution was calculated, given a 5% significance level (Figure 3.7 lower panel). While already 95 percent of the random samples are indistinguishable for a sample size of 10, the mean p-values of 0.5 are indicative of a high false discovery rate. However, randomly choosing 550 microarray samples guarantees that all the samples are indistinguishable with mean p-values of 0.93 making it a reasonable sample size for selecting the reference. The low and further decreasing average standard deviation of the p-values, on the other hand, suggests that the choice of the samples is less important than the sample size, since the result is based on 100,000 randomly chosen subsets.

CHAPTER 4 Results

This chapter details the results of the developed methodologies for reconstructing gene regulatory networks from transcriptomics data and prior knowledge networks followed by the integration of gene-level accessibility for further refinement. Validation of the improved context-specificity of the reconstructed networks compared to previously developed methods is analyzed by quantifying the amount of cell type specific protein-DNA interactions. A differential network analysis approach is presented for identifying cellular stability determinants and candidate combinations of instructive factors that was applied to the prioritization of drugs for inducing desired cellular transitions. Subsequently, the method for predicting gene-level accessibility is validated and integrated into the computational pipeline for reconstructing condition specific gene regulatory networks. With it, a combination of instructive factors was predicted for inducing the cellular conversion of adipocytes into osteoblasts. Finally, the developed methodology for discretizing transcriptomics data is validated and its ability to compare multiple cell types is highlighted. All presented approaches build on the work presented in the Methods chapter.

The 'Differential Network Reconstruction for Establishing Disease-Gene-Drug Relationships' section (4.1) details the reconstruction and validation of condition specific networks from prior knowledge networks and gene expression data. Further, strategies for identifying candidate instructive factors of cellular conversions are shown and their utilization in prioritizing drugs is illustrated. Drug prioritizations are validated with known drug screening examples and predictions are made for two autoimmune diseases. Two major obstacles in reconstructing gene regulatory networks are addressed in the subsequent sections. First, the 'Prediction of Chromatin Accessibility in Gene-Regulatory Regions from Transcriptomics Data' section (4.2) presents a machine-learning based methodology for identifying genes in accessible and inaccessible chromatin domains in the absence of epigenetic datasets. With it, an integrated approach for identifying instructive factors of the cellular conversion from adipocytes into osteoblasts is presented that utilizes cell type specific transcriptomics data, predicted gene-level accessibility and prior knowledge networks (4.3). At last, in the 'Reference-based Discretization of Gene Expression Data' section (4.4), the discretization of gene expression data is addressed allowing the comparison of multiple cell types and conditions. A novel, more accurate approach is presented and validated against current state-of-the-art methods.

4.1 Differential Network Reconstruction for Establishing Disease-Gene-Drug Relationships

A method for reconstructing condition specific gene regulatory networks was developed that addresses the limitations of current methodologies (see Chapter 1 for details). In particular, the developed method infers a Boolean gene regulatory network that requires only a prior knowledge network (PKN) and a discretized gene expression profile as input (see in Methods section 3.2). Its utility is demonstrated by proposing a selection scheme for candidate instructive factors given two condition specific GRNs and subsequent prioritization of drugs targeting these candidate factors. The analysis of the network inference method consists of four parts: the validation with condition specific ChIP-seq data, the comparison of the reconstructed network against other methods, the recovery of compounds from known drug screening experiments and the proposal of new compounds for treating Systemic Lupus Erythematosus and Rheumatoid Arthritis, two autoimmune diseases.

4.1.1 Network Reconstruction and Validation

The developed method for reconstructing condition specific gene regulatory networks relies on the discretized transcriptomics profile of cell or tissue types and a prior knowledge network containing directed interactions among genes. Here, the PKN is compiled from MetaCore™ from Thomson Reuters, a proprietary database of manually curated and experimentally validated gene-gene interactions observed in various cell lines, cell types or tissues. The selected interactions are constrained to be implicated in the direct regulation of the target gene for explicitly excluding indirect effects. To obtain reconstructed gene regulatory networks, the method prunes the PKN such that it is consistent with the observed, discretized phenotype, i.e. it is a point-attractor in the Boolean network (see Methods section 3.2).

For validating the network reconstruction approach, two main analyses were conducted showing that the vast majority of cell type specific interactions are retained after pruning and that the reconstruction is more accurate compared to other, similar methods. A gold standard dataset of condition specific interactions and phenotypes was obtained for four highly studied cell lines included in the Encyclopedia Of DNA Elements (ENCODE) (Dunham *et al.*, 2012). In order to increase the diversity of datasets included in the analysis, the selected cell lines are composed of embryonic stem cells (H1-hESC), B-lymphocytes (GM12878), leukemic erythromyeloblastoids (K562) and hepatocellular carcinoma cells (HepG2). A phenotypic representation of each cell line was gathered

Table 4.1. ChIP-seq validated interactions of condition specific network reconstruction algorithm

	HepG2/ GM		HepG2/ H1		HepG2/ K562		GM/ H1		GM/ K562		H1/ K562	
	HepG2	GM	HepG2	H1	HepG2	K562	GM	H1	GM	K562	H1	K562
Raw	92	36	122	20	74	0	2	0	88	18	4	4
Pruned	86	30	111	13	71	0	2	0	79	13	4	3
Re-tained (in %)	93.5	83.3	91	65	95.9	-	100	-	89.8	72.2	100	75

Analysis of retained ChIP-seq validated interactions after differential network reconstruction. Raw values correspond to the number of interactions in the prior knowledge while pruned values denote the number of retained interactions after network reconstruction. Pairwise analysis of four cell lines shows average retention of 89.6%.

from quantified gene expression data of Duke Affy Exon experiments also contained in ENCODE (Table S1). All cell lines were pairwise combined, resulting in six different examples, for obtaining discretized phenotypes by differential expression analysis with strict cutoffs (p-value < 0.001 and fold change > 4). Cell line specific interactions where obtained associating genomic regions enriched in aligned reads of chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments to the promoter regions of genes. Up to 122 identified interactions are contained in the prior knowledge networks of the different examples that serve as validation of context specificity. Of note, due to the differential expression analysis, the number of validated interactions does not only depend on the cell line the experiment had been conducted in but also on the cell line it is compared to.

After network reconstruction the analysis revealed that, on average, 89.6% of ChIP-seq validated interactions are preserved, which demonstrates that the algorithm can reconstruct fairly reliable networks (Table 4.1). In addition, the reconstructed networks show highly variable characteristics of common and phenotype-specific interactions, in the six examples (Table 4.2). On average, 18.8% of the interactions are phenotype specific underlining that a unique GRN topology is not sufficient to accurately model both phenotypes.

Despite the ability of preserving validated interactions after network reconstruction, it is essential to demonstrate the increased accuracy of the method compared to other, state-of-the-art approaches. In order to provide a fair comparison, two methods, CellNOptR (Terfve *et al.*, 2012) and SignetTrainer (Melas *et al.*, 2013), were selected that aim at the reconstruction of directed, signed networks from PKNs, as

Table 4.2. Regulatory interaction statistics of reconstructed networks

	HepG2/ GM	HepG2/ H1	HepG2/ K562	GM/ H1	GM/ K562	H1/ K562
Common	424	608	430	298	662	594
Phenotype 1 (total)	530	824	520	324	999	859
Phenotype 1 (specific)	106	216	90	26	337	265
Phenotype 1 (ratio)	20.00%	26.20%	17.30%	8.00%	33.70%	30.80%
Phenotype 2 (total)	508	707	498	323	819	720
Phenotype 2 (specific)	84	99	68	25	157	126
Phenotype 2 (ratio)	16.50%	14.00%	13.70%	8.30%	19.20%	17.50%

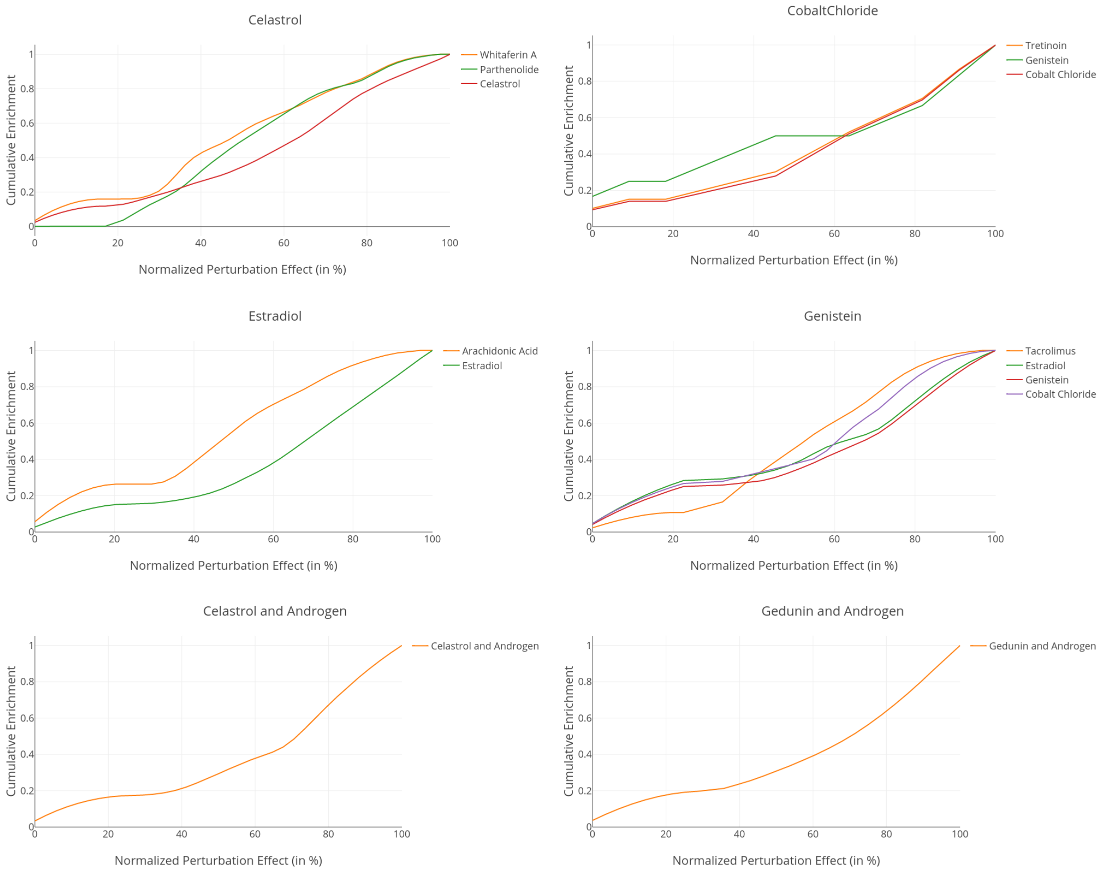
Differential interaction statistics of the reconstructed networks in six examples. While most of the interactions are common to both networks, around 18.8% of interactions are, on average, specific to the individual networks. Comparison of cancer cell lines (K562 and HepG2) to normal cell lines (GM12878 and H1) typically yields higher ratio of network specific interactions than the comparison of normal cells or cancer cells among themselves.

well. In the comparison, the number of preserved validated ChIP-seq interactions and the agreement of the booleanized phenotypes was assessed using the same cell lines as for the network validation. However, both CellNOptR and SignetTrainer were unable to handle prior knowledge networks including more than 300 interactions. The six examples were therefore subsampled to networks encompassing between 11 and 78 genes and 27 to 164 interactions. Additionally, a manually curated, self-consistent core network of MEP (myeloid erythroid progenitor) cell fate commitment was included that does not require pruning (Doré and Crispino, 2011). The results of this assessment confirm that the proposed methodology for reconstructing condition specific gene regulatory networks resembles the given phenotypes more accurately. On average, 94% of genes in the network showed consistent expression levels, with respect to the discretized gene expression profile, compared to 48% and 88% obtained by CellNOptR and SignetTrainer, respectively. Furthermore, the preservation of ChIP-Seq interactions is significantly elevated (94%) in comparison with CellNOptR (58%) and SignetTrainer (83%).

4.1.2 Inference and Validation of Disease-Gene-Drug Relationships

The utility of condition specific gene regulatory networks lies in its ability to systematically assess the effect of perturbations to the cellular system through model simulation. Network stability determinants such as positive and negative feedback loops have been experimentally validated in living cells and play an important role in the

Figure 4.1. Drug enrichment in six validation examples



Drug enrichment results in six validation examples of drug-induced phenotypes from the CMap. The cumulative enrichments of the drug effects in the multitarget combinations are shown. *In silico* effects of the multitarget combinations are depicted on the x-axis and are normalized with respect to the highest observed effect of any combination. Drugs more specific to genes in multitarget combinations inducing high phenotypic changes are favorable. Due to potential combinatorial effects of multiple drugs, the examples of celastrol/gedunin and androgen are not compared to other drugs.

formation of attractors (Bateman, 1998). Therefore, the strategy for inferring candidate instructive factors of desired cellular conversions is based on the identification of common stability determinants and differentially regulated genes of the reconstructed networks (see Methods section 3.3). These candidate factors were subsequently contrasted with known compound effects to systematically prioritize drugs for inducing cellular conversions. In particular, the strategy consists of three parts. First, drugs having reported effects on candidate instructive factors were identified from the Comparative Toxicogenomics Database (Davis *et al.*, 2014), a compendium of reported drug modes of action. Second, millions of combinations of candidate factors, coined as multitarget combinations, were simulated as perturbations to the network attractor corresponding to the cellular phenotype. Finally, simulated multitarget combinations

Table 4.3. Comparison of simulated drugs in validation examples

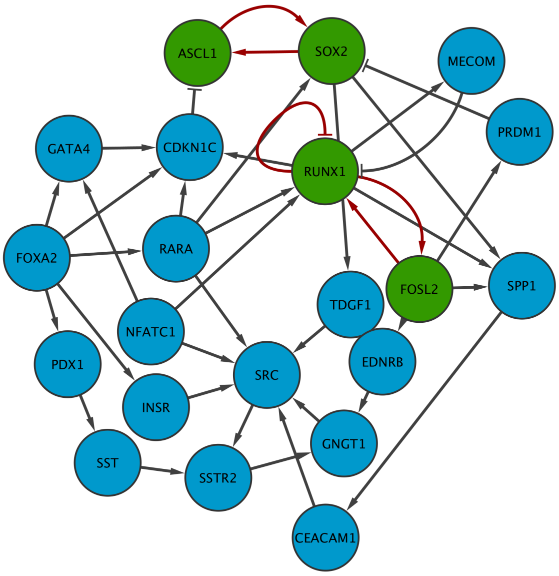
Case	Drug	AUC	Difference from AUC of uniform distribution
Celastrol+Androgen	**Celastrol+Androgen**	**37.193**	**25.61%**
Gedunin+Androgen	**Gedunin+Androgen**	**37.225**	**25.55%**
Estradiol	Arachidonic acid	54.950	-0.99%
	Estradiol	**36.887**	**26.22%**
Genistein	Cobalt chloride	47.188	5.62%
	Estradiol	44.426	11.15%
	Genistein	**41.468**	**17.06%**
	Tacrolimus	46.877	0.62%
Cobalt chloride	**Cobalt chloride**	**36.364**	**27.27%**
	Genistein	43.940	12.12%
	Tretinoin	37.586	24.83%
Celastrol	**Celastrol**	**42.076**	**15.85%**
	Parthenolide	46.688	0.66%
	Whitaferin A	52.067	-0.41%

Enrichment results of drugs for six drug-induced phenotypes. In all cases, the area under the cumulative enrichment curve (AUC) is lower, i.e. more favorable, compared to other drugs inducing similar phenotypes (highlighted in bold). The difference from an AUC of a uniform distribution shows the non-randomness of the enrichments. Combinations of drugs where not compared to other drugs, but show significant differences from an uninformed distribution.

were linked to compounds through computing the enrichment of compound targets in these combinations (see Methods section 3.3 for details).

To underline the utility of the developed strategy for linking genes and drugs to induce cellular transitions, six examples were selected from the Connectivity Map (CMap) (Lamb *et al.*, 2006), a database embodying gene expression changes upon chemical perturbation. These examples include LNCap and MCF7 cells treated with celastrol, cobalt chloride, estradiol and genistein as well as cells treated with multiple compounds at the same time, such as celastrol + androgen and gedunin + androgen. To obtain a conclusive validation, drugs showing similar differential expression patterns according to the CMap were selected for comparison. After reconstructing condition specific gene regulatory networks for the phenotypes before and after drug induction, the areas under the cumulative multitarget enrichment curves were compared to rank the drugs according to their predicted efficacy in generating the compound-induced phenotype.

Figure 4.2. Core network of untreated cells in case of cobalt chloride



The core network of the control phenotype is not responsive to the perturbation of candidate instructive factors. Only four genes are identified (green) that are contained in common stability determinants (red), the negative autoregulation of RUNX1 and two positive circuits between RUNX1 and FOSL2, and ASCL1 and SOX2. The poor circuit coverage of the network and many activating interactions result in high stability.

Figure 4.1 depicts the cumulative enrichment curves of drug effects according to *in silico* perturbation effects. Recall that normalized perturbation effects refer here to the number of induced changes to the network attractor of non-perturbed genes normalized by the highest observed effect. In all studied examples, the applied drug was correctly predicted using the proposed strategy and show significant differences from the AUCs of a uniform distribution corresponding to uninformed strategy (see Table 4.3). However, two important remarks must be made. On one hand, the combinatorial effect of celastrol/gedunin and androgen underlines the utility in the presence of drug combinations. Nevertheless, in general, the combinatorial effects of multiple compounds need to be taken into account but can be neglected in case of celastrol/gedunin and androgen as they do not possess reported antagonistic effects with respect to the genes in the network. However, in order to avoid inconclusive results, no comparison to other drug combinations has been carried out in these examples. Despite the absence of comparable drugs, a decrease of more than 25% of the area under curves with respect to an uninformative uniform distribution and the increased enrichment in high ranking multitarget combinations indicate the predicted efficacy (see Table 4.3).

On the other hand, the normalized perturbation effects are comparably low in case of cobalt chloride. *In silico* perturbation of all identified candidate instructive factors (RUNX1, FOSL2, ASCL1 and SOX2) leads to an attractors that resembles only 35% of the compound-induced gene expression pattern. Of note, ASCL1 and SOX2 are pioneer factors whose perturbation is likely involving major changes of the accessible chromatin landscape. However, this cannot be reflected by the current approach, which solely relies on the utilization of transcriptomics data and prior knowledge networks. By inspecting the reconstructed network of the cellular phenotype before drug induction, it can be seen that the majority of the network consists of regulatory interactions activating their target genes (Figure 4.2). Therefore, changes to the gene expression pattern can only be modulated by down-regulation of the activators, which cannot be achieved through targeting the genes in the network stability determinants alone.

4.1.3. Prediction of Candidate Compounds for Treating Autoimmune Diseases

Table 4.4. Comparison of drug enrichments for Systemic Lupus and Rheumatoid Arthritis

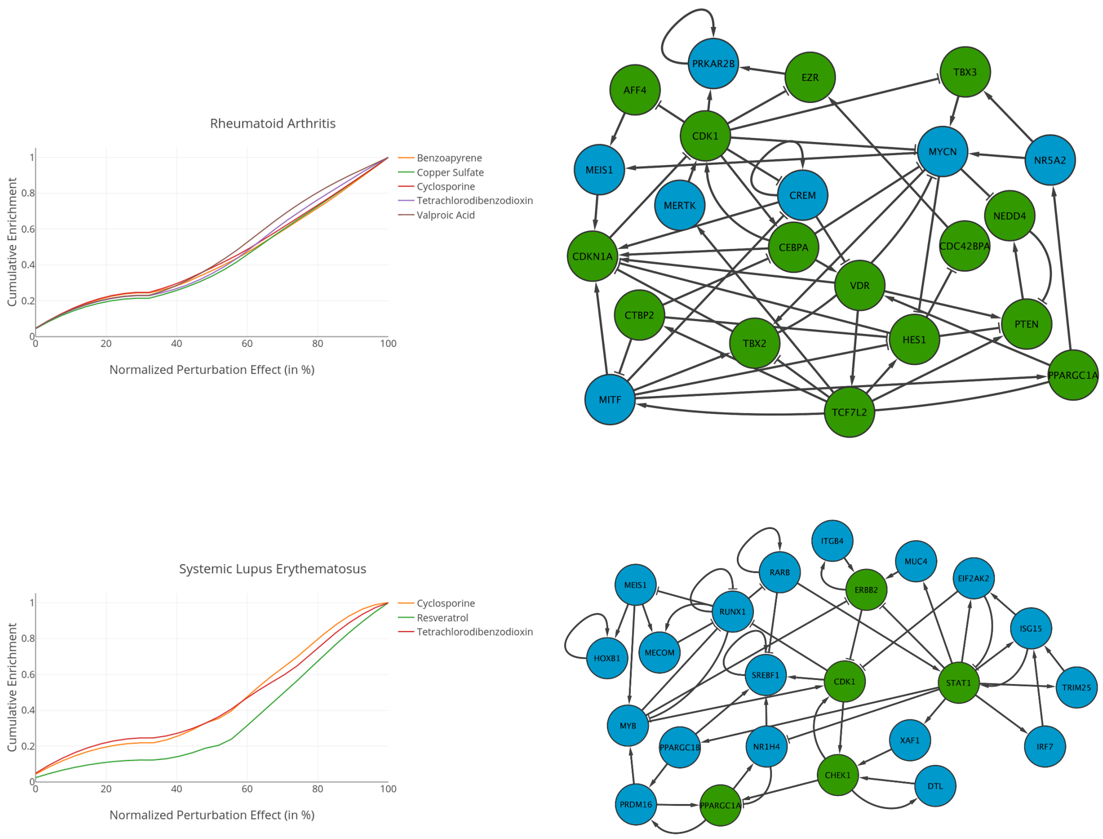
Case	Drug	AUC	Difference from AUC of uniform distribution
Systemic Lupus	Cyclosporine	43.416	13.17%
	Resveratrol	**33.030**	**33.94%**
	Tetrachlorodibenzodioxin	43.001	14.00%
Rheumatoid Arthritis	Benzo(a)pyrene	42.473	15.05%
	Copper sulfate	**41.351**	**17.30%**
	Cyclosporine	43.460	13.08%
	Tetrachlorodibenzodioxin	43.156	13.69%
	Valproic acid	45.340	9.32%

Comparison of drug enrichments for candidate drugs in Systemic Lupus and Rheumatoid Arthritis by means of their area under the curve. Lower AUC values correspond to more pronounced enrichment in high scoring multitarget combinations and are more favorable. The difference to the AUC of a uniform distribution shows that the enrichment is non-random. For Rheumatoid Arthritis, all drugs except for Valproic acid show similar enrichments while Resveratrol is clearly superior for Systemic Lupus. Prioritized drugs are highlighted in bold.

The validated drug prioritization strategy was subsequently applied to Systemic Lupus and Rheumatoid Arthritis for predicting compounds that revert the disease phenotype. Following this approach, networks were reconstructed for disease and control phenotypes of both diseases and candidate instructive factors were predicted whose multitarget combinations were able to revert the disease phenotype *in silico*. Information about drugs having reported effects on these genes was obtained from the Comparative Toxicogenomics Database (Davis *et al.*, 2014).

Rheumatoid Arthritis was analyzed on the basis of healthy and pathologic B cells (GEO: GSE4588). Twenty-seven candidate instructive factors were identified, including TCF7L2, which is associated to Arthritis (Mota *et al.*, 2012), and CDKN1A whose down-regulation results in an increased risk for developing autoimmune diseases (Perlman *et al.*, 2003). The drugs having the most reported effects on the candidate factors are benzo(a)pyrene (14 targets), copper sulfate (12 targets), Tetrachlorodibenzodioxin (TCDD, 12 targets), valproic acid (12 targets) and cyclosporine (10 targets). According to the area under the enrichment curve (Table 4.4), copper sulfate has the most pronounced effect on the phenotype, which is further supported by 53% reversion of the pathological phenotype upon *in silico* perturbation of all known drug effects. Additionally, copper sulfate targets most of the genes in the core disease network (see Figure 4.3) and as such constitutes a novel, predicted treatment for Rheumatoid Athritis whose therapeutic effect needs to be further elucidated (Fernández-Madrid, 1998). On the other hand, cyclosporine is a broadly applied drug for treating arthritis (Wells *et al.*, 1998), but shows a less pronounced effect with respect to the analysis.

Figure 4.3. Drug prioritization results for Systemic Lupus and Rheumatoid Arthritis



Cumulative enrichment distributions of candidate drugs for treating the pathologic phenotypes of Systemic Lupus Erythematosus and Rheumatoid Arthritis in B cells. Normalized perturbation effects refer to the absolute changes induced by multitarget combinations normalized by the highest observed effect. Drugs more highly enriched in multitarget combinations reverting the disease-phenotype the most are more favorable compared to others. (Upper Panel) Similar enrichment patterns are found for Cyclosporine, Copper sulfate and Benzoapyrene. Copper sulfate shows a pronounced *in silico* effect as it has reported effects on almost the complete disease network (green nodes). Perturbation of reported effects of Benzoapyrene results in a cyclic attractor and is therefore not considered. (Lower Panel) For Systemic Lupus in B cells, Resveratrol is the top-ranking drug having reported effects on five genes in the core disease network (green nodes). The perturbed genes are covering most of the common circuits in the network.

Similar to the analysis conducted for Rheumatoid Arthritis, healthy and disease phenotypes of Systemic Lupus in B cells have been obtained (GEO: GSE4588). Eleven candidate instructive factors were identified among which STAT1 constitutes a marker for onset and progression of the disease pathology (Liang *et al.*, 2014). Moreover, a positive feedback loop containing STAT1, IRF7 and ISG15 has been detected in the reconstructed networks of the pathological and control phenotypes (see Figure 4.3). This feedback seems to be of particular importance for the activity of Systemic Lupus. IRF7 was functionally associated with Systemic Lupus and results in increased type I interferon (IFN) induction upon up-regulation (Xu *et al.*, 2012). Several IFN inducible genes, such as ISG15, were significantly up-regulated in the disease pathology and serve as markers of disease activity (Feng *et al.*, 2006).

The drugs having the most reported effects among these candidate factors were found to be tetrachlorodibenzodioxin (TCDD, 8 targets), cyclosporine (5 targets) and resveratrol (5 targets). Among these drugs, resveratrol shows a considerably more favorable effect, i.e. lower area under the enrichment curve (see Figure 4.3), and additionally reverts 59.4% of the pathologic gene expression program upon *in silico* simulation its reported effects. Experimental evidence in human macrophages supports this prediction, since resveratrol has been shown to act as an antiatherogenic agent, i.e. it counters the formation of fatty deposits in the arteries (Reiss *et al.*, 2012). Even though TCDD shows a less favorable enrichment pattern, it has been found to act immunosuppressive to SLE mice and significantly decreased the symptoms (Li and McMurray, 2009). Interestingly, cyclosporine has significant effect on the disease pathology of SLE, which has been proven in clinical trials (Caccavo *et al.*, 1997), but does not turn out to be the most effective drug according to the analysis.

4.2 Prediction of Chromatin Accessibility in Gene-Regulatory Regions from Transcriptomics Data

Cellular phenotypes are shaped by the complex interplay of transcriptional and epigenetic regulatory interactions. The dynamic epigenetic landscape allows for establishing distinct transcriptional regulatory interactions through the modulation of chromatin accessibility and activity. Therefore, it is of utmost importance to pinpoint the accessible gene-regulatory genomic regions that give rise to a particular phenotype. Several experimental procedures, such as DNase I hypersensitive site sequencing (DNase-seq), formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) and the assay for transposase-accessible chromatin using sequencing (ATAC-seq), have been developed for profiling the accessibility landscapes in different cell types (Buenrostro *et al.*, 2013; Song *et al.*, 2011). However, current downstream computational tools for the identification of enriched genomic sites, i.e. peak callers, fail to provide reliable results. Based on a previous study, the overlap of the four most widely used peak calling algorithms (Hotspot, F-Seq, ZINBA and MACS) applied to the same dataset amounts to only 11% of all identified regions (Koohy *et al.*, 2014). In addition, the configuration of peak calling parameters has significant effects on the identification of these regions whereby a dataset-dependent optimal setting is usually unknown (Koohy *et al.*, 2014). Especially the control of the false discovery rate is fundamental for balancing false negative and false positive peaks. Here, more stringent cutoffs may result in increased false negatives while less stringent cutoffs render an elevated number of false positive peaks.

Besides the limitations of the current computational pipeline, the absence of experimental chromatin conformation assays hinders the systematic study of gene regulation and its modulation by the accessibility landscape. For example, the pronounced effect of pioneer factors on the formation of stable, compound-induced phenotypes cannot be described without additional information about epigenetic changes (see section 4.1.2). Hence, a machine-learning approach was developed that predicts gene-level accessibility from transcriptomics data in the absence of experimental data and can assist in the identification of optimal peak calling parameters if chromatin accessibility assays are available.

4.2.1 Machine-Learning Model Construction and Cross-Validation

The methodology for predicting chromatin accessibility is based on a stacked classification tree model, specially designed for utilizing transcriptomics data as an input (see Methods section 3.4.2 for details). In order to train the model, a collection of 18 distinct RNA-seq gene expression profiles and corresponding chromatin accessibility assays (DNase-seq) from various cell types and cell lines was compiled and homogenously processed. Chromatin accessibility data was aligned to human genome version 19 (hg19) and subsequently screened for enriched genomic regions using Hotspot (John *et al.*, 2011) with standard false discovery rate of 1%. It is important to note that the training datasets consists of several cancer cell lines typically containing several structural variations (Moncunill *et al.*, 2014; Malhotra *et al.*, 2013). Thus, experimental sequencing data of these samples should be aligned to their specific sequenced genome instead of the reference genome to account for these genetic variations. However, the unavailability of these sequenced genomes only leaves the possibility of aligning it to a reference genome. To facilitate the utilization of the model predictions in the context of gene regulatory network reconstruction, a gene-level accessibility measure was obtained. A previous study reported distinct DNase-seq peak locations depending on the expression level, i.e. highly expressed genes predominantly show peaks around their transcription start site (TSS) while medium and lowly expressed genes contain peaks in their gene bodies (He *et al.*, 2014). Consequently, a gene is called accessible if its promoter or gene body is overlapping a peak identified by Hotspot and inaccessible otherwise.

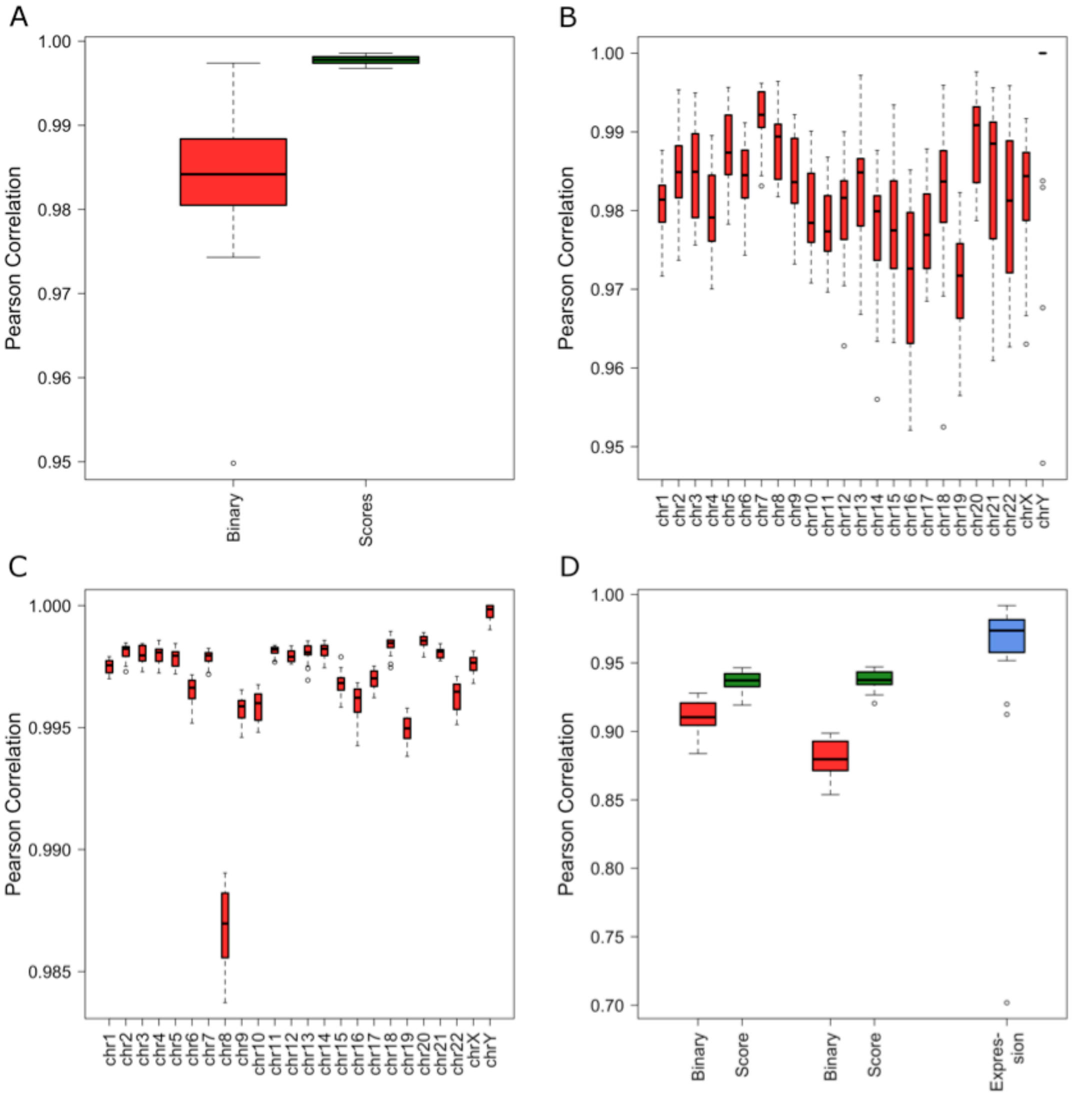
After the model was conceived and an appropriate training dataset was selected, its predictive power and generalizability to unseen samples was analyzed by means of leave-one-out cross-validation. More specifically, leave-one-out cross-validation was performed horizontally and vertically by partitioning the complete dataset either by sample or by chromosome. The model was then trained with all but one partition while the remaining one was left out for prediction. The correlation of the predictions of the reduced model with those of the full model then quantifies the generalizability to unseen data.

The horizontal comparison of the full model against the cross-validation model by means of binary gene-level accessibility and assigned confidence scores of the predictions reveal strong individual correlations of at least 0.95 (Figure 4.4A). On average, the predictions of both models are highly concordant showing correlations of 0.984 and 0.997 for binary gene-level accessibility assignments and their corresponding confidence scores, respectively. These results are further confirmed by vertical cross-validation in which all but one chromosome was used for training while the predictions were carried out on the hold out chromosome. The obtained correlations of predicted binary gene-level accessibility (Figure 4.4B) and confidence scores (Figure 4.4C) of the full and hold-out model are similar to those of the horizontal cross-validation. While some chromosomes exhibit greater correlation variations, a minimum concordance of 0.947 supports the insensibility of the model towards the training dataset. Of note, the correlations of the confidence scores in chromosome 8 appear lower than all others. However, since all obtained correlations are above 0.98, these differences are insignificant.

Owing to the cellular heterogeneity leading to variations across different gene expression measurements of the same cell type, the reproducibility of predictions in the presence of gene expression replicates was assessed (Figure 4.4D). A reliable second expression sample was collected for 17 cell types or cell lines in the training dataset. The available replicate of HMEC cells was of poor quality showing no expression for most of the genes and was consequently excluded from further analysis. All other replicates where highly consistent (median correlation of 0.97) with HeLa-S3 cells being a notable outlier (correlation 0.7). The hierarchical classification tree model was then trained with the set of first (second) replicates to predict the accessibility of both replicates. Subsequently, the correlations of the gene-level assignments and confidence scores were assessed (Figure 4.4D, left and middle boxes). In the optimal case, similar correlations than in the horizontal cross-validation analysis were expected. However, even though the obtained correlations of at least 0.85 (median: 0.91 and 0.88, respectively) indicate a strong concordance between the predictions of the model trained with different replicates, a significant difference was observed. Given that the confidence scores do not exhibit median correlation differences, a plausible explanation of the correlation differences in predicted gene-level accessibility assignments is that the optimal confidence score thresholds, above which a gene is predicted to be accessible, are distinct in both cases. However, due to the inability of identifying optimal thresholds for unseen data and the obtained strong correlations, the standard threshold, which assigns the gene to the accessibility status having the highest confidence, is kept throughout the following analyses.

Overall, the results of the cross-validated model highlight the insensitivity of predictions to the training dataset and its generalizability to unseen data providing confidence in the accurate model design for predicting gene-level accessibility.

Figure 4.4. Correlation analysis of cross-validated and full model predictions



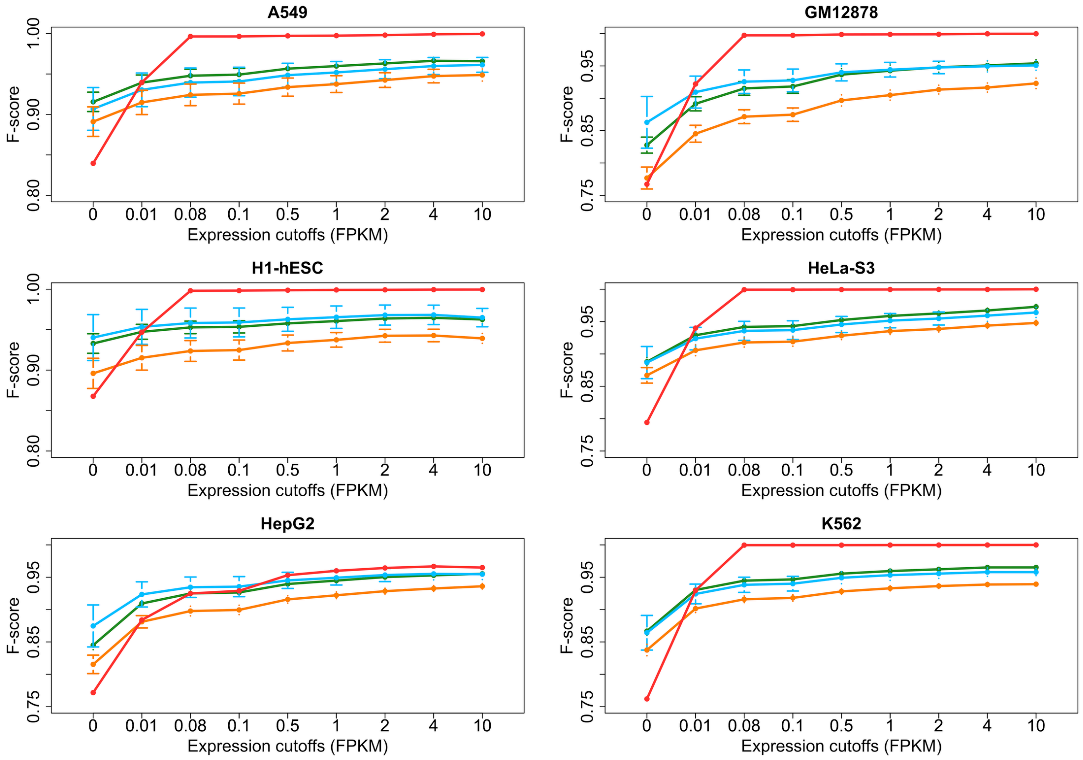
Analysis results of cross-validation assays correlating predictions from the full model trained with 18 samples and leave-one-out models are represented by boxplots. Whiskers extend to 1.5 times the interquartile range and outliers are depicted as circles. (A) Pearson correlation of binary gene-level predictions and confidence scores show highly reproducible results with values greater than 0.95. (B) Cross-validation by training the model with all but one chromosome and predicting the remaining one yields similar performance to the sample-based cross-validation having correlations greater than 0.94 in all cases. (C) Comparison of the correlation of scores shows an almost perfect relationship through correlations greater than 0.98. (D) Reproducibility analysis for multiple samples of the same cell lines. Training with the first (left) and second (right) replicates while predicting gene-level accessibility of both shows high correlations between the predicted values. Correlations of raw gene expression replicates are depicted in the blue box plot.

4.2.2 Comparison of predicted gene-level chromatin accessibility with traditional peak calling algorithms

Despite the previous quantitative evaluation of the conceived model, it is inevitable to assess its predictive power with respect to experimental evidences and compare it to current downstream computational tools for processing chromatin accessibility assays based on a gold standard dataset. Since DNase-seq experiments that were processed with traditional peak calling algorithms were used for training the model, a distinct validation set was compiled. First, between 31 to 100 transcription factor binding site (TFBS) ChIP-seq experiments were obtained from ENCODE ((Dunham *et al.*, 2012), Table S2) in six cell lines (A549, GM12878, H1-hESC, HeLa-S3, HepG2 and K562) for pinpointing accessible regions. Genes overlapping with at least one TFBS are defined as accessible, due to the high overlap of TFBS and DNase hypersensitive sites (Song *et al.*, 2011). Inaccessible genes, on the other hand, cannot be reliably determined by TFBS ChIP-seq experiments, since it is impossible to decide whether the absence of binding events is caused by incorrect downstream processing, the unavailability of experimental data for other TFs that could bind this region or the gene's location in inaccessible chromatin. Therefore, inaccessible genes were determined based on ChromHMM (Ernst and Kellis, 2012), a computational tool for annotating genomic regions based on histone modification ChIP-seq experiments. Precompiled datasets for the six cell lines were obtained from the Roadmap Epigenomics project (Roadmap Epigenomics *et al.*, 2015) including annotations dependent on five histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3). Heterochromatin was then defined as regions enriched in either a combination of H3K4me3, H3K36me3 (Chantalat *et al.*, 2011) and H3K9me3 indicating ZNF genes and repeat regions targeted by heterochromatin proteins (Vogel *et al.*, 2006; Blahnik *et al.*, 2011) or H3K9me3 alone. Genes that overlap at least one genomic region enriched in one of the two aforementioned categories are defined as inaccessible.

For comparing the developed machine learning approach against traditional peak calling methods, transcriptomics and experimental DNase-seq data was obtained for the six cell lines in the gold standard dataset. DNase-seq assays were processed by MACS (Zhang *et al.*, 2008), F-Seq (Alan P Boyle *et al.*, 2008) and Hotspot (John *et al.*, 2011) with varying false discovery rates or Z-score cutoffs to identify genes located in accessible and inaccessible regions. Identified peaks were processed as described for model training. Predictions were obtained by applying the model to corresponding RNA-seq datasets and subsequently compared to the assignments from MACS, F-Seq and Hotspot by means of the harmonic mean of precision and recall, the F_1- score. Since the developed model utilizes transcriptomics data, it is of utmost importance to assess the predictive power for different expression ranges. Therefore, different lower expression cutoffs were applied to define sets of genes whose expression is greater than the threshold (Figure 4.5).

Figure 4.5. Comparison of stacked classification tree model with peak calling methods



Comparison of gene-level accessibility predictions (red) with processed experimental data by F-Seq (blue), MACS (orange) and Hotspot (green) in six gold standard datasets by means of F-scores. Performance is measured for different expression thresholds including only genes expressed above the cutoff. Dots represent mean values and corresponding confidence intervals are shown for different replicates and parameter settings (FDR or z-score). Predictions on the full dataset have on average 0.075 lower scores. Performance is gradually increasing for higher cutoffs with similar performance achieved when excluding non-expressed genes (0 FPKM).

Analysis of the complete datasets, corresponding to a lower threshold of 0 FPKM, suggests a strong predictive power of the developed method but shows slightly reduced performance compared to traditional peak calling approaches (average difference between 0.055 and 0.094). However, in contrast to F-Seq, MACS and Hotspot, the performance gradually increases until F_1- scores of 0.999 are reached for genes more expressed than 0.08 FPKM. This trend is observed in all gold standard datasets except HepG2. Of note, this cutoff cannot be regarded as a threshold for determining expressed genes and is substantially lower than previously reported values

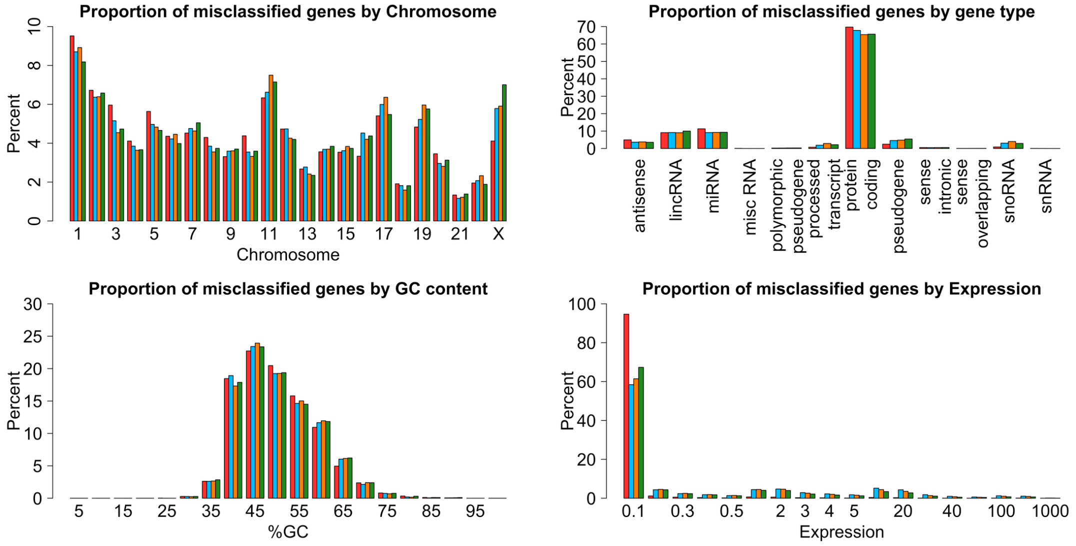
(Haltaufderhyde and Oancea, 2014; Rau *et al.*, 2013; Trakhtenberg *et al.*, 2016). Thus, the visible robust performance is not due to the trivial assertion that expressed genes should be, generally, located in accessible chromatin domains. Importantly, remind that the F_1-score is defined as the harmonic mean of precision and recall i.e. the ratios of truly accessible predictions over all genes predicted to be accessible and all accessible genes, respectively. Therefore, it is independent of the amount of accessible and inaccessible chromatin regions. This implies that the F_1-scores of embryonic stem cells (H1-hESC), which contain more accessible chromatin regions than differentiated cells, do not differ significantly from the other cell lines.

Even though the developed machine learning approach yields comparable performance to any peak calling methodology, which relies on experimental data, the utilization of gene expression as a predictor of chromatin accessibility might introduce unexpected biases in the predictions. In particular, three potential sources can be distinguished:

- ***GC content of genes.*** The GC content of genes is negatively correlated with their methylation level (Meissner et al., 2008), which in turn influences gene expression. In particular, high methylation levels are indicative of low or no expression even though the gene is located in accessible chromatin.
- ***Gene expression.*** The hierarchical classification tree model uses gene expression as a predictor while peak calling algorithms neglect transcription.
- ***Gene type.*** Genes of different types differ in their expression, e.g. transcripts of protein-coding genes are typically more abundant than those of non-coding genes (Djebali et al., 2012). Especially non-coding genes are largely lowly expressed (Iyer et al., 2015) (< 1 FPKM) and, as such, might be more often misclassified.

The analysis of the proportion of misclassified genes for each potential bias source revealed gene expression as the only significant difference (Figure 4.6). More specifically, the maximum difference in the proportion of misclassified genes between predictions and peak calling methods amounts to only 3% per chromosome and 7% per gene type. Differences in genes by their GC content are below 1% and as such negligible. However, as already suggested by the evaluation of predictions on the gold standard dataset, gene expression is a significant source of bias. 94% of misclassified genes are expressed below 0.1 FPKM, which is significantly higher in comparison to F-Seq (58.4%), MACS (61.4%) and Hotspot (67.3%), respectively. A possible explanation for this result is the deterministic behavior of the model. When using only gene expression or its derivatives as predictors for accessibility, the same expression value is classified as either accessible or inaccessible. Especially non-expressed genes (0 FPKM) show a high misclassification rate because of the indistinguishability of accessible and inaccessible chromatin regions. Finally, the discrimination of genes by chromosomal location was not expected to yield significant biases in the prediction of gene-level accessible and was analyzed as a negative control. The results of this assessment confirmed the unbiased predictions.

Figure 4.6. Assessment of potential sources of bias in the prediction of gene-level accessibility

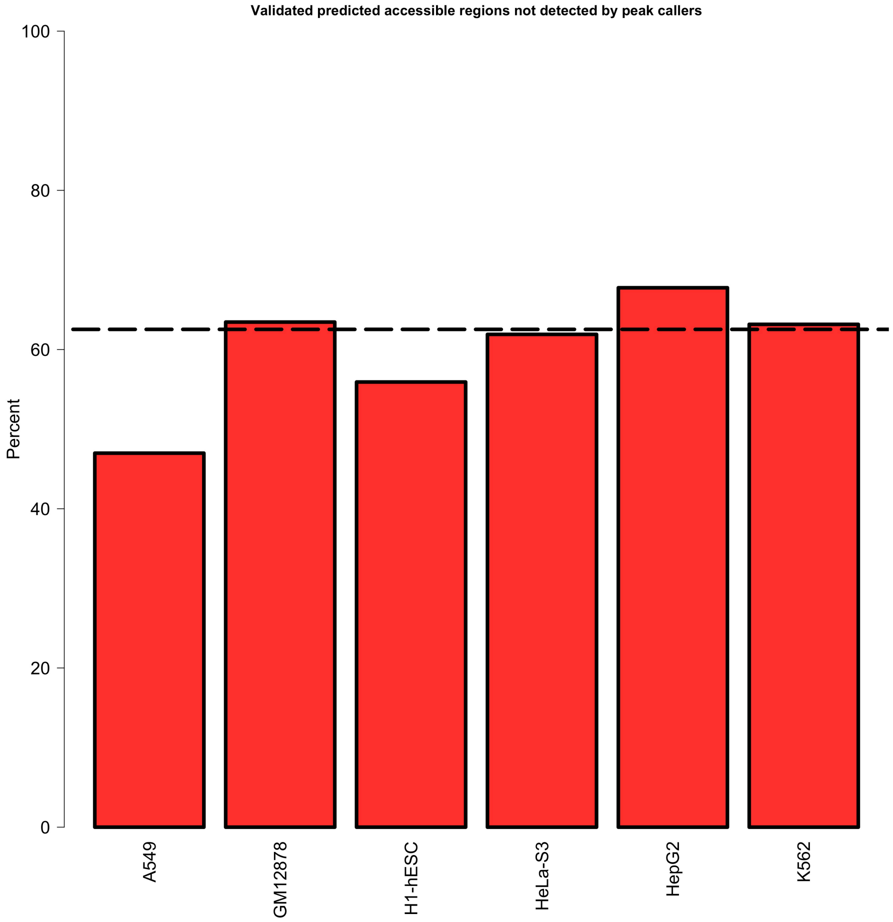


Potentially influencing factors of the predictions (red bars) were assessed and compared against assignments of F-Seq (blue bars), MACS (orange bars) and Hotspot (green bars). Neither the chromosomal location nor gene type nor the GC content of misclassified genes distinguishes predictions from observations. However, by predicting accessibility based on gene expression (bottom right histogram), we significantly reduce the number of misclassified genes in regions expressed above 0.1 FPKM (6% compared to 41.6%, 38.6% and 37.2%). Due to the deterministic prediction, non-expressed genes are overrepresented in the range of 0 to 0.1 FPKM (first bar, 38% out of 94%).

At last, further investigation was conducted with respect to the number of genes predicted to be accessible but declared to be inaccessible by downstream processing of DNase-seq assays. Interestingly, between 47% and 67% (median: 62%) of these genes in the six gold standard cell lines were bound by a transcription factor and were therefore accessible (Figure 4.7). Since the gold standard dataset consists of only 31 to 100 transcription factor binding site ChIP-seq experiments, these numbers are supposedly increasing with more available datasets.

Overall, the developed methodology is able to accurately classify genes located in accessible and inaccessible chromatin regions based on transcriptomics data when experimental DNase-seq data is unavailable. Additionally, the accordingly derived gene-level accessibility is readily usable for integration into the gene regulatory network framework and provides a reliable proxy for identifying genes that are actively regulated. Experimental DNase-seq data is, however, still favorable if peak calling parameters could be optimized to render less false negatives assignments, due to its pronounced ability of distinguishing the chromatin accessibility of non-expressed genes.

Figure 4.7. Validated predicted accessible regions not detected by peak callers



Percentage of validated accessible genes that are predicted by the model and not detected by peak calling methods. Validation was performed on the basis of the TFBS ChIP-seq experiments in the gold standard datasets from ENCODE. Bars represent the percentage of genes bound by transcription factors with respect to all predicted accessible genes that are not detected by peak callers. Between 49% (A549) and 69% of predictions could be validated (median 62%, dashed line).

4.2.3 Optimizing peak calling parameters with model predictions

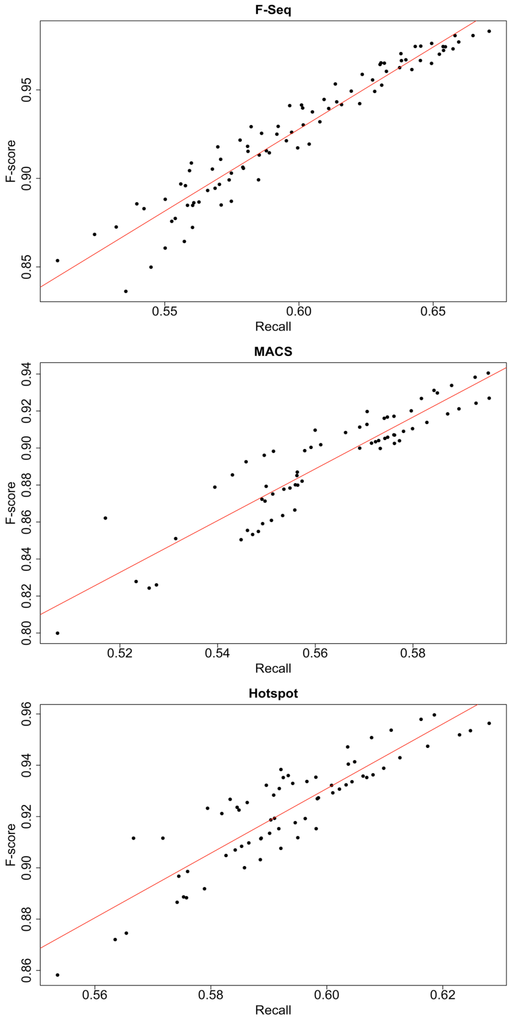
Obtaining functional insights from chromatin accessibility assay peak locations and reliable gene-level assignments of non-expressed genes are major obstacles, which cannot be overcome by the sole use of gene expression. Instead, other experimental assays could be used to improve the predictions or the parameters for downstream computational tools could be optimized. In the following, the latter approach will be

Table 4.5. False Discovery Rate thresholds for peak calling methods

Peak Caller	Thresholds
F-Seq (version 1.84)	0.5, 1, 2, 3, 4, 5, 6, 7
MACS (version 2.1.0.20140616)	0.01, 0.05, 0.075, 0.1, 0.15, 0.2
Hotspot (version 5.1)	0.01, 0.05, 0.075, 0.1, 0.15, 0.2

For MACS and Hotspot, thresholds correspond to false discovery rate (FDR) and in case of F-Seq to z-score thresholds. Z-score thresholds are negatively correlated with FDR. Thus, small values correspond to high FDR thresholds and large values to low FDR thresholds.

Figure 4.8. Dependence between Recall-measure and F-score



Regression analysis provides evidence for the strong linear relationship of F-scores on the gold standard datasets and the proposed recall measure (adj. R squared of 0.92, 0.84 and 0.79). Dots represent distinct parameterization, cell line, and replicate settings.

addressed. In particular, the conducted analysis focuses on the optimization of the false discovery rate, an important parameter in all peak calling methodologies (Koohy *et al.*, 2014), through the employment of predicted gene-level accessibility..

The validation results in the six gold standard datasets highlighted the accuracy of the machine-learning model predictions for genes expressed above 0.01 FPKM. Thus, peak calling parameterizations resulting in the highest resemblance of recovered accessible genes with respect to genes predicted to be accessible should render the most reliable results. More specifically, this translates to the recall of peak caller derived accessibility assignments with respect to model predictions. This assertion was tested for F-Seq, Hotspot and MACS with varying false discovery rate (Z-score) parameters (Table 4.5) and evaluated against the F_1- score in the six gold standard cell lines for each parameterization. Of note, the called peaks are again transformed to gene-level accessibility, as described earlier in this section. The results of the analysis reveal a strongly positive, linear relationship between the F_1- scores and the proposed recall measure. Adjusted coefficients of determination of 0.92, 0.84 and 0.79 for F-Seq, MACS and Hotspot, respectively, indicate that the recall measure explains the vast majority of variation in the F_1- scores (Figure 4.8). However, for practically determining the best parameterizations, the rankings based on the recall measure and F_1- score have to agree. Therefore, rankings were obtained for the individual measures and pairwise compared. This intuitively corresponds to the orthogonal projection of the dots in Figure 4.8 onto the regression line. As expected from the visual representation in Figure 4.8, the rankings agree for all peak calling methodologies, which provides support for the suitability of the proposed recall measure for parameter optimization.

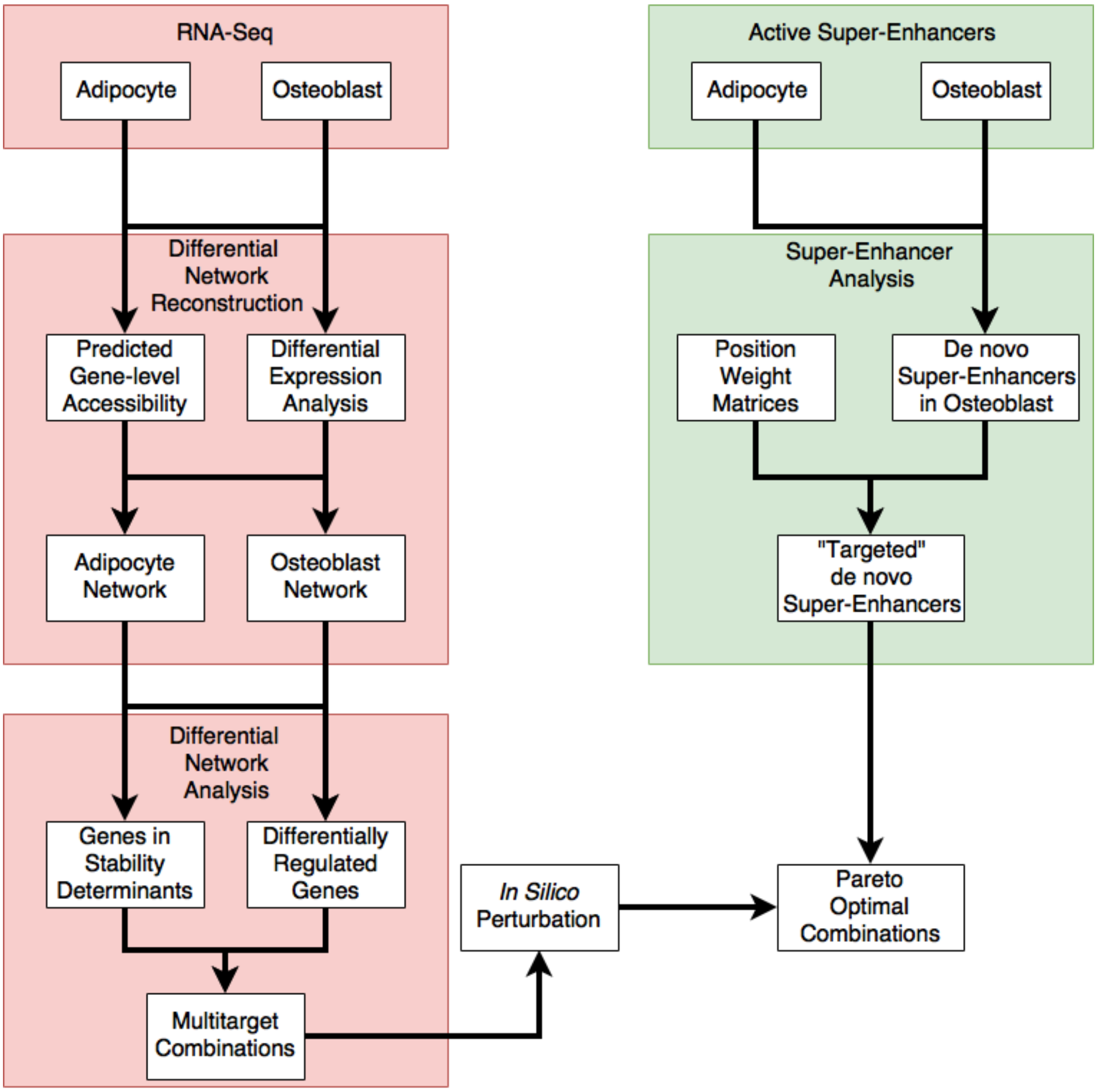
Overall, if experimental data is available, the developed methodology can be used for optimizing parameters of downstream computational tools leading to more accurate assignments of non-expressed genes.

4.3 Integrating Epigenetics, Transcriptomics and Prior Knowledge Networks for Predicting Candidate Instructive Factors in Adipocyte to Osteoblast Conversion

The methodology for predicting gene-level accessibility can be readily integrated into the framework for reconstructing condition-specific gene regulatory networks. In the attractor state, inaccessible genes cannot regulate other genes, due to the absence of expression, and cannot be regulated by other genes. Therefore, these genes do not play an active role in the stabilization of the attractor and should be excluded from the during network reconstruction in a condition dependent manner. More specifically, the attractor of the condition-specific networks should only be explained by accessible genes that can be actively regulated while inaccessible genes need to be removed from the networks.

In order to highlight the utility of the proposed condition-specific differential network analysis, gene-level accessibility is integrated into the existing framework for predicting multitarget combinations inducing the cellular conversion of adipocytes into osteoblasts. An integrative analysis approach is presented that involves the reconstruction of condition-specific networks as well as the identification of genes targeting osteoblast-specific super-enhancers (Figure 4.9).

Figure 4.9. Workflow for predicting instructive factors of adipocyte to osteoblast conversion


RNA-Seq
Adipocyte
Osteoblast
Differential Network Reconstruction
Predicted Gene-level Accessibility
Differential Expression Analysis
Adipocyte Network
Osteoblast Network
Differential Network Analysis
Genes in Stability Determinants
Differentially Regulated Genes
Multitarget Combinations
In Silico Perturbation
Active Super-Enhancers
Adipocyte
Osteoblast
Super-Enhancer Analysis
Position Weight Matrices
De novo Super-Enhancers in Osteoblast
"Targeted" de novo Super-Enhancers
Pareto Optimal Combinations

The workflow for predicting instructive factors for adipocyte to osteoblast conversion consists of two parts. First, cell type specific networks were reconstructed and optimal combinations of genes were identified by differential network analysis (red). Then, super enhancers unique to osteoblasts were obtained and the transcription factors with predicted binding sites in them were identified by position weight matrices (green). The pareto optimal solutions of both assessments constitute the final set of predictions.

4.3.1 Differential Network Reconstruction and Identification of Candidate Instructive Factors

Transcriptome profiling of mouse ST2 multipotent bone marrow stromal cells (MSCs) differentiated into adipocytes and osteoblasts was performed after 15 days by RNA sequencing (differentiation was conducted following the protocol in (Gerard *et al.*, 2017)) and differentially expressed genes were detected ($\log_2$ FC >= 1). A prior knowledge network of direct gene regulatory interactions between differentially expressed genes was extracted from MetaCore™ and subset to interactions contributing to transcriptional regulation. In total, 147 regulatory interactions among 55 transcription factors were retrieved from which cell type specific networks were reconstructed.

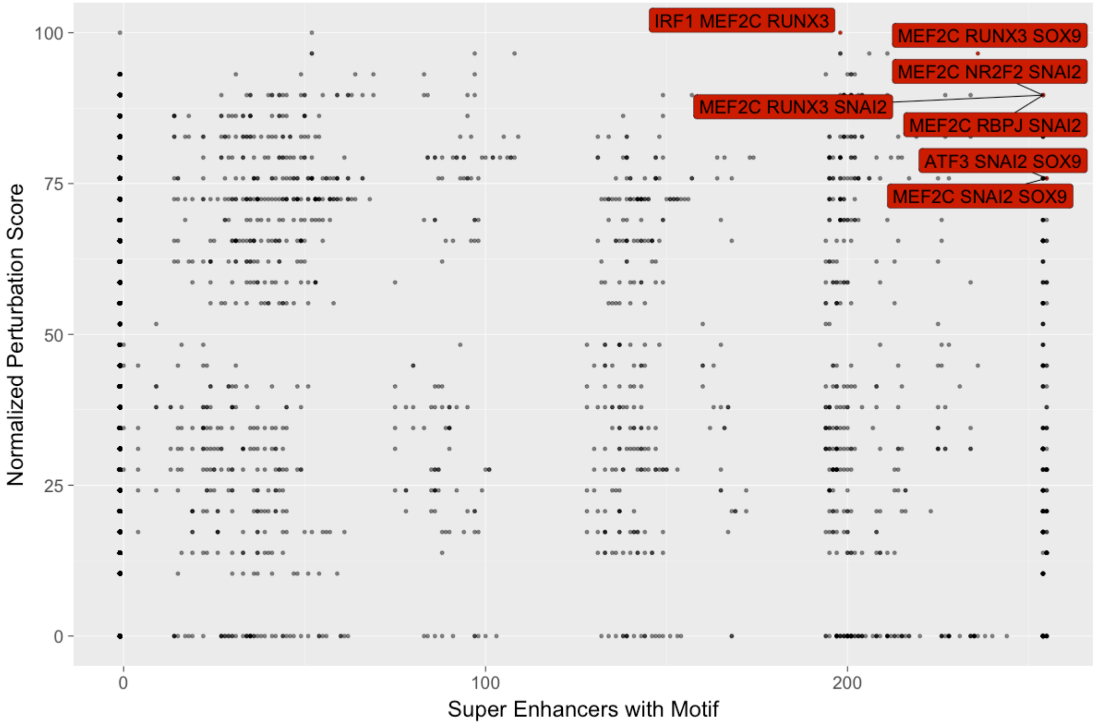
Due to the absence of experimental chromatin accessibility data for the differentiated cells, the methodology for predicting gene-level accessibility was applied to the transcriptomics profiles. The stacked classification tree model was validated in the previous section using solely human data. Therefore, in order to apply the method to expression data from mouse cells, a training dataset was compiled from ENCODE (Dunham *et al.*, 2012) consisting of 14 homogeneously processed transcriptomics and DNase-seq samples of various cell lines, cell types and tissues. All transcription factors more expressed than 1 FPKM or those predicted to be accessible were defined to be accessible. With it, inaccessible genes in adipocytes and osteoblasts were removed from the corresponding networks during network reconstruction.

Following this rationale, cell type specific networks were obtained for both cell types and candidate instructive factors for inducing the transition from adipocytes to osteoblasts were derived from elementary feedback loops in the common network parts and transcription factors under differential regulation. Due to the dense transcriptional regulation in the two networks, 38 transcription factors are selected as candidate factors.

4.3.2 Prediction of Instructive Factors for Adipocyte to Osteoblast conversion

The identified candidate instructive factors were combined into multitarget combinations of up to three transcription factors and used for *in silico* perturbation of the adipocyte network. In addition, active super-enhancers were identified in adipocytes and osteoblasts as dense regions of active enhancers marked by H3K27ac. Due to the pronounced effect of super-enhancers on cell identity (Hnisz *et al.*, 2013), the 330 genes that overlap 256 osteoblast specific super-enhancers are presumed to play an important role in the cellular conversion. Therefore, potential regulators for each osteoblast specific super-enhancer were identified by transcription factor binding site predictions with MOODS (Korhonen *et al.*, 2009) using position weight matrices from HOCOMOCO (Kulakovskiy *et al.*, 2016).

Figure 4.10. Predicted pareto-optimal combinations for inducing adipocyte to osteoblast conversion



Combinations of candidate instructive factors were scored based on their ability to induce the osteoblast phenotype upon *in silico* perturbation and predicted targeting of osteoblast-specific super-enhancers. Annotated red dots are the pareto-optimal combinations. Darker dots represent multiple combinations having the same score in both assessments. A normalized perturbation score of 100 corresponds to 81% induction of the core adipocyte network.

Pareto-optimal multitarget combinations were derived based on the predicted number of targeted osteoblast specific super-enhancers and the *in silico* induction of the osteoblast phenotype upon perturbation (Figure 4.10). In particular, pareto-optimality describes a combination in which one ranking criterion cannot be improved without lowering the other criterion. Markedly, all but one of the seven pareto-optimal combinations contain myocyte enhancer factor 2c (MEF2C) highlighting its predicted pronounced effect in the induction of osteoblasts. A previous study provides further support to this finding showing that MEF2C deficiency impedes osteoblast differentiation, extracellular matrix mineralization and osteoblast specific gene expression (Stephens *et al.*, 2011). Due to the pronounced effect of the multitarget combination of interferon regulatory factor 1 (IRF1), MEF2C and runt-related transcription factor 3 (RUNX3) on the network attractor upon *in silico* perturbation (81% reversion of the core adipocyte network) while potentially regulating 76% of the osteoblast specific super-enhancers, it was selected as a suitable candidate for the induction of osteoblasts. More specifically, perturbation of the network with this combination involves the down-regulation of IRF1 together with the up-regulation of both MEF2C and RUNX3. From a practical point of view, micro RNAs can efficiently modulate down-regulation of IRF1 through post-transcriptional regulation. Especially miR-23a and miR-24-2, which are both contained in the micro RNA cluster mirn-23a, are predicted targets of IRF1, according to miRDB (Wong and Wang, 2015). Further support to this prediction is provided by a previous study that identified IRF1 as a direct target of miR-23a resulting in significantly decreased expression (Liu *et al.*, 2013).

The potential efficacy of the identified multitarget combination is supported by previous studies. RUNX3-deficient mice show a decreased number of active osteoblasts and bone formation deficiencies, which provides evidence that RUNX3 is an integral component in proper osteogenesis (Bauer *et al.*, 2015). On the other hand, miR-23a, together with miR-23b, is a vital regulator of osteoblast and adipocyte cell fate determination (Guo *et al.*, 2016). In particular, it has been shown that overexpression of these micro RNAs promotes osteoblast differentiation in mesenchymal stem cells while their down-regulation promotes adipocyte differentiation. This observation is confirmed by the transcriptomics data from MSC-derived adipocytes and osteoblasts used in this analysis, since miR-23a/b are undetectable in adipocytes while being clearly expressed in osteoblasts (143 and 85 FPKM, respectively).

4.3.3 Experimental Validation of Predicted Instructive Factors

Despite the support provided to the predictions by previous studies of the individual factors, it is necessary to validate the predicted multitarget combination experimentally. Mouse ST2 MSCs from Whitlock-Witte BC8 long-term bone marrow culture will be cultured and differentiated into adipocytes as described in a previous study (Gerard *et al.*, 2017). More specifically, a differentiation medium consisting of growth medium, 5 µg\mL insulin (Sigma-Aldrich, I9278), 0.5 mM isobutylmethylxanthine (IBMX) (Sigma-Aldrich, I5879) and 0.25 µM dexamethasone (DEXA) (Sigma-Aldrich, D4902) will be added for two days. After 48 hours, the medium will be exchanged with another differentiation medium consisting of growth medium, 5 µg/mL insulin (Sigma-Aldrich, I9278) and 500 nM rosiglitazone (RGZ) (Sigma-Aldrich, R2408). The medium will be replaced every two days until 15 days of differentiation.

Lentiviral vectors, designed as LV-EF1A-mirn23a-CMV-MEF2C-T2A-RUNX3-T2A-eGFP, were produced. EF1A and CMV are constitutive promoters in mammalian cells showing high expression in many cell types (Qin *et al.*, 2010). These sequences serve as gene promoters to allow the active transcription of mirn-23a, MEF2C, RUNX3 and GFP. For expressing multiple genes with the same promoter, T2A sequences were added as separators. With it, polypeptide cleavage is mediated by skipping the formation of peptide bonds and, thus, allowing the individual translation of MEF2C, RUNX3 and GFP (Liu *et al.*, 2017). In this setting, lentiviral vectors are utilized due to their ability to transduce non-dividing cells, like the adipocytes used for the predictions of the multitarget combination.

Vectors will be transduced to the differentiated adipocytes deprived from the differentiation medium and compared to cells also deprived from the differentiation medium but transduced with a control virus containing a GFP reporter only. Successful transduction will be validated by the expression of the GFP reporter contained in the vectors allowing for a visual assessment of the transduction efficiency and for fluorescence activated cell sorting (FACS), if needed.

Validation of the successful conversion will be organized in two parts. First, cells before and after virus transduction will be stained with Oil Red O and alkaline phosphatase to measure the decrease of lipid droplets and the increase of alkaline phosphatase, which is expressed in mineralized tissues (Golub and Boesze-Battaglia, 2007), respectively. Second, qPCR of different osteoblast marker genes, such as RUNX2 and SP7, will be performed. These experiments will be conducted during the course of two weeks post-transduction for observing the transient changes in marker gene expression as well as alkaline phosphatase activity and lipid droplet abundance.

4.4 Reference-based Discretization of Gene Expression Data

Although the utility of the developed methodology for reconstructing condition specific gene regulatory networks has been highlighted by applications to the prediction of candidate instructive factors and drugs that are implicated in reversing disease phenotypes, a significant disadvantage was identified. So far, all presented analyses involved the comparison of only two cellular phenotypes by differential expression analysis. However, this creates an isolated view on the two conditions under study (Hudson *et al.*, 2012) while detecting significant differences, but lacks the possibility of comparing multiple cell types or conditions. For example, when considering the gene expression patterns of three cell types A, B and C, the discretized gene expression of A is dependent on the comparison to either B or C. Therefore, the reconstructed network is dependent on B or C, as well, and does not allow for the reconstruction of more than two condition specific networks with mutually consistent phenotypes.

Unlike differential expression analysis, other methods have been developed for providing discretized values consistent with multiple conditions. These pseudo-global approaches rely on the combined analysis of multiple time points, cell types or

Table 4.6. Methods for comparison against RefBool

Method	Discretization	Statistical significance	Reference
BiKmeans	Ternary	No	(Li *et al.*, 2010)
BascA	Binary	Per gene	(Müssel *et al.*, 2016)
BascB	Binary	Per gene	(Müssel *et al.*, 2016)
kMeans (Basc toolkit)	Ternary	Per gene	(Müssel *et al.*, 2016)
TascA	Ternary	Per gene	(Müssel *et al.*, 2016)
EFD	Binary	No	(Catlett, 1991; Dougherty *et al.*, 1995; Kerber, 1992)
EWD	Binary	No	(Catlett, 1991; Dougherty *et al.*, 1995; Kerber, 1992)
Gallo et al.	Binary	No	(Gallo *et al.*, 2011)
kMeans	Ternary	No	(Macqueen, 1967)
MeanPlusEstDev	Ternary	No	(Madeira and Oliveira, 2005)

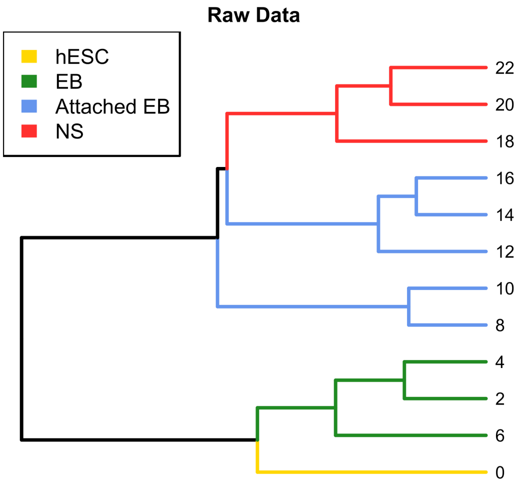
Current methods for discretizing transcriptomics data. These approaches either binarize expression values or include a third state corresponding to intermediate expression (ternary discretization). Only the most recent developments of the Basc-Toolkit (including BascA, BascB, kMeans (basc) and TascA) assign statistical significance scores to the obtained discretizations.

conditions. However, all of these methods have certain limitations such as the lack of statistical support or the sensitivity to changes in the analyzed datasets. In order to overcome the distinct limitations of differential expression analysis and pseudo-global approaches, *RefBool* was developed. Based on a background distribution of gene expression profiles, it classifies expression values into active and inactive states and provides the statistical significance for each classification. This chapter details the validation of *RefBool* against 10 other discretizaton approaches on the basis of a qualitative and quantitative analysis (see Table 4.6 for an overview of the different methods). A dataset of 675 RNA-seq samples of cancer cell lines served as the background distribution ((Klijn *et al.*, 2014), Table S3). Time-series RNA-seq measurements of 12 equally-spaced timepoints from neuroepithelial differentiation experiments were selected as a validation dataset, which is comprised of 17088 genes for which *RefBool* contains a background distribution ((Qiao *et al.*, 2015), Table S3). This dataset reflects the induced differentiation of human embryonic stem cells and undergoes distinct stages that can be subsumed in three categories, embryonic bodies (EBs), embryonic bodies attached to the cultureware (attached EBs) and neuronal sphere (NS).

4.4.1 Qualitative analysis of discretized neuroepithelial differentiation measurements

Despite the sole separation of continuous gene expression measurements into active and inactive states, discretization methods should still preserve descriptive characteristics of the dataset. One important characteristic of time series measurements is the order of samples that plays an important role to determine causal changes in the regulation of genes. In the differentiation experiment of embryonic stem cells into neuroepithelial cells, distinct stages of the development become evident when clustering the samples hierarchically so that time-points with similar transcriptomics profile are close to each other (Figure 4.11).

Figure 4.11. Hierarchical clustering of raw expression

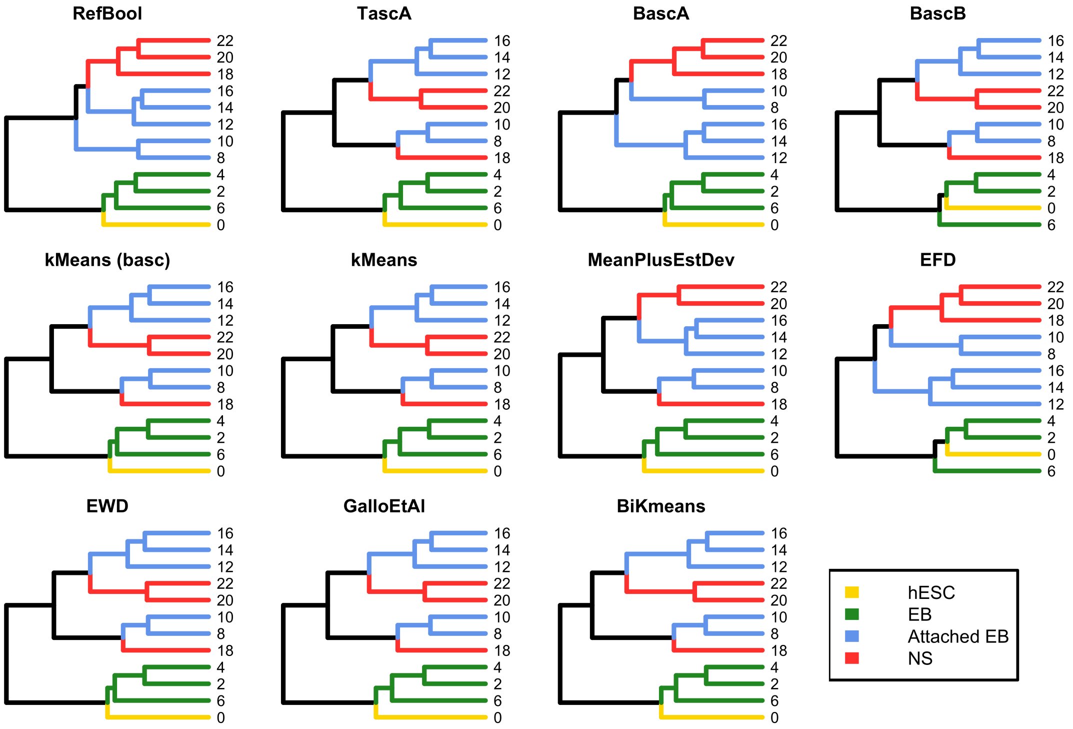


Hierarchical clustering of gene expression data resembles distinct developmental stages in the differentiation of human embryonic stem cells into neuroepithelial cells. (hESC: human embryonic stem cells; EB: embryonic bodies; attached EB: embryonic bodies attached to feeder-free medium; NS: neural sphere corresponding to neuroepithelial cells)

After discretization of the whole dataset with RefBool and all other methods, hierarchical clustering was performed using the same criterion as for the continuous expression data (Figure 4.12). Evidently, *RefBool* is the only approach that is able to resemble the clustering of the continuous data and correctly groups the different stages during the differentiation. The most common clustering error is the relation of the early neuronal sphere stage (day 18) to early attached embryonic bodies (days 8 and 10) by TascA, BascB, kMeans (basc), kMeans, MeanPlusEstDev, EWD, GalloEtAl and BiKmeans. Less common errors constitute the incorrect distinction of embryonic bodies from embryonic stem cells by BascB and EFD or the relation of early attached embryonic bodies (days 8 and 10) to neuronal sphere cells (days 18, 20 and 22) by BascA and EFD.

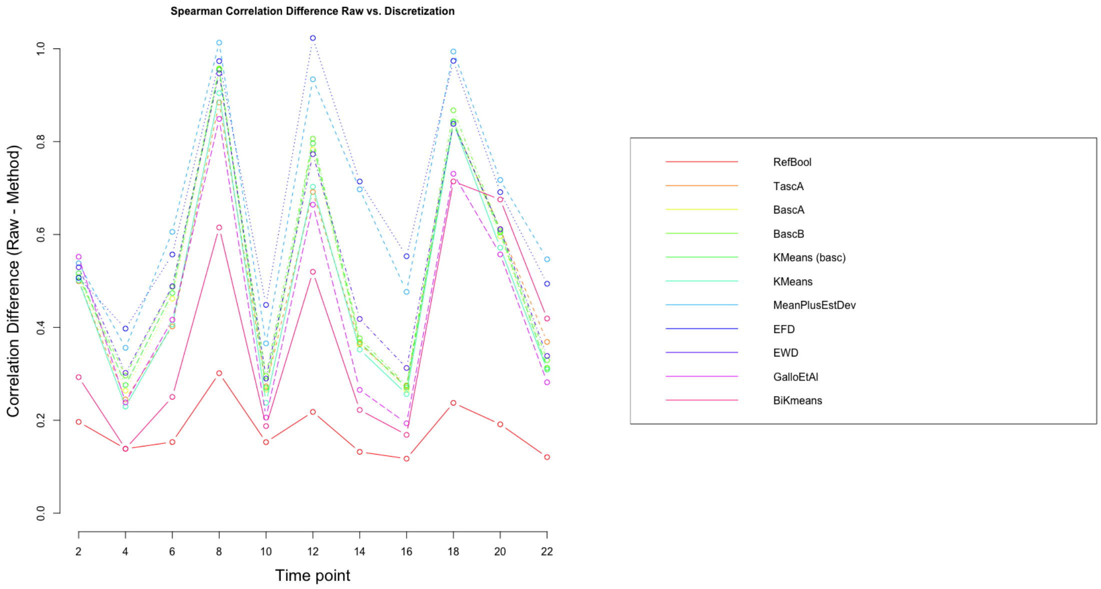
RefBool's superior descriptive ability is further supported by the analysis of the correlation differences between discretized and continuous expression in successive samples. Of note, spearman correlation was employed for this assessment due to its applicability to ordinal numbers. On average, the correlation differences between the continuous and discretized data by *RefBool* amounts to only 0.17 whereas the differences of all other methods amount to at least 0.38 (Figure 4.13). The significantly increased concordance between successive timepoints thus provides a plausible explanation of the higher clustering resemblance.

Figure 4.12. Hierarchical clustering of discretized data



Hierarchical clustering of discretized gene expression data provided by different methods shows that only *RefBool* is able to accurately recapitulate the ordering of the raw data. Most other methods wrongly assign the early neuronal sphere stage (day 18) to the early attached embryoid body state (days 8 and 10). (hESC: human embryonic stem cells; EB: embryonic bodies; attached EB: embryonic bodies attached to feeder-free medium; NS: neural sphere corresponding to neuroepithelial cells)

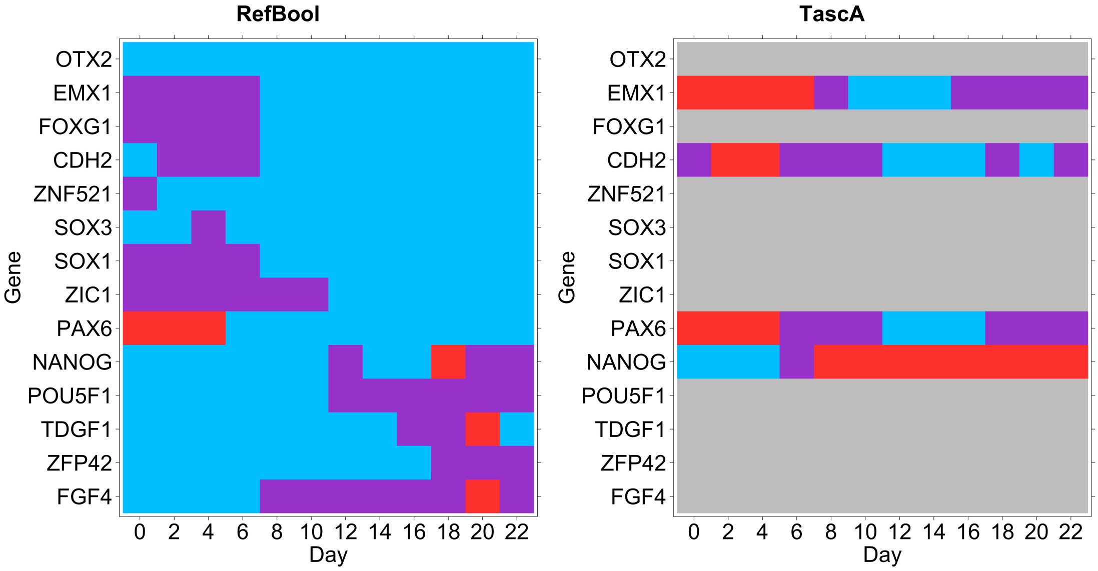
Figure 4.13. Correlation differences between continuous and discretized values of successive time points



Correlation differences of discretized and raw data of adjacent time points show that *RefBool* most accurately resembles the trends in gene expression data. All other methods have similar error trends. This indicates a more plausible temporal ordering of successive time points.

Due to the fact that *RefBool* was developed for discretizing gene expression measurements, it is crucial to assess its ability to accurately identify active genes in different conditions or cell types. In this regard, the study of known marker genes of the different stages in the neuroepithelial differentiation dataset is of particular importance. The considered genes in this assessment are inner cell mass markers (FGF4, ZFP42, TDGF1, POU5F1 and NANOG) that should be active during the first six days of the differentiation as well as neuroectodermal (PAX6, ZIC1, SOX1, SOX3, ZNF521 and CDH2) and forebrain markers (FOXG1, EMX1 and OTX2) that should become activated at later stages of the differentiation. The TascA method was selected for this comparison, since it is the most recent advancement in the field of gene expression discretization.

Figure 4.14. Comparison of RefBool and TascA on known marker genes



Comparison of discretized marker genes obtained by *RefBool* and TascA (red: inactive; violet: intermediate expression; blue: active, grey: insignificant). TascA only obtains significant discretization for EMX1, CDH2, PAX6 and NANOG while determining all other genes insignificant (grey). In contrast, discretizations from *RefBool* indicate a patterning consistent with the marker genes.

Similar to the results of the cluster resemblance analysis, *RefBool* and TascA show significant differences in the discretization of marker genes (Figure 4.14). First and foremost, TascA is only able to obtain significant discretizations for four of the fourteen genes (false discovery rate cutoff: 0.05) while *RefBool* classifies most of the expression values in different conditions to be either active or inactive (false discovery rate cutoff: 0.05). The inner cell mass markers NANOG, POU5F1, TDGF1, ZFP42 and FGF4 are constitutively active throughout the first six days of the differentiation and are subsequently gradually down-regulated. On the other hand, the activity of OTX2 throughout the whole differentiation process is seemingly unexpected owing to its role as a forebrain marker. However, an extensive study of the role of OTX2 in embryonic stem cell determination validated its expression and necessity in ESC maintenance and embryoid body formation (Acampora *et al.*, 2013). Similarly, SOX3 and CDH2 were found to play a fundamental role in maintaining pluripotency (Pieters and van Roy, 2014; Abdelalim *et al.*, 2014) and are therefore likely to be correctly classified in the embryonic stem cell state. Strikingly, the early activation of ZNF521 identified by

RefBool was shown to be essential and sufficient for neural differentiation in mice (Pieters and van Roy, 2014). The remaining forebrain and neuroectodermal marker genes were found to be active after six days of differentiation and were, thus, correctly classified.

TascA, in contrast, provides significant discretizations for only four genes. Recall that TascA assesses the significance of the discretization per gene while *RefBool* performs more fine-grained significance analysis per expression value. Most importantly, TascA does not classify PAX6 and EMX1 as constitutively active at the later stages of differentiation but identifies a rather dispersed pattern. Similarly, the functional importance of CDH2 in embryonic stem cells is not captured, as it is not classified to be active. The expression of NANOG seems to be the only reasonably discretized pattern being active in embryonic stem cells and embryoid bodies and inactive (down-regulated) during the formation of neuroepithelial cells.

Overall, *RefBool* shows clear advantage in the qualitative description of phenotypes with respect to the accurate recapitulation of developmental processes and the discretization of known marker genes. Especially the dynamic expression of important genes for maintenance of embryonic stem cells is accurately captured and overall highlights the descriptive power of the approach even when applying strict significance thresholds.

4.4.2 Quantitative Assessment of Discretization with RefBool

Apart from the increased descriptive ability of *RefBool,* it is important to analyze and compare the discretized gene expression profiles quantitatively against the results of other methods. There exist two different categories for validating cluster consistency, external and internal clustering indices. External clustering indices rely on additional information that validates the correct clustering while internal indices quantify the clustering consistency based on features derived from the original dataset. Due to the absence of a ground truth for determining whether genes are active or inactive, the following analysis quantifies consistency based on internal indices. More specifically, the internal clustering indices utilized in this assessment are Ksq_DetW (Marriott, 1971), Ray-Turi (Ray and Turi, 1999), SD (Halkidi *et al.*, 2001), Trace_W (Edwards and Cavalli-Sforza, 1965), Trace_WiB (Friedman and Rubin, 1967), Wemmert Gancarski (WEMMERT *et al.*, 2000) and Xie-Beni (Xie and Beni, 1991), which are implemented in the clusterCrit R-package (Desgraupes, 2013). All of these metrics assess either the within-cluster scattering, the between-cluster separation or a weighted combination of both. An overview of how the different indices can be compared, i.e. whether the minimum or maximum value constitutes the best clustering, can be found in Table 4.7.

Table 4.7. Rules for determining best clustering

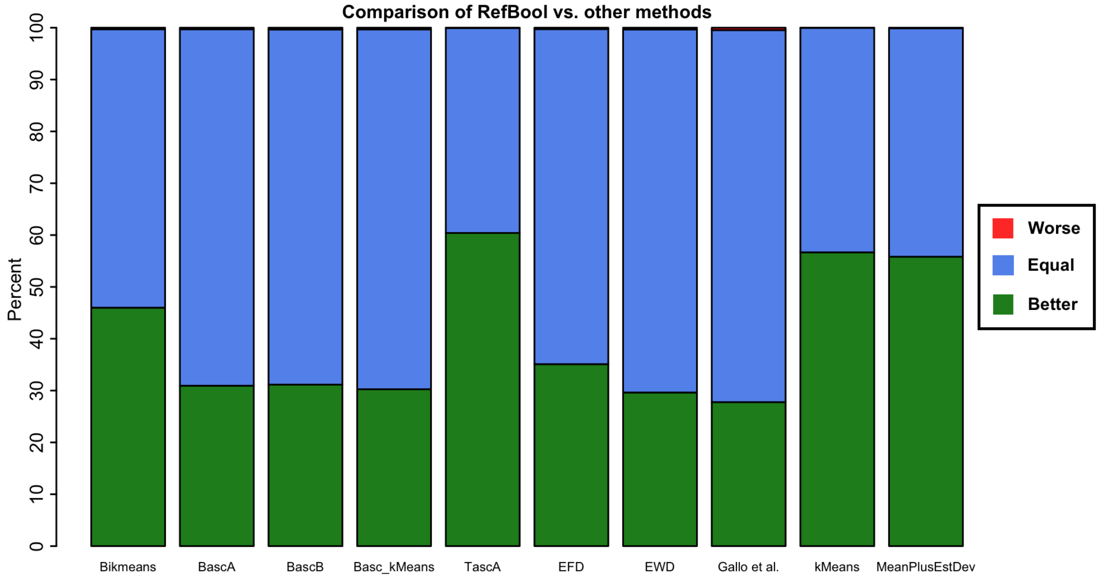
Index	Rule for best clustering
Ksq_DetW	Max
Ray-Turi	Min
SD	Min
Trace_W	Max
Trace_WiB	Max
Wemmert Gancarski	Max
Xie-Beni	Min

Each cluster consistency index used for validating the discretized datasets is assigned a rule for comparing different values. 'Min' means that the minimum value is the superior while 'Max' shows that the maximum consistency value is regarded as the best score.

All clustering indices were calculated on a per gene basis for the discretizations obtained by *RefBool* and the other approaches previously presented in the qualitative comparison. In particular, the clustering indices are calculated for each gene separately on the basis of the neuroepithelial RNA-seq dataset. *RefBool* is then said to perform better (worse) in clustering a gene if more clustering indices are better (worse) with respect to the rules in Table 4.7. Importantly, some indices might not be able to assess the quality due to the composition of the expression values for each gene and are therefore excluded from the comparison. In case of ties in the number of superior and inferior clustering indices, a third category indicating the equality of the clustering will be described. The results of this assessment are summarized in Figure 4.15. It is evident that *RefBool* outperforms all other methods and classifies at least 27.7% of genes more consistently than other methods (mean: 40.4%). On the other hand, on average, the discretization of only 44 genes (0.26%) is worse when compared to existing approaches. Of note, the performance against TascA, the most recently introduced method for discretizing gene expression data, is highly elevated and results in 60.4% higher cluster consistency.

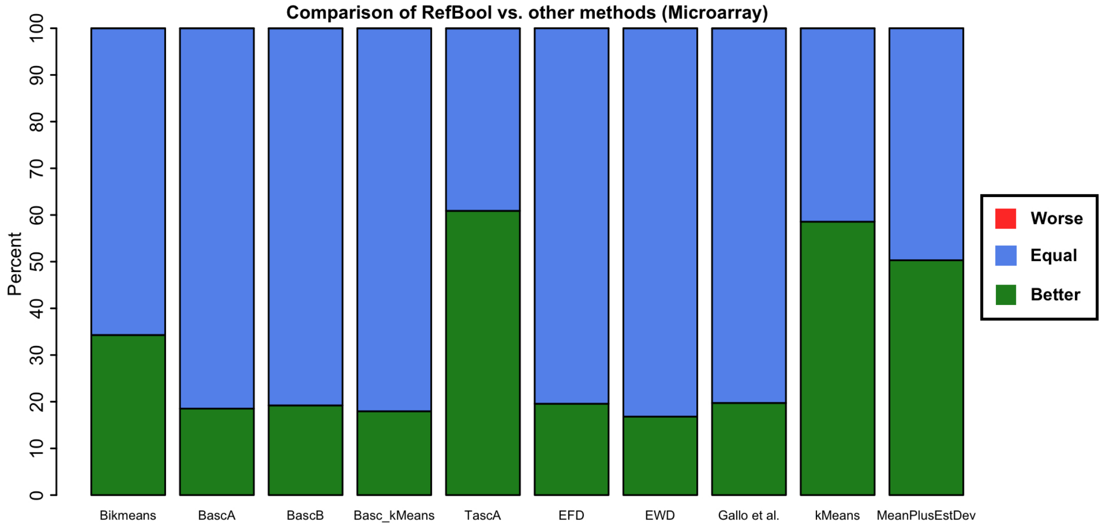
In order to support the increased accuracy of *RefBool* and to investigate the generalizability of the approach to other experimental transcriptional profiling techniques, the same quantitative assessment was performed for microarray expression data. A compiled dataset of 27887 microarray samples ((Torrente *et al.*, 2016), Table S3), which comprises 47000 probesets, served as the basis of the previously detailed analysis. The first ten samples of this set were used for discretization and cluster consistency assessments while the remaining samples built the reference distribution for *RefBool*. Evidently, the obtained results substantiate the improvements of the developed reference-based discretization (Figure 4.16). However, even though *RefBool* performs similarly against TascA, having 60.8% more consistent clusterings, the improvements against other methods decreased compared to RNA-seq data. Better discretizations are obtained in at least 16.8% of probesets (mean: 31.6%) compared to only 0.01% less consistent clusterings totaling 5 out of 47000 probesets.

Figure 4.15. Quantitative comparison of RefBool against other methods in clustering RNA-seq data



Quantitative comparison of *RefBool* against other state-of-the-art methods on the basis of internal clustering indices. The assessment was conducted on 17088 genes in 12 RNA-seq measurements of a neuroepithelial differentiation study. Evidently, *RefBool* provides the most consistent clustering with, on average, 40.4% better (green) and 0.26% worse (red) discretizations per gene. All other genes resulted in an equal number of superior and inferior consistency scores (blue).

Figure 4.16. Clustering consistency assessment on microarray data.



Quantitative comparison of *RefBool* against other state-of-the-art methods on the basis of internal clustering indices for 47000 microarray probesets in 27887 samples. Evidently, *RefBool* provides the most consistent clustering with, on average, 31.6% better (green) and 0.01% worse (red) discretizations per gene. The remaining genes showed no differences in the number of superior and inferior clustering indices (blue).

Overall, the developed reference-based discretization approach provides better qualitative and quantitative results than other currently used state-of-the-art methods for classifying genes as active or inactive based on transcriptomics data. Especially the predominantly correct classification of biologically relevant marker genes in the context of neuroepithelial differentiation supports the application in the context of gene regulatory network inference.

4.4 Chapter Summary

This chapter described the results of the proposed Boolean condition-specific gene regulatory network inference method, including a strategy for identifying candidate instructive factors for cellular conversions and the selection of multitarget combinations thereof. The strategy was applied to disease-control case studies of Systemic Lupus and Rheumatoid Arthritis where potential therapeutic compounds were selected based on their reported effects in multitarget combinations. Validation of the reconstructed networks with transcription factor binding site ChIP-seq data revealed an elevated condition specificity in comparison to other state-of-the-art methodologies by retaining most of the experimentally determined regulatory interactions.

To overcome the widespread absence of condition specific epigenetic information, a machine-learning approach for predicting gene-level chromatin accessibility from transcriptomics data was developed that can be readily integrated into the Boolean gene regulatory network framework. A thorough comparison against traditional computational downstream analysis tools for chromatin accessibility assays supports the reliability of the predictions. In addition, the predicted gene-level accessibility can be utilized for obtaining more reliable peak calling parameters by increasing the number of correctly identified accessible chromatin regions.

With it, a multitarget combination of candidate instructive factors for inducing the cellular conversion of adipocytes into osteoblasts was derived from an integrative network reconstruction and analysis approach involving transcriptomics, epigenetics and prior knowledge networks.

Finally, *RefBool* was developed, a reference-based method for discretizing gene expression data that does not require multiple replicates of the same condition and provides statistical confidence to each discretized expression value. In contrast to differential expression analysis, *RefBool* provides an absolute discretization independent of the condition it is compared to, thus allowing for the comparison of multiple cell types or conditions. Moreover, quatitative and qualitative analyses highlight the increased performance in comparison to current methodologies, which is underlined by an accurate resemblance of marker gene expression in induced neuroepithelial differentiation.

In the Discussion chapter these methods and their results are considered in the context of the challenges in gene regulatory network inference that complicate the identification of instructive factors for cellular conversions. Also, the particular advantages and disadvantages of the different methodologies are outlined, and future directions are pinpointed to provide more realistic gene regulatory networks and ultimately more efficient combinations of instructive factors.

CHAPTER 5 Discussion

Cellular reprogramming is perhaps the most promising strategy towards patient-specific therapeutic interventions for treating disease pathologies. After the groundbreaking finding that the cellular conversion of somatic cells into pluripotent stem cells can be induced by ectopic expression of only four transcription factors (Yamanaka factors), great efforts have been devoted to understand the underlying gene regulatory network (Li and Belmonte, 2017). As a result, a complex circuitry of intertwined transcriptional and epigenetic regulation was identified that orchestrates the activation of the pluripotency program of cells. In particular, two of the four identified transcription factors, i.e. SOX2 and OCT4, are mutually regulating themselves (Kashyap *et al.*, 2009) for keeping up their expression in an equilibrium state while initiating the rearrangement of the global chromatin conformation (Soufi *et al.*, 2012). Cooperation of transcription factors eventually elicit the activation of genes necessary and sufficient for establishing and maintaining pluripotency (Soufi *et al.*, 2015).

Besides the instructive factors for inducing pluripotency, other combinations of genes have been identified and implicated in the direct cellular conversion of cell types belonging to distinct lineages (Ieda *et al.*, 2010; Szabo *et al.*, 2010; Son *et al.*, 2011). However, due to the limited understanding of the underlying the gene regulatory network and the dynamic formation of accessible and inaccessible chromatin regions, which eventually enable or inhibit transcription, current screening strategies for novel combinations of instructive factors are time consuming and expensive. In particular, the most commonly followed strategy is a combinatorial approach with which the Yamanaka factors were identified (Takahashi and Yamanaka, 2016). Starting with a list of candidate transcription factors or microRNAs that are assumed to regulate the cell fate or are involved in differentiation, the list is gradually decreased until a minimal set of factors is identified that induce the desired transition.

The development of computational tools constitutes a promising alternative to the above-mentioned approach and enables the a priori *in silico* evaluation of candidate instructive factors. With such computational approaches, both screening times and experimental costs could be significantly reduced by preselecting the most reassuring combinations. Due to the availability of high throughput sequencing technologies for quantifying RNA abundance and the resulting ability of integrating transcriptional information from large amounts of diverse cell types, this area of research has gained increasing attention.

Of these developed methods for identifying instructive factors of cellular conversions, network based approaches are of special interest, since they offer the possibility of understanding the regulatory mechanisms leading to stabilization and destabilization of cellular phenotypes. However, current methodologies suffer from certain important limitations. First, the input to those methods is mostly static and, thus, cannot be customized towards new, unseen experimental data (Rackham *et al.*, 2016) or, second, demand substantial amounts of data for tailoring it to the desired cell types (Cahan *et al.*, 2014). At last, network models represent different cell types or conditions within the same topology (Crespo *et al.*, 2013; Cahan *et al.*, 2014) even though experimental evidence verified the existence of differential regulatory interactions (Song *et al.*, 2011). Although it is difficult to assess the correspondence of reconstructed gene regulatory networks with the real regulatory circuitry, these methods might misrepresent the regulatory interactions of at least one phenotype.

The differential network analysis approach presented in this thesis is justified by the large-scale profiling of epigenetic data, such as chromatin accessibility and covalent histone modifications, that revealed the structural differences of active regulatory elements in diverse cell types (Ernst and Kellis, 2012; Song *et al.*, 2011). It is, therefore, rational to attempt the integration of an epigenetic layer into the gene regulatory network for explaining the differential regulatory mechanisms modulating distinct phenotypes. With current technology of epigenetic and transcriptional profiling as well as transcription factor target identification, the reconstruction of cell type or condition specific networks became feasible. An accurate inference of the gene regulatory networks seems eventually necessary for enhancing the understanding of the complex transcriptional landscape to identify important genes whose deregulation induces a desired cellular conversion.

5.1 Cell-type specific network reconstruction and differential network analysis

The differential network analysis and reconstruction approach proposed in this thesis accounted for the differences of cellular gene regulatory network topologies between different cell types and addressed two distinct aspects. First, the reconstruction of cell type specific networks considers the regulatory differences in the network topology that might be responsible for stabilizing the cellular phenotype. Second, the premise for the approach is the general applicability to diverse cell types and cellular conditions. Thus, it only requires a single transcriptional readout for the initial and final cell type as well as a prior knowledge network obtained from proprietary or public databases. Modeling gene regulation within the Boolean framework enabled the joint realization of these goals. While Boolean networks are a rather coarse approximation of the regulatory behavior, their utility in describing cellular phenotypes and transitions upon perturbation has been highlighted previously (Albert and Othmer, 2003).

Besides the topological insights obtained from differential network analysis, it enabled the identification of candidate instructive factors for inducing desired cellular conversions. In the context of chemically induced alterations of the gene regulatory networks, important genes have been pinpointed whose perturbation is prone to induce the phenotype after treatment. In the context of drug prioritization, the phenotypic changes induced by Cobalt chloride include the down-regulation of two reported pioneer factors, ASCL1, SOX2 (Vierbuchen *et al.*, 2010; Soufi *et al.*, 2012). As a consequence, large-scale epigenetic changes can be expected that in turn influence the transcriptional regulation of other genes in the network. Even without including epigenetic information in this analysis, the developed differential network analysis methodology was able to pinpoint these genes as candidates for inducing the observed changes. The accuracy in recognizing important stability determinants is further underlined by the proposed prioritization scheme for selecting compounds inducing a desired cellular transition. In all examples of induced changes with known compounds, the drug target enrichment in *in silico* perturbations was shown to be predictive of the correct compound. The sensitivity of this analysis has been underlined by the ability to correctly predict the used compounds while showing comparable enrichment results for drugs inducing similar phenotypes, i.e. the drugs are predicted to have similar functions. It further provides support for the qualitative ability of the network to prioritize combinations of candidate instructive factors that are able to induce a desired cellular transition. In particular, the enrichment of drug effects in simulated multitarget combinations solely relies on the simulated effects of perturbations to the network attractor. Therefore, a correct drug prioritization requires the network to identify essential regulators and reflect their importance in the network dynamics.

After validating the differential network analysis pipeline for prioritizing drugs according to their expected efficacy based on known drug effects, its application to Rheumatoid Arthritis and Systemic Lupus Erythematodes, two autoimmune diseases, provided continued support for the sensitivity of the approach. In both cases, current state-of-the-art therapeutics were predicted alongside potentially novel candidates. However, since the method is only based on previously reported phenotypic changes induced by the enriched drugs, their efficacy cannot be concluded. Other important aspect such as toxicity, concentration and susceptibility of the cells for uptake are disregarded. Therefore, this approach solely enables the prioritization of promising drugs for further screening. Nevertheless, the results of the analysis suggest that the developed analysis approach constitutes a helpful tool for reducing the number of screened drugs significantly.

The particular advantages and limitations of the developed differential network reconstruction methodology, which constitutes the basis of the drug prioritization strategy, are discussed in the remainder of this section.

5.1.1 Advantages of this approach

The presented approach offers several advantages in the identification of combinations of instructive factors for inducing cellular transitions. First, by using available biological knowledge about gene regulatory interactions in form of prior knowledge networks, an appropriate search space for the network reconstruction process can be defined. The wealth of manually curated, functional regulatory interactions that have been obtained in the past through a combination of DNA binding experiments (ChIP-seq) and gene knockouts have, thus far, not been systematically exploited in the reconstruction of gene regulatory networks. However, it constitutes an indispensable resource in the search for accurate network models, particularly due to the infeasibility of large scale *de novo* network. For instance, the approximately 1700 known or predicted transcription factors (Zhang *et al.*, 2015) alone could, in theory, establish 2,890,000 regulatory interactions among themselves, which would require vast amounts of heterogeneous datasets to stratify for actual and unreal interactions. By using public or proprietary databases, like MetaCore™, that already contain manually curated and experimentally validated information about gene regulatory interactions observed in diverse cell or tissue types, the search space can be significantly reduced to a few thousand interactions. Furthermore, networks reconstructed from prior knowledge databases, using the proposed differential network reconstruction methodology, retain most of the cell type specific experimentally validated interactions, which provides support to the chosen strategy. However, it is of note that the utilization of DNA binding experiments for validation can only support the existence of interactions rather than their absence. There is a wealth of reasons why downstream computational tools do not detect certain interactions. On one hand, the antibody used for the experimental assay should show high sensitivity against the target protein while not being cross-reactive with other related protein family members (Kidder *et al.*, 2011). However, this cannot be guaranteed for all proteins. On the other hand, computational methods for detecting regions significantly enriched in reads require the setting of various parameters, such as the false discovery rate (FDR). Generally, lower false discovery rates render less but more certain interactions while higher thresholds result in more false positives (Thomas *et al.*, 2017). Since an optimal parameter setting is not a priori known, the datasets used for validating the condition specificity of the reconstructed networks were processed with low false discovery rates for increasing the certainty in the observed interactions.

Secondly, the differential network analysis greatly helps in pinpointing important network structures, such as regulatory feedback loops and genes under differential regulation, for reducing the search space of candidate instructive factors. Previous approaches mainly relied on the identification of positive feedback loops (Crespo *et al.*, 2013) or hub genes (Cahan *et al.*, 2014) in the network structure. However, the combination of positive and negative feedback loops in Boolean networks could lead to network stabilization (Remy and Ruet, 2008), since they are embedded in a greater network structure controlling them. Furthermore, negative feedback loops have also been reported in real biological networks and occur, for example, in the transcriptional control of P53 (Brown *et al.*, 2009). Therefore, the perturbation of negative feedback loops in the reconstructed networks has to be considered for identifying candidate instructive factors. On the other hand, the proposed context specific gene regulatory network reconstruction allows for the selection of genes under differential regulation, i.e. genes whose set of regulators differ in the two networks. These genes constitute important perturbation targets, as their discretized expression cannot be explained with one network topology. Support for this assertion is provided by the selected candidate instructive factors for reverting the disease phenotypes of Systemic Lupus Erythematodes and Rheumatoid Arthritis. In particular, STAT1, a marker gene for onset and activity of Systemic Lupus (Liang *et al.*, 2014), as well as TCF7L2 and CDKN1A, which are associated to the Rheumatoid Arthritis (Mota *et al.*, 2012; Perlman *et al.*, 2003), have been identified as differentially regulated.

Lastly, the proposed network reconstruction approach is able to readily integrate gene-level accessibility for removing genes in inaccessible chromatin domains, which cannot be regulated and cannot regulate other genes. Thus, the context specificity of the networks is expected to increase. In contrast, other methodologies modeling multiple phenotypes within a single network topology do not allow for the integration of cell type specific information, since it requires distinct alterations to the network topology.

5.1.2 Limitations

Despite the advantages of the developed methodology, network reconstruction within a Boolean framework has certain limitations. For instance, it requires the choice of logic rules determining when a gene is expressed given the state of its regulators. In living systems, these logic rules are gene-specific and are determined by several aspects. For example, binding sites of different regulators that are sufficiently close to each other are likely to be mutually exclusive, i.e. only one regulator can bind. However, in contrast to this complexity, the reconstructed networks in this thesis simplified this notion by using a majority rule, which determines the gene's state based on whether most regulators are active or inactive. Even though cellular gene regulatory networks possess a robust, redundant design (Macneil and Walhout, 2011), which supports the choice of a majority rule, and all regulatory interactions in the prior knowledge networks have been experimentally validated, the choice of the same logic rule for each gene constitutes a simplification. However, this issue is not unique to Boolean modeling but pertains to all logic modeling frameworks such as ordinary differential equations. This leaves it an important direction for future research. However, due to the conferred robustness and redundancy of the majority rule, it constitutes a reasonable assumption for reconstructing gene regulatory networks.

Another important aspect constitutes the utilization of available biological knowledge for reconstructing gene regulatory networks. On one hand, this approach limits the search space to a tractable amount of potential interactions, as described earlier in this chapter, which is an important advantage over methods without prior knowledge. On the other hand, however, the impossibility of *de novo* interaction identification creates a significant dependence of the networks on the knowledge base. Therefore, the utilization of a reliable database is essential for accurate network reconstruction. Previous approaches relying on prior knowledge networks of regulatory interactions use ChIP-seq derived interactions whose mode of action is undetermined (CellNet (Cahan *et al.*, 2014)) or text-mining based approaches such as the ResNet mammalian database (Crespo *et al.*, 2013). In contrast, MetaCore™, which has been used throughout this thesis, contains manually curated interactions whose mode of action was predominantly experimentally validated and as such constitutes a valuable and accurate resource of prior knowledge networks.

Finally, the proposed approach for integrating epigenetics information currently only involves the static removal of genes in inaccessible chromatin domains. However, such a network cannot explain the dynamic changes that must be induced during the desired cellular transitions. Nonetheless, the epigenetic information provides additional support for the selection of stability determinants and differentially regulated genes during network analysis and, ultimately, the deduction of candidate instructive factors.

5.2 Prediction of gene-level chromatin accessibility

The accessibility landscape of cells plays an important role in the transcriptional regulation of genes, since active regulatory elements, such as enhancers and promoters, were identified to be located in accessible chromatin domains (A P Boyle *et al.*, 2008; Song *et al.*, 2011; Thurman *et al.*, 2012). In order to identify these active regulatory elements, several experimental assays have been utilized, e.g. DNase-seq, FAIRE-seq and ATAC-seq (Song *et al.*, 2011; Buenrostro *et al.*, 2013), and required the development of downstream computational tools, i.e. peak callers, for detecting genomic regions significantly enriched in aligned reads. However, a previous study comparing the most widely used peak calling methodologies, Hotspot (John *et al.*, 2011), MACS (Zhang *et al.*, 2008), F-Seq (Alan P Boyle *et al.*, 2008) and ZINBA (Rashid *et al.*, 2011), found that the overlap of detected peaks amounts to only 11% (Koohy *et al.*, 2014). In addition, the reliability of these methodologies is heavily dependent on the particular parameter settings, which are usually not known a priori (Koohy *et al.*, 2014). In contrast to these methodological issues, the availability of chromatin accessibility assays is still limited for specialized cell types or conditions.

In order to address these limitations, a machine-learning framework has been developed that predicts chromatin accessibility based on transcriptomics data. Importantly, instead of predicting the actual enriched genomic regions, the method aims at identifying the gene-level accessibility, i.e. whether a gene is located in accessible chromatin domains, for readily integrating the information in the proposed Boolean network model. Since machine-learning approaches rely on the identification of patterns in known datasets, the use of computationally processed accessible chromatin assays was unavoidable, despite their known limitations. The choice for experimental DNase-seq datasets that were processed with current downstream computational tools relied on two considerations. First, DNase-seq data shows a higher signal-to-noise ratio compared to FAIRE-seq (Tsompana and Buck, 2014), which makes the identification of peaks more reliable. Second, Hotspot shows the least sensitivity to its parameter settings (Koohy *et al.*, 2014) and was, thus, chosen for the identification of DNase hypersensitive sites.

Based on cell type specific gold standard datasets including transcription factor binding site ChIP-seq and heterochromatin regions defined by a combination of three histone modifications, H3K4me3, H3K9me3 and H3K36me3, the conceived framework was evaluated. Importantly, the definition of heterochromatin regions in these datasets does not include polycomb repressed sites, defined by H3K27ac, which are usually considered to be epigenetically silenced. An integrative analysis of 111 human epigenomes revealed that polycomb repressed regions are moderately enriched in DNase-seq reads and, thus, might contain DNase hypersensitives sites reflecting accessible chromatin (Roadmap Epigenomics *et al.*, 2015). Therefore, polycomb repressed regions are not considered for the composition of the gold standard datasets. This implies that polycomb repressed sites are prone to misclassifications, due to the co-occurrence of DNase hypersensitive sites and epigenetic silencing in certain cases.

The particular advantages and limitations of the developed machine-learning framework will be discussed in the following.

5.2.1 Advantages of the chromatin accessibility prediction

The evident advantage of the proposed machine-learning framework is its ability to provide gene-level accessibility assignments of genes in the absence of experimental chromatin accessibility assays. Validation against F-Seq, MACS and Hotspot in six gold standard datasets suggested a comparable accuracy of the predictions without predictive biases to gene types or the GC content of genes. Especially genes more expressed than 0.08 FPKM were significantly more accurately classified by the developed methodology. This expression level is significantly lower than typical thresholds for defining expressed genes (Haltaufderhyde and Oancea, 2014; Rau *et al.*, 2013; Trakhtenberg *et al.*, 2016) such that the increased accuracy cannot be explained by the trivial assignment of expressed genes to accessible chromatin domains.

Furthermore, the method is able to identify the optimal set of peak calling parameters in the presence of experimental data. Importantly, the optimization of peak calling parameters does not require prior knowledge about certain accessible or inaccessible genomic regions, but solely relies on the comparison to the predictions. Of note, the applicability of the proposed recall measure together with the results in the six gold standard examples suggests that peak callers typically impose too stringent false discovery rates, which is in accordance with previous results, suggesting that acceptable true positive rates require less stringent false discovery rates (Koohy *et al.*, 2014).

Besides the prediction of accessible chromatin from transcriptomics data alone, the design of the stacked classification tree model enables the integration of other predictors. In particular, the integration of chromatin conformation capture data from, for example, Hi-C experiments (Lieberman-Aiden *et al.*, 2009) could further increase the predictive power of the methodology by providing information about spatial organization of chromatin.

In the context of Boolean network reconstruction, the prediction of whether genes are located in accessible or inaccessible chromatin domains allows for a more precise identification of the regulatory interactions during network reconstruction. More specifically, it enables the distinction of genes that are transcriptionally repressed from those that are epigenetically repressed and is expected to increase the condition specificity of the reconstructed networks.

5.2.2 Limitations

The analysis of the predictive accuracy revealed that the performance of the developed machine-learning approach is reduced when considering all genes contained in the gold standard datasets. More specifically, non-expressed genes (0 FPKM) cannot be reliably determined. Since only transcriptomics data is utilized for the predictions and due to the deterministic behavior of the stacked classification tree model, all of these genes are classified as being inaccessible. In particular, deterministic machine-learning approaches relate the same input gene expression with the same chromatin accessibility assignment such that this limitation can only be overcome by integrating additional predictors such as chromatin conformation capture experiments. As already mentioned in the last section, the design of the proposed methodology enables the integration of additional experimental data as a predictor for chromatin accessibility, which could address this issue. However, its use for predicting gene-level accessibility of the differentially expressed transcription factors in adipocytes and osteoblasts is not impeded. In particular, the only differentially expressed transcription factor showing no expression is TCF21, which has no reported interaction with any other differentially expressed TF and is, thus, not included in the prior knowledge network.

Finally, it is worth noting that the developed methodology cannot be readily extended to predict active regulatory regions, as defined by peak callers, from transcriptomics data. Gene expression measurements from, for example, RNA-seq experiments do not provide the sufficient resolution for pinpointing regions significantly enriched in aligned reads from chromatin accessibility assays. However, the proposed approach is able to assist in a more reliable definition of the accessible chromatin landscape by optimizing the parameters of current peak calling methodologies. Nevertheless, this limitation does not relate to the primary aim of predicting gene-level accessibility for integration into the Boolean network formalism.

5.3 Reference-based discretization of transcriptomics data

Computational processing of experimentally derived gene expression data is vital for drawing conclusions about transcriptional differences among cells or conditions. Numerous biological insights have been obtained from the wealth of data that has been generated from microarray and RNA-seq experiments. Most notably, the derivation of cell type and condition specific gene signatures allowed for the identification of enriched pathways and biological process (Palmer *et al.*, 2006; Ashburner *et al.*, 2000; Ko *et al.*, 2013). Great efforts have been devoted to the development of more accurate methodologies for carrying out enrichment analysis relying on previously defined gene signatures (Subramanian *et al.*, 2005; Huang *et al.*, 2009; Eden *et al.*, 2009) while the necessity for the concurrent development of tools producing these signatures has been largely disregarded. Like the development of gene signatures, discretization of gene expression data is vital for the logic modeling of gene regulation with Boolean networks.

To date, the most widely used methods for discretizing gene expression data are based on differential expression analysis. Generally speaking, these approaches aim at identifying significantly different genes in two conditions based on location shifts in the expression distributions and, thus, create an isolated view of the two conditions under study (Hudson *et al.*, 2012). However, the presence of more than two conditions requires their pairwise comparison and cannot be studied as whole. This is caused by the lack of transitivity of differential expression analysis and can be illustrated by a simple scenario. Given three conditions A, B and C, if a gene is down-regulated when comparing A to B and up-regulated in the comparison of B to C, no conclusion can be drawn about its status in the comparison of A to C.

Several approaches have been developed and applied to allow for the consistent discretization of multiple gene expression profiles (Catlett, 1991; Dougherty *et al.*, 1995; Gallo *et al.*, 2011; Madeira and Oliveira, 2005; Müssel *et al.*, 2016). However, all of these methods possess important limitations. First, most of the approaches derive a single gene expression threshold and classify genes to be active or inactive depending on whether their expression is greater or lower than the threshold, respectively, regardless of the statistical support (Macqueen, 1967; Catlett, 1991; Gallo *et al.*, 2011; Madeira and Oliveira, 2005). More recent advances included the assessment of the statistical significance on the separation of active and inactive genes (Müssel *et al.*, 2016), but are limited in their application. In particular, the combined analysis of gene expression datasets from different cell types or unequally spaced time-series datasets results in unquantifiable biases of the results. Generally, all developed approaches share the limitation of being sensitive to changes in the datasets used for discretization, since they do not take into account the distinct expression distributions of genes across different cell types.

The developed methodology, *RefBool*, addresses these issues and makes actively use of already available transcriptomics data for providing a reference-based discretization of gene expression values. Its particular advantages and limitations will be discussed in the following.

5.3.1 Advantages of *RefBool*

RefBool's main contribution is the integration of biological knowledge in the discretization of gene expression data and considers the distinct expression characteristics of individual genes (Djebali *et al.*, 2012). Through the compilation of a homogenously processed set of transcriptomics samples in distinct cell types, each gene is associated to an individual distribution that eventually enables the discretization of expression values without requiring gene expression replicates. In addition, rather than assigning no or per gene significance of the discretized values, *RefBool* computes the significance of the discretized expression values individually based on the distribution provided in the reference library. This way, it is possible to dissect which samples include significantly active and inactive genes and which result in insignificant results.

Another advantage of the approach is that the discretized expression values are more reliable compared to currently employed methodologies both qualitatively and quantitatively. The discretization of marker genes clearly showed the accurate detection of active genes important for distinct stages of neuroepithelial differentiation. They have been largely confirmed by previously published studies that particularly illustrated their necessity for maintaining cellular identity. In addition, the comparison to current methods highlighted the improved performance with respect to cluster consistency. On average, the discretization of more than 30% of genes is more consistent compared to any other method.

In the context of Boolean network modeling, it allows the reconstruction of cell type or condition specific networks that is not dependent on the comparison to other conditions. Due to the requirement of a single transcriptomics sample without replicates, cellular heterogeneity measured by single-cell RNA-seq could be effectively addressed. Cellular differences in the transcriptional and epigenetic regulation have been previously implicated in the response to drugs for therapeutic intervention, making several subpopulations responsive while others are in a drug-tolerant state (Altschuler and Wu, 2010). Therefore, the identification of instructive factors for cellular conversions as well as the response to cellular perturbations is likely dependent on the individual configuration of each cell. Thus, modeling these differences could provide important insights to predict more efficient instructive factors for cellular conversions.

5.3.2 Limitations

Despite the demonstrated improvements in discretizing gene expression data, the utilization of available transcriptomics datasets as a reference has limitations. In particular, the reference dataset must resemble the global expression distribution closely for obtaining reliable results. However, these real distributions are unknown, which hinders an assessment of how suitable the selected reference is. Due to the impossibility of quantifying those differences, it is important to include as many samples as possible coming from diverse cell and tissue types and covering the whole organism. Nonetheless, a weaker condition for the reference samples is their indistinguishability upon random selection to identify a minimal number of gene expression profiles, which has the same distribution as the complete reference. Following this approach, a minimum of 550 samples was identified to be sufficient as a reference set. The 675 RNA-seq samples used throughout the qualitative and quantitative comparison with other methodologies is thus expected to yield comparable results as if the global distribution would be known.

Another important aspect to consider is the experimental data that was used for estimating the background distribution and comparing *RefBool* against other methods. In particular, all datasets were obtained from bulk RNA-seq experiments in which the RNA of multiple, potentially heterogeneous cells was sequenced. Therefore, the quantified expression values correspond to the average expression of all cells, which hampers the interpretation of discretized gene expression values. This especially concerns later stages in the comparison of *RefBool* to other methods based on a neuroepithelial differentiation dataset including the formation of embryoid bodies (EBs) and neural tube-like rosettes. Embryoid bodies, for example, are composed of multiple cells that concurrently differentiate or remain undifferentiated and are, thus, a heterogeneous population. A previous report exemplified this heterogeneity by examining the expression of OCT4 and identified significant differences in individual cells within the same EB and between different EBs (Wilson *et al.*, 2014). Therefore, the marker genes of EB and neural tube-like rosettes, which were proposed in literature and have been used for comparing *RefBool* with other methods, are not necessarily expressed in all individual cells. However, due to their high average expression, an accurate discretization approach should classify these genes to be active on the population level. Of note, *RefBool* can be readily applied to single cell sequencing datasets, given that the background samples cover the different gene expression ranges, to directly address the heterogeneity of cellular populations.

5.4 Future Work/Outlook

Due to the aforementioned limitations of the network reconstruction process and the increasing knowledge about transcriptional and epigenetic mechanisms, subsequent research could focus on two main lines. First, incorporating gene-specific logic rules considering the cooperative effects of genes on the activation and repression of genes. Second, by integrating the dynamic regulation of accessible and inaccessible chromatin domains into the transcriptional regulatory network, the prediction of instructive factors might be significantly refined. In addition, the use of proteomics data for defining the activity genes could be beneficial to overcome the use of gene expression data as a proxy for protein abundance.

Logic Rules

Elucidating the cooperative behavior of genes in the activation or repression of transcription remains a major objective. In this context, cell type specific enhancers, identified by acetylation of lysine 27 on histone 3, have been shown to be an integral part of the regulatory landscape for expressing lineage-determining TFs (Lara-Astiaso *et al.*, 2014; Monticelli and Natoli, 2017). Especially in the context of Boolean gene regulatory networks, the integration of enhancers is vital for the formulation of logic rules. Since general transcription factors and RNA polymerase II are recruited to active enhancers where the pre-initiation complex is formed and transferred to the promoter, they are seemingly necessary for transcription at high levels (Koch *et al.*, 2011). For this reason, a gene should only be active in the Boolean framework if at least one of its enhancers is bound by a transcription factor. However, practically, the availability of enhancer specific interactions in biological knowledge bases needs to be analyzed for assessing the feasibility of distinguishing the genomic binding locations in reported interactions.

Data Integration

Understanding the interplay of epigenetic and transcriptional changes plays a vital role in enhancing the identification of more efficient combinations of instructive factors both *in vitro* and *in vivo*. Especially epigenetic alterations are gaining increasing attention in the treatment of disease pathologies due to their pronounced effects on the phenotypes (Heerboth *et al.*, 2014). For instance, cardiovascular diseases, such as atherosclerosis, are attributed to epigenetic dysregulations of key genes. There, estrogen receptors ESR1 and ESR2, which are usually expressed, are stably silenced through hypermethylation in vascular smooth muscle cells leading to the loss of atheroprotective estrogen effects (Heerboth *et al.*, 2014). However, epigenetic restructuring does not only occur in the context of disease development and progression, but also during normal development and cellular conversions, in general. A notable example constitutes the addition of vitamin C in the induction of pluripotent stem cells promoting histone demethylation and thereby increasing the efficiency of the conversion (Esteban *et al.*, 2010; Eid and Abdel-Rehim, 2016). Thus, the integration of different regulatory layers into the framework for network reconstruction promises the identification of more efficient instructive factors.

As a first step, gene-level chromatin accessibility information, as predicted by the presented machine learning approach, can be overlaid to exclude inaccessible genes, as has been done in the prediction of instructive factors for inducing the transition of adipocytes into osteoblasts. Without additional information, this approach cannot explain how genes dynamic formation of accessible or inaccessible chromatin domains. However, a directed mechanism has been reported for inducing accessible chromatin through the binding of pioneer transcription factors to unprogrammed chromatin domains, i.e. in which histones are not marked by activating or repressing covalent modifications (Iwafuchi-Doi and Zaret, 2016). Even though the molecular mechanisms of pioneer factors are not yet understood, it can be hypothesized that a reported interaction in prior knowledge networks first triggers chromatin decompression, provided that the target gene is located in inaccessible domains, and subsequently regulates its target transcriptionally. Thus, including this information leads to a dynamic rewiring of the gene regulatory network in which a pioneer factor makes a gene accessible and allows for additional regulation of non-pioneer TFs.

Based on these considerations, a prototype was conceived that reconstructs cell type specific networks containing only interactions originating from active genes. Nonetheless, inactive genes are still included in the networks as terminal nodes whose expression is explained by transcriptional repression while inaccessible genes are omitted. This conceived model is otherwise based on the same considerations as the method presented in this thesis, but overcomes the necessity for representing distinct phenotypes as attractors of the same network. Due to the absence of interactions originating from repressed genes, the network is inherently stable upon perturbation. Therefore, dynamics can be introduced through topological changes induced by copying interactions from the final to the initial network whenever a gene becomes active. Interactions to inaccessible genes are only included in case a pioneer factor was activated that is a reported activator or repressor of these genes and assuming its target becomes readily accessible for other factors, as well. On the other hand, the formation of inaccessible genes could be induced probabilistically in each simulation step, since the temporal ordering of the events is not known a priori. With it, each perturbation needs to be simulated multiple times to obtain a distribution of *in silico* induced phenotypic changes on the network.

The primary drawback of this approach is in the computational resources needed for scoring the perturbations. A hypothetical scenario could contain networks with 1000 genes and the simultaneous perturbation of up to three genes. Without preselecting candidate instructive factors, more than 166 million combinations need to be simulated and, due to the probabilistic closing of genes, repeated several times. If each simulation takes one second and each combination is repeated 100 times, the analysis would require more than 520 years on a single computing core. This hypothetical example underpins the need for other pre-selection schemes taking into account the dynamic topology of the networks. Therefore, the proposed methodology remains a prototype needing further research to obtain insights into its predictive abilities.

Proteomics

Many transcriptional regulatory processes require the abundance of genes in the form of proteins that bind to DNA. For that, the messenger RNA molecules are translated into proteins by ribosomes through the recruitment of complementary tRNA codons carrying amino acids. Like transcription, the translation of mRNA to proteins is regulated through various processes, such as microRNA binding or phosphorylation and inactivation of the translational initiation factor eIF-2 (Sonenberg and Hinnebusch, 2009). Due to this additional layer of regulation, there is no trivial correspondence between protein and mRNA abundance, which probably impedes the appropriate use of gene expression as a proxy for protein abundance (Maier *et al.*, 2009). Correlation analysis of mRNA and protein abundance in mouse xenograft models of ovarian cancer highlights this issue and found no significant correlations between the two measures (Koussounadis *et al.*, 2015). The same study, however, also identified significantly elevated correlations for differentially expressed genes, which supports the decision of a Boolean modeling framework based on differential expression analysis. Nevertheless, evaluating the correspondence with respect to discretization by *RefBool* is important to verify the suitability of absolute discretization, as well. In case no significant correlation can be detected, the use of proteomics data that can be translated to network states could overcome this impediment.

5.5 Conclusion

The advent of cellular conversions through the ectopic expression of cell type specific transcription factors constituted an important step towards regenerative medicine. Continued efforts in elucidating the transcriptional and epigenetic landscapes have led to the identification of numerous target gene combinations that are able to induce the conversion towards a desired cell type. Predominantly, these approaches rely on cell type specific expert knowledge and the sequential interpretation of experimental evidences. Here, computational approaches that systematically analyze and integrate diverse sources of available biological information are beneficial to the establishment of instructive factors inducing novel cellular transitions. Knowing the regulatory networks and their differences in the modulation of distinct gene expression programs will enable the selection of crucial cell fate regulators, the simulation of cellular responses to gene perturbations, and ultimately an increase in efficiency and fidelity of cellular conversions.

Both the absence of experimental data and adequate computational tools for processing them as well as the simplifications of existing models hindered the orderly identification of instructive factors. The goal of this thesis was to address some of these issues. The contributions made in this thesis are that:

- **Condition specific Boolean gene regulatory networks can be reconstructed** from transcriptomics data and prior knowledge networks that account for the distinct transcriptional and epigenetic landscapes of cells. The developed methodology largely captures validated interactions, which have been identified by transcription factor specific ChIP-seq experiments. This provides a topological validation of the networks, which has direct consequences on the dynamic behavior of the model. The elevated condition specificity of reconstructed networks is further supported by the comparison to other, similar methods showing decreased recoveries of discretized phenotypes and experimentally validated interactions.

- **Using a knowledgebase of biological information makes network reconstruction tractable with minimal data requirements**. Large data requirements hindered the development of a general tool for inferring instructive factors of cellular conversions. Most current tools either require substantial amounts of experimental input data, which is largely unavailable, or do not allow for the specification of condition specific data other than gene expression. The presented methodology for condition specific network reconstruction provides a framework that only requires a discretized gene expression profile and prior knowledge networks as a minimal requirement while additional, condition specific epigenetic information can be readily integrated.
- **Candidate instructive factors can be defined by differential network analysis**. In the search for combinations of instructive factors having the potential to induce a desired cellular transition upon perturbation, it is essential to prioritize and select a tractable number of candidate genes. Topological comparison of two networks by means of common regulatory feedback loops and differentially regulated genes identifies important cellular regulators. With it, important genes implicated in maintenance and progression of Systemic Lupus Erythematosus and Rheumatoid Arthritis as well as in shaping the cellular identity of osteoblasts could be identified.
- **Drugs can be prioritized by *in silico* perturbation of candidate instructive factors**. The simulated effects of multitarget combination perturbations establish a condition for prioritizing drugs that are able to induce desired phenotypes. Applying the proposed strategy might substantially reduce the amount of compounds that need to be screened for identifying new therapeutic interventions by repurposing known drugs. However, the efficacy of prioritized drugs dependence on many other factors such as their toxicity or affinity and specificity to molecular targets.
- **Transcriptional comparison of multiple cell types can be enhanced** by taking into account the distinct expression distribution of individual genes. Typically, transcriptional profiles are processed by means of differential expression analysis to obtain insights into significant differences in mRNA abundance. This approach, however, creates an isolated view of the two

conditions under study and does not allow for the comparison of multiple cell types or conditions. *RefBool* integrates available transcriptomics data for providing accurate assignments of whether genes are active or inactive in a particular condition. With it, the comparison of multiple cell types becomes feasible and can be applied to all analysis involving the discretization of gene expression data such as gene set and overrepresentation enrichment analysis. Together with the developed methodology for gene regulatory network reconstruction, it enables the consistent comparison of multiple condition specific networks.

- **Transcriptomics data is indicative of gene-level accessibility** and is equally predictive as experimental chromatin accessibility data processed with current downstream computational tools. The developed machine-learning framework for predicting gene-level accessibility enables the distinction of transcriptionally and epigenetically silenced genes for better understanding the cellular gene regulatory network in the absence of experimental data. In addition, parameter settings of current peak calling methods can be optimized for improving their ability to detect active regulatory regions throughout the genome. As such, the developed methodology is able to overcome the main limitation of current downstream computational tools and provides predictions that can be readily integrated into the condition specific gene regulatory network reconstruction.

This thesis provided an innovative framework for reconstructing condition specific gene regulatory networks that approaches current impediments through the integration of available prior knowledge networks, statistical and machine learning methodologies and the acknowledgment of the cellular differences in transcriptional regulation, which are modulated by the chromatin accessibility landscape. The assumption that distinct networks underlie phenotypic stabilization and its implementation in the current approach is an integral component for future integration of additional layers of condition specific information, such as histone modifications and chromatin remodeling by pioneer factors. Finally, the proposed framework is able to identify candidate instructive factors and prioritize their combination in inducing desired cellular transitions, which will be useful in assisting the establishment of new therapeutic approaches in regenerative medicine.

References

Abdelalim,E.M. *et al.* (2014) The SOX transcription factors as key players in pluripotent stem cells. *Stem Cells Dev.*, **23**, 2687–99.

Acampora,D. *et al.* (2013) Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development*, **140**, 43–55.

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **19**, 716–723.

Albert,R. and Othmer,H.G. (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. *J. Theor. Biol.*, **223**, 1–18.

Altschuler,S.J. and Wu,L.F. (2010) Cellular Heterogeneity: Do Differences Make a Difference? *Cell*, **141**, 559–563.

Amano,T. *et al.* (2009) Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, **16**, 47–57.

Amar,D. *et al.* (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.*, **9**, e1002955.

Apostolou,E. and Hochedlinger,K. (2013) Chromatin dynamics during cellular reprogramming. *Nature*, **502**, 462–471.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Atala,A. (2012) Regenerative medicine strategies. *J. Pediatr. Surg.*, **47**, 17–28.

Atala,A. *et al.* (2006) Tissue-engineered autologous bladders for patients needing cystoplasty. *Lancet (London, England)*, **367**, 1241–6.

Bai,F. *et al.* (2015) Directed Differentiation of Embryonic Stem Cells Into Cardiomyocytes by Bacterial Injection of Defined Transcription Factors. *Sci. Rep.*, **5**, 15014.

Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–95.

Barman,S. and Kwon,Y.-K. (2017) A novel mutual information-based Boolean network inference method from time-series gene expression data. *PLoS One*, **12**, e0171097.

Barozzi,I. *et al.* (2014) Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol. Cell*, **54**, 844–857.

Bateman,E. (1998) Autoregulation of eukaryotic transcription factors. *Prog. Nucleic Acid Res. Mol. Biol.*, **60**, 133–68.

Bauer,O. *et al.* (2015) Loss of osteoblast Runx3 produces severe congenital osteopenia. *Mol. Cell. Biol.*, **35**, 1097–109.

Ben-Dor,A. *et al.* Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–97.

Bhar,A. *et al.* (2013) Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. *Algorithms Mol. Biol.*, **8**, 9.

Bian,Q. and Cahan,P. (2016) Computational Tools for Stem Cell Biology. *Trends Biotechnol.*, **34**, 993–1009.

Blahnik,K.R. *et al.* (2011) Characterization of the Contradictory Chromatin Signatures at the 3′ Exons of Zinc Finger Genes. *PLoS One*, **6**, e17121.

Blake,W.J. *et al.* (2003) Noise in eukaryotic gene expression. *Nature*, **422**, 633–637.

Bonzanni,N. *et al.* (2013) Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics*, **29**, i80–i88.

Bossard,P. and Zaret,K.S. (1998) GATA transcription factors as potentiators of gut endoderm differentiation. *Development*, **125**, 4909–17.

Boyle,A.P. *et al.* (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–8.

Boyle,A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

Brown,C.J. *et al.* (2009) Awakening guardian angels: drugging the p53 pathway. *Nat. Rev. Cancer*, **9**, 862–73.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, **10**, 1213–1218.

Burke,J.F. *et al.* (1981) Successful use of a physiologically acceptable artificial skin in the treatment of extensive burn injury. *Ann. Surg.*, **194**, 413–28.

Burridge,P.W. *et al.* (2014) Chemically defined generation of human cardiomyocytes. *Nat. Methods*, **11**, 855–60.

Caccavo,D. *et al.* (1997) Long-term treatment of systemic lupus erythematosus with cyclosporin A.

Cahan,P. *et al.* (2014) CellNet: network biology applied to stem cell engineering. *Cell*, **158**, 903–15.

Caiazzo,M. *et al.* (2011) Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature*, **476**, 224–7.

Calzone,L. *et al.* (2010) Mathematical Modelling of Cell-Fate Decision in Response to Death Receptor Engagement. *PLoS Comput. Biol.*, **6**, e1000702.

Catlett,J. (1991) On changing continuous attributes into ordered discrete attributes. In, Kodratoff,Y. (ed), *Machine Learning --- EWSL-91: European Working Session on Learning Porto, Portugal, March 6--8, 1991 Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 164–178.

Chantalat,S. *et al.* (2011) Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res.*, **21**, 1426–1437.

Cirillo,L.A. *et al.* (2002) Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell*, **9**, 279–89.

Clark,S.J. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 781.

Crespo,I. *et al.* (2013) Detecting cellular reprogramming determinants by differential stability analysis of gene regulatory networks. *BMC Syst. Biol.*, **7**, 140.

D'Alessio,A.C. *et al.* (2015) A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. *Stem cell reports*, **5**, 763–75.

D'Arcy,P. *et al.* (2011) Inhibition of proteasome deubiquitinating activity as a new cancer therapy. *Nat. Med.*, **17**, 1636–1640.

Davis, a. P. *et al.* (2014) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–D920.

Davis,F.P. and Eddy,S.R. (2013) Transcription factors that convert adult cell identity are differentially polycomb repressed. *PLoS One*, **8**, e63407.

Davis,R.L. *et al.* (1987) Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, **51**, 987–1000.

Deb,K. *et al.* (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, **6**, 182–197.

Desgraupes,B. (2013) Clustering Indices.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Doré,L.C. and Crispino,J.D. (2011) Transcription factor networks in erythroid cell and megakaryocyte development. *Blood*, **118**, 231–9.

Dorier,J. *et al.* (2016) Boolean regulatory network reconstruction using literature based knowledge with a genetic algorithm optimization method. *BMC Bioinformatics*, **17**, 410.

Dougherty,J. *et al.* (1995) Supervised and Unsupervised Discretization of Continuous Features. In, *MACHINE LEARNING: PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE*. Morgan Kaufmann, pp. 194–202.

Dunham,I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Dupont,C. *et al.* (2009) Epigenetics: definition, mechanisms and clinical perspective. *Semin. Reprod. Med.*, **27**, 351–7.

Durillo,J.J. and Nebro,A.J. (2011) JMetal: A Java framework for multi-objective optimization. *Adv. Eng. Softw.*, **42**, 760–771.

Eden,E. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Edwards,A.W.F. and Cavalli-Sforza,L.L. (1965) A Method for Cluster Analysis. *Biometrics*, **21**, 362.

Eid,W. and Abdel-Rehim,W. (2016) Vitamin C promotes pluripotency of human induced pluripotent stem cells via the histone demethylase JARID1A. *Biol. Chem.*, **397**, 1205–1213.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 14863–8.

Eldar,A. and Elowitz,M.B. (2010) Functional roles for noise in genetic circuits. *Nature*, **467**, 167–173.

Elowitz,M.B. (2002) Stochastic Gene Expression in a Single Cell. *Science (80-.).*, **297**, 1183–1186.

ENCODE Datasets ENCODE DNase-Seq experiments.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, **9**, 215–216.

Esteban,M.A. *et al.* (2010) Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. *Cell Stem Cell*, **6**, 71–9.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Faure,A.J. *et al.* (2017) Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Syst.*, **5**, 471–484.e4.

Feng,R. *et al.* (2008) PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 6057–62.

Feng,X. *et al.* (2006) Association of increased interferon-inducible gene expression with disease activity and lupus nephritis in patients with systemic lupus erythematosus. *Arthritis Rheum.*, **54**, 2951–2962.

Fernández-Madrid,F. (1998) Zinc and copper in the treatment of rheumatic diseases. In, Rainsford,K.D. *et al.* (eds), *Copper and Zinc in Inflammatory and Degenerative Diseases*. Springer Netherlands, Dordrecht, pp. 125–137.

Fiannaca,A. *et al.* (2015) Analysis of miRNA expression profiles in breast cancer using biclustering. *BMC Bioinformatics*, **16 Suppl 4**, S7.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–25.

Friedman,H.P. and Rubin,J. (1967) On Some Invariant Criteria for Grouping Data. *J. Am. Stat. Assoc.*, **62**, 1159.

Gallo,C.A. *et al.* (2011) Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinformatics*, **12**, 123.

Gaspar-Maia,A. *et al.* (2011) Open chromatin in pluripotency and reprogramming. *Nat Rev Mol Cell Biol*, **12**, 36–47.

Gaszner,M. and Felsenfeld,G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.*, **7**, 703–13.

Gerard,D. *et al.* (2017) Temporal epigenomic profiling identifies AHR as dynamic super-enhancer controlled regulator of mesenchymal multipotency. *bioRxiv*.

Ghisletti,S. *et al.* (2010) Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, **32**, 317–28.

Glass,L. and Kauffman,S.A. (1973) The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.*, **39**, 103–29.

Goh,K.-I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 8685–8690.

Golub,E.E. and Boesze-Battaglia,K. (2007) The role of alkaline phosphatase in mineralization. *Curr. Opin. Orthop.*, **18**, 444–448.

Gouzé,J.-L. (1998) Positive and Negative Circuits in Dynamical Systems. *J. Biol. Syst.*, **6**, 11–15.

Grieco,L. *et al.* (2013) Integrative Modelling of the Influence of MAPK Network on Cancer Cell Fate Decision. *PLoS Comput. Biol.*, **9**, e1003286.

Gualdi,R. *et al.* (1996) Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes Dev.*, **10**, 1670–82.

Guo,Q. *et al.* (2016) miR-23a/b regulates the balance between osteoblast and adipocyte differentiation in bone marrow mesenchymal stem cells. *Bone Res.*, **4**, 16022.

Halkidi,M. *et al.* (2001) On clustering validation techniques. *J. Intell. Inf. Syst.*, **17**, 107–145.

Haltaufderhyde,K.D. and Oancea,E. (2014) Genome-wide transcriptome analysis of human epidermal melanocytes. *Genomics*, **104**, 482–489.

Hassan,A.H. *et al.* (2002) Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell*, **111**, 369–79.

Haury,A.-C. *et al.* (2012) TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.*, **6**, 145.

Hayward,P. *et al.* (2008) Wnt/Notch signalling and information processing during development. *Development*, **135**, 411–24.

He,Y. *et al.* (2014) Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. *Front. Genet.*, **5**.

Heerboth,S. *et al.* (2014) Use of epigenetic drugs in disease: an overview. *Genet. Epigenet.*, **6**, 9–19.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–89.

Hesselberth,J.R. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, **6**, 283–289.

Hieronymus,H. *et al.* (2006) Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell*, **10**, 321–330.

Hikichi,T. *et al.* (2013) Transcription factors interfering with dedifferentiation induce cell type-specific transcriptional profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 6412–7.

Hnisz,D. *et al.* (2013) Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, **155**, 934–947.

Hu,G. and Agarwal,P. (2009) Human disease-drug network based on genomic expression profiles. *PLoS One*, **4**, e6536.

Hu,L. *et al.* (2018) Mesenchymal Stem Cells: Cell Fate Decision to Osteoblast or Adipocyte and Application in Osteoporosis Treatment. *Int. J. Mol. Sci.*, **19**.

Hu,N. *et al.* (2008) Abnormal histone modification patterns in lupus CD4+ T cells. *J. Rheumatol.*, **35**, 804–10.

Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Huang,P. *et al.* (2011) Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature*, **475**, 386–9.

Hudson,N.J. *et al.* (2012) Beyond differential expression: the quest for causal mutations and effector molecules. *BMC Genomics*, **13**, 356.

Ieda,M. *et al.* (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, **142**, 375–86.

Iwafuchi-Doi,M. and Zaret,K.S. (2016) Cell fate control by pioneer transcription factors. *Development*, **143**, 1833–7.

Iwafuchi-Doi,M. and Zaret,K.S. (2014) Pioneer transcription factors in cell reprogramming. *Genes Dev.*, **28**, 2679–92.

Iyer,M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–80.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–2.

John,S. *et al.* (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.

Johnson,D.B. (1975) Finding All the Elementary Circuits of a Directed Graph. *SIAM J. Comput.*, **4**, 77–84.

Jonsson,P.F. and Bates,P. a. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**, 2291–2297.

Jung,S. *et al.* (2017) RefBool: a reference-based algorithm for discretizing gene expression data. *Bioinformatics*, **33**, 1953–1962.

Kadauke,S. *et al.* (2012) Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1. *Cell*, **150**, 725–37.

Kærn,M. *et al.* (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, **6**, 451–464.

Kaiser,L.R. (1992) The future of multihospital systems. *Top. Health Care Financ.*, **18**, 32–45.

Kashyap,V. *et al.* (2009) Regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the NANOG, OCT4, and SOX2 pluripotency transcription factors with polycomb repressive complexes and stem cell microRNAs. *Stem Cells Dev.*, **18**, 1093–108.

Kasowski,M. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–2.

Kauffman,S. a (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.

Kaul,A. *et al.* (2016) Systemic lupus erythematosus. *Nat. Rev. Dis. Prim.*, **2**, 16039.

Kerber,R. (1992) ChiMerge: Discretization of Numeric Attributes. In, *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92. AAAI Press, pp. 123–128.

Kidder,B.L. *et al.* (2011) ChIP-Seq: technical considerations for obtaining high-quality data. *Nat. Immunol.*, **12**, 918–22.

Kilpinen,H. *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–7.

Klijn,C. *et al.* (2014) A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.*, **33**, 306–312.

Knuth,D.E. (1964) backus normal form vs. Backus Naur form. *Commun. ACM*, **7**, 735–736.

Ko,Y. *et al.* (2013) Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proc. Natl. Acad. Sci.*, **110**, 3095–3100.

Koch,F. *et al.* (2011) Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.*, **18**, 956–63.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–58.

Koohy,H. *et al.* (2014) A Comparison of Peak Callers Used for DNase-Seq Data. *PLoS One*, **9**, e96303.

Korhonen,J. *et al.* (2009) MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, **25**, 3181–2.

Kosti,I. *et al.* (2016) Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci. Rep.*, **6**, 24799.

Kostka,D. and Spang,R. (2004) Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, **20 Suppl 1**, i194-9.

Koussounadis,A. *et al.* (2015) Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.*, **5**, 10775.

Kulakovskiy,I. V *et al.* (2016) HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, **44**, D116-25.

Kulessa,H. *et al.* (1995) GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev.*, **9**, 1250–62.

Lai,Y. *et al.* (2004) A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*, **20**, 3146–55.

Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer*, **7**, 54–60.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Lara-Astiaso,D. *et al.* (2014) Immunogenetics. Chromatin state dynamics during blood formation. *Science (80-.).*, **345**, 943–949.

Lèbre,S. *et al.* (2010) Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst. Biol.*, **4**, 130.

Levine,A.J. and Berger,S.L. (2017) The interplay between epigenetic changes and the p53 protein in stem cells. *Genes Dev.*, **31**, 1195–1201.

Li,J. and McMurray,R.W. (2009) Effects of chronic exposure to DDT and TCDD on disease activity in murine systemic lupus erythematosus. *Lupus*, **18**, 941–949.

Li,M. and Belmonte,J.C.I. (2017) Ground rules of the pluripotency gene regulatory network. *Nat. Rev. Genet.*, **18**, 180–191.

Li,Y. *et al.* (2010) Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. *BMC Bioinformatics*, **11**, 520.

Liang,Y. *et al.* (2014) Association of signaling transducers and activators of transcription 1 and systemic lupus erythematosus. *Autoimmunity*, **6934**, 1–5.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93.

Liu,L. *et al.* (2015) Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res.*, **43**, 3873–85.

Liu,X. *et al.* (2013) miR-23a targets interferon regulatory factor 1 and modulates cellular proliferation and paclitaxel-induced apoptosis in gastric adenocarcinoma cells. *PLoS One*, **8**, e64707.

Liu,Z. *et al.* (2017) Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Sci. Rep.*, **7**, 2193.

Liu,Z.-P. (2015) Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data. *Curr. Genomics*, **16**, 3–22.

Locke,J.C.W. *et al.* (2005) Extension of a genetic network model by iterative experimentation and mathematical analysis. *Mol. Syst. Biol.*, **1**, 2005.0013.

Lu,Q. *et al.* (2005) Demethylation of the same promoter sequence increases CD70 expression in lupus T cells and T cells treated with lupus-inducing drugs. *J. Immunol.*, **174**, 6212–9.

Macneil,L.T. and Walhout,A.J.M. (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.*, **21**, 645–57.

Macqueen,J. (1967) Some methods for classification and analysis of multivariate observations. In, *In 5-th Berkeley Symposium on Mathematical Statistics and Probability.*, pp. 281–297.

Madeira,S.C. and Oliveira,A.L. (2005) A Linear Time Biclustering Algorithm for Time Series Gene Expression Data. In, Casadio,R. and Myers,G. (eds), *Algorithms in Bioinformatics: 5th International Workshop, WABI 2005, Mallorca, Spain, October 3-6, 2005. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 39–52.

Maffioletti,S.M. *et al.* (2015) Efficient derivation and inducible differentiation of expandable skeletal myogenic cells from human ES and patient-specific iPS cells. *Nat. Protoc.*, **10**, 941–58.

Mahalanobis,P.C. (1936) On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, **2**, 49–55.

Maier,T. *et al.* (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett.*, **583**, 3966–73.

Malhotra,A. *et al.* (2013) Chromosomal structural variations during progression of a prostate epithelial cell line to a malignant metastatic state inactivate the NF2, NIPSNAP1, UGT2B17, and LPIN2 genes. *Cancer Biol. Ther.*, **14**, 840–852.

Mani,S. and Cooper,G.F. (2004) Causal discovery using a Bayesian local causal discovery algorithm. *Stud. Health Technol. Inform.*, **107**, 731–5.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Marriott,F.H.C. (1971) Practical Problems in a Method of Cluster Analysis. *Biometrics*, **27**, 501.

Mason,C. and Dunnill,P. (2008) A brief definition of regenerative medicine. *Regen. Med.*, **3**, 1–5.

McCullagh,E. *et al.* (2009) Not all quiet on the noise front. *Nat. Chem. Biol.*, **5**, 699–704.

McVicker,G. *et al.* (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747–9.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, **72**, 417–473.

Meissner,A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–70.

Melas,I.N. *et al.* (2013) Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Comput. Biol.*, **9**, e1003204.

Mentzen,W.I. *et al.* (2009) Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumor. *BMC Genomics*, **10**, 601.

Mercer,T.R. *et al.* (2013) DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nat Genet*, **45**, 852–859.

Moncunill,V. *et al.* (2014) Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.*, **32**, 1106–1112.

Monticelli,S. and Natoli,G. (2017) Transcriptional determination and functional specificity of myeloid cells: making sense of diversity. *Nat. Rev. Immunol.*, **17**, 595–607.

Moris,N. *et al.* (2016) Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.*, **17**, 693–703.

Morris,S.A. (2016) Direct lineage reprogramming via pioneer factors; a detour through developmental gene regulatory networks. *Development*, **143**, 2696–705.

Morris,S.A. *et al.* (2014) Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell*, **158**, 889–902.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–8.

Mota,L.M.H. Da *et al.* (2012) Lack of association between the CC genotype of the rs7903146 polymorphism in the TCF7L2 gene and rheumatoid arthritis. *Rev. Bras. Reumatol.*, **52**, 523–528.

Mujtaba,S. *et al.* (2007) Structure and acetyl-lysine recognition of the bromodomain. *Oncogene*, **26**, 5521–7.

Müssel,C. *et al.* (2016) BiTrinA—multiscale binarization and trinarization with quality analysis. *Bioinformatics*, **32**, 465–468.

Nachman,I. *et al.* (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, **20**, i248–i256.

Neph,S. *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–86.

Nepomuceno,N. *et al.* (2007) Evolutionary Computation in Combinatorial Optimization Blum,C. and Ochoa,G. (eds) Springer Berlin Heidelberg, Berlin, Heidelberg.

Ng,H.-H. and Surani,M.A. (2011) The transcriptional and signalling networks of pluripotency. *Nat. Cell Biol.*, **13**, 490–6.

Niwa,H. *et al.* (2000) Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.*, **24**, 372–6.

Le Novère,N. (2015) Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.*, **16**, 146–58.

van Oevelen,C. *et al.* (2015) C/EBPα Activates Pre-existing and De Novo Macrophage Enhancers during Induced Pre-B Cell Transdifferentiation and Myelopoiesis. *Stem cell reports*, **5**, 232–47.

Palmer,C. *et al.* (2006) Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, **7**, 115.

Paulsson,J. (2005) Models of stochastic gene expression. *Phys. Life Rev.*, **2**, 157–175.

Perlman,H. *et al.* (2003) IL-6 and matrix metalloproteinase-1 are regulated by the cyclin-dependent kinase inhibitor p21 in synovial fibroblasts. *J. Immunol.*, **170**, 838–845.

Pieters,T. and van Roy,F. (2014) Role of cell-cell adhesion complexes in embryonic stem cell biology. *J. Cell Sci.*, **127**, 2603–13.

Plahte,E. *et al.* (1995) Feedback Loops, Stability and Multistationarity in Dynamical Systems. *J. Biol. Syst.*, **3**, 409–413.

Polynikis,A. *et al.* (2009) Comparing different ODE modelling approaches for gene regulatory networks. *J. Theor. Biol.*, **261**, 511–530.

Qiao,Y. *et al.* (2015) AF9 promotes hESC neural differentiation through recruiting TET2 to neurodevelopmental gene loci for methylcytosine hydroxylation. *Cell Discov.*, **1**, 15017.

Qin,J.Y. *et al.* (2010) Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. *PLoS One*, **5**, e10611.

Rackham,O.J.L. *et al.* (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.*, **48**, 331–5.

Radman-Livaja,M. and Rando,O.J. (2010) Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.*, **339**, 258–66.

Raj,A. and van Oudenaarden,A. (2008) Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, **135**, 216–226.

Rashid,N.U. *et al.* (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.*, **12**, R67.

Rau,A. *et al.* (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**, 2146–2152.

Ray,S. and Turi,R.H. (1999) Determination of number of clusters in K-means clustering and application in colour segmentation. In, *The 4th International Conference on Advances in Pattern Recognition and Digital Techniques.*, pp. 137–143.

Reiss,A.B. *et al.* (2012) Resveratrol Counters Pro-Atherogenic Effects of Systemic Lupus Erythematosus and Rheumatoid Arthritis Plasma On Cholesterol Efflux in Human Macrophages [abstract]. *Arthritis Rheum.*, **64 Suppl 1**, 924.

Remy,E. and Ruet,P. (2008) From minimal signed circuits to the dynamics of Boolean regulatory networks. In, *Bioinformatics*.

Roadmap Epigenomics,C. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Rohs,R. *et al.* (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–53.

Ropers,D. *et al.* (2006) Qualitative simulation of the carbon starvation response in Escherichia coli. *Biosystems.*, **84**, 124–52.

Rosmarin,A.G. *et al.* (2005) Transcriptional regulation in myelopoiesis: Hematopoietic fate choice, myeloid differentiation, and leukemogenesis. *Exp. Hematol.*, **33**, 131–43.

Sabo,P.J. *et al.* (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci.*, **101**, 16837–16842.

Said,M.R. *et al.* (2004) Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 18006–11.

Sampogna,G. *et al.* (2015) Regenerative medicine: Historical roots and potential strategies in modern medicine. *J. Microsc. Ultrastruct.*, **3**, 101–107.

Schadt,E.E. *et al.* (2009) A network view of disease and compound screening. *Nat. Rev. Drug Discov.*, **8**, 286–295.

Seiffert,C. *et al.* (2010) RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, **40**, 185–197.

Sekiya,S. and Suzuki,A. (2011) Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature*, **475**, 390–3.

Shaltouki,A. *et al.* (2013) Efficient generation of astrocytes from human pluripotent stem cells in defined conditions. *Stem Cells*, **31**, 941–52.

Shen,C.N. *et al.* (2000) Molecular basis of transdifferentiation of pancreas to liver. *Nat. Cell Biol.*, **2**, 879–87.

Shi,Y. *et al.* (2012) Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat. Protoc.*, **7**, 1836–46.

Shlyueva,D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–86.

Snoussi,E.H. (1998) Necessary Conditions for Multistationarity and Stable Periodicity. *J. Biol. Syst.*, **6**, 3–9.

van Someren,E.P. *et al.* (2006) Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics*, **22**, 477–84.

Son,E.Y. *et al.* (2011) Conversion of mouse and human fibroblasts into functional spinal motor neurons. *Cell Stem Cell*, **9**, 205–18.

Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, **136**, 731–745.

Song,L. *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, **21**, 1757–1767.

Soufi,A. *et al.* (2012) Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell*, **151**, 994–1004.

Soufi,A. *et al.* (2015) Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*, **161**, 555–568.

Soulè,C. (2003) Graphic Requirements for Multistationarity. *Complexus*, **1**, 123–133.

Starzl,T.E. (2000) History of clinical transplantation. *World J. Surg.*, **24**, 759–82.

Stephens,A.S. *et al.* (2011) Myocyte enhancer factor 2c, an osteoblast transcription factor identified by dimethyl sulfoxide (DMSO)-enhanced mineralization. *J. Biol. Chem.*, **286**, 30071–86.

Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–60.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102**, 15545–15550.

Syeed,N. *et al.* (2010) Mutational profile of the CAV-1 gene in breast cancer cases in the ethnic Kashmiri population. *Asian Pacific J. Cancer Prev.*, **11**, 1099–1105.

Szabo,E. *et al.* (2010) Direct conversion of human fibroblasts to multilineage blood progenitors. *Nature*, **468**, 521–6.

Takahashi,K. *et al.* (2007) Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell*, **131**, 861–872.

Takahashi,K. and Yamanaka,S. (2016) A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.*, **17**, 183–93.

Takahashi,K. and Yamanaka,S. (2015) A developmental framework for induced pluripotency. *Development*, **142**, 3274–85.

Takahashi,K. and Yamanaka,S. (2006) Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, **126**, 663–676.

Takenaka,T. (2001) Classical vs reverse pharmacology in drug discovery. *BJU Int.*, **88 Suppl 2**, 7-10-50.

Terfve,C. *et al.* (2012) CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst. Biol.*, **6**, 133.

Thomas,R. *et al.* (2017) Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.*, **18**, 441–450.

Thomas,R. (1981) On the Relation Between the Logical Structure of Systems and Their Ability to Generate Multiple Steady States or Sustained Oscillations. In, Della Dora,J. *et al.* (eds), *Numerical methods in the study of critical phenomena*, Springer Series in Synergetics. Springer Berlin Heidelberg, pp. 180–193.

Thomas,R. (1994) The role of feedback circuits: Positive feedback circuits are a necessary condition for positive real eigenvalues of the Jacobian matrix.

Berichte der Bunsengesellschaft für Phys. Chemie, **98**, 1148–1151.

Thomson,M. *et al.* (2011) Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell*, **145**, 875–89.

Thurman,R.E. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

Torrente,A. *et al.* (2016) Identification of Cancer Related Genes Using a Comprehensive Map of Human Gene Expression. *PLoS One*, **11**, e0157484.

Trakhtenberg,E.F. *et al.* (2016) Cell types differ in global coordination of splicing and proportion of highly expressed genes. *Sci. Rep.*, **6**, 32249.

Tsompana,M. and Buck,M.J. (2014) Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*, **7**, 33.

Uygun,B.E. *et al.* (2010) Organ reengineering through development of a transplantable recellularized liver graft using decellularized liver matrix. *Nat. Med.*, **16**, 814–20.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Vierbuchen,T. *et al.* (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, **463**, 1035–41.

Vijesh,N. *et al.* (2013) Modeling of gene regulatory networks: A review. *J. Biomed. Sci. Eng.*, **6**, 223–231.

Vogel,M.J. *et al.* (2006) Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res.*, **16**, 1493–1504.

Waddington,C.H. (1957) The Strategy of the Genes Allen and Unwin, London.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Wapinski,O.L. *et al.* (2013) Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell*, **155**, 621–35.

Wei,G. *et al.* (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell*, **10**, 331–342.

Wells,G.A. *et al.* (1998) Cyclosporine for treating rheumatoid arthritis. *Cochrane Database Syst. Rev.*, CD001083.

WEMMERT,C. *et al.* (2000) A COLLABORATIVE APPROACH TO COMBINE MULTIPLE LEARNING METHODS. *Int. J. Artif. Intell. Tools*, **9**, 59–78.

Wilks,S.S. (1938) The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Stat.*, 60–62.

Wilson,J.L. *et al.* (2014) Single-cell analysis of embryoid body heterogeneity using microfluidic trapping array. *Biomed. Microdevices*, **16**, 79–90.

Wittmann,D.M. *et al.* (2009) Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. *BMC Syst. Biol.*, **3**, 98.

Wong,N. and Wang,X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.*, **43**, D146-52.

Wooller,S.K. *et al.* (2017) Bioinformatics in translational drug discovery. *Biosci. Rep.*, **37**, BSR20160180.

Wu,F. *et al.* (2015) Two transcription factors, Pou4f2 and Isl1, are sufficient to specify the retinal ganglion cell fate. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, E1559-68.

Xia,X. (2017) Bioinformatics and Drug Discovery. *Curr. Top. Med. Chem.*, **17**, 1709–1726.

Xie,X.L. and Beni,G. (1991) A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**, 841–847.

Xu,H. *et al.* (2014) Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput. Biol.*, **10**, e1003777.

Xu,W.D. *et al.* (2012) IRF7, a functional factor associates with systemic lupus erythematosus. *Cytokine*, **58**, 317–320.

Yagil,G. and Yagil,E. (1971) On the Relation between Effector Concentration and the Rate of Induced Enzyme Synthesis. *Biophys. J.*, **11**, 11–27.

Young,R.A. (2011) Control of the embryonic stem cell state. *Cell*, **144**, 940–54.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, **68**, 49–67.

Zhang,H.-M. *et al.* (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.*, **43**, D76-81.

Zhang,T. and Kraus,W.L. (2010) SIRT1-dependent regulation of chromatin and transcription: linking NAD(+) metabolism and signaling to the control of cellular functions. *Biochim. Biophys. Acta*, **1804**, 1666–75.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Zhang,Y. *et al.* (2013) Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron*, **78**, 785–98.

Appendix

Supplementary Tables

Table S1. Gene expression datasets for network inference validation

Cell line	GEO accessions
GM12878	GSM993481, GSM993482, GSM993483
H1-hESC	GSM993497, GSM993498, GSM993499, GSM993500
HepG2	GSM993552, GSM993554, GSM993556, GSM993557
K562	GSM993518, GSM993519, GSM993520

Table S2. ENCODE filenames of transcription factor ChIP-seq data for defining accessible chromatin regions

A549	GM12878	H1-hESC	HeLa-S3	HepG2	K562
wgEncodeHaibTfbsA549Atf3V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Atf2sc81188V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescAtf2sc81188V0422111PkRep1.broadPeak	wgEncodeHaibTfbsHelas3GabpPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsHepg2Atf3V0416101PkRep1.broadPeak	wgEncodeAwgTfbsHaibK562Bcl3Pcr1xUniPk.narrowPeak
wgEncodeHaibTfbsA549Bcl3V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Atf3Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescAtf3V0416102PkRep1.broadPeak	wgEncodeHaibTfbsHelas3NrsfPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsHepg2Bhlhe40V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Atf3V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549Creb1sc240V0416102Dex100nmPkRep1.broadPeak	wgEncodeHaibTfbsGm12878BatfPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescBcl11aPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsHelas3Pol2Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsHepg2Cebpbsc150V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Bclaf101388Pcr1xPkRep1.broadPeak
wgEncodeHaibTfbsA549E2f6V0422111PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Bcl11aPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescCreb1sc240V0422111PkRep1.broadPeak	wgEncodeHaibTfbsHelas3Taf1Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsHepg2Cebpdsc636V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Cbx3sc101004V0422111PkRep1.broadPeak
wgEncodeHaibTfbsA549Elf1V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Bcl3V0416101PkRep1.broadPeak	wgEncodeHaibTfbsH1hescCtcfsc5916V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Ap2alphaStdPk.narrowPeak	wgEncodeHaibTfbsHepg2Creb1sc240V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Cebpbsc150V0422111PkRep1.broadPeak
wgEncodeHaibTfbsA549Ets1V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Bclaf101388V0416101PkRep1.broadPeak	wgEncodeHaibTfbsH1hescE2f6V0422111PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Ap2gammaStdPk.narrowPeak	wgEncodeHaibTfbsHepg2Ctcfsc5916V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Cebpdsc636V0422111PkRep1.broadPeak
wgEncodeHaibTfbsA549Fosl2V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Cebpbsc150V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescEgr1V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Baf155IggmusPk.narrowPeak	wgEncodeHaibTfbsHepg2Elf1sc631V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Creb1sc240V0422111PkRep1.broadPeak
wgEncodeHai	wgEncodeHaib	wgEncodeHaib	wgEncodeSyd	wgEncodeHaib	wgEncodeHaib

bTfbsA549Foxa2V0416102Etoh02PkRep1.broadPeak	TfbsGm12878Creb1sc240V0422111PkRep1.broadPeak	TfbsH1hescFosl1sc183V0416102PkRep1.broadPeak	hTfbsHelas3Baf170IggmusPk.narrowPeak	TfbsHepg2Fosl2V0416101PkRep1.broadPeak	TfbsK562Ctcflsc98982V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549GabpV0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Ebfsc137065Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescGabpPcr1xPkRep1.broadPeak	wgEncodeSydhTfbsHelas3Bdp1StdPk.narrowPeak	wgEncodeHaibTfbsHepg2Foxa1sc101058V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Egr1V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549Gata3V0422111PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Egr1V0416101PkRep1.broadPeak	wgEncodeHaibTfbsH1hescHdac2sc6296V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Brca1a300IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Foxa1sc6553V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Elf1sc631V0416102PkRep1.broadPeak
wgEncodeHaibTfbsA549JundV0416102Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Elf1sc631V0416101PkRep1.broadPeak	wgEncodeHaibTfbsH1hescJundV0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Brf1StdPk.narrowPeak	wgEncodeHaibTfbsHepg2Foxa2sc6554V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Ets1V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549MaxV0422111PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Ets1Pcr1xPkRep1V2.broadPeak	wgEncodeHaibTfbsH1hescMaxV0422111PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Brf2StdPk.narrowPeak	wgEncodeHaibTfbsHepg2GabpPcr2xPkRep1.broadPeak	wgEncodeHaibTfbsK562Fosl1sc183V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549NrsfV0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Foxm1sc502V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescNanogsc33759V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Brg1IggmusPk.narrowPeak	wgEncodeHaibTfbsHepg2Hdac2sc6296V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562GabpV0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549P300V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878GabpPcr2xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescNrsfV0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3CebpbIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Hnf4asc8987V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Gata2sc267Pcr1xPkRep1.broadPeak
wgEncodeHaibTfbsA549Pbx3V0422111PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Irf4sc6059Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescP300V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3CfosStdPk.narrowPeak	wgEncodeHaibTfbsHepg2Hnf4gsc6558V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Hdac2sc6296V0416102PkRep1.broadPeak
wgEncodeHaibTfbsA549Rad21V0422111PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Mef2aPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescPol2V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Chd2IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2JundPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsK562MaxV0416102PkRep1.broadPeak
wgEncodeHaibTfbsA549Sin3ak20V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Mef2csc13268V0416101PkRep1.broadPeak	wgEncodeHaibTfbsH1hescPou5f1sc9081V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3CjunIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2MaxV0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Mef2aV0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549Six5V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Mta3sc81325V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescRad21V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3CmycStdPk.narrowPeak	wgEncodeHaibTfbsHepg2Mbd4sc271530V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Nr2f2sc271940V0422111PkRep1.broadPeak
wgEncodeHai	wgEncodeHaib	wgEncodeHaib	wgEncodeSyd	wgEncodeHaib	wgEncodeHaib

bTfbsA549Sp1V0422111Etoh02PkRep1.broadPeak	TfbsGm12878Nfatc1sc17834V0422111PkRep1.broadPeak	TfbsH1hescRxraV0416102PkRep1.broadPeak	hTfbsHelas3Corestsc30189IggrabPk.narrowPeak	TfbsHepg2Mybl2sc81192V0422111PkRep1.broadPeak	TfbsK562NrsfV0416102PkRep1.broadPeak
wgEncodeHaibTfbsA549Taf1V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Nficsc81335V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescSix5Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsHelas3E2f1StdPk.narrowPeak	wgEncodeHaibTfbsHepg2Nficsc81335V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Pmlsc71910V0422111PkRep1.broadPeak
wgEncodeHaibTfbsA549Tcf12V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878NrsfPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescSp1Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsHelas3E2f4StdPk.narrowPeak	wgEncodeHaibTfbsHepg2Nr2f2sc271940V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Pu1Pcr1xPkRep1.broadPeak
wgEncodeHaibTfbsA549Tead4sc101184V0422111PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Pax5c20Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescSp2V0422111PkRep1.broadPeak	wgEncodeSydhTfbsHelas3E2f6StdPk.narrowPeak	wgEncodeHaibTfbsHepg2NrsfV0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Rad21V0416102PkRep1.broadPeak
wgEncodeHaibTfbsA549Usf1V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Pbx3Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescSp4v20V0422111PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Elk112771IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2P300V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Sin3ak20V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549Yy1cV0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Pmlsc71910V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescSrfPcr1xPkRep1.broadPeak	wgEncodeSydhTfbsHelas3Elk4UcdPk.narrowPeak	wgEncodeHaibTfbsHepg2Pol2Pcr2xPkRep1.broadPeak	wgEncodeHaibTfbsK562Six5V0416101PkRep1.broadPeak
wgEncodeHaibTfbsA549Zbtb33V0422111Etoh02PkRep1.broadPeak	wgEncodeHaibTfbsGm12878Pol2Pcr2xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescTaf1V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Gcn5StdPk.narrowPeak	wgEncodeHaibTfbsHepg2Rad21V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Sp1Pcr1xPkRep1.broadPeak
wgEncodeSydhTfbsA549Bhlhe40IggrabPk.narrowPeak	wgEncodeHaibTfbsGm12878Pou2f2Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescTaf7sc101167V0416102PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Gtf2f1ab28179IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2RxraPcr1xPkRep1.broadPeak	wgEncodeHaibTfbsK562Sp2sc643V0416102PkRep1.broadPeak
wgEncodeSydhTfbsA549CebpbIggrabPk.narrowPeak	wgEncodeHaibTfbsGm12878Pu1Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsH1hescTcf12Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsHelas3Hae2f1StdPk.narrowPeak	wgEncodeHaibTfbsHepg2Sin3ak20Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsK562SrfV0416101PkRep1.broadPeak
wgEncodeSydhTfbsA549CmycIggrabPk.narrowPeak	wgEncodeHaibTfbsGm12878Rad21V0416101PkRep1.broadPeak	wgEncodeHaibTfbsH1hescTead4sc101184V0422111PkRep1.broadPeak	wgEncodeSydhTfbsHelas3Hcfc1nb10068209IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Sp1Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsK562Stat5asc74442V0422111PkRep1.broadPeak
wgEncodeSydhTfbsA549CtcfbIggrabPk.narrowPeak	wgEncodeHaibTfbsGm12878Runx3sc101553V0422111PkRep1.broadPeak	wgEncodeHaibTfbsH1hescUsf1Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsHelas3Ini1IggmusPk.narrowPeak	wgEncodeHaibTfbsHepg2Sp2V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Taf1V0416101PkRep1.broadPeak
wgEncodeSyd	wgEncodeHaib	wgEncodeHaib	wgEncodeSyd	wgEncodeHaib	wgEncodeHaib

hTfbsA549MaxIggrabPk.narrowPeak	TfbsGm12878RxraPcr1xPkRep1.broadPeak	TfbsH1hescYy1sc281V0416102PkRep1.broadPeak	hTfbsHelas3Irf3IggrabPk.narrowPeak	TfbsHepg2SrfV0416101PkRep1.broadPeak	TfbsK562Taf7sc101167V0416101PkRep1.broadPeak
wgEncodeSydhTfbsA549Pol2s2IggrabPk.narrowPeak	wgEncodeHaibTfbsGm12878Six5Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescBach1sc14700IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3JundIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Taf1Pcr2xPkRep1.broadPeak	wgEncodeHaibTfbsK562Tead4sc101184V0422111PkRep1.broadPeak
wgEncodeSydhTfbsA549Rad21IggrabPk.narrowPeak	wgEncodeHaibTfbsGm12878Sp1Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescBrca1IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3MafkIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Tcf12Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsK562Thap1sc98174V0416101PkRep1.broadPeak
	wgEncodeHaibTfbsGm12878SrfPcr2xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescCebpbIggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3MaxIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Tead4sc101184V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Trim28sc81411V0422111PkRep1.broadPeak
	wgEncodeHaibTfbsGm12878Stat5asc74442V0422111PkRep1.broadPeak	wgEncodeSydhTfbsH1hescChd1a301218aIggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3MaxStdPk.narrowPeak	wgEncodeHaibTfbsHepg2Usf1Pcr1xPkRep1.broadPeak	wgEncodeHaibTfbsK562Usf1V0416101PkRep1.broadPeak
	wgEncodeHaibTfbsGm12878Taf1Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescChd2IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Mazab85725IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Yy1sc281V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Yy1V0416102PkRep1.broadPeak
	wgEncodeHaibTfbsGm12878Tcf12Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescCjunIggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Mxi1af4185IggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Zbtb33V0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Yy1sc281V0416101PkRep1.broadPeak
	wgEncodeHaibTfbsGm12878Tcf3Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescCmycIggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3NfyaIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Zbtb7aV0416101PkRep1.broadPeak	wgEncodeHaibTfbsK562Zbtb33Pcr1xPkRep1.broadPeak
	wgEncodeHaibTfbsGm12878Usf1Pcr2xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescCtbp2UcdPk.narrowPeak	wgEncodeSydhTfbsHelas3NfybIggrabPk.narrowPeak	wgEncodeHaibTfbsHepg2Zeb1V0422111PkRep1.broadPeak	wgEncodeHaibTfbsK562Zbtb7asc34508V0416101PkRep1.broadPeak
	wgEncodeHaibTfbsGm12878Yy1sc281Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescGtf2f1IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Nrf1IggmusPk.narrowPeak	wgEncodeSydhTfbsHepg2Arid3anb100279IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Arid3asc8821IggrabPk.narrowPeak
	wgEncodeHaibTfbsGm12878Zbtb33Pcr1xPkRep1.broadPeak	wgEncodeSydhTfbsH1hescMafkIggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3P300sc584sc584IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2Bhlhe40cIggrabPk.narrowPeak	wgEncodeSydhTfbsK562Atf106325StdPk.narrowPeak
	wgEncodeHaibTfbsGm12878	wgEncodeSydhTfbsH1hesc	wgEncodeSydhTfbsHelas3Po	wgEncodeSydhTfbsHepg2Br	wgEncodeSydhTfbsK562Bac

	Zeb1sc25388V0416102PkRep1.broadPeak	Mxi1IggrabPk.narrowPeak	l2StdPk.narrowPeak	ca1a300IggrabPk.narrowPeak	h1sc14700IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Bhlhe40cIggmusPk.narrowPeak	wgEncodeSydhTfbsH1hescNrf1IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Pol2s2IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2CebpbIggrabPk.narrowPeak	wgEncodeSydhTfbsK562Bdp1StdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Brca1a300IggmusPk.narrowPeak	wgEncodeSydhTfbsH1hescRfx5200401194IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Prdm19115IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2CebpzIggrabPk.narrowPeak	wgEncodeSydhTfbsK562Bhlhe40nb100IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Cdpsc6327IggmusPk.narrowPeak	wgEncodeSydhTfbsH1hescSin3anb6001263IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Rad21IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2Chd2ab68301IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Brf1StdPk.narrowPeak
	wgEncodeSydhTfbsGm12878CfosStdPk.narrowPeak	wgEncodeSydhTfbsH1hescSuz12UcdPk.narrowPeak	wgEncodeSydhTfbsHelas3Rfx5200401194IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2CjunIggrabPk.narrowPeak	wgEncodeSydhTfbsK562Brf2StdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Chd1a301218aIggmusPk.narrowPeak	wgEncodeSydhTfbsH1hescTbpIggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Rpc155StdPk.narrowPeak	wgEncodeSydhTfbsHepg2Corestsc30189IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Brg1IggmusPk.narrowPeak
	wgEncodeSydhTfbsGm12878Chd2ab68301IggmusPk.narrowPeak	wgEncodeSydhTfbsH1hescUsf2IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Smc3ab9263IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2Irf3IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Ccnt2StdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Corestsc30189IggmusPk.narrowPeak	wgEncodeSydhTfbsH1hescZnf143IggrabPk.narrowPeak	wgEncodeSydhTfbsHelas3Spt20StdPk.narrowPeak	wgEncodeSydhTfbsHepg2JundIggrabPk.narrowPeak	wgEncodeSydhTfbsK562Cdpsc6327IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Ctcfsc15914c20StdPk.narrowPeak	wgEncodeSydhTfbsH1hescZnf274m01UcdPk.narrowPeak	wgEncodeSydhTfbsHelas3Stat1Ifng30StdPk.narrowPeak	wgEncodeSydhTfbsHepg2Maffm8194IggrabPk.narrowPeak	wgEncodeSydhTfbsK562CfosStdPk.narrowPeak
	wgEncodeSydhTfbsGm12878E2f4IggmusPk.narrowPeak		wgEncodeSydhTfbsHelas3Stat3IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2Mafkab50322IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Chd2ab68301IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Elk112771IggmusPk.narrowPeak		wgEncodeSydhTfbsHelas3TbpIggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2Mafksc477IggrabPk.narrowPeak	wgEncodeSydhTfbsK562CjunIggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878ErraIggrabPk.narrowPeak		wgEncodeSydhTfbsHelas3Tcf7l2UcdPk.narrowPeak	wgEncodeSydhTfbsHepg2MaxIggrabPk.narrowPeak	wgEncodeSydhTfbsK562MycIggrabPk.narrowPeak

	wgEncodeSydhTfbsGm12878Gcn5StdPk.narrowPeak		wgEncodeSydhTfbsHelas3Tf3c110StdPk.narrowPeak	wgEncodeSydhTfbsHepg2Mazab85725IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Corestab24166IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Ikzf1iknuclaStdPk.narrowPeak		wgEncodeSydhTfbsHelas3Tr4StdPk.narrowPeak	wgEncodeSydhTfbsHepg2Mxi1StdPk.narrowPeak	wgEncodeSydhTfbsK562Corestsc30189IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Irf3IggmusPk.narrowPeak		wgEncodeSydhTfbsHelas3Usf2IggmusPk.narrowPeak	wgEncodeSydhTfbsHepg2Nrf1IggrabPk.narrowPeak	wgEncodeSydhTfbsK562CtcfbIggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878JundIggrabPk.narrowPeak		wgEncodeSydhTfbsHelas3Zkscan1hpa006672IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2P300sc582IggrabPk.narrowPeak	wgEncodeSydhTfbsK562E2f4UcdPk.narrowPeak
	wgEncodeSydhTfbsGm12878MafkIggmusPk.narrowPeak		wgEncodeSydhTfbsHelas3Znf143IggrabPk.narrowPeak	wgEncodeSydhTfbsHepg2Pol2IggrabPk.narrowPeak	wgEncodeSydhTfbsK562E2f6UcdPk.narrowPeak
	wgEncodeSydhTfbsGm12878MaxIggmusPk.narrowPeak		wgEncodeSydhTfbsHelas3Znf274UcdPk.narrowPeak	wgEncodeSydhTfbsHepg2Rad21IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Elk112771IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Mazab85725IggmusPk.narrowPeak		wgEncodeSydhTfbsHelas3Zzz3StdPk.narrowPeak	wgEncodeSydhTfbsHepg2Rfx5200401194IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Gata1UcdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Mxi1IggmusPk.narrowPeak		wgEncodeUwTfbsHelas3CtcfStdPkRep1.narrowPeak	wgEncodeSydhTfbsHepg2Smc3ab9263IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Gtf2bStdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Nfe2sc22827StdPk.narrowPeak			wgEncodeSydhTfbsHepg2TbpIggrabPk.narrowPeak	wgEncodeSydhTfbsK562Gtf2f1ab28179IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878NfkbTnfaIggrabPk.narrowPeak			wgEncodeSydhTfbsHepg2Tcf7l2UcdPk.narrowPeak	wgEncodeSydhTfbsK562Hcfc1nb10068209IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878NfyaIggmusPk.narrowPeak			wgEncodeSydhTfbsHepg2Tr4UcdPk.narrowPeak	wgEncodeSydhTfbsK562Hmgn3StdPk.narrowPeak
	wgEncodeSydhTfbsGm12878NfybIggmusPk.narrowPeak			wgEncodeSydhTfbsHepg2Usf2IggrabPk.narrowPeak	wgEncodeSydhTfbsK562Ini1IggmusPk.narrowPeak
	wgEncodeSydhTfbsGm1287			wgEncodeSydhTfbsHepg2Zn	wgEncodeSydhTfbsK562Jun

	8Nrf1IggmusPk.narrowPeak	f274UcdPk.narrowPeak	dIggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878P300IggmusPk.narrowPeak		wgEncodeSydhTfbsK562Kap1UcdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Pol3StdPk.narrowPeak		wgEncodeSydhTfbsK562MaffIggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Rfx5200401194IggmusPk.narrowPeak		wgEncodeSydhTfbsK562Mafkab50322IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Sin3anb6001263IggmusPk.narrowPeak		wgEncodeSydhTfbsK562Mazab85725IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Smc3ab9263IggmusPk.narrowPeak		wgEncodeSydhTfbsK562Mxi1af4185IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Spt20StdPk.narrowPeak		wgEncodeSydhTfbsK562NelfeStdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Srebp1IggrabPk.narrowPeak		wgEncodeSydhTfbsK562Nfe2StdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Srebp2IggrabPk.narrowPeak		wgEncodeSydhTfbsK562NfyaStdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Stat1StdPk.narrowPeak		wgEncodeSydhTfbsK562NfybStdPk.narrowPeak
	wgEncodeSydhTfbsGm12878Stat3IggmusPk.narrowPeak		wgEncodeSydhTfbsK562Nrf1IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878Tblr1ab24550IggmusPk.narrowPeak		wgEncodeSydhTfbsK562P300IggrabPk.narrowPeak
	wgEncodeSydhTfbsGm12878TbpIggmusPk.narrowPeak		wgEncodeSydhTfbsK562Pol2IggmusPk.narrowPeak

wgEncodeSydhTfbsGm12878Tr4StdPk.narrowPeak	wgEncodeSydhTfbsK562Pol3StdPk.narrowPeak
wgEncodeSydhTfbsGm12878Usf2IggmusPk.narrowPeak	wgEncodeSydhTfbsK562Rfx5IggrabPk.narrowPeak
wgEncodeSydhTfbsGm12878WhipIggmusPk.narrowPeak	wgEncodeSydhTfbsK562Rpc155StdPk.narrowPeak
wgEncodeSydhTfbsGm12878Znf143166181apStdPk.narrowPeak	wgEncodeSydhTfbsK562Setdb1UcdPk.narrowPeak
wgEncodeSydhTfbsGm12878Znf274StdPk.narrowPeak	wgEncodeSydhTfbsK562Sirt6StdPk.narrowPeak
wgEncodeSydhTfbsGm12878Znf384hpa004051IggmusPk.narrowPeak	wgEncodeSydhTfbsK562Smc3ab9263IggrabPk.narrowPeak
wgEncodeSydhTfbsGm12878Zzz3StdPk.narrowPeak	wgEncodeSydhTfbsK562Tal1sc12984IggmusPk.narrowPeak
	wgEncodeSydhTfbsK562Tblr1ab24550IggrabPk.narrowPeak
	wgEncodeSydhTfbsK562Tblr1nb600270IggrabPk.narrowPeak
	wgEncodeSydhTfbsK562TbpIggmusPk.narrowPeak
	wgEncodeSydhTfbsK562Tf3c110StdPk.narrowPeak
	wgEncodeSydhTfbsK562Tr4UcdPk.narrowPeak
	wgEncodeSydhTfbsK562Ubfsc13125Iggmu

sPk.narrowPeak
wgEncodeSydhTfbsK562Ubtfsab1404509IggmusPk.narrowPeak
wgEncodeSydhTfbsK562Usf2IggrabPk.narrowPeak
wgEncodeSydhTfbsK562Xrcc4StdPk.narrowPeak
wgEncodeSydhTfbsK562Zc3h11anb10074650IggrabPk.narrowPeak
wgEncodeSydhTfbsK562Znf143IggrabPk.narrowPeak
wgEncodeSydhTfbsK562Znf263UcdPk.narrowPeak
wgEncodeSydhTfbsK562Znf274UcdPk.narrowPeak
wgEncodeSydhTfbsK562Znf384hpa004051IggrabPk.narrowPeak
wgEncodeSydhTfbsK562Znfmizdcp1ab65767IggrabPk.narrowPeak

Table S3. Utilized datasets for reference-based discretization of gene expression data

Dataset	Reference	Accession
Cancer cell line background distribution	(Klijn *et al.*, 2014)	E-MTAB-2706 (Array Express)
Neuroepithelial differentiation	(Qiao *et al.*, 2015)	GSE68396 (GEO)
Microarray background distribution	(Torrente *et al.*, 2016)	E-MTAB-3732 (Array Express)

Published Papers

S. Zickenrott, V. E. Angarica, B. B. Upadhyaya and A. del Sol (2016), "Prediction of disease-gene-drug relationships following a differential network analysis", Cell Death and Disease, **7**

S. Okawa, S. Nicklas, **S. Zickenrott**, J. C. Schwamborn and A. del Sol (2016), “A generalized gene-regulatory network model of stem cell differentiation for predicting lineage specifiers”, Stem Cell Reports, **7(3)**, 307-315

S. Jung, A. Hartmann and A. del Sol (2017), “RefBool: a reference-based algorithm for discretizing gene expression data”, Bioinformatics, **33(13)**, 1953-1962

S. Jung, V. E. Angarica, M. A. Andrade-Navarro, N. J. Buckley and A. del Sol (2017), “Prediction of Chromatin Accessibility in Gene-Regulatory Regions from Transcriptomics Data”, Scientific Reports, **7**