

Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Put forward by
M. Sc. Christoph Kommer
Born in: Ravensburg
Oral Examination: 10.10.2018

STATISTICAL LEARNING BASED INFERENCE AND
ANALYSIS OF EPIGENETIC REGULATORY NETWORK
TOPOLOGIES IN T-HELPER CELLS

Referees: Prof. Dr. Thomas Höfer
Prof. Dr. Ursula Kummer

Zusammenfassung

Die verlässliche statistische Inferenz von epigenetischen regulatorischen Netzwerken, die das Zellschicksal bei Säugetieren bestimmen, ist eine äußerst anspruchsvolle Aufgabe. In dieser Arbeit behandeln wir diese Problemstellung im Rahmen von Differenzierungsentscheidungen von T-Helferzellen (Th Zellen), von denen gezeigt werden konnte, dass sie ein Kontinuum von differenzierten Zuständen in Abhängigkeit verschiedener Zytokinsignale annehmen können. Um die zugrundeliegenden regulatorischen Netzwerke zu bestimmen, führen wir eine neuartige Methode zur Inferenz epigenetisch regulatorischer Netzwerktopologien ein, die auf Methoden des statistischen Lernens basiert.

Zunächst bestimmen wir, mithilfe eines Hidden Markov Modells, Chromatinzustände die auf Histonmodifikationsmustern in naiven und differenzierten Th1, Th2 und gemischten Th1/2 Zuständen basieren. Diese Zustände werden durch externe Zytokinstimuli und die Gendosis des Master-Transkriptionsfaktors Tbet (*Tbx21*) bestimmt. Danach führen wir ein lineares multivariates Korrelationsmaß ein, welches der Zuordnung von Enhancern zu ihren Zielgenen dient. Dieses Maß wird anhand eines Satzes von bekannten Enhancern gelernt. Diese Analyse wird verfeinert durch die Anwendung partieller Korrelationen, um direkte von indirekten Effekten zu unterscheiden. Bei der Anwendung dieser Methode auf unsere Daten bestätigen wir zum einen bekannte Enhancer und erhalten zum anderen eine genomweite Zuordnung zwischen Enhancern und Genen. Dies erweitern wir zudem auf die Korrelation repressiver regulatorischer Elemente mit Genexpressionen.

Des Weiteren untersuchen wir Enhancer, die differentiell exprimierte Th1 und Th2 spezifische Transkripte regulieren. Mithilfe von Prädiktoren, die auf Methoden des maschinellen Lernens basieren, identifizieren wir Th1 und Th2 spezifische Enhancer-Klassen und solche repressiver Zustände, die durch ihre Reaktionsmuster auf Zytokinstimuli und auf die Dosis von Tbet charakterisiert werden. Außerdem verwenden wir Chromatin-Immünpräzipitationsdaten von Transkriptionsfaktoren, um die transkriptionelle regulatorische Logik, die die Aktivität der Enhancer-Klassen bestimmt, zu definieren.

Schlussendlich kombinieren wir die Zuordnungen von Enhancern zu ihren Zielgenen und sowohl die regulatorische Enhancerlogik als auch die von inhibitorischen Elementen, um ein bipartites epigenetisches Netzwerk zu erhalten. Die Netzwerkarchitektur basiert dabei sowohl auf Enhancer-Klassen und repressiven Zustandsklassen als auch auf Genen und Transkriptionsfaktoren, was zu gewichteten Multi-Digraphen führt. Die Netzwerktopologie offenbart ausgeprägte unterscheidbare Strukturen, die mit einer Funktionalität für Th1, Th2 und Hybrid-Zellen identifiziert werden können. Außerdem analysieren wir Multiplex-Netzwerke, was zu zellspezifischen Topologien führt. Aus diesen Analysen erhalten wir charakteristische Beiträge von einzelnen Knoten des jeweiligen Netzwerks. Mithilfe von Random Walks auf Multi-Digraphen gewinnen wir Informationen über metastabile Prozesse, die den beobachteten Systemen zugrunde liegen.

Zusammenfassend präsentieren wir eine robuste quantitative Methode, um Chromatinzustände und Genaktivität einander zuzuweisen und um epigenetische Netzwerke durch die Bestimmung von Transkriptionsfaktorregulierung von Enhancern zu lernen. Diese Vorgehensweise ist auf eine Vielzahl von Systemen anwendbar.

Abstract

The reliable statistical inference of epigenetic regulatory networks that govern mammalian cell fates is very challenging. In this thesis we study this question for the differentiation decisions of T-helper (Th) cells, which have recently been shown to adopt a continuum of differentiated states in response to cytokine signals. To infer the underlying regulatory networks we introduce a novel framework for the inference of epigenetic regulatory network topologies based on statistical learning.

First, we infer, via a Hidden Markov Model, chromatin states based on histone modification patterns in naïve Th cells and differentiated Th1, Th2 and mixed Th1/2 states; these states are controlled by external cytokine stimuli and the gene dose of the Th1 master transcription factor Tbet (*Tbx21*). We then introduce a linear multivariate correlation measure for mapping enhancers to their target genes, which is parametrized on a training set of known enhancers. This analysis is refined further by the application of partial correlations to distinguish direct from indirect effects. Applying this approach to our data, we recover known enhancers and obtain a genome-wide enhancer-gene mapping. We also extend this to the correlation of repressive regulatory elements with gene expression.

Next, we focus on the enhancers that regulate differentially expressed Th1 and Th2 specific transcripts. Building machine learning based predictors, we identify Th1 and Th2 specific enhancer and repressive state classes characterized by their response patterns to cytokine stimuli and Tbet dose. In turn, we use chromatin immunoprecipitation data of transcription factors to define the transcriptional regulatory logic governing the activities of the enhancer classes.

Finally, we combine enhancer-target gene maps and enhancer regulatory logic as well as inhibitory elements to infer a bipartite epigenetic network. The network architecture builds on enhancer and repressive state classes as well as on genes and transcription factors leading to a weighted multidigraph. The network topology reveals distinct community structures related to Th1, Th2 and hybrid functionality. We furthermore analyse multiplex networks resulting in condition-specific topologies. From these analyses we obtain unique contributions of distinct network nodes. Utilizing random walks on multidigraphs we extract metastable processes underlying the observed system.

In conclusion we present a robust quantitative framework for mapping chromatin states to gene activity, and, by factoring in transcription factor regulation of enhancers, inferring epigenetic regulatory networks. This methodology is applicable to a wide range of systems.

"It's..."

The It's Man
MONTY PYTHON'S FLYING CIRCUS

Contents

I	Introduction & Motivation	1
II	Fundamentals & Literature Review	5
II.1	Biological preliminaries	5
II.1.1	T-helper cells	5
II.1.2	Enhancers and epigenetic regulation of transcription	7
II.1.3	Gene regulatory networks	11
II.2	Graph and Network Theory	13
	Semantics	14
	Topological properties	15
II.3	Machine Learning	18
II.3.1	Decision Trees and Random Forests	18
II.3.2	Hidden Markov Models	20
III	Data sets: description and analysis	24
III.1	Underlying experimental data sets	24
III.2	Analysis of experimental data	26
III.2.1	Histone modifications	26
	Data quality control	26
	Alignment	26
	Peak search	27
	Methods	27
	Results	29
III.2.2	Gene expression	31
	Data quality control	31
	Alignment	31
	Differential expression analysis	32
	Methods	32
	Results	33
	Absolute expression analysis	34
	Methods	34
	Results	34
III.2.3	Data pre-processing pipeline	37
III.3	Discussion & Summary	38

IV	Inference of chromatin states in T-helper cells	39
IV.1	Pattern recognition of epigenetic states	39
IV.1.1	Method	40
IV.1.2	Results	44
	Model parameters	44
	Chromatin state annotation examples	46
IV.2	Discussion & Summary	51
V	Epigenetic landscape inference by implementation of a multivariate correlation measure model	53
V.1	A parametrized multivariate correlation measure	53
V.1.1	Optimization of correlation measure	53
	Method	54
	Discussion	57
V.2	Computational implementation	59
V.2.1	Preprocessing	60
V.2.2	Input	60
V.2.3	Correlation algorithm	62
	Segmentation of chromatin state elements	62
	Merging of statistically similar elements	64
	Actual correlation and output	65
V.2.4	Summary	66
V.3	Results	67
V.3.1	<i>Ifnγ</i>	68
V.3.2	<i>Tbx21</i>	71
V.3.3	Additional notable gene loci	73
V.3.4	Th2 cytokine locus	74
V.4	Partial correlations	77
V.5	Inference of inhibition	80
V.6	Prediction of gene expression	82
V.7	Discussion & Summary	84
VI	Establishing cell-type-specific enhancer classes	86
VI.1	Introducing a typology of enhancer states	86
VI.2	Inter-class specificity of enhancer types	87
VI.2.1	Method	88
VI.2.2	Results	89
VI.2.3	Classification of transcript specificity	91
	Results	91
VI.3	Intra-class specificity of enhancer state classes	92
VI.3.1	Method	92
VI.3.2	Results	94
VI.4	Co-occurrence of enhancer state classes	97
VI.5	Discussion & Summary	97

VII	Epigenetic network inference and analysis	99
VII.1	Typology of underlying network and adjacency matrix	99
VII.1.1	Definition of network components	99
	Nodes	100
	Edges	101
	Inference of TF binding at CSCs	102
	Adjacency matrix	102
VII.2	Network analysis	104
VII.2.1	Full CSC-gene network	104
VII.2.2	Network properties and metrics	105
	Degree distribution and hubs	105
	Alternative centrality measures	108
	Motifs and loops	112
	Attack tolerance	113
VII.2.3	Core CSC-TF network	114
VII.2.4	Validation of known and prediction of novel TF connections . .	115
VII.2.5	Network communities	118
	Methods	118
	Results	118
VII.2.6	Multiplex networks	121
VII.2.7	Differential network analysis	126
VII.3	Random walks on weighted multiplex multidigraphs	130
	Features of the eigenvalue spectra of stochastic transition matrices	134
	Metastability of the epigenetic Tbet/Gata3 network motif . . .	141
VII.4	Discussion & Summary: Topology and function	145
VIII	Future directions & improvements	148
IX	Summary	152
IX.1	General	152
IX.2	Originality of work	155
A	Additional mathematical background	157
A.1	Definitions and explanations	157
A.1.1	Statistics	157
	Model selection criteria	157
	Adjusted p-value	158
A.1.2	Partial correlation	158
A.1.3	Perron-Frobenius-Theorem for ergodic Markov chains	159
A.2	Mathematical Notation	160
B	Supplementary Figures	161
C	Supplementary Tables	173

D	Code documentation	227
D.1	Algorithmic commands	227
D.2	Correlation algorithm	227
D.3	Class-specificity computation	231
D.4	Computational dependencies & packages	231
	Acronyms	233
	Bibliography	235
	List of Figures	265
	List of Tables	269

CHAPTER I

Introduction & Motivation

The rise of analytical and computational methodology provided important insights into a great variety of complex cell biological processes in the last decades leading to novel approaches in dealing with experimental biological data. Especially concerning the rise of high throughput sequencing technologies the importance of reliable mathematical and computational approaches to deal with highly heterogeneous data sets on a genome-wide scale cannot be overstated. This leads to the possibility of globally exploring e.g. gene expression, transcription factor binding or epigenetic changes to DNA sequences.

Especially in the context of epigenetic regulation of transcription constituting a highly complex field on its own little is known concerning the respective epigenetic landscape and the mapping of candidate elements actually regulating the transcription of single genes from within this landscape in particular cell types. This is obviously a vital question for the actual determination of cell fate as the genome-wide epigenetic landscape can change drastically between different cell types and even under slight perturbations. The knowledge of the underlying regulatory behaviour hence is necessary for the construction of sufficiently complex regulatory networks in order to capture the unique underlying topology of a respective cell lineage under certain environmental conditions and external stimuli. In turn this serves as the basis of building reliable mathematical models since very reduced models often do not capture the exact behaviour observed in the underlying experiments. This is not only important for the prediction of the expression of individual genes but also in terms of the global stability behaviour of a certain cell type.

An experimental system of particular interest for the investigation of its epigenetic regulatory network structure is given by T-helper cells of type I or type II (see section [II.1.1](#)), commonly called Th1 or Th2 cells, belonging to the class of main drivers of the adaptive immune system. In recent years it became an increasingly well established viewpoint to extend the classical dichotomy of a binary T-helper cell fate to a larger variety of distinct lineages, which furthermore exhibit plasticity under certain environmental conditions. This even led to the discovery of stable long-lived steady states in between these lineages forming so-called hybrids as in the case of Th1/2. Furthermore experimental evidence suggests [[14](#), [111](#), [142](#), [255](#), [256](#)] that an even larger amount of stable steady states might exist in between the binary cell fates depending on external stimuli and cell culture conditions adding valuable information to the plasticity behaviour w.r.t. the classical lineages. It turns out that simple regulatory motifs, such as toggle switches, are only to a limited extent able to de-

scribe these circumstances appropriately especially w.r.t. a larger number of possible steady states. A remedy to this issue might lie in the unique topology of the regulatory network underlying transcriptional regulation in Th1 and Th2 cells.

It turns out that the reliable quantitative inference of gene regulatory networks that account for epigenetic processes as well such as enhancer regulation or equivalently inhibitory actions w.r.t. gene expression is a complex problem that also calls for reliable quantitative mathematical and computational approaches. As we want to focus on an unbiased data-driven approach methodology from the field of statistical learning is an essential element in determining the components of a respective epigenetic network. One not only faces the problem of e.g. finding viable candidates for enhancers or epigenetic regulatory elements in general but also of finding one-to-one relations between those elements and the genes, which are regulated by them, and approaching this in a quantitative way. Another important question in this context actually is if there is a possibility to distinguish these elements w.r.t. their specific regulatory task, i.e. if a certain element plays an important role in a certain regulatory context as for example enhancer activity can be highly cell condition dependent. Hence the information if certain elements differ in their activity patterns is vital for a description of the regulatory landscape and can yield important insight on cell-specificity or lineage determination. Finally the question is to what extent we can recover Th1 and Th2 lineage-specificity from a resulting network and to what extent the underlying topology provides information on stable steady states. In addition to that epigenetic regulation can give insight into new regulatory relations between genes and more specifically between transcription factors themselves that are usually not taken into account by simple regulatory motifs. Another question is if those investigations can be extended even further to condition-specific subnetworks that exhibit themselves unique steady state topologies. In that context the formation of certain structural entities such as communities of genes or *cis*-regulatory elements are able to provide information on lineage specificity as well as on the particular importance of certain regulatory players that can influence a certain cell-specific process more than others.

We point out that there are little straightforward or even generally established approaches to deal with the above stated issues. Although there are methods to e.g. partition the DNA into so-called chromatin state elements there exist only little data-based quantitative analysis methods of co-regulation of these elements with gene expression or more specifically of inferring a one-to-one mapping quantitatively. To our best knowledge especially a quantitative data-driven measure for regulation via histone modification data is still completely missing. The same holds true for an analysis of epigenetic regulators in terms of their regulatory logic especially w.r.t. cell-specificity. Furthermore the analysis of GRNs only rarely includes advanced analysis concepts such as community detection or investigations on metastability, which are important in discovering topological properties. Especially in the context of Th1 and Th2 there is no actual formulation of an epigenetic enhancer network including enhancers as nodes themselves as well as inhibitory elements and hence making predictions about relevant *cis*-regulatory entities on a genome-wide scale. All of these are issues that have to be approached in order to gain a deeper understanding of unravelling plasticity as well as lineage-specifying properties in the topology of

the regulatory structure in Th1 and Th2 cells.

The goal of this work hence shall be to establish a unique and novel computational and mathematical methodology for inference of epigenetic networks in general and for Th1 and Th2 cells in particular. Additionally the underlying network topologies in the model system are investigated leading to valuable implications for stable steady states of differentiated cells. We also want to follow up on investigations into multistable steady states of models of enhancer networks in general, which serve as a starting point for discussion of the breaking of the bistable Th1/Th2 cell dichotomy. It is important to note that the whole methodology is adaptable to a wide range of model systems with ease and efficiently gives insight into the architecture of regulatory enhancer and repressor control in gene networks.

The following questions will be among those to be addressed in this work:

- How does the epigenetic landscape look like, especially w.r.t to enhancers in Th1 and Th2 cells as well as in Th1/2 hybrids, on a genome-wide scale?
- Is there a quantitative way to robustly infer a one-to-one mapping between enhancers and genes?
- What is the effect of a master transcription factor such as Tbet on the epigenetic landscape?
- Do enhancers differ in their regulatory logic and what impact does this have?
- What is the structure and the topology of the underlying epigenetic networks?
- Do we find unique topologies that play a major part in terminal cell differentiation including hybrid cells and do we find hints on multistability?

The layout of the thesis will hence be as follows: In chapter **II** we will introduce fundamental concepts and give a literature overview on the current research status in the fields relevant for this work. Chapter **III** will introduce the underlying experimental data sets in Th1, Th2 and hybrid Th1/2 cells with additional perturbations and cover their analysis in some detail as this is a crucial step in explaining how subsequent results are obtained. The following chapter **IV** focusses on epigenetic pattern detection in the respective experimental conditions w.r.t. histone modifications. This will employ the usage of so-called Hidden Markov Models, which results in a candidate detection of e.g. enhancer and repressive states. As these results only solve as a prior we are aiming on a quantitative method in order to refine these predictions in a sophisticated way. To this end we introduce a multivariate parametrized histone modification model, which is learned from the underlying data for correlation with gene expression in chapter **V**. In addition we also implement a sophisticated correlation algorithm in order to extract unique regulatory segments statistically leading to graded correlating elements. We test the viability of the method at different genomic loci and make genome wide predictions for the epigenetic landscape around notable Th1 and Th2 genes. This includes a discussion on the impact a master transcription factor like Tbet can have on the epigenetic landscape in Th1 cells. We also discuss the

potential of distinct regulatory entities such as different types of enhancer activation patterns depending on the cell conditions, which results in uniqueness w.r.t. lineage specificity in chapter VI. This leads to predictions that can be made on the specificity of a certain gene transcript based on its respective epigenetic surroundings. To this end we also introduce a new measure that accounts for cell-specificity of a certain epigenetic element but is in general applicable to any sort of classification problem. Finally we give a full account on the inference of actual epigenetic networks in Th1 and Th2 cells and discuss their topological properties in chapter VII. This includes the determination of resulting regulatory relations between notable transcription factors revealing new regulatory patterns as well as the inference of epigenetic as well as genetic clusters within the network. We will extend this even further to condition-specific multiplex networks and differential networks and finally propose node relevance rankings being unique to the respective topologies. The chapter ends with a discussion on metastability inference via random walks on directed multi-digraphs with implications on Th1/Th2 plasticity. Future perspectives will be given in VIII also discussing the implications on further investigations of multistability.

CHAPTER II

Fundamentals & Literature Review

II.1 Biological preliminaries

To set up the foundations for the observed experimental system especially focussing on T-helper cells and their epigenetic landscape, we will discuss in the following basal foundations on T-helper cell and enhancer biology as well as the current research status on gene regulatory and epigenetic networks. For later discussion in the main part of this thesis we will also introduce additional important concepts, which will be used throughout the following work.

II.1.1 T-helper cells

In general *T-cells*, also called *T-lymphocytes*, are the key players of adaptive cell-mediated immunity. The name stems from their maturation origin in the thymus. They consist of a class of subtypes, which fulfill different immune response functionality. Among the most notable subtypes are the supersets of *effector* and *memory* T-cells with their subsets of *cytotoxic* (or *killer*) as well as *helper* T-cells (see e.g. [5, 229]). Effector T-cells are responsible for the short-term immune response, while memory T-cells provide a long-term protection upon subsequent infection. Mature T-helper cells in particular express the surface glycoprotein CD4 which leads to their denomination as belonging to the class of CD4+ T-cells. The name “helper” cell stems from the fact that they assist other lymphocytes such as cytotoxic T cells, B cells or macrophages in their activation process by secretion of specific cytokines, hence playing an important role not only in the cellular but indirectly in the humoral immune response as well. In addition there also exist *regulatory* T-cells (T_{reg})¹ [173], which suppress effector T-cells and are part of CD4+ cells as well.

T-helper cells can in turn themselves be divided into functionally different effector lineages or subtypes, depending on their specific cytokine secretion profiles. Naïve T-helper cells represent immature undifferentiated basal T-helper cells which have never seen an antigen from some antigen-presenting cell via their T-cell receptors (TCR) for recognition of a foreign pathogen, according to which they eventually differentiate to antigen-specific T-helper cells. The classic bistable differentiation paradigm of naïve T-helper cells into those of type 1 and type 2 (Th1 and Th2)

¹formerly frequently being termed suppressor cell [181] and often denoted as iT_{reg} for *induced*.

[1, 227, 230] has been extended in recent years to even more terminally differentiated subtypes which are called Th17 [137] or follicular T-helper cells (Tfh)² [79] and recently introducing Th9 and Th22 cells [262, 284]. These subtypes characterized by long-lived stable cell populations exhibit distinct phenotypes with characteristic cytokine expression profiles as well as with so-called specific master transcription factors (TFs) as shown in table II.1 being thought as necessary and sufficient for their respective cell fate [151, 152, 159, 164, 204, 270, 288, 298, 367].

	Th1	Th2	Th17	T _{reg}	Tfh
Master TF	Tbet	Gata3	ROR γ t	Foxp3	Bcl-6
Other TFs	STAT4	STAT6	STAT3	STAT5	
	STAT1				
Effector	<i>Ifn</i> γ	<i>Il4</i>	<i>Il17</i>	<i>Il10</i>	<i>Il21</i>
Cytokines		<i>Il5</i>	<i>Il22</i>	<i>TGF</i> β	
		<i>Il13</i>	<i>Il21</i>		
			<i>Il25</i>		

Table II.1: Master TFs and signature cytokines for the most notable T-helper and CD4+ subtypes.

It has e.g. been shown in vitro as well as in vivo that upon Tbet knock-out T-helper cells show significant defects in Th1 differentiation [300] whereas the same holds true for Gata3 knock-out concerning Th2 cell differentiation [370]. Extending the classic Th1/Th2 dichotomy the larger range of phenotypically stable subtypes forms a more refined picture w.r.t. immune responses and autoimmunity³ [132].

In addition to their central role in cytokine signalling signal transducers and activators of transcription (STATs) have been found to play a major part in activation of TF-coding genes and a large range of lineage-specifying loci [247] especially in the context of epigenetic regulation [314] extending the exclusive regulatory impact of master TFs [242].

In general cytokines can be separated into being pro-inflammatory or anti-inflammatory. Th1 cells, playing a major part in cellular immune responses, are often associated with pro-inflammatory tasks, e.g. in the context of autoimmunity [299], while Th2 cells being important e.g. in helminth infections are associated with anti-inflammatory tasks and are linked to allergic responses [311, 322]. As an optimality principle a balance between respectively counteracting and mutually exclusive Th1 and Th2 responses is desirable for a functioning immune response. Yet these mutual exclusive viewpoints have been challenged over the years (see e.g. [233, 280, 297]), which is where for example other T-helper subtypes come into play.

Yet another important aspect in this context is the notion that emerged in recent years that extends the above concepts even more by additionally showing flexibility between T-helper subtypes leading to the possibility of reprogramming from one subtype to another as well as exhibiting plasticity even resulting in mixed hybrid phenotypes [42, 180, 234, 248]. This view departs now from the classic concept mutual

²We note that the categorization of Tfh cells as a distinct subtype is a matter of ongoing debate [68].

³We note for completion that selective cytokine production can be exerted by other cell types as well such as by innate lymphoid cell lineages [95].

exclusive terminally differentiated lineages. This is partially due to the fact that although specific signature genes are to a certain extent exclusive to a distinct subtype there still exists in many cases some residual expression in other subtypes, like e.g. in the case of *Il10* or even *STAT4* [66, 349]. This leads e.g. to the possibility of (partial) reprogramming of *Tbx21*⁺ and *Ifn γ* ⁺ Th1 cells to *Gata3*⁺ cells and vice versa (see e.g. [39, 142, 255]) maintaining steady states in memory phase. In addition to these Th1/2 hybrids [160] analogous cell hybrids have also been demonstrated experimentally for Th17/Th1, Th17/*T_{reg}* as well as Th17/Th2 (see e.g. [131, 260]). Yet arguably the plasticity w.r.t. the classic antagonistic Th1-Th2 dichotomy is among the most interesting cases when it comes to reprogramming and plasticity phenomena. Recently also a continuum of hybrid Th1/2 was proposed as well depending on the respective cytokine environment (see e.g. [14, 111, 255]) leading to important questions of the underlying regulatory mechanisms in Th1/Th2 systems and leading to the need of extending ordinary models of bifurcation extensively.

In order to shed more light on these processes a crucial point is to elucidate the regulatory mechanisms for gene transcription. At this point epigenetic modifications come into play extending the limited picture of merely investigating protein-coding genes. Although TF binding in T-helper cells has been investigated in some detail [334, 369] it has also been shown that not all genes bound by a certain TF are in fact regulated by it [369] and not all genes that require a certain TF do exhibit binding of that particular TF [334]. The investigation of epigenetic landscapes of T-helper cells hence forms an important field of investigation and will be sketched at the end of the following section.

II.1.2 Enhancers and epigenetic regulation of transcription

In eukaryotes there exists a large number of epigenetic mechanisms which affect gene regulation without altering the DNA sequence itself [67]. The means to achieve this is mostly discussed in terms of post-translational modification (PTM) of histone octamers and DNA methylation in combination with transcription factor binding. While DNA methylation of CpG is associated with silencing of transcription (see e.g. [80]) in eukaryotes a key player of gene transcription has been shown to be given by *cis*-regulatory elements⁴ like promoters, enhancers, silencers or insulators, possessing the ability to e.g. enhance or silence transcription of a certain gene⁵ [208, 346]. In fact they can be very specific to a certain cell type or environmental condition [10, 148, 362]. Among the most prominent elements are so-called *enhancers*, which are defined as being non-protein coding *cis*-regulatory DNA sequences that can themselves lie upstream, downstream or even in the intronic region of the gene whose promoter it is supposed to regulate positively [60, 257, 289]. An enhancer element subsequently facilitates the binding of transcription activating TFs to sequence motifs located at the enhancer, which in turn influences enhancer activity [333]. In other words the likelihood for transcription of the respective gene is eventually higher than

⁴*Cis*-regulatory elements facilitate intramolecular interactions opposed to intermolecular ones of *trans*-regulatory elements such as TFs.

⁵We note that e.g. transcriptional enhancers and silencers also appear in prokaryotes yet are not the main drivers of transcription [193].

without an enhancer regulating the respective gene promoter. Enhancers have been shown to lie up to 1 Mbp away from the gene it eventually regulates [257] and even are capable of regulating multiple genes at the same time. It is believed that several hundreds of thousands of enhancers (see ENCODE [303] enhancer identification study in humans (e.g.[144])) might exist in the human genome opposing around 20.000 protein encoding genes [110]. Hence a far larger number of cell-specific transcription enhancing regulatory elements regulate a smaller number of genes. Enhancers exhibit the important feature to bind large varieties of lineage-specifying as well as general transcription factors, which can bind to enhancer sites via cooperative action or via so-called pioneering factors that open up the closed heterochromatin of DNA wrapped tightly around the respective nucleosomes and providing the euchromatin necessary for *cis*-regulatory events inducing transcription [60].

Attempts to assign a certain sequence code to enhancers in general have failed, which calls for other indicative mechanisms of enhancer existence. Since in order to ensure TF binding one needs nucleosome-depleted chromatin one such marker is given by so-called DNase hypersensitive sites which are highly sensitive e.g. for the DNase I enzyme [48, 327]. Yet more importantly modification of histone tails especially in the case of histone 3 lysine 1 monomethylation (H3K4me1) has been first associated with *cis*-regulatory regions on a genome-wide scale [145]. While H3K4me1 is not entirely exclusive to enhancers it is often a prerequisite for later nucleosomal depletion. In that case the mere appearance of H3K4me1 at an enhancer poising the element for possible transcriptional activity coins the term of so-called *poised enhancers*⁶ [43, 78, 144].

The recruited TFs at enhancers yet only are able to influence the transcriptional process via the help of coactivator proteins. Among the most notable ones are so called histone acetyltransferases (HATs), which transfer acetyl groups to histone tails with histone acetylation in general being associated with the formation of euchromatin [15] by lowering the overall charge on the respective nucleosome. The most notable HAT in this regard is p300/CBP [128, 178] which has been used frequently for enhancer identification⁷. Among its most important acetylated targets is histone 3 lysine 27 leading to H3K27ac [169][253] being the most notable marker of enhancer activity. This is especially due to the fact that in certain cases poised enhancers can bind p300/CBP although not yet actively influencing transcription [267]. This observation leads to the conclusion that H3K27ac presents a more reliable indicator of enhancer activity compared to p300. In comparison with H3K4me1 it is acknowledged that the latter is rather supposed to be a prerequisite for enhancer activity by subsequently acquiring H3K27ac [43], while after potential loss of its activity mark H3K4me1 frequently pertains at enhancer sites.

Other notable histone modifications associated with important *cis*-regulatory functionality are e.g. H3K4me3 appearing predominantly at active promoters [72]⁸,

⁶Sometimes poised enhancers also comprise H3K27me3 in the literature yet we will rather stick in these cases to the term *bivalent* enhancer.

⁷A notable study in T-helper cells can be found in [314].

⁸It has been furthermore shown that variation in cell-specific gene expression can be more ade-

H3K9me3 correlating with gene silencing [29] and H3K27me3 with gene repression [60]. While histone modifications have been generally shown to form broad peaks on the DNA [25, 329] especially H3K27me3 stands out in this regard. The reason for this is thought to be the stabilization of the post-translational modifications during cell differentiation [32].

Additionally for the case of H3K27me3 simultaneous occurrence in promoter regions with H3K4me3 has been reported frequently resulting in so-called *bivalent* promoters exhibiting activating as well as repressive potential and has been shown to be especially important for pluripotency e.g. in embryogenesis [318]. The same has also been shown for enhancers with the antagonists being H3K4me1 and H3K27me3 (e.g. [31, 318]).

The actual mechanism of enhancing transcription has been assumed to be performed by the formation of 3D contact DNA loops between the enhancer itself and the respective promoter facilitated by insulator-associated CTCF and cohesin binding in combination with a so-called mediator complex [177]. Activator TFs then interact with this mediator complex leading among other things to RNA polymerase II recruitment and finally to gene transcription. DNA-looping after all has been found to be even sufficient for the transcription of genes [86].

The realization of such loops can now be achieved via actual protein complex binding [257] or as well via diffusion of the TF and protein complexes to the promoter. The formation of these chromatin loops has been experimentally shown by chromatin conformation capture assays (3C) or similar methods (4C, 5C or Hi-C) (see e.g. [83]). In addition to this conformation capture and especially Hi-C data has elucidated that enhancer-promoter interactions are mainly restricted to so-called *topologically associating domains* (TADs) (see e.g. [89, 174, 240, 263, 275]). TADs, forming larger 3D loops than the above mentioned enhancer-promoter connections and also often comprising a large number of neighbouring genes, are widely conserved over different cell types as well as species [175]. Again CTCF and cohesin binding significantly correlates with TAD formation and it has been shown that the removal of a TAD boundary enables new enhancer-promoter interactions to form [175, 277].

Also the processing of transcriptional information by enhancers still constitutes a very debated subject in general considering either so-called *enhanceosome* or *flexible billboard* models [17, 293] where in the former case transcription is achieved via cooperative TF action at an enhancer site while in the latter case separate functional units within one enhancer regulate gene expression independently.

Turning to T-helper cells the relevance of H3K4me3 as well as of the repressive mark H3K27me3 was shown in [334] especially in their bivalent combination in concert with master transcription factor binding of Tbet and Gata3 resulting in a switch between Th1 and Th2 cell fates respectively depending on the respective cytokine environment [179]. Among important studies of enhancers in Th1 and Th2 cells is the Th1-specific *Ifn γ* locus, being devoid of enhancer looping in naïve T-helper cells, extensively described in [19, 74]. Another example of enhancer regulation in CD4+ cells, although not being as well annotated as the *Ifn γ* case, is the Th2 cytokine locus with Th2-specific interleukins like *Il4*, *Il5* and *Il13* relying mostly on DNase HS I data (see

quately described by enhancer marks than with gene promoters [246].

[13, 206, 292, 302, 344]). A general account of epigenetic regulation in CD4+ cells can be found in [237, 344]. In Th1 cells it has been claimed that Tbet only exhibits limited ability when it comes to shaping the enhancer landscape, which is rather supposed to be driven by STATs [314]. To what extent this really holds true will be part of our subsequent investigations. Although given the above evidence a concise mapping of enhancers in CD4+ cells and especially in Th1 and Th2 cells is still very fragmentary and necessarily calls for genome-wide accounting.

Another widely discussed concept concerning enhancers and cell-specificity is the concept of so-called *super-enhancers*, which have been found to occur around a number of genes in different cell types (most notably embryonic stem cells) that drive cell identity [154, 155, 226, 341]. Most commonly super-enhancers are identified via ChIP-Seq (chromatin immunoprecipitation DNA sequencing) data either for p300/CBP or H3K27ac requiring them to lie above a certain cutoff elevation in the exponential rise of the ranked integrated read load (see e.g. [264]). An extensive account of super-enhancers can be found in databases such as dbSUPER [184] yet algorithmic implementations for automatic determination in epigenetic omics data are e.g. given in HOMER⁹ [147].

Although genome-wide analyses via massive amounts of high-throughput data became increasingly popular in recent years (e.g. the combination of ChIP- and RNA-Seq data [12]) and epigenomes have been extensively mapped by the ENCODE and NIH Roadmaps Epigenomics projects (see e.g. [33, 144]) one still faces the basic problem in determining epigenetic landscapes specific to certain experimental conditions and of how exactly to determine unique epigenetic states such as enhancers in general and assign them to certain genes. Due to the potential differential regulation of epigenetic landscapes this is of utmost importance in the elucidation of pluripotency and reprogramming properties of distinct cell types including specific underlying regulatory networks (see e.g. [197, 325]). While early approaches included mapping of so-called conserved non-coding sequences (CNS) [56] not all CNS sites correspond to enhancers and many enhancers do not exhibit any or only slight conservation across different species [257, 283]. Various approaches have been developed over the years ranging from naïvely facilitating overlaps in DNase HSI and p300 binding data in combination with TF and histone modification data [78, 217, 320, 342] to more ingenious approaches via statistical assignment of epigenetic states and unsupervised machine learning state predictions from histone modification peak data (see e.g. [81, 103, 115, 187, 220, 225, 338]). Although genome-wide epigenetic state or enhancer predictions are made possible via these methods unique mappings between these regulators and genes are rarely made in a quantitative way. Most commonly only a nearest-neighbour assignment is performed (see e.g. [104, 163, 315]) completely omitting the possibility of an enhancer being located several genes away from its actual target. While restrictions to TADs and the facilitation of e.g. conformation capture methods or even concrete experimental validation of a certain enhancer provides the possibility of either narrowing down the respective mapping possibilities or actually prove them individually a genome-wide quantitative predictive approach is still a hard task to be solved and is of special interest in constructing reliable

⁹<http://homer.ucsd.edu/homer/>

regulatory networks, which we will discuss in the following.

II.1.3 Gene regulatory networks

As we have indicated gene regulation can take place on a variety of different levels e.g. on the genome, epigenome, transcriptome or proteome level. Depending on the experimental context and on the questions asked different layers have to be integrated in order to obtain regulatory circuits resulting in a so-called *gene regulatory network* (GRN) in which genes, among other regulatory players, represent the system's nodes, often via gene expression measurements. Especially the rise in importance of large varieties of omics or high-throughput measurement data over a broad range of species and cell-types calls for reliable approaches in network inference as well as in network analysis and modelling (see e.g. [219]).

Although GRN nodes often represent the expression of a certain gene this does not necessarily have to be the case as nodes are generally rather interpreted as some arbitrary functional entity and can as well be represented by different regulatory processes other than genes¹⁰ whereas network edges represent regulatory interactions.

Obviously the underlying experimental data sets determine a crucial ingredient in the inference of a GRN of a certain biological system. In general this also determines if one observes a dynamic (in the case of time-series data) or a static network¹¹. A crucial feature to achieve viable GRN reconstruction is to have perturbation experiments, e.g. gene knock-out, in order to detect regulatory relations between nodes in a GRN (see e.g. [98, 141, 219]). Naturally the quality of the data as well as the integration of different data sets can influence the accuracy of the resulting GRN considerably (see e.g. [122, 138, 141, 205, 326]). In general the respective system structure with its regulatory connections as well as the respective parameters or edge weights can be subsequently learned from the underlying data. The task of proper GRN inference hence boils down to finding an optimal network topology w.r.t. considered nodes and edge connectivity, which is able to explain the underlying data best and optionally take existing knowledge to some extent into account.

There exists a large range of individually tailored approaches for the inference of a certain GRN architecture, e.g. correlation networks [295], information theory or regression-based networks as well as probabilistic models or neural networks (see e.g. [65, 182, 271, 306, 323, 324, 365]). Furthermore models can be either deterministic or stochastic and lead to directed or un-directed graphs.

Correlation GRNs most frequently depend on coexpression of a set of genes by determining their respective pairwise correlation coefficients. This naturally can lead to a large number of false positive relations, which can be remedied to some extent by approaches such as partial correlation networks [170, 202, 272, 319, 361, 372]. Obviously another shortcoming of a naïve treatment of correlation networks without further data integration only yields undirected graphs. Nevertheless a rather popular method is presented by the so-called weighted gene co-expression network analysis

¹⁰as in general in transcriptional regulation including *cis*- as well as *trans*-regulators

¹¹Since we will be mainly interested in steady-state behaviour within the T-helper cell context in our subsequent analysis we will disregard dynamic network models.

(WGCNA) [198] posing an intuitive means to obtain weighted GRNs.

Information theory based networks (see e.g. [55]) utilize measures such as Euclidean distances, mutual information or maximum entropy yet in general also only providing undirected graphs.

Popular regression-based methods include in their simplest form e.g. Boolean networks which use the structure of a Boolean gene response¹² in a certain experimental condition in order to infer a regulatory logic (see e.g. [46, 176, 195]). The obvious shortcoming of these approaches is that in general one cannot easily distinguish between graded or continuous gene expression responses since the modelling threshold for a gene to be “on” or “off” depends sensitively on the chosen Boolean response function. Continuous versions for dependence estimation between nodes are given by linear as well as by non-linear regression methods (see e.g. [121, 141, 196, 238, 245, 278]).

Considering probabilistic graphical models such as Gaussian (see e.g. [123]) or Bayesian networks [122, 336]) gene structure and parameter relations are inferred based on random variable distributions. In the case of Bayesian models problems can arise with the choice of a proper prior acyclic graph model. Additionally feedback loop structures are only allowed in dynamic Bayesian networks in contrast to static ones [141, 259, 316, 360].

Obviously an important concern in inferring GRNs is which nodes to include for the system under consideration. To this end important features within a potential GRN are frequently extracted via co-expression clustering¹³ [44, 317] or differential expression analysis methods (see e.g. [326]), which is an essential step in data-based feature selection. Additional information on GRNs can be gained by selection of putative or well-known TFs (e.g. via CHIP-Seq data) of the system and include significantly regulated targets [30] or to alternatively allow for computational TFBS inference [168, 301]. Additionally frequently GRN properties such as sparseness, requiring that only a small number of genes, i.e. hubs, act as TFs [16, 44, 285] and scale-freeness [45, 244], increasing the robustness of the underlying topology as is observed in biological systems [26, 172], are imposed on reverse-engineering networks. Yet although sparseness and even more so scale-freeness emerges as a frequent organizing principle [16, 26, 172, 244, 358] it is not always evident if one can impose these requirements on the system under consideration.

Recurring network motifs are also an integral part of GRNs [224] and often act as a simplifying assumption to model e.g. stability in binary cell differentiation via feedback-loop motifs such as bistable toggle switches or even for tristability via the MISA (mutual-inhibition and self-activation) motif [11, 130]. One of the most prominent examples is given by the Gata1/PU.1 system in erythroid/myeloid differentiation (e.g. [364]) but has been shown for a variety of other systems such as the Tbet/Gata3 system for Th1/Th2 cell differentiation (see e.g. [160, 221]), which can be shown to exhibit at least three possible long-lived steady states. It has also been shown that certain systems are even able to exhibit higher multistability based on their respective underlying network motif as in the case of multisite phosphorylation

¹²i.e. a gene is either expressed or not

¹³For a general account on clustering methods see e.g. [348].

[307] and microbial signalling systems [192] with additional accounts being given e.g. in [149, 150, 183]. A general account on important basic motifs in transcriptional regulation is e.g. given in [6, 112, 119, 291].

Epigenetic regulatory networks also gained more and more attention recently facilitating the potential to incorporate different levels of regulatory transcriptional logic leading to unique network topologies and exhibiting a mechanistic way to infer feedback loops in networks. Of special interest in this regard are enhancer based GRNs (see e.g. [62, 217, 274]), which among other things are facilitated to infer transcription factor networks such as e.g. in CD8+ T cells [357]. Although preliminary attempts have been made to unravel actual epigenetic histone modification networks [140, 202, 215, 258, 359], histone modifications and especially epigenetic states are in general not directly included as regulatory entities within GRNs but rather indirectly via TF binding, which obviously is a shortcoming when trying to elucidate the role of distinct binding sites within GRNs.

As indicated above basic GRNs including well-known TFs also have been investigated in the context of T-helper cell networks in order to decipher lineage-specification (see [261]). Although having been widely acknowledged that master TFs in general are the main players in T-helper cells such as Tbet and Gata3 in the Th1/Th2 system¹⁴, recently this view has been challenged by proposing a major role of STATs within the respective GRN especially w.r.t. the underlying epigenetic landscape [314]. Moreover epigenetic regulation plays an increasing role of inferring regulatory elements in T-helper cells in order to obtain topologies enabling a functional explanation of plasticity properties [288, 309]. Although various efforts have been made in elucidating epigenetic regulatory elements in T-helper cells (see e.g. [153, 282, 309, 314, 344]) there has actually never been a genome-wide account on regulatory epigenetic or more specifically enhancer networks in Th1 and Th2 cells let alone an actual inferred network. Hence in order to unravel the underlying regulatory logic in Th1 and Th2 cells and to shed some light onto the formation of hybrid Th1/2 cell states and potential multistability or even a continuum of hybrid steady states [14, 97, 111] we will focus especially on inferring and analysing epigenetic regulatory networks in this particular system.

II.2 Graph and Network Theory

In the following we are going to introduce some basic notions of graph and network theory in order to set the foundation for chapter VII covering not only semantics but also advanced concepts such as community detection and multilayer networks. General introductions to graph theory can be found in [88][41] while for network theory we refer the reader to [22, 38, 84, 85, 107, 108, 236]. We furthermore note that although graph and network theory share most of their respective concepts a certain graph can be realized in a variety of different networks describing a range of distinct systems.

¹⁴An account of the MISA motif in Th1/Th2 cells is e.g. given in [221].

Semantics

A graph is generally defined via a pair of sets $G = (V, E)$ where the set V denotes the graph *vertices*, in the network context more frequently called *nodes*, and E denotes the *edge* set, in networks often called *links*. G' is called a subgraph of G if $G' \subseteq G$. Graphs can be grouped into different families ranging from complete graphs where each vertex is connected to every other vertex in a graph via an edge, to cyclic, star and *bipartite* graphs¹⁵. Bipartite graphs consist of two disjoint and independent vertex sets V and W with an edge only connecting vertices in-between the two sets. An important property of bipartite graphs is that they cannot exhibit odd-length cycles¹⁶.

Another distinction can be drawn between directed (so-called *digraphs*) and undirected graphs, where in the former case an edge between two vertices has a direction leading to ordered vertex pairs $\{v_i, v_j\} \neq \{v_j, v_i\}$. If one also allows for multiple edges to exist between a pair of vertices, leading to a so-called *multigraph* where E denotes the multiset of edges, the directed version is called a *multi-digraph* instead of a simple digraph. A simple graph is furthermore distinguishing itself from other graphs by not allowing for self-loops to exist¹⁷.

In the case where each instance from the set of edges obtains a numerical value we call the graph a *weighted graph* with edge weight w_{ij} for a vertex pair $\{v_i, v_j\}$.

Furthermore graphs can be either connected, where there exists a path between any two vertices in the graph, or disconnected in which case unreachable vertices exist. In the case of a directed graph further distinctions can be made concerning weak or strong connectivity, while in the former case one only requires to obtain a connected graph when replacing all directed edges with undirected ones and in the latter one requires to find a directed path between every pair of two vertices.

The so-called *adjacency matrix* presents the most popular way to algebraically represent the structure of a graph, i.e. denoting if vertices within a graph are adjacent or not. While for undirected graphs the adjacency matrix is symmetric this is not true for di-graphs. An adjacency matrix with elements \mathcal{A}_{ij} is a square $|V| \times |V|$ matrix where in the case of a weighted graph the adjacency matrix elements read

$$\mathcal{A}_{ij} = \begin{cases} w_{ij} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}$$

In the case of multigraphs the entries in the adjacency matrix in general denote the number of multi-edges between any pair of vertices whereas in a weighted multi-graph usually the sum over the multiset of each pair is taken (see e.g. [235, 353]). For completion we note that another concept is given by the so-called Laplacian matrix, which relates the so-called diagonal degree matrix and the adjacency matrix by $\mathcal{L} = \mathcal{D} - \mathcal{A}$ and is often used in determining the spanning trees and hence the partitioning of a graph.

¹⁵An extensive account can be found e.g. in [41]

¹⁶We will see in section VII.1 that we are indeed dealing with bipartite graphs excluding odd-length cycles.

¹⁷Also bipartite graphs by definition do not exhibit self-loops.

Topological properties

The topology of a graph is an essential feature in understanding the respective system described by the graph and exhibits unique characteristics revealing insights about the underlying graph structure. Of important note in this regard are graph *isomorphisms*, which form edge-preserving bijections between two graphs G and G' , hence preserving the underlying network structure. This is not only important to keep in mind when checking if two graphs are structurally equivalent but also when considering different layout structures such as hierarchical, circular or so-called force-directed depictions (see e.g. [279]) in order to reveal topological features such as vertex clusters intuitively.

Among the most important properties of a graph is the *degree* of each vertex v , i.e. $k(v)$, which denotes the number of incident edges at a certain vertex. This then defines a *degree distribution* over the full set of vertices V . In the case of a directed graph one further distinguishes between the *in-degree* $k^-(v)$ and the *out-degree* $k^+(v)$ of a vertex being given by the number of incoming and outgoing edges respectively. If $k^+(v_i) = 0$ then the vertex is said to be a sink while for $k^-(v_i) = 0$ the vertex is called a source. Furthermore the sum of the in-degrees of a directed graph is always equal to the sum of out-degrees, i.e. $\sum_i k^+(v_i) = \sum_i k^-(v_i)$. For a weighted graph the degree of vertex i is generally given by the sum over all incident weights $k(v_i) = \sum_j w_{ij}$.

For some graphs the underlying degree-distribution $P(k)$ for the number of nodes belonging to degree k is found to approximately follow a power-law hence obeying

$$P(k) \propto k^{-\gamma}$$

with γ being the degree exponent. In cases where no such heavy-tail distribution is applicable a Poissonian distribution is employed.

In the case of power-law distributions the graph contains a low number of nodes with a degree significantly exceeding that of all other nodes hence exhibiting high connectivity. These nodes are coined *hubs*. In random networks hubs do not exist. They generally serve as a connection for low-degree nodes in the network and ensure small path lengths, which is why they serve as important features for distributing the information flow within the network. Hence such networks are equivalently called small-world or even ultra-small-world networks [22] although they are most commonly referred to as *scale-free* networks. Additionally hubs are crucial for attack robustness within the network. While elimination of low degree nodes does not have significant effects on the topology of the network targeted removal of several hubs leads to the disintegration of the network and its topological features.

Depending on the degree exponent one usually distinguishes between different network regimes (see [22]):

$$\begin{array}{ll} \gamma < 2 & \text{anomalous regime} \\ 2 < \gamma < 3 & \text{scale-free regime} \\ 3 < \gamma & \text{random network regime} \end{array}$$

while in the random network regime, scale-free networks are practically indistinguishable from random networks and in the anomalous regime scale-free networks can only exist when e.g. exhibiting multi-edges.

Scale-freeness has gained great importance in the context of random networks by introducing mechanisms of generating the scale-free property. For general random graphs the most important models for network generation are given by the *Erdős-Rényi* (see e.g. [4, 236]) and the *Watts-Strogatz model* [332] where in the most basic version a given set of vertices V is iteratively connected by adding more and more edges between the vertices at random following a certain probability distribution. Yet these graphs do not exhibit any power laws in their vertex degrees but rather follow Poissonian distributions. In contrast to this the *Barabási-Albert model* relies on the concepts of network growth, meaning addition of nodes over time, and preferential attachment, meaning that highly connected nodes have a higher probability of receiving new edges, in order to generate power laws (see e.g. [22]). Both of these features are also frequently observed in real networks [21]. In recent years there emerged quite some debate on the viability of scale-freeness as a real emergent property in many real networks stating that not all networks that are claimed to be scale-free truly are [54, 71, 185].

Nevertheless already weak scale-free graphs exhibit important properties irrespective of their denomination since even weak scale-freeness leads to important characteristics such as the appearance of hubs and a high fault-tolerance w.r.t. node removal leading to highly resilient networks. Depending on the power-law exponent hub removal can lead to sets of disconnected graphs. This property is also studied extensively in the context of so-called *percolation*. The interesting quantity in this context is the critical percolation threshold p_c being a measure of network failure or disintegration corresponding to a phase transition within the respective network (see e.g. [22, 268]). It denotes the occupation probability of network nodes or the fraction of nodes that have to be kept until the phase transition from functionality to non-functionality sets in. It has been shown that for a large amount of scale-free networks the percolation threshold is actually $p_c = 0$ meaning that nearly all nodes would have to be removed and hence never leading to actual fragmentation into smaller disjoint sets of nodes [73]. For a detailed account on scale-free networks including site-percolation see e.g. [22, 210]¹⁸.

In order to assess vertex importance in a graph, e.g. in the case of the above mentioned hubs, one can consider a range of graph *centrality* measures. The most widely known and used are *degree centrality*¹⁹, *closeness centrality*, *betweenness centrality*, *eigenvector centrality* and *Katz centrality* [84, 236].

The degree centrality is just given by the respective degree of a node. The closeness centrality assesses the distance of all shortest paths between a specific node and all other nodes, hence indicating to what extent the node is centrally located. For the betweenness centrality one observes shortest paths between pairs of nodes in the underlying network and assesses how often a certain node mediates all of these shortest paths. The eigenvector centrality is an extension of the degree centrality also taking the importance of links to high-scoring nodes into account²⁰. The Katz centrality is a

¹⁸We note that whereas percolation generally describes cluster growth via adding nodes the robustness of networks is rather described by inverse percolation via node removal where the fraction of removed nodes is given by $f = 1 - p$ leading to the above interpretation of the percolation threshold.

¹⁹Again in the case of di-graphs one has to consider *in-degree centrality* and *out-degree centrality* separately.

²⁰In the case of digraphs the eigenvector centrality can be interpreted as an extension to the in-degree

variation of the eigenvector centrality also giving importance to highly linked nodes, which might not be linked to high-scoring ones²¹. As one can see these different centrality measures all have different interpretations and also yield different results for a respective vertex ranking hence they have to be handled with care w.r.t. their interpretability. We also note that in large networks especially the leading ranked nodes w.r.t. a centrality measure exhibit significant relevance while for low ranked nodes this may not necessarily be the case and they might in fact be underestimated in some cases (see e.g. [203, 290]). This emphasizes even more the need for comparative analysis w.r.t. multiple centrality scores.

While some structural features such as strongly connected components can be readily highlighted by centrality measures one is often interested in unravelling topological substructures such as relatively higher connected subsets of nodes compared to their surroundings. This leads to the emergence of subclusters exhibiting strong regulation between their own respective vertices and weak connections between the clusters themselves. Several methods have emerged over the years in order to find clusters within networks (for an extensive account see e.g. [116, 136, 166]) such as *hierarchical clustering*, *spectral clustering*²² or even employing dynamic methods such as *random walks* in order to assess the information flow within parts of the network. Further distinctions are usually made in discovering disjoint or overlapping communities depending on the need for strictness or fuzziness of the respective identification. Additional complexity emerges in directed compared to undirected networks, which is due to the asymmetry of the underlying adjacency matrix in the case of digraphs [116]. In recent years community detection methods also show rising importance in the context of GRNs (e.g. [61, 343]).

Further insight into real-world networks can be provided by so-called *multidimensional* and more specifically by *multilayer* networks (see e.g. [37, 82, 186]). The terminology of these networks has been found to be quite ambiguous yet we use the most widely accepted definitions. In such a case a weighted graph is given by a quadruplet of sets $G = (V, E, D, W)$ where D indicates the set of network layers and W the set of weighted edges between the layer dimensions. The corresponding one-dimensional multilayer adjacency matrices are hence transformed to an adjacency tensor of size $(|V| \times |D|) \times (|V| \times |D|)$. In general arbitrary edges can exist between the nodes of different multilayers yet a special case is given by so-called multiplex networks where nodes from one dimension cannot influence nodes from another dimension. Even though sometimes multiplex networks are visualized with edges between different dimensions they are only used to track the existence of each node. This rather leads to a adjacency tensor of size $(|V| \times |V|) \times D$ with matrix elements \mathcal{A}_{ij}^d . A potential shortcoming of the use of multiplex instead of general weighted multilayer networks can yet be that differential changes cannot be readily included between different dimensions. In order to account for this one can furthermore investigate differential

centrality.

²¹An extension to this going even further is given by the PageRank centrality, which will be discussed in detail in section VII.3

²²Spectral clustering depends on the eigenvalue spectrum of the underlying network, hence the name. A popular implementation is e.g. given by the Perron cluster analysis PCCA+ algorithm [87].

networks between pairwise multilayer dimensions in order to unravel potential topological changes [126, 212, 249]. An extensive account of problems and properties of multilayer networks are given in [37, 186] and in [27, 28, 34, 63, 228] for multiplex networks.

II.3 Machine Learning

Machine Learning has emerged in the last decades as a set of viable computational and statistical methods in order to make data-driven predictions from sets of test or learning samples on sets of test samples. In general one can distinguish between *supervised* and *unsupervised learning* methods depending on if one wants to find already known or new structures in the underlying data given by so-called *instances* with a set of *feature variables*. While the former method needs some sort of annotation, usually denoted by a quantity called the *target* or *response variable* consisting of a set of classes, the latter rather infers the response variable on its own. While supervised learning methods infer the relationship between input and output in the case of unsupervised learning one obtains a concise description of the underlying data, i.e. certain patterns, with related probability distributions. This is obviously important when no structure is known a priori. Among the most common problems in machine learning are classification, regression, clustering and dimensionality reduction. A general introduction into a wide variety of machine learning concepts is given in [23, 36, 139, 190, 231].

In the following we are going to elaborate on two of these concepts²³ that will be central to some of the thesis's methods and results.

II.3.1 Decision Trees and Random Forests

Generally decision trees are still among the most popular methods in supervised classification (discrete target variable) and regression (continuous target variable) problems often being even preferred to comparable methods like support vector machines (SVM), logistic or Lasso regression or neural networks depending on the application²⁴ (see e.g. [139]). One of the main advantages of decision trees clearly is their white-box interpretability when it comes to prediction as well as feature extraction. Apart from that classification can be established very fast, decision trees are fairly robust with respect to outliers and the methods are non-parametric, hence minimizing the a-priori model assumptions, and as well circumventing the extraction of unimodal training features (see e.g. [165]).

In general binary rules are applied to obtain a certain target value. This is achieved via performing binary splits at non-leaf nodes within the tree corresponding to a certain feature variable with the resulting branches representing the decision results and the leaves representing the respective target values. The input data for learning a decision tree is generally of the form of a feature matrix with entries \mathcal{M}_{ij} for i instances

²³one is supervised the other unsupervised

²⁴According to the *no-free-lunch theorem* there is no preferred optimization method when averaged over all class of problems, although there is some debate on that in the machine learning community [340, 347]. Still this means that certain problems call for specific methods.

and j features and the dependent target variable vector with entries Y_i representing the associated class or target of each instance i .

A simple decision tree now recursively partitions the full data set into smaller subsets of instances by a top-down method in order to end up with members of the same class in the same tree nodes.

The question on which feature variable results in the most informative partitioning can be answered via the choice of an appropriate *splitting function*. From information theory there exist several choices the most popular of which are *information gain*²⁵, depending on the Shannon entropy²⁶, and *Gini impurity* for classification trees. The information gain is given by the decrease in entropy after splitting (see e.g. [216])

$$IG(X|Y) = H(X) - H(X|Y) \quad (\text{II.1})$$

$$= - \sum_{i=1}^{\mathcal{C}_X} p(x_i) \log p(x_i) + \sum_{j=1}^{\mathcal{C}_Y} p(y_j) \sum_{i=1}^{\mathcal{C}_Y} p(x_i|y_i) \log p_i(x_i|y_i) \quad (\text{II.2})$$

where X denotes the parent node (before the split), Y the child nodes (after the split) and $H(X|Y)$ the conditional entropy with \mathcal{C}_X being the classes in the parent node and \mathcal{C}_Y being the classes in the child node.

On the other hand the Gini impurity at a certain node k reads

$$\mathcal{I}_{\text{Gini},k} = \sum_{i=1}^{\mathcal{C}_k} (p(x_i)(1 - p(x_i)))_k \quad (\text{II.3})$$

and hence tells the mislabelling probability of a randomly chosen and randomly labelled instance from the underlying sample.

The splitting functions hence act as an importance measure for all features in distinguishing best between different classes. A relevant feature in this regard ends up in the root of the tree. In the case of the Gini impurity a small value is hence associated with all instances in the subsets being of the same class after splitting at a certain node and a large value when all classes occur with equal probability over all instances in the subsets. A visualization of this circumstance can be found in Fig. II.1. Splitting is usually performed until a certain low impurity threshold is met. Otherwise for a large number of features splitting continues until the impurity is zero.

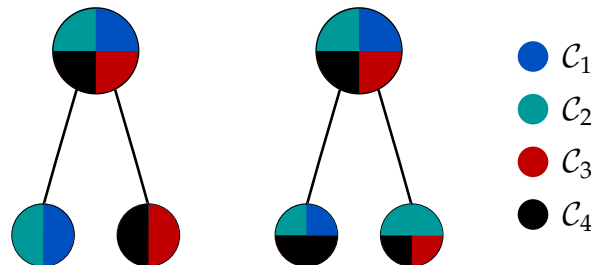


Figure II.1: Impurity visualization of two possible splittings w.r.t. to different features. Each leaf node consist of a subset of different class distributions w.r.t. to the underlying classes \mathcal{C}_i . The left splitting produces a less mixed leaf with a lower impurity compared to the right splitting.

²⁵alternatively *Kullback-Leibler divergence* or mutual information

²⁶For a thermodynamically motivated introduction of entropy see e.g. [191].

The main competitive computational models in decision tree learning are given by e.g. CART and C4.5 (see e.g. [139]).

Among the main problems of simple decision trees is their tendency to overfit the underlying data with over-complex trees, which can be due to a too large set of features, i.e. exhibiting small bias but large variance²⁷, which is in many cases solved by pruning, i.e. removing features with little predictive power. For the above reasons the obtained classifiers tend to be quite noisy. To this end improvements on ordinary decision trees have been developed, which aim at reducing this shortcoming extensively, leading to ensemble methods that are realized either via so-called boosting methods (e.g. gradient or adaptive) [117, 139] or via bagging or bootstrap aggregation [50].

The most popular example surely is given by *Random Forests* [7, 51] employing bagging to generate an ensemble of de-correlated decision trees by randomly selecting a subset from the original data with replacement, i.e. bootstrapping the data and in the end averaging over the ensemble or taking the majority vote as in the case of classification. For each subset an individual decision tree is learned. The averaging procedure reduces the noisiness of the individual trees drastically yet at the same time keeping the small bias of the individual trees [139]. In addition to tree bagging random forests employ the method of feature bagging, i.e. selecting a random feature subset in each tree, in order to avoid correlation of individual trees. This prevents important features from reoccurring in every single individual tree [156]. By avoiding overfitting random forest methods are able to deal particularly well with so-called “small n , large p ” problems, i.e. having a small number of samples and a large number of features (or predictors) without the need for feature pruning. In addition random forest methods have been shown to outperform not only decision trees but also methods like SVM or even neural networks in various classification studies (see e.g. [40, 113, 165]).

Improving even further on the bagging methods applied in random forest one can additionally randomize the feature split values for each randomly drawn feature instead of selecting a feature based on an optimal split value. This leads to a further decrease in variance at the cost of a slight bias increase w.r.t. tree splits themselves. The result is called *Extremely Randomized Trees (ERT)* implemented in the ExtraTrees algorithm [124]. In contrast to other ensemble methods it also makes use of the full learning sample for tree growth. Especially for noisy data the ERT method has been shown to be advantageous and outperforms other tree-based ensemble methods as e.g. in the case of bias-variance trade-off (see e.g. [124]). We will make use of this method later in chapter VI.

II.3.2 Hidden Markov Models

A *Hidden Markov Model (HMM)* is an unsupervised learning method and a special case of a so-called dynamic Bayesian network being a probabilistic graphical model. It is especially suited for sequential data not being independent and identically dis-

²⁷The bias-variance trade-off aims at the incommensurability of vanishing bias in combination with vanishing variance (see e.g. [139, 165]).

tributed (*i.i.d.*). History independence of sequential information leads to the assumption of a model exhibiting the Markov property, i.e. that a future probability distribution only depends on the currently observed state (see. e.g. [23, 36]). This leads to a discrete first order Markov chain, which is represented as a directed tree graph. If one furthermore introduces discrete *latent* or *hidden variables* to the *Markov chain* one obtains a so-called state-space model, which forms the underlying structure of the HMM²⁸. A depiction of an exemplary Markov chain for an HMM can be seen in Fig.II.2.

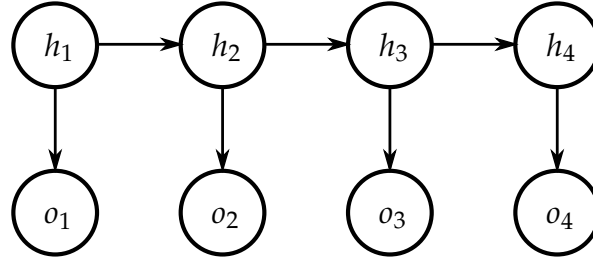


Figure II.2: Exemplary scheme of a Markov chain with four sequence steps underlying an HMM with observations o_i and hidden variables h_i .

The observables o_i of the system are then related to the latent variables h_i via so-called *emission probabilities*²⁹ given by $p(o_i|h_i)$ whereas the sequence itself is defined by *transition probabilities* $p(h_i|h_{i-1})$ defining a joint probability given by

$$p(h, o) = p(o_1|h_1)p(h_1) \prod_{i=2}^N p(o_i|h_i)p(h_i|h_{i-1}) \quad (\text{II.4})$$

with the distributions for $i = 1$ defining the *initial distributions* and a total of N steps (see e.g. [23]). Hence the full parametrization is given by this joint distribution through the set of all unique emission and transition probabilities. Further assumptions for HMMs apart from the Markov property are stationarity and observation independence. Stationarity assumes that transition probabilities are the same for every timepoint whereas observation independence states an observable at a certain timepoint is statistically independent from an output at another timepoint.

Learning a candidate set of locally optimal parameters³⁰ for the HMM is generally achieved by an iterative expectation maximization (EM) procedure via the so-called *Baum-Welch* (forward-backward) algorithm. In order to perform the learning step the model structure has to be already known in advance i.e. the number of hidden states h and observations o . Hence parameter estimation depends on a given number of observation sequences. The Baum-Welch algorithm basically aims at maximizing the log-likelihood of the joint distribution in II.4 w.r.t. the parameter set $\theta = \{\Theta, \Sigma, \pi\}$ where Θ denotes the transition probability matrix, π the initial state probabilities and

²⁸The following derivations will follow along the lines of introductions such as [23, 36].

²⁹sometimes also called observation probabilities

³⁰Local optimality has to be assumed since we face a non-convex optimization problem depending on the initial parameters.

Σ the emission probability matrix. The log-likelihood for the joint distribution reads

$$\mathcal{L} = \ln \sum_o p(h, o | \theta) \quad (\text{II.5})$$

$$= \ln p(h_1, \pi) + \sum_{i=1}^N \ln \sum_o p(o_i | h_i, \Sigma) + \sum_{i=2}^N \ln p(h_i | h_{i-1}, \Theta). \quad (\text{II.6})$$

The optimization basically amounts to taking

$$\frac{\partial \ln p(h, o | \theta)}{\partial \Theta} = 0 \rightarrow \Theta_{jk} = \frac{\sum_{i=2}^N \langle h_{i,j} h_{i-1,k} \rangle}{\sum_{i=2}^N \langle h_{i-1,k} \rangle} \quad (\text{II.7})$$

$$\frac{\partial \ln p(h, o | \theta)}{\partial \pi} = 0 \rightarrow \pi_j = \langle h_{1,j} \rangle \quad (\text{II.8})$$

$$\frac{\partial \ln p(h, o | \theta)}{\partial \Sigma} = 0 \rightarrow \Sigma_{jk} = \frac{\sum_{i=1}^N o_{i,j} \langle h_{i,k} \rangle}{\sum_{i=1}^N \langle h_{i,k} \rangle} \quad (\text{II.9})$$

where details can be found e.g. in [23, 36, 125].

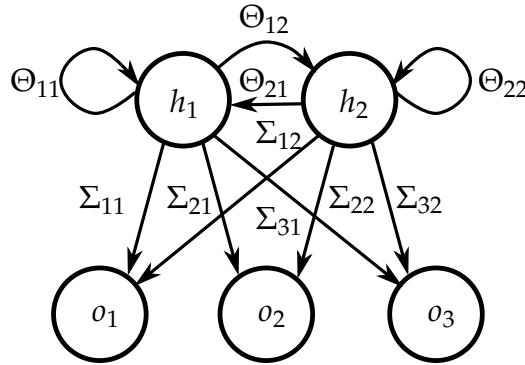


Figure II.3: Schematic depiction of probabilistic parameters with emission probabilities Σ_{jk} and transition probabilities Θ_{jk} as defined before. The indices now denote the number of the respective observable or hidden state respectively.

The resulting parameters from this so-called maximization step now depend on the expectation values of the hidden states $\langle h \rangle$, which are then computed by the actual forward-backward procedure of which a full derivation is given in [23, 36, 125]. The procedure is then performed with updated hidden state expectation values until convergence of the log-likelihood. In an HMM the full set of parameters is hence inferred via the learning step given a number of observations and hidden states and estimating an appropriate initial distribution from the underlying test data. An exemplary schematic depiction of the parameter set is given in Fig.II.3 for two hidden states and three observables.

Although there exists a wide range of extensions to the standard discrete first-order HMM such as tree-structured HMMs, Bayesian HMMs and continuous versions as in the case of Gaussian distributions with so-called Kalman filters we refer for these cases to the standard literature as e.g. [23, 36].

HMMs have a wide range of applications including but not limited to speech and hand-writing recognition, object tracking but also in sequential biological data as in

DNA sequence analysis and epigenetic state detection, which will be of utmost interest for the work conducted in the following (see e.g. [23, 94, 106, 338]). In this case observables are generally given by e.g. ChIP-Seq data on epigenetic marks such as histone modifications while the hidden states are combinations of such marks, i.e. chromatin states, whose sequence on the DNA is facilitated in obtaining the parametrization of the underlying HMM learning process [104][103].

In recent years chromatin state detection advanced rapidly with the use of machine learning methods among which different realizations using variations of HMMs are frequently employed [338]. Among the most popular ones for static data sets is ChromHMM [104][103], which is especially tailored to ChIP-Seq data input concatenating a variety of input cell conditions. This implementation was also employed in the ENCODE and Roadmap Epigenomics projects for de novo chromatin state detection [158, 194]. Other implementations include EpiCSeq, modelling raw read counts as opposed to ChromHMM [218], histoneHMM making use of differential peak changes [143], hierarchical HMMs in diHMM [220], tree HMMs in TreeHMM [35] or HMM-like dynamic Bayesian models as implemented in Segway [157]. We will especially make use of the most widely known implementation, i.e. ChromHMM, despite its limitation of merely modelling binary observables via the presence or absence of certain chromatin marks and hence losing quantitative i.e. differential peak information in the process³¹. We will face this shortcoming later on in chapter V with our own implementation of a parametrized correlation algorithm. The reason for the usage of ChromHMM is also general comparability to results e.g. obtained by the large genome consortia such as ENCODE.

³¹That can be e.g. the ratio between H3K4me1 and H3K4me3, which can be important in distinguishing promoter from enhancer states.

CHAPTER III

Data sets: description and analysis

In this chapter we are going to describe and analyse the underlying experimental data sets which will be used for the subsequent inference of the properties of the epigenetic landscape in Th1 and Th2 cells as well as for the inference and analysis of the resulting network(s) which is connected to unique network topological features. Furthermore we will show in detail how the cumbersome data analysis is performed since these steps are quite tricky and are among the largest pitfalls when it comes to understanding how subsequent results are obtained. Also a detailed analysis of the raw and pre-processed data sets contributes significantly to the understanding of the data and later on will justify why certain analytical measures were taken and how the introduction of a large body of computational, statistical and modelling frameworks can be motivated.

III.1 Underlying experimental data sets

The experimental data¹ we are going to investigate *in silico* for the Th1/Th2 system consists of a mixture of different cytokine culture conditions w.r.t. terminal Th1 and Th2 differentiation as well as of different levels of Tbet dose, which is of special interest being the canonical Th1 master regulator. Hence we do not only investigate the genotypic differences between Th1 and Th2 cells themselves but also the effect of perturbation of one of the main drivers of the system². The experimental procedure itself is depicted in Fig.III.1.

Naïve LCMV-specific (lymphocytic choriomeningitis virus) CD4+ cells were taken from LCMV-TCR^{tg}Thy1.1⁺ donor mice being subject to perturbations in Tbet dose w.r.t. allele occurrence, i.e. Tbx21^{+/+}, Tbx21^{+/-} and Tbx21^{-/-}, leading to a gradient in expression of Tbx21 influenced gene targets, which will be shown in due course. Three days before viral infection with LCMV these cells were retransferred into uninfected wild-type recipient mice. LCMV was used as an immune response trigger in order to increase the number of virus-specific CD4+ T cells. Ten days post

¹Cell cultures have been performed by Ahmed Hegazy in the group of Prof. Max Löhning at Charité and German Rheumatism Research Center (DRFZ) in Berlin, Germany, whereas ChIP-Seq and RNA-Seq experiments have been performed by Qin Zhang from the group of Prof. Thomas Höfer at the German Cancer Research Center (DKFZ) in Heidelberg, Germany.

²We note at this point that Gata3 specificity was not investigated since this posed substantial issues when it came to extracting a significant amount of cell material under the conditions of interest.

infection Thy1.1⁺ donor Th1 cells were isolated in their effector phase and after sorting kept under in vitro neutral and Th2 polarizing culture conditions respectively. The neutral conditions are characterized by the addition of α -IL4, α -IL12 and α -Ifn γ , hence not contributing to any differentiation program, where α is short for “anti”, hence blocking the respective cytokine. The Th2 polarizing culture conditions are achieved by the addition of IL-4, α -IL12 and α -Ifn γ , being Th2-specific. This yields in the end a total of eight different cell conditions namely

Th1 neutral	Th2	ex vivo
Tbx21 ^{+/+} Th1	Tbx21 ^{+/+} Th1	Naïve
Tbx21 ^{+/-} Th1	Tbx21 ^{+/-} Th1	
Tbx21 ^{-/-} Th1	Tbx21 ^{-/-} Th1	
	Naïve	

Table III.1: Experimental treatment conditions for the considered data sets.

The cells which are kept under Th2 polarization conditions exhibit properties of mixed Th1/2 phenotypes with hybrid gene expression as well as hybrid epigenetic landscape. This occurrence of plasticity can be shown to be stable and long-lived in memory phase even after one month after retransferring them into uninfected wild-type mice. The same also holds true for the unique phenotypic properties w.r.t. perturbed Tbet dose under neutral as well as under Th2 polarization conditions. Also the lower the level of Tbet the higher the plasticity potential.

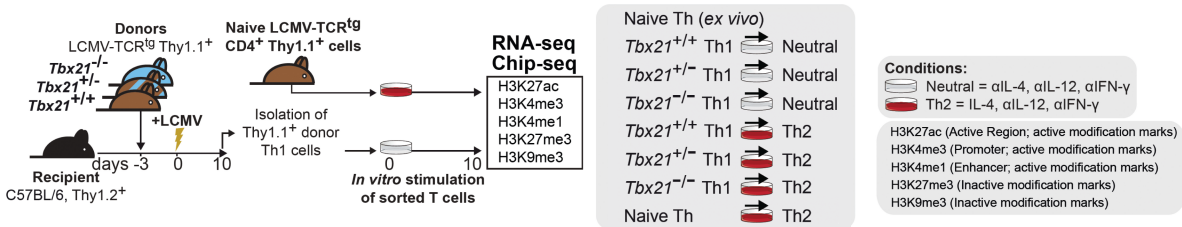


Figure III.1: Mouse model used for underlying Th1 and Th2 cell conditions with different Tbet dose as well as for different histone modifications. All of this is done under Th1-neutral and Th2 culture conditions as defined in the upper right box (adapted with kind permission from Hegazy *et al.*, *in prep.*).

In the following the differentiated Th1 cells under neutral conditions will be abbreviated as Th1 cells and the Th1 cells under Th2 culture conditions will be abbreviated as Th1/2 hybrid cells since they effectively represent Th1 \rightarrow Th2 reprogramming properties as will become evident in due course. An exception are obviously the naïve cells in Th2 conditions, which represent the Th2 control as well as the ex vivo naïve cells, which will keep these labels respectively.

The cell cultures were then sequenced, i.e. RNA-Seq as well as ChIP-Seq is performed for all eight conditions with two biological replicates each³. In the case of ChIP-Seq antibodies for different histone modifications were used for each of the above cell conditions. In order to be able to investigate a broad range of epigenetic profiles five important histone modifications were investigated, which represent marks

³A general introduction on ChIP and RNA sequencing methodology can be e.g. found in [118, 252, 328].

associated with certain epigenetic functionality (see section II.1.2) as shown in table III.2.

histone modification	associated with
H3K4me1	enhancer
H3K4me3	promoter
H3K27ac	active region
H3K27me3	repression
H3K9me3	Polycomb silencing

Table III.2: List of histone modifications used in the experiments.

In the following we are going to assess the data quality output from the experiments and establish a pre-processing pipeline before evaluating preliminary properties of the different cell conditions regarding gene expression and the respective epigenetic landscape.

III.2 Analysis of experimental data

In the field of bioinformatic analysis of high-throughput sequencing data there exists a wide range of different computational methods, the naïve usage of which can influence the outcome of the processed data dramatically. Additionally workflows for ChIP-Seq and RNA-Seq analyses differ considerably (see e.g. [75, 76, 232]). In the following we are going to justify the usage of different methods in the subsequent computational pipeline and discuss the result on basis of the aforementioned data sets and its implication for further analysis and modelling.

III.2.1 Histone modifications

Data quality control

For the ChIP-Seq data sets we obtain a total of two biological replicates with two technical replicates each. For each of those the data quality can be checked with respect to read length, number of reads or purity filtering and most importantly w.r.t. per base sequence quality based on the so-called Illumina Phred+33 quality score⁴. Single read sequencing was performed with a read length of 51 bp. On average around 90% of all reads are purity filtered. An exemplary depiction of the per base sequence quality and sequence length distribution of a replicate is shown in Fig.B.1. We find for all replicates and conditions that the base quality score is lying within the highest quality region. Hence we do not have to trim the read length, which is usually done to prevent the occurrence of subsequent alignment errors.

Alignment

Raw ChIP-sequencing data sets like the ones for histone modifications consist of a text-format (ASCII) based method to store nucleotide sequences for reads including

⁴see https://www.illumina.com/documents/products/whitepapers/whitepaper_datacompression.pdf for details

quality scores. The read sequences have to be aligned to the reference genome first in order to obtain spatial meaning. There exists a multitude of different sequence aligners incorporating a variety of implementation methods for alignment for different purposes. Additionally there are different reference genome builds available from which to choose. In our case we settled for the most recent mouse reference genome build mm10⁵. Furthermore we are interested in an alignment method which is able to deal particularly well with short read sequences as well as providing a fast and memory-efficient algorithm for which the *Burrows-Wheeler transform* is considered to be a viable candidate (see e.g. [58]). Hence we employed the use of the popular Bowtie algorithm [199] for the alignment of our ChIP-Seq data sets. For the resulting aligned reads we perform additional data quality assessment via correlation of the respective conditions and histone modifications. To this end the genome is partitioned into 5 kb windows. Correlations are now performed genome-wide between all samples on top of which we perform hierarchical clustering. The result is depicted in Fig.B.3 in the appendix. We find that all histone modifications are found in distinct clusters and the cell conditions are found in neighbouring bins of the hierarchical clustering as well, which confirms the high quality of the data.

Peak search

Peak calling is an important step in determining statistically significant enrichment of reads or tags in ChIP-Seq data. Especially the question of calculating the genome-wide signal-to-noise ratio is of utmost importance in order to subtract a statistically insignificant background of ChIP-Seq reads. From general experience peak calling is a crucial step in data-preprocessing since the choice of a certain peak calling algorithm as well as its respective statistical assumptions and input parameters can have large impact on the subsequent downstream analysis (see e.g. [294, 305]).

Methods For our purposes we have to consider peak calling algorithms which can deal equally well with narrow and broad peak structures at the same time. Especially in the case of H3K27me3 one finds very broad profiles sometimes even ranging over distances of several genes while for H3K4me1 and H3K27ac the peak profiles tend to be narrower, in some cases even encompassing only several nucleosomes⁶. An exemplary depiction of this is shown in Fig.III.2 in the appendix around the *Ifn γ* gene. The important point to consider here is that peaks are not truncated at some low threshold so in order not to end up with a huge number of small fragmented peaks one has to take small gaps with low or even no signal into account over which a peak detection algorithm can integrate. Hence depending on the data that is considered one has to find an appropriate combination between sliding window size for detecting read enrichment and gap size. Also we have to note that diffuse extended signals in the case of histone modifications tend to lack saturation, unlike transcription factor binding sites. Hence the determination of a viable background bias model is of utmost importance.

⁵build version GRCm38.p4 was used – see http://sep2015.archive.ensembl.org/Mus_musculus/Info/Annotation

⁶Although in some cases, as we will see, the peak domains can be broad as well compared e.g. to TF binding ChIP-Seq data.

Broad, diffuse peaks which are found especially in the case of repressive histone modifications marks or as well in the case of extended “super-enhancer” regions are especially well dealt with in implementations like SICER [363]. The crucial difference to ordinary sharp, narrow peak searching methods is that the respective sliding window is not fixed. Rather it is only used for scanning the genome and identifying clusters of so-called read islands first. Furthermore small gaps between islands are allowed such that broad structures are not fragmented too heavily. We preferred SICER to other popular peak searching algorithms like MACS since at the time of evaluation SICER was the most advanced algorithm in being able to deal with narrow and broad peaks reliably at the same time (see e.g. [350, 363]) and avoiding short-comings like the default fragmentation of peaks into smaller subpeaks.

The eligibility of reads within a local scanning window w is assessed by a score s of reads l . The lower limit for determining if a certain window is eligible or not depends on a Poissonian random read background model. The limit on the lower bound of read counts l_0 is set by a p -value p_0 according to

$$\sum_{l=l_0}^{\infty} P(l, \lambda) \leq p_0$$

while the score s of a local window is determined by

$$s(l) = -\ln P(l, \lambda)$$

with the Poisson distribution $P(l, \lambda)$ where $\lambda = w \cdot N/L$ and N denotes the number of library reads and L the genome length. Including a certain gap length g over which integration between different local windows is allowed, one can finally find a scoring function for candidate peak domain islands by aggregating the individual scores of neighbouring local windows maximally separated by g . After performing the random background island evaluation for all samples each sample is then compared with the respectively equally evaluated control sample to further reduce signal bias. For this last evaluation one has to specify a certain **FDR** as a statistical cutoff for successful peak island calls. All of this eventually leads to the search for significantly enriched peak domains with variable length opposed to just using a fixed-length window. Further details on the computation can be found in [363]⁷. The peak calling parameters are listed in appendix **D** and motivated in the following.

After having performed read alignment we retain so-called strand asymmetry since read fragments are sequenced for both strands. For this reason strand-specific peaks experience shifts away from the true peak mean starting from the 5' end. To account for this issue one needs to specify the original ChIP-seq average *sonication fragment size* d from which the tag shift being half the fragment size is determined. In our experimental data set we had $d = 150$ bp, hence the shift being 75 bp. For our customized pipeline we were iterating the peak calls for variable gap lengths as well as false discovery rates in terms of the background model. The window size was chosen to be 200 bp being roughly more than the size of a single nucleosome. We did not allow for multiple read mappings and hence set the threshold of redundant

⁷We note that the potentially large width of the islands leads to further stabilization w.r.t. the background model additionally reducing possible sampling variabilities at the nucleosome level.

reads to one. Varying gap sizes were important in accounting for broad peaks away from gene bodies as well as at the same time for smaller peak fragments in promoter regions. From the gathered results we assumed that FDRs of 0.01 were a viable trade-off choice in not losing too much peak information over the whole genome.

In addition to fixing the above constraints on statistical peak domain evaluation we merge technical replicates beforehand in order to increase the data load for more reliable statistics. After the estimation of the appropriate background model based on the control library input we obtain a list of significantly estimated islands based on the aforementioned FDR. For all of these significant islands we respectively obtain the integrated number of redundancy-removed reads.

For further downstream evaluation of the histone modification peaks (e.g. in section IV.1) we are considering gap lengths of $g = 600$ bp for H3K4me1, H3K4me3, H3K27ac and H3K27me3 since peak islands are especially broad for these modifications (see e.g. [294, 350, 356] and Fig.III.2) and $g = 200$ bp for H3K4me3 since promoter regions are thought to be more fine-grained. Finally we obtain the background subtracted reads for each unit read bin of the length of the sliding window w and normalize them w.r.t. library size according to

$$r'_i = \frac{r_i}{10^6 \cdot \sum_i r_i}$$

where r_i denotes the unnormalized reads in some unit bin i . The result is given in reads per million (RPM). From this point onwards these results are used for further downstream analyses.

Results As an exemplary result we turn to a notable gene locus in Th1 cells with its respective histone modification peak profiles representing the statistically significant noise-reduced epigenetic landscape in Fig.III.2. *Ifn γ* being one of the most important Th1-specific cytokines [49] with one of the best annotated loci in Th1 cells [19] presents a viable candidate for observing distinct differences in the resulting histone modification peak structure. We show all histone modifications of one biological replicate for several experimental conditions measured. All histone modifications enrichments follow the above mentioned peak calling conditions w.r.t. island determination ($w = 200$, $g = 600$, FDR= 0.01 – except H3K4me3 with $g = 200$).

We can confirm our expectations that for some conditions like Tbet^{+/+}Th1, Tbet^{+/+}Th1/2 or Tbet^{+/-}Th1 the H3K4me1 peak structures are rather broad and pronounced. This does not come as a surprise since in these conditions we can expect high enhancer occurrence. For these conditions the same also holds true to a little lesser extent for H3K27ac, which indicates activity since the acetylation is responsible for opening up the heterochromatin. We also find large overlaps between these two histone modifications, which indicates the occurrence of active enhancers in these regions for the respective conditions. On the other hand we also find broad pronounced domains for H3K27me3 in Th2 cells, which are also already occurring in naïve and Tbet knock-out conditions. This can be readily expected as well since H3K27me3 is associated with gene repression, which is the case for *Ifn γ* to a different extent in those conditions being especially prominent in the opposing cell program of Th1, namely Th2. We also find occurrences of H3K9me3 to be quite rare. We will assess this later statistically in a genome-wide manner in more detail. The

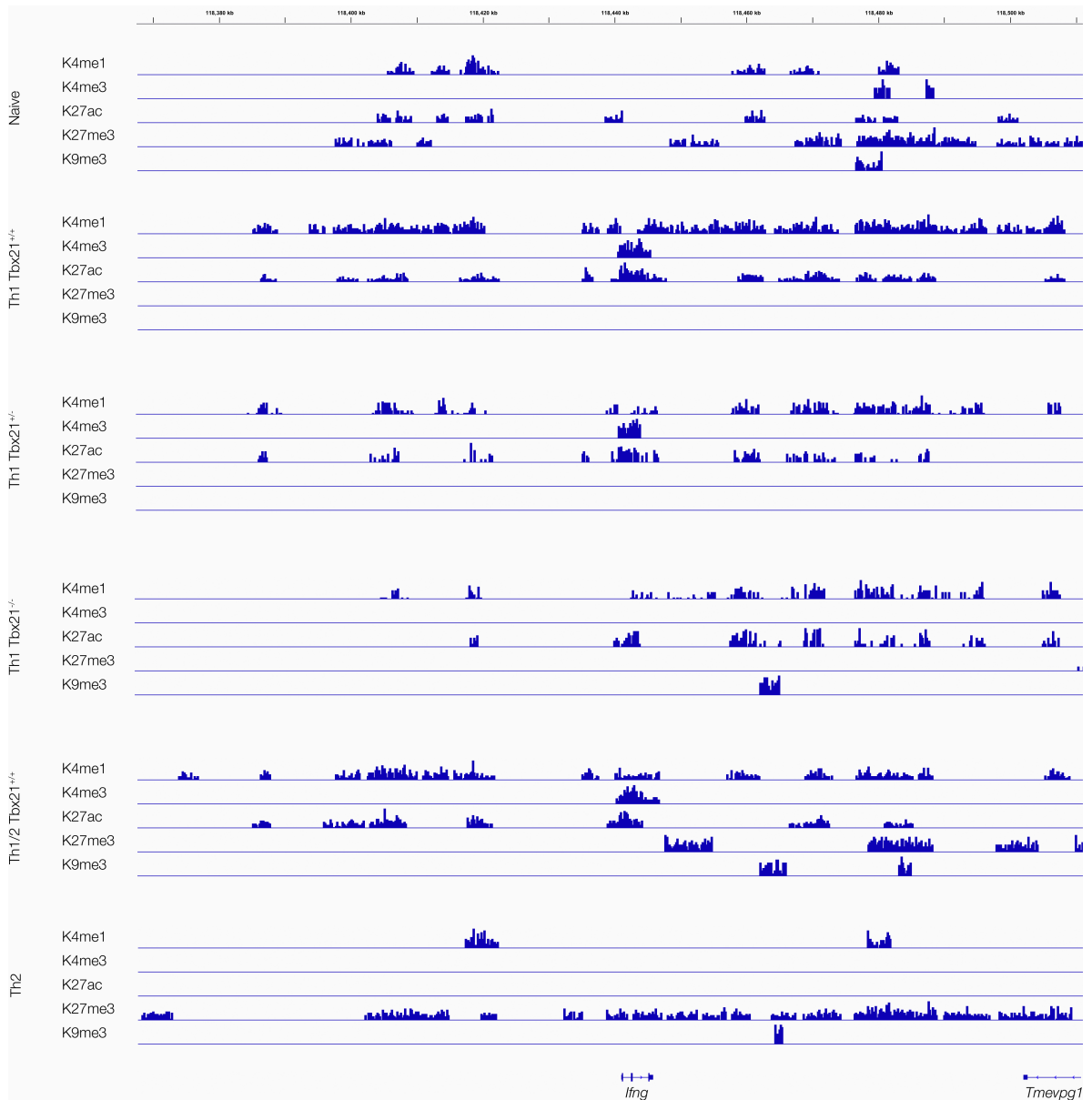


Figure III.2: Epigenetic histone modification peak landscape results around *Ifn γ* for selected experimental conditions and replicates with peak calling parameters as specified in appendix D.

indicator for promoter activity, H3K4me3, shows indeed distinct occurrence around the *Ifn γ* promoter naturally in $Tbet^{+/+}$ Th1 and $Tbet^{+/+}$ Th1/2 conditions but also in $Tbet$ heterozygous cells. Hence we find not only a cytokine dependency but also a $Tbet$ dose specificity w.r.t. this mark. We see that we obtain distinct connected peak structures exhibiting different levels of modification over different experimental conditions with varying peak widths and heights. This becomes specifically apparent for $Tbet$ dose variations. We can furthermore confirm that for certain cell conditions certain histone modification peaks, which are expected not to occur in these conditions indeed vanish and hence former statistically irrelevant residuals in the raw data are removed. We will come back to a more detailed investigation of the *Ifn γ* locus in due course.

III.2.2 Gene expression

Data quality control

In accordance with the previous case of the histone modification data we again investigate the quality reports for all technical replicates for the RNA-Seq data. The paired-end sequenced RNA reads are now longer than in the case of the histone modifications with a length of 101 bp. Again approximately 90% of all reads were purity filtered. Based on the Phred+33 score the per base sequence quality is again exemplified in Fig.B.2 while the per base sequence quality amount for each sample found to be lying in the high quality region ($\text{Phred}+33 \geq 30$) is also always around 90%. Hence trimming of the samples can be neglected.

Alignment

Since we will not focus on de novo alignment of RNA reads and the discovery of novel transcripts we are interested in efficient and reliable short-read alignments to a reference transcriptome. Since in the experimental RNA-Seq protocol paired-end reads were sequenced we will not run into multi-mapping problems (see e.g. [328]) since the alignment of read pairs invokes boundary conditions on how close read sequences can be to each other as well as on the mapping order.

Additionally in the case of RNA-Seq data the alignment protocols have to incorporate splicing events as well, which for years has led to problems in development of efficient and fast mappings to genomic transcriptomes. Furthermore not only splicing events have to be accounted for but also sequence mismatches as well as insertions and deletions. In order to account for different splicing variants during alignment, especially w.r.t. sensitivity in aligning reads over splice junctions we consider the STAR algorithm [91], which aligns read pairs directly to a reference correcting for splicing events. The accuracy of STAR has been reported to exceed most contending algorithms [101] The corresponding alignment parameters can be found in appendix D.

Concluding summary statistics of the performed alignment on our data sets reveal that the average mapped read length for paired-end reads was around 200 bp in contrast to an input length of 202 bp while the percentage of uniquely mapped reads was always around 75% being a reasonable amount for further downstream processing. Furthermore we note that paired-end reads are eventually evaluated as one read, avoiding double counting. Multi-mapping occurrences were quite infrequent being significantly below 10%.

Since the alignment itself only maps reads to genomic positions we are still in need of extracting read counts as well as attributing these counts to specific genomic annotations, such as for example transcripts. First of all we consider the union of exons for a certain gene transcript from the genome build version GRCm38.p4⁸. Ambiguous reads that map to several transcripts are omitted. For efficient annotated read counting we use the Python-based HTseq-algorithm [9]. The specific command options are again given in appendix D. Eventually we obtain a total of 52.734 unique ENSEMBL transcript mappings in form of a raw count matrix, which will be used as an input for further detailed expression analysis.

⁸see http://sep2015.archive.ensembl.org/Mus_musculus/Info/Annotation

Differential expression analysis

Methods Starting from the inferred raw count matrix with elements \mathcal{K}_{ij} with i being the ENSEMBL gene transcript and j being the replicate of a certain cell condition we begin by estimating the dispersion for each gene transcript. Furthermore we are interested in the respective fold changes of those transcripts between samples. What is usually done is just naively applying the null hypothesis that the logarithmic fold change (LFC) of the expression of a certain gene transcript is zero. Neither a p -value ranking nor mere selection by LFC might yield the desired set of either up- or down-regulated genes, since in the former case the LFC might be too low, while in the latter case low count estimates have too high intrinsic noise. Furthermore the fold change for a certain transcript also depends heavily on library size. To resolve these problems we follow the statistical procedures introduced in the DESeq2 R-package⁹ being laid out in detail in [214]. Along these lines the count matrix \mathcal{K} is assumed to follow a negative binomial distribution with a fitted mean μ_{ij} fulfilling

$$\mu_{ij} = s_j q_{ij}$$

where s_j is denoting a scaling factor for sample j and q_{ij} being proportional to the concentration of reads. The size factor s_j is itself determined via taking the median of the ratios of all sample gene counts and their individual geometric mean g_i via

$$s_j = \text{median}_j \left(\frac{\mathcal{K}_{ij}}{g_i} \right).$$

This basically accounts for different sequencing depths for all of the considered samples. The variability between different replicates is accounted for by determining a dispersion factor α_i for each gene transcript i correcting for noise in Wald-tests of LFC estimates between samples. This relates the variance and mean of the count matrix in the following fashion

$$\text{Var}(\mathcal{K}_{ij}) = E[(\mathcal{K}_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \cdot \mu_{ij}^2. \quad (\text{III.1})$$

We observe two things: first, the dispersion factors α_i measure the weight of increasing residual dispersion. Second the variance grows with the mean itself. Furthermore it turns out that the dispersion estimates vary heavily with the sample mean. Performing a fit through these dispersion estimates the “true” dispersion value is obtained. Via a Bayesian method this fitted dispersion value is approached by shrinkage of the measured dispersion values towards the fitted value, which just means that the residuals are minimized by a Bayesian maximum likelihood a posteriori estimator (MAP). In the end too high dispersion estimates are thus lowered while too low estimates are raised at the same time avoiding type I errors in underestimating dispersion when having only two sample replicates as in our case. Further heteroscedasticity occurs in the LFC estimation procedure, since the variability of low count gene LFCs is higher than for high count genes as can be seen straightforwardly. Via a MAP procedure the LFCs are then shrunken with a bias towards zero, s.t. low count genes experience lower variability in their LFCs. This leads also to the fact that low count genes have very low LFCs overall as could be readily expected. As a last step a Wald

⁹<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

test is performed on the MAP LFC estimates and their standard errors yielding in combination with multiple testing (see section A) a final statistical evaluation of the corrected LFCs. In addition gene count outliers are accounted for by Cook's distance and removed from the samples.

Results We exemplify the above LFC analysis with the comparison between Th1 and Th2 cells (Tbet^{+/+}Th1 vs. Th2 control) in Fig.III.3. First of all we visualize the MAP dispersion estimates for all gene transcripts. Additionally we see that for very high mean values of normalized counts over all samples extreme LFCs become less significant and are hence excluded by this method from further downstream analyses. We filter the LFC results w.r.t. adjusted p -values being subject to $p_{adj} < 0.01$ and sort w.r.t. LFC.

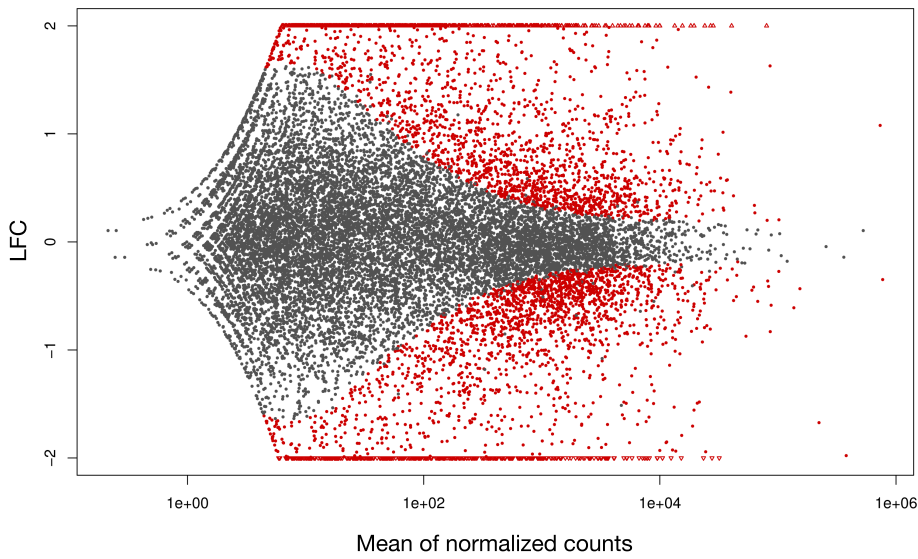


Figure III.3: LFC transcript expression plot of differential regulation between Th2 and Th1 wild-type cells. In red statistically insignificant LFCs are shown that are excluded from further analyses.

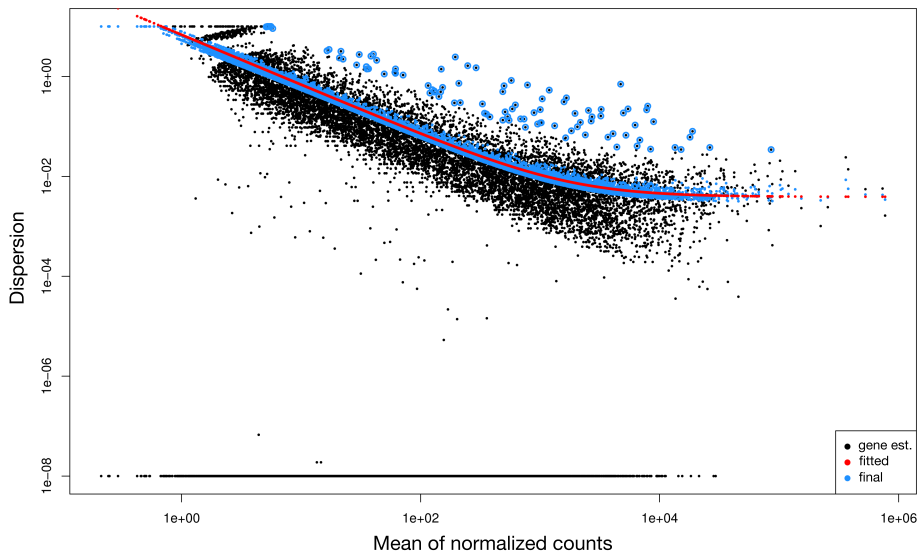


Figure III.4: Dispersion estimate plot for the differential expression between Th2 and Th1 cells with the black dots denoting the estimate for every single gene transcript, the red line denoting the fitted dispersion and the blue dots representing the MAP dispersion estimates. Black dots with blue rings denote outliers and are hence excluded from the analysis.

As expected we find a wide range of upregulated Th1 genes and down-regulated Th2 genes. A full list of the respective Top 50 upregulated Th1 genes w.r.t. Th2 and vice versa can be found in table C.1 and C.2. This will be of further importance in our subsequent analysis as we will find differences in cell-specific behaviour partly on basis of these upregulated genes in the respective cell conditions as determined by these analyses. To this end we compare the list of upregulated Th1 and Th2 genes with published Th1- and Th2-specific genes (see [334]) and extract from that a subset of significant LFCs ending up with 46 Th1 and 50 Th2 gene transcripts in total (see table C.3 for full list) to be used for further downstream analysis.

Absolute expression analysis

Methods Again we follow one of the main statistical procedures as proposed in [214] and [8]. The bottom line of getting absolute gene expression values is basically the question of normalization. More specifically gene expression inferred from sequencing data is not homoscedastic but rather heteroscedastic. This means in detail that variances differ for different gene expression values according to

$$\text{Var}(\xi_i | x_{ij}) = \sigma_i^2, \quad \forall j \in 1, 2, \dots, n$$

with ξ_i denoting the error of observation i , x_{ij} the set of j variables for observation i and σ_i^2 the observation-dependent variance. We note that for regression models like ordinary least squares the variance is assumed to be constant. From Eq.III.1 follows that the absolute value of the residuals grows with the mean of gene expression and we obtain the respective relation for the concrete case of a count matrix \mathcal{K}_{ij} for i genes and j samples. In order to obtain appropriately normalized gene expression data being homoscedastic one applies a so-called *variance stabilizing transformation (VST)* to the count matrix \mathcal{K} . In general a VST is a transformation $h(\mathcal{K})$ of the random variable \mathcal{K} with constant variance. A possible description in terms of the variance depending on the mean is (see e.g. [8, 93])

$$h(\mathcal{K}) = \int_{\mathcal{K}} \frac{d\mu}{\sqrt{\text{Var}(\mu)}} \quad (\text{III.2})$$

Detailed information on how the dispersion relationship $\alpha(\mu)$ is fitted on the underlying data in order to obtain the explicit VST can again be found in [8, 214].

Results Applying this VST relation to all count values of the matrix \mathcal{K}_{ij} we obtain normalized gene expression values, which do not grow with variance anymore. This behaviour is depicted in Fig.III.5.

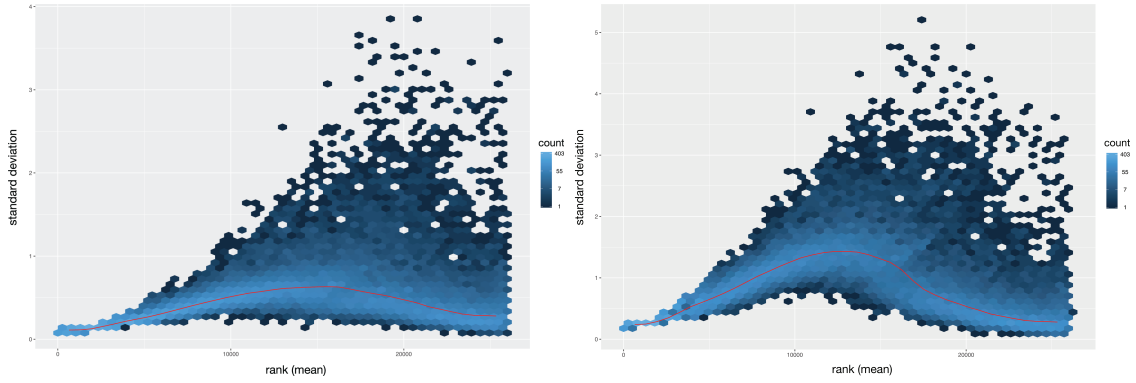


Figure III.5: Standard deviations for absolute gene transcript expression values determined via a VST (left) and via a classical library size normalization (right). The mean standard-deviation is indicated by the red line.

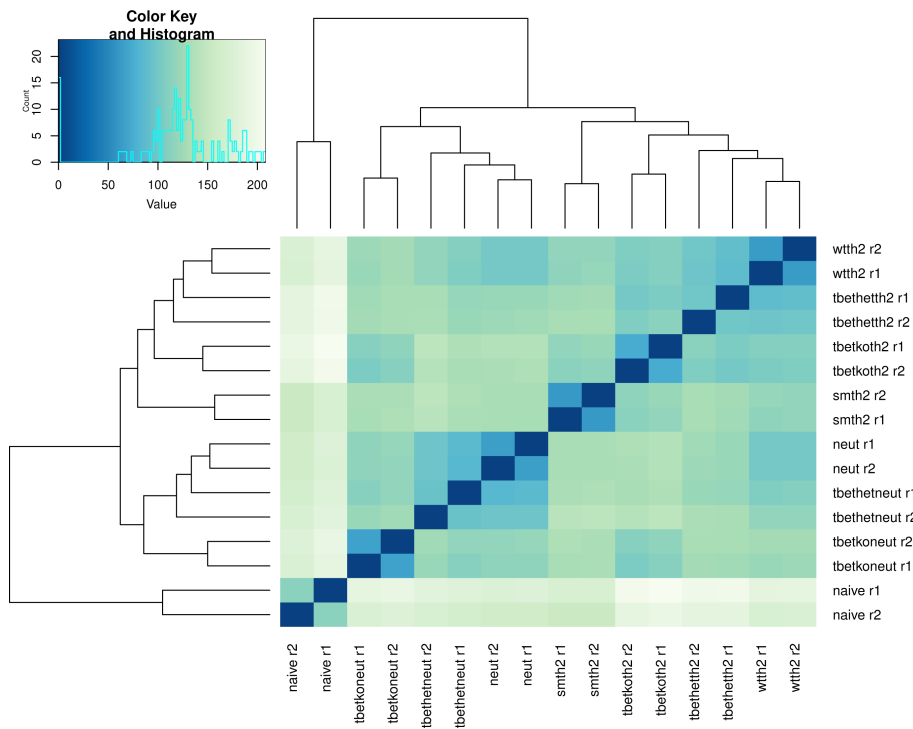


Figure III.6: Hierarchical clustering of the Euclidean distance VST count matrix for all RNA-Seq samples. *Neut* labels neutral conditions and *Th2* labels Th2 culture conditions. *SMTh2* denotes the Th2 control and *r1* and *r2* denote the two considered replicates respectively.

Employing the variance stabilized counts we investigate our RNA-Seq data by clustering the Euclidean sample distances. By doing so we are able to infer similarities between different samples and replicates and hence obtain another means to test if the VST was successful. For this we determine the elements of the distance matrix of the VST count matrix with elements $\mathcal{K}_{VST;ij}$. The result of this procedure is shown in Fig.III.6 employing hierarchical clustering. As can be readily expected the VST method is able to discriminate extremely well between the respective sample phenotypes not only clustering replicates correctly but also discriminating between Th1- and Th2-specific cell conditions. Furthermore the Euclidean distances are also high in between same condition replicates. We also find that the phenotypical differences between Tbet^{+/+}Th1 and Tbet^{+/-}Th1 are comparatively low. In addition we observe a

clustering of the underlying samples by considering the highest expressed VST genes in Fig.B.4, which is again fully consistent with the Euclidean distance method. Furthermore we investigate this important difference w.r.t. normalization of absolute expression values by doing a principal component analysis (PCA) of the variance stabilized samples in comparison to the heteroscedastic data. As it is always the case with a PCA we scan the uncorrelated variable space after performing an orthogonal transformation, which are called the principal components. Starting with the component that accounts for the largest amount of variance in the data (commonly called PC1 – here 48% variance) and comparing this to the PC with the second highest variance (PC2 – here 25% variance) as seen in Fig.III.7 we already find at this level that the VST is able to capture expected clustering behaviour between the different samples.

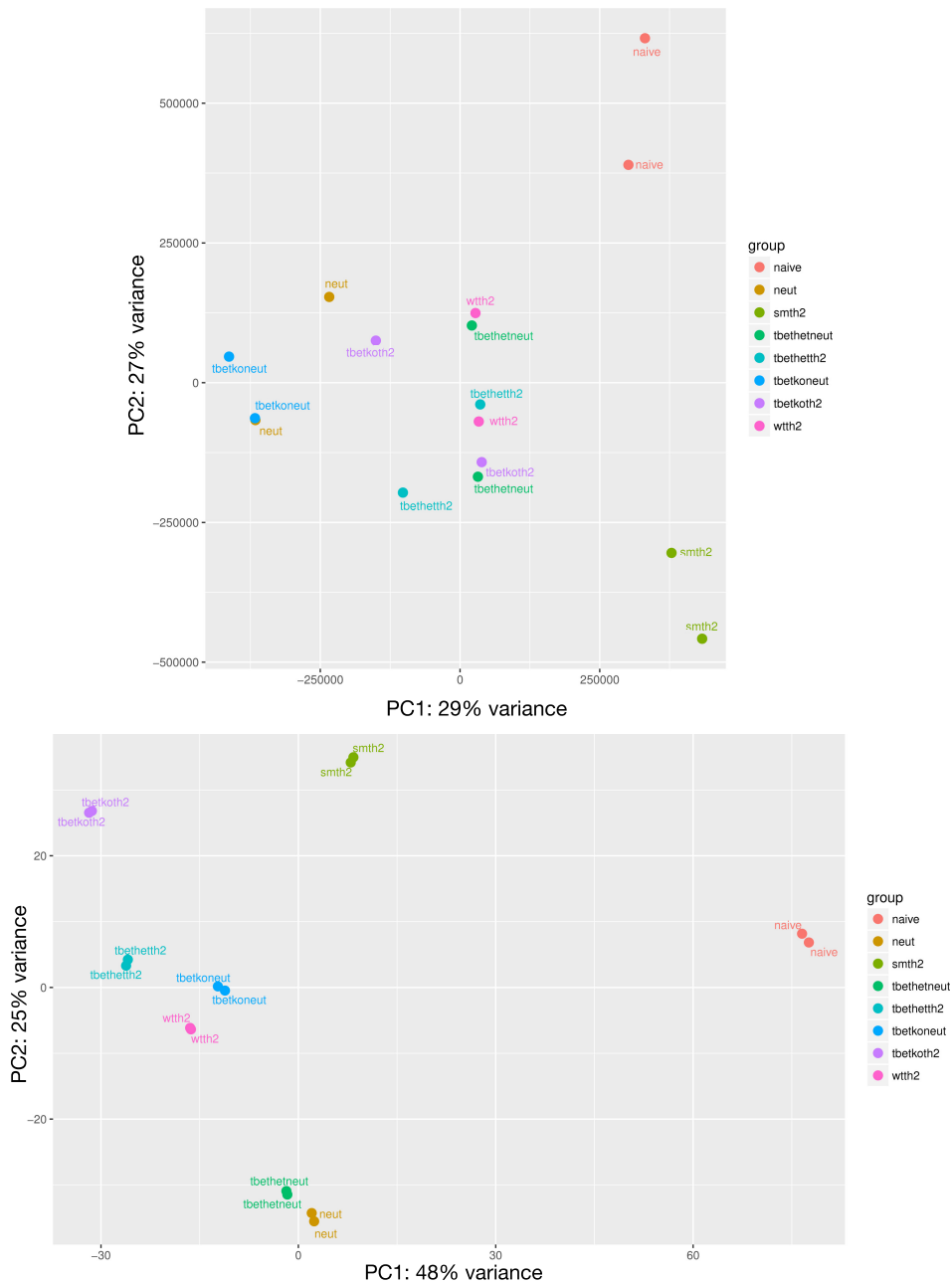


Figure III.7: Principal component analysis of all sample replicates for absolute values obtained by naive normalization (top) in comparison to VST normalized values (bottom) with sample labellings as before.

This is in stark contrast to the unstabilized counts where the phase space shows heavy mixing of the samples for both PCs. Up to this point the contentual importance of the principal components is an abstract one. We can infer some meaning on the interpretation of the PCs by observing the variable space containing the actual samples. We find that the PC1 variable is able to discriminate particularly well w.r.t. Tbet-dose and between neutral and Th2 polarizing conditions. This becomes especially evident when going to lower PC1 values where the ranking is $Tbet^{+/+}Th1 \rightarrow Tbet^{+/-}Th1 \rightarrow Tbet^{-/-}Th1$ followed by the Th1/2 conditions. The second PC is able to distinguish even better between Th1 and Th2 conditions and even shows functional similarities between the Th2 control and $Tbet^{-/-}Th1/2$ as well as between $Tbet^{+/+}Th1/2$ and $Tbet^{-/-}Th1$, which can also be expected. A particular good distinction can now be made from the combination of these first two PCs, which separate the naïve and the Th2 control samples from the Th1 cells under neutral and Th2 culture conditions. In Fig.B.5 we also show the third and fourth PC for the sake of completeness. We learn that although still contributing 15% and 7% to the total variance PC3 only puts additional focus on the Tbet knock-out conditions while already the interpretation of PC4 is not straightforward anymore.

From the above analysis we find that the VST leads to significantly low variability between the two replicates in all conditions and separates the Tbet dose grading as well as the cytokine condition dependencies accordingly. Yet the same is not true for the untransformed gene expression values. For these reasons we adopt the VST count matrix with entries $\mathcal{K}_{VST;ij}$ for further downstream analyses.

III.2.3 Data pre-processing pipeline

In the schematic depiction shown in Fig.III.8 we summarize the computational pipeline for the analysis of the sequencing data.

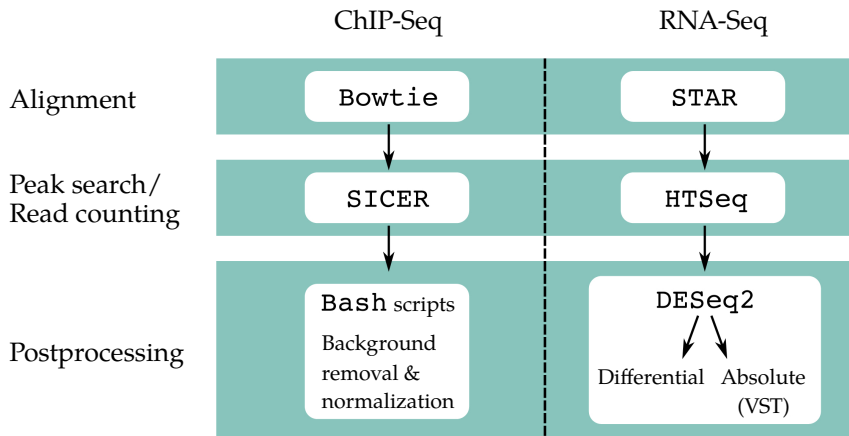


Figure III.8: Pre-processing pipeline for histone modification ChIP-Seq data and RNA-Seq data samples.

After quality assignment the ChIP-Seq data are aligned with Bowtie to the mouse genome build mm10 and subsequently analyzed with the peak search algorithm SICER for broad and narrow peaks. Afterwards the obtained peak files with background-subtracted histone modifications are normalized with respect to library size via bash-scripts. After checking for read trimming to enhance read quality the RNA-seq data are aligned to mm10 via the spliced-aligner STAR. The aligned reads are then

mapped to transcript isoforms where only unambiguous reads are counted via HTseq. Finally we perform different normalization procedures for differential gene expression analysis and for obtaining absolute values for further downstream analyses. For this we mainly use the DESeq2 implementation for variance stabilization of gene expression counts as well as applying a Wald-statistic for LFC significance computation.

III.3 Discussion & Summary

We have pre-processed and pre-analyzed the underlying histone modification ChIP-Seq as well as RNA-Seq data of an experimental setup consisting of long-lived Th1 cells from LCMV infected mice being subject to varying Tbet dose conditions as well as partially reprogrammed hybrid Th1→Th2 cells exhibiting plasticity, which were obtained from Th1 cells exposed to Th2 polarizing conditions. In addition we investigated naïve and Th2 control samples. We saw that besides differences in cytokine dose the underlying experimental system shows distinct gradings in Tbet dose. This dual dose dependency leads to different plasticity levels, hence Th2 control and Tbet^{-/-}Th1/2 conditions as well as Tbet^{+/+}Th1/2 and Tbet^{-/-}Th1 show similarities w.r.t. their genotypes. This was shown e.g. in a subsequent PCA. After data quality assessment we laid out a pre-processing workflow which on the one hand is able to deal with broad as well as with narrow histone modification peaks and on the other hand yields robust absolute gene transcript count data. We also checked for the consistency of the replicates and the conditions of the gene expression data, yielding clear distinctions arising on the basis of the histone modification landscape as well on the basis of the respective genotypes. Also we assessed peak variability around *Ifnγ* exhibiting changes in height and width for notable marks depending both on cytokine signal and on Tbet dose. We saw that in the case of absolute gene transcript values extreme care had to be taken with respect to normalization which was resolved by utilizing a VST method leading to homoscedastic data. Furthermore we extracted a range of top-ranked Th1- and Th2-specific genes by differential expression analysis which was merged with known genes from literature and which will be used later on in downstream analyses for obtaining an epigenetic network.

CHAPTER IV

Inference of chromatin states in T-helper cells

IV.1 Pattern recognition of epigenetic states

In the last chapter we saw that around notable genes of interest we can find epigenetic landscapes of different histone modifications, which not only exhibit peaks of different shapes w.r.t. to their height but also w.r.t. to their width and hence w.r.t. integrated read count. Yet we also find that different histone modifications do not necessarily behave in the same way, i.e. H3K27ac might not strictly follow the profile changes in H3K4me1 while at the same time repressive histone marks can behave complementary in one way or another. This leads to the observation that the phase-space of profile combinations of peaks of different histone modifications at specific points on the DNA theoretically gets immensely large. We are interested in finding specific histone modifications patterns such that we can assign an epigenetic state to a specific histone modification peak combination for an a priori arbitrary number of histone marks.

One of the obvious remaining questions is how many of these combinations are of utmost interest based on the underlying data and the theoretical questions we are posing, namely investigating activation and inhibition of genes. Since in the case of transcriptional activation where we want to find enhancer structures we already know that states, which include a significant amount of H3K4me1 w.r.t. to the rest of the genome should be able to appear in combination with H3K27ac while at the same time excluding the occurrence of repressive or silencing marks or alternatively alone. The question of other combinations is at this point a combinatorial one depending on the number of the observed histone marks as well as on the quantitative significant differences over multiple samples when several peaks appear at one position. We could for example imagine that we might find instances of histone mark combinations where one modification appears to a lesser extent than others. On a genome-wide scale taking into consideration different samples the question is now if such a pattern (or from now on termed *chromatin state*) occurs frequently enough to be considered significant or if a similar pattern occurs more often and the appearance of the former would hence be classified as the latter.

Basically an arbitrary amount of fine-grained subclasses can be inferred taking into consideration all possible peak combinatorics. We hence learn that we have the problem of finding a categorial annotation based on a given model with a certain

number of expected patterns, which also depend on their immediate surroundings. This means it should be less probable to find a repressive chromatin state next to an enhancing chromatin state in one cell condition while at the same time the probability for finding a state only exhibiting H3K4me1 next to one exhibiting H3K4me1 in combination with H3K27ac is more probable since in this case the H3K27ac peak might just have ended at some point. Furthermore we do not have any a priori expectation about how and when an enhancer state or a repressive chromatin state might occur at a certain position apart from knowing what histone marks to expect or not to expect. This is a problem commonly resolved by unsupervised machine learning methods, which are of special importance in inferring hidden variables or patterns of unlabelled data sets. Since there are several of such methods that are designed to cope with these kind of problems, we note that when observing sequential dependencies what comes immediately to mind are so-called Markov models or more generally dynamic Bayesian networks. For this reason we already discussed basic properties of a special case of dynamic Bayesian networks i.e. Hidden Markov Models (HMM) in section II.3.2.

IV.1.1 Method

The basic questions regarding an HMM now are: what are the observables in our data sets and what is the interpretation of the yet to be inferred hidden variables or states? Observables are in our case clearly the sets of significantly called peak islands for the entirety of the measured histone modifications. At a certain position on the DNA the respective observable is the overlap of the existing peak signals, which are observed as binary entities, i.e. as being present or not. The hidden state variable is the respective underlying pattern, which is generated by this overlap. The probabilities of assigning a certain state to an observable depends on the sequence of observables, hence generating a certain output from a hidden state (i.e. the emission probability) or switching from one state to the other and creating a different output (i.e. the transition probability). In our case this just means observing the sequential order of the peak structure combinations at subsequent nucleosomes for different samples.

A popular implementation of an HMM, which can readily cope with peak structured files, is the JAVA-based ChromHMM algorithm [103], which basically assigns the meaning of observables as discussed before to the data structure provided from the data processing steps above. The basic input for the inference of the underlying HMM hence is the peak call information for all the histone marks and all the conditions of interest. Additionally one has to specify the fixed number of hidden states according to which the model is trained. Obviously a whole range of models may be feasible for our particular system as well as for the questions we want to answer with it. Since we have a total of five histone marks at our disposal we should have a model with at least five hidden states, s.t. we allow for the possibility of every mark occurring in its own unique state, although this does not necessarily have to be the case. We also consider at least every possible pairwise combination of histone marks to be possible, s.t. we should also consider at least a maximum of $5^2 = 25$ hidden states. In order to classify the whole genome for every cell condition the different multivariate models first have to be learned w.r.t. their parametric relations. The parameter estimation procedure for the emission and transition probability distributions is based

on an expectation-maximization approach via the Baum-Welch algorithm with convergence to some arbitrary maximum likelihood value which is due to the fact that the initial log-likelihoods differ. The state likelihood is sequentially updated after every chromosome starting from an initial set of parameter values. As a training sample all chromosomes are used for 200 learning iterations until convergence. The respective input commands are given in appendix D.

The respective hidden states are obtained by first segmenting the genome into subsequent 200 bp bins, which are roughly the size of a single nucleosome. From the significantly called peak input the bin information is binarized s.t. a histone mark peak is labelled as existent or not in the respective bin. We use the formerly specified peak call parameters in this step (see section III.2.1). This binarization is then used as an input for the learning procedure. In order to find a model that offers a sufficiently low number of states to capture all relevant biological features we started with a randomly initialized 25 state model and compared it with models down to 5 states. The parameter results of all HMMs are shown in Fig. IV.1. A general penalized log-likelihood of all the state parameters was achieved by Bayesian information criterion (BIC) scores with a BIC penalty of $\ln(1.09 \cdot 10^6) \approx 18.5$. In table C.5 we can see the BIC results of all the differently seeded HMMs.

We find that the BIC (as well as alternatively the Akaike information criterion (AIC), which is more weakly penalized and hence discarded in this case) is not able to penalize the models sufficiently according to their large differences in increasing log-likelihood. This is due to the fact that for $k' > k$, where k and k' are different amounts of free parameters, we always obtain $\Delta\text{BIC} = \text{BIC}_k - \text{BIC}_{k'} \gg 2$, while the same holds true for ΔAIC . Basically this means that the higher the number of parameters the “better” the model¹. Especially for a low number of states it has been shown before [134, 158] that the BIC or for that matter the AIC alone are not well discriminating scores. Since also other publications (see e.g. [104, 105, 158, 194]) suggest a state number between 10 and 25 for a low number of histone marks we have to resort to a different quantitative method to infer a viable number of model states. In order to do this we prune the highest state model (in our case 25 states) down to the minimal reasonable number of states (5 states). During this process we want to find the lowest number of hidden states which still capture a sufficient biological interpretatory content. This means in particular that “relevant” states are still preserved when removing transition probabilities. “Relevant” here means that we are looking for states that do not give us solely a mixed emission probability, i.e. a fine-grading w.r.t. a mixture of probabilities, like states 12 or 23 in the 25 state model, but rather a distinct state with high emission probabilities like states 2 or 11 in the 25 state model. In Fig. IV.2 we see the heatmap of the correlations between a state of the highest state model and the recovery of that state somewhere within a lower state model.

We are looking for a plateau in the correlations where the removal of another model state does not change the recovery of the states from the full state model anymore. Also since we do not want the information from the most complex model to drop below a certain threshold some cut-off has to be applied. We find that around a state number of 16 states the information content of the hidden states is quite sta-

¹Hence the model with more parameters is always preferred (see section A).

Emission Parameters

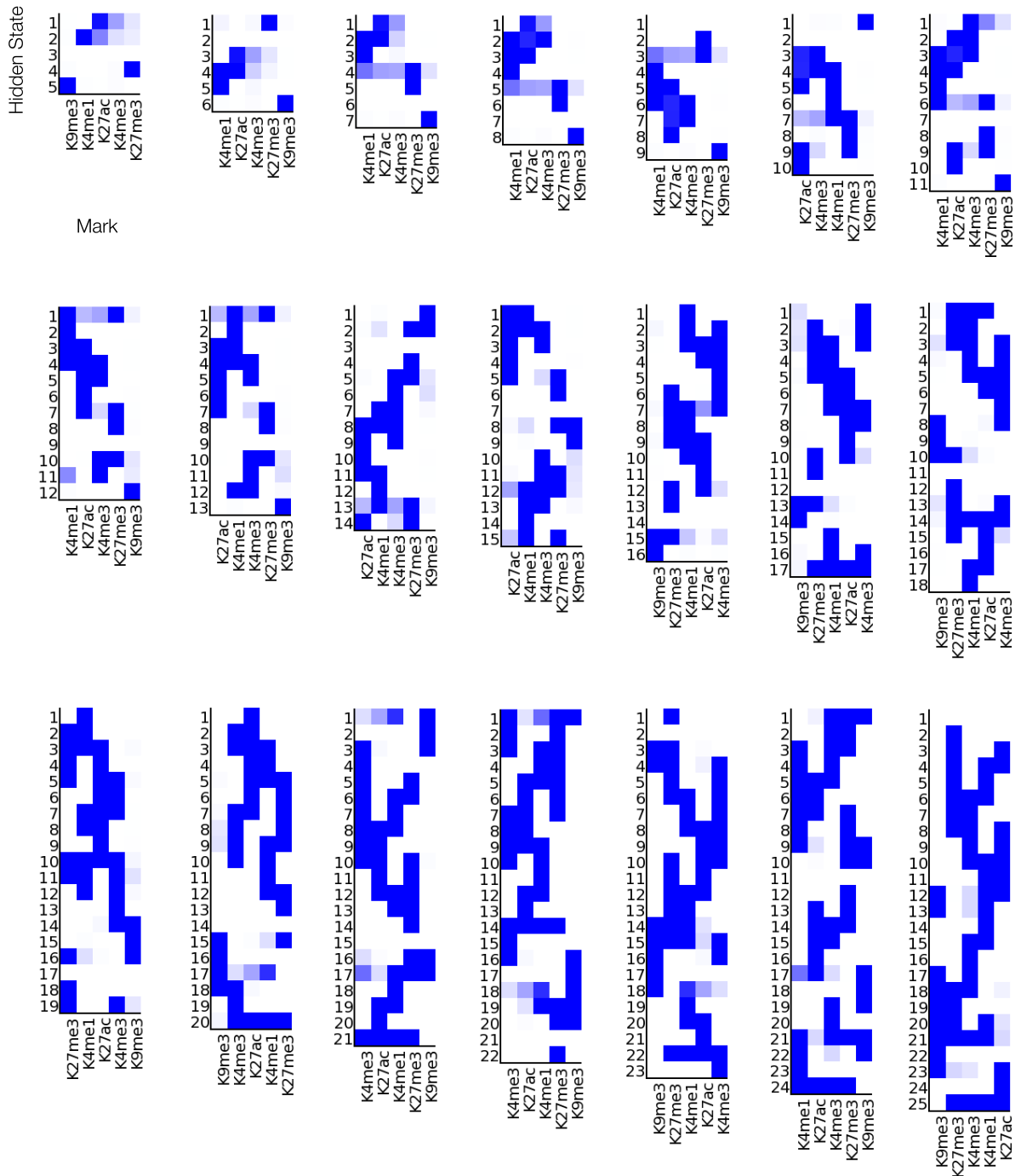


Figure IV.1: Emission probabilities for models with five up to 25 hidden states for the underlying five histone modification marks. The shades of blue represent the emission probability values of a histone mark within each state and hence are $\in [0, 1]$.

ble, while going to lower state numbers state information starts to decrease again².

²This can be seen in the marginal changes of the correlation heatmap values from a higher to a lower state HMM.

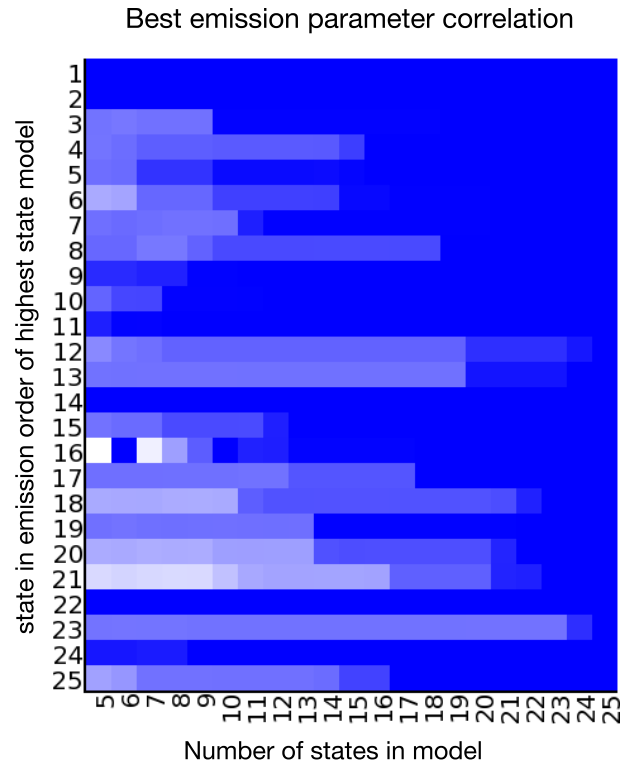


Figure IV.2: Best correlation value between emission parameters of the states in the 25-state model (y-axis) and a fixed number of model states equal or smaller than 25 down to a 5-state model (x-axis).

It turns out that hence a model with 16 hidden states presents a viable choice for a trade-off between including all the relevant information and a high model likelihood. Checking in Fig. IV.2 we would be interested in removing mixed states with low emission probabilities of a certain mark as often as possible without removing the information of another state completely. This would e.g. correspond to state 15 in the 16-state-model. Yet this state keeps reoccurring in similar forms in lower state-models as well with comparable emission probability and can hence be not eliminated. Higher state models on the other hand increase the fine-graining again. We can see exactly that with the example of state 15 from the 16 state model: in the 17-state model this state gets split into the more refined states 2, 3 and 13, which does not help in terms of interpretability of this state at all. Also the number of low emission probability mark states increases slightly when going from a 16- to a 17-state HMM. From now on the 16-state HMM will hence be considered.

IV.1.2 Results

Model parameters

In Fig. IV.3 we summarize the final emission and transition probabilities of the 16-state model with assigned annotation of the respective states. They form a complete set of estimated model parameters.

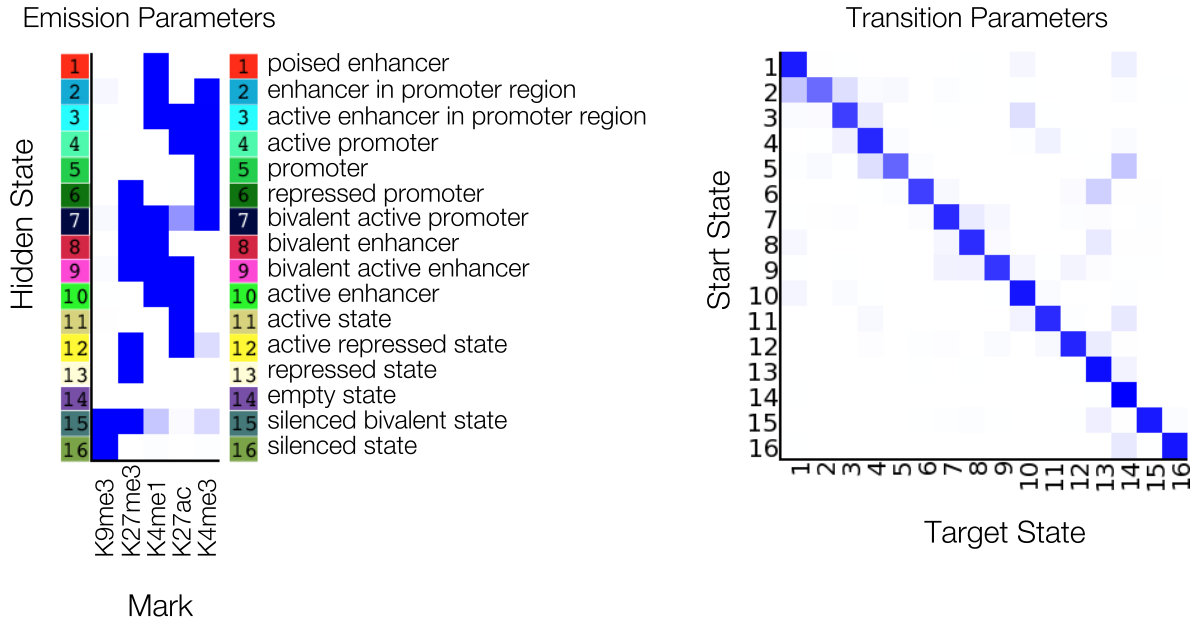


Figure IV.3: 16-state HMM parameters with emission probabilities (left) and transition probabilities (right). Depending on the respective histone mark combinations we assign an epigenetic interpretation to each chromatin state.

The interpretation of the emission probability heatmap is quite straightforward: we find a probability for every trained histone mark of occurring in the respective hidden state. It is of important note that for each individual state these mark emission probabilities are independent from each other. More specifically the probability distribution is a product of independent Bernoulli random variables (see [103]). The transition probabilities on the other hand are responsible for preventing an estimation bias for individual states, since the occurrence of a certain state also depends on its own neighbourhood, hence implementing Markovian properties in order to stabilize e.g. diffuse overlapped peak structures against small variations in the histone signals. In addition to that the transition probabilities also determine the relation between subsequent states hence define which state is more probable to follow another state given a certain observable signal combination. Alternatively it might be as well advantageous within a certain hidden state sequence not to switch to another hidden state but to remain in the original one depending on the underlying peak data.

We also find a corresponding percentage of each state occurring over the whole genome in each condition, which can be seen in table C.4. We find that state 14 in the model has the highest occurrence, which basically means that regions exhibiting none of the investigated marks make up most of the genome as can be expected. This is tagged as an empty chromatin state from now on. The other leading states concerning genome percentage are states 1, 10 and 13 which will indeed be of utmost importance since they are indicators of activation as well as of inhibition. Already

having a priori knowledge about histone modifications being marks of a certain epigenetic functionality (see section II.1.2) we annotate each inferred chromatin state in that fashion, which can be seen in Fig.IV.3. The non-empty states can be roughly grouped into four main clusters, s.t. we obtain enhancing (containing a significant amount of H3K4me1) and repressing/silencing (containing a significant amount of H3K27me3 or H3K9me3 respectively) states as well as those corresponding to promoter regions (containing H3K4me3). In addition we obtain a group of states which exhibit enhancing as well as repressing/silencing histone marks at the same time. These features are termed bivalent states since they represent “undecided” primed states for two opposing functional programs. They can basically unfold fully enhancing or repressing functionality under slightly different cell conditions. We also find that there are several promoter regions that carry enhancing or repressing marks, mainly depending on (or rather causing) gene expression activity.

Additional interesting information on the interpretation of chromatin states can be obtained by checking the overlap enrichment of significant functional chromatin elements such as CpG islands as well as the positional dependency of the individual states with respect to TSSs and TESs. A cell condition dependent depiction is shown via fold enrichment heatmaps in Fig.IV.4 for the Tbet^{+/+}Th1 condition.

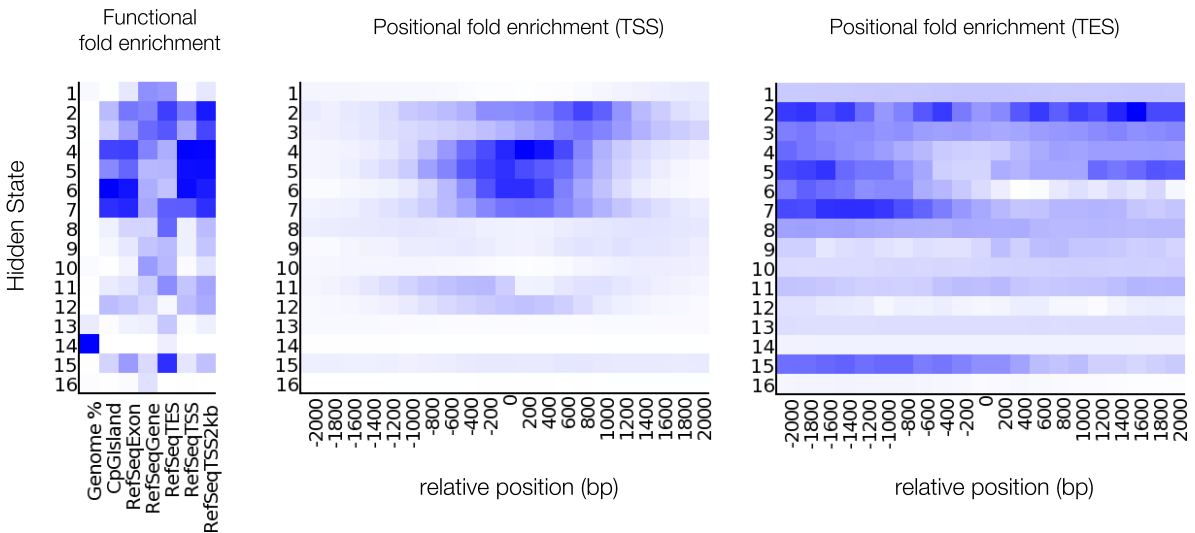


Figure IV.4: Functional fold enrichment of the 16-state HMM w.r.t. each individual state (left). The references are taken from the RefSeq annotation. The middle and right figure depict the fold enrichment w.r.t. the relative position of the TSS and TES respectively.

What is shown is the overlap enrichment in terms of fold change of a certain state overlapping a feature w.r.t. the whole genome. For the positional enrichment the fold enrichment is calculated with respect to all RefSeq genes. As already suspected we find for the TSS high enrichment especially of states 2-7 which are all related to the promoter mark H3K4me3. These states are also highly relevant w.r.t. occurrence of CpG islands. Concerning the TES we find a more distributed situation over all states with state 15 as a mixed state with silencing marks standing out. Yet especially within RefSeq genes and TES enhancer states like 1 and 10 start to appear more clearly indicating the existence of intragenic enhancers.

Chromatin state annotation examples

Turning the attention first to active promoter regions we observe two prime examples for Th1 and Th2 cells namely the promoter regions of the master regulators *Tbx21* and *Gata3*, which is shown in Fig. IV.5.

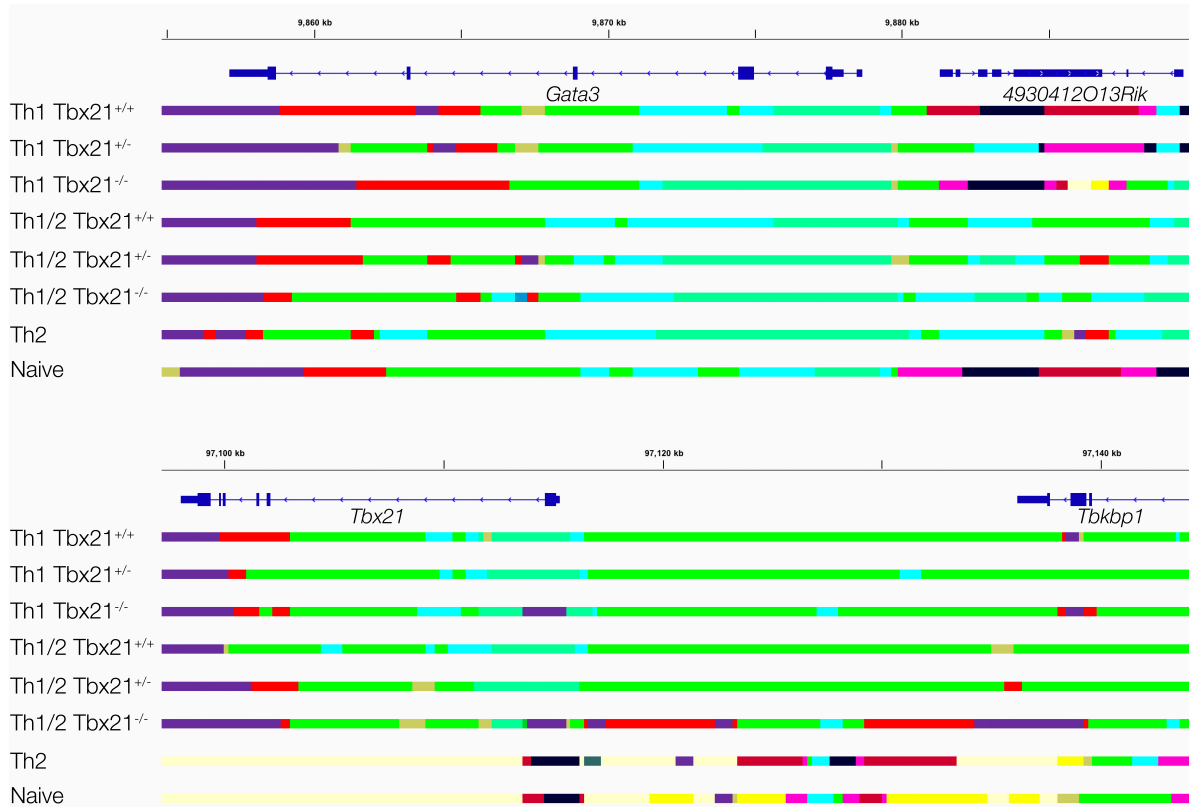


Figure IV.5: Chromatin state landscape of *Gata3* and *Tbx21* according to the 16-state HMM with state color-coding as defined in Fig. IV.3.

In agreement with canonical knowledge and with our own RNA-seq results we find a regulatory landscape around the *Gata3* promoter in Th2 cells that significantly exhibits the promoter-associated histone modification H3K4me3. In addition we find the activity mark H3K27ac. We find the same for the *Tbx21* promoter in Th1 cells. This results in the appearance of the chromatin states 3 and 4 in our model in the regions of the promoter and the TSS. Their interpretation hence would be active promoters in the case of state 4 and active promoters with enhancer marks or enhancers in promoter regions for state 3. In the case of *Tbx21* we see that the promoter and TSS region mainly consists of chromatin states 3, 4 and 10, yet in the case of Tbet knock-out we find in Th1 and Th2 culture conditions that the state changes to state 14 indicating empty chromatin with respect to our histone modification marks. This makes complete sense concerning the fact that we consider a Tbet knock-out. Interestingly the naïve and Th2 control conditions show the appearance of states 7 and 8, which are coined as bivalent enhancer/promoter states. Now the present chromatin states hence contain a significant amount of H3K27me3 repressing the activity of the promoter region. Hence we find already on the level of the HMM that chromatin states might switch between different functional states according to their combined

histone modification pattern from one condition to another, which is in our case dependent on cytokine signal and Tbet dose. When turning to *Gata3* on the other hand we see that no significant changes of chromatin states can be observed in the promoter region. If we now want to find differential changes between the conditions we will have to integrate the modification count somehow, which we will address later on.

Of utmost interest to the subsequent analysis will be the detection of enhancer states which positively regulate gene expression as we already discussed shortly in section II.1.2. Around *Tbx21* we find an extended chromatin state region ranging from its intragenic region to the intragenic region of the next-nearest gene upstream, *Tbkbp1*. For most of the conditions under consideration, especially for the Th1 cells under neutral conditions but also for the wild-type and heterozygous Tbet hybrid Th1/2 cells we mainly observe state 10, which in our model corresponds to an active enhancer chromatin state. The spatial extension of the state suggests a so-called *super-enhancer*, as already indicated in e.g. [154], the existence of which shall be discussed in due course. We also find that these active enhancers disappear in Tbet^{-/-}Th1/2 cells as well as in naïve and Th2 control conditions. The chromatin states in these conditions span a wide range from poised enhancers to bivalent and repressed states. While under Tbet knock-out in hybrid Th1→Th2 conditions a formerly active enhancer state (e.g. under Th1 conditions) loses its active mark and we obtain a poised enhancer, the situation looks completely different when observing the transition from naïve to differentiated Th1 cells. In naïve cells much of the upstream *Tbx21* region still is either in a repressed chromatin state either with or without an active histone mark present or in a bivalent active or inactive state with an enhancer mark already present. In the Th1 cell condition these states additionally gain either H3K4me1 and/or H3K27ac and they lose the repressive H3K27me3 mark. We observe that heterochromatin silencing as indicated by the H3K9me3 mark does not play a role at this particular example locus. We will find out later on that when performing a genome-wide analysis on Th1- and Th2-specific genes H3K9me3 also very rarely occurs. This can be also confirmed w.r.t. its overall state-specific occurrence via states 15 and 16 of approximately 0.6% of the genome over all cell conditions.

As mentioned earlier one of the best annotated and studied gene loci in Th1 cells is *Ifnγ*, a cytokine important in the function of the innate and – more importantly for our data – of the adaptive immune system. As is shown for example in [19] the *Ifnγ* locus consists of a multitude of validated regulatory elements taking the function of enhancers. We focus on so-called conserved non-coding sequences (CNS) that have been shown to be sufficient for part³ of the enhancer functionality (see e.g. [56]). The *Ifnγ* locus is depicted in Fig.IV.6 including the positions of the validated CNS sites (see [19]). In addition to the HMM classification of chromatin states for the considered cell conditions we also show again the peak files for selected conditions of one of the two biological replicates. As an important marker for enhancer activity (see section II.1.2) we also include published ChIP-Seq data of histone acetyltransferase p300 [314] for Th1 cells as well as for Th2 cells as a reference.

³We note that not all CNS sites necessarily have non-redundant functionality in all cell types. See for example [74].

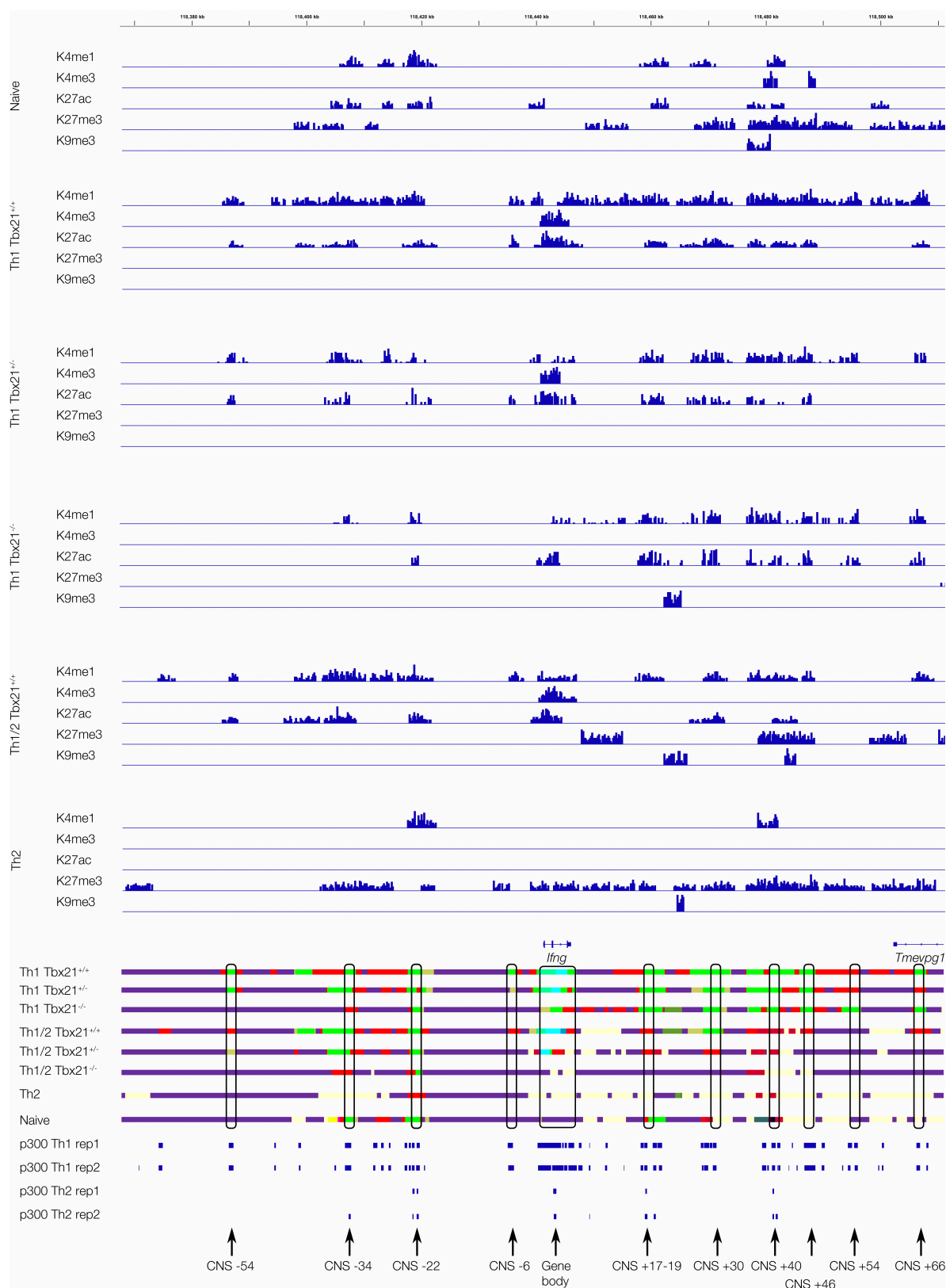


Figure IV.6: Annotated *Ifn γ* locus with colour-coded HMM state segmentation as well as depiction of investigated CNS sites from literature [19]. Also shown are again selected histone modification peak landscape results as well as an annotation of p300 binding peaks [314].

We find a multitude of distinct delimited regions containing in one or several cell conditions either an active enhancer state (state 10) or a poised enhancer containing no significant H3K27ac accumulation. Of high importance are obviously those active enhancer states that are found in Tbet^{+/+}Th1 conditions since *Ifn γ* is a hallmark of these cell types. Furthermore we find in the RNA-Seq data that *Ifn γ* is most highly expressed under these conditions as can be readily expected. From the histone modification peaks we already observe a strong increase of H3K4me1 from naïve to Th1 cells on average over the whole gene locus.

Simultaneously H3K27me3 disappears in Th1 cells, while in general it becomes more dominant going to Th1/2 and even more so when considering Th2 cells. We find an analogous behaviour when turning to the inferred HMM states: from Th1 to Th2 cells we see active enhancer states disappear while repressed states appear. Also in the intragenic region we find active promoter states, in some cases also exhibiting enhancer marks, which disappear in naïve cells and even become repressive states in Th2 cells. Downstream of *Ifn γ* at CNS+40 we even find a superposition of enhancing and repressing marks forming a bivalent state in Th1/2 hybrid conditions.

The HMM classification is able to reproduce and validate basic enhancer features of the aforementioned CNS sites. For wild-type Tbet Th1 cells we find active enhancer states at all CNS locations except CNS+54 where H3K27ac was not significant enough to qualify as an active enhancer. Its classification is hence a poised enhancer. The CNS sites also overlap with p300 binding sites from Th1 cells, while we also observe some overlap of repressive states in Th2 cells with p300 binding sites in the Th2 condition like e.g. in the gene body, at CNS-34, CNS+17-19 or at CNS+40.

We also observe strong dependence of the chromatin state classification on Tbet dose, especially w.r.t. enhancer activity in accordance with *Ifn γ* expression. This is prominently seen at the CNS-6 enhancer. Here Tbet heterozygosity leads to a loss of the H3K4me1 mark leading to the solely active state 11, while under Tbet knock-out the region is classified as an empty chromatin state. Similar phenomena can be observed in the gene body where we not only find a decrease in H3K4me1 and H3K27ac, which is also reflected in the state classification, but as well with respect to the promoter mark H3K4me3. Furthermore we can confirm analogous behaviour in hybrid Th1/2 cells where we also find dependence on Tbet dose as can be readily seen in the gene body or e.g. at CNS+30. More generally we observe less enhancer activity in Th1 and hybrid Th1/2 cells with diminished Tbet dose. Hence enhancer states do not only change their activity state w.r.t. cytokine stimuli but also w.r.t. different Tbet dose conditions turning to poised, bivalent or even repressive states. This prominently shows that Tbet in fact does affect the epigenetic landscape, at least at the example locus of *Ifn γ* . This analysis also undermines the importance of the H3K27me3 mark for the definition of an enhancer state. From the point of view of the inferred HMM chromatin states an enhancer is hence not only defined by its activity in the cell condition of highest expression of the gene to which it supposedly belongs, but also to some yet to be determined extent on the quality of the chromatin state change for different stimulation and dose conditions as well as on the amount of the repressive histone mark H3K27me3 in an antagonistic condition.

We also investigated p300 binding sites in general around notable Th1 and Th2 genes (which are listed in table C.3) as well as their overlap with active or poised enhancer states as found by the HMM. In Fig. IV.7 we see that a notable amount of p300

binding sites indeed has significant overlap with active or poised enhancer states in wild-type Th1 and Th2 cells as well as in Th1/2 conditions. We delimited the region for determining associated enhancers around those genes to so-called *topologically associating domains* (TADs) (see section II.1.2) from published data (see [89])⁴. We found 1092, 1011 and 1217 poised or active enhancer elements in these TADs in Th1, Th2 and Th1/2 cells respectively, where 62%, 60% and 63% of these contained at least one p300 binding peak in combined ChIP-seq data of differentiated Th1 and Th2 cells, thus supporting the HMM model. Yet we note that we also find it to be neither necessary for an active enhancer to appear at a p300 binding site nor compulsory for p300 to be a preliminary for enhancer occurrence as confirmed e.g. in [267]. Furthermore some HMM enhancer sites also bind p300 in Th1 and Th2 cells at the same time. Some of those instances appear to be enhancers that don't change their activity state significantly w.r.t. cytokine stimulus or Tbet dose, while some other instances exhibit repressive chromatin states in one of the two cell conditions. One prominent example for the latter is CNS+17-19 downstream of *Ifn γ* .

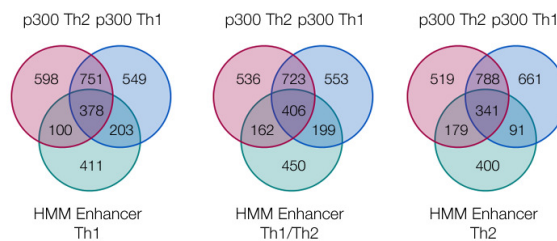


Figure IV.7: Venn diagram depiction of overlap of p300 binding sites in Th1 and Th2 cells with HMM enhancer states in Th1, Th1/2 and Th2 cells respectively.

Performing additional genome-wide analyses we rarely find instances of promoter states which do not occur in the immediate vicinity of any annotated RefSeq TSS. Nevertheless an extreme example with certain interesting implications is provided upstream of the gene *Il1rl1* with currently debated functionality in Th1 and Th2 cells alike [3, 129]⁵. At its TSS we find a promoter state (HMM state 3) denoting an active promoter with the inclusion of enhancer marks in Th2 cells. This has been reported before as being a distal gene promoter (see e.g. [3]). Yet at -36 kb upstream of this TSS we find another promoter state being active in Th1 wild-type cells hence being responsible for the opposing cell program. This is depicted in Fig.IV.8. It will be the focus of future experimental research to elucidate the question if such an instance inferred by an HMM alone can be a viable candidate for an alternate TSS for a gene like *Il1rl1*.

⁴The data-sets on embryonic stem cell topological domains in mouse can be found at <http://chromosome.sdsc.edu/mouse/hi-c/download.html>

⁵We also find comparable situations upstream of *Il21* having reported functionality in Th17 cells as well as e.g. at -11.5 kb upstream of *Tbx21*.

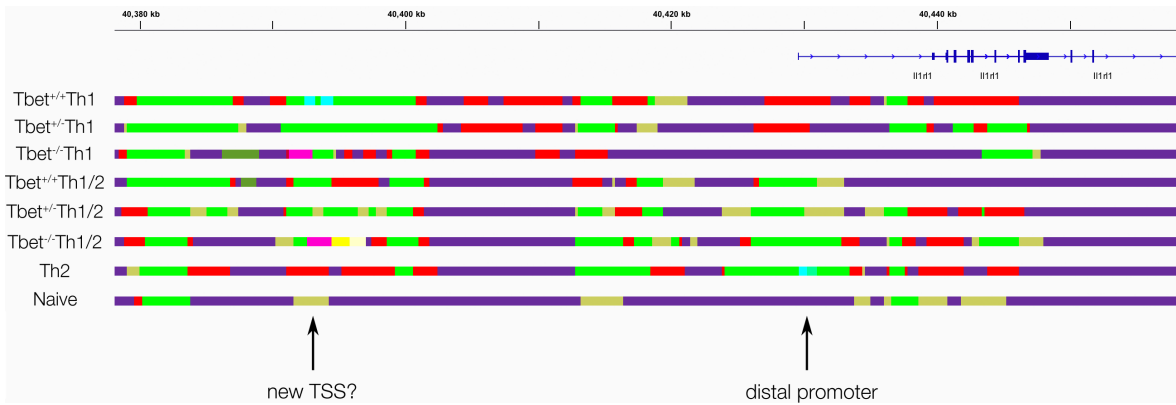


Figure IV.8: The *Il1r1* locus with the inferred HMM segmentation. The Th2 condition clearly shows a distal promoter state while 38 kb upstream we find another promoter state in the Th1 wild-type condition probably regulating the opposing cell program.

In the following we will nevertheless mainly focus on active and poised enhancer states as described for the above loci as well as on repressive chromatin states, which as we have seen form in distinct patterns, which was formerly barely recognizable from just inspecting histone modification peaks alone.

IV.2 Discussion & Summary

We have inferred the chromatin landscape w.r.t. five notable histone marks on a genome-wide scale in the previously described cell conditions using a Hidden Markov Model implemented in ChromHMM. By learning hidden state patterns from a combination of these histone marks based on their genomic sequential peak occurrence we tested several models with different state numbers and compared them based on their ability to reproduce a certain state information in all of the other models respectively. From this we obtain a minimal model consisting of 16 chromatin states, which are assigned genome-wide to all experimental conditions. For the following discussion we will adopt the following definitions for the inferred hidden states:

DEFINITION: Chromatin State

Be a significantly categorized superposition of histone modifications present at some position on the DNA we call a model-dependent irreducible pattern of such a superposition a chromatin state \mathcal{S} . The number of chromatin states depends on the underlying fine-graining of the statistical model. From an HMM point of view a chromatin state is determined by the individual emission parameters of the underlying observables in combination with the number of hidden states under consideration.

We furthermore define an enhancer state by

DEFINITION: Enhancer State

A chromatin state is called an enhancer state \mathcal{E} if we find a sufficiently high emission probability for H3K4me1 as well as a vanishing emission probability for H3K27me3.

while a repressive state is from now on defined by

DEFINITION: Repressive State

A chromatin state is called a repressive state \mathcal{R} if we find a sufficiently high emission probability for H3K27me3 as well as a vanishing emission probability for H3K4me1.

In the case where H3K4me1 and H3K27me3 appear significantly at the same time we coin these chromatin states bivalent.

We also find that the different chromatin states change their respective activity or functionality between cell conditions by switching in-between states. We observe two main drivers for changing chromatin state functionality like e.g. enhancer or promoter activity. This is on the one hand based on cytokine dependency as can be readily expected, since certain gene loci as e.g. *Ifn γ* are strongly T-helper cell-specific. On the other hand a strong effect is exerted by Tbet dose as well. We will follow up on this discussion later on since this will have enormous effect on chromatin state specificity for the ultimate epigenetic network underlying Th1 and Th2 cell differentiation.

In conclusion to the ChromHMM implementation and the resulting epigenetic state landscape we have to note that a shortcoming is the Boolean analysis of the previously analyzed peak structure of histone modifications. As a result state parameters are solely computed on the basis of the existence of a certain mark and hence only incorporate binary information. What is completely missing at this point in the analysis is any information on differential distinctions between different conditions as well as between states that are being classified identically yet differ in their respective modification load. Furthermore histone modifications are thought of as being independent from each other, which is realized in the HMM by assuming the observables of the respective hidden states to be independent Bernoulli random variables. Yet for an actual chromatin state these independent variables act in concert for its respective very definition as e.g. in the case of active enhancers. In order to shed light on these issues and to make predictions about similar behaviour w.r.t. gene expression we address the points in the subsequent chapter about the inference of a parametrized correlation measure for epigenetic transcriptional regulation.

CHAPTER V

Epigenetic landscape inference by implementation of a multivariate correlation measure model

V.1 A parametrized multivariate correlation measure

We already saw in the last chapter that there exists a wide range of possible variations of chromatin states in between different cell lines w.r.t. cytokine stimuli as well as w.r.t. Tbet dose. In other cases putative active enhancer elements seem to act constitutively over several or in some cases all cell conditions under consideration. Up to this point we only have knowledge about HMM chromatin states acting e.g. as viable indicating priors for enhancers, i.e. for epigenetic activation of gene expression. We want to find out at this point if for a certain element being tagged in some cell condition as an enhancer state we can find a co-regulating behaviour with gene expression itself, s.t. a prior for a possible causal relation can be inferred. Hence we are going to focus on enhancer states since they potentially present a well investigated way of describing gene activation. To this end we want to find a robust way of correlating enhancers with genes and preferably make a one-to-one mapping of Th1 and Th2 regulating enhancer-gene pairings. This will be done via the before analyzed RNA-Seq expression and histone modification data sets. Furthermore we will find in due course that this picture can be extended to repressive states as well.

V.1.1 Optimization of correlation measure

From first principles it is unclear if there are several histone modification that are co-regulated in the same way w.r.t. cytokine signal or Tbet dose and also in correspondence with the respective gene to which a certain chromatin state supposedly belongs¹. Later on the question of a one-to-one mapping between a chromatin segment or state and a specific gene will be addressed in more detail. Since we are focusing on positive correlations and gene regulation from proximal as well as from distal chromatin elements we have to take enhancer elements in general into consideration, meaning not only far away elements but also also intragenic enhancer states

¹Quite often in literature enhancers are just mapped to the next-nearest downstream gene (see e.g. [315]), yet this has not necessarily to be the case since gene-related enhancers have been shown to appear downstream of a TSS (see e.g. [257]) or even several genes away as well.

fulfilling the definition of \mathcal{E} , which also includes the addition of the promoter mark H3K4me3. In Fig.V.2 we show the integrated histone modification load for one exemplary enhancer state upstream of *Ifn γ* – CNS-34. We find that while the expression of *Ifn γ* increases we find an increase in the enhancer mark H3K4me1 as expected as well as in the active mark H3K27ac. At the same time we find a prominent decrease in the repressive modification H3K27me3. The promoter mark H3K4me3 is non-existent at this site – as being the case for most distal sites – which suggests exclusion from the analysis of distal chromatin states. The same holds true for H3K9me3, which as we have already discussed rarely occurs genomewide overall, and with an even lower frequency around Th1 and Th2 genes of interest. We already saw that active enhancer states, which overlap with CNS sites at the *Ifn γ* locus as well, involve both H3K4me1 and H3K27ac. Hence we are in need of a combination of both marks in order to obtain robust correlations. At the same time we observe that in several cases while enhancers states can disappear chromatin states containing H3K27me3 appear in the opposing cell program. One such example was CNS-54. From this we hypothesize that the repressive mark H3K27me3 plays an important role in enhancer regulation. We will validate this for a set of notable enhancers in the following.

Method

In order to obtain a robust measure for correlation we consider a set of experimentally validated enhancer sites in Th1 and Th2 cells. At these enhancer sites we parametrize a combined multivariate measure of H3K4me1, H3K27ac and H3K27me3. We assume that an adequate multivariate measure for correlation should exhibit the ability to maximize the correlations with the respectively attributed genes. Without loss of generality we also assume that correlations with enhancers should be maximized simultaneously in contrast to independent maximizations that would yield different parametrizations for every enhancer independently². We propose the following linear parametrization measure \mathcal{M} for histone modifications

$$\mathcal{M} = \text{H3K4me1} + a \cdot \text{H3K27ac} + b \cdot \text{H3K27me3} \quad (\text{V.1})$$

with free parameters a and b . Without loss of generality we set the prefactor of H3K4me1 to one, which is a gauge freedom of the linear equation. Non-linearities are not considered at this point since we have no reason to assume that the three considered histone modifications are in some specific way necessarily coupled to each other or that there is some evidence of some assisted loading recruitment mechanism between several histone modifications. This also manifests in the fact that in our 16-state HMM we find poised enhancer states, as well as active or repressed states with significant peaks, without the need for another histone modification to occur simultaneously or even at the same site in a different cell condition.

We already noted that H3K27me3 decreases at enhancer sites with increasing gene expression while H3K27ac behaves in accordance with H3K4me1. This leads

²Independent maximizations are furthermore undesirable if we want to obtain only one parametrization since only some enhancers might profit from this method with high correlations while at the same time this lowers the correlation of other enhancers.

to boundary conditions on the sign of the parameters a and b s.t. we require

$$a > 0 \quad \wedge \quad b < 0. \quad (\text{V.2})$$

For some learning sample of enhancer sites \mathcal{E}_i we hence obtain the following optimization problem:

We demand a maximization of the sum of the correlations of the histone modifications at enhancer sites \mathcal{E}_i with the expression of the respective gene \mathcal{G}_i :

$$\max(\mathcal{O}) \quad \text{w/} \quad \mathcal{O} = \sum_i \text{corr}_j(\mathcal{E}_{ij}, \mathcal{G}_{ij}) \quad (\text{V.3})$$

where \mathcal{O} denotes the objective function. Also we set $\mathcal{E}_{ij} \equiv \mathcal{M}_{ij}$ s.t.

$$\mathcal{O} = \sum_i \text{corr}(\text{H3K4me1}_{ij} + a \cdot \text{H3K27ac}_{ij} + b \cdot \text{H3K27me3}_{ij}, \mathcal{G}_{ij}) \quad (\text{V.4})$$

Expressions like H3K4me1_{ij} denote integrated read counts for an enhancer site i in condition j . Correlations are performed over conditions j . To make computationally meaningful predictions we have to minimize the negative objective function and hence compute

$$\min(-\mathcal{O}) \quad \text{s.t.} \quad a > 0 \quad \& \quad b < 0. \quad (\text{V.5})$$

We also consider Pearson correlations in contrast to Spearman correlations. The reasoning for this is as follows: first of all for Pearson correlations the relation between input and response is assumed to be linear while for Spearman the only assumption is monotonicity. While increasing the number of data points by considering replicates independently in order to obtain more significant results (for higher sample numbers the significance increases for high correlation values) we often encounter slight non-monotonicity arising from differences between replicates. This means it can be the case that we obtain slightly higher modification for one replicate accompanied by lower gene expression w.r.t. the other replicate and vice versa. For better comparability and to avoid outliers we also only consider samples exhibiting differences in Tbet dose. An example for this behaviour can again be observed for H3K4me1 in Fig.V.2. Spearman's rank correlation coefficient ρ hence might decrease due such possible artifacts while Pearson's correlation coefficient \mathcal{R} yields higher values. We also show the respective Pearson and Spearman correlations for the following analysis in Fig.B.6 from which we see that Pearson correlation indeed performs better. Furthermore we do not assume a non-linear relationship between gene expression and histone modification at this point, which is also reflected in the correlation measure w.r.t. which we perform our optimization.

The learning sample for the parametrization of the multivariate correlation measure is depicted in table C.6³. Most of the enhancer sites are found at the *Ifn γ* locus exhibiting the best annotation in T-helper cells in mice. For optimization we employ the `fminsearch` and `fmincon` routines in MATLAB since we face a smooth non-linear optimization problem for random search initial conditions in the interval $[-1, 1]$. The

³For more details on the individual enhancers in the learning sample we refer to the following studies [19, 171, 179, 344, 351].

optimization is then performed 1000 times for different initial conditions. For the underlying training sample we obtain the following parameter values, which are stable w.r.t. initial conditions

$$a \approx 1.24 \quad b \approx -2.82. \quad (\text{V.6})$$

This parametrization, which we will justify now, will be adopted for subsequent analyses from now on. In order to account for stability w.r.t. the training sample we apply resampling techniques. For a weak resampling we first use the jackknife method corresponding to a leave-one-out resampling. This yields the following mean and standard deviation

$$a = 1.2562 \pm 0.2034 \quad b = -2.8417 \pm 0.3243. \quad (\text{V.7})$$

We obtain the confidence intervals of the respective parameters via the distribution's quantiles. The quantile z^* of a Gaussian distribution is in general calculated (see e.g. [77]) via

$$z^* = 1 - \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \quad (\text{V.8})$$

where Φ denotes the cumulative distribution function of a Gaussian and α as usual denotes the significance level. For the 95% quantile, for a Gaussian being equivalent to a p-value of 0.05, we obtain the 95% confidence interval by

$$\text{CI}_{0.95} = \pm z_{0.975}^* \cdot \frac{\sigma}{\sqrt{n}} \quad (\text{V.9})$$

with n being the sample size. For a Gaussian distribution we have $z_{0.975}^* = 1.96$ (see e.g. [339]). This yields for the two parameters

$$\begin{aligned} a_{\text{CI}_{0.95}} &= [1.1478; 1.3646] & \sigma_{a_{\text{CI}_{0.95}}} &= [0.1503; 0.3148] \\ b_{\text{CI}_{0.95}} &= [-3.0145, -2.6689] & \sigma_{b_{\text{CI}_{0.95}}} &= [0.23957, 0.5019] \end{aligned}$$

At the same time bootstrapping the sample 1000 times yields an empirical distribution where the first two moments represent the mean and the standard deviation which reads

$$a = 1.3168 \pm 0.6601 \quad b = -2.8916 \pm 1.1238 \quad (\text{V.10})$$

respectively. Applying a bootstrapping procedure [96] the 95% confidence intervals lie at

$$\begin{aligned} a_{\text{CI}_{0.95}} &= [1.2746, 1.3590] & \sigma_{a_{\text{CI}_{0.95}}} &= [0.6316, 0.69128] \\ b_{\text{CI}_{0.95}} &= [-2.9634, -2.8197] & \sigma_{b_{\text{CI}_{0.95}}} &= [1.0753, 1.1770] \end{aligned}$$

respectively. We see from the resampling statistics of the two free parameters a and b that the initial estimates from the full training sample lie well within the respective confidence bounds w.r.t. their errors. Additionally we observe a slight skewness of the distribution for the parameter a indicating that because of its lower positive mean and relatively high standard deviation it tends more often to slightly larger values. This can be seen in Fig.V.1.

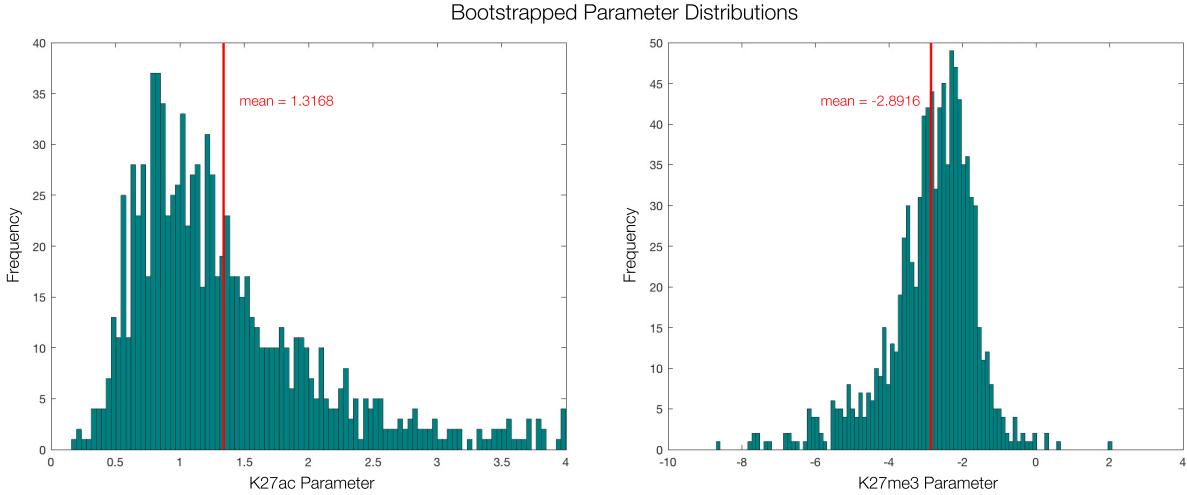


Figure V.1: Parameter distributions for a (left) and b (right) obtained from an $n = 1000$ bootstrap sampling procedure. Assuming an approximate Gaussian distribution we label the respective means with the red lines respectively.

The resulting multivariate histone modification measure that is hence assumed from now on reads⁴

$$\mathcal{M} = \text{H3K4me1} + 1.24 \cdot \text{H3K27ac} - 2.82 \cdot \text{H3K27me3}. \quad (\text{V.11})$$

Discussion

By and large jackknife and bootstrapping results are comparable for the underlying training set yielding mean values and error bounds for the parameters that represent consistent results for bounded and unbounded optimization routines being independent of perturbations in the initial starting values. The order of magnitude of the absolute values of the parameters is also the same while H3K27ac is comparable to H3K4me1 in its impact on correlation. The parametrized coefficient of H3K27me3 in the correlation measure, if present at an enhancer, seems to have a slightly higher absolute impact yet contributing negatively to the correlation measure. This supports the evidence that H3K27me3 is a highly important histone mark for the definition of enhancer activity. In general if the repressive mark starts occurring in one of the opposing cell conditions of a labelled enhancer we already conjectured earlier that this is a far stronger statement than observing an enhancer that solely loses its active mark H3K27ac. We saw in Fig. IV.6 that not only do we find active enhancers that are “switched off” due to a perturbation in Tbet dose or w.r.t. different cytokine stimuli, but we also find active enhancers that switch to a repressive chromatin state and hence exhibit significant amounts of H3K27me3. These enhancers play a special role according to the parametrized correlation measure since H3K27me3 influences their correlation with gene expression to a large extent.

To investigate these findings further we single out a prominent enhancer example of the underlying training set namely the *Ifn γ* enhancer CNS-34. We find overlaps

⁴We note that we neglect error bounds on the parameters in the computational framework later on in order to reduce computation time for genome-wide computations significantly. Yet this surely represents an aspect that will be included in future implementations.

of p300 binding sites at this location with active enhancer states in the HMM in all Tbet wild-type and heterozygous conditions as well as in the naïve T-helper cells as can be seen in Fig.IV.6. The knock-out conditions already exhibit loss of the active mark, classifying the region as a purely poised enhancer. In Th2 cells yet we find a repressive HMM state that solely exhibits H3K27me3.

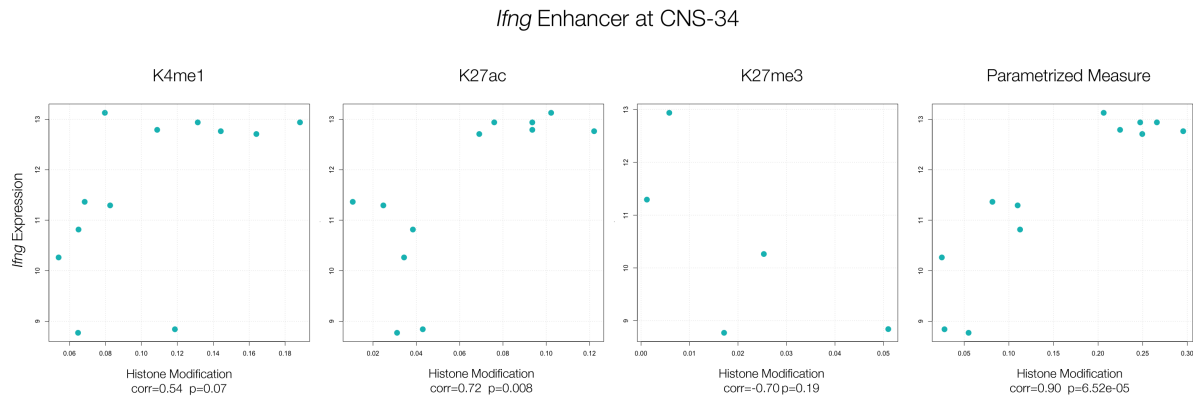


Figure V.2: Dependence of *Ifng* gene expression on H3K4me1, H3K27ac, H3K27me3 and the combined parametrized measure respectively at the *Ifng* enhancer segment at CNS-34. In all cases we observe non-monotonicity being among the reasons for the usage of Pearson correlations. The respective correlations values are indicated in combination with their respective p -values.

In Fig.V.2 we show the Pearson correlations for the individual modifications with expression of *Ifng* as well as with the obtained parametrized correlation measure of CNS-34. As can be readily expected H3K4me1 and H3K27ac individually correlate positively with gene expression with high statistical significance. At the same time we find significant negative correlation with H3K27me3. This can be also seen in the peak structure in Fig.IV.6 where we find pronounced peak signals in Th2 cells exhibiting the lowest gene expression values for *Ifng*. Turning now to the parametrized correlation measure we find that the situation becomes even more distinct. Not only does the correlation itself increase significantly, especially opposed to the traditional enhancer marks H3K4me1 and H3K27ac, which themselves only show comparably poor correlation values, but it also becomes more distinct and statistically significant.

In order to get more insight into the generality of this behaviour we redo the same analysis for all enhancers of the training sample obtaining the results shown in Fig.V.3. Here the enhancer elements appear in descending order of their correlation from the parametrized multivariate correlation measure. From the frequency distribution of the respective correlations we find that for the whole training sample the parametrized combined histone modifications yield higher correlations peaking slightly below $\mathcal{R} = 0.8$. Also the distribution is comparably narrow. The distributions for H3K27ac and H3K4me1 both peak at roughly $\mathcal{R} = 0.6$ being considerably lower while the active mark H3K27ac falls off much quicker to lower correlation values.

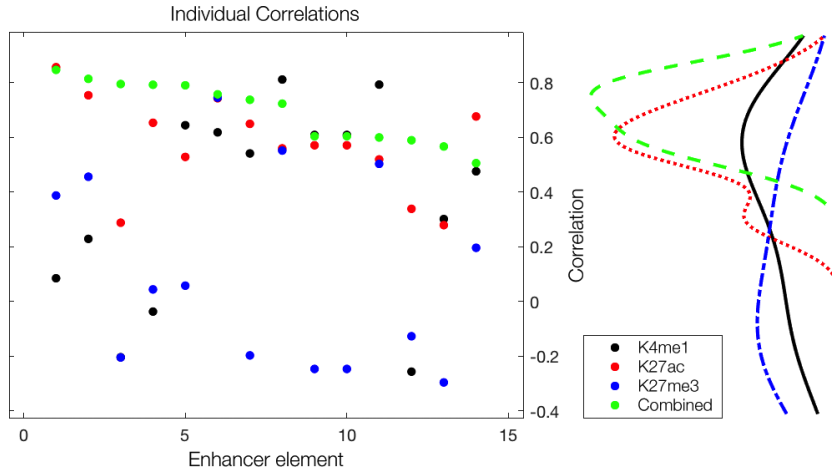


Figure V.3: Pearson correlation values of all enhancer elements from the training sample rank ordered decreasingly by their respective correlation via the parametrized correlation measure. Additionally we show the respective correlations of all individual histone modifications separately. On the right hand side we depict the distribution of these correlation values showing a distinct narrow peak with a considerably higher maximum than of all individual histone modifications.

For the classic enhancer mark H3K4me1 we find more low-valued correlations at enhancer instances up to the point where no correlation can be inferred and we even find slight anti-correlation (mostly being statistically insignificant). Anti-correlation is a phenomenon which can even be found on a genome-wide scale and will be discussed in more detail later on. This is to some extent due to the high variability of histone modifications over cell conditions in some enhancer areas where a higher number of data points (be it replicates and/or cell conditions) might instead yield positive correlation results and turn negative correlations into positive ones, hence improving on these results. We note additionally that although bootstrapping yields a significant validation of the full training set’s parameter estimation result additional experimentally validated enhancers that might turn up in future experiments will increase the ability to tighten the constraints on the parametrized correlation measure further.

We conclude that from the optimization of the above objective function we have hence obtained a robust parametrization w.r.t. initial conditions as well as random subsampling of the underlying training set of Th1 and Th2 enhancers which will be used in the subsequent analysis. We will see shortly how this can be applied in general to a genome-wide analysis and what implications these results have for the creation of a regulatory network of Th1 and Th2 cells.

V.2 Computational implementation

To this end we will exemplify in the following the power of the methodology of the parametrized multivariate histone modification correlation measure on basis of the underlying HMM for different epigenetic feature patterns. This should comprise regions with distinct well separated features as well as regions with broad, extended features (e.g. superenhancers). Furthermore we are interested in small uniquely defined regions where epigenetic features are thought to contribute directly only to the

expression of a single gene as well as in crowded regions w.r.t. the number of occurring genes where such a distinction cannot be straightforwardly made. In order to investigate the analytical implications of the results we will hence focus on some specific Th1 and Th2 cell loci exhibiting the exemplary cases stated above. The proper computational analysis on a genome-wide scale yet faces several intricate problems and hence conceptual refinements that have to be clarified first, which we will lay out in the following. For this we set up a novel correlation algorithm included in a specific computational framework implemented in R and bash. Among the reasons for the implementation in those programming languages was to have a foundation for an easy-to-use framework, which is not restricted to usage only when having a strong programming background but is also application oriented with regards to the general biology community.

V.2.1 Preprocessing

In order to obtain a proper correlation result the data naturally has to be prepared appropriately first. This is necessary since we are not solely interested in the called peak islands where background subtraction has already happened but we want to investigate all possible regions where read calls have been made – also those with insignificant peaks for which the SICER routine does not provide normalized results. This is necessary since a peak might occur in one condition but not in another one. Still we are in need of the reads from the non-peak condition in order to appropriately correlate anything. Apart from the aforementioned normalization w.r.t. library size we have to subtract the read background from the respective control files. In the following we also make the assumption that replicates will be treated independently from each other in order to stabilize the correlation fit as already mentioned during the optimization procedure. This means that two replicates with slightly different gene expression as well as histone modification values are preferred to a mean estimation, hence obtaining a bigger sample, being beneficial for significant Pearson correlations. For a genome-wide correlation analysis we will only investigate data points that exhibit a grading in Tbet dose since Th2 control as well as naïve conditions in many cases exhibit heavy outliers and therefore turn out not to be robustly comparable for a range of instances not being part of the learning sample. For data preparation we scan the binned genome with a window of 200 bp for all replicates and all conditions obtaining a discrete grid on which we can perform possible correlations. This is implemented in a script called `histmodsegmentation.R`.

V.2.2 Input

The algorithm has to be supplied with a list of ENSEMBL transcript IDs. The respective transcripts are then checked w.r.t. their genomic locations from which corresponding topologically associating domains from published data [89] are determined. These TADs are used as a prior for determining enhancer-gene associations using the parametrized correlation measure. The underlying TAD data in mice has been obtained via Hi-C experiments in embryonic stem cells with a 40 kb bin resolution⁵. One of

⁵See mouse embryonic stem cell topological domains on <http://chromosome.sdsc.edu/mouse/hi-c/download.html>

the main findings from [89] being widely acknowledged is that the boundaries of the respective TADs are to a large extent cell-type-invariant. Hence we assume TADs on large scales to hold also true for our experimental data sets. Furthermore it has been found that enhancer-promoter interactions in general do not cross TAD boundaries, hence we restrict correlations solely to these domains (see e.g. [240, 263, 275]). Additionally we supply the respective parametrized weighting of the different histone modifications as specified by the optimization of the multivariate correlation measure in Eq.V.11. In general the algorithm can be supplied with any arbitrary parametrized linear or non-linear combination of any arbitrary number of histone marks. This has to be specified beforehand. Hence the whole procedure is highly generalizable.

To improve computation time for the segmented modification bins and their parametrized measure the data sets are separated into smaller subsets and parallelized and subsequently assigned with the respective parametrization. This is implemented in `Parametrization.R`. We note that the whole correlation routine which will be subsequently performed is based on the aforementioned HMM chromatin state segmentation, s.t. different chromatin states can be observed independently from each other. One of the main reasons for this is the reduction of computation time of the respective routines as we can already make some preselection w.r.t. the chromatin state patterns we are interested in, e.g. enhancers. The main correlation routine, which is explained in more detail below, hence uses as an additional input the HMM chromatin states that are investigated in the correlation analysis as well as the cell condition in which these states are supposed to occur for subsequent analysis. As a most general case we might be interested in a genome-wide analysis of all states and all gene transcripts available for some species, which leads to a fine-grained genome-wide correlation pattern. On the other hand we might for example only be interested in enhancer states or in states with repressive marks that are exhibited in some particular condition. This then defines the segments that are considered for correlation. Not only can we consider the unification of chromatin states $S_i \cup S_j$ for $i \neq j$ over different conditions but we can also specify only the overlap $S_i \cap S_j$ for $i \neq j$ as a minimal consensus set for some chromatin state. For example we might be interested in all poised or active enhancer states (e.g. states 1, 2, 3, 10) that only occur in Th2 or hybrid Tbet^{+/+}Th1/2 cells. In order to see this more clearly we again turn to the respective HMM segmentation around *Ifn γ* in Fig.IV.6. We find that many enhancer states overlap over several conditions as can be e.g. observed at the CNS-34 enhancer. Also some enhancer states switch to a different chromatin state and some instances extend over a wider range in a particular condition. All of these regions would qualify as being relevant candidates for positive gene regulation. Later on this will lead to cell type-specific definitions for the respective chromatin state correlations.

We already note here that the definition of this predefined segmentation also affects the correlation results to some extent. Additional inputs for the correlation algorithm are the resolution under which enhancer states are correlated, which is related to this issue. This basically means that a chromatin state of interest, which has a certain width on the DNA, is being partitioned according to this value in order to obtain a reasonable fine-graining in correlation values. Optionally one can specify a statistical threshold that merges statistically similar neighbouring correlation elements at the end of the analysis. In addition to this one can also indicate if a computational transcription factor binding site search should be performed or not. The general ex-

ecution command of the correlation algorithm with options to be specified reads

```

1 HistoneCorrelation.sh <gene_transcript> <histone_mod_file_ending> [options]
2
3 #with [options]:
4
5 -r [1=unification of regions, 2=overlap of regions]
6 -s #save data for later analysis
7 -p #preload saved data
8 -f #make correlation figures for each significant element
9 -c #input of multiple transcripts via specified file: -c <multiple transcripts file>
10 -d #specify domains for transcripts individually: -d availableIDs.txt
11 -m #merge segments; optionally specify -m <mean> <sd> with numerical arguments,
    otherwise default is used
12 -t #TFBS analysis; usage: -t <motif file>
13 -g #do gene expression model if only one transcript is correlated: -g [1=linear model,
    2=linear exponential model, 3=linear logistic model]
14 --partial #calculate partial correlations for a supplied list of transcripts at a
    certain locus via --partial <transcript file>
15 --hmmstates #customize choice used for correlation by vector listing like: --hmmstates
    [X Y Z] with X,Y,Z, etc. being HMM state numbers
16 --significance #set significance cutoff for correlation: --significance <corr> <pvalue>
17 --resolution #specify upper bound resolution numerically
18 --verbose #detailed information on the computation process

```

We will explain some of the above stated algorithmic implementations in more detail now.

V.2.3 Correlation algorithm

The full correlation algorithm⁶ is executed by the bash-script `HistoneCorrelation.sh` that incorporates the pre-processing of the ChIP-Seq data as described above. This is then followed by the actual correlation procedure which is implemented in `Correlation.R`. The functions used by the algorithm are themselves implemented in `FuncPipeline.R`. We will discuss some important main aspects of the algorithm which are highly important for understanding the subsequent results.

Segmentation of chromatin state elements

As the HMM result of classified chromatin states is used as a prior for correlating epigenetic regions of interest with the expression of a certain gene transcript we find a huge amount of instances of chromatin states spanning several kb. This essentially means that in such a case one integrates over the full modification count within these states, hence one only obtains an effective mean value approximation for correlation. This implies coarse-graining up to a level where significant correlations of some segments might vanish completely since the mean values over sufficiently large regions might become approximately equal for certain cell conditions. The minimal size of a chromatin fragment always is 200 bp as specified by the HMM. Yet for segmentation of larger HMM states we have to specify an upper bound in chromatin state resolution.

⁶consisting of roughly 2000 lines of code

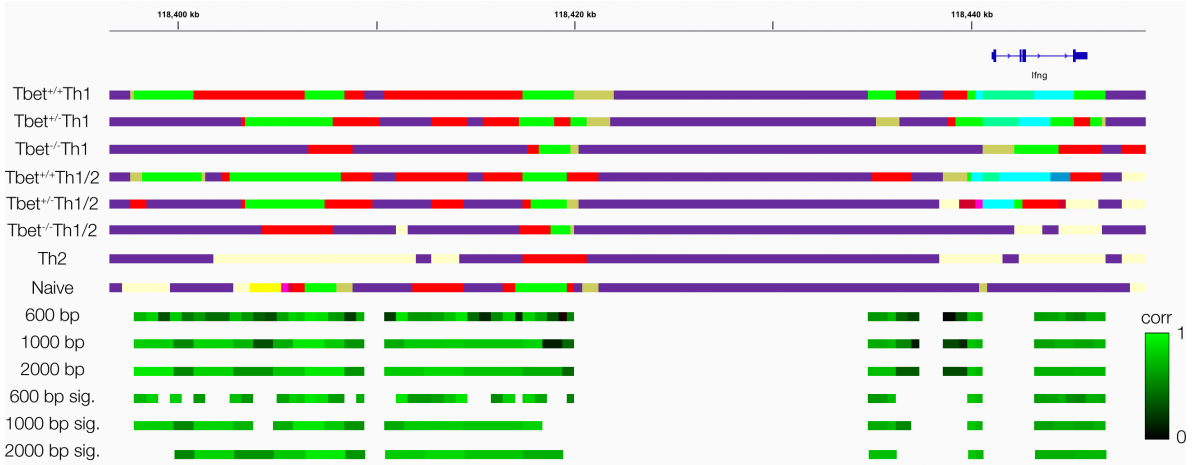


Figure V.4: Correlations of enhancer state elements at the *Ifng* locus for different upper bound resolutions in different shades of green corresponding to different correlation coefficients as indicated in the legend. The difference w.r.t. significant correlations is shown as well.

In Fig. V.4 we exemplify the differences in upper bounds on the length of chromatin elements. To this end we depict part of the *Ifng* locus where only enhancer states occurring in the Tbet^{+/+}Th1 condition were considered for correlation using the correlation measure from Eq. V.11. Beneath the HMM state lines we show all correlations for upper bound resolutions of initial chromatin state sizes of 600 bp, 1000 bp, and 2000 kb in shades of green where bright green denotes high correlation and dark green denotes low correlation. We clearly see that the width of the different new segments gets larger with increasing upper bound resolution. In addition we also observe changes in correlation values for different resolutions. For the upper bound resolution of 600 bp we still observe a very fine grained correlation substructure with distinct high correlation peaks while higher resolutions yield broader highly correlating structures.

We can furthermore introduce a significance level on correlations that has to be met in order to qualify as a significant correlation. Let us assume a correlation $\mathcal{R} > 0.5$ and a corresponding p -value of $p < 0.1$. We see that in the case of a resolution of a maximum of 600 bp significantly more elements vanish while for a higher upper bound resolution more chromatin state segments pass the significance threshold.

For further downstream analysis we will employ a combination of different upper-bound resolutions in order to achieve a trade-off between fine-graining and obtaining a large amount of statistically significant chromatin state elements. Hence for actual gene locus analyses we will merge the 600 bp and 2000 bp results.

The computation of the respective histone modifications in each new segment is performed as follows: We find that in some cases the value within a certain histone modification bin m_b as specified in the input data might be split or that several formerly smaller input data bins might lie subsequently within a larger chromatin state segment s as defined by the upper resolution bound. In this case these values are summed by their respectively weighted overlapping length. We thus obtain the respective parametrized histone modification value m_s within a newly determined

chromatin state segment s from its former bin values m_{b_i}

$$m_s = \sum_{i:|b_i \cap s| \neq 0} \frac{|b_i \cap s|}{b_i} \cdot m_{b_i}, \quad (\text{V.12})$$

where the vertical bars denote the length of the segment.

We additionally note that in Fig.V.4 we already employed an additional feature of the algorithm namely the merging of sufficiently similar neighbouring elements, which we will introduce now.

Merging of statistically similar elements

What is in general meant by the determination of “statistically similar elements” is a quantitative procedure to check whether neighbouring elements should be merged due to their similarity w.r.t. their histone modification profiles over different conditions c . This naturally leads to a reduction in the number of individual fragments and additionally reduces the number of very narrow correlation segments with potentially low significance. It is important to note that we only want to achieve a fine-graining of individual segments that does not unnecessarily break down all elements to a level where the experimental data is not accurate enough to yield statistically meaningful results and fluctuations in read numbers might get too large. We hence merge on basis of similarities. This is done by determining the mean and the standard deviation of the difference of the parametrized histone modification measure over all conditions between neighbouring elements. In order to perform a meaningful comparison of neighbouring fragments we first of all convert the histone modification value within a segment of length $|s|$ to densities by

$$d_s = \frac{m_s \cdot 100}{|s|}. \quad (\text{V.13})$$

For every segmented element with running index i we now determine the tuple

$$\{\overline{\Delta d_s}, \sigma_{\Delta d_s}\}_{ij} \quad (\text{V.14})$$

being the mean and the standard deviation of the difference of the parametrized histone modification densities for neighbouring segments s_i and s_j of all conditions c . The mean e.g. is obtained via

$$\overline{\Delta d_{s_{ij}}} = \frac{\sum_c (d_{s_{c,i}} - d_{s_{c,j}})}{c}. \quad (\text{V.15})$$

In more detail segment i is the left neighbour of segment j .

We can now define boundaries within which neighbouring fragments have to lie in order to be merged. This can be specified as an additional input to the correlation algorithm as well. To this end we choose an exemplary gene locus to learn the mean and standard deviation distributions. In our case we again chose the *Ifn γ* locus to determine the learning sample distribution. From this we determine the inner 0.5-quantiles of the two distributions representing the median. Thus we assume similarity of a total of half of the neighbouring segments within the training sample. From

this we obtain for our correlation workflow based on the correlated parametrized histone modification segments the following boundaries on the similarity of neighbouring elements ij

$$\{\overline{\Delta d_s} \approx 0.025, \sigma_{\Delta d_s} \approx 0.05\}_{ij} \quad (\text{V.16})$$

Individual segments lying within these boundaries are iteratively merged. As soon as this merging condition is met the respective elements are being merged according to

$$d_s^* = \frac{d_{s_i} \cdot |s_i| + d_{s_{i-1}} \cdot |s_{i-1}|}{|s_i| + |s_{i-1}|} \quad (\text{V.17})$$

Here again the vertical bars denote lengths of segments s_i . In effect we weighted each neighbour with its respective length and normalized the whole density in the end to its combined final length. This is repeated with the adjacent neighbours of the subsequently resulting element d_s^* until the merging condition is not met anymore. This code is furthermore shown in appendix D.

If we perform the splitting and merging procedures for different HMM states we obviously obtain different resulting segments. We show one exemplary result where we perform correlations for all possible HMM states at the *Ifn γ* locus as well as for enhancer states only. The result is shown in Fig.B.7. For several correlating segments we now find differences in their spatial extension as well as in their respective correlation and p -values. This becomes even clearer when only observing all correlating segments fulfilling the significance threshold. This is obviously due to the fact that the merging for the correlation of all HMM states also includes HMM states that are next to enhancer states depending on the statistical merging condition also leading to an absorption of these segments. Hence to some extent the recovery of statistically significant elements is also dependent on the considered HMM states⁷.

Actual correlation and output

The actual Pearson correlation is subsequently performed via the parametrized histone modification value within each iterated merged⁸ segment m_s^* , which is straightforwardly obtained via Eq.V.13. Each segment is in addition to its Pearson correlation value \mathcal{R} and its p -value colour-coded in shades of green for positive correlation values and shades of red for negative correlations. Bright colours indicate high absolute values while dark colours indicate low values. We obtain output files including all segments as specified by the chromatin state input and the respective conditions and its results as well as only those segments that satisfy the cut-off of $\mathcal{R} > 0.5$ and $p < 0.1$. Furthermore we obtain the workspace with all segments and parametrized histone modification values as well as normalized gene transcript expression values for the considered regions in the workspace file `CorrelationData.RData` for further postprocessing. The commands for the correlations are given in appendix D.

⁷Since we can also determine if a unification or an intersection of states over certain conditions has to be performed the resulting segments m_s^* also depend on this specification.

⁸The merging itself is optional, yet always performed in our subsequent analysis.

V.2.4 Summary

The general workflow of the correlation algorithm is summarized in Fig.V.5.

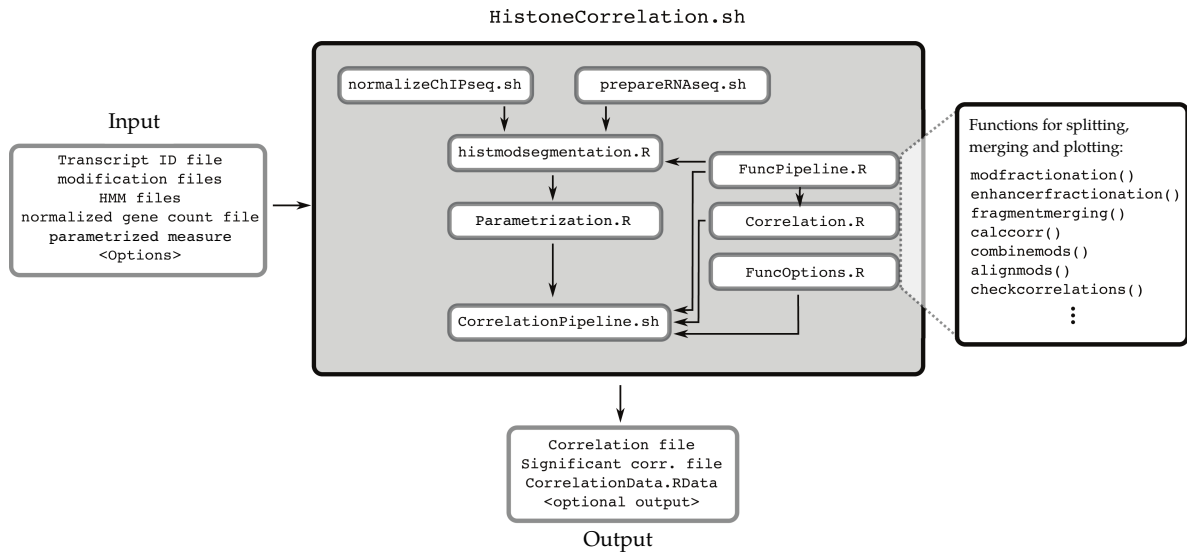


Figure V.5: Algorithmic flowchart of the correlation analysis. After specifying the input options as well as a transcript file `HistoneCorrelation.sh` is executed performing all needed normalization and data alignment procedures as well as performing the parametrization w.r.t. the specified histone modification measure. In the actual correlation procedure, `CorrelationPipeline.sh`, the above specified chromatin state element segmentation with subsequent statistical merging and calculation of significant correlations is performed.

The way we apply the correlation algorithm in general from now on can be hence summarized as follows:

At the beginning the data has to be preprocessed, i.e. the ChIP-Seq data is normalized and the data frames are prepared such that corresponding read bins are identified between replicates and over conditions. As input conditions for correlation we only use the Tbet-dose graded conditions for better comparability, hence we correlate in total six conditions with two replicates each. Then we specify the chromatin states that should be used for correlation based on the HMM. Alternatively we can always use the full set of states, hence performing correlation for all regions. In table V.1 we illustrate which HMM states are considered for which functional annotation within our correlation algorithm. Then we specify under which conditions these states have to occur and if the intersection or union over several conditions has to be performed⁹. In our case we are also interested in which condition a certain state preferentially occurs depending on a certain cell-specific transcript. This determines the further analysis of a certain chromatin feature like e.g. an enhancer. These specifications are given in table V.1 as well and will be applied in the following examples.

Furthermore a list of gene transcripts has to be specified beforehand which automatically fixes the correlation domain for each transcript individually via experimentally validated TADs (see [89]). Correlations can hence only be performed within these domains although arbitrary regions can be manually supplied via a separate file¹⁰ using a unique grep identifier for the transcript ID.

As we already saw the upper bounds on the resolution of a correlating segment has to be specified as well. Since we found above that the choice of this upper bound

⁹In the following we will always apply the union.

¹⁰`availableIDs.txt`

	Th1 transcript	Th2 transcript
activating feature	1, 3, 10 Tbet ^{+/+} Th1	1, 3, 10 Th2, Tbet ^{+/+} Th1/2
inhibiting feature	7, 8, 12, 13 Th2, Tbet ^{+/+} Th1/2	7, 8, 12, 13 Tbet ^{+/+} Th1

Table V.1: Intrinsic logic for the correlation algorithm to specify a notable activating or inhibiting chromatin feature like an enhancer or a repressive state for correlation with a Th1 or a Th2 transcript from the HMM results. For activation we include poised and active enhancers as well as active enhancers in promoter regions while for inhibiting features we consider repressed as well as bivalent states. Most importantly the conditions in which these states have to occur for correlation are specified with inhibiting states always occurring in the opposing cell condition of the respective transcript specificity. By default the union of these conditions is considered.

can influence the similarity concerning the merging of neighbouring elements. This can result in different segment widths and also in smaller segments experiencing drops in their significance level. To account for this we consider two different upper bounds on resolution, namely 600 bp and 2000 bp, and merge the respective significantly correlating elements afterwards.

The correlation procedure is then performed by the combination of the segmentation via the specified resolution and the subsequent iterative merging of similar neighbouring segments. The conditions on the merging procedure can as well be supplied optionally with a customized mean value and standard deviation. The correlation output yields quantitative statistical results for each final segment m_s^* as well as a colour-coding for subsequent inspection. In the following analyses we only consider chromatin segments that fulfill the above specified significance cut-offs.

For further details on the algorithm including computational dependencies on previously released packages as well as their versions see appendix D.

V.3 Results

In the following we want to exemplify the application of the above described method at some prominent genomic locations. To this end we apply the correlation algorithm as described above to TADs containing only one gene as well as to some containing multiple genes. We will shortly see that in the latter case complications arise w.r.t. unambiguously mapping epigenetic features to genes. In addition we are interested in narrow as well as broad chromatin states in order to find out which regions are preferentially co-regulated with gene expression. Especially two genetic loci are of utmost interest for Th1 and Th2 cells as they are thought of acting as their respective master regulators, namely *Tbx21*, producing the TF generally known as T-bet, and *Gata3*. But first of all we focus again on the well-annotated *Ifn γ* gene locus in order to validate the computational method.

V.3.1 *Ifn* γ

The *Ifn* γ locus, which has also been utilized for the training of the multivariate correlation measure¹¹, is probably one of the most obvious examples when it comes to epigenetic annotation in mouse T-helper cells (see e.g. [20]). The result of the correlation procedure is depicted in Fig.V.6.

We again show the HMM results for all experimental conditions but include now all significant correlations that fulfill $\mathcal{R} > 0.5$. Since we are mainly interested in positive regulation of gene transcription we focus on enhancer regions and since *Ifn* γ is a notable Th1 gene the respective conditions for enhancer states in Th1 cells as listed in table V.1 have to be met. Additionally we included independently published ChIP-Seq data in Th1 and Th2 cells for STAT1 (Th1), STAT4 (Th1), STAT6 (Th2), Tbet (Th1) and Gata3 (Th1 and Th2) [233, 314, 334, 335]. In the following analyses we will make intense use of these data sets in order to gain additional information on the chromatin state properties.

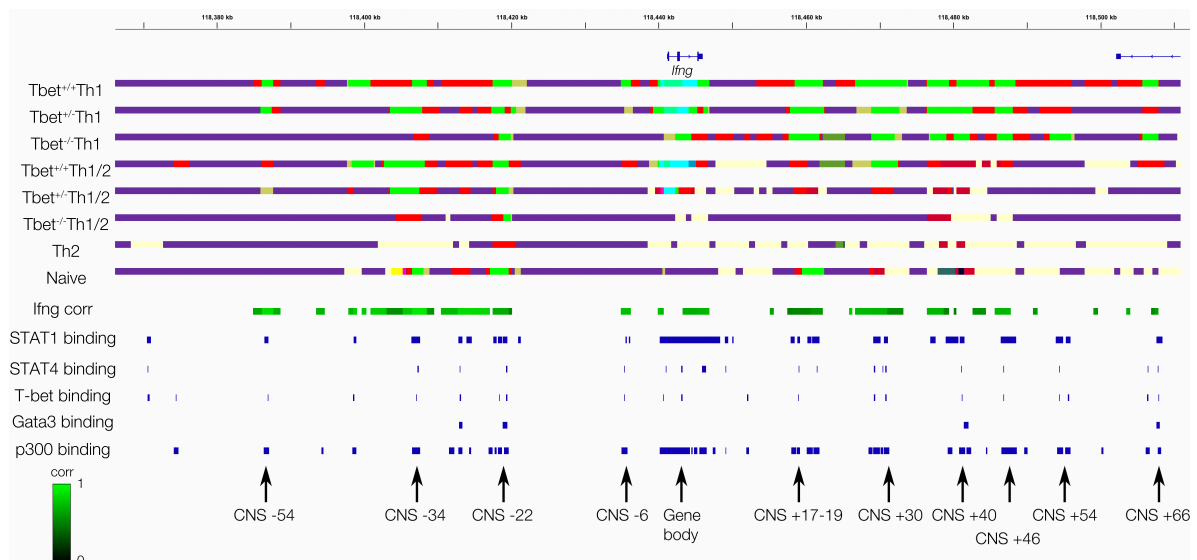


Figure V.6: Significant correlations of enhancer state elements at the *Ifn* γ locus for combined 600 bp and 2000 bp resolutions. We also show TF ChIP-Seq binding data from independent publications [233, 314, 334, 335] as well as binding of p300 in Th1 cells and the locations of the aforementioned CNS sites.

We can see right away that there is an obvious colour-grading in many of the neighbouring significant correlations. This can be notably seen at e.g. CNS-54 or CNS-34. We hence obtain a peak-like structure in the correlation results themselves, which leads to the conclusion that there are parts of connected enhancer structures that are more co-regulated with gene expression than others. Especially in the case of CNS-54 (see Fig.V.7 for a close-up) we find that the segment overlapping with the

¹¹Testing this locus again is not a circular argument since the correlation of some elements can result non-trivially in some value below or above the aforementioned significance threshold, which is partially due to the intricate splitting and merging procedures for each element. Hence trivial re-occurrence of a certain enhancer is not guaranteed as we will see in due course. Additionally the segmentation of the correlation algorithm is different from that of the learning sample. The latter rather observes one single extended segment for each instance.

active enhancer state in the $Tbet^{+/+}$ Th1 condition is highly regulated with the expression of $Ifn\gamma$ ($\mathcal{R} \approx 0.87$, $p = 0.0003$) while its flanking segments are only of the order of $\mathcal{R} \approx 0.65$.

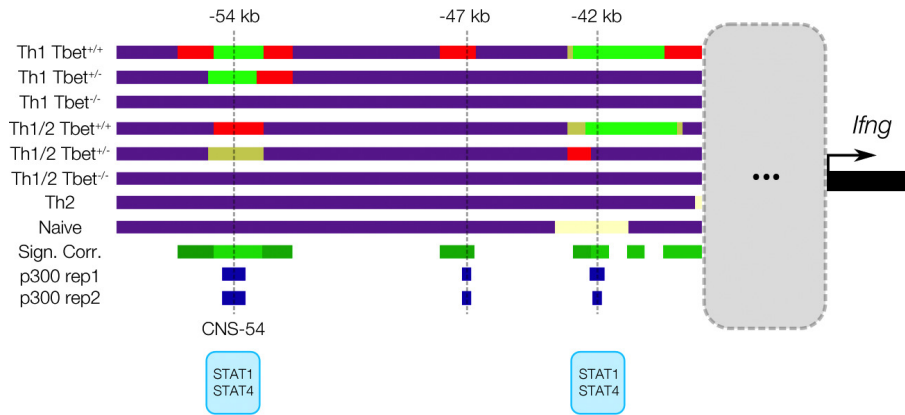


Figure V.7: Close-up of an upstream region of $Ifn\gamma$ showing the validation of the CNS-54 enhancer element as well as the prediction of new enhancer elements at -47 kb and -42 kb.

We see that we recover all CNS enhancer sites apart from CNS+54 which did not pass the significance requirements. This already represents a striking accordance with experimental observations. We also find that there appear to be a lower number of newly inferred correlating segments in comparison with validated elements. This is also exemplified in Fig. V.7. Apart from the already known site CNS-54 having an intersection with a p300 binding site in Th1 cell conditions we find two more p300 binding sites downstream of that element, which are located at -47 kb upstream of the TSS. From the HMM we can also infer that there is at least a poised enhancer state in $Tbet^{+/+}$ Th1 cells that vanishes in all other conditions under consideration. In this case we already find significant positive correlation ($\mathcal{R} \approx 0.70$) with $Ifn\gamma$ expression. In contrast to studies, which only screen enhancers for H3K27ac opposed to a simultaneous occurrence of H3K4me1 as a substitute for p300 binding site occurrence (see e.g. [154]), we find that even poised enhancers correlate in some cases significantly with gene expression. This does not contradict our previous claim that an active enhancer might be an interesting prior for positive regulation of gene expression. Yet it tells us that if the activity mark might be missing due not fulfilling statistical tests w.r.t. peak calling and hence the absence of an active chromatin state these regions might yet be of interest w.r.t. positive regulation. On the other hand we note that w.r.t. the ChIP-Seq TF binding data at hand we cannot confirm significant binding of any of these TFs in that regio which would at least call for more detailed screening or the need to employ binding information of different TFs not considered here in order to validate this candidate for an enhancer site. For later analysis such elements will be only considered if they exhibit ChIP-Seq binding from the respective data sets.

More apparently the element at -42 kb is an example of a prediction of a new-found active enhancer element. We find high significant correlation overlapping with p300 binding sites as well as with STAT1 and STAT4 binding, which is in accordance with the case of the already known element CNS-54. As before we note the dose dependence on $Tbet$ concerning the activity of this enhancer state and the additional

occurrence of the repressive state in the naïve cell condition, yet again undermining the importance of the repressive mark H3K27me3 for correlation. Actually this effect can be observed at many significantly and highly correlating enhancer segments around *Ifn γ* . Of special note for this phenomenon are e.g. the CNS elements at -34 kb but also all CNS elements downstream of the gene as well as the gene body itself where we also correlated the enhancer/promoter signature of chromatin state 3. All of these examples exhibit an additional heavy dependence on H3K27me3 as could be already seen in Fig. V.3 since the CNS sites were included in the enhancer learning sample. It is exactly these segments including H3K27me3 which are now among the regions that significantly correlate most often according to our implementation.

We note that because of the intricate computation based on statistical significance features, spanning from peak calling to the HMM segmentation and to the segmentation and merging procedure employed in the correlation algorithm and finally to the limits imposed on the correlation significance, if we find gaps in between significantly correlating segments that these gaps are significant as well and mark a distinct separation from neighbouring elements. In some cases as e.g. in the flanking regions around CNS-34 the graded peak-like correlation distribution can be seen as a large connected enhancer element with different contributions to gene regulation. On the other hand there are distinctions like in-between CNS-34 and the neighbouring newly inferred segment at -42 kb which significantly separate these two segments. Having stated this we find three new enhancer elements, i.e. at -42 kb, +36 kb and +42 kb, that are labelled by an active enhancer state in the *Tbet*^{+/+} Th1 conditions, exhibit p300 binding in this condition and in two of three cases also contain ChIP-seq bindings.

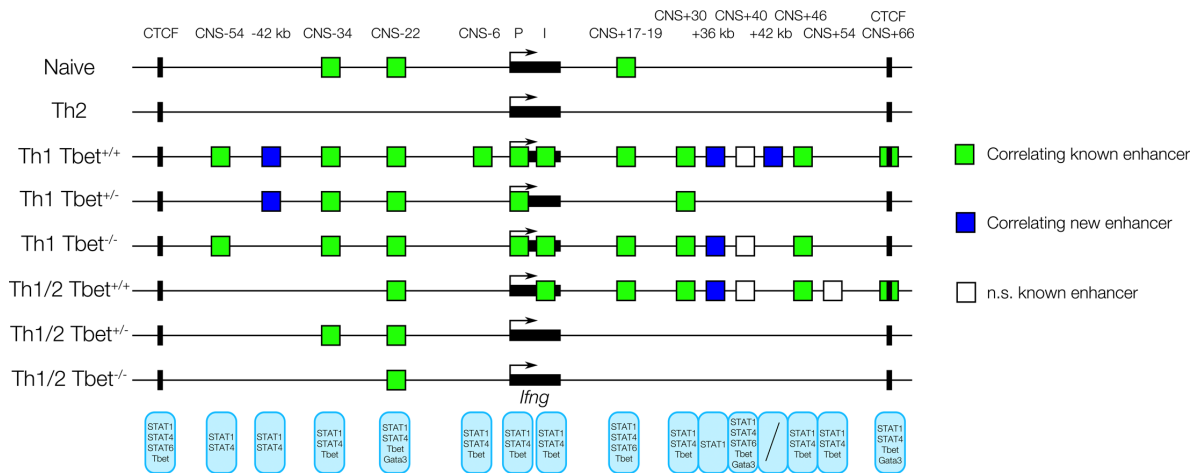


Figure V.8: Schematic depiction of significantly correlating connected enhancer segments at the *Ifn γ* locus including its regulatory enhancer activity logic. Additionally TF binding occurrences from ChIP-Seq data are noted for each enhancer instance. Every square depicts an active enhancer in the HMM scheme. In green we see instances which are significantly correlated and already known enhancers, in blue we depict significantly correlated but as of yet unknown enhancers and in white we depict known enhancers which could not be significantly reproduced within our algorithmic approach.

Finally in Fig. V.8 we give a summary of the whole *Ifn γ* locus analysis only containing significantly correlating elements and summarizing connected correlation elements without gaps into one entity. Depicted are occurrences of active enhancer

states over all conditions at these locations. Additionally we distinguish between known enhancer locations, newly inferred enhancer locations and those that could not significantly be reproduced within our analysis hence not exhibiting a measurable coregulation effect between the histone modifications and gene expression. We see that after the correlation analysis we are left with different entities w.r.t. changes in enhancer activity. This means that active enhancers are either existent or non-existent in certain cell conditions. This is again observed w.r.t. Tbet dose as well as cytokine dependence. On basis of the HMM we note that there are $16^8 \approx 4.30 \cdot 10^9$ combinatorial possibilities of state combinations for a 16 state model of which only 4538 are realized at significantly correlating enhancer states around notable Th1 and Th2 genes. For a binary categorization as depicted in Fig.V.8 this number can be reduced already considerably to $2^8 = 256$ possible combinations¹². At this point it will suffice to note that the number of realized state combinations reduces even more after applying the parametrized multivariate correlation model. We will quantify later on in how far this is true for the set of Th1 and Th2 genes and what implications come with this finding.

V.3.2 *Tbx21*

The importance of the lineage-specifying transcription factor T-bet and its corresponding gene *Tbx21* has been acknowledged for quite a long time and although doubts on its functionality as a single master regulator for Th1 differentiation have come up in recent years (see e.g. [242]) its importance for the Th1 cell differentiation program is undisputed (see e.g. [167, 179]). Although we will come back to its importance concerning gene regulatory networks in due course we want to elucidate its epigenetic landscape in more detail first. We already have seen the heavy impact Tbet dose can have on enhancer activity with the prominent example of *Ifn γ* not only for general chromatin state changes across conditions but even more so at correlating enhancer locations due to the impact of the different histone modifications. It has been reported in several publications [154, 313, 345] that the *Tbx21* locus is controlled by a so-called super-enhancer a short definition of which can be found in section II.1.2. The occurrence of this super-enhancer as well as collection of other examples is listed in the extensive super-enhancer database dbSUPER for the mouse genome [184].

As can be seen in Fig.IV.5 we indeed find an extended active enhancer state in Th1 cells extending from the *Tbx21* gene body up to the next-nearest upstream gene *Tbkbp1* – in fact even further than the TSS of the latter only shortly interrupted by a short active promoter state segment. We observe that the active enhancer state does not change its activity as prominently as e.g. *Ifn γ* . In this case a complete knock-out of *Tbx21* changes the activity state only in some instances in Th1 cells, while only under Th2 culture conditions this is achieved more distinctly. A nearly complete switch-off of enhancer activity and even the appearance of repressive states can be observed in the Th2 control and in naïve conditions. Yet without some differential measure on the respective histone peaks as it is achieved by our correlation algorithm we cannot disentangle how the epigenetic landscape is regulated w.r.t. gene expression. Especially the constitutive enhancer activity over many conditions is harder to interpret

¹²In chapter VI we will extend this to a ternary classification scheme with $3^8 = 6561$ possible combinations with 670 actual realizations.

from first principles keeping in mind the T-bet dose dependency.

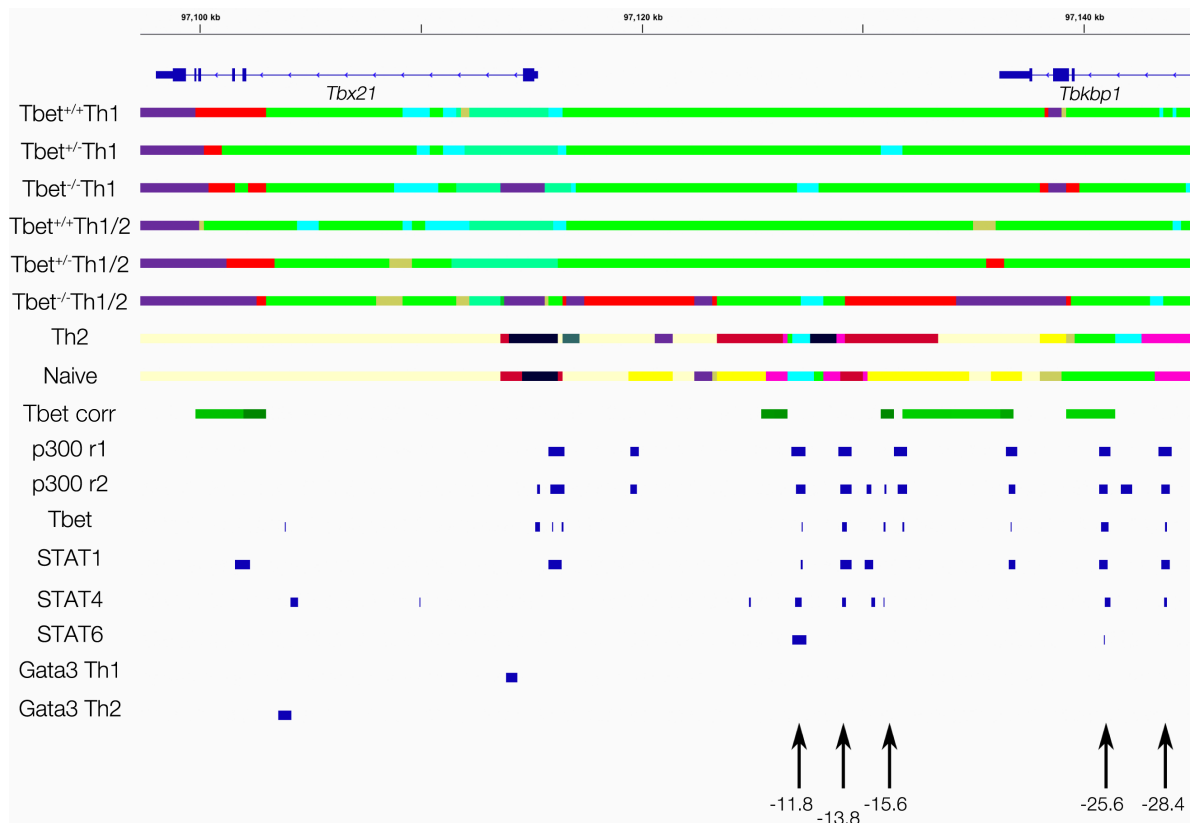


Figure V.9: Annotated *Tbx21* locus according to our algorithmic correlation procedure. Again we only show significant correlations and also include TFs binding sites. The colour coding is also the same as before. Shown are experimentally relevant cis-regulatory sites in units of kb upstream of *Tbx21*. We find that most of these sites are recovered with our method.

After applying the correlation algorithm we are left with only a handful of significantly correlating regions, all of which overlap with p300 binding sites as well as with different transcription factors, yet not incorporating all of the actual binding occurrences. From the corresponding peak file (not shown) we find that only in some small regions the respective combination of histone peaks decreases significantly with decreasing Tbet dose as well as with cytokine dose s.t. only a handful of peaks are regulated differentially. Several of these segments have also been reported in different publications [179, 233] yielding reasonable candidates for cis-regulatory elements in Th1 cells in human cells as well as in mice. In Fig.V.9 we label the interesting candidate sites, some of which are recovered with our analysis.

Although some correlation values are only slightly above our significance threshold it is still remarkable that the analysis is able to recover these sites from multiple histone mark traces, which in this particular case are constitutively regulated over a broad range of different cell conditions, especially w.r.t. Tbet dose itself. Spotting these instances just by eye-inspection or even with the HMM alone is obviously unfeasible and a naïve approach in just observing individual histone modifications also misses most of these features completely while at the same time attributing higher importance to features that might not be quantitatively supported by the underlying histone modification data sets.

We conclude that the putative super-enhancer at *Tbx21*, although fulfilling the general definition, consists to a large extent of segments that are constitutively regulated over varied Tbet dose and cytokine conditions and hence are not regulated in the same way as *Tbx21* expression itself. Yet we find more distinct segments that break up the super-enhancer w.r.t. to their regulatory potential which are co-regulated more significantly with *Tbx21* expression. These regions are already among putative cis-regulatory candidates for the gene. More interestingly the constitutive non-correlating regions might yet form a basis for maintenance or recruitment of respective enhancer function although more experimental evidence for this hypothesis is required. The regulatory enhancer activity pattern over different experimental conditions within the significantly correlating segments is finally depicted in Fig.V.10. Here we still recover certain elements exhibiting constitutive enhancer activity whereas regulatory changes are unveiled by the correlation algorithm, which otherwise could not be readily observed on the HMM basis alone.

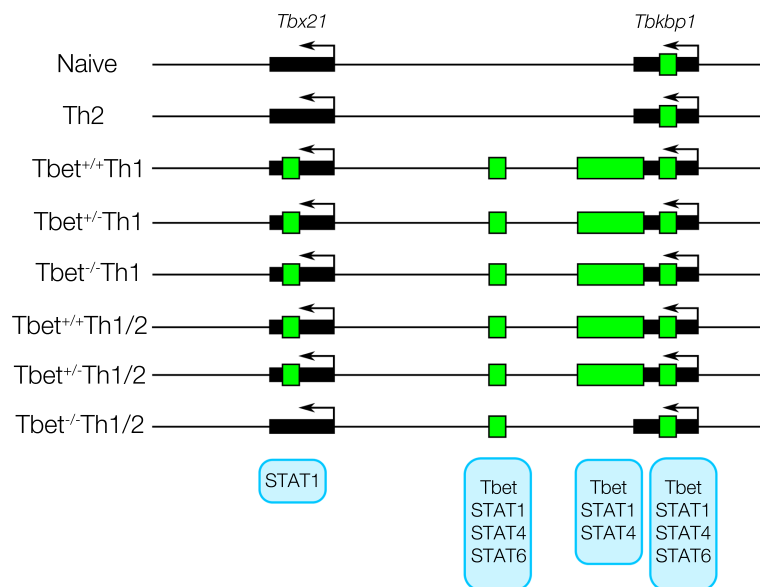


Figure V.10: Schematic depiction of significantly correlating enhancer elements for the *Tbx21* locus with active enhancer states appearing in a certain condition within the correlating segment.

V.3.3 Additional notable gene loci

For completeness we annotate additional important gene loci in Th1 and Th2 cells with our computational method. In appendix B we find schematic depictions of significantly correlating enhancers around *STAT1*, *STAT4*, *STAT6* and *Gata3* as well as around a housekeeping gene like *GAPDH* as a control reference.

For the *Gata3* locus we find multiple significantly correlating enhancer sites including an extended site at 288 – 309 kb downstream of the TSS. In addition just like in the case of *Tbx21* we also find an extended enhancer upstream extending into the next-nearest gene body at -2 kb until -10 kb. Interestingly most of these enhancer

sites also bind either *Gata3* and/or *STAT6*. Apart from one constitutive enhancer at +519 kb all enhancers vanish in wild-type Th1 conditions and some of them in heterozygous *Tbet* Th1 conditions and/or in *Tbet* knock-out Th1 conditions as well. Some of them additionally exhibit *STAT1* and *Tbet* binding indicating a repressor role of the TFs potentially contributing to *Gata3* repression. We will discuss in due course.

In case of *STAT1* and *STAT4* we find many distinct fragmented enhancer segments near the TSS or lying well within the genes itself. We find that most of them bind *STAT1*, *STAT4* as well as *Tbet*, while we also observe occasional *STAT6* and *Gata3* binding in the Th2 conditions where in most cases the respective enhancers are switched off, indicating inhibition. The *STAT6* gene is more ambivalent since we find only one intronic significantly correlating enhancer, which is yet active in Th1 cells as well as several downstream enhancers, which are constitutively active while only their respective histone peak profiles change. At the same time we find that those enhancers bind Th1 specific TFs as well, also indicating a potential important role in inhibition as we will validate later when considering the respective networks.

As a last check we also investigate *GAPDH* as being an example of a T-helper cell housekeeping gene (see [24, 135]). We only find one enhancer site in the promoter region being constitutively active and as well not correlating at all, which confirms our expectation of it having housekeeping functionality even on the level of our epigenetic analysis.

V.3.4 Th2 cytokine locus

Another prominent example in T-helper cells is the so-called Th2 cytokine locus, which has been already studied in quite a number of publications (see e.g. [114, 207, 344]), most of which yet still struggle to distinctly map the epigenetic landscape to specific genes at that locus. The problem here, which is also the case for most TADs including Th1 or Th2 cell-specific genes, is that the locus is crowded, i.e. it contains multiple genes which might or might not be co-regulated. That means that although one can readily identify a large number of enhancer-associated epigenetic states the question is which enhancer maps to which gene exactly and, even if such a one-to-one mapping could be found, are there enhancers that co-regulate several genes at once?

First we want to state the results that are obtained via the HMM and subsequently via the correlation algorithm. We will quickly run into several problems as there will be some ambiguity in the results as can be readily expected for the above stated reasons. In the next section we will address these issues and propose a resolution for many of these ambiguities.

In Fig. V.11 we depict the Th2 cytokine locus including relevant Th2 genes like *Il4*, *Il5*, *Il13*, *Sept8* or *Rad50*¹³ as well as the correlation of enhancer states as defined in table V.1 for Th2 cells with different gene transcripts respectively. We observe that within the cytokine locus as defined by the respective TAD there are some putative enhancers defined by the HMM states that seem to behave in the same way hence seeming to be co-regulated with several gene transcripts, while others are not and

¹³There is actually quite some debate on the functional specificity of *Rad50* for the differentiation into Th1 or Th2 cells (see e.g. [114, 207, 344]).



Figure V.11: Significant correlations of enhancer segments with several notable Th2-specific transcripts within the Th2 cytokine locus on chromosome 11. Most notably we find several co-regulated enhancer regions w.r.t. multiple transcripts at the same time. This is true for positive as well as for negative correlation values.

some segments seem to significantly correlate with some gene transcripts while some others do not make the imposed statistical requirements. Furthermore we observe some enhancers which exhibit negative correlation with some gene transcript yet not with others. This seems quite strange at first. We will elucidate in the following what possibly went wrong here.

First of all we especially look at the gene transcripts that behave in a similar fashion hence clustering them hierarchically. The result of this is shown in Fig. V.12. The colour coding shows a normalization w.r.t. the Z-score of the respective significant correlation values w.r.t. one specific gene transcript. We find that some enhancer regions are co-regulated positively as well as negatively especially for the gene transcripts of *Il4*, *Il5*, *Il13*, *Sept8* and *Rad50*. Hence we would expect co-regulation of expressions of those gene transcripts as well, while we would expect e.g. negative correlation with a gene like *Irf1*. Hence gene expression clustering is shown in Fig. V.13. On basis of clustering the correlation matrix of VST expression values hierarchically we confirm close proximity of the above mentioned transcripts. Hence a considerable amount of co-regulation of the enhancer states can be attributed to the co-regulation of the gene transcripts themselves.

Moreover we still observe differences in correlating a certain enhancer segment with multiple transcripts of the same gene. Examples are *Il4-001* and *Il4-003* where several enhancer states correlate with one transcript yet not significantly with the other. The same holds true for several other cases.

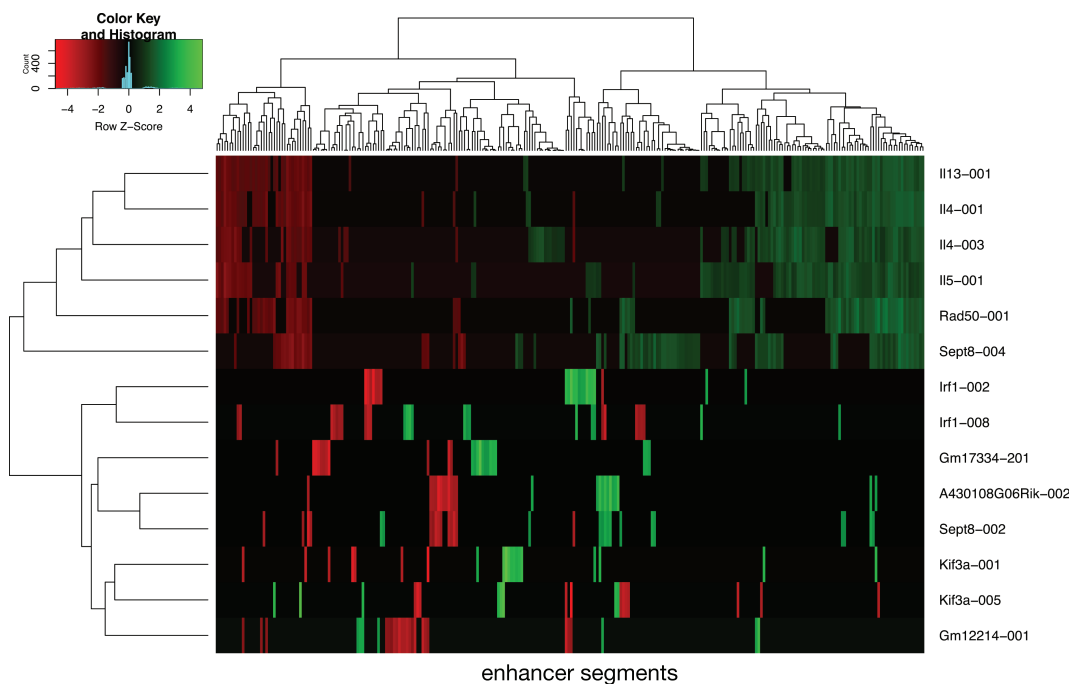


Figure V.12: Hierarchical clustering of significant enhancer segment correlations with gene transcripts at the Th2 cytokine locus. The colour-coding shows the respective row Z-scores.

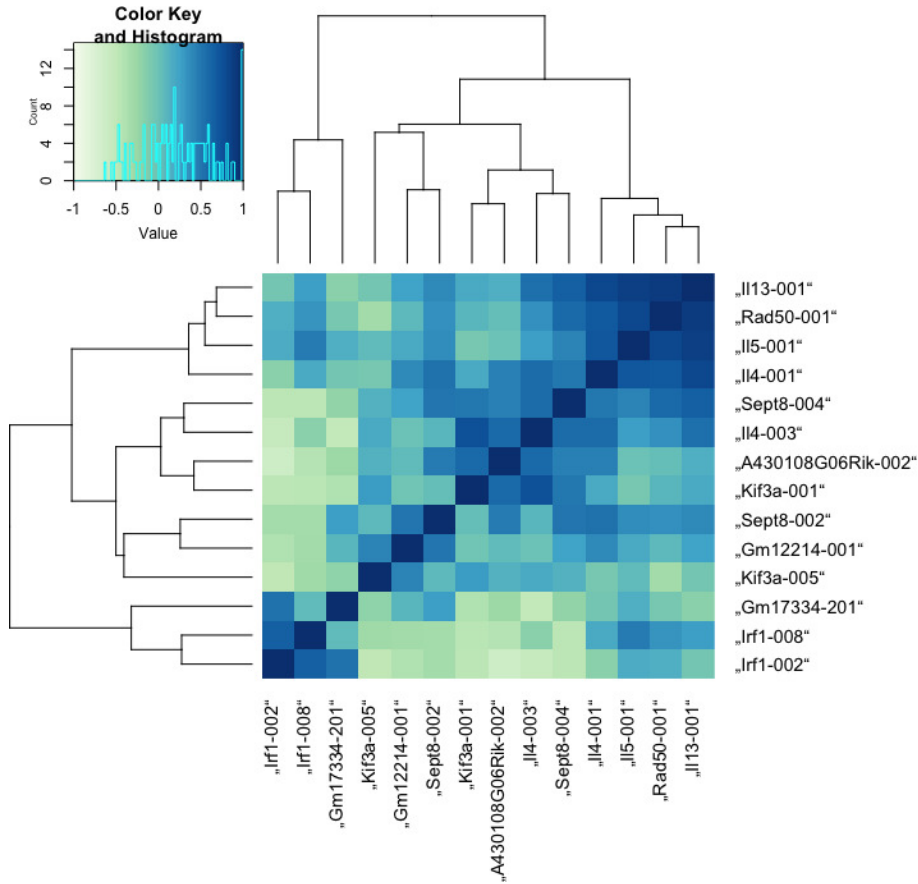


Figure V.13: Clustering of the correlation matrix of VST-normalized expression values of Th2-specific transcripts in the Th2 cytokine cluster.

Also we find some enhancer instances where we observe negative correlation with one transcript while we find positive correlation with another one. One explanation for this occasional occurrence of negative correlations indicates that the considered enhancer is co-regulated with some gene correlating negatively with the expression of the transcript that correlates in turn negatively with the respective enhancer¹⁴. This is at least an explanation for those cases where we find clear positive correlation with another gene transcript. Yet we should nevertheless be skeptical about the occurrence of these negative correlations. At the same time it is questionable if some enhancers are really co-regulated w.r.t. several transcripts or if this is rather due to some intrinsic residual information not taken into account. This will be investigated in more detail now.

V.4 Partial correlations

The question we can ask ourselves is if the co-regulation found in the correlations is “real” in some sense or if this is in fact an artifact that can be removed by refined sta-

¹⁴We note here that because of the definition of our parametrized correlation measure negative correlation of an enhancer cannot serve as an indicator of inhibition since this would mean that in the condition of highest gene expression we should observe a rise in H3K27me3 and a decline in H3K4me1. This would yet not yield an active enhancer state in this condition as was already used as a prior for enhancer detection. Hence this results in a contradiction.

tistical analyses. Furthermore we want to answer the question if there is some way to obtain a one-to-one mapping between enhancers and genes in crowded domains. For this we employ partial correlations with only residuals between some random variables being correlated with another random variable. This is done in order to remove any possible correlations between the initial set of random variables. In our case this would be the set of transcripts that seem to be co-regulated by some set of enhancers. Formally the partial correlation between two random variables A and B having removed the effect by a class of random “control” variables $C \dots N$ with N being the total number of variables is given by Eq.A.2. In our case A is the parametrized histone measure, B is the transcript to be correlated with and the vector $C \dots N$ contains the other co-regulated transcripts for which we correct the correlation itself.

We focus first on the region positively correlating with the transcripts under consideration. We find that after application of partial correlations about half of the correlating segments can be attributed to one specific transcript while for some the statistical analysis is inconclusive and many others even fall below the imposed significance threshold and hence drop out of the analysis. Nevertheless this provides us with a good indicator for a larger enhancer segment to be attributed more clearly to a certain transcript. In Fig.V.14 we show the association of all significantly correlating elements around the interleukin transcripts in the cytokine cluster after applying partial correlations w.r.t. the set of gene transcripts for which the elements are co-regulated¹⁵. We find that actually most of the significantly correlating enhancer segments from the co-up-regulated enhancer cluster after partial correlation are associated with *Il4-003*, while the rest of the co-regulated segments can only be associated with either *Sept8-004*, *Il5-001* or *Rad50-001*. All other transcripts exhibit either a lower ranking with respect to their partial correlation with the respective enhancer segments or appear to be not significant. This is shown for an exemplary enhancer segment in table C.7.

Some correlations yet still remain associated with several transcripts at the same time and hence cannot be uniquely mapped. In general the case of multiple mappings is applied if correlations are similar within a range of $\Delta\mathcal{R} = 0.2$ and if for these cases we obtain $p < 0.1$ and a relative factor between p -values of $p_i/p_j > 3$ for $p_i > p_j$.

Furthermore the significance threshold for partial correlations is of utmost interest in this respect. Let us turn to the co-regulation cluster of negative correlations for the considered gene transcripts. What we find here is that all negative correlations in fact vanish since all of them fall below the correlation threshold, hence being statistically insignificantly correlated, and thus missing the significance threshold. The resulting p -value and partial correlation value distribution for the negative correlation cluster is depicted in Fig.V.15. We also recover the two positively correlated instances as depicted in Fig.V.14.

In conclusion we find that after removing correlation between transcripts themselves and only considering correlations of enhancer elements with the respective residuals a large number of naïvely correlated enhancer segments either vanish or can be finally mapped conclusively to a particular gene. Only in some cases a final distinction cannot be readily made yet the number of candidate transcripts can

¹⁵That is all gene transcripts for which we show the correlation tracks in Fig.V.14.

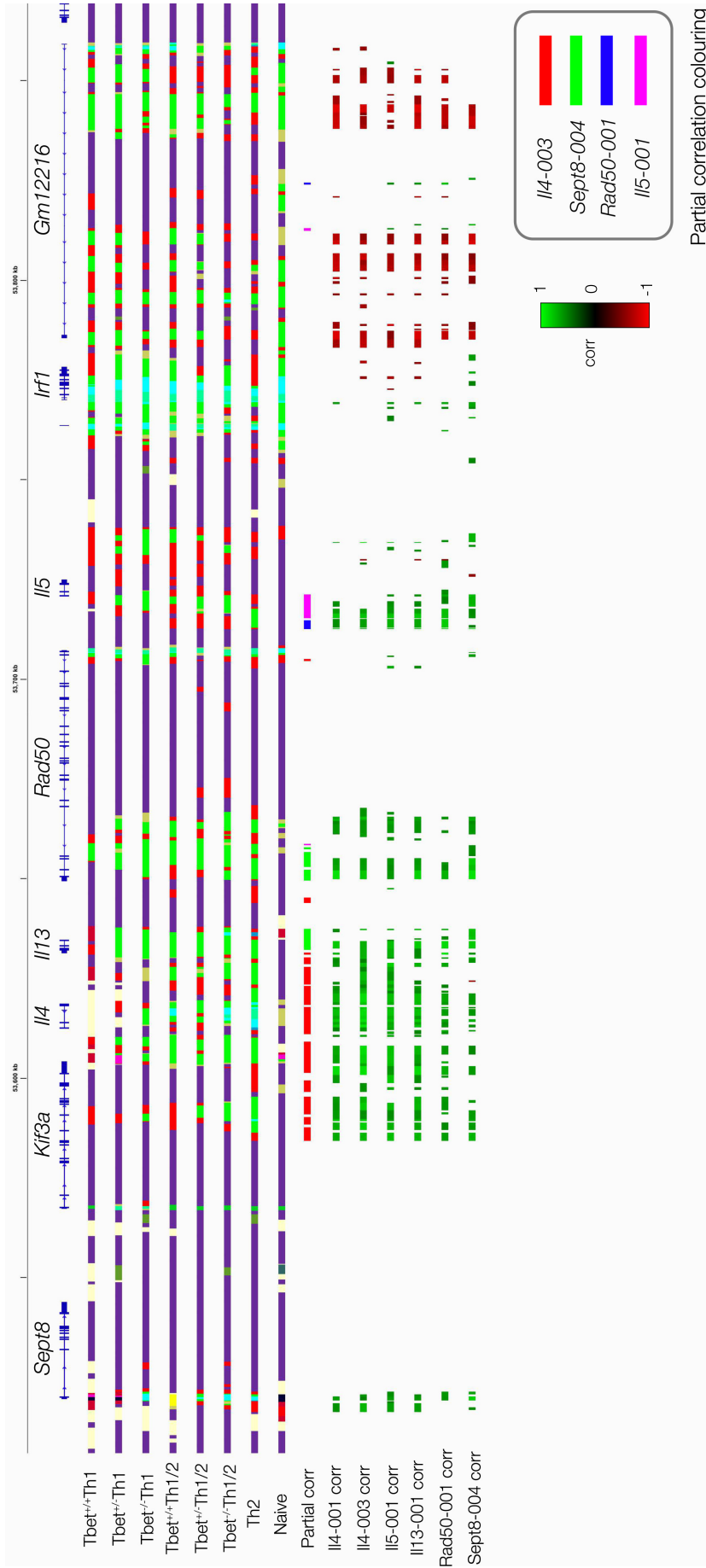


Figure V.14: Significant remaining partial correlations at the Th2 cytokine locus. We colour-coded the respective most significant attribution of each enhancer segment to a certain transcript.

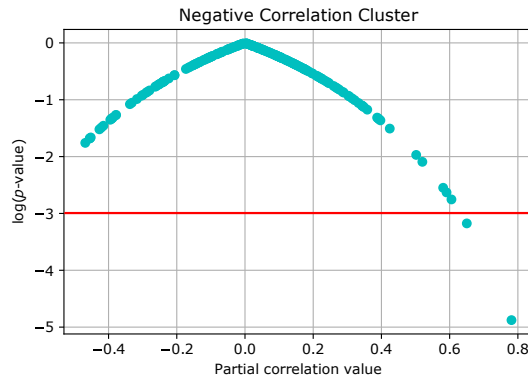


Figure V.15: Phase-space of partial correlations and p -values for all enhancer segments from the negative co-regulated correlation cluster with all the respectively co-regulated gene transcripts. The red line denotes $p = 0.05$.

be significantly reduced. We find that the application of partial correlation removes the significance of negative correlations with certain transcripts in most cases completely and leads to the conclusion that the occurrence of negative correlations mainly seem to be artifacts if no corresponding positively correlating transcript can be found which in turn anti-correlates with the other transcript. Finally we are able to make conclusive predictions for enhancer-gene mappings on TADs in general.

V.5 Inference of inhibition

In the case studies for the correlation algorithm above we have mostly considered enhancer states being primers for positive gene transcript regulation. At significantly correlating enhancer sites we also investigated the potential of finding binding of TFs as observed in respective cell-specific ChIP-Seq data sets. We also find that Th1-specific TFs like Tbet, STAT1 and STAT4 bind preferentially at Th1 enhancers (see definition in table V.1), while STAT6 and Gata3 do the same at Th2 enhancers as can be expected. Additionally TFs can take the function of a repressor by decreasing enhancer activity and hence gene expression (see e.g. [273]).

In order to study inhibition we find ourselves ending up with two different possibilities: either we focus on repressor activity by TF binding at enhancers where depending on the binding context a certain enhancer can be repressed by a certain TF while in another setting it can still be activated by another TF. The alternative method is to focus on repressive chromatin states. Yet here we might run into a stalemate situation since in some cases an enhancer turns to a repressive state. From an enhancer point of view we find activation of some gene in the respective condition yet in the opposing cell differentiation program we find a inhibition with increasing H3K27me3 and decreasing H3K4me1 and H3K27ac, equally decreasing gene expression. How do we decide in this situation which definition applies to the chromatin state element?

Luckily there are other cases in which we do not find an enhancer state at the same position as a repressive state in different conditions . Let us consider for exam-

ple the CTCF binding location upstream of *Ifn γ* which is also delimiting the TAD (see Fig. V.6). Here we find empty chromatin states in all conditions but in Th2 cells where a repressive state appears. Additionally we find Th2-specific TF binding via STAT6 at this location in combination with significant correlation of this particular chromatin segment with *Ifn γ* expression. Since the correlation measure includes H3K27me3 with a negative weight increasing H3K27me3 at the same time means decreasing the parametrized measure. If we obtain positive significant correlation we can hence assume co-regulation with gene expression and we find that the correlation measure is also equally applicable to repressive states. Hence a positively correlating repressive state is in accordance with the inhibition of gene expression. Because of this we assign to a repressive state the task of acting in an inhibitory way on gene expression in contrast to an enhancer state actively regulating gene expression as soon as it fulfills the following conditions¹⁶:

1. A repressive state is found in accordance with table V.1
2. We find TF binding in the cell condition of the repressive state

Up to now we furthermore only considered positive binding of TFs at certain chromatin states, hence TFs fulfilled the role of an activator, i.e. binding in the same condition of a positively correlating enhancer or a positively correlating repressive state. In addition to this we also consider TF binding in the condition of the opposing differentiation program, hence potentially acting as a repressor. For this label to be valid e.g. an active enhancer has to be switched off by TF binding. The same obviously can happen to a repressive state as well which would in turn correspond to an enhancing feature again. This is due to the inhibition of an inhibiting state, which results in activation. More prominently in the ambiguous case of finding an enhancer and a repressive state in the same segment if the only occurring TF binding belongs to the enhancer-specific cell condition, hence to the opposing condition of the repressive state then obviously this segment as a whole is characterized rather as an enhancer \mathcal{E} than a repressive state \mathcal{R} . Hence we obtain the following decision table

	Th1 TF	Th2 TF
\mathcal{E} Th1	\mathcal{A} TF \rightarrow +	\mathcal{R} TF \rightarrow -
\mathcal{E} Th2	\mathcal{R} TF \rightarrow -	\mathcal{A} TF \rightarrow +
\mathcal{R} Th1	\mathcal{A} TF \rightarrow -	\mathcal{R} TF \rightarrow +
\mathcal{R} Th2	\mathcal{R} TF \rightarrow +	\mathcal{A} TF \rightarrow -

where “ \mathcal{R} TF” denotes repressor binding and “ \mathcal{A} TF” an activator binding of a TF within the respective state. We note that the necessary condition for a repressor to exist is that the enhancing or repressive state has to disappear in the condition of the respective TF binding¹⁷. The plus and minus always denotes effective activation or inhibition respectively.

¹⁶We note that we do not explicitly exclude here the simultaneous occurrence of enhancers and repressive states and assume that if TF binding in Th1 as well as Th2 conditions that are characteristic of the respective differentiation programs exist that the meaning indeed is ambiguous and hence depends on the context.

¹⁷To exemplify this we take for example the case of an enhancer in Th1 with an Th2 TF binding. If the enhancer state does not vanish in Th2 we end up with the case of an enhancer in Th2 with a Th2 TF, which is again an activator.

Applying this to a genome-wide analysis of Th1- and Th2-specific gene transcripts as shown in table C.3 we find a total of 159 repressor instances as opposed to a total of 1570 activator instances while we find a total of 5975 significantly correlating enhancer segments vs. a total of 4545 inhibiting state segments.

The respective combination of these enhancing and repressive states with activators and repressors will be of special importance concerning a full account of activating and inhibiting edges within regulatory networks later on.

V.6 Prediction of gene expression

We also implemented a basic *linear model*¹⁸ of gene regulation by additive enhancer modification via our parametrized histone modification measure which automatically selects positively correlating enhancer segments and estimates the linear model parameters. For this simple gene regulation model we assumed the following simple relation:

$$g_i = a \cdot \mathcal{H}_{ij} \cdot c_j + b, \quad (\text{V.18})$$

with free parameters a and b . Here g_i denotes an element of the gene expression vector \vec{g} with i being the respective cell condition¹⁹. Furthermore c_j is an element of a correlation weight vector which assigns a weight to every enhancer segment around the gene under consideration depending on the resulting correlation value, which we inferred earlier. This is an assumption necessary to estimate the enhancer-specific deviations in the free parameter a , hence reducing the parameter space substantially. The length of index j depends now on the number of independently called significantly correlating enhancer segments. The matrix with elements \mathcal{H}_{ij} is a “conditions \times enhancer segments” matrix containing the parametrized modification measure values for each segment.

We again choose *Ifn γ* as the model locus²⁰ obtaining as a χ^2 -fit result

$$a = 0.039 \pm 0.007 \quad b = 10.625 \pm 0.289 \quad \chi_r^2 = 0.735 \quad (\text{V.19})$$

hence 73.5% of the variance can be explained with the model fit. Obviously the parameters have to be re-fitted for each gene transcript.

In order to investigate the predictive power of the parametrization we also fitted only half of the data by predicting gene expression of one replicate with the other replicate. Subsequently the predicted results are fitted against the experimental observations, which is depicted in Fig.V.16.

¹⁸For an extensive review of different models of enhancer-gene regulation see e.g. [64].

¹⁹Again we only investigate the Tbet dose dependent conditions in Th1 and Th1/2 conditions.

²⁰We assume that a gene expression model should be tested at the most well researched and hence reliable locus in Th1 and Th2 cells, which is why *Ifn γ* is again chosen as *pars pro toto*.

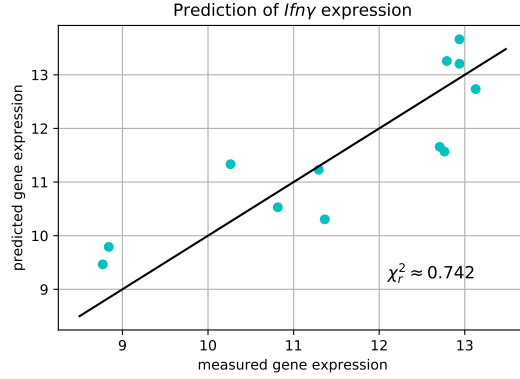


Figure V.16: Linear fitting result of predicted vs. measured *Ifn γ* expression where only one replicate was respectively used for fitting the other. The results are given by the VST normalized gene expression values.

The reduced chi-square statistic in this case yields $\chi_r^2 \approx 0.742$ hence we find good accordance with the actual prediction of the data given the fact that only half of the actual data was used for fitting.

Furthermore we also investigated a *linear-exponential model* capturing potential multiplicative behaviour of neighbouring enhancer elements. This follows the model in [64] allowing for sub- and super-additive relations within TADs. The model in this case reads

$$g_i = e^{a \cdot \mathcal{H}_{ij} \cdot c_j + b} \quad (\text{V.20})$$

being simply the exponential of the linear model. In this case we obtain for the *Ifn γ* locus

$$a = 0.0057 \pm 0.0009 \quad b = 2.1081 \pm 0.0542 \quad \chi_r^2 = 0.79 \quad (\text{V.21})$$

with a slightly higher reduced chi-square value. Yet if we again fit the predicted vs. the measured gene expression results in this case we actually obtain worse accordance with $\chi_r^2 \approx 0.707$.

Extending the analysis to a *linear-logistic model* of the form (see again [64])

$$g_i = \frac{g_{\max}}{1 + e^{-(a \cdot \mathcal{H}_{ij} \cdot c_j + b)}} \quad (\text{V.22})$$

with g_{\max} denoting the maximum expression value, we only obtain $\chi_r^2 \approx 0.714$ for the prediction vs. measurement comparison. Hence we conclude that already the linear model is able to sufficiently describe the measured gene expression data whereas a linear-exponential or linear-logistic model accounting for enhancer-enhancer interactions or asymptotic gene expression behaviour is not able to increase the predictive power for the example of the well-investigated *Ifn γ* locus.

V.7 Discussion & Summary

In this chapter we have primarily investigated the necessities for and consequences of a robust correlation model for enhancer-gene regulation on TADs. We motivated the need for a coupling of histone modifications for correlation. To this end we introduced a general version of a linear parametrizable histone modification measure which we trained on a set of experimentally validated enhancers. We obtained a robust measure which was also confirmed by bootstrapping the training sample and inferred that the contribution of three important histone modifications, i.e. H3K4me1, H3K27ac and H3K27me3, being a classical repressive mark, are of the same order, while the latter contributes negatively. Yet furthermore we also find that the appearance of the repressive mark in the cell conditions with lowest gene expression have a slightly larger impact on correlation than the classic enhancing marks themselves, which is an astonishing finding on its own. The resulting parametrized histone modification measure is now defined by Eq.V.11. In fact we find that the inferred measure increases correlation values as well as their significance notably compared to individually correlating histone modifications with gene expression. We note that a subsequent experimental validation of newly inferred enhancer sites can in turn lead to an improvement on the parametrized measure by extending the learning sample iteratively.

In order to perform meaningful correlations on TADs using the HMM segmentation as a prior for chromatin state candidates we developed a correlation algorithm based on the histone modification measure which appropriately segments the genome into uniquely identifiable units taking into consideration chromatin state overlaps. Furthermore statistically similar segments are merged in comparison to neighbouring elements. The final algorithmic procedure is capable of dealing with an arbitrary correlation measure for individual as well as multiple transcripts on their respective TADs and allows for customizable merging procedures, resolutions in correlation as well as chromatin state selection over an arbitrary number of samples.

Testing the algorithm at the well-investigated *Ifn γ* locus we not only recover a significantly large amount of experimentally validated enhancers passing our correlation significance threshold but we also find a smaller number of hitherto unknown enhancer sites which also exhibit enhancer HMM states as well as p300 and Th1-specific TF binding in condition-specific experimental data sets.

We furthermore investigated additional loci of notable Th1 and Th2 genes and mapped their epigenetic enhancer landscape. We especially found that in some cases only parts of the HMM enhancer states significantly correlate with transcript expression, which can be most prominently observed at the *Tbx21* super-enhancer. There we find that although being switched off or even repressed in naïve and Th2 control conditions the Tbet dose sensitivity as well as the cytokine dependency w.r.t. hybrid cell conditions is not reflected in the HMM landscape but rather in the peak structure at the super-enhancer. With our computational approach we are able to recover distinct fragments which are co-regulated exceedingly stronger with transcript expression than their immediate environment. This results in breaking up the structure of the super-enhancer concerning gene regulation and leads to the hypothesis that parts of the super-enhancer might rather play an important role in maintaining enhancer activity in the whole region itself concerning recruitment than actually in

regulation.

Turning to crowded TADs with densely located gene promoters like in the case of the Th2 cytokine locus we quickly run into problems of uniquely identifying enhancer-gene mappings via traditional correlation procedures. Additionally we find in this case large clusters of co-regulated enhancer segments, correlating not only positively but also negatively. In order to resolve these issues we investigated partial correlations of co-regulated enhancer segments and found that not only a large part of the investigated segments can be uniquely mapped to distinct transcripts but also the negatively correlated instances completely vanish since after applying partial correlations the correlation values as well as their significances themselves drop under a certain threshold becoming statistically insignificant. It turns out that a large amount of potentially negative correlation values can be removed in this way if no other positively correlating transcript in the vicinity is available on the TAD.

Furthermore we propose the inference of inhibition by investigating locations of repressive states in opposing cell conditions of the respective transcript specificity where additionally TF binding can be found. Moreover TF binding in the opposing cell condition upon removal of either an active enhancer state or an repressive state acts itself as a repressor. Hence depending on activator or repressor binding at a positively correlating enhancer or respectively a repressive state we obtain either effective activation or repression.

We also exploit significantly correlating enhancer segments which are uniquely mapped to a certain transcript for the prediction of gene expression. Although we have to parametrize a certain gene regulation function for every gene separately depending on the respective histone modification content within each enhancer segment we can pre-parametrize the respective function depending on the correlations of each individual segments. This considerably reduces the number of parameters for different models. We investigated a simple linear model as well as a linear exponential and a linear-logistic model around $I_{fn}\gamma$ for reasons of continuity and reliability from which we found that the linear model performs best considering the fact that only two parameters were fitted. We also obtain surprisingly good accordance between predicted and measured gene expression values. We note at this point that obviously more data points as well as a more detailed investigation w.r.t. pre-parametrization could improve the model fits. It would be especially interesting to obtain priors for models which extend the mere usage of correlation values and are able to fit arbitrary gene loci only on basis of their surrounding enhancer activity patterns. If this is possible still remains to be elucidated but could in principle be achieved for a genome-wide Th1 and Th2 transcript investigation w.r.t. their epigenetic landscape. We will investigate regulatory epigenetic patterns in more detail in the following chapter.

CHAPTER VI

Establishing cell-type-specific enhancer classes




We have already seen that there are certain differences between enhancer instances w.r.t. their activity state according to different cell conditions being subject to different cytokine and Tbet dose dependencies. Hence some enhancer becomes inactive in a particular cell condition while another enhancer stays active in the same condition. We saw this particularly clearly for the schematic enhancer activity depictions for e.g. *Ifn γ* in Fig.V.8. In short chromatin states change all the time for various experimental conditions and these changes can be different for every position on the DNA. Furthermore we note that not the full set of combinatorial possibilities of chromatin states according to the HMM is realized for Th1 and Th2 transcript loci of interest but rather a considerably smaller subset of 3322 state combinations at significantly correlating enhancer segments and 1997 at significantly correlating repressive state segments.

There are certain question we can ask at this point. Can we determine patterns of certain Tbet-dose dependent and cytokine dependent enhancers that reoccur to some extent around certain co-regulated genes and can we even predict if some gene is falling into a certain cell-specific co-regulation category by just looking e.g. at the enhancer activity changing pattern around the gene itself? If we would be able to do this we could obtain a probability estimate for a certain gene belonging to a certain differentiation program.

VI.1 Introducing a typology of enhancer states

As already mentioned before the chromatin state combinatorics can be reduced substantially by e.g. turning to a binary classification. The binarization obviously has to be performed w.r.t. some “interesting” reference state or subset of states. This can for example be done by considering if an enhancer state is present in some condition or restrict this further to enhancer state activity. We have argued earlier that of special importance for the definition of enhancer states especially w.r.t. correlation of state segments are not only the classical histone modifications H3K4me1 and H3K27ac but also the repressive mark H3K27me3 occurring in numerous instances in the opposing cell condition. This also implies the occasional presence of repressive states.

Because of this we propose a ternary state classification allowing for three different state conditions:

-  = active enhancer¹
-  = repressive state²
-  = $\neg (\text{blue} \vee \text{red})$

Hence we obtain ternary state classes including enhancer activity as well as repressive states. These classes consist of a state classification for each cell condition hence obtaining a total of $3^8 = 6561$ possible combinations of which only 670 are realized at significant correlations of enhancer states around the previously analyzed set of Th1- and Th2-specific genes. This classification is not restricted to giving information about gene activation but it can be equally applied to gene inhibition as introduced above.

Formally a general definition of a so-called *chromatin state class* will be from now on:

DEFINITION: Chromatin State Class

A general *chromatin state class* (CSC) is defined as a certain combinatorial realization of chromatin states, as e.g. defined by an HMM, over a set of – in our case – different experimental conditions. The full set of CSCs is given by all possible combinations of these chromatin states.

More specific cases of a CSC will in our case include enhancer state classes (ESCs) and repressive state classes (RSCs), which form a subset of all CSCs. An ESC has to fulfill the condition of containing an enhancer state \mathcal{E} in a certain condition of interest while an RSC has to fulfill the condition of containing a repressive state \mathcal{R} in a certain condition of interest (according to table V.1). More clearly stated: a state class represents a combination of states at a certain position for different experimental conditions. Obviously these states have been assigned their respective meaning from the correlation analysis introduced in the previous chapter.

We will investigate the implications of this typology in the following analysis.

VI.2 Inter-class specificity of enhancer types

We want to find out if there is some kind of underlying hierarchy to the above state class typology w.r.t. its importances in distinguishing between two or even more cell types. More precisely this will be performed w.r.t. ESCs and RSCs. Given the case that such a ranking can be found the question is if there is some subset of e.g. ESCs that has more predictive power than others for making that distinction. For this we are also in need of a statistically powerful and robust deduction method in order to reverse-engineer the contributions of a certain state class to a certain cell type – essentially shifting the focus from a prediction to a fitting problem. This is where supervised learning comes into play.

¹HMM state 10

²HMM state 12 or 13

VI.2.1 Method

A well-known and easily interpretable approach for this task is the usage of decision tree based classifiers. This contrasts with Support-Vector-Machines (SVM) or even neural networks that are in general less flexible and slower in learning certain classifications (see e.g. [40, 113, 165]). Additionally decision trees can deal with arbitrary large numbers of classes and are less outlier sensitive than for example ordinary SVM methods [36]. Yet one of the most striking advantages of decision trees apart from easy white-box interpretability is the straightforward identification of the underlying feature importances for classification.

We have already discussed various variants of decision tree methods with differing levels of sophistication in section II.3.1. Common to all of them is the structure of the input data. Although we are generally interested in occurrence patterns of chromatin state classes we choose as a particular example the lower level hierarchy class of enhancers. For our purposes we again select as a learning sample the previously inferred set of Th1- and Th2-specific genes and extract all ESCs, denoted as \mathcal{E}_j in the following, from the subset of significantly correlating segments. In addition to this we weight every single enhancer instance k with the width of the respective correlation segment denoted by $|s_k^*|$. Obviously the maximal number of instances k can differ considerably from transcript to transcript leading to a matrix of segments k_{ij} to be considered for each transcript i and each enhancer state class j . The resulting weighting coefficient will be called w_{jk} . Every ESC is now treated as a feature or predictor variable for a certain set of gene transcripts. As we are dealing with supervised classification every gene itself is assigned to a certain cell-specificity superset. The whole class hierarchy from histone modification peak overlaps to chromatin state classes and gene classes can be summarized as in Fig.VI.1.

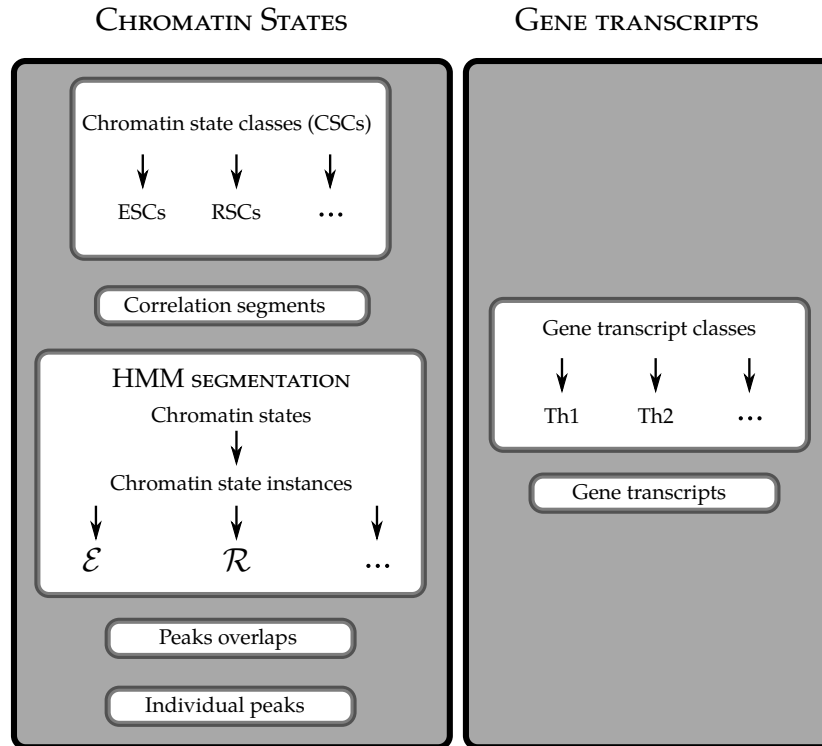


Figure VI.1: Class hierarchy for chromatin states and gene transcripts.

For simplification we start as before with only two gene classes, namely the set of Th1- and Th2-associated gene transcripts. Yet the method is the same for an arbitrary number of gene classes³. Multiple instances k of a certain enhancer class around a specific transcript additionally receive a sum over their individual weights. Hence we obtain the following sets of weighted class features

$$\left\{ \left(\sum_k w_{jk} \right)_{\mathcal{E}_j} \right\}_{\mathcal{G}} \quad (\text{VI.1})$$

around some gene transcript \mathcal{G} belonging to gene class \mathcal{C} . Finally we obtain weighted transcript-feature matrix elements \mathcal{M}_{ij} , which read

$$\mathcal{M}_{ij} = \sum_{\mathbf{k}} w_{ij\mathbf{k}} = \begin{pmatrix} \sum_{\mathbf{k}} (w_{1,1})_{\mathbf{k}} & \sum_{\mathbf{k}} (w_{1,2})_{\mathbf{k}} & \cdots & \sum_{\mathbf{k}} (w_{1,j})_{\mathbf{k}} \\ \sum_{\mathbf{k}} (w_{2,1})_{\mathbf{k}} & \sum_{\mathbf{k}} (w_{2,2})_{\mathbf{k}} & \cdots & \sum_{\mathbf{k}} (w_{2,j})_{\mathbf{k}} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{\mathbf{k}} (w_{i,1})_{\mathbf{k}} & \sum_{\mathbf{k}} (w_{i,2})_{\mathbf{k}} & \cdots & \sum_{\mathbf{k}} (w_{i,j})_{\mathbf{k}} \end{pmatrix} \quad (\text{VI.2})$$

where $\mathbf{k} \equiv k_{ij}$ denotes the individual number of instances of each feature j around transcript i . Additionally we specify a class vector with elements v_i containing the information about every gene transcript from the training sample being associated with a certain class \mathcal{C} . This is the target variable vector, which in classification problems consists of discrete numerical or categorical variables. We will focus on an advanced decision tree method as described in section II.3.1, namely *Extremely Randomized Trees* (ERT), which like Random Forest represents an ensemble learning method yet with the addition that the decision splits are randomized as well.

As we also saw in section II.3.1 the choice of a best split in order to obtain some measure of homogeneity of the target/response variable depends on the splitting function. Hence we can either apply the so-called Gini impurity or information gain as a split estimator⁴. For later analysis we choose the Gini impurity which is partially due to the fact that it is more commonly known in decision tree analysis and requires less computational resources since it is not logarithmic. Yet for completeness we list the highest ranked results of both metrics in table C.8 in the appendix from which we see that the top-ranked results are in very good accordance.

VI.2.2 Results

In Fig.VI.2 we find the top-ranked results for the enhancer class features \mathcal{E}_j as indicated by their respective Gini impurity⁵. We can interpret these values basically as an importance measure in their ability to distinguish between the classes \mathcal{C} included in the response vector \vec{v} . We will call this ability *inter-class specificity*. Among the highest ranking features, i.e. the enhancer state classes \mathcal{E}_j , are enhancers that are prominently expected to be found around Th1 genes – for example those active in Tbet^{+/+}Th1 alone or in combination with Tbet^{+/-}Th1 – as well as those around Th2 genes – for

³This could be in general any differentially expressed set of genes being unique to some specific cell condition.

⁴It turns out that the results between both metrics are negligible [269].

⁵The full list is given in table C.9 in the appendix.

example those active in Th2 cells alone – but there are also features that could not be readily expected being ranked that high. Especially the first two highest ranked features are additionally active in Tbet^{+/+}Th1/2 cells hence exhibit a high ability to distinguish between the different transcript classes, which is apparently even more prominent due to their ability of maintaining enhancer activity under reprogramming culture conditions. We note that this is quite an astonishing result in its own right.

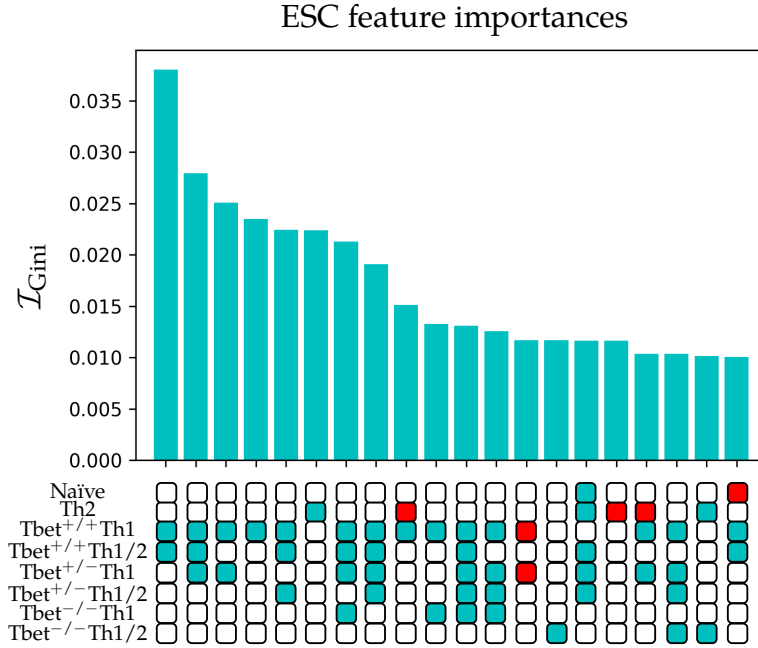


Figure VI.2: Feature ranking w.r.t. Gini impurity of all ESC features. The features are labelled by the above introduced colour-coding. For clarity we only depict the top 20 ranked features.

To check the robustness of the Gini impurity ranking we employed a jackknife method on all the transcripts used for the training of the decision tree. We find that on average the ranking stays mostly the same especially concerning the top-ranked features. In fact we recover the top 10 feature ranks in all cases and the top 20 features with an accuracy of 93.44%. We will follow up more rigorously on the dependencies of the feature ranking with respect to sample removal from the training set with cross-validation in the following.

In Fig. VI.3 we additionally show the Gini impurities for the top 20 repressive state combinations. As could be expected we find highly ranked RSCs which solely exhibit repressive states in either Th2 cell conditions or in wild-type Th1 cells. Additionally we yet find RSCs with repressive states occurring in Tbet knock-out conditions and in Th1/2 conditions as well. Quite interestingly the most highly ranked RSCs all do not include an enhancer state in any other cell condition. Hence we see the significantly top-ranked class-distinguishing RSCs do to large extent not include those state classes which had potential ambiguity overlaps of at the same time exhibiting active enhancer states removed but rather those which are characterized as acting solely repressively. Hence this shows that for ESCs repressive states or repressive mark residuals play a significant role, while on the other hand for distinct RSCs active enhancer states are to some extent redundant.

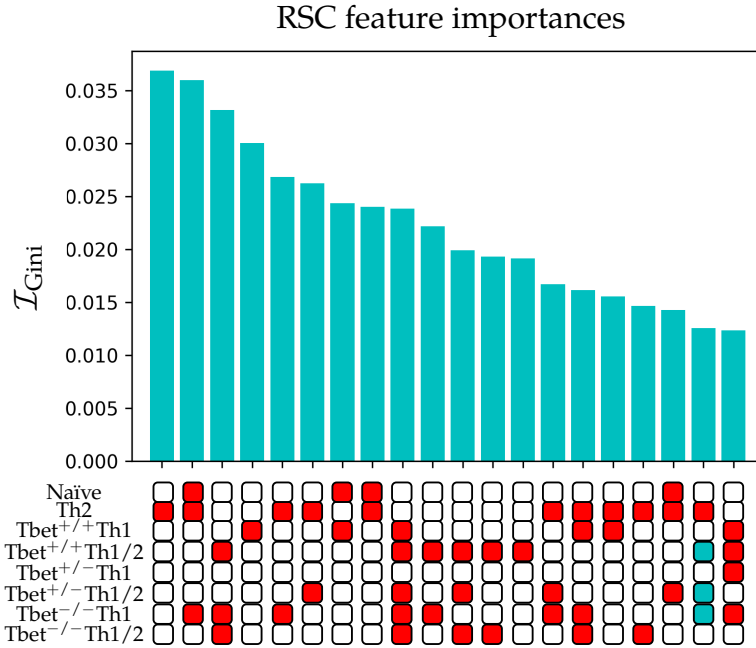


Figure VI.3: Feature ranking w.r.t. Gini impurity of all RSC features. The features are again labelled by the above introduced colour-coding and again we only depict the top 20 ranked features for clarity.

VI.2.3 Classification of transcript specificity

Obviously the above inference of the Gini impurity measure from the underlying learning sample can be applied in turn to predictions of the class a certain gene transcript belongs to, hence in our case labelling a transcript as being associated rather with Th1 or Th2 differentiated cells. In this case the occurrence of certain ESCs determines statistically to which class the transcript is assigned.

Results

When we perform this classification on our full training set again we obtain a classification accuracy score of 100%. This is not a trivial result since there might still be some inconclusive locus w.r.t. its weighted feature occurrences in their entirety not being highly predictive of the respective gene class. In order to find out a little bit more about classification accuracy we hence have to resort to resampling methods again.

To this end we employ leave-one-out cross-validation (LOOCV) to test how the prediction accuracy changes when one observation is left out of the training sample repeatedly. LOOCV is applied in contrast to leave-p-out cross-validation (LpOCV) since if we only leave out ca. 10% of the data of our training sample this becomes unfeasible w.r.t. computation time since the model has to validated $\binom{77}{8} \approx 2.1 \times 10^{10}$ times⁶, hence we stick to LOOCV. From LOOCV we obtain a prediction accuracy of 87%.

There are actually two main possible reasons for misclassification in the LOOCV case. The first one is rooted in the fact that a certain gene transcript might be es-

⁶LOOCV with 77 model validations already has a computation time of approximately 22 minutes on a 4GHz Intel i7 (4790K) with 32GB 1600 MHz DDR3 RAM.

essential for the correct inference of the respective tree ensemble parametrization. In fact we find that especially important transcripts like *STAT6-001* or *STAT1-007* can themselves not be categorized correctly if they are left out of the learning sample. This means that their epigenetic landscape imposes a strong statistical characteristic feature imprint on their own transcript-specificity. These transcripts with their epigenetic landscape are hence indispensable for correct annotation w.r.t. gene classes.

Another reason for misclassification due to LOOCV might yet as well be that we rather now found the true classification of the transcript which was concealed due to its initial misclassification and a possible uniqueness w.r.t. its epigenetic landscape. This can impose a bias on the parametrized tree method which in turn classifies a transcript with a similar epigenetic imprint always the same way leading to false positives. In contrast with the STAT-transcripts above for which independent validation of their cell-specificity exists (see e.g. [270, 371]) we find misclassified transcripts like *Eomes-003* with a prediction probability of 87.5 % of which the role in Th1 and Th2 differentiation is still being debated (see e.g. [99, 100, 213, 352]). We will come back to this possible re-classification of transcripts in the context of network clusters later on.

We note here that in principle the classification via ERT based on our training sample can be applied to any arbitrary gene transcript in question of which the surrounding epigenetic landscape has been analyzed w.r.t. a weighting of enhancer state classes \mathcal{E}_j . Furthermore the process described above can be applied to any arbitrary cell type for which a transcript learning sample classification is possible as soon as a significant number of appropriately graded cell conditions are available in order to observe some change in the epigenetic landscape, i.e. chromatin state class features can be inferred. Also this is obviously not restricted to enhancer state classes but can be applied to any arbitrary state pattern at hand as we have also indicated for RSCs above.

VI.3 Intra-class specificity of enhancer state classes

VI.3.1 Method

Although we find ESCs which are able to distinguish between opposing cell differentiation states, it would be even more important to be able not only to tell their inter-class specificity but also pin down in what way this ranking can be translated to some cell-specificity with respect to the different gene classes themselves. Hence we would like to determine which features rather classify Th1-specific gene transcripts and are hence Th1-specific themselves and which ones rather belong to Th2, hence establishing the respective epigenetic landscape of a gene transcript as a prior for its cell-specificity. This is what we call *intra-class specificity* from now on. At the same time we want to keep the intrinsic predictive ranked power of the Gini impurity. To this end we propose a novel intra-class specificity measure \mathcal{I}_C which acts as a weight for the already determined Gini impurity $\mathcal{I}_{\text{Gini}}$. Hence we want to obtain a modified intra-class Gini impurity $\mathcal{I}_{\text{Gini}}^*$ which reads

$$\mathcal{I}_{\text{Gini}}^* = \mathcal{I}_C \cdot \mathcal{I}_{\text{Gini}}. \quad (\text{VI.3})$$

The intra-class-specificity measure is now defined as

$$\mathcal{I}_{\mathcal{C},j} = \mathcal{N}_j \cdot \frac{\frac{\sum_i^{m \in \mathcal{C}} \mathcal{M}_{ij}(\mathcal{C})}{m \in \mathcal{C}}}{\frac{\sum_{i,j}^{m,n \in \mathcal{C}} \mathcal{M}_{ij}(\mathcal{C})}{n \in \mathcal{C}}} \cdot \frac{\frac{\sum_{i,j}^{m,n \notin \mathcal{C}} \mathcal{M}_{ij}(\neg \mathcal{C})}{n \notin \mathcal{C}}}{\frac{\sum_i^{m \notin \mathcal{C}} \mathcal{M}_{ij}(\neg \mathcal{C})}{m \notin \mathcal{C}}} \quad (\text{VI.4})$$

In principle we introduce a class-specific weighting measure with proper inter- and intra-class normalization (similar class-specificity measure definitions which are either less sophisticated or not applicable for our purposes can be found e.g. in [276, 368]). In order to do this right we have to account for all parameters w.r.t. uniqueness of class of the different features as well as ensure comparability between all dimensions of the transcript-feature matrix with elements \mathcal{M}_{ij} .

Herein the index j denotes the feature number with a maximum at n for a certain class, \mathcal{C} denotes the respective class (here either Th1 or Th2) and the maximal number of instances of a class is delimited by m . This normalization and averaging procedure takes all the weighted matrix entries from the feature matrix \mathcal{M}_{ij} belonging to a certain class \mathcal{C} and a certain feature – being a ternary state combination – and averages them for every feature separately. This is then normalized by the total weight of an average instance. This additionally has to be normalized by the same measure for all other classes to make it independent of the total number of instances and classes. We also multiply this by the number of instances where we find an entry larger than zero for each feature j which we call \mathcal{N}_j . In the case of a binary class categorization this yields two complimentary feature rankings for each class separately.

In a more readable way the intra-class-specificity measure is determined by

$$\mathcal{I}_{\mathcal{C},j} = \frac{\frac{\text{average weight for class } \mathcal{C} \text{ and feature } j}{\text{all combined weights for class } \mathcal{C} \text{ averaged over all class features}}}{\frac{\frac{\text{average weight for all other classes and features } j}{\text{all combined weights for all other classes}}}$$

additionally weighted by the number of instances where an entry is found for feature j . We also chose as a normalization of the respective class \mathcal{C} the same expression for all features $\notin \mathcal{C}$. All of this ensures that the resulting modified impurity measure is comparable between each feature within a class and in addition between features of different classes being responsible for introducing a class-specificity distinction. For two gene transcript classes this is furthermore pretty straightforward yielding two fully complementary intra-class-specificity feature rankings. The underlying Python code used for the computation of the following results is shown in appendix D.

VI.3.2 Results

We can observe the respectively top-ranked results of the modified Gini impurity weighted features in Fig. VI.4 for Th1-specific as well as Th2-specific genes while a full account is given in tables C.10 and C.11.

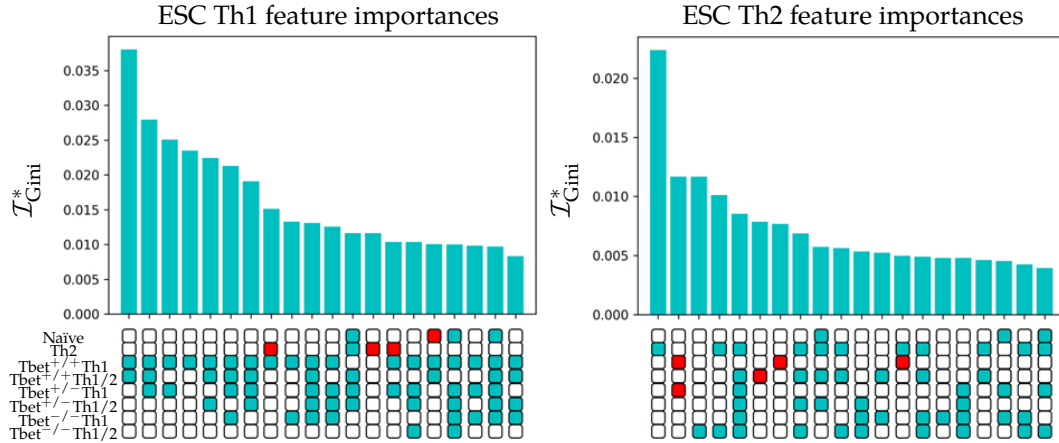


Figure VI.4: Top 20 ranked ESC features for Th1 (left) as well as Th2 transcripts (right). The ranking was obtained via the application of the intra-class Gini impurity.

We find that 280 features are considered to be Th1-specific while 390 are considered to be Th2-specific. For Th1-specificity we find multiple astonishing results. The highest ranked ESC contains active enhancer states only in $Tbet^{+/+}Th1$ conditions as well as in the corresponding Th1/2 conditions. Hence Tbet activity on both alleles is the best prior for gene-transcript Th1-specificity. This is closely followed by classes which exhibit enhancer activity in heterozygous Tbet conditions as well. Quite obvious is the high-ranking occurrence of enhancer states only being active in classic Th1 cell conditions. We also find a high Th1 importance for the ESC with an active enhancer in wild-type Tbet Th1 cell conditions as well as a repressive state in Th2 conditions. In the case of the ESC at ranking position 13 we see that it contains a repressive state in Th2 cells yet no active enhancer state in any of the conditions. This is due to our typology of observing enhancer activity, since we also correlate poised, hence non-active, enhancer states with gene expression s.t. ESCs without any active enhancer state occur among ESCs as well. This basically shows that there is a high occurrence of these significantly correlating state classes around Th1 genes compared to Th2 genes, possessing significant predictive power as well.

Furthermore we can annotate the respective CSCs w.r.t. their response to either Tbet or cytokine dose as we might have already seen. It turns out that the leading Th1-specific features are those that show heavy Tbet dose dependencies. This strengthens the view of Tbet as being a major player in Th1 differentiation. Additionally these results imply a quite astonishing finding, namely that without the quantification through the modified Gini impurity the naïve expectation would be that cytokine dependence would be of higher importance than Tbet dose for regulation of gene expression w.r.t. a certain differentiation path. This is apparently only partially the case since Tbet is at least of comparable importance within the respective ranking. A general distinction can be observed especially when it comes to a mixture of both dependencies. Let us for example consider the second-highest ranked feature for Th1 specificity. We see that the Tbet dose dependency is in turn dependent

on the cytokine dose. This manifests itself in the stability of the active enhancer state under neutral culture conditions vanishing only for Tbet knock-out conditions while vanishing already for heterozygous cells in Th2 polarizing environments.

We note that some ESCs w.r.t. their combinations of chromatin states appear to be somewhat harder to interpret. Examples are the ones at position 9, where enhancer activity is switched off in Tbet^{+/-}Th1 cells but on again in Tbet^{-/-}Th1, or at position 12, where enhancer activity is switched off in Tbet^{+/+}Th1 cells but we instead find active enhancers in Th2 cells. In the former case we find poised enhancers in Tbet^{+/-}Th1 conditions that were not classified as being active while in the latter case this holds true for the Tbet^{+/+}Th1 condition. There are several possible reasons for that. One is that the HMM categorizes several flanking regions of active enhancer state segments as poised since H3K27ac peaks are more pronounced in the center of the accumulation of several segments. A penalization depending on some Markovian property might help in this case. Via this one could introduce a weighting depending on next-nearest neighbours of active enhancer peaks and hence include the transition probability dependence of poised and active enhancer states into the classification method introduced here. Hence in the case of position 12 where we also observe enhancer activity in the Th2 control we can assume for this reason⁷ that this ESC represents rather a constitutively active enhancer state, which is in fact heavily dependent on Tbet knock-out. Another viable explanation is that the distinction into Th1- and Th2-specific transcripts is still too crude and especially reprogramming from Th1→Th2 has to be taken into account. In this case we hypothesize such an ESC would drop out of the Th1-specificity ranking and rather reoccur in a Th1/2 ranking. We will have a more detailed look into this ESC distinction when we investigate differential networks in section VII.2.7.

Turning to Th2-specificity we see a similar picture yet cytokine dependency plays a larger role here. This can be readily expected since Th2 genes might be only to a lower extent depend on Tbet dose while we could for example expect a similar dose dependency on a Th2-specific master transcription factor like Gata3 if such data was available. In our case we observe Tbet dependencies for slightly lower modified Gini impurities and we also see that rather repression of Th1 cells with high Tbet dose can play a large role.

The results for the RSCs are shown in Fig.B.13. From a total of 387 features 179 are classified as being Th1-specific while 208 are Th2-specific. As can be readily expected we find among Th1 transcript-specific RSCs those that exhibit repressive states in the Th2 cell conditions while the opposite is true for Th2 transcript-specific RSCs. In the case of Th1 feature importances we also find that repressive states in Th1/2 conditions especially play an important role in determining the Th1 transcript class. Rather surprisingly we also find a high ranked importance of an RSC with repressive states in Th1 as well as in Th2 cells which occur quite frequently around Th2 gene transcripts. This suggests that the Th2 transcripts under consideration sometimes also exhibit repressive states in their characteristic cell conditions hence repression also seems to be relevant in those cases. In conclusion also the intra-class-specific results

⁷For the sake of the argument we accept the above statement that the wild-type Tbet condition exhibits poised enhancer states. This has to be the case since if they occur around Th1-labelled transcripts the necessary condition for an element to be correlated was that it had to be an enhancer HMM state (see table V.1).

for RSCs confirm our former expectations and provide us with a ranked importance measure for class distinction of RSCs.

As a follow-up we asked how many of the top-ranked features have to be taken into account in order to yield a certain percentage in prediction power as indicated by the Gini impurity⁸ and how this quantity is influenced by the amount of considered chromatin state classes. In table VI.1 we find a summary of specific percentages of the total Gini impurity with corresponding number of considered best ranked Th1 and Th2 states. For this we used the ranking of the unmodified impurity first and mapped these results on the class-specificities.

Gini impurity %	# Th1 ESCs	# Th2 ESCs	# Th1 RSCs	# Th2 RSCs
99.5%	230	285	150	165
95%	140	160	95	100
90%	100	120	75	75
85%	80	95	60	60
75%	55	62	42	42

Table VI.1: Total amount of top ranked Th1 and Th2 ESCs and RSCs respectively in order to reproduce a certain percentage of total Gini impurity.

In the case of the Top 20 features for ESCs we obtain percentages of 47% and 43% and for RSCs both times 56% for Th1 and Th2 respectively.

We see that in order to approximate an optimal prediction accuracy we can nevertheless exclude certain features that don't contribute significantly to transcript class prediction. This is interesting since there seem to be certain state combinations and hence significantly correlating enhancer elements that are practically irrelevant when it comes to their significance in mutually exclusive differentiation programs. For a significance cut-off of 99.5% this enables us to exclude the 154 lowest ranked ESCs and 72 RSCs from further analysis. At the same time if we include the respective amount of highly-ranked features a larger amount of particularly hard-to-classify genes will be labelled correctly.

The nature of the above analysis suggests that the enhancer landscape of Th1 and Th2 genes consists of a variety of enhancer state combinations obeying a regulatory logic that is specific for the respective differentiation program. Not only can we infer this ESC (or more generally CSC) specificity, but we can also predict the cell-specificity of a certain gene with high confidence via a transcript-specific class probability based solely on its enhancer landscape. This is as well dependent on the amount of considered highly ranked ESCs. The validity of our method can in turn be confirmed in terms of the differential expression of respective genes from our RNA-Seq data sets.

⁸The Gini impurity is normalized to unity.

VI.4 Co-occurrence of enhancer state classes

Ultimately we are as well interested in a co-occurrence quantification of several of the top-ranked transcript-class-specific enhancer types. We exemplify this via a frequentist approach for the training set of Th1- and Th2-specific genes. In order to do this we have to compute the pairwise conditional probability for distinct instances of all significantly correlating enhancer state classes. Pretty straightforwardly we just determine the conditional probability

$$p(\mathcal{E}_i|\mathcal{E}_j) = \frac{p(\mathcal{E}_i \cap \mathcal{E}_j)}{p(\mathcal{E}_j)}$$

$$\forall i \neq j$$

where we consider pairwise occurrences within each TAD. This results in 75.344 conditional probabilities. In order to remove frequently occurring residuals we only focus on the Top 20 Th1- and Th2-specific ESCs respectively, which leaves us with 1186 combinations. From this we obtain a conditional probability ranking shown in table C.12. Among the highest ranking conditional co-occurrence patterns are especially Th1-specific ESCs with \mathcal{E}_i exhibiting an active enhancer only in Tbet^{+/+}Th1 cells. Among them are most notably ESCs also exhibiting active enhancers in heterozygous Tbet Th1 conditions or including repressive states in Th2 cells. The ranking also confirms the mutually exclusive specificity of ESCs preferably occurring with other highly ranked cell-specific ESCs. This also imposes boundaries on recruiting mechanisms of individual enhancers since a large variety of combinations appears quite frequently at gene loci playing a key role in cell differentiation.

VI.5 Discussion & Summary

We have investigated the regulatory activity changes of chromatin state patterns over different experimental conditions for significantly correlating enhancer and repressive state segments around notable Th1 and Th2 transcripts. To this end we introduced a ternary state classification considering enhancer state activity as well as the appearance of repressive states with a switch-like logic. Considering these changes in chromatin state activity w.r.t. environmental changes this naturally results in a class hierarchy wherein simple enhancer states are abstracted based on their regulatory logic as being realized instances of so-called enhancer-state classes. This unique and to our knowledge yet unregarded viewpoint of enhancers belonging to certain subclasses depending on the investigated experimental stimuli presents a novel approach to the investigation of epigenetic landscapes. Enhancer or repressive state instances or more generally chromatin state instances hence belong to a certain functional group fulfilling certain regulatory tasks depending on the respective cell context. Hence from now on we focus on the set of all chromatin state classes in order to investigate functional particularities.

Furthermore we find that the regulatory state class logic around transcripts, which are special to a certain cell type, are indicators for the specificity of a certain gene transcript. This is plainly due to the fact that a certain regulatory logic is hypothesized

of being unique for a certain cell type. In order to show this and to elucidate if certain types of CSCs indeed provide a large regulatory potential w.r.t. classification of cell-specificity we investigated the epigenetic CSC landscapes of a training sample of Th1- and Th2-specific transcripts. This resulted in an inter-class-specific ranking where we found that certain ESCs as well as RSCs are more suitable for distinguishing between Th1- and Th2- transcripts. Moreover in order to investigate intra-class-specificity, i.e. associating a transcript with either Th1 or Th2 statistically, we introduced a novel intra-class-specificity measure in Eq.VI.4, which results in a hierarchy ranking of ESCs and RSCs for Th1 and Th2 transcripts. In the case of two classes the results are mutually exclusive while the method is in general applicable for an arbitrary number of cell classes. Among obvious CSC candidates for Th1 or Th2 cells we also found highly ranked CSCs which could not be readily inferred as providing a large classification potential by naïvely approaching the subject. More specifically this leads to a categorization of CSCs within the ranking to be more cytokine or rather more Tbet dose dependent. Especially in the case of Th1 transcripts we find the highest ranking ESC to be heavily Tbet dose dependent emphasizing the role of Tbet in Th1 regulation even further.

This general approach leads to the implication that in principle one can probabilistically determine the classification of a certain gene with the regulation of a certain experimental condition based on the fitted model from an underlying learning sample just by looking at the positively correlating CSCs at their locus. This was also investigated by a LOOCV method to show the robustness of the approach. We obtain a prediction accuracy of 87% revealing that on the one hand certain regulatory landscapes within the training set are rather unique like in the case of some *STAT* transcripts while other like *Eomes* lack unambiguous CSCs. Obviously these issues can be resolved by considering a larger training set but we furthermore conjecture that possible mis-classification apart from additional prior biases concerning false positives in the training sample could be also remedied by extending the set of transcript classes. Quite naturally classification could be refined further if we find genes specific to another cell condition, i.e. wild-type Th1/2, on which a new model can be learned. This will be the subject of future endeavours in classifying more precisely the condition-specificity of gene transcripts.

Apart from that we determined the amount of top ranked CSCs to be considered in order to obtain a certain level of total Gini impurity and hence purity of a branch within the ERT method. From this we find that a certain amount of low-ranked CSCs is indeed negligible for correct transcript classification.

In addition we also investigated the conditional probabilities of finding certain pairs of CSCs co-occurring at Th1 or Th2 loci. From these results we obtain further information for predicting just by investigation of a low number of e.g. ESCs around a gene promoter which regulatory elements are expected to act in concert at the locus and hence providing an additional prior for gene transcript specificity.

CHAPTER VII

Epigenetic network inference and analysis

We have already discussed the basic terminology as well as some preliminary important quantities that are commonly used in graph and network theory in section II.2. By analyzing the underlying epigenetic network topology in Th1 and Th2 cell types we aim at elucidating the role not only of certain gene transcripts w.r.t. their direct dynamic interaction but also via higher order motifs and topologies that only act indirectly via epigenetic states. This leads to the possibility of finding as of yet unknown relations between TFs and gene transcripts in general w.r.t. activation and inhibition. Additionally from the deduction of cell-specificity of enhancer state combinations as well as the prediction of a classification of a certain gene transcript as discussed in detail in the last chapter, this leads to the question if these results can be confirmed by investigation of the underlying network topology. We will address this as well as further questions concerning cell-specificity in the context of monolayer and multilayer networks in the following. To this end we will also investigate cell-specific topologies via differential networks and infer the respective community structures. In order to investigate steady state subnetworks we finally discuss the implications of random walks w.r.t. different network topologies and also an epigenetic regulatory extension of the classical mutual-inhibition/autoactivation tristable motif in this context.

For computation and visualization purposes we employ a mixture of the Python package NetworkX as well as Cytoscape and cytoscape.js.

VII.1 Typology of underlying network and adjacency matrix

VII.1.1 Definition of network components

We have argued earlier that we want to extend the concept of ordinary gene regulatory networks by including regulation via epigenetic states and hence potentially extending the topological structure of the underlying T-helper cell network. We already inferred activation of gene transcripts by individual enhancer instances by means of identifying significantly co-regulated enhancer state instances according to our correlation model in chapter V. The respectively obtained Pearson correlations potentially act as normalized weights w.r.t. activation of a certain gene transcripts, hence exhibiting inter-transcript comparability. On the other hand we also defined inhibition via

significantly correlating repressive states in addition to repressor binding at a certain chromatin state instance. These findings serve now as a starting point of a definition of the components of the desired epigenetic network.

Nodes

Previously we extended the concept of significantly correlating enhancers to a typology of chromatin state classes exhibiting functional differences w.r.t. their response to Tbet dose as well as cytokine dose. Furthermore we saw that these CSCs facilitate differences in predictive relevance w.r.t. Th1 and Th2 cell-specificity. Not only does this classification introduce a means of valuable insight regarding transcript regulation and predictive power for cell-specificity but it also considerably simplifies the state space concerning epigenetic interactions. Hence we propose the usage of ESCs and RSCs as regulatory nodes in the network in addition to the set of Th1 and Th2 genes. Furthermore for simplicity we restrict ourselves to genes instead of gene transcripts by integrating the information from all transcripts of a specific gene into a single entity respectively¹. This also resolves the problem of dealing with general TF binding data in the network as well.

The reduction of network complexity to a level of epigenetic state classes taking the function as distinct nodes within a network is to our knowledge an unprecedented way of looking at epigenetic gene regulation and can be employed without loss of generality and applied to other experimental systems as well. The reason for this conceptual shift is that we consider regulatory entities that already exhibit differences in their regulatory structure w.r.t. the considered cell conditions from the experiment. This shifts the focus from the very special distinct viewpoint of uniquely realized enhancer instances to a more ensemble-focussed approach. The advantage of this view is that it is done on a statistically larger scale where it does not necessarily matter if an enhancer has been called at a certain position but rather how the whole set of enhancers behaving in a similar way under certain experimental cell conditions regulate the genomic landscape as a whole. This is of particular importance when asking for example how enhancers being present in a certain cell condition are affected by a change in cytokine conditions or e.g. by a change in Tbet dose and how the network is turned is affected by this phenotypic change on an epigenetic level and hence also by the respective regulation via TFs.

For these reasons we propose the following node definition:

DEFINITION: Network node

We define the set of network nodes \mathcal{N}_i to consist of Th1- and Th2- specific genes as well as of significantly correlating ESCs and RSCs corresponding to the respective transcripts as obtained in chapter V. We note that the epigenetic nodes represent general chromatin state classes opposed to individually realized CSC instances.

¹This results in 64 distinct gene nodes.

Edges

If we only consider the aforementioned weighting of activating and inhibitory edges where a CSC regulates a certain gene we only obtain a unidirectional graph. We note that direct edges between CSC nodes are forbidden since we do not consider regulation between ESCs and/or RSCs. The same also holds true for direct regulation between genes. Furthermore these edges are weighted by the respective Pearson correlation coefficient between a certain CSC instance and the respective gene. Yet we observe that every transcript can be regulated by several correlated enhancer segments from the same CSC at the same time and we also defined a node to be a general CSC in contrast to a specific CSC instance. We thus obtain a potential activation of a certain gene by multiple edges coming from one and the same enhancer or repressive state class node. Such graphs are in general called *multigraphs*. In the particular case of directed edges such a graph is furthermore called *multidigraph* or *quiver*. Since we have to deal with a weighted multidigraph we have to find a way how to treat the combination of edge weights between every two nodes. A straightforward way to deal with this is by adding all available weights between a pairwise set of nodes $\{\mathcal{N}_i, \mathcal{N}_j\}$ for $i \neq j$. Alternatively one could introduce a weighting w.r.t. contribution either by width of the respective state class segment or rather by the mean of the parametrized histone measure over the conditions under consideration². Yet for the sake of considering the respective particular instances of state classes as comparable entities we employ the simplest definition of multigraph weighting, since it is nevertheless questionable e.g. if it is more or less important for a certain segment to be more or less extended than another one. In this case the priority could either be at the narrow distinct realization of a state as it is generally assumed when considering p300 peaks or one rather accepts that if a sufficiently high correlation value of a broad segment from a similarity merging of neighbouring elements can indeed be achieved that this is an even stronger argument for the segment to contribute more importantly to the regulation of the gene. Since we do not have any evidence at hand to support either of the two presented viewpoints we hence assume equal contribution from every realization of a particular chromatin state class instance. As a consequence of this we define the weighting \mathcal{W}_{ij} of a directed *multi-edge* between a set of nodes $\{\mathcal{N}_i, \mathcal{N}_j\}$, where i denotes a chromatin state and j denotes a gene transcript as

$$\mathcal{W}_{ij} = \sum_k w_{ijk} = \sum_k \mathcal{R}(\mathcal{E}_{ik}, \mathcal{G}_j) \quad (\text{VII.1})$$

where w_{ijk} denotes the k individual edge weights going from $i \rightarrow j$. At this point the underlying network in general includes activating and inhibiting edges yet we cannot obtain any interesting network motifs such as loops being due to the fact that the graph is unidirectional and direct CSC-CSC and gene-gene interactions are prohibited, since some mediation is always assumed. This leads to a bipartite graph with a set of CSC nodes and a set of gene nodes between which interactions are assumed. To this end we also want to infer a normalized statistical importance score of TF binding at chromatin state class instances.

²We are not aware of any other more sophisticated methods.

Inference of TF binding at CSCs

In order to assign a binding weight of a TF to a certain CSC we assess binding frequencies³ of the considered set of TFs (introduced in section V.3.1) around the set of the Th1 and Th2 genes of interest. We also considered TF binding up to 400 bp away from a CSC. These binding frequencies then form a training set for inferring classification importance scores as before in chapter VI making once more use of the ERT algorithm. To this end we again employ the modified Gini impurity from Eq.VI.4. In this case the number of classes at hand is considerably larger, i.e. given by the maximum number of chromatin state classes. The resulting Gini impurities are exemplified for ESC-specificity in table C.13 in the appendix.

This way of obtaining a viable weighting of binding occurrence at general CSCs is fully legitimate since the modified Gini impurity as defined previously presents a normalized relative measure of the contribution of a certain TF binding within a certain CSC. It also already accounts for an ensemble averaged weighting w.r.t. occurrence frequency. For every CSC a directed edge is hence obtained leading again to multi-edges being directed from TFs to CSCs. The weighting in this case reads

$$\mathcal{W}_{ji} = \sum_k w_{jik} = \sum_k \mathcal{I}_{\text{Gini},jik}^* \quad (\text{VII.2})$$

We note here that the individual modified Gini impurity $\mathcal{I}_{\text{Gini},ji}^*$ for a certain CSC with index i is also normalized to one hence the individual edge weights \mathcal{W}_{ij} and \mathcal{W}_{ji} are comparable. On this basis including the above edge definition from CSCs to genes we end up with a *weighted multidigraph* with the following general edge definition

DEFINITION: Network edge

An effective network edge is defined to be directed and to present a possible contraction of multiple directed edges itself if multiple connection between a pair of nodes is present. Additionally every edge obtains an edge weight \mathcal{W} . Each separate edge is unidirectional whereas for each edge weight \mathcal{W}_{ij} as defined in Eq.VII.1 we always allow for the existence of a corresponding multi-edge with edge weight \mathcal{W}_{ji} as defined in Eq.VII.2 having an opposed orientation, hence introducing bi-directionality. This forms the basis of a weighted multidigraph.

Adjacency matrix

From the above node and edge definitions we have seen that we have to consider an adjacency matrix for a weighted multidigraph. Its adjacency matrix with entries \mathcal{A}_{ij} can be constructed in general from the edge weights \mathcal{W}_{ij} and \mathcal{W}_{ji} . Starting from individual chromatin segments, which formed the output of our correlation algorithm, the general form of the corresponding adjacency matrix reads

$$\mathcal{A}_{ij} = \sum_{k,l} \mathcal{T}_{ik} \mathcal{C}_{kl} \mathcal{G}_{lj} \quad (\text{VII.3})$$

³In general read counts or scores at specific binding sites also present a viable option yet since the data sets stem from different sources the data pre-processing is rather heterogeneous and hence these values are not straightforwardly comparable without additional information.

We call the matrix with elements \mathcal{G}_{lj} *gene-transcript matrix* with l denoting all considered individual chromatin state instance segments and j denoting all gene transcripts as well as all occurring CSCs. The entries \mathcal{C}_{kl} form the so-called *chromatin-state-class matrix* with l as before and k spanning all CSCs and finally the elements \mathcal{T}_{ik} form the *transcription-factor-binding matrix* with k as before and i denoting CSCs as well as all TFs⁴. From this we obtain an easy-to-implement mapping of chromatin segment correlations with gene transcripts, which are contracted based on their frequencies to CSCs, which are then in turn mapped to TF binding occurrences. Essentially the resulting adjacency matrix can be presented in the following form:

$$\mathcal{A} = \begin{array}{c} \text{TFs} \\ \text{CSCs} \end{array} \begin{array}{cc} \text{Genes} & \text{CSCs} \\ \left(\begin{array}{cc} 0 & \mathcal{A}_1 \\ \mathcal{A}_2 & 0 \end{array} \right) \end{array} = \begin{array}{c} \text{TFs} \\ \text{CSCs} \end{array} \begin{array}{cc} \text{Genes} & \text{CSCs} \\ \left(\begin{array}{cc} 0 & \mathcal{W}_{ji} \\ \mathcal{W}_{ij} & 0 \end{array} \right) \end{array} \quad (\text{VII.4})$$

where we note that obviously according to the previous definitions $\mathcal{W}_{ij} \neq \mathcal{W}_{ji}^T$. Hence we obtain a non-symmetric hollow adjacency matrix. This has to be the case since edges are only possible from genes to CSCs and back. The result of this straightforward assumption is that the shortest connection between two genes has to be at least via one CSC. Finally we note that CSCs do not merely consist of ESCs and TF activation but also of RSCs as well as TF repression. Hence the adjacency matrix consists of activating and inhibiting edges. This can be distinguished within the adjacency matrix via either assigning negative signs to all inhibiting edge weights or rather extend the adjacency matrix to include a so-called regulation dimension k in which case the adjacency matrix becomes \mathcal{A}_{ijk} . In this case k just consists of the set of values {activation, inhibition}. In most cases we will make use of the latter in order to guarantee a non-negative matrix which has to be interpreted element-wise making edge weights comparable between activation and inhibition as well.

⁴As specified from the available ChIP-Seq data sets

VII.2 Network analysis

VII.2.1 Full CSC-gene network

From the above definition of the adjacency matrix we obtain the network depicted in Fig.VII.1 containing all significantly called CSCs as well as all previously considered Th1 and Th2 transcripts.

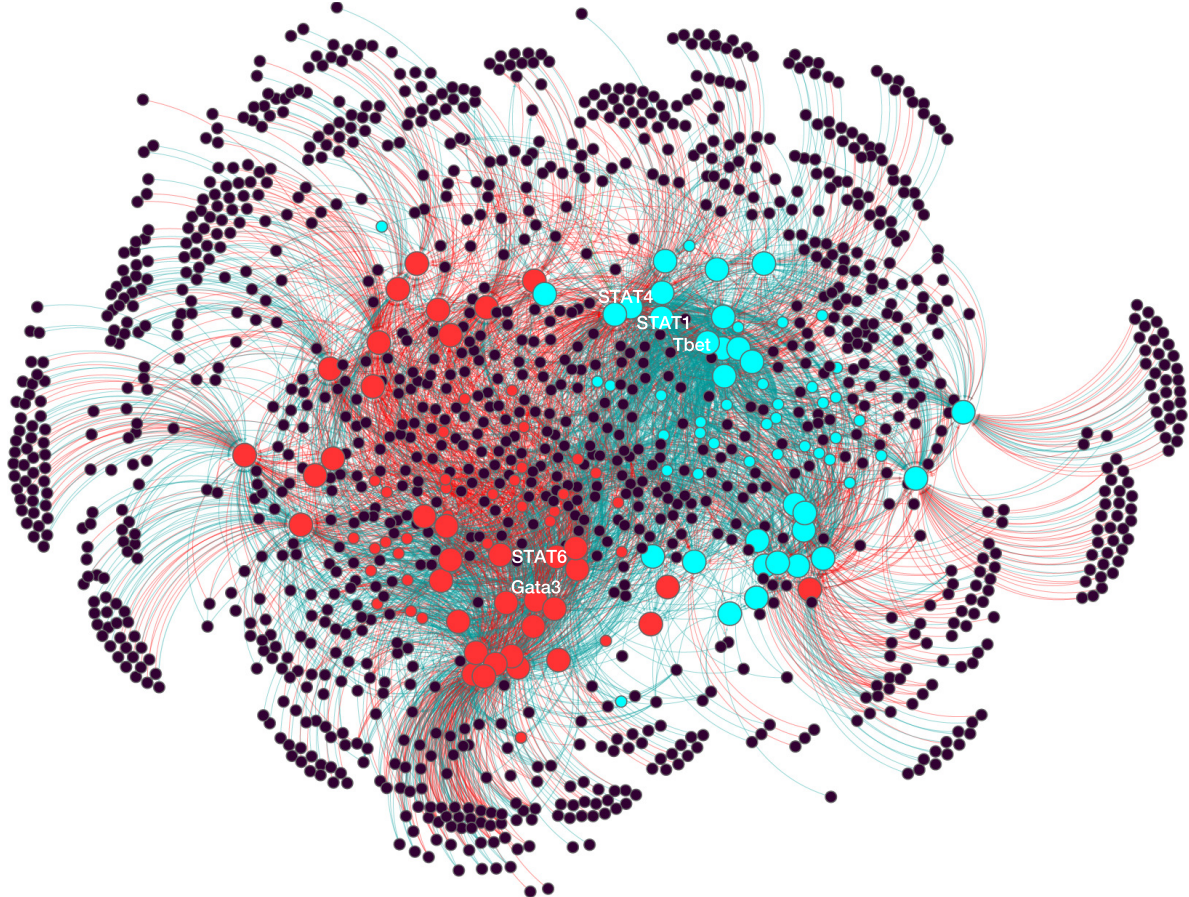


Figure VII.1: Full force-directed CSC-gene network resulting from Eq.VII.4. In cyan we depict the Th1-specific nodes as determined from the ERT learning sample, while in red we find the Th2-specific ones. Large nodes depict gene transcripts while small nodes denote CSCs. In case of CSC nodes we only coloured the top 20 cell-specific CSCs respectively – the rest is shown in black. Cyan edge colouring depicts activation while red edge labelling depicts inhibition. The edge weighting is proportional to the respective edge width, which becomes more pronounced in the core of the network. For simplicity we only denote the TFs specifically where the names are in each case overlapping with the respectively associated node.

The network depiction itself is chosen to be a force-directed graph drawing. Attractive and repulsive forces act on network nodes via a physical simulation in order to avoid edge crossings as well as overlaps of groups of nodes with high interconnectivity. The resulting network state is found at mechanical equilibrium where node positions cannot change w.r.t. spring-like edge attraction and Coulomb-like vertex repulsion. We hence obtain an intuitive visualization of the topology of the network already forming specific clusters only due to vertex connectivity via respective edge weights. Thus the data itself encoded in the adjacency matrix with entries \mathcal{A}_{ijk} controls the visualized output.

We visualize Th1- and Th2-cell-specificity according to the ERT method in cyan and red respectively, while large dots label genes and small dots CSCs. The labelling

in the case of CSCs is only done w.r.t. the top-ranked CSCs. Activation and inhibition is encoded in cyan and red edges respectively with the edge strength being visualized with differing widths.

We find a distinct separation of Th1 and Th2 gene transcripts validating the a priori classification from literature and RNA-Seq data as well as the transcript classification from section VI.2.3. The same holds true for the top-ranked CSCs that were classified as Th1- or Th2-specific respectively. Apart from a large part of CSCs that are tightly connected within the core of the network we additionally observe a larger number of detached smaller CSC clusters in the periphery, which are not tightly bound and hence do not contribute to many gene transcripts and in some cases are even not significantly regulated by some TF. This already leads to the hypothesis that we obtain a network, which consists of an inner part being approximately fully connected and an outer rim, which is merely partially connected. Furthermore we find that the respective Th1 and Th2 TFs can be found in close proximity respectively with especially STAT1 and STAT4 lying right next to each other. Additionally we note that STAT4 and STAT6 occur at the boundaries of their respective cell-specific sections indicating an important role in mediation between the two major parts.

We observe that although we employed a very reduced concept w.r.t. the definition of nodes by only considering CSCs we obtain a network that is highly connected. This can be quantified by a variety of network metrics as will be investigated in the following. A full summary of some basic network properties can be found in table C.14.

VII.2.2 Network properties and metrics

Degree distribution and hubs

We already introduced in section II.2 the degree k of a node \mathcal{N}_i which for a directed graphs results in two distributions, namely the in-degree distribution $P_{\text{in}}(k)$ and the out-degree distribution $P_{\text{out}}(k)$ of the whole set of graph vertices. Additionally since we consider weighted digraphs we also have to consider the weighted in-degree as well as out-degree of each node, being represented by the sum of weights of all incoming and outgoing edges respectively. In Fig.VII.2 we show the in- and out-degree distribution on a log-log scale with corresponding fitted power-law functions of the form $k^{-\gamma}$.

For the fitting procedure we excluded instances for $k = 0$ obtaining the following fitting parameters for the exponents:

$$\gamma_{\text{in}} \approx 0.98 \quad \gamma_{\text{out}} \approx 1.45. \quad (\text{VII.5})$$

We observe striking differences for the two different distributions. On the one hand this stems from the fact that some CSCs only serve as input functions for certain genes while no significant ChIP-Seq binding can be found in those cases. On the other hand the out-degree distribution exhibits several nodes with $k = 0$ which is due to the fact that not all genes serve as TFs and hence are not attributed as binding somewhere themselves. Turning to the in-degree the amount of nodes being subject to $k = 0$ is even larger. The reason for this is that there are several CSCs not experiencing any binding events from the TFs under consideration at all. Thus we exclude these nodes

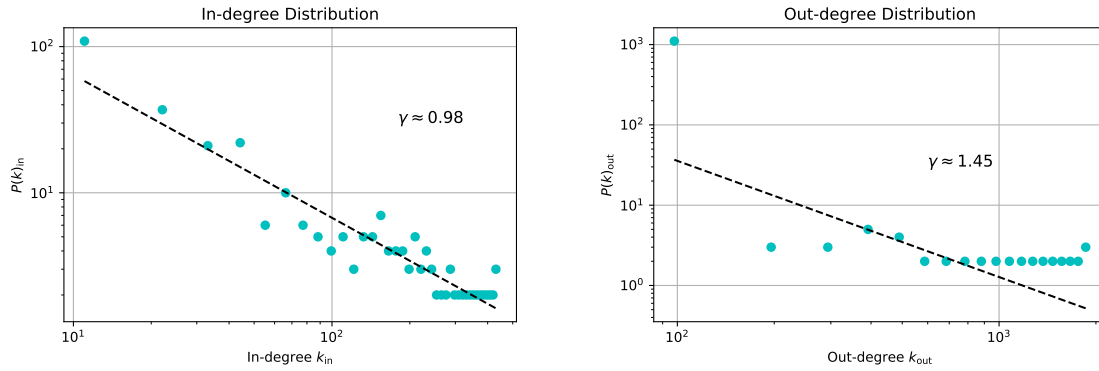


Figure VII.2: In- and out-degree distributions.

from the analysis step. The above power-law fits suggest that the underlying network topology is represented by a weak *scale-free* graph.

Although most investigated scale-free networks are found to exhibit higher scaling exponents compared to ours⁵ (see [22, 54, 71, 286, 308]) we still have reason to believe that we nevertheless find a similar tendency being weakly scale-free or heavy-tailed. Nevertheless we observe multiple features of real scale-free networks which contradicts the notion of a random network structure⁶. Among the reasons for our deviations are that we observe a bias from intermediate towards low degrees because of the low number of reliable TF data at hand, rendering the connectivity information at least incomplete. Nevertheless since the TFs under consideration are believed to be among the most relevant binding factors we find that they exhibit in combination with some frequently occurring ESCs the highest out-degrees within the network representing so-called hubs, which are notable features of scale-free graphs. The large values within the in-degree distribution on the other hand are dominated by genes themselves since they on average exhibit a larger amount of different CSCs contributing to their regulation than a certain TF w.r.t. binding to a CSC.

Nevertheless the main reason for the scaling exponent γ to lie in our case in a low regime – also called *anomalous regime* – is that we feature a multidigraph where we absorbed all multi-edges into one edge resulting from the sum of all multi-edge weights. This significantly lowers the degree of certain nodes. This also resolves problems of effective scaling exponents with $\gamma < 2$. One such problem is that if edges and nodes are added subsequently to such a network the hub-connecting edges would at some point exceed the number of network nodes. Without accounting for multi-edges a hub could at some point not connect to any more nodes. Yet since we already account for multi-edges we are also able to find a scale-free network within the anomalous regime.

In order to obtain a node-specific view we can simply visualize the *indegree* and *outdegree centralities* independently of each other which is shown in Fig.VII.3 and in Fig.B.14. From this we see straightforwardly that in the case of the highest ranked

⁵In fact most scale-free networks exhibit a scaling exponent of $2 < \gamma < 3$ being usually coined *ultra small-world* networks [22].

⁶There is clearly some debate on this issue as can for example be seen in [54, 71, 185] and the question breaks down to semantics regarding the fact if a network is characterized as being weakly or strongly scale-free.

in-degree centralities we find, quite unsurprisingly, the genes themselves followed by several CSCs from the central network. In the case of the out-degree again most CSCs from the core reoccur yet now also some peripheral CSC clusters come in, in one case serving as an input for Th1-specific genes (lower left part of figure) in another one forming an input for Th2-specific genes exclusively (middle right part of figure). Additionally the highest out-degrees are given by the TFs as well as by some mediating CSCs as well as some cell-specific CSCs. The actual candidates are shown in table C.15 respectively.

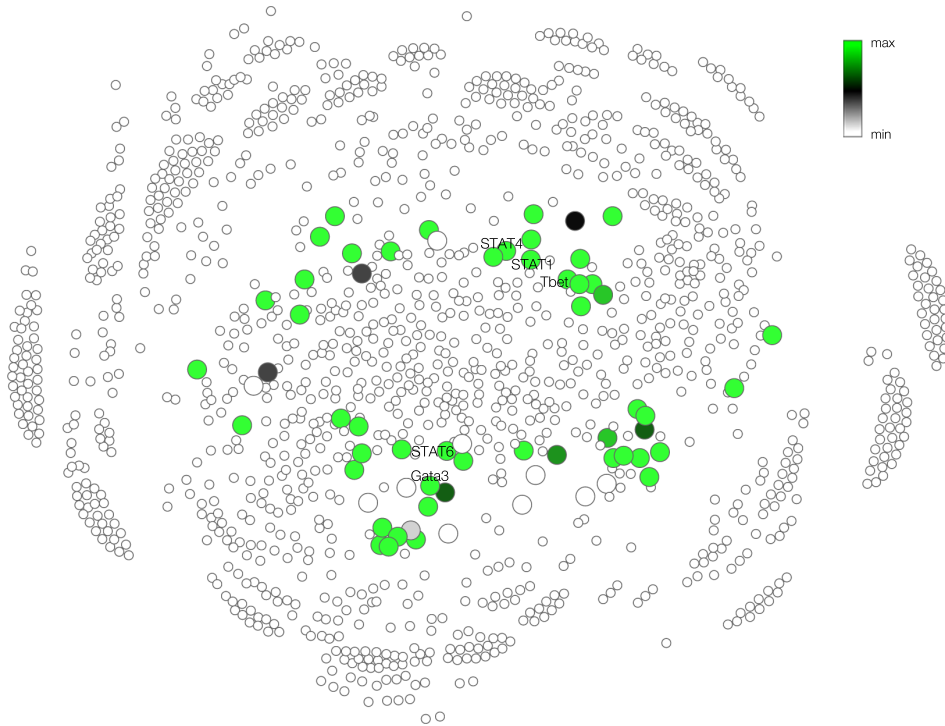


Figure VII.3: In-degree centrality for the full weighted multi-digraph. The colour-scheme depends on the maximum and minimum in-degree centrality values. As previously we depict only TF node labels for simplicity.

As we have already mentioned in section II.2 hubs are a consequence of power-laws in the degree distribution of the underlying network. We can see clearly in the in- and out-degree distributions that only a handful of nodes exhibit connectivity that significantly exceed the average degree. Unsurprisingly the network out-hubs are represented mainly by the TF nodes which exhibit high out-connectivity because of their ability to bind at a large number of CSCs. They are followed by some Th1- and Th2-specific CSCs which frequently occur and are also among the most relevant CSCs according to our inter- and intra-cell-specificity Gini impurity ranking in section VI.3. Examples are ESCs with active enhancers in wild-type Th1 conditions and in Th2 conditions respectively. In the case of the in-degree we find that hubs are again given by TFs but now as well by other notable Th1- and Th2-specific genes and as well by some CSCs. Since we established the notion of a weighted in- and out-degree we find that some CSCs are frequently regulated by a large number of TFs on the one hand, while there are some genes that are regulated by a very large number of CSCs. Since we observe multidigraphs this can mean regulation by a high number of only some CSCs or by a respectively low number of many different CSCs. For the in-degree

gene hubs we can conclude that there are several genes which are themselves highly regulated by a significant amount of CSC instances.

Quite naturally the full network quickly exhibits decreasing connectivity when removing several of the in- and out-degree hubs, especially concerning the TFs, which is also due to the principle of preferential attachment⁷. We also find preferential attachment within our network quite naturally w.r.t. frequently occurring Th1- or Th2-specific CSCs as well as w.r.t. the considered TFs. Hence although exhibiting high connectivity hubs in general pose a vulnerability risk to the network w.r.t. targeted attack, which we will investigate in due course. On the other hand we also obtain resilience w.r.t. perturbations in certain CSCs including complete removal. This is a key feature which we will recover in particular concerning genes involved in sub-networks when considering multiplex networks, which are condition-specific, later on. This is to be expected since most genes are recovered in hybrid Th1/2 conditions as well as in conditions with different Tbet dose dependency. Yet when removing certain Th1- or Th2-specific hubs from the network the network flow, which will be investigated later on in the context of random walkers, can change and hence also the importance weighting of a certain node. We also can assume that this will be in some way proportional to gene expression itself assuming e.g. a linear model of gene regulation by enhancers as in section V.6.

Yet since degree-distributions and hence also the degree-centralities do not exhibit a unique one-to-one relation w.r.t. network topology [92] there are certain other features related to network structure as well as to information flow that have to be taken into account, which we will investigate in the following.

Alternative centrality measures

There are several other popular centrality measures apart from the trivial degree centrality that are usually investigated in network theory which can give important insight into the importance ranking of a certain network node (see e.g. [236]). We already introduced some in section II.2. In Figs.VII.4–VII.7 we depict the so-called betweenness, closeness, eigenvector and Katz centralities of the nodes in our weighted multi-digraph for which the top 100 results are also listed in table C.16. The highest top ranked nodes for the betweenness as well as for the closeness centralities overlap to some extent and notably include all TFs being hubs in the network as well as some high ranked CSCs among which we also find Th1- and Th2-specific ESCs with high importance. The most obvious difference yet is that the betweenness centrality imposes a very strict boundary on some nodes exhibiting high relevance while the closeness centrality includes a high amount of nodes most of which are located in the core of the network and mainly including CSCs. In contrast to this betweenness completely excludes the network periphery. This is not surprising since in the case of the betweenness centrality only shortest paths are taken into account passing through the respective node \mathcal{N}_i , which are furthermore weighted by the sum of the respective edge strengths extending the purely topological centrality aspect. In this regard the betweenness centrality acts as a measure of mediating information between sets of

⁷We note that the same scale-free properties can be achieved by different mechanisms (e.g. fitness models (see e.g. [59]) not being restricted to preferential attachment also being a result from different distinct network topologies.

nodes within the network. We find that especially the TFs and some CSCs fulfill this task.

Closeness centrality on the other hand only can have a meaningful definition for nodes satisfying $k_{\text{out}} > 0$, excluding non-TF-coding genes, since it represents the relevance of a node w.r.t. to spreading information representing the starting point of that spread at the same time. We find that quite obviously the network core represents highly important nodes which also transmit information between the left and right network part, which can be associated with Th1- and Th2- specific tasks. Furthermore we find small CSC clusters at the beginning of the periphery in both cell-specific parts of the network which also contribute to an effective information spread themselves.

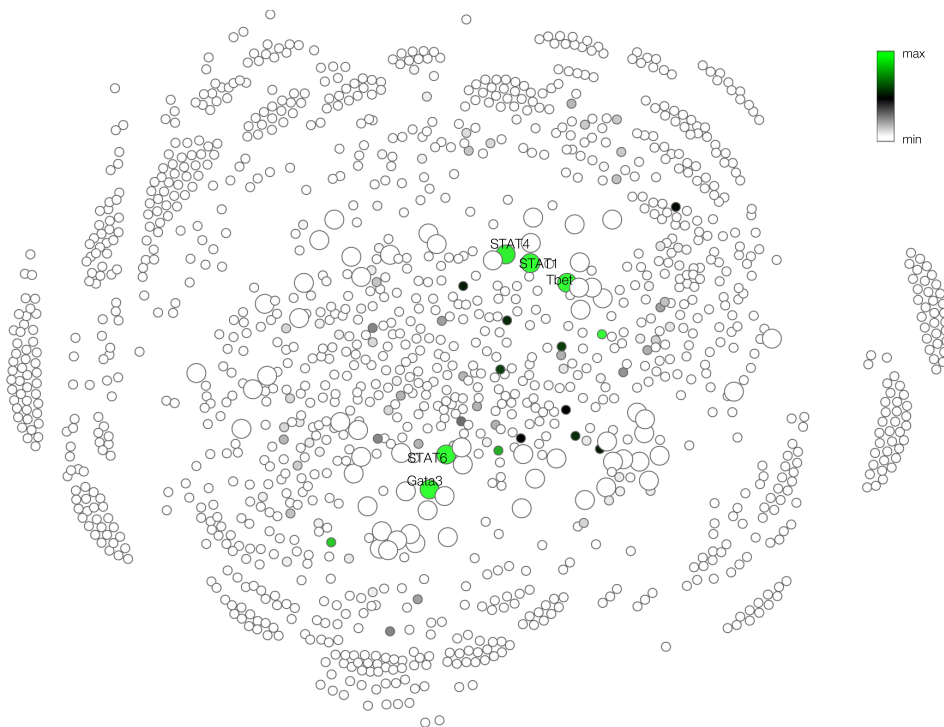


Figure VII.4: Betweenness centrality for the full weighted multi-digraph.

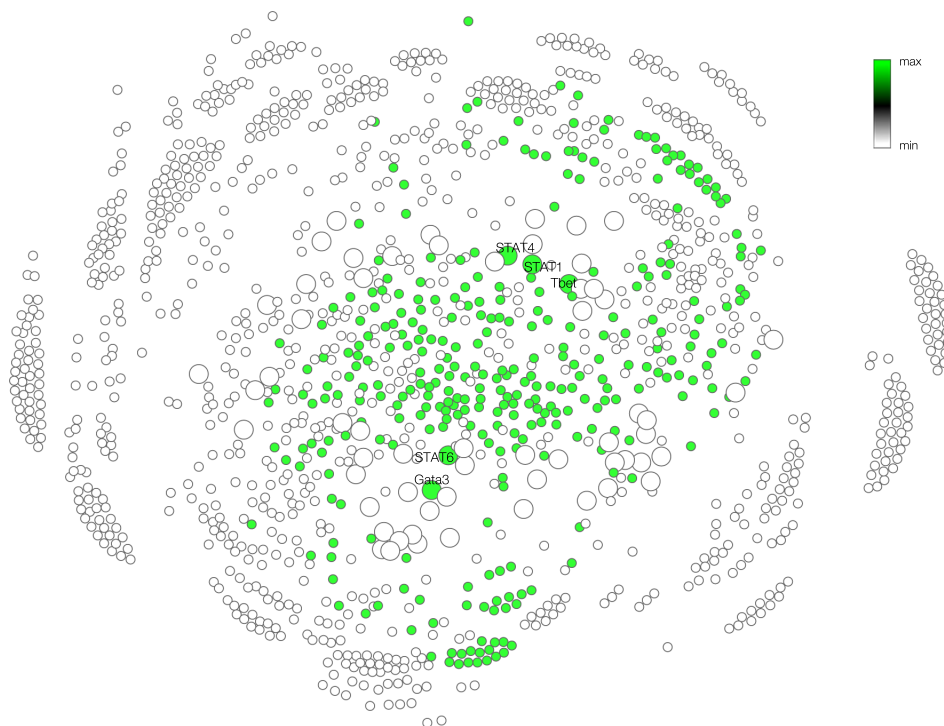


Figure VII.5: Closeness centrality for the full weighted multi-digraph.

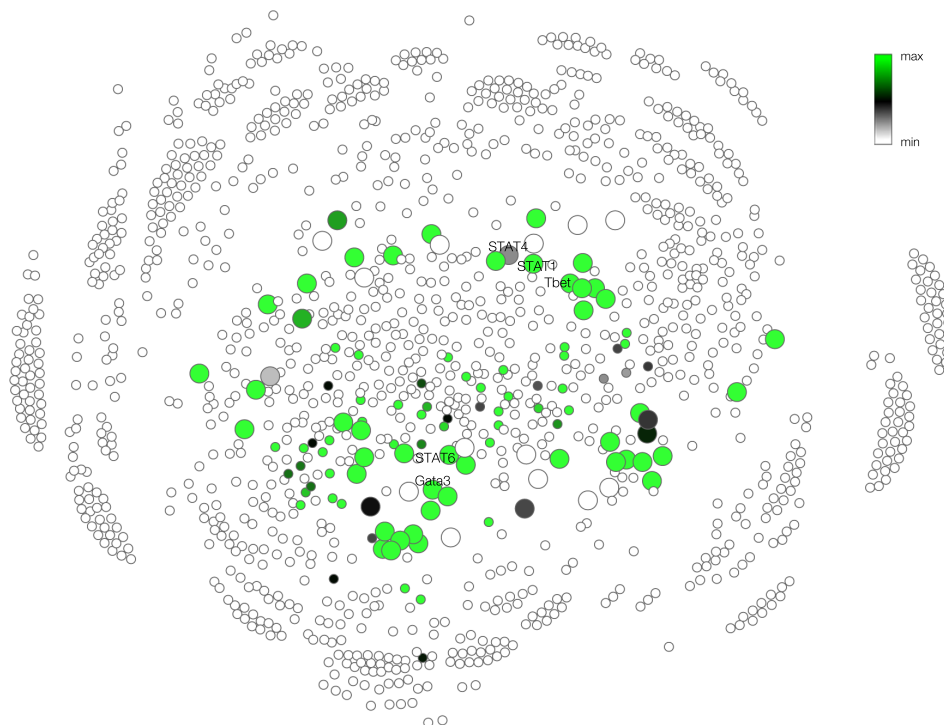


Figure VII.6: Eigenvector centrality for the full weighted multi-digraph.

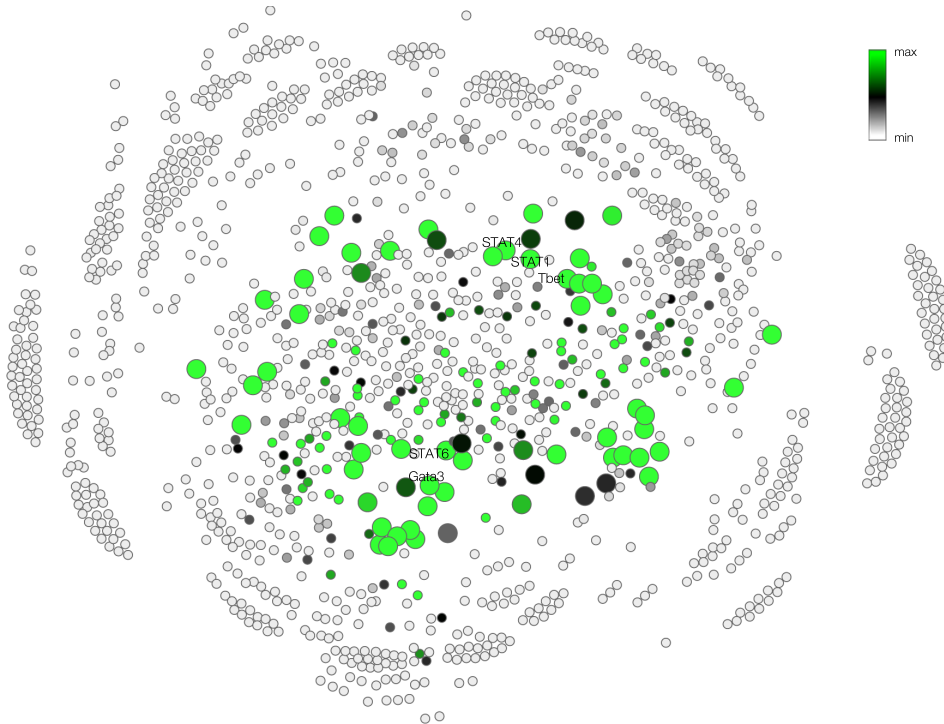


Figure VII.7: Katz centrality for the full weighted multi-digraph.

The eigenvector centrality on the other hand is an extension of the aforementioned in-degree centrality. Not only does it account for the contribution from other nodes to a particular node it furthermore takes the importance of neighbouring connected nodes into account. We especially observe that in contrast to the in-degree centrality we find a lower importance assigned to nodes in the periphery while nodes in the inner part of the network and especially CSCs receive a far higher centrality score. This is due to the fact that in addition to the in-degree importance of genes themselves, receiving large contributions from different CSCs, the CSCs are considered to have high importance since they are linked to these highly interlinked nodes. In this case among high-ranked nodes like the TFs we also find important nodes like *Ifn γ* on the Th1 side but also several genes from the Th2 cytokine cluster as well as highly ranked Th1- and Th2- specific ESCs and RSCs.

While the Katz centrality is related to the eigenvector centrality it extends the above concept also to nodes with no incoming edges themselves or in turn to those which receive input from nodes with no incoming edges. Hence Katz centrality also gives a non-zero centrality score to nodes which are not strongly connected including such highly linked nodes that are neglected in the eigenvector centrality analysis. For the Katz centrality the weighting barely changes as it is practically the same as in the eigenvector case. Only a very low number of nodes is minimally affected and non-zero centralities vanish completely. For us this is nevertheless a proof of principle with regard to the importance weighting of nodes within our network and confirms the above results.

We find that the most notable changes in assigning relevance to a network node appear between closeness, betweenness and variants of degree centrality, to which at this point we also count eigenvector and Katz centrality. In conclusion from between-

ness centrality we learn the nature of nodes with a highly mediating potential, while closeness centrality on the other hand yields nodes with a high potential for information spread themselves. Hence we find that there are many nodes apart from the standard network hubs experiencing high connectivity that act as potential targets in destabilizing information flow within the network. On the other hand this shows that the network is highly interconnected being a key feature of scale-free networks. Especially CSCs play a notable role for this property. The eigenvector and Katz centralities as an extension to the in-degree centrality now showcase the characteristics of each node w.r.t. receiving a certain information amount themselves, which is furthermore weighted by the importance scores of other highly ranked interconnected nodes that connect with those. In this case we find a reduced core network of nodes which consist of all genes within the network with varying contributions but also of certain CSCs that play a special role in regulation of these genes. In this case we do not only recover highly ranked Th1- and Th2-specific CSCs but also some intermediate ones being highly responsible for co-regulation of both cell-types, which in some cases are constitutively in an active enhancer or in a repressive state respectively and contribute to ESCs as well as RSCs. In other cases we also recover ESCs which are only switched on in the Th1 \rightarrow Th2 conditions. We will investigate these hybrid CSCs as well as highly interconnected genes associated with them in more detail in due course.

Motifs and loops

In the full network autoactivation loops are found to be very frequent. Hence we investigate their occurrences for some notable TFs, namely for T-bet and Gata3. In the case of Gata3 we find 34 possible autoactivation loops where we demand that only one mediating node, i.e. a CSC, occurs. If we do not consider statistical TF binding occurrences but actual ones as in Fig. IV.5 we still obtain seven distinct loops.

For Tbet we find 15 general statistical CSC autoactivation loops. For actual binding occurrences we still recover three loops as can be seen schematically in Fig. V.10.

Obviously we find that these loops become even more frequent when we also consider higher orders which are mediated by at least one more TF and CSC. For the time being we do not go further into this direction yet we note that this might be a worthwhile investigation when considering stability properties of Tbet and Gata3 expression in more detail.

In the case of mutual inhibition between both genes we obtain one CSC instance for Gata3 inhibition of Tbet with also one actual binding occurrence. In the reverse case where Tbet inhibits Gata3 we obtain three CSC instances while we also find three actual binding occurrences.

Quite obviously the differences between actual binding occurrences arising in the ChIP-Seq data sets and the statistical binding for autoactivation loops comes from the fact that the latter is a genome-wide measure in assessing statistical binding frequencies in general in all CSCs. Hence within the network we find certain loops between CSCs and TFs which might not always occur as a distinct binding in the actual data. Yet at the same time this rather accounts for the general mediation strength of a certain connection between two network components.

In addition to the above autoactivation and mutual inhibition assessment for the master transcription factors in Th1 and Th2 cells we also investigate the most common motifs occurring in the full network. Considering e.g. 4-node subgraphs we find 62.824 feed-forward loops (FFL) and 1673 autoactivation loops rendering these events in the full network quite frequent. Autoactivation loops on a 2-edge basis can obviously only occur for TFs as described above of which we find a total of 93 instances.

In order to investigate the statistically most frequently occurring network motifs we employ the FANMOD tool [337] which builds on the RAND-ESU subgraph enumeration algorithm, and is able to cope with edge colouring – being important for inhibiting edges – being in addition extremely fast in annotating the underlying minimal motifs. We performed the algorithm for subgraphs of four nodes. We obtain a Z-score based ranking comparing occurrence frequencies in the actual network w.r.t. random occurrences. From this we extract a top 10 ranking of subgraph motifs, which can be seen in Fig.B.15. As we can see for 4-node subgraphs the most frequently occurring motifs are coherent inhibition FFLs as well as different forms of coherent double inhibition FFLs also incorporating bidirectional edges. Quite notably most of the highest ranked subgraphs include inhibition. For 3- or 5-node subgraphs we obviously do not find any complete graph motif since we defined our network to mediate between any genes only via a CSC⁸. So any occurring loops have to be generated via modulo two nodes. We did not check for 6 node-subgraphs since for sampling of 1000 subgraphs alone this would result in a total subgraph enumeration of $2.5 \cdot 10^{10}$ instances.

Attack tolerance

We have previously mentioned that on the one hand the removal w.r.t. hubs or also w.r.t. nodes with too low out-degree can either have a large effect in the former case or no effect in the latter on the construction of the network itself. More specifically this is of particular interest when one wants to investigate the resilience of a certain network via so-called *percolation*, i.e. the random removal of a certain amount of nodes within the network in combination with its incoming and outgoing edges. The question which is usually posed is to find the so-called percolation threshold up to which nodes have to be removed in order to induce a phase transition and the network breaks up into smaller detached network components (see e.g. [22]). Yet this becomes quite problematic for directed graphs following a power-law degree $\lambda < 2$ since for $\gamma < 3$ we find a critical threshold of a fraction of removed nodes $f_c \rightarrow 1$ hence always showing some residual resilience and never breaking up completely unless nearly all nodes are removed (see [22, 268]). Hence we only focus on the effects the removal of major hubs, i.e. targeted attacks, have on the network topologies themselves.

We do this exemplary for the two master transcription factors Tbet and Gata3 as well as for STAT1. The results can be seen in Figs.B.16 and B.17. Upon removal of Tbet and Gata3 at the same time we clearly see that the Th1 and Th2 parts of the network are still preserved and we only observe a slight shift in the affiliation of some genes and a more precise localisation of some CSCs between both cell-specificities such as an ESC with an active enhancer in Th2 as well as in Th1 wild-type conditions.

⁸2-node graphs are fully accounted for via the above TF-CSC loop investigation.

Nevertheless we also observe a larger separation of STAT1 and STAT4 compared to before. Additionally we also observe the detachment of a class of CSCs being obviously unique to the two TFs. The same also holds true for the removal of STAT1 where we only observe the detachment of one node while still the Th1 and Th2 parts remain intact and do not considerably mix. From this we already see that even upon the removal of one or two major TF hubs in the network the overall network topology including the preservation of the Th1-Th2 dichotomy stays intact. Pushing node removal even further by removing approximately two thirds of all genes from the network we find the disintegration of a large amount of CSCs (indeed more than 900) while the Th1 and Th2 separation begins to break slowly. This also suggests that the underlying network indeed exhibits strong Th1- and Th2- specificity even after the removal of central hubs resulting indeed in strong resilience. On the other hand the removal of nodes results in the decrease of certain network motifs such as loops and, quite obviously, as soon as all TFs are gone, so is the potential for autoactivation loops. The view naturally changes if TFs are thought of as being necessary for ESC or RSC existence. In that case all CSCs would be removed that need the TFs for their respective activity and the network disintegrates faster.

VII.2.3 Core CSC-TF network

In addition to the full network consisting of 1.339 nodes we additionally consider a simpler core network in order to show only interactions of the main network components. To this end we include the top-10 ESCs and RSCs for Th1- and Th2-specific transcripts as well as all considered TFs. The result is shown in Fig.VII.8. We again obtain an utterly distinct separation of Th1- and Th2-characteristic gene transcripts as well as the predicted CSCs. We find that the classification obtained by the ERT method from section VI.3 can be naturally recovered already on this simple level independently validating the respective class-specificity.

We find several mediating components that bridge the gap between Th1- and Th2-specific clusters. In fact a prominent role in this case is played by ESCs exhibiting active enhancers in Th1/2 conditions as well as by RSCs exhibiting repressive states in Th1, Th2 as well as in Th1/2 conditions. We also find that nearly all connections in promoting exchange between these co-regulatory Th1 and Th2 modules are inhibiting as can be readily expected from mutual inhibition of Tbet and Gata3.

The two mixing Th2-specific CSCs (red circles) at the top of the Th1 cluster exhibit no active enhancer states in any condition and respectively a repressive state in Th1 conditions (left CSC) and one in the Th1/2 hybrid condition. The appearance of Th2-CSC on the left in the Th1 subnetwork can additionally be justified of it being an RSC with repressive states being activated by the binding of Th1-specific TFs.

Furthermore we observe that activating regulation in this core network exhibits the highest edge strengths in comparison to inhibiting edges marked in red.

One of the most important questions that now arise when observing the core network is obviously that of how the TFs are now effectively connected to each other and how that compares to prior knowledge on Th1 and Th2 cell networks. This will be analyzed in the following.

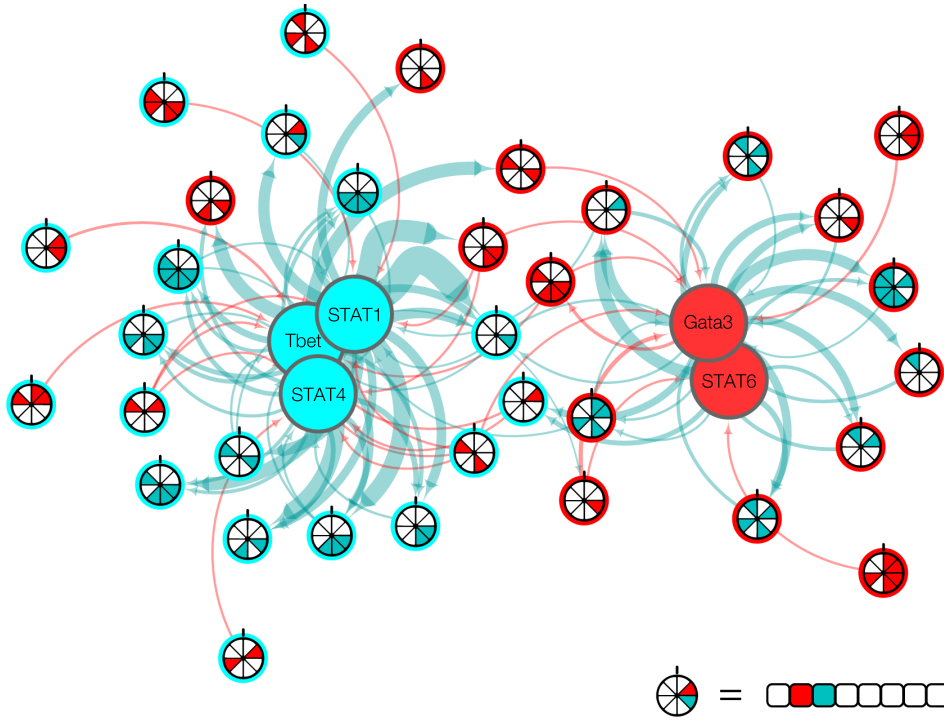


Figure VII.8: Core CSC-TF network with the respective ternary state labelling for every CSC with the same ordering as before. We denote Th1-specific CSCs with a cyan circle and Th2-specific ones with a red circle. Furthermore RSCs can be identified by exhibiting a red directed edge to some TF.

VII.2.4 Validation of known and prediction of novel TF connections

We remove now all CSC nodes from the TF network in such a way that they are effectively dressed. Hence we only observe the resulting TF interaction network. This is done by contracting the respective multi-edges. We note that there can be several connections from one TF to another that hence have to be accounted for, which are mediated via multiple different CSCs. Every CSC mediation is treated linearly independent from the others leading to a sum over all CSC mediations between a pair of TFs. We note that there are four different combinations possible concerning effective activation and inhibition of two consecutive edges. This can be seen in the following table:

	activation	inhibition
activation	activation	inhibition
inhibition	inhibition	activation

We accordingly obtain the following adjacency matrix for the resulting TF network

$$\mathcal{A}_{ij, \mathcal{C}_{\text{eff}}}^{\text{TF}} = \sum_{k \in \mathcal{C}_{\text{eff}}} \mathcal{W}_{ik} \mathcal{W}_{kj}. \quad (\text{VII.6})$$

where i denotes the first TF that interacts with a second TF j . The label k denotes the respective CSC and \mathcal{W} denotes the multi-edge weight. Since we can have effective inhibition being mediated from one TF to another via a particular CSC while

another one might mediate effective activation we define two separate matrices for the two different states $\mathcal{C}_{\text{eff}} = \{\text{activation, inhibition}\}$. Otherwise we would run into the problem that we would have to just add “simultaneous” activation and inhibition, yet these functions are mediated in our framework by completely different processes and hence are independent of each other. This leads to the definition of two adjacency matrices depending on the effective activation state of the mediating CSCs. The result of this is shown in Fig. VII.9.

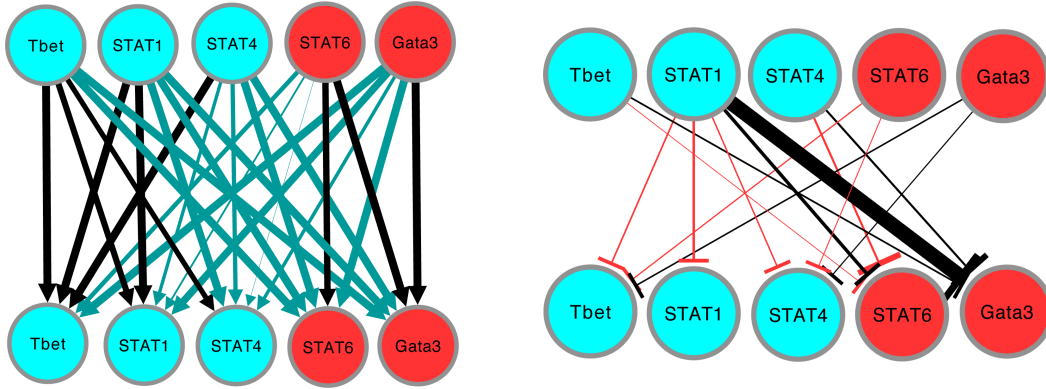


Figure VII.9: TF activation (left) and inhibition network (right) after removing the CSC layer. The edge strength is proportional to the respective edge weight from the effective adjacency matrix $\mathcal{A}_{ij, \mathcal{C}_{\text{eff}}}^{\text{TF}}$. Black edges represent known TF connections while cyan and red ones denote new predictions concerning activation and inhibition respectively.

For depiction purposes we scale the edge weights w.r.t. the largest absolute value for each figure separately. In addition we marked known validated connections in black and newly inferred connections in green. The already known connections are taken from [162, 256, 312, 369]. Considering only activating connections we can especially validate well-known connections such as Tbet being autoactivated by Tbet as well as receiving significant contributions via STAT1 and STAT4. Furthermore we can confirm autoactivation of Gata3 as well as activation by STAT6. In addition to that all TFs autoactivate themselves. Yet we also find activating connections of different order between many TFs also interchangeably mixing between Th1- and Th2-specific TFs. Yet some connections have a statistically significantly low contribution such as STAT6 activating any Th1-specific TF or Gata3 activating STAT1 or STAT4. On the other hand we also find significant high active regulation of STAT6, Gata3 and Tbet by TFs from the respectively opposing cell program. This was a feature not observed in the core network shown above. This is due to the fact that this regulation is mediated by cell-type unspecific ESCs, which are also ranked very low in the class-specificity ranking method. These rather generic constitutive ESCs w.r.t. the investigated cell conditions do not exhibit any active enhancer state in any condition or at least only in naïve cells yet they still significantly correlate with gene expression. We reasoned earlier that such ESCs frequently occur next to CSCs with active enhancer states quite often including poised enhancer states as can be seen from the transition probability of the HMM (see section IV.1).

In order to remove possible artifacts we also consider the respective TF activation network without those ESCs. The result is shown in Fig. VII.10. We find that considerably clearer distinctions can be made between the different TF interactions also removing most of the statistically determined contribution of activation between Tbet and Gata3 respectively. We also find contributions of STAT1 to STAT4 as well as

autoactivation of STAT4, which makes sense since they are all Th1-specific. Still we retain larger contributions of STAT1 and STAT4 to STAT6 respectively, which cannot be readily removed. This might indicate that there is some cross-activation between different cell-types on the level of STATs.

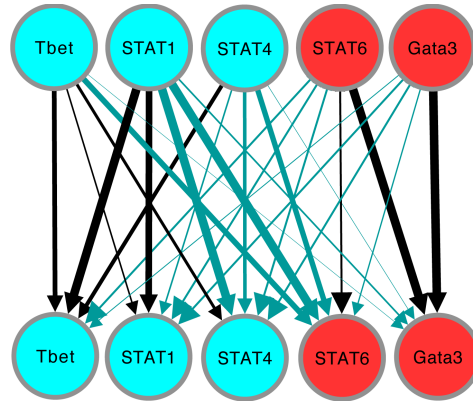


Figure VII.10: TF activation network with ESC layer removed. In addition also low-ranked ESCs were removed in this case.

Turning to the inhibition network we also confirm the well-known mutual inhibitions between T-bet and Gata3. In addition to this we also find several inhibiting relations between Th1- and Th2-specific TFs respectively, such as e.g. between STAT1 and Gata3 having by far the largest contribution or STAT6 and Tbet. Interestingly enough we also predict in addition to that inhibition of Tbet by STAT1 as well as possible auto-inhibition of STAT1. Interestingly enough the RSCs being the cause of this interaction exhibit repressive states in Th1 cells where STAT1 binding respectively occurs. Hence these inhibiting interactions are non-redundant w.r.t. the general node definition of the respective CSC being responsible for it. This means that w.r.t. a statistical point of view of this RSC there is a significant non-zero binding relevance of STAT1 at an inhibiting chromatin state element class regulating the respective CSC.

Although this interpretation of CSCs poses a novel view on epigenetic networks including numerous advantages w.r.t. their analysis there are still some disputable aspects of this approach. An example would be that since we employed an ensemble approach to TF interconnectivity we might have some redundancy especially in weakly interacting connections. Some of these interactions have been shown to be somewhat counterintuitive. The reason for their occurrence was that the edge weights \mathcal{W} themselves can be interpreted as relative probability weights due to the ensemble-based inference via the class-specificity method. Furthermore we hypothesized that a certain TF binding in a condition where we find an active enhancer acts positively on the activation of the ESC. This still would have to be confirmed via ChIP-Seq binding data of these TFs for all of our experimental conditions. Furthermore we would have to rule out first that for the co-binding of several TFs at a specific ESC we do find an activation which is due to the either linear independent or cooperative action of all these TFs and none of the TFs act as a weak repressor in turn regulating the activation of the ESC.

Yet since a great number of the above interactions involve CSCs with respective TF bindings which are clearly non-redundant we propose that there is actually an exceedingly large number of TF interactions to be found in the network, many of which

were unknown up to this point especially considering activation. Also quite a range of the inferred connections are poorly investigated and only reported e.g. in [256, 312] supporting several of our predictions. In addition w.r.t. the ensemble-focussed definition of our network we consider all of the interactions shown in Fig.VII.10 and in the inhibition TF network in Fig.VII.9 considering similar CSCs to behave the same and in consequence to this also to regulate the respective genes analogously.

VII.2.5 Network communities

Methods

One of the most revealing aspects of networks in general is the detection of community structure. A community⁹ is defined by a sufficiently dense clustering of neighbouring nodes, hence exhibiting high intra-connectivity. In addition to this dense internal community structure a distinction between communities is made by relatively sparse communal inter-connectivity. We already saw in the construction of the full CSC-gene as well as in the core network that there is already a visual distinction possible between Th1- and Th2-related genes and CSCs. We will quantify this observation in the following via community detection. Since there are in general several distinct possibilities for the definition of communities as well as their inferred number¹⁰ we compare several detection methods. Communities are in general determined via an unsupervised number of clusters, hence methods similar to k-means are not considered¹¹. We focus in particular on *hierarchical edge removal (HER)*¹², *random walk information flow (RWIF)*¹³ and *spectral clustering (SC)*, which are briefly described in section II.2. We consider all of these for the special case of undirectedness, which is due to the algorithmic implementation, yet still exhibiting the relevant connectivity features we are interested in. The results are depicted in Fig.VII.11 as a selection and in tables C.17-C.24 for all methods.

Results

We see that there is some consensus to be found in distinguishing between Th1- and Th2-specific genes as could be already expected. This can be seen e.g. for the first two modules for HER and SC (with $k = 5$). We observe that our previous classification is retrieved nearly without exceptions¹⁴. In the case of HER we find six distinct sub-communities where we find stronger association with a subset of Th2 genes between each other than with the rest. This is especially true for *Gata3* as well as for members

⁹We will use the terms community, module and cluster interchangeably.

¹⁰This happens e.g. along the lines of general clustering algorithms with a certain number k of cluster seeds.

¹¹The only exception to this rule will be the spectral clustering method since the automatic initialisation of cluster centers, which follows a spectral eigenvalue decomposition, yields an unfeasibly large number of clusters. Hence in this case we set $k = 5$ as well as $k = 10$.

¹²Based on the Girvan-Newman algorithm [127] and implemented in the GLay algorithm [296].

¹³Depending on the Markov-Cluster algorithm (MCL.) [102]

¹⁴The only exceptions in the SC case are *Dpysl3* occurring in the Th2 cluster and *Itga1* in the Th1 cluster. We also find the formerly classified Th2 ESCs which in one case exhibits no active enhancer in any condition and in the other case only an active enhancer in naïve conditions to appear in the Th1 cluster.

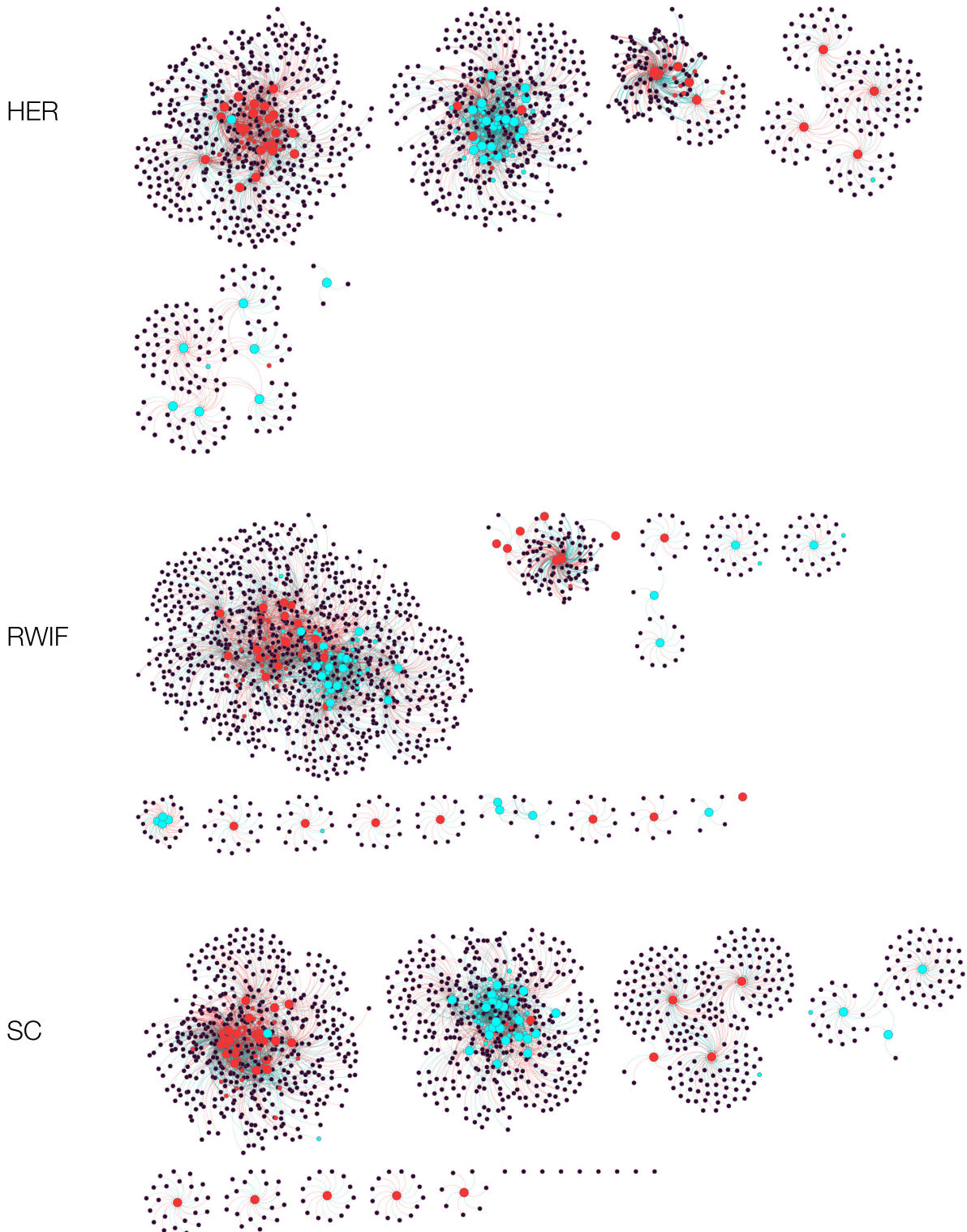


Figure VII.11: Resulting community structures according to different community algorithms. We show the results of the HER, RWIF (Granularity 1.8) and SC ($k = 5$) methods. More detailed association of distinct nodes with each cluster can be found in tables [C.17-C.24](#).

of the Th2 cytokine cluster. On the other hand we find that since a weaker association of those genes also w.r.t. the Th1 cluster exists and *STAT6* negatively regulates several

of those components, *STAT6* gets stuck to the periphery of the Th1 cluster hence acting as a strong mediator. We also find another detached Th1 cluster including *Runx3* and *Eomes* as well indicating a distinct role of those genes compared with classical Th1 genes like *STAT1* or *Tbx21*. We also found separate indications concerning *Eomes* via the intra-class specificity classification (see VI.3).

For SC we also find distinct clusters containing either Th2 or Th1 genes. In one case this features *Il10*, in another one e.g. *Il2*, *Ccr2* and *Ccl5*. When increasing the number of cluster centers to $k = 10$ (see Fig.B.18) we retrieve much of this structure yet some more fragmentation can be observed as well e.g. via a further detachment of several Th2-specific nodes into new clusters. From this we see that there are several subcommunities which are even more strongly connected and occur when increasing the cluster resolution. We can for example observe which particular subcluster w.r.t. genes is responsible for providing certain CSCs with a certain specificity like in the case of module 5 in Fig.VII.11. On the other hand we also start to see that some mixing is occurring between several network components being otherwise loosely connected with other subcommunities. Hence even classical Th2 genes like *STAT6* can again appear on the periphery of the main Th1 cluster indicating as well several mediating possibilities w.r.t. Th1-specific genes. Interestingly enough when performing the same procedure while only including ESCs and activating edges *STAT6* again ends up in the classical Th2 cluster. This means that in the former case the inhibition of Th1 outweighs the activation of Th2, which for example is not true for other hallmark genes such as *Gata3*.

In contrast to this we find for the RWIF method a larger number of modules compared to the other methods, which can be influenced by the so-called granularity or inflation parameter r , which essentially determines the connectivity between strong and weak edges within the network after some point in the ongoing Markov chain flow. The larger this factor the more the differences between communities get pronounced and hence more modules can form. Therefore we compare results for $r = 1.8$ with results for $r = 2.0$ (see Fig.B.19). First of all we find that for low values the main network structure is still retained indicating again tight connectivity between all components¹⁵, which was already indicated when we investigated attack tolerance of the network. From the perspective of a random walk flow this basically means that a random walker will get anywhere in the whole network quite quickly. In addition to that we find several detached elements like the Th2 cytokine cluster forming again a tightly connected subcommunity as well as several detached Th1 and Th2 genes, which are more strongly regulated by their respective CSCs. For larger granularities the network subcommunities become even more fine-grained providing mainly interesting results for direct community interaction. Here we also recover e.g. strong interaction between *STAT1* and *STAT4* which gets lost for lower granularities. More interestingly we also obtain in the fine-grained case a flow mediation cluster between Th1 and Th2 also including a high amount of CSCs with high Th1/2 activity. We will come back to Th1/2 networks in the next chapter.

In conclusion we observe that the different clustering methods not only serve different purposes by yielding a range of particular results, but they also confirm a large variety of previous observations. In case of the HER and SC method we clearly re-

¹⁵This makes it in fact harder to find distinct large-scale clusters, which is also due to the low scaling exponent of the in- and out-degrees.

cover the Th1- and Th2-specificity of several components. Nevertheless we find – depending on the number of clusters – that certain subcommunities are more tightly related than others. This is for example true for the Th2 cytokine cluster. On the other hand we find that certain key players like STAT6 play quite a large role in mediating inhibition to Th1. Additionally we obtain several clusters which are detached from the larger Th1 and Th2 clusters and which contain genes with mediating functionality. Examples for this are e.g. *Eomes* and *Runx3* [90, 99, 100, 189, 213, 352]. We see that the intra-cluster connectivity in the case of HER for these genes is statistically higher compared to their coupling to other classic Th1- or Th2-specific nodes. Yet since there obviously still exists coupling to the pure Th1- and Th2-clusters respectively a viable possibility for the functional role of the clusters would be indeed to mediate between pure Th1- and Th2-cell phenotypes. They might as well play a central role in hybrid cell conditions. In the case of RWIF we find a tight intraconnectivity of the whole network for low granularities resulting from low inflation. Increasing these differences we nevertheless recover some hybrid Th1/2 clusters at the cost of a large number of detached small gene clusters indicating possible candidates for direct strong regulation. We find that community detection for a weighted multi-digraph with an adjacency matrix as defined in Eq. VII.4 is an ambiguous and hard-to-solve problem and hence the definition of effective meta-nodes is not straightforward. That is also why the above results are so diverse. In addition this already hints at the shortcomings of modelling the Th1/Th2 network as an effectively dressed simple MISA motif. Because of the above results we would rather propose a coupling of a variety of motifs which seems essential to provide a complete possible Th1/Th2 multistability information (see e.g. [52, 250, 251]).

Yet since we want to learn more about condition-specificity, which cannot be solely unravelled by community detection, we turn now to so-called multilayer networks.

VII.2.6 Multiplex networks

The above discussion focuses on a network view which collapses all existing information from multiple experimental conditions via an ensemble-centric approach to one single graph, hence being called unidimensional network. Yet in this case the condition-specific perturbations w.r.t. Tbet dose or cytokine condition get lost or are at least dressed in terms of correlation values as well as the specific gene instances used within the network itself. A way out of this shortcoming is presented in the form of so-called *multilayer networks*¹⁶. Hence we want to map the information as it is used in the full network above to condition-specific subnetworks that provide only the condition relevant information. This leads to condition-dependent adjacency matrices with elements \mathcal{A}_{ij}^d where d denotes the cell condition or dimension of the respective network layer. We want to discuss the nature of these adjacency matrices in the following.

Considering all layers as a whole in order to include inter-layer connections there are two general ways the system can be observed. Either one considers weighted multilayer networks, which include weighted connections between the network nodes of

¹⁶We note that the terminology provides many ambiguities in the way the underlying terms are used within the network theory community itself. We refer here especially to the distinctions made in e.g. [186].

different layer dimension, or one considers so-called *multiplex* networks, which only track if a certain node can be mapped from one layer to another. When applying the former several complications arise and even when considering the latter careful scrutiny has to be applied. For instance for weighted multilayer definitions we would be in need of defining edges between identical transcript nodes as well as edges between identical CSC nodes and even a mixture between them. There is no straightforward solution to this. A probable remedy could yet involve the inter-dimensional edges to denote the respective activity changes, which would mean the normalized gene expression value for gene transcripts and the parametrized measure for every CSC. Yet in order not to complicate things even further we are considering so-called *multiplex* networks, which just track the existence of a certain node over various dimensions for the time being.

The question at this point is how a condition-specific network would look like in our case and what the requirement has to be for a node or an edge to be removed from the condition-specific network. Because of our ternary CSC definition we remove an enhancer state class from the network if the enhancer gets turned off in a certain condition and we do the same with an inhibiting state class if the repressive state vanishes in the respective condition. In addition we remove the Tbet-node in both Tbet^{-/-} conditions.

For visualization purposes we keep the general arrangement of the nodes over all conditions. We hence obtain the condition-specific multilayers forming a multiplex network with eight distinct layers as depicted in Fig.VII.12. We indeed can now confirm several hypotheses. First of all we strikingly discover that in case of the Tbet^{+/+}Th1 conditions we find CSCs mainly appearing on the Th1 side of the network while in the Th2 control case we find the same for the CSCs on the Th2 side. For naïve cells we observe CSCs which are equally distributed over the whole network although in comparison to the full network we also find a significant level of sparsity w.r.t. CSC nodes. This does not come as a surprise since we also expect different CSCs to come into play for different conditions and in fact we see that this also implies a far larger number. We also observe a grading in the appearance of Th1-specific CSCs upon gradual removal of Tbet while at the same time the number of CSCs on the Th2 side increases. The same can be observed as well for the Th1/2 conditions w.r.t. Tbet grading while compared to the pure Th1 conditions the number of CSCs around the Th2 part is significantly higher.

Being accompanied by the (dis)appearance of certain CSCs in different conditions is also the change in edge connectivity in certain regions of the underlying network. In distinction between naïve and Th1/2 wild-type conditions where we observed a CSC distribution over the whole network range for the former we see that the connectivity mainly focuses on the lower Th2 network part for the latter with genes like STAT6, Gata3 and the Th2 cytokine cluster. This also differs from the Th2 control condition. Also for the Tbet dose graded conditions we find a connectivity shift from the Th1 part to the Th2 part with connectivities ending up on medium levels in both areas in the Tbet knock-out conditions respectively. This is in contrast to the wild-type Tbet Th1 and Th2 control conditions respectively.

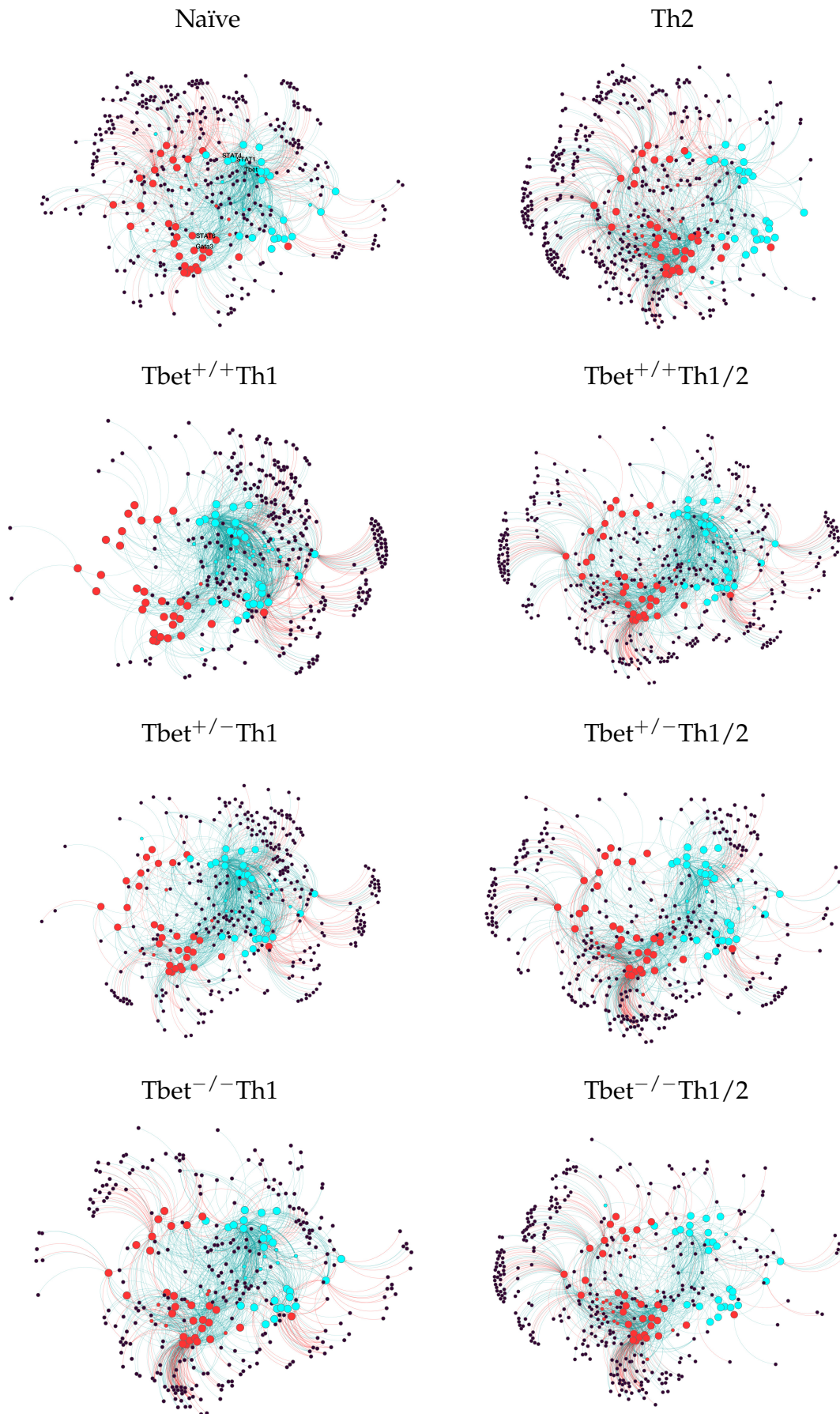


Figure VII.12: Depiction of all considered multiplex network dimensions with fixed node positions w.r.t. the full network.

We also observe certain structural similarities between $Tbet^{-/-}$ Th1/2 and Th2 network dimensions of which we already assumed the epigenetic imprint to be similar. This is now confirmed especially w.r.t. to CSCs on the Th2 side of the network. Yet we find that the connectivities in the latter case are higher in the upper Th2 cluster part. At this point we still have to investigate the topological and functional uniqueness of the hybrid cell conditions. We will start with such an undertaking by dropping the static treatment of nodes w.r.t. the full reference network.

This is due to the fact that in contrast to e.g. geographical networks where nodes are naturally bound by their location this is in general not the case for an epigenetic network where only pseudo-distances exist in contrast to physical distances. Hence we want to find out visually which subnetworks are deleted and which in turn occur. In general in a force-directed network the respective nodes in each layer might rearrange according to their connectivity in a dynamical way. For that reason we also want to investigate some newly formed sub-communities due to the newly found sparsity of the respective network layers. We exemplify this in Fig.VII.13 with the HER method for the $Tbet^{+/+}$ Th1/2 and the Th2 control networks. For the Th2 condition we observe tightly regulated modules the first two of which mainly consist exclusively of Th2 regulated or Th1 regulated genes respectively. Since the Th2-specific part of the subnetwork is especially strongly regulated we also obtain several modules consisting only of Th2-specific nodes as can be expected. The reason for this is that the stronger regulation of this part of the subnetwork also leads to more distinction in its fine-graining structure indicating that *STAT6*, *Gata3* and *Il10* all have their own tightly regulated substructure compared to most neighbouring genes within the full network context. For the hybrid Th1/2 condition on the other hand we most notably find a large co-regulation cluster of Th1- and Th2-specific genes as well as CSCs. We not only find important Th1 genes like *Tbet*, *STAT1* or *STAT4* but also important Th2 genes like *Gata3* or *STAT6*. As well we observe new strong interconnected sub-clusters where e.g. *Gata3* and *STAT6* bind exceedingly strongly to *Runx3*. In addition to that a separate module emerges in which *Eomes* is attached to a whole cluster of Th2-specific genes. This new clustering is especially due to the network reordering, which can be easily observed if we do not keep the nodes fixed relative to a reference network but again apply a force-directed algorithm e.g. to the Th1/2 subnetwork. This is shown in Fig.B.20. We observe that in comparison to Fig.VII.12 certain nodes are now more closely connected than before resulting in the above clustering.

In conclusion the analysis of multiplex networks results in the discovery of expected as well as new substructure and topologies. First of all since certain CSCs are not only Th1- or Th2-specific but strongly condition-specific in general we recover this fact also in the network substructure w.r.t. different layer dimensions. When shifting from Th1 to Th2 conditions we hence not only observe a gradual shift from CSCs appearing in the Th1 part of the network to those appearing in the Th2 part but we also find a gradual change in edge connectivity. This can be observed as well for a graded *Tbet* dose dependency for neutral as well as for Th2 polarizing conditions. This leads to the suggestion that the removal of *Tbet* to a certain level results in mixed cell conditions where we observe the most notable similarities between the network topologies of the $Tbet^{-/-}$ Th1/2 and the Th2 control condition. Finally we find that each multilayer exhibits a unique topology and also unique network properties leading to a completely new network structure for each instance and as well resulting in

new subcommunities within each network layer. Hence the structural and functional connectivity between certain genes changes all the time whereas other gene clusters remain mostly unperturbed from these inter-layer differences.

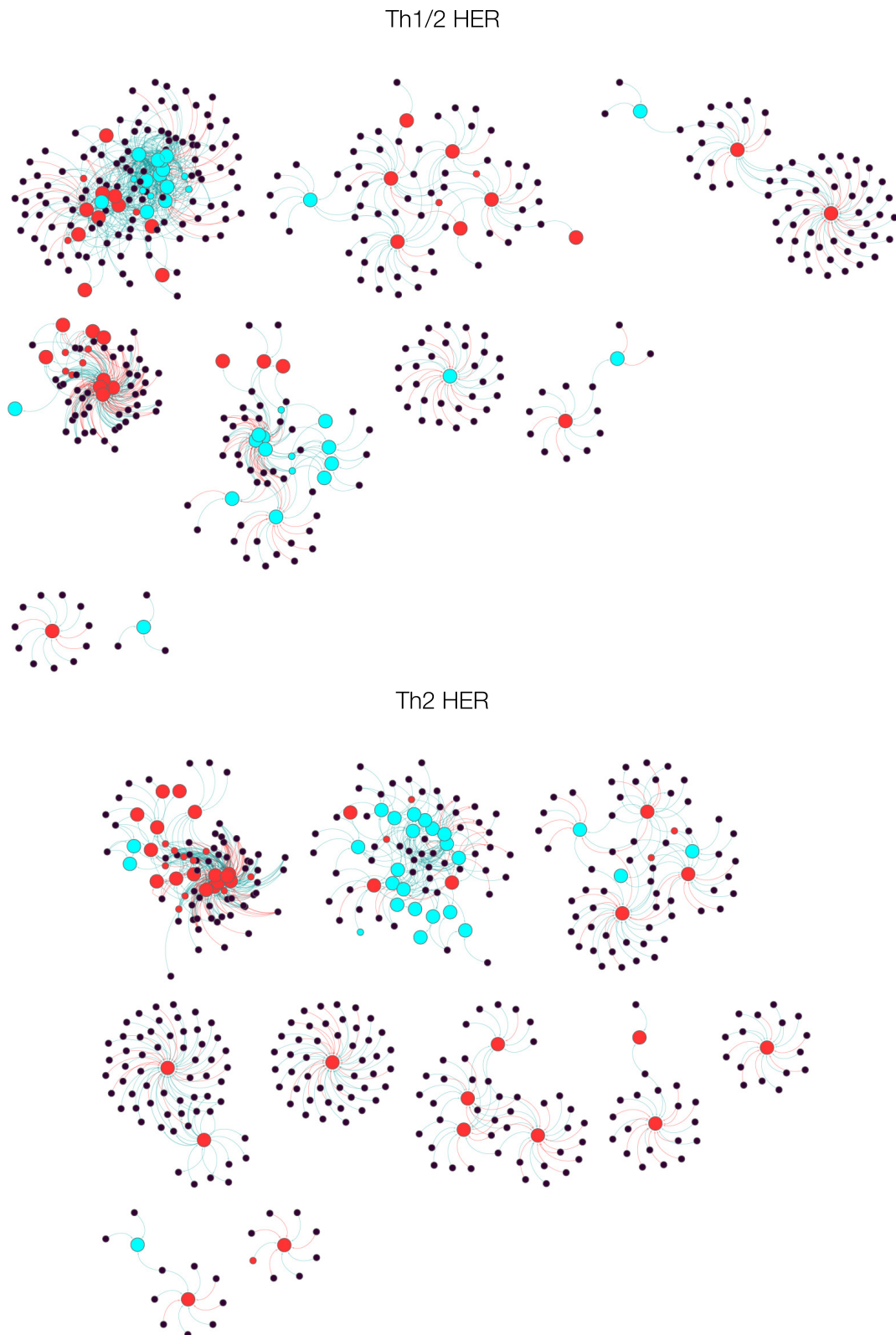


Figure VII.13: Resulting community structures according to the HER method for two exemplary multiplex layers, i.e. $Tbet^{+/+}$ Th1/2 and Th2 control.

In order to investigate the inter-layer changes even more closely we will now turn to the differential analysis of the layers of the multiplex network.

VII.2.7 Differential network analysis

Another way of looking at multiplex networks providing also features in a far easier approach compared to weighted multilayer networks is to investigate differential regulation between pairwise network dimensions. If we do not allow for the existence of interdimensional edges in multilayer networks between different node instances but only for node bijectivity we get analogous results by considering up- and downregulation of edge weights. We consider this in particular for Th2 control vs. $Tbet^{+/+}$ Th1 as well as for $Tbet^{-/-}$ Th1 vs. $Tbet^{+/+}$ Th1, $Tbet^{+/+}$ Th1/2 vs. $Tbet^{+/+}$ Th1 (*Diff1*) and $Tbet^{+/+}$ Th1/2 vs. Th2 control (*Diff2*). In addition we always depict the betweenness centralities of all networks at all nodes. The results are shown in Figs. VII.14–VII.17. Up- and down-regulation are shown in dark-green and red respectively and the nodes are fixed in relation to the full network as before.

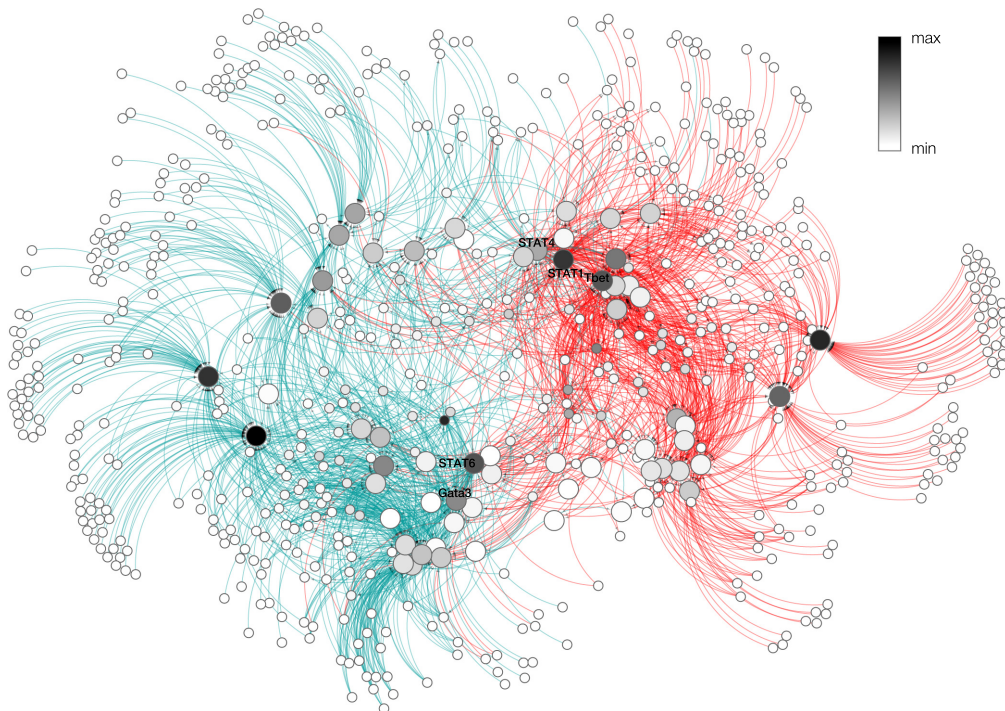


Figure VII.14: Differential network of Th2 control w.r.t. $Tbet^{+/+}$ Th1. The shades of grey of the nodes represent different betweenness centralities while green edges denote upregulation and red edges denote downregulation.

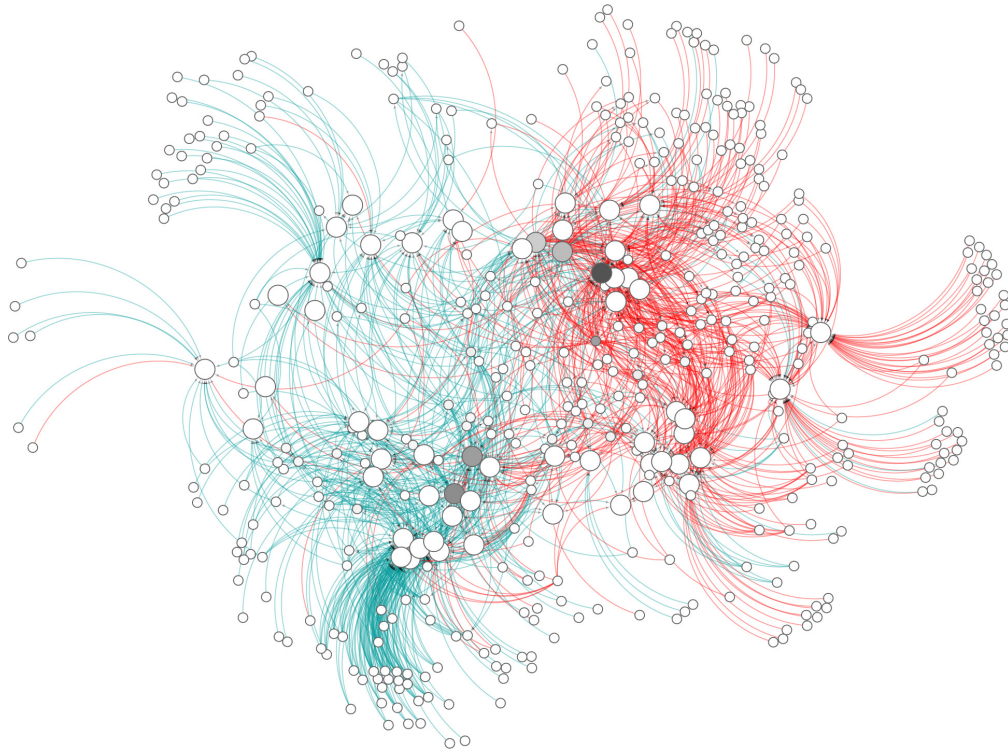


Figure VII.15: Differential network of $Tbet^{-/-}$ Th1 vs. $Tbet^{+/+}$ Th1. The shades of grey of the nodes represent different betweenness centralities while green edges denote upregulation and red edges denote downregulation.



Figure VII.16: Differential network of $Tbet^{+/+}$ Th1/2 vs. $Tbet^{+/+}$ Th1 (*Diff1*). The shades of grey of the nodes represent different betweenness centralities while green edges denote upregulation and red edges denote downregulation.

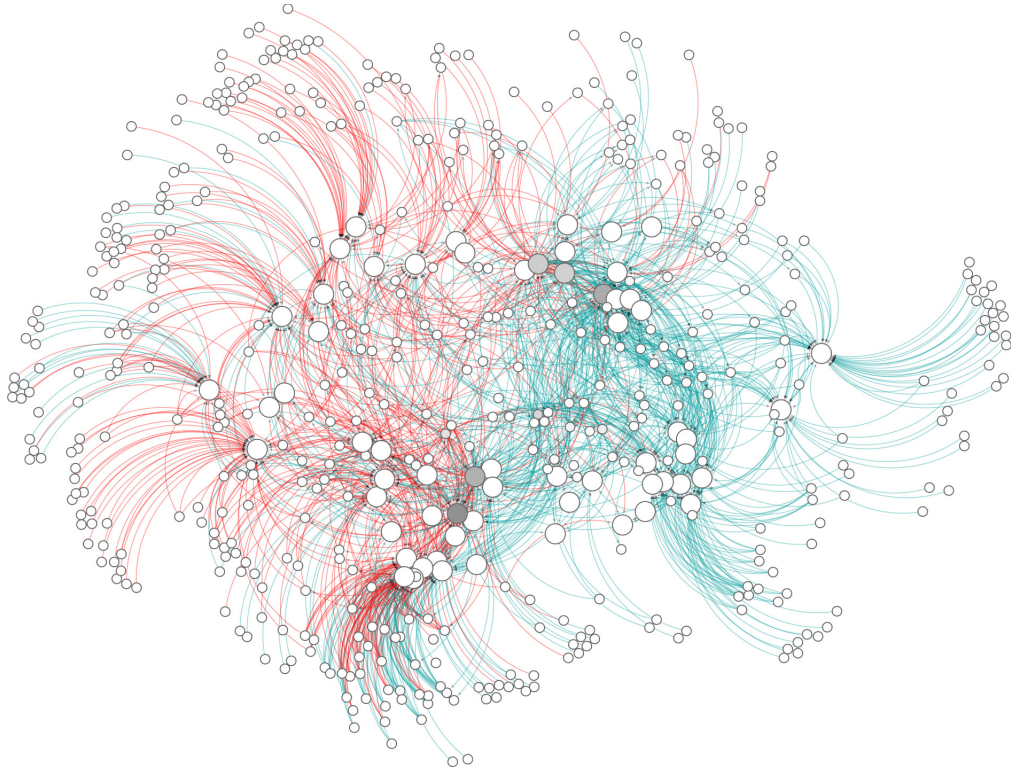


Figure VII.17: Differential network of $Tbet^{+/+}$ Th1/2 vs. Th2 control (*Diff2*). The shades of grey of the nodes represent different betweenness centralities while green edges denote upregulation and red edges denote downregulation.

For the direct Th2 vs. Th1 comparison in Fig. VII.14 we can confirm certain expected processes such as up-regulation of general connections associated with Th2-specific features and down-regulation of Th1-specific features. We also find a high betweenness centrality of a large amount of Th1- and Th2-specific genes. The highest betweenness centralities are listed for the up- and down-regulation processes independently in table C.25. This again confirms our previous hypotheses and is a valuable consistency check.

For the case of Tbet knock-out in Th1 cells we most notably find that genes as well as CSCs in the Th1 part of the network are heavily down-regulated w.r.t. wild-type Tbet cells. On the other hand Th2-specific genes and CSC experience slight up-regulation. We also find strong down-regulation of enhancers which are in particular sensitive to Tbet, which is another important consistency check. Of utmost interest is also the finding that the TFs within the network experience the highest betweenness centrality which in turn means that they are the ones suffering most from Tbet knock-out. This holds quite trivially true for Tbet itself since it is removed from the knock-out network, yet the levels of all other TFs are of comparable order.

More interestingly we try to elucidate the differences between hybrid Th1/2 cells and classic Th1 and Th2 cells respectively. As before we find the highest betweenness centralities for the respective TFs, yet we also find in *Diff1* as well as in *Diff2* that there is strong activating regulation by CSCs specific to the Th1/2 condition, which are activating Th1-specific genes in the *Diff2* case and Th2-specific genes in the *Diff1* case. In comparing *Diff1* and *Diff2* themselves we also find that certain CSCs are rather unique to regulation when going from Th1 to Th1/2 or from Th2 to Th1/2. For example we see in the Th2-specific network part that in the case of *Diff1* other regulating CSCs

come into play than in the *Diff2* case where mainly the down-regulation of the whole upper part is relevant. On the other hand in the *Diff1* case mainly up-regulation of the most distant part in comparison to the Th1-specific network part (especially w.r.t. *Tbet*, *STAT1* and *STAT4*) is of relevance. The same line of argument also holds true for the Th1-specific network part.

Topological changes can be generally inferred by either investigating by how much a certain edge changes between conditions such that some motifs contribute more strongly or more weakly for that matter or one just investigates the topological changes that are introduced via edge deletion or addition. If we do that for the *Diff1* network we find the differential edge addition and deletion networks in Fig.B.21 with the node shades now depicting the in-degree centrality. We find that there is barely any difference concerning edge connectivity to be found compared to the earlier analysis from Fig.VII.16 hence we conclude that in fact an exceedingly large number of edges is respectively added and deleted in between the different network layers, and do not merely exhibit some minor changes in their respective weights.

The fundamental question to be asked at this point is which nodes are most important in distinguishing Th1/2 hybrid cells from classic Th1 or Th2 cells and if we can infer predictions on stability in between these three conditions. In order to do this we have to investigate which nodes characterize the new conditions best. Obviously this includes the respective condition-specific CSCs occurring as new nodes in one network dimension relative to another one, yet this is trivial. The degree to which they can achieve this is certainly determined by their respective connectivity to other genes. Hence we can just investigate the degree centralities of all nodes which are differentially regulated in the *Diff1* and *Diff2* networks and rank them. The respective rankings yet differ completely for the in- and out-degrees respectively since for the in-degrees we mostly find relevant genes (i.e. nodes that are being regulated) while for the out-degrees we find mostly CSCs (i.e. nodes that actively regulate). Such a ranking is already indicated in Fig.B.21 and additionally listed in table C.26. Up to this point we have already evaluated to some extent the importance of certain CSCs to the full network yet we now want to obtain a robust ranking method for node importance. We have also seen that certain shortcomings of the in-degree centrality are resolved by the eigenvector as well as the Katz centralities, yet we still end up with the problem of certain nodes receiving a high centrality ranking just by the fact of linking to other nodes with high centrality. Although this takes into account that a certain node has to have linkers with high centrality it might not be desirable that if that linker has a high out-degree that all linked nodes obtain high centrality themselves. In the following we will discuss an alternative approach to infer node relevance w.r.t. to cell-condition-specificity.

VII.3 Random walks on weighted multiplex multigraphs

An intricate and revealing way to investigate node importance and especially a shift in node importance in multiplex networks is by considering random walks on these networks via so-called Markov chains (see e.g. [241]). This is also closely related to the determination of cluster centers in spectral clustering [87, 321] and to the investigation of metastable network states considering the so-called *spectral gap* [47, 321]. Further information on the topic of random walks on networks can be e.g. found in [211, 222, 354, 366].

From the respective adjacency matrices \mathcal{A}_{ij}^d for the different dimensions of the multiplex network we want to determine in a next step the corresponding transition matrices with elements \mathcal{P}_{ij}^d . We call them *raw multiplex transition matrices* for the time being. We will see in due course why this is important. Since we are considering a weighted directed multidigraph we consider the following corresponding general relation for the elements of a one-step raw transition matrix

$$\mathcal{P}_{ij} = k_{\text{out},i}^{-1} \cdot |\mathcal{W}_{ij}|. \quad (\text{VII.7})$$

transitioning from node i to node j via a probability being determined by the absolute weight between those nodes and being normalized by the weighted out-degree of node i . More precisely this defines a time-discrete time-forward random walk. A random walk of N steps starting at i and ending at j is now obtained by N multiplications of the transition matrix with itself, hence computing \mathcal{P}_{ij}^N . Extending this to $N \rightarrow \infty$ we are interested in obtaining the steady state distribution of the probabilities that the random walker is found at a certain node. This can be solved as an eigenvalue problem of the respective transition matrix. Due to the fact that the underlying transition matrices are non-symmetric we get a set of real as well as complex eigenvalues. We note furthermore that due to the differences in in- and out-degree of each node we get in general different eigenvalue spectra for the time-forward and its corresponding backward-time transition matrix. This is a notable feature of directed networks and Markov chains in general (see e.g. [70, 120, 241]).¹⁷ The so-called *Perron-Frobenius theorem for ergodic Markov chains* (see e.g. [201, 223] as well as section A) furthermore states that a stochastic irreducible transition matrix with elements \mathcal{P}_{ij} has a unique eigenvalue $\lambda_1 = 1$ providing an upper bound for the set of all eigenvalues. This maximum eigenvalue is often called the *Perron root*. This implies a transformation via the transition matrix under which the corresponding eigenvector π_j does not change:

$$\pi_j^T \mathcal{P}_{ij} = \pi_j^T. \quad (\text{VII.8})$$

The left eigenvector π^T is hence generally called the stationary distribution of \mathcal{P} , since it is an invariant probability measure of the Markov chain. The stationary distribution is a unique characteristic of a Markov chain and in our case of a specific

¹⁷We note that one of the usual approaches to resolve this is to just transform the directed to an undirected graph, yet this is in general not permissible and yields different results since the random walker becomes time-reversible. We investigate only the time-forward component since we want to obtain the relevance of each node by considering flow distribution.

network not only assigning a steady-state distribution to the probability of finding a random walker at node j but also assigning this steady-state distribution to every step in the Markov process, hence being a robust analytical estimator for the frequency a random walker is found at a certain node. This means we consider the Markov chain to be time-homogeneous. We can interpret this frequency as an importance rank measure for every node in the network. In the case where such a stationary distribution exists we are able to rank all network nodes time-independently w.r.t. to their respective stochastic importance within the network [47].

Yet if we determine plainly the raw transition matrix \mathcal{P} for the full network as well as for the different dimensions of the multiplex network we run into problems quickly. First of all the naïve construction of the raw transition matrix does not yield a stochastic transition matrix. This is due to the fact that there are nodes in the network with $k_{\text{out}} = 0$ leading to rows where all entries are zero. This obviously opposes the definition of a stochastic matrix (see e.g. [200]).

In addition to that the raw transition matrix is reducible, which means it can be brought into a triangular form. This is due to the fact that there exist sinks in the underlying network hence a random walker can get trapped in a certain state/node, which means that also the respective Markov chain is reducible. This is clearly an unwanted feature and has to be accounted for. This reducibility is on the one hand a result from nodes having the property $k_{\text{out}} = 0$ but on the other hand as well from those with $k_{\text{in}} = 0$. Stated differently: a Markov chain is irreducible if any node can be reached from every other node in the network, for which the following statement is necessary:

$$\forall i : (k_{\text{in},i} \neq 0 \quad \wedge \quad k_{\text{out},i} \neq 0)$$

Both of the above statements independently lead to the non-applicability of the Perron-Frobenius theorem mentioned before and subsequently prohibit the existence of a stationary distribution. Hence we need a resolution for these issues.

Stochasticity of \mathcal{P} can be quite easily ensured if we replace each row with all zeros in the raw transition matrix with

$$\frac{\mathbf{1}^T}{N} \tag{VII.9}$$

where $\mathbf{1}^T$ is a row vector of all ones and N is the order of the raw transition matrix hence providing a proper normalization. We call the resulting stochastic transition matrix $\tilde{\mathcal{P}}$.

At this point it might still be the case that certain nodes as well as small subnetworks are only connected via a unidirectional edge s.t. sinks still exist. This is obviously true in our case since we still consider nodes with $k_{\text{in}} = 0$. This contradicts the requirements that have to be met for irreducibility. Hence some regularization of $\tilde{\mathcal{P}}_{ij}$ has to be employed to ensure irreducibility. There exists a very prominent solution to this specific problem which forms the basis of the popular PageRank algorithm (see e.g. [53, 200])¹⁸. The basic solution is to randomly switch to another node by

¹⁸PageRank among other methods measures the relative importance of webpages within the Google search routine [53].

introducing stochastic perturbations into the system. In our case this accounts for the incompleteness of our TF binding data, since we only included TF data of five of the considered 64 genes. This makes it possible to say that there is always a low but nevertheless existing probability that a pair of nodes is connected via a directed edge. This also introduces possible residual binding events between CSCs directly. The same obviously holds true for genes¹⁹. This perturbation matrix is now added to the stochastic matrix with elements $\tilde{\mathcal{P}}_{ij}$. The irreducible stochastic matrix $\tilde{\tilde{\mathcal{P}}}$ is now obtained via the convex convolution

$$\tilde{\tilde{\mathcal{P}}}_{ij} = \alpha \tilde{\mathcal{P}}_{ij} + (1 - \alpha) \epsilon_{ij}. \quad (\text{VII.10})$$

in accordance with [200]. Here α is a scalar prefactor determining the mixing probability of the stochastic transition matrix with the perturbation matrix ϵ with $\alpha \in [0, 1]$. In the simplest case the perturbation matrix is just a matrix of all ones normalized by the order N of $\tilde{\mathcal{P}}$: $\epsilon = \frac{\mathbf{1}\mathbf{1}^T}{N}$. A more advanced method replaces $\frac{\mathbf{1}\mathbf{1}^T}{N}$ by a customizable probability vector \mathbf{v}^T where weights can be respectively assigned to each transition separately²⁰. In the following we do not want to introduce any bias on stochastic switching probabilities and assume the former definition. Additionally we set $\alpha = 0.85$ which is the consensus literature value (see e.g. [201]) among a large tested parameter range.

The resulting stochastic irreducible matrix with elements $\tilde{\tilde{\mathcal{P}}}_{ij}$ hence fulfills the requirements set by the *Perron-Frobenius theorem* by construction as we now indeed consider ergodic Markov chains. Hence we have to obtain a maximal eigenvalue of one and because of that a stationary distribution π_j exists, which is the corresponding eigenvector to $\lambda_1 = 1$. We determine this eigenvector for the full network as well as exemplary for the wild-type Th1 and Th1/2 multiplex network dimensions after determining their respective stochastic irreducible matrices separately and list the highest rank-ordered result in table C.27. The stationary distribution is also depicted with node colour-labelling in Fig. VII.18 for the full network. We find that in general gene nodes are ranked highest which is due to the fact that they receive in most cases the largest amount of incoming edge weights compared to CSCs. That means that more nodes are pointing in general to a certain gene than to a certain CSC hence ranking them higher. For the full network we find a mixture of highly ranked Th1 and Th2 genes whereas for the Th1 network layer Th1 genes are dominant as can be expected and for the Th1/2 condition we again find a mixture of Th1 and Th2 genes. At this point this is not much of a revelation. Yet for the CSCs the results naturally begin to differ. The full network is now very much dominated by CSCs which are dominant in Th1 as well as in Th2 cells, but also by unspecific CSCs such as ESCs which are constitutively inactive or only exhibit active enhancers in naïve conditions. For the Th1 network we find e.g. high relevance of ESCs which show active enhancer states in Tbet^{+/+}Th1 conditions and those which in addition exhibit

¹⁹We can think of this as direct binding of a TF to a gene promoter. In the case of genes not acting as TFs this would implicate indirect binding via mediating TFs, which are dressed in such a connection.

²⁰In contrast to introducing a stochastic switching matrix perturbation between all nodes in the network there are also weaker irreducibility assumptions that can be introduced yet they always lead to biases on preferential stochastic switching, which we do not consider (see [200]).

active enhancers in $Tbet^{+/-}$ Th1 conditions. In the case of Th1/2 hybrid conditions we now obtain a condition-specific ranking as well for the CSCs. In this case we find the leading CSCs to be ESCs, which are constitutively active over all conditions, closely followed by ESCs, which are active only in Th1/2 wild-type conditions, as can be expected.

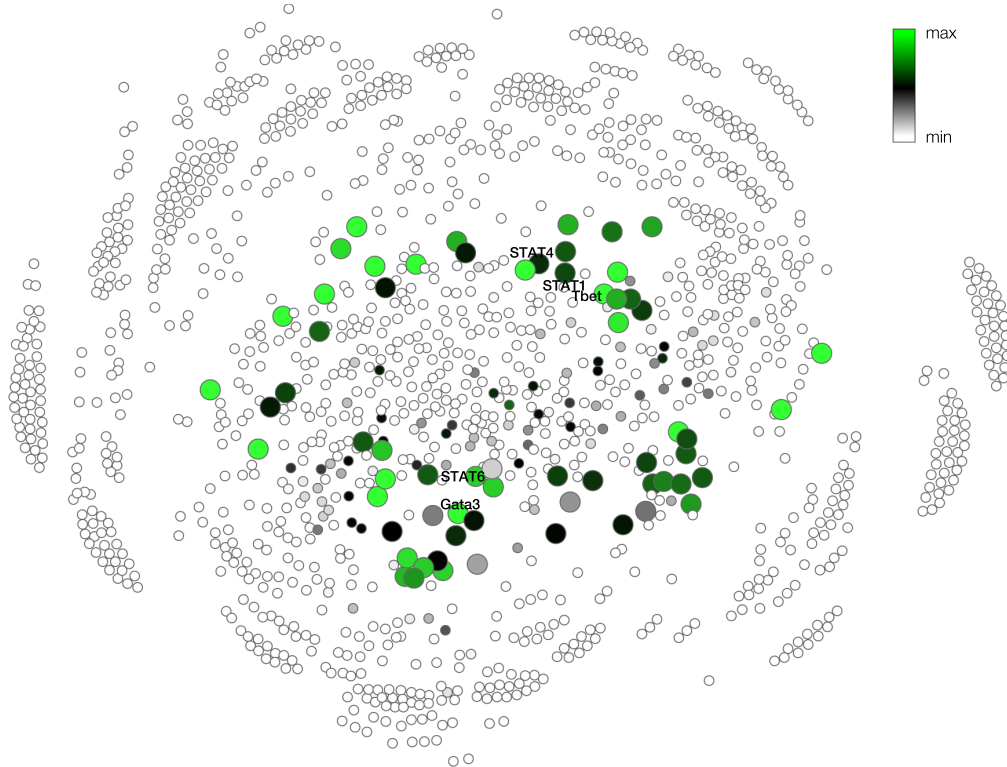


Figure VII.18: Stationary distribution of nodes in the full CSC-gene network with fixed node positions compared to earlier depictions. The stationary distribution probability is encoded in the colour-labelling of the nodes.

Since the gene ranking of the multilayer dimensions shows that in some cases only slight changes w.r.t. ranking position of a certain gene can be observed we apply the above method to the differentiated networks as well, which we obtained in the last section in order to obtain more distinguished results. In this case we additionally discriminate between up- and down-regulation of the respective networks. The result is exemplified for the *Diff1* and *Diff2* networks and can be found in table C.28. We naturally find a large difference between up- and down-regulation of node importance as well as certain notable differences even between down-regulation of *Diff1* and up-regulation of *Diff2* as well as for up-regulation of *Diff1* and down-regulation of *Diff2*²¹. While for the up-regulation of *Diff1* we naturally obtain a high-ranking of mostly Th2 genes for *Diff2* we actually obtain a mixture of Th1 and Th2 genes the latter of which is due to heavy inhibition of the respective Th2 genes. Considering down-regulation we find for *Diff2* mainly Th2 genes, yet among the top-ranked ones we do not find either *Gata3* or *STAT6* but rather candidates like *Lrrc32*, which is thought to play a key role in regulatory T-helper cells [266] or *Cyp11a1*, which has been associated with phenotypic maintenance in Th2 cells before [243]. Accordingly we find

²¹We compare these two pairings respectively since in one case we approach the Th1/2 hybrid network from the Th1 side and in the other case from the Th2 side.

for the *Diff1* network w.r.t. downregulation that the main focus is on *Ifn γ* as well as on genes like *Smpdl3b*, which are associated with inhibition of TLR signalling [146], or *Ccl5* followed by several others most of which also play key roles in the wild-type Th1 subnetwork. When cross-comparing up- and down-regulation between *Diff1* and *Diff2* we also obtain slightly different rankings for the Th1 and Th2 genes respectively suggesting that a different set of cell-specific genes are up-regulated when observing a Th1 \rightarrow Th2 transition than are down-regulated when going from Th2 to Th1. In the former case this mainly involves *Gata3*, *Il4*, *Il5*, *Il10* and *Il13* while in the latter a completely different set of Th2 genes is down-regulated. To some extent the same holds true for Th1 specific genes.

For the regulation of CSCs we compare both up-regulation results where we find that also different Th1/2-unique CSCs are switched on depending on which differential transition is observed. For *Diff1* this mainly includes ESCs which exhibit active enhancers in the Th2 condition, while for *Diff2* we do not only find ESCs with active enhancers in the Th1 wild-type condition but rather more which contain active enhancers in the Th1/2 conditions but not in any Th1 condition.

We observe that with such a rank-ordering we not only introduce the possibility to obtain a condition-specific importance assignment for CSCs as we have shown before with the ERT method in section VI.3 but we now also include genes and can even compare in between these entities. Furthermore we are not anymore restricted to a particular fully differentiated condition like a Th1 or a Th2 cell but are able to use any condition for which we can construct a unique network. This naturally can be applied to any arbitrary treatment condition. Although the node influence in some cases is hard to compare for different conditions with slightly increasing or decreasing importance with respect to another node we find that if we consider up- and down-regulation for differential networks of different network dimensions we can elucidate more clearly what impact a certain node obtains from one condition compared to another. Compared to a mere in-degree or eigenvector centrality analysis we obtain a more robust ranking of node importances also taking into account that inputs from other nodes are diluted over all outgoing edges of all of these other nodes. This means that e.g. a certain gene is less weighted in its importance if a certain CSC, pointing to many genes, points to this particular gene. Although several of these highly ranked candidates provide hints of specificity w.r.t. a certain treatment condition the only tendency for driving hybrid Th1/2 cell conditions can be obtained via analysis of differential regulation of the respective network dimensions. In order to obtain some insight into the inference of possible multistable network states we have to consider possible dynamics of the underlying long-lived steady-state conditions we considered up to this point merely via the stationary distribution of the Markov process. We will extend this now as well to lower order steady state processes.

Features of the eigenvalue spectra of stochastic transition matrices

In the above discussion we only observed the largest eigenvalue of the underlying stochastic irreducible transition matrices. Obviously we get a full spectrum of eigenvalues λ_i for $i = 1, \dots, N$ with N being the respective order of each network's adjacency matrix. Of special interest e.g. for methods like spectral clustering are the leading eigenvalues of which their corresponding eigenvectors cluster the nodes of the network representing a certain microstate into larger macrostates or modules (see

e.g. [87]). This was already implicitly applied in section VII.2.5 for a fixed number of macrostates or clusters, defined by that amount of leading eigenvalues. One can identify a lifetime-scale with every eigenvalue according to

$$t_{L,k} = -\frac{t}{\ln|\lambda_k|} \propto -\frac{1}{\ln|\lambda_k|} \quad (\text{VII.11})$$

where t denotes the lag time of the Markov process (see [47]). The approach to a stationary process for $\lambda_1 = 1$ hence corresponds to an infinite lifetime while other modules are rather finite and hence called *metastable*. Depending on the size of the eigenvalue we obtain a spectrum of lifetimes defining submodules which are characterized by the corresponding eigenvectors. From the point of view of the stationary distribution the other eigenvalues can be interpreted as perturbations which, depending on their respective lifetime, die out at some point in time. A particular interesting aspect hence is to try to separate the eigenvalue spectrum into eigenvalues resulting in fast and slow processes. A separation of such time scales is marked by a so-called *spectral gap*. One can now investigate the eigenvalue spectrum and infer the C largest eigenvalues including λ_1 for which we find $\lambda_C \gg \lambda_{C+n}$ for $n \geq 1$. The set of these largest eigenvalues is called *Perron cluster* the amount of which determines the number of the (relatively long-lived) metastable states of the system (see e.g. [47]).

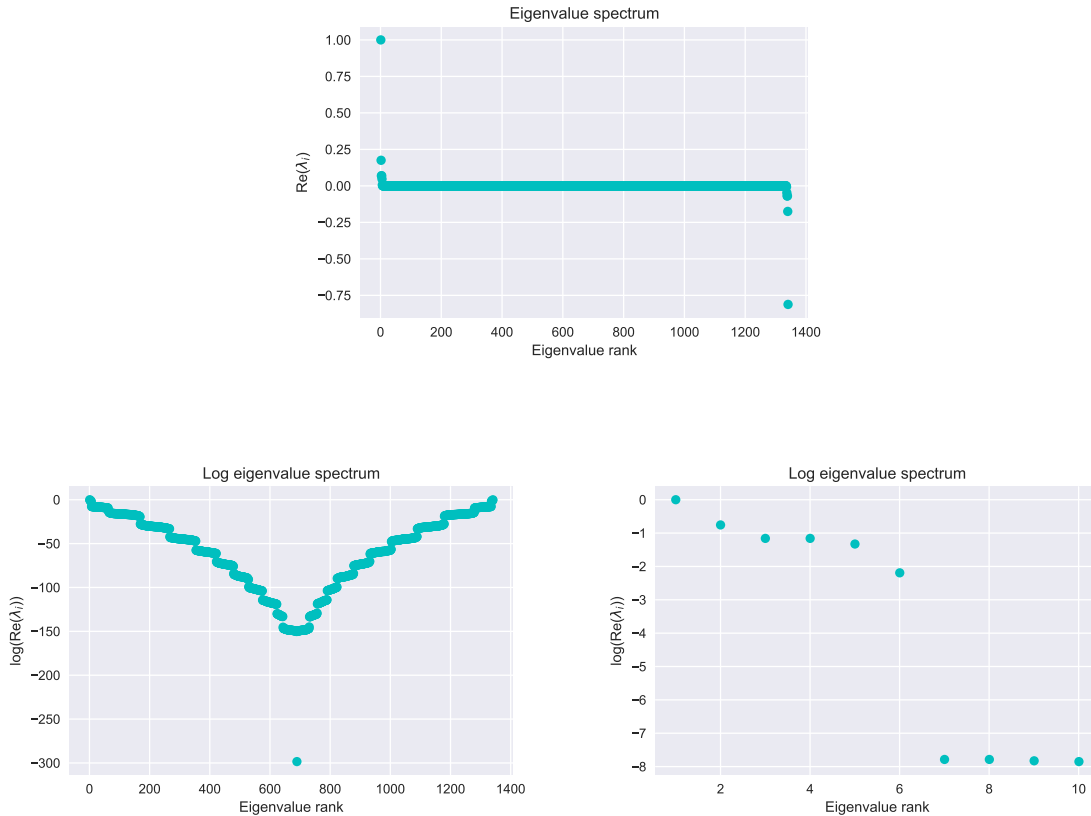


Figure VII.19: Eigenvalue spectrum of the full CSC-gene network (top). For detailed investigation of the spectral gap we also depict the logarithmic real absolute eigenvalues (bottom left) as well as a close-up on the leading values (bottom right).

The eigenvalue spectrum of the full network is depicted in Fig. VII.19. A first separation of timescales can be observed directly after λ_1 where the drop in the spectrum

is about a factor of five. Yet if we observe the absolute real parts of the spectrum on a logarithmic scale we find that there is a considerable drop after the sixth eigenvalue corresponding to more than five orders of magnitude. Hence although the time-scales of the metastable states do not even approach a stationary distribution we obtain a small class of eigenvectors, which exhibits lifetimes on a considerably larger scale than all other eigenvectors. This is obviously due to the high connectivity within the network itself s.t. most metastable states can diffuse quickly. The second and third eigenvectors are depicted in Fig. VII.20 to exemplify the underlying structure.

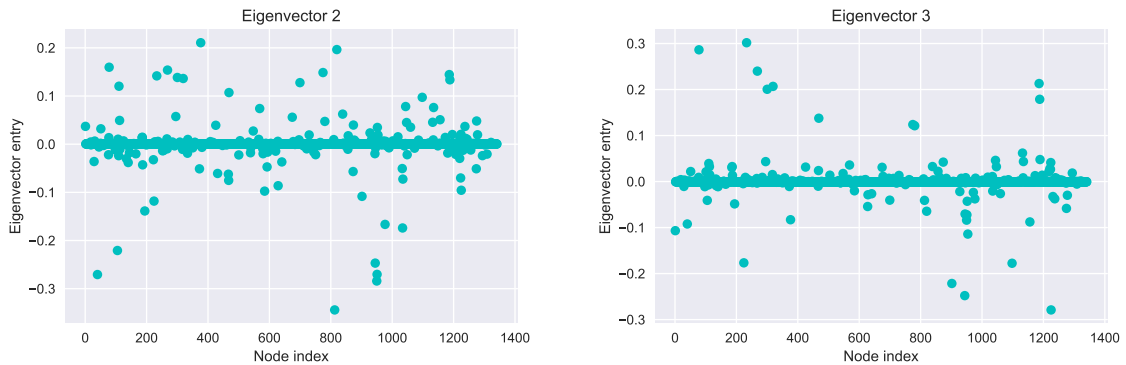


Figure VII.20: Visualization of all eigenvector components of the second and third eigenvector of the full network. Even at this level one can clearly see the emergence of similar behaviour of certain network nodes depending on the respective eigenvector.

Not only do we observe changes in absolute values but also w.r.t. their sign with a certain eigenvector and for different eigenvector entries of the same network node. These sign changes between different nodes indicate an exchange of flow for a random walker between these nodes (see [47, 69]) and hence we see that different eigenvectors associated with different timescales are responsible for the exchange of clusters of nodes, the number of which is defined by the number of dominant eigenvalues C^{22}

In fact this can be derived from an analogue from transfer operator theory. It turns out that the stochastic irreducible transition matrix is a discrete approximation of a continuous transfer operator. At the same time the eigenvectors correspond to the respective eigenfunctions in the continuous case. Eigenvectors and eigenfunctions can be mapped onto each other by Fourier expansions, hence the entries in the eigenvectors correspond to the respective Fourier coefficients expressed in the orthonormal basis of in our case nodes within the network, i.e. genes and CSCs. From this point of view we see that we have opposing contributions to the discrete analogue of the eigenfunction namely the eigenvector as an approximation depending on the respective modes (for further details on this viewpoint see e.g. [161, 188, 265]).

Accordingly different processes, encoded in different eigenvectors, are contributing on different time-scales.

²²In fact certain types of spectral cluster analyses like the Perron cluster analysis (PCCA) employ the nature of the sign changes in the eigenvectors for clustering nodes. Respectively splits of clusters are performed for each leading eigenvector subsequently by dividing nodes according to them having positive or negative eigenvector components (see e.g. [47]).

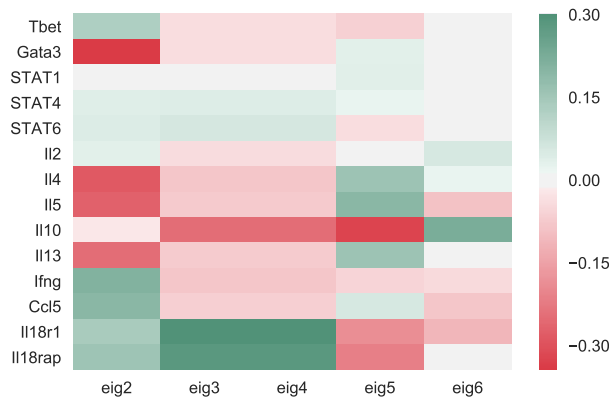


Figure VII.21: Heatmap of the leading eigenvector components (from λ_2 to λ_6 , denoted as eig2 to eig6) of the full network of notable Th1 and Th2 genes.

In Fig. VII.21 we depict the eigenvector entries for a set of selected genes. For the second eigenvector λ_2 we observe significant deviations from the stationary distribution where especially a significant shift in node importance is observed for *Gata3*, *Il4*, *Il5* and *Ifng* among others. More specifically we find an emergent separation between genes like *Tbet* and *Ifng*, being up-regulated w.r.t. node importance and *Gata3*, *Il4* and *Il5* being down-regulated²³. This occurrence of different signs within the higher order eigenvectors hence leads to a node importance flow or shift between these genes. We find that this signature hence approximates the Th1- and Th2-specificity. Going to even lower metastable processes with corresponding eigenvalues we also find similar Th1- and Th2-specific behaviour for λ_5 . Concerning λ_2 we also observe some mixing between Th1 and Th2 genes like in the case of *STAT4* and *STAT6*. Additionally we find processes with mixed gene importance for λ_3 and λ_4 which for the genes under consideration look the same. Hence at this point we can recover hints at hybrid Th1/2 processes.

We already see at this level that metastable processes appear on finite timescales which either show separation of Th1 and Th2 genes or already correspond to a mixing of different Th1- and Th2- nodes within the network, the extent of which can be readily determined via a spectral eigenvalue decomposition. We are now able to make additional predictions via the assessment of perturbations in the importance of genes that might also reflect subsequently in their gene expression profiles. Via this approach we infer a respective contribution of CSCs to these metastable network states.

If we apply this evaluation to multiplex network dimensions we also find hints for the existence of long-lived metastability. This is exemplified for the wild-type Th1/2 hybrid condition in Fig. VII.22.

We observe a large drop in the eigenvalue spectrum after λ_7 . In Fig. VII.23 we again show the eigenvector entries for the genes analyzed before. For λ_2 we find metastable up-regulation of *Gata3* and Th2-specific interleukins at the cost of the Th1-specific genes as well as *STAT6*, hence mostly recovering the Th2 part of the network.

²³We note at this point that the choice of the sign of the eigenvector is arbitrary and hence the denominations of up- and downregulation in this context are interchangeable.

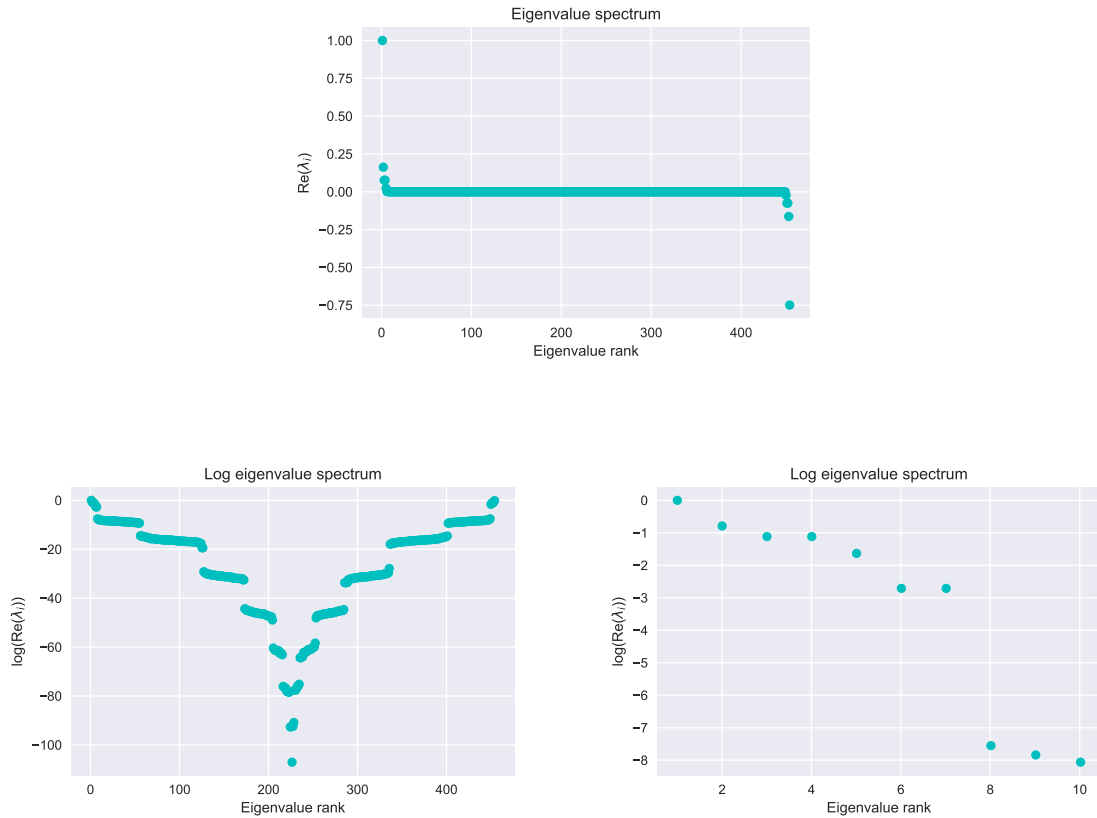


Figure VII.22: Eigenvalue spectrum of the $Tbet^{+/+}Th1/2$ multiplex dimension (top). For detailed investigation of the spectral gap we also depict the logarithmic real absolute eigenvalues (bottom left) as well as a close-up on the leading values (bottom right).

We find λ_3 and λ_4 as well as λ_6 and λ_7 to fulfill the same metastable processes for all genes. In the former case we find up-regulation of a subset of Th1 genes in combination with *Gata3* at the cost of decreasing importance in *STAT4* and *STAT6* the same happens at the slight cost of several Th2 interleukins. We also find a metastable state for λ_5 in which *Il10* is heavily differentially regulated w.r.t. to everything else with the other genes being either unperturbed or experiencing an importance gain in the case of the remaining interleukins.

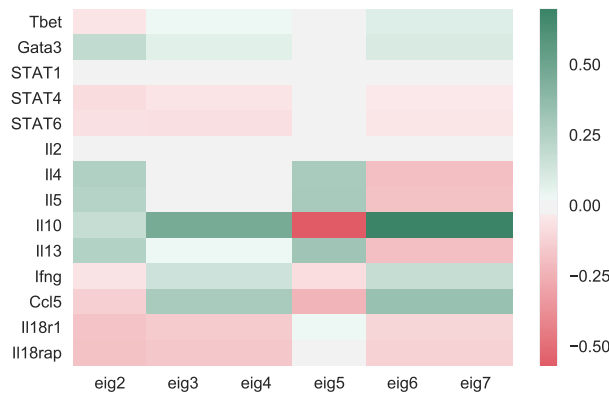


Figure VII.23: Heatmap of the leading eigenvector components (from λ_2 to λ_7) of the $Tbet^{+/+}Th1/2$ multiplex dimension of notable Th1 and Th2 genes.

The implications of this analysis are quite interesting since we find perturbations w.r.t. node regulation that emerge in metastable processes in hybrid Th1/2 cells indicating additional cell conditions as deviations from the stationary distribution of this particular long-lived steady state. Hence perturbing this network with cytokine and TF doses as indicated in the respective eigenvectors might exhibit potential power to obtain network states away from the before determined mixed phenotype. This could also be achieved by extracting the highest ranked CSCs for each metastable state and target such CSCs specifically around genes which exhibit significant deviations in the respective eigenvector as well.

Further analysis of the differential network condition makes the above statements even more pronounced. Let us for example take *Diff1*. We again distinguish between up- and down-regulation, corresponding to a flow between components of opposing sign, for both of which the logarithmic eigenvalue-spectra with the respective spectral gaps are depicted in Fig. VII.24.

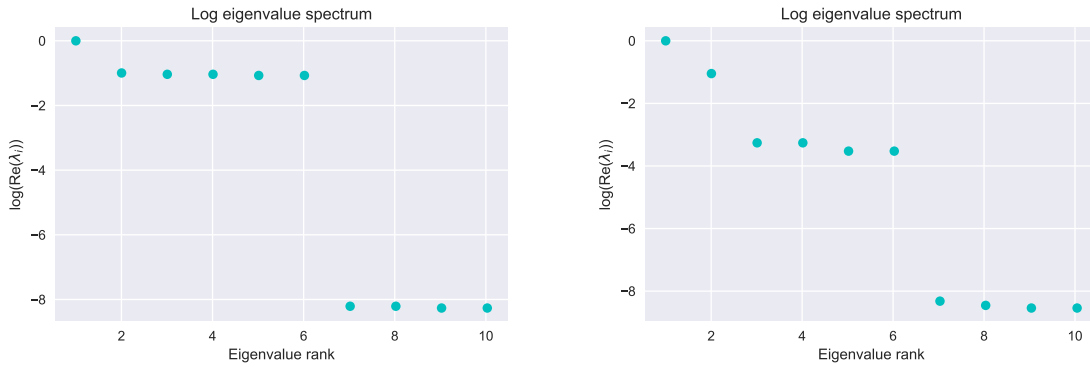


Figure VII.24: Leading logarithmic real eigenvalue parts of the upregulated section of the *Diff1* network (left) as well as of the downregulated section (right).

We find spectral gaps in both cases after λ_6 leading to the gene-specific eigenvector components shown in Fig. VII.25. In the case of the up-regulated part of *Diff1* we find that e.g. for λ_2 a metastable state emerges in between Th1 and Th1/2 that exhibits an exchange between several Th2 genes and different Th1 genes in addition to *STAT6*²⁴. Nevertheless a mixture of Th1- and Th2-specific genes coexists w.r.t. node importance. Especially the interleukin differences are even more distinct for the λ_3 and λ_4 eigenvalues, whereas no other distinctions for the genes under consideration can be observed. Furthermore for these eigenvalues especially the STATs and the master TFs do not contribute at all to the metastable process. We also observe pairwise correspondence of lower eigenvectors in the observed genes where especially λ_5 and λ_6 take the complementary part to λ_2 and the λ_3 - λ_4 -pairing for several genes. Hence we not only end up with metastable states approaching the Th1 condition more but also others which approach the Th2 condition more.

Although we find a very small spectral gap for the down-regulation network after λ_2 we also included for the sake of comparison lower order eigenvectors for the down-regulation processes of *Diff1* yet our main focus will be on λ_2 . In this case we find a separation in importance between e.g. *STAT1*, *STAT4* and *Ifn γ* . The latter also clearly promotes the transition from Th1 to Th1/2 w.r.t. down-regulation via a

²⁴We check that in this case regulation w.r.t. *STAT6* corresponds to inhibition.

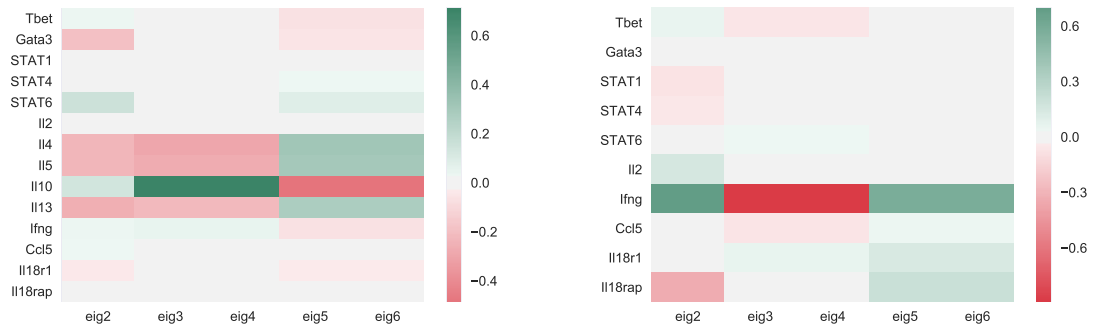


Figure VII.25: Heatmaps of the leading eigenvector components (from λ_2 to λ_7) of the *Diff1* network of notable Th1 and Th2 genes for upregulated connections (left) as well as for downregulated connections (right).

metastable state. This is pretty intuitive considering the high importance of *Ifn γ* for the Th1 cell-type being regulated by very Th1-specific CSCs most of which vanish when going to the Th1/2 condition. If we would still include even more short-lived metastable states we find that other processes come into play excluding the participation of *STAT1* and *STAT4* and rather focussing again mainly on *Ifn γ* as well as on other Th1-related genes, among which we also find *Tbx21*. Interestingly enough we observe that *Ifn γ* is among the main drivers of metastability in down-regulation processes going from a Th1 to Th1/2 steady state, which is especially interesting in considering its respective regulatory enhancer landscape w.r.t. e.g. specific enhancer knock-outs. The predictions will pose among other things will this be an important point to be considered when investigating reprogramming effects.

We see from the above analysis of random walks on the respective networks that we are able to extract metastable states via a spectral eigenvalue decomposition of the underlying epigenetic network structure. We have seen that we can obtain additional information from the corresponding eigenvectors, which is otherwise implicitly used in spectral clustering methods²⁵ and builds on similar structural principles as the PageRank algorithm. This information manifests itself in attributing changes within the flow of node importances between specific genes as well as CSCs leading to metastable processes. This can be applied to the full network as well as to unique multiplex dimensions as well as to differential processes approximating dynamical flow behaviour on the network grid. We note that we can infer unique contributions of network nodes which correspond to higher order processes that reach beyond an ordinary stationary node importance distribution and to metastable states corresponding to processes being tightly regulated on a finite time-scale. The knowledge of these processes provides further insight into functional relations being possibly responsible for a large range of steady states providing information on new phenotypes.

We will now finally investigate to what extent we can obtain new insight into the regulation of the classical Tbet/Gata3 MISA motif if we reduce the degrees of freedom of the network drastically and only include CSCs relevant for their respective regulation.

²⁵yet in this case for random walks

Metastability of the epigenetic Tbet/Gata3 network motif

For performing the eigenvalue spectral analysis of the Tbet/Gata3 MISA motif we find several possibilities. Either we can use the reduced network w.r.t. the original full network and only include Tbet, Gata3 and all connected CSCs or we reduce this even further and only include those CSC instances that are significantly correlated at the respective loci. Then we still have the possibility only to include those that in fact bind Gata3 and Tbet directly, all which have a non-zero statistical weight of binding either of the two TFs or include all significantly correlating CSC instances no matter if they bind any of the two TFs. In the following we exemplify the analysis for the full set of CSCs which have non-vanishing statistical binding occurrence of Tbet and Gata3. The result for the dominant eigenvalues is shown in Fig. VII.26 from which we deduce that only λ_2 can lead to a relatively long-lived metastable state.

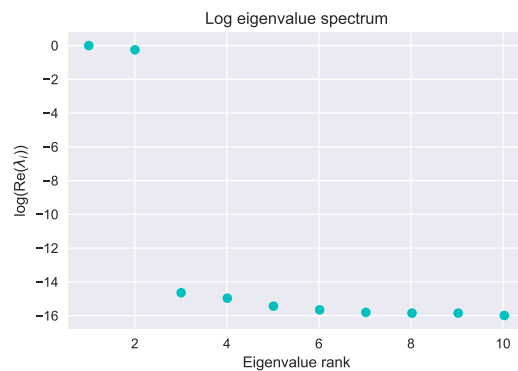


Figure VII.26: Close-up of the leading logarithmic eigenvalues of the reduced epigenetic Tbet-Gata3 network.

The largest positive and negative components of λ_2 are shown in table C.29. Determining the leading eigenvector λ_1 , which corresponds to the PageRank vector, we find a high ranking of both *Tbx21* and *Gata3* with slightly higher importance of *Gata3* since it is binding more CSCs. This is then followed by a mixture of Th1- and Th2-specific CSCs as well as by several completely unspecific instances which bind both TFs. Turning our attention to λ_2 we yet already observe a clear shift away from the stationary distribution with a majority of Th2-specific ESCs in combination with *Gata3* experiencing an upshift in node importance at the cost of Th1-specific ESCs including *Tbx21*²⁶.

Another alternative method for the analysis of metastability of the MISA motif is the direct integration of combinatorial CSC possibilities for Tbet and Gata3, representing an underlying network for combined gene regulation. For this we only consider CSC instances which directly bind Tbet and/or Gata3 at each locus. For Tbet this results in three different binding sites while for Gata3 we obtain seven different CSC instances. We do this according to Fig. V.10 and B.8²⁷. Combining all binding possibilities for both genes at the same time we obtain a total of $2^{10} = 1024$ microstates

²⁶Again because of the arbitrariness of the sign these views are interchangeable.

²⁷Since several of the respective ESCs provide inhibitory potential by (at least statistically) binding repressors in the respectively opposing cell condition we assume for simplicity an active enhancer being switched off directly or indirectly by *Gata3* or to be effectively repressed in its activity for the

and hence a total of 1024 eigenvalues of the according irreducible transition matrix. A microstate is written as

$$T_{abc}G_{defghij} \quad (\text{VII.12})$$

while all indices $\{abcdefghij\} \in \{0,1\}$ and T denotes Tbet and G denotes Gata3. The transition probabilities between each pair of microstates are determined by the simultaneous probability that every CSC instance separately is switched on or off or stays in its respective state, which means we only consider only binary decisions as opposed to ternary for simplicity. These frequency statistics are evaluated on basis of each CSC instance over all experimental conditions individually. More specifically we can take enhancer a , which is the first Tbet binding Tbet enhancer upstream of *Tbx21* in Fig.V.10, and we find that for this ESC there are the following enhancer activity²⁸ transition probabilities over all observed cellular conditions:

$$\begin{aligned} \mathcal{P}_{0 \rightarrow 0} &= \frac{2}{8} \cdot \frac{2}{8} = \frac{1}{16} \\ \mathcal{P}_{0 \rightarrow 1} &= \frac{2}{8} \cdot \frac{6}{8} = \frac{3}{16} = \mathcal{P}_{1 \rightarrow 0} \\ \mathcal{P}_{1 \rightarrow 1} &= \frac{6}{8} \cdot \frac{6}{8} = \frac{9}{16} \end{aligned}$$

where we always find $\mathcal{P}_{0 \rightarrow 1} = \mathcal{P}_{1 \rightarrow 0}$ for symmetry reasons. For a transition like $T_{000}G_{0000000} \rightarrow T_{000}G_{0000001}$ we have to consider the combined probability over all ten individual ESC transition probability instances respective indices that are taking place within this particular transition. This is done for every combinatorially possible transition where only one ESC instance changes activity at a time which results in 10240 edges, the complexity of which again showcases why we do not consider ternary transitions.

From this we obtain again a raw transition matrix which is being transformed to a stochastic irreducible transition matrix. This results in the dominant eigenvalues shown in Fig.VII.27 with corresponding positive and negative eigenvector components with highest value microstates shown in table C.30. The stationary distribution reveals that the most highly ranked nodes always include an active enhancer around *Tbx21* whereas the amount of *Gata3* enhancer is quite variable yielding the statistically highest ranked state/node $T_{111}G_{0011001}$ with all three *Tbx21* enhancers active yet only the third, fourth and seventh *Gata3* enhancer being active the rest of which are assumed inactive or – for simplicity – repressed. This corresponds to high expression levels of *Tbx21* and intermediate levels of *Gata3*²⁹.

time being. Hence inactivity of an ESC in some condition (i.e. the absence of an active enhancer state) corresponds to repression in this picture.

²⁸An enhancer is active or not.

²⁹Rough quantitative estimates on this can e.g. be provided by respective gene expression fits for conditions where the mere existence of an active enhancer yields associated histone modification values in our data and for missing combinations a respective fit value would be taken along the lines of Eq.V.18. Yet at this point we will not further go into this direction which will be a matter of future investigations.

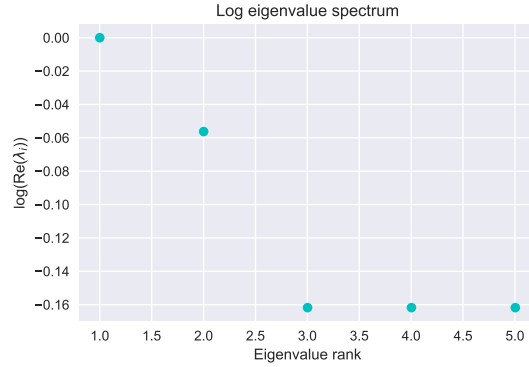


Figure VII.27: Close-up of the leading logarithmic eigenvalues of the epigenetic Tbet-Gata3 MISA motif.

At the same time we infer only one relatively long-lived metastable state with several consecutive states being comparably short-lived. The largest deviations for each component of the eigenvector denote the exchange happening between the respective microstates. The amount of ESCs that are switched on are considered as priors for the expression of a certain gene. For simplicity CSCs are assumed to be contributing to transcription in a distance-independent way.

Averaged over the top 5 positive and negative results³⁰ in λ_2 we find an information flow between microstates including a relatively higher amount of active *Gata3* enhancers and those with a relatively lower amount of active *Tbx21* enhancers. This trend changes a little bit for λ_3 where we find an exchange between high *Tbx21* enhancer amount microstates with slightly lower amount ones yet with corresponding lower *Gata3* enhancer amount microstates with higher ones. In contrast to considering only the most up- or down-regulated enhancer configuration³¹ we consider the set of leading up- and down-regulated configurations from which we can extract an overall relative amount of gene activity assuming that enhancers contribute linearly to gene expression³² depending on the amount of active enhancers. For λ_2 the exchange happens between microstates with relative gene expression values w.r.t. to the possible maximum of $\{Tbx21 = 100\%, Gata3 \approx 57\%\}$ and $\{Tbx21 \approx 93\%, Gata3 \approx 37\%\}$ averaged over the top 5 high- and low-ranked microstates³³. For λ_3 on the other hand the flow appears between averaged values of $\{Tbx21 \approx 93\%, Gata3 \approx 46\%\}$ and $\{Tbx21 \approx 80\%, Gata3 \approx 54\%\}$. This preliminary analysis can be performed for all subsequent eigenvectors to corresponding processes we yet find to be suppressed.

The tendency towards more fine-grained distinctions in *Gata3* can be expected since we did not assume the number of binding sites to be equally distributed for both genes in the MISA motif. This suppresses the detailed occurrence of an equally distributed Th1/2 hybrid state, which occurs only as a significantly upregulated state

³⁰This can in addition be weighted by the respective importance gain/loss from the corresponding eigenvector entry.

³¹This would resemble the line of argument made in [69].

³²Again we note that this could be quantified even more stringently by investigation of predicted gene expression values along the lines of Eq. V.18 which we yet will not delve deeper into at this point.

³³Since the number of all possible activity states of *Tbx21* for five microstates is 15 and for *Gata3* 35 one can naively count the total amount of active enhancer states in all five microstates and obtain this rough estimate.

in heavily suppressed processes³⁴. In extension to the analysis performed with the example of statistically inferred ESCs around Tbet and Gata3 we would still have to include an equal number of potential binding sites for both genes also considering indirect loops or instead reducing the *Gata3* binding locus for an equal binding site number distribution. In our case the bias towards more *Gata3* binding sites equally results in the leading metastable states experiencing a higher variance in *Gata3* expression resulting in several metastable enhancer gene configurations depending more strongly on *Gata3*. We nevertheless conclude that already on basis of the simple enhancer-gene configuration network we are able to infer metastability w.r.t. gene expression in combination with a simple linear gene expression model.

³⁴For this condition to be fulfilled all binding sites have to exhibit active enhancer states, i.e. the microstate $T_{111}G_{1111111}$.

VII.4 Discussion & Summary: Topology and function

We have investigated the nature of a specific class of combined epigenetic gene regulation networks on different levels. To this end we introduced a specific definition w.r.t. graph vertices, which consisted of relevant Th1 and Th2 genes as well as earlier defined CSCs in order to disentangle a class-specific functionality of certain epigenetic state combinations w.r.t. different treatment conditions as well as the functionality of certain genes involved in the definition of long-lived steady state phenotypes. This served as an extension to the earlier class-specificity analysis established by means of an ERT method. With the reduced graph concept of not treating every enhancer or repressive state as a uniquely realized instance but rather as belonging to a certain functionality class we defined regulatory connections between nodes that manifest themselves in a weighted multi-digraph.

Although having established a very reduced network concept w.r.t. the number of nodes and connections we still observe high interconnectivity which manifests itself in a low network diameter of $d = 4$, exhibiting small-world properties with only some CSCs appearing in the periphery hence regulating only a small number of genes specifically. These features also result in a weak scale-free behaviour w.r.t. in- and out-degrees although some of this is due that the network does not provide full information on TF binding. Hence only genes which are defined as TFs from the use of the available ChIP-Seq data sets possess out-going edges. We also found that especially the multigraph definition leads to a dampening of the scale-free behaviour and at the same time corrects for problems occurring from low scaling exponents. Still we find a notable number of network hubs consisting of TFs and several CSCs in the case of the out-degree distribution and mainly of TFs and other notable Th1 and Th2 genes for the in-degree distribution. As a result of lower scaling degrees the network shows high levels of resilience upon hub removal.

From the analysis of different centrality measures we were able to extract some preliminary information on node importance with varying interpretability. For our purposes we found especially the betweenness and eigenvector or Katz centrality measures to be of high importance. We also found that genes themselves play a generally larger role than CSCs with only some CSCs exhibiting high regulatory potential within the whole network itself and being able to change the phenotypic structure as a whole significantly. Nevertheless we also find an importance ranking in gene nodes themselves with TFs quite naturally occurring among the highest ranked genes in combination with which the removal of certain CSCs are obviously able to change at least the gene expression values of certain gene groups. Yet at this level we note that even slight variations w.r.t. node removal will, in combination with the definition of a stochastic matrix, in most cases lead to a different stationary network distribution and hence to a new characteristic steady state. Already with a jackknife node removal this leads to a maximum number of stationary network states that correspond to the amount of nodes within the full network, hence in principle leading to a quasi-continuum in terms of network stability.

Additionally we find a large number of direct autoactivation loops³⁵ of important TFs like Tbet and Gata3. Statistically among the most frequent motifs in the full network we mostly find coherent FFLs that both include inhibition and activation. We

³⁵mediated by a CSC

argue that for more than 5-node subgraphs the size of the network renders subnetwork computations unfeasible.

By eliminating the CSCs as gene mediators from the network and only focusing on TFs themselves we are furthermore able to infer direct statistically feasible activating and inhibiting connections between TFs themselves. We can not only confirm well established connections but w.r.t. to our network structure are also able to infer a range of new connections many of which are only suspected or completely unknown to our knowledge. Yet we also found that in some cases we cannot infer a one-to-one mapping condition of a TF binding directly at another TF locus since the mediation via CSCs infers the binding probability of a TF at a certain CSC genome-wide for Th1 and Th2 genes and hence also the reduced TF network only provides statistical information in contrast to actual binding. This naturally has to be taken into consideration.

Via community detection methods we were subsequently interested in functional clusters within the network itself. Even by crude inspection we were able to observe a distinction between Th1- and Th2-specific network parts, which is already evident on the basis of a reduced core network. Yet we also find regions of overlapping functionality providing a regulatory basis of mediation in between the classically differentiated functional parts of the network. To quantify these functional groups more strictly we inferred subcommunities unique to the underlying network topology which reproduce Th1- and Th2-specificity on different levels, yet also hint on hybridisation to which not only Th1/2-specific CSCs contribute differently but also certain previously classified Th1 and Th2 genes with still debated functionality. We note that this is for some subcommunities heavily dependent on the applied algorithm and parameters and has to be interpreted with care. Yet we also find nevertheless a consensus clustering particularly distinguishing between classical Th1- and Th2-specific nodes. Hence certain network nodes form tightly connected subclusters which are not only relevant for classic T-helper cell types but also for intermediate states. The attribution if a certain cluster belongs more or less to Th1 or Th2 can now either be made via the aforementioned classification or also on basis of their respective location in the network as well as in combination with their respective betweenness centralities.

We turned our focus on condition-specific networks by inferring a multiplex network by including only dimension-characteristic CSCs within each multilayer. This leads to a network topology unique to every condition. We find a condition-specific reordering of nodes due to their changing intraconnectivity, which leads to new subclusters unique to their respective topology and in some cases also to a heavier mixing of Th1- and Th2-specific nodes. If we keep the node position fixed we also find that only certain CSC subclusters within each network dimension are actively contributing as is obviously intended by construction. Yet more surprisingly this also corresponds to unique regions within the network which also confirms the structural and topological uniqueness and importance of CSCs including their definitions themselves.

Pushing this approach a little further we also investigated differential changes between the different multiplex dimensions in order to find the degree of up- and down-regulation between different dimension pairs. Especially the investigation w.r.t. the hybrid Th1/2 condition showed the increasing importance of certain network nodes

and regulations in order to enhance or suppress either certain Th1- or Th2-specific features and also being able to investigate an importance ranking of these nodes based on centrality measures. By adding or removing certain network nodes during up- and down-regulation we note that we again end up in conditions in between classical hybrids and the fully differentiated Th1 and Th2 cells. We also introduced the measures how to perform this in theory.

In order to obtain a more robust ranking measure w.r.t. node importance in the full network but also in the different condition-specific network dimensions as well as in differential networks, we finally investigated the implications of random walks performed on these networks. To this end we defined stochastic transition matrices and determined the respective stationary distributions. We hence obtain a topology-characteristic node ranking which is now not only able to distinguish between genes or CSCs separately but between all nodes in combination. We finally can not only confirm most results from the ERT class-specificity ranking within the condition-specific multiplex dimensions but also are able to pinpoint important players in up- and down-regulating processes in between different conditions. Integrating higher-order effects by investigating the whole eigenvalue spectrum of each network topology we are even able to determine medium- and long-lived metastable states occurring during the relaxation of a certain network to its stationary distribution. This unravels relatively tightly connected small-scale exchange processes between certain groups of genes which were grouped already earlier similarly by spectral clustering methods. Via this approach we can even quantify for differential networks which metastable states can be reached when performing a transition between two different steady states adding a dynamical aspect to the steady-state networks.

In conclusion we find distinct general and condition-specific network topologies with distinct functionalities which we can not only pinpoint via classical community clustering methods but to which we even can assign a robust node importance ranking in order to determine which genes or CSC contribute to a certain process in a relative manner. From this we can even elucidate the potential role of network nodes in metastable processes, which influence the respective function of a certain phenotype.

CHAPTER VIII

Future directions & improvements

As a final endeavour we want to give a short outlook on future directions of the work and potential starting points for extending the methods developed herein.

Concerning underlying data sets we investigated RNA-Seq as well as ChIP-Seq data in this work which is highly specific in the regard that it is dependent on Th1 and Th2 cells as well as on Tbet dose. As the methods such as the parametrization of the histone modification measure as well as the correlation algorithm and the inference of importance of CSCs with regulatory logic based on the underlying samples is highly generalizable it will be of general interest to test these methods on independent data sets also from different cell types. Not only will this provide important insights e.g. in the case of different master TF dose specificity such as Gata3 in Th2 cells or respectively others in different T-helper cell lineages but also with respect to STATs or other important network players. Such investigations will yield additional valuable information on T-helper cell plasticity such as e.g. in the case of Th1/Th17. Another question in this regard can be if epigenetic networks obtained in those systems provide comparable network topology in order to achieve similar effects or if their behaviour differs entirely.

Concerning the applicability of the HMM it would be certainly revealing to extend the binary statistical peak detection to continuous peak combinations with different heights, potentially further narrowing down the detection of possible enhancer states in certain regions. Yet since our correlation procedure already accounts for narrower regions of enhancers or in general chromatin states significantly contributing to gene expression it would surely provide an interesting mode of validating the correlation method since changes in peak structure assigning a certain hidden state to a specific position should be reflected in the significant correlations as well.

Additionally actual validations of certain enhancer regions around notable genes like *Tbx21*, *Gata3* or at the Th2 cytokine cluster would be an important step in strongly confirming our learning-based statistical inference methods and their results experimentally. A viable first step in this direction could be the usage of conformation capture methods for the same cell conditions that we investigated such as 3C, 5C or Hi-C depending on the scope one wants to achieve. In general although specific enhancer-promoter binding events would be already very informative a genome-wide approach naturally would provide more flexibility for the full range of Th1 and

Th2 gene loci.

Turning to the correlation measure model additional experimental validations of enhancers can be easily integrated in the parametrization of the correlation measure leading to an increasingly robust model based on an growing enhancer learning sample on which the optimization procedure is trained. Hence such iterative procedures provide a feedback workflow via updated learning. Furthermore a parametrized model including error bounds of the parameters for the estimation of subsequent significant correlations still calls for an efficient implementation for genome-wide transcript analysis. The correlation algorithm itself provides a computational framework which can be easily extended as well e.g. by incorporating further data base information from ENCODE or the Roadmap Epigenomics projects such as e.g. DNase HSI data into the workflow. In order to provide the research community with a stand-alone version of the algorithmic procedure or in form of a computational package it might be advantageous to generalize the algorithm to other input data different from HMM states but also to provide the ability to process input from different chromatin state inference algorithms automatically. Yet manual labelling is possible in principle at the moment while modifications in order to achieve this have to be made before running the code.

For the inference of the underlying methods we relied on a limited set of TFs in Th1 and Th2 cells leading to a bias in TF-CSC binding towards the investigated dominant TFs in Th1 and Th2 cells. Obviously accounting for a larger range of such TFs will yield a more complete picture of the underlying network topology. Furthermore the ERT results on TF binding at CSCs at the moment rely on the presence or absence of a TF at a certain CSC. This can be quantified even more robustly by considering the normalized read counts from ChIP-Seq binding at different CSCs. In doing so one can obtain a data-driven statistical measure on TF binding at CSCs which also takes the read load from the underlying data into account. In addition the intra-class specificity method for CSCs could profit even more by integrating more classes into the procedure. This could be for example achieved by extracting leading nodes from the stationary distribution of the *Diff1* and the *Diff2* network that are comparably low-ranked within the Th1 and Th2 networks respectively and classify them as Th1/2 genes. It will be a matter of further investigation if a classification for steady states in between two extremes such as Th1 or Th2 can effectively lead to better results.

With regard to community detection within the underlying networks the methods still have to be extended for directed networks including the application of consensus clustering to obtain more robust results. Since as of yet implementations for this kind of problem are still rare this will naturally pose an important point for further investigations.

Concerning the random walk methods on the weighted multidigraphs under consideration the results could be improved even further by not only considering positive edge weights in order to obtain stochastic irreducible matrices but also by distinguishing explicitly w.r.t. inhibitory edges by e.g. decreasing transition probabilities of associated flows in the Markov chain. In contrast to this as of yet we treat activating and inhibitory edges as being equal by only considering the flow over possible

connections within the respective network. This is a reasonable thing to do as we only want to rank node importance¹. Yet w.r.t. the underlying network dynamics this would be of great importance.

As we have already indicated at the end of section VII.3 we can also investigate subnetworks of certain genes and observe the activity of all CSCs that exhibit edges with those genes respectively. This can be obviously extended beyond the aforementioned reduced MISA motif and also include e.g. STATs but more importantly also RSCs². Since we have only investigated random walk metastability as well as the stationary distribution of the according stochastic matrix associated with transition probabilities that were determined via a frequentist approach one can furthermore ask for individual gene regulation functions (GRFs) that are produced via the respective significant CSCs. In order to do this one will have to parametrize all respective possible transitions between activity and inactivity combinations of all regulatory CSCs. This follows the line of thought established in [2, 109] for non-equilibrium systems. In this case all viable activity states for gene expression are assessed, i.e. all experimental condition where the gene is expressed with corresponding activity patterns of all regulating CSCs. This basically forms a graph partitioning into subgraphs that contribute to gene expression itself via the determination of all sets of spanning trees or arborescences [2, 355] that can achieve the respective activity microstate over a series of state transitions. The resulting partition function corresponds to the GRF for the respective gene. In [2] the spanning tree model was derived for a full account of all possible of such microstates regarding TF binding of two TFs. In our case obvious problems arise concerning graph complexity since we usually have more than two CSCs around each gene. Also it depends on the model complexity e.g. if one additionally considers looping or more sophisticated models of transcriptional activation such as telegraph or refractory state models [209, 254]. These models can be drastically reduced by including only the observed CSC activity states, which reduces the number of nodes of the transcriptional activation network. Yet this will in turn lead to problems with the linear Laplacian framework as proposed in [2] resulting in non-linear transitions since several activity states can change in one step. In general for all transitions to be considered one needs a robust way to estimate the transition parameters between CSC activity microstates around the respective gene. An approach to this has been already proposed via the MISA motif. Nevertheless we note that this parameter estimation is not always straightforward and also can be different for each enhancer site. Alternatively one can sample the space of all not straightforwardly inferable transitions. One can now e.g. ask for all possible steady states that are achievable via certain parameter combinations. Furthermore one can even try and solve an optimization problem, i.e. under which parameter combinations for the full partition function, i.e. the GRF, a fixed maximum number of steady states can occur. Since this corresponds to a root finding problem one can utilize that an upper bound for the maximum number of possible positive real roots is given by *Sturm's theorem* in combination with *Descartes's rule of signs* [304]. In spite of the problems arising with the parametrization of a full system of CSCs in principle one can determine GRFs

¹In our case we always have to check explicitly for each metastable process which connections appear in the actual activation and inhibition network between the respective nodes.

²In our case we only included inhibition by absence of activity for simplicity.

for different genes such as *Tbx21* and *Gata3* and via the epigenetic landscape and the CSCs around these genes infer their respective steady states and hence also the stability properties not only of small motifs like in the case of the MISA motif but also for increasingly larger subnetworks such as the core network mentioned in VII.2.3.

We also suspect that by calculating effective GRFs for certain subsystems and providing a coupling between these systems the combination of both might give additional insight into the existence of multiple long-lived steady (or at least metastable) states (see e.g. [52, 250, 251]). Work into this direction has been already going on for some time in the context of coupled community-structures oscillator networks producing so-called chimera states (see e.g. [281, 287, 310, 373]) being important in neuronal brain networks. Under certain conditions one hence obtains a potential landscape with multiple local minima [239] similar to the final states in the famous Waddington landscape analogy from epigenetics. The question at this point is if similar behaviour can be equivalently achieved in the epigenetic networks under consideration in this thesis or if coupled GRFs can produce this behaviour under realistic circumstances. Obviously this will provide an extensive body of future research.

We have seen that although we have already achieved a great deal on robustly inferring the epigenetic landscape including their underlying epigenetic networks in Th1 and Th2 cells there are many ways to investigate the methodology even further especially including experimental validations and different experimental systems as well as improving certain computational details. We also see that especially the investigations w.r.t. stability properties of subnetworks and small motifs have found a new quantitative starting point from which to build on w.r.t. results obtained in this work.

CHAPTER IX

Summary

IX.1 General

We are now able to answer the problems and questions that arose in the beginning in the introduction and summarize the discoveries that unfolded during the process chronologically.

We started this work in chapter III by introducing and analysing the underlying histone modification ChIP-Seq and RNA-Seq data sets of different T-helper cell conditions w.r.t. cytokine as well as Tbet dose resulting in differentiated Th1, Th1/2, Th2 as well as naïve cell types. To this end we established the bioinformatical pipeline for further post-processing of the data sets.

Utilizing these results in chapter IV we mapped the epigenetic landscape in all experimental conditions by first determining so-called chromatin states via a HMM implemented in ChromHMM. The resulting model included a total of 16 hidden states, corresponding either to promoter, enhancer, bivalent or repressive/silenced states. We found at different genetic loci that not only are these chromatin states in many cases highly condition specific but also their activity state changes quite frequently depending not only on cytokine but as well on Tbet dose. This especially favours the actual importance of Tbet, coming back to one of the initial questions, on the epigenetic landscape in Th1 and Th1/2 cells contrary to reports such as in [314]. In the end we obtained an annotation not only of chromatin states in general but also of enhancer states, showing significant overlap with p300¹ and repressive states as well in different Th1 and Th2 conditions. We also found that active enhancer states can become not only poised in certain conditions, e.g. under Tbet knock-out, but acquire a bivalent state or even become fully repressed.

In order to map chromatin states and specifically enhancers to respective gene candidates we introduced in chapter V a quantitative correlation measure based not only on the classic histone modification enhancer marks H3K4me1 and H3K27ac but also on the repressive mark H3K27me3, which is frequently occurring in the opposing cell condition of the respectively occurring active enhancer state. Based on a learning sample of well known experimentally validated enhancers and by optimizing an appropriate objective function we found a stable parametrization which provides the

¹yet not being limited to this HAT

ability to stabilize and maximize the correlation of enhancer and equally of repressive states with gene expression. We were able to show that this novel measure outperforms the usage of every individual histone modification in a robust way. This leads to the conclusion that not only H3K4me1 and H3K27ac are important markers for enhancers but the appearance of the repressive mark H3K27me3 in respectively opposing cell conditions is an equally important indicator of the gene regulatory correlate². Additionally we implemented a correlation algorithm from scratch which is able to deal with chromatin state input data as a prior for selecting respective correlation regions and utilizes an arbitrary (in our case the learned) histone modification measure for correlation calculation. The algorithm also contains a sophisticated splitting and merging procedure for unravelling substructures in correlations reflecting the underlying peak structure and treating neighbouring elements due to a statistical similarity criterion. We tested the procedure on a variety of genomic loci on TADs including the well-annotated *Ifn γ* locus and were able to reproduce nearly all previously known enhancer sites as well as a smaller number of new ones. The results were also integrated with p300 binding data as well as with published ChIP-Seq of the master TFs and notable STATs. We also showed that in the case of a superenhancer like the one at *Tbx21* only small segments were actually co-regulated with gene expression. In the case of crowded loci like the Th2 cytokine locus we were able to elucidate a one-to-one mapping for a large number of significantly correlating enhancer segments via the facilitation of partial correlations. This led to an annotation of the locus with a concrete distinct mapping of several already known enhancer segments of which the distinct functionality was either simply unclear or a coregulation of all genes at the locus was proposed. We were also able to infer a logic for inhibitory action of significantly correlating chromatin states and predicted the expression of genes with different models of enhancer regulation based on the parametrized histone modification measure. This results in unravelling the unique epigenetic landscape in Th1 and Th2 conditions.

Since we found that chromatin states in general and enhancer states in particular follow different regulatory patterns w.r.t. their activity state we introduced a ternary state classification in chapter VI leading to so-called chromatin state classes (CSCs)³ which appear in a variety of combinations around certain genes they were initially mapped to. We investigated a large set of Th1 and Th2 specific transcripts and found via the ERT procedure an importance ranking of CSCs in distinguishing between Th1- and Th2-specific transcripts. We also introduced a measure which is able to disentangle the respective relevance contribution of each of these CSCs either to the Th1 or Th2 condition. This resulted in an intra-class specificity ranking, including the introduction of a novel class-specificity measure, of respectively Th1- and Th2-specific enhancers and repressive states depending on their regulatory logic. These CSCs can be furthermore grouped into being mostly cytokine or Tbet dose responsive. Furthermore we also found a co-occurrence probability of pairwise combinations of these cell-specific CSCs leading to a robust predictor in furthermore specifying the cell-

²We find that the algorithm to some extent resolves a known shortcoming of ChromHMM, i.e. only considering presence/absence of modification in a certain region, by introducing the peak structure through the back door via correlations.

³Depending on the context we exemplary defined enhancer or repressive state classes.

specificity of a certain transcript by just mapping its respective epigenetic landscape and observing its regulatory logic.

Following up on the intra-class specificity approach we were furthermore able to assign an importance weight of a TF binding at a certain specific CSC hence leading to an epigenetic network containing bidirectional edges, based on the ERT method. The epigenetic network introduced in chapter VII of Th1 and Th2 cells was uniquely defined with gene and CSC nodes reducing the complexity of the network significantly and also putting additional focus on the unique regulatory logic of ESC and RSC elements. By defining an appropriate adjacency matrix we obtain an epigenetic network defined by a weighted multi-digraph. We found that although the concept of network architecture is a very reduced one we obtain a topology exhibiting high connectivity. We observe that the network consists of a small number of high in- and out-degree nodes i.e. hubs and is highly resilient to node removal exhibiting weak scale-free properties. Reducing the network to an effective version only including TF as nodes we obtain new modes of TF regulation not only in the case of activation but also concerning inhibition. We are also able to infer auto-activation and mutual inhibition loops based on the adjacency of the respective TF nodes considering the underlying epigenetic bipartite CSC logic.

Furthermore we inferred community structures in the network, depending on the respective method leading to different substructures but overall distinguishing between Th1- and Th2-cell specificity not only for genes themselves but also for CSCs, hence confirming results from the intra-class-specificity analysis. We also found sub-clusters which rather fulfill the task of mediating between the Th1 and the Th2 part. We extended these investigations by considering so-called multiplex networks by only including condition-specific CSCs. We found that different parts in these networks are responsible for different condition-specific regulatory tasks leading to completely unique topologies. Considering differential networks between the respective multiplex dimensions we were even able to unravel dynamic properties in considering up- and down-regulation from one steady state to another indicating that certain regulatory functions and especially nodes contribute more to a certain transition than others.

In order to take this even one step further we were asking which nodes are more or less relevant not only in the full network and the multiplex dimensions but also in differential regulation. To this end we introduced a stochastic irreducible transition matrix for each network individually as well as for the full network and determined their stationary distributions. This corresponds to the long-term behaviour of a random walker on a network being described via a Markov chain. From this we were able to rank the importance contribution of each node for each network respectively and hence make predictions on the relevance of each particular node on each process. Furthermore these investigations also led to the inference of metastable processes indicating a separation in relevance of particular nodes leading to a specific information flow in the frame of a random walker. This was possible not only for the full network recovering pure Th1- and Th2-specific processes but also in condition-specific networks uncovering processes e.g. unique to Th1/2 hybrids. By including hybrids one can even find metastable processes occurring during the process of up- and down-regulation leading to possible predictions of cell states in between the classic Th1 and

Th2 dichotomy as well as the Th1/2 hybrid state. We argued that these findings could stimulate the search for new long-lived steady states not only based on the respective node importance and their unique topology but also by future investigations of CSC-regulatory microstates as in the case of the MISA motif or by determining gene regulatory functions as outlined in chapter VIII.

IX.2 Originality of work

The question of how this work relates in contrast to other research in the field and which contributions have been made by the methods and the analysis established herein can finally be answered.

Not only is this to our best knowledge the first time the epigenome in Th1 and Th2 cells in mice has been mapped genomewide w.r.t. the respective specific enhancer landscape by a chromatin state segmentation assigning and extending this by also considering repression but we determined a unique quantitative method via the parametrized histone modification measure for inferring significant co-regulation of enhancers with genes. In doing so we find that not only the classical histone marks have to be considered for enhancer activity but also the appearance of the repressive mark H3K27me3 when occurring in adverse cell conditions. Hence the growing absence of H3K27me3 is a comparably good estimator of enhancer activity at the same time as H3K4me1 and H3K27ac yet we note that all three modification marks have to be considered at the same time according to our inferred parametrization. We extended this approach even further to partial correlations and provided a computational framework which implements a sophisticated statistical approach to narrow down the prior indicators of transcriptional regulation from the HMM and assign them a quantitative correlate, which has been shown to show extraordinarily good performance in well-known regions. We note that this presents a non-trivial problem. In addition to this we are even able to make genome-wide predictions *in silico* based on the respective learning sample. This method now does not rely on nearest neighbour association of enhancers and promoters but is without any additional assumption on interaction able to specify co-regulated enhancer states with gene expression with the interacting regions being merely restricted to TADs. We not only found that as it is already known cytokine dose has significant impact on the alteration of the epigenetic landscape in Th1 and Th2 cells but this work furthermore strengthens the importance of Tbet dose in Th1 and Th1/2 hybrid cells also with special focus on the enhancer and repressive state landscape. This extends the prevailing view that the influence on Tbet on epigenetic regulation in these cell types is rather low.

We additionally propose a dose-dependent ternary regulatory logic dependent on the underlying experimental conditions again including enhancer activity, the loss of the activity mark as well as the acquisition of a repressive mark, which results in an enhancer landscape, and more generally in a chromatin state landscape, with enhancer states exhibiting different functionality considering their activity. To our best knowledge this has never been proposed before and definitively not w.r.t. Tbet dose as well as in hybrid Th1/2 cells. By this approach we are able to not only infer regulatory enhancer classes (or CSCs) that are indicative of a certain cell-specificity but also to predict cell-specificity of transcripts other than those that the ERT model was

trained on. In order to assess this intra-class specificity as we coined it we developed a novel measure to achieve this cell-specificity. Also to our knowledge such a definition of an intra-class specificity measure does not exist at this point. This whole approach results in a set of reliable predictions and is highly generalizable also to a larger number of classes and obviously also to other differentially regulated experimental systems.

This can be also seen from the network perspective where we were able to validate the respective predictions by community clustering recovering CSCs, which were classified before in the respective cell type. We also chose a novel approach to GRN architecture for which we found no correspondence in the literature. This consisted of inferring a bipartite graph where gene regulation is always mediated via enhancers or by repressive states not only including activation but also inhibition. More specifically the concept of the adjacency matrix is also unique in the regard that we did not choose actual enhancer instances as regulatory entities but rather their regulatory classes. Hence the underlying structure is not only based on specific regulation by special enhancer instances, which might or might not be actually validated afterwards, but rather on their respective logic decreasing the necessity of a specific enhancer instance to be completely accurately predicted. The bottom line of this is that regulation is not only inherent in the edges but also in the nodes themselves. We provide the first full account on epigenetic enhancer networks in Th1 and Th2 cells including these entities themselves as well as the first account of condition specific multiplex networks in those cells not only for the classical lineages but also for hybrid cells. We were furthermore able to pin down topological properties of the respective networks and not only investigate a node importance ranking via the stationary distribution of a random walker but additionally also inferred metastable properties relations that have been rarely investigated in this field.

We conclude that we present a series of novel quantitative data-based methods employing statistical learning procedures and developed not only a computational framework but also an unprecedented viewpoint on epigenetic regulatory logic resulting in unique network topologies. Among the special strengths of the above stated novelties are that they are all highly generalizable to every arbitrary question of differential regulation although they have been developed for the example of the Th1/Th2 system. We continued our investigations by employing analyses of these networks which are rarely used in this field and also provide a general way in order to obtain crucial information on important players in differential network regulation serving as a strong starting point for further investigations into multistable plasticity properties of Th1 and Th2 cells.

APPENDIX A

Additional mathematical background

In the following we will introduce some preliminary statistical quantities which are used in some of the chapters as well as a glossary of selected mathematical notations throughout this work.

A.1 Definitions and explanations

A.1.1 Statistics

Model selection criteria

In multi-regression problems issues often arise about a bias-variance trade-off resulting either in underfitting or respectively in overfitting. This trade-off is usually a matter of balancing fit and complexity (see e.g. [330, 331]). Measures for this issue are provided by model selection criteria the most popular of which are the so-called *Bayesian information criterion* (BIC) as well as the *Akaike information criterion* (AIC). A general version of the BIC reads

$$\text{BIC} = -2\ln(L) + k \cdot \ln(N)$$

where $\ln(L) \equiv \mathcal{L}$ is the log-likelihood value of the model, k denotes the number of employed free parameters and N is the number of samples.

The AIC instead reads

$$\text{AIC} = -2\ln(L) + 2k$$

omitting the consideration of the sample size in penalizing the number of parameters as a model selection criterion. Hence in general the BIC penalizes more complex models more strongly¹. If one now compares different models with each other the one having the lowest AIC or BIC is chosen. The question up to which point differences in the information criteria of two models yet are relevant for choosing the model with a lower score has to be given by a rule-of-thumb relevance cut-off. Hence if the difference between the best model and another candidate model is e.g. ΔBIC

¹This is already valid for a sample size of $N = 8$ which one can easily check.

$\in [0, 2)$ there is significant reason to rather support the candidate model [57]. The same holds true for the AIC. We furthermore note that both measures have to be taken as a relative criterion yielding for a low score merely a model which is better than all other models in comparison. The model can still be bad in general.

Adjusted p-value

The so-called adjusted p-value p_{adj} is a corrected version of the regular p -value, which gives a rejection level of the underlying null hypothesis. If a certain amount of null hypothesis tests for a large sample are carried out one faces the so-called *multiple testing* problem as the number of false-positives increases with the number of hypothesis tests being carried out at the same time. This happens e.g. in the case of differential gene expression as carried out in chapter III where hypothesis tests for the differential expression of all known ENSEMBL transcripts have to be carried out at the same time. There are two popular methods that provide a solution to this problem, i.e. the *Bonferroni* method and the *Benjamini-Hochberg* method (see standard textbooks such as [330]). In the former method if N is the number of multiple tests we obtain the adjusted p -value by

$$p_{\text{adj}} = N \cdot p,$$

which is a rather conservative estimation. A more relaxed one is given by the Benjamini-Hochberg method, which yields for sample i

$$p_{\text{adj},i} = \min_{r \in \{i, \dots, N\}} \left(\min \left(\frac{N}{r} \cdot p_r, 1 \right) \right).$$

where r is the rank of the p -value in the whole sample when sorted in ascending order, the inner minimum ensures p -values to be smaller than one and the outer one ensuring the same ordering as for the unadjusted p -values. This is actually the method employed for the results in tables C.1 and C.2.

A.1.2 Partial correlation

Considering the pairwise correlation of a number of variables greater than two the mutual dependence of the variables on each other has to be disentangled first in order to remove spurious correlations of all other variables on any each pair. The solution to such an issue is given by so-called *partial correlations*. One can for example consider three variables A , B and C where a potential correlation between A and B is only mediated via the variable C . In this case the true partial (first-order) correlation coefficient between A and B given C is given by

$$\rho_{AB|C} = \frac{\rho_{AB} - \rho_{AC}\rho_{BC}}{\sqrt{(1 - \rho_{AC}^2)(1 - \rho_{BC}^2)}}. \quad (\text{A.1})$$

where ρ_{ij} denote the ordinary zero-order correlation coefficients as e.g. in the case of Pearson or Spearman rank correlation. In the case of N variables we have to

determine the partial correlation of order $(N - 2)$, which is defined recursively by lower order partial correlation coefficients:

$$\rho_{AB|C\dots N} = \frac{\rho_{AB|D\dots N} - \rho_{AC|D\dots N}\rho_{BC|D\dots N}}{\sqrt{(1 - \rho_{AC|D\dots N}^2)(1 - \rho_{BC|D\dots N}^2)}}. \quad (\text{A.2})$$

For relations to multiple regression including a definition via residuals see e.g. [18].

A.1.3 Perron-Frobenius-Theorem for ergodic Markov chains

Since ergodic Markov chains are also called irreducible the following version of the Perron-Frobenius theorem for irreducible stochastic matrices also holds for ergodic Markov chains in general:

Theorem (Perron-Frobenius) *Be P an irreducible, aperiodic stochastic matrix, then it has a largest unique single eigenvalue $\lambda_1 = 1$ with corresponding left eigenvector π_1 . Furthermore π_1 can be chosen to have only positive entries.*

Further information on the formulation including a proof can be e.g. found in [133, 223].

A.2 Mathematical Notation

Notation	Meaning
\mathcal{R}	Pearson correlation coefficient
\mathcal{A}_{ij}	Elements of adjacency matrix
\mathcal{A}_{ij}^d	Elements of adjacency matrix of multiplex dimension d
\mathcal{W}_{ij}	Elements of multi-edge weight matrix
\mathcal{P}_{ij}	Elements of raw transition matrix
$\tilde{\mathcal{P}}_{ij}$	Elements of stochastic transition matrix
$\tilde{\tilde{\mathcal{P}}}_{ij}$	Elements of stochastic irreducible transition matrix
π_j	Elements of stationary distribution vector
k_{in}, k^-	In-degree
k_{out}, k^+	Out-degree
γ	Scaling exponent
\mathcal{E}_j	Enhancer state class (feature) instance
$\mathcal{I}_{\text{Gini}}^*$	Intra-class Gini impurity
$\mathcal{I}_{\mathcal{C}}$	Intra-class measure
$\mathcal{I}_{\text{Gini}}$	Gini impurity
\mathcal{L}	Log-likelihood
v_i, \mathcal{N}_i	Vertex, node
$\theta = \{\Theta, \Sigma, \pi\}$	HMM parameters = {transition probabilities, emission probabilities, initial state probabilities}
$ s $	length of segment
$\langle \cdot \rangle$	mean of element \cdot

APPENDIX B

Supplementary Figures

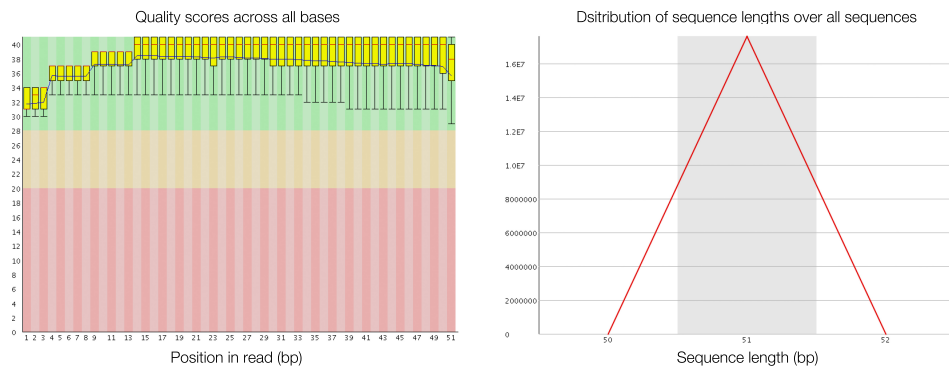


Figure B.1: Per base sequence quality (left) and sequence length distribution over all sequences (right) for one ChIP-Seq sample of H3K4me1 under Tbet^{+/+}Th1 conditions.

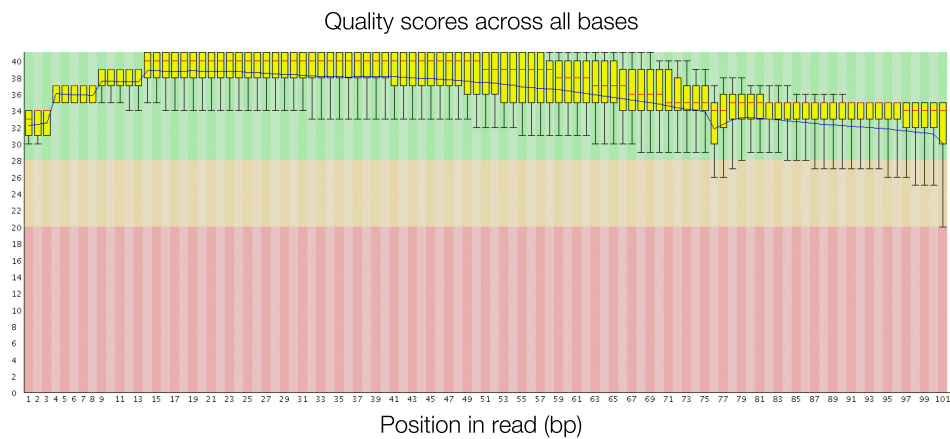


Figure B.2: Per base sequence quality for one RNA-Seq sample for Tbet^{+/+}Th1 conditions.

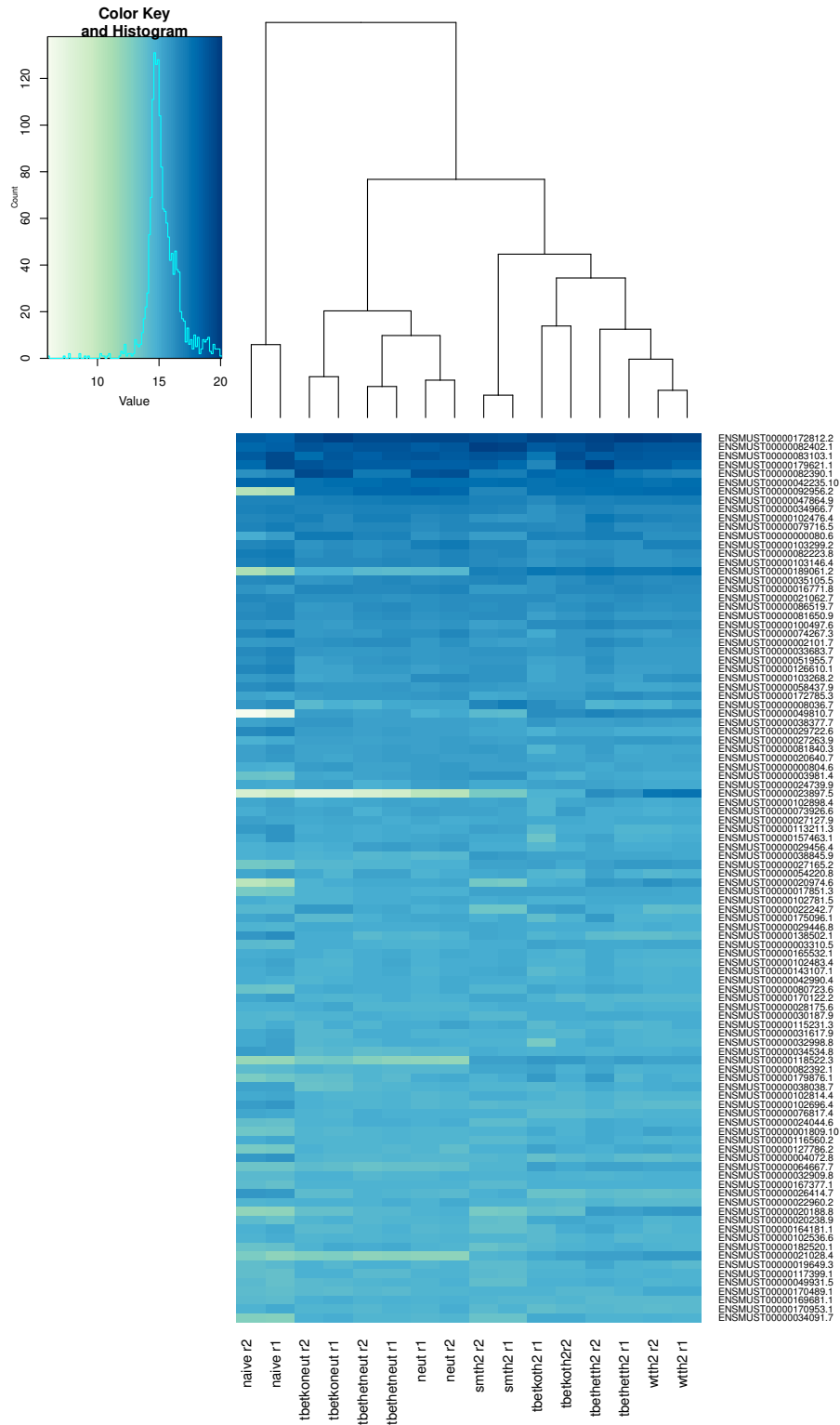


Figure B.4: Hierarchical clustering of a subsample of VST normalized genes for all RNA-Seq samples.

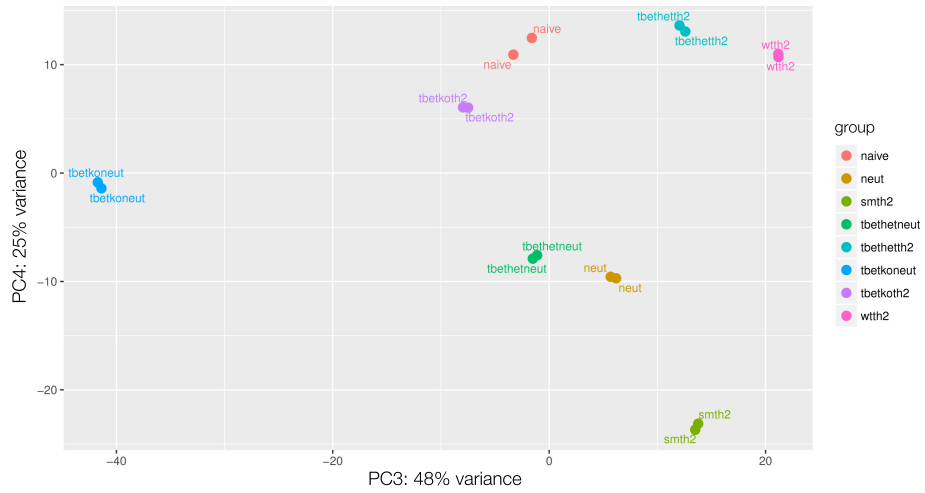


Figure B.5: PCA analysis components PC3 and PC4 for the VST case.

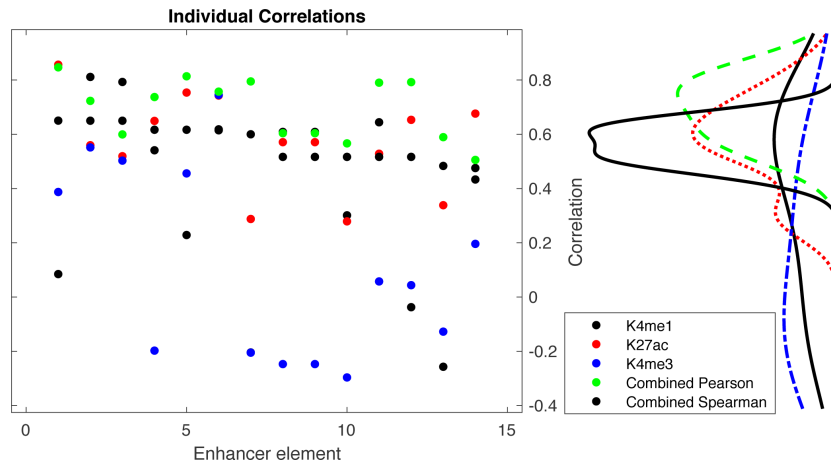


Figure B.6: Pearson and Spearman correlations for the parametrized histone measure in comparison to the individual histone modification Pearson correlations.

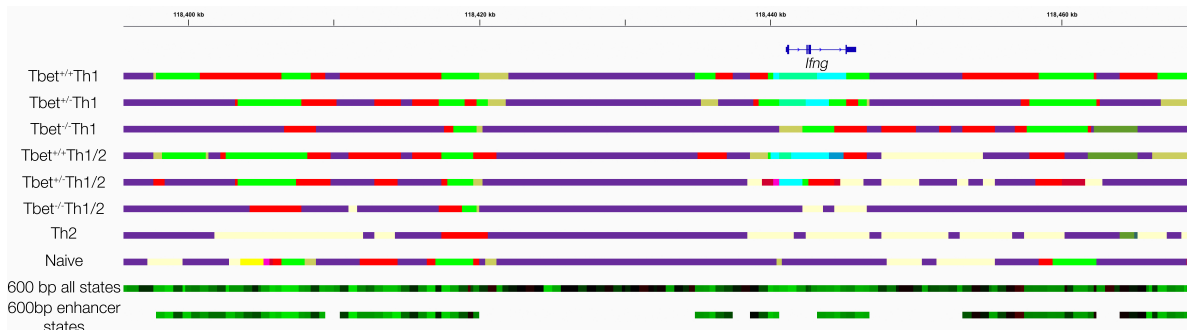


Figure B.7: Results from the correlation algorithm for a resolution of 600 bp for all HMM states as well as for enhancer states only. We observe slight deviations in the merging of statistically similar elements.

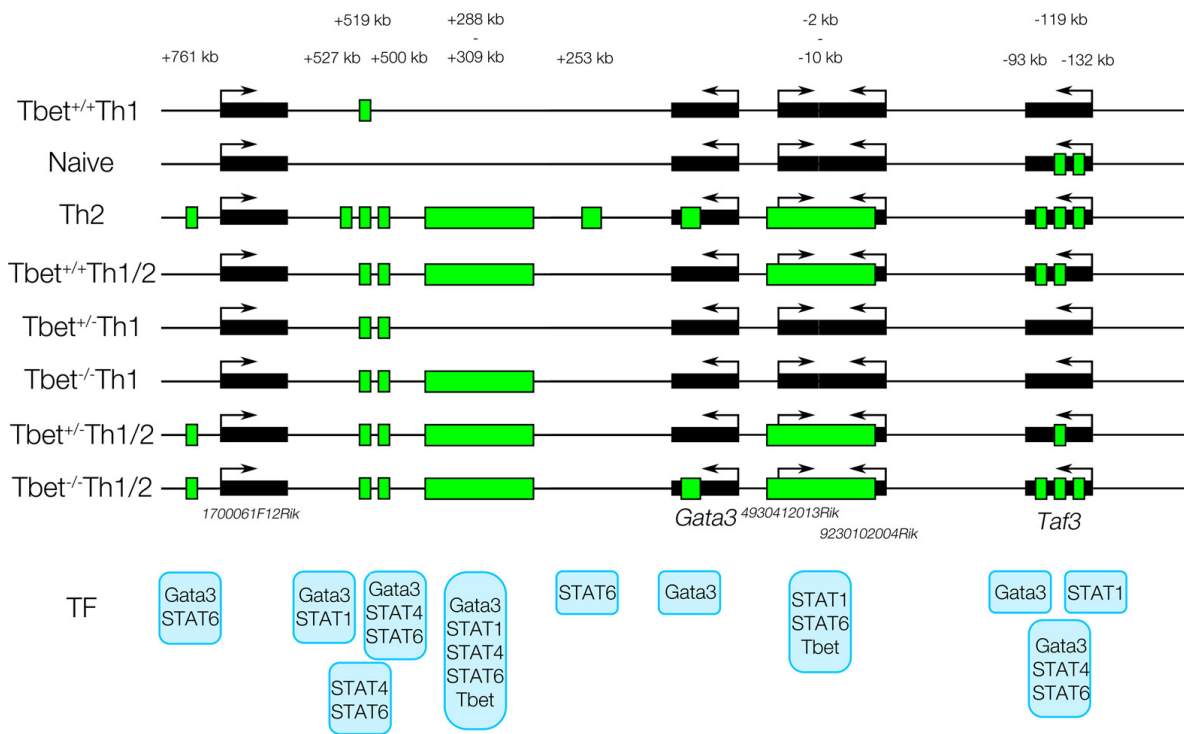


Figure B.8: Schematic depiction of significantly correlating activating elements at the *Gata3* locus unveiling the enhancer activity regulation.

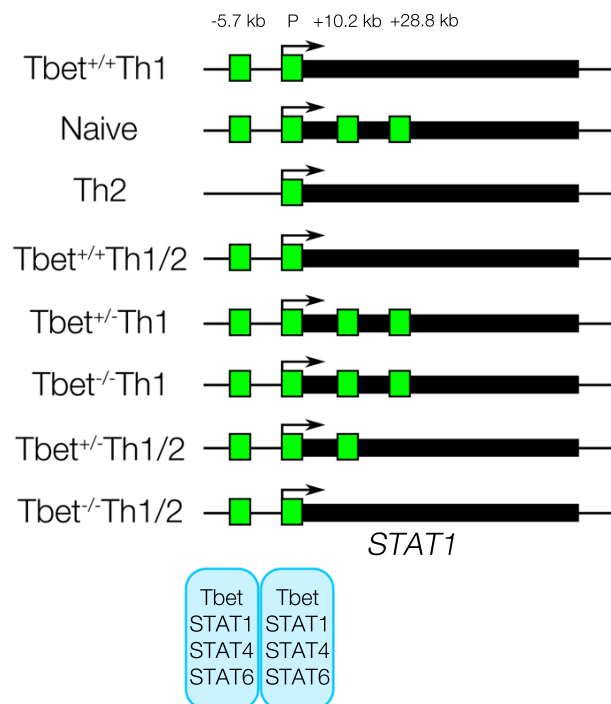


Figure B.9: Schematic depiction of significantly correlating activating elements at the *STAT1* locus unveiling the enhancer activity regulation.

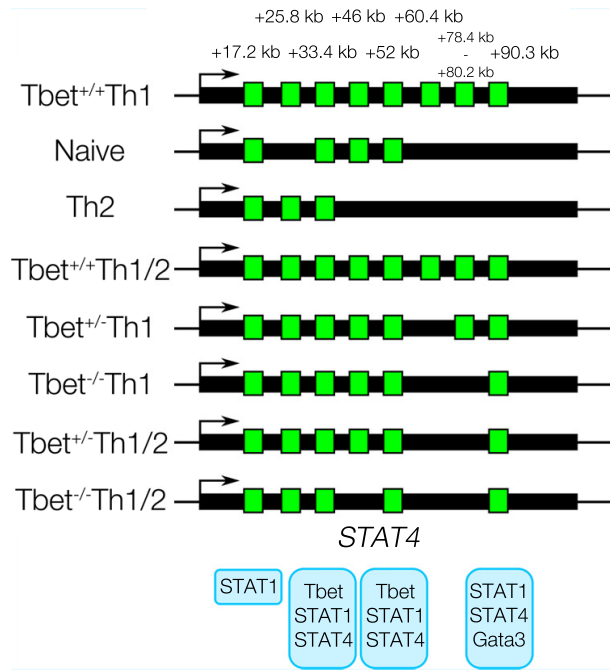


Figure B.10: Schematic depiction of significantly correlating activating elements at the *STAT4* locus unveiling the enhancer activity regulation.

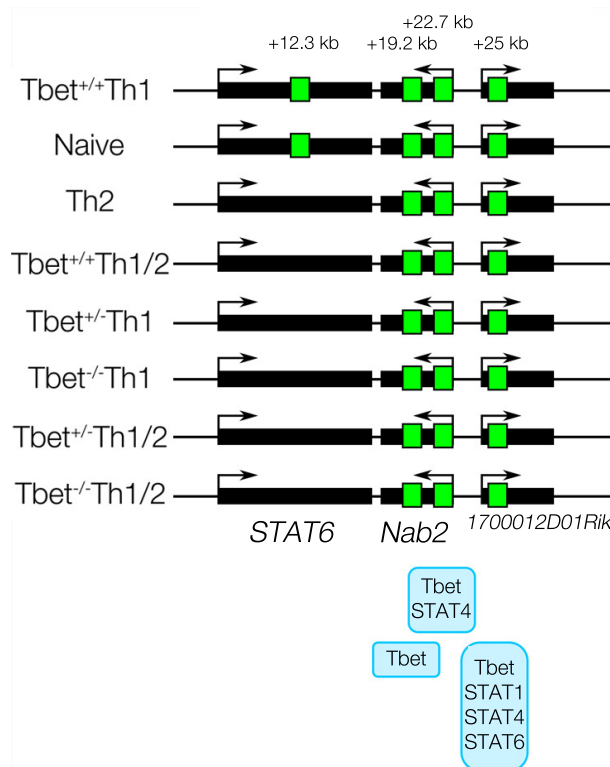


Figure B.11: Schematic depiction of significantly correlating activating elements at the *STAT6* locus unveiling the enhancer activity regulation.

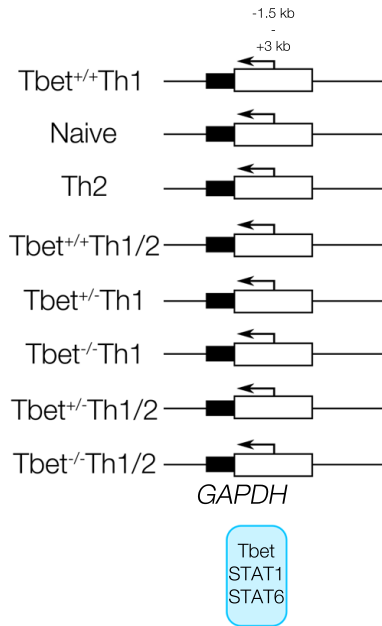


Figure B.12: Schematic depiction of significantly correlating activating elements at the *GAPDH* locus unveiling the enhancer activity regulation. In white we denote active enhancer elements which do not exhibit any significant correlation according to our correlation method as can be expected for a housekeeping gene.

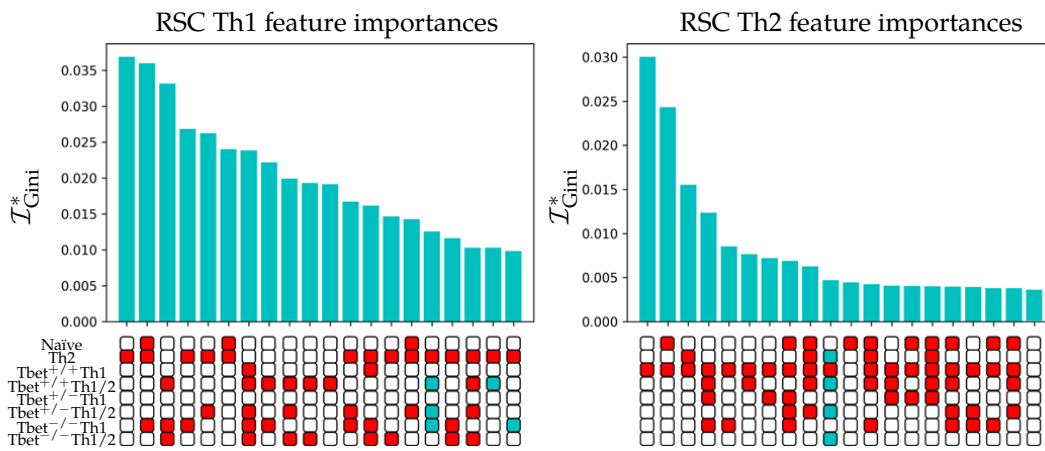


Figure B.13: Top 20 ranked RSC features for Th1 (left) as well as Th2 transcripts (right). The ranking was obtained via the application of the intra-class Gini impurity.

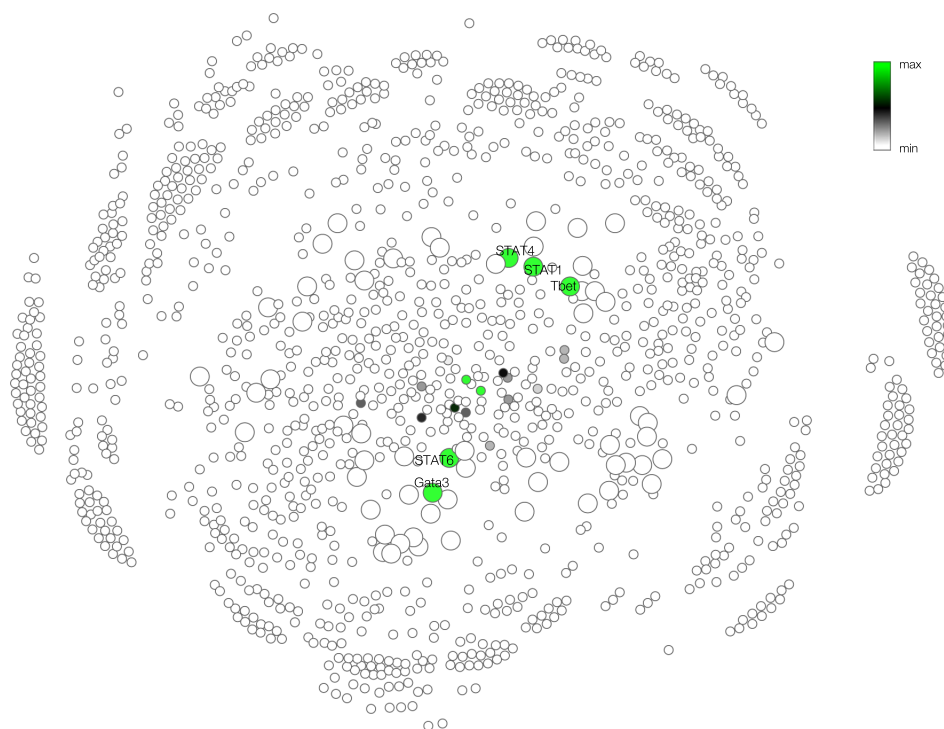


Figure B.14: Out-degree centrality for the full weighted multi-digraph. The colour-scheme depends on the maximum and minimum in-degree centrality values. As previously we depict only TF node labels for simplicity.

Subgraph	Frequency	Random mean frequency	Random standard deviation	Z-score
	0.002748%	3.0916e-05%	2.2603e-07	120.21
	0.002016%	1.1315e-05%	1.7298e-07	115.89
	0.00099872%	2.3266e-06%	8.6095e-08	115.73
	0.049743%	0.0018082%	5.0943e-06	94.097
	7.4438e-05%	1.1889e-07%	8.3643e-09	88.853
	0.065574%	0.0038364%	7.7499e-06	79.663
	0.001954%	1.0607e-05%	2.9112e-07	66.756
	0.00063893%	7.9203e-06%	9.4621e-08	66.688
	0.00044663%	1.434e-06%	6.919e-08	64.344
	0.00079401%	4.825e-06%	1.2403e-07	63.627
	0.0031016%	8.5833e-05%	4.8813e-07	61.782

Figure B.15: Top-ranked directed 4-node subgraphs w.r.t. Z-score. Green edges indicate activation and red edges indicate inhibition. We also list the respective occurrence frequencies of each subgraph and compare them to the mean frequency for a set of 100 random graphs including the respective standard deviation. From this we obtain the associated Z-scores.

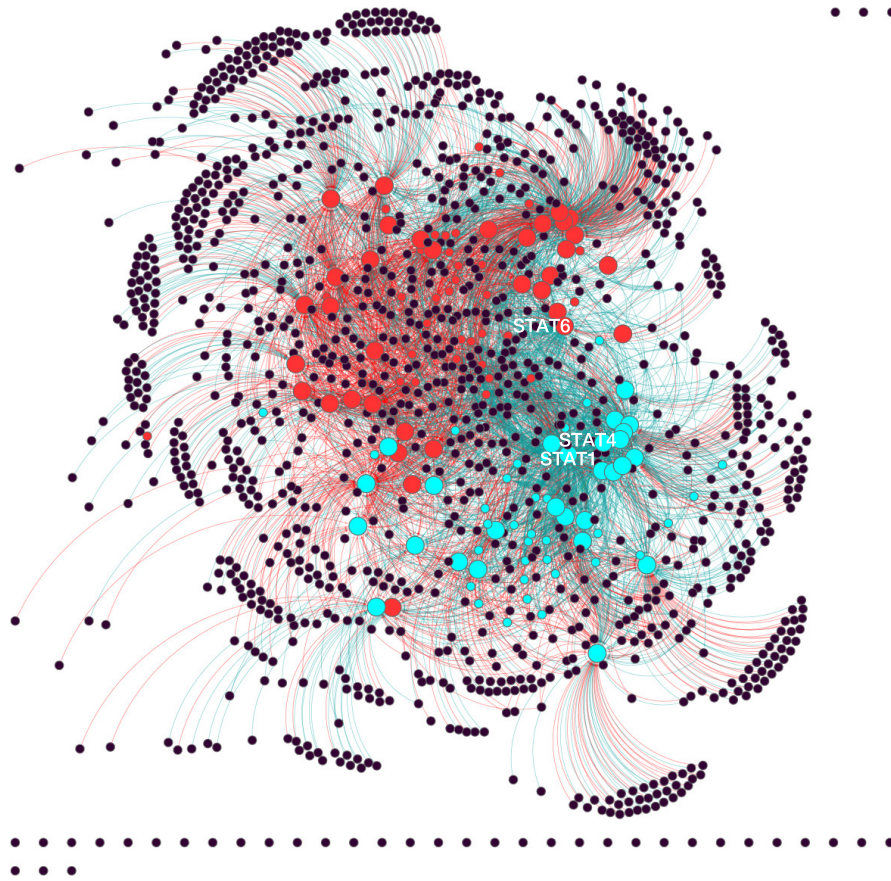


Figure B.16: Rearranged full network after the deletion of the two Th1 and Th2 master TFs Tbet and Gata3.

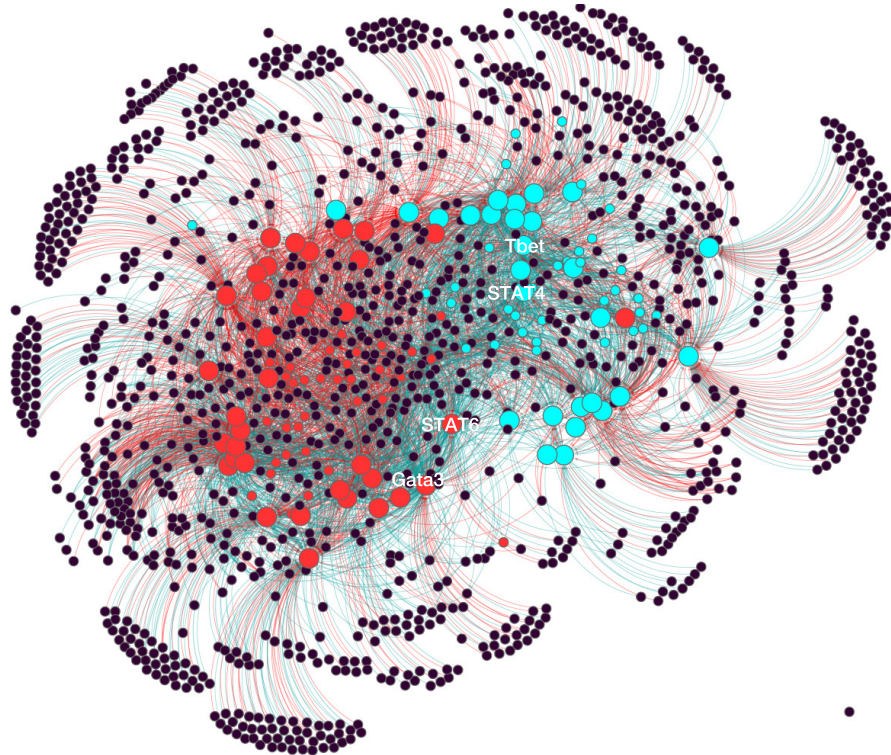


Figure B.17: Rearranged full network after the deletion of STAT1 corresponding to a STAT1 knockout of the system.

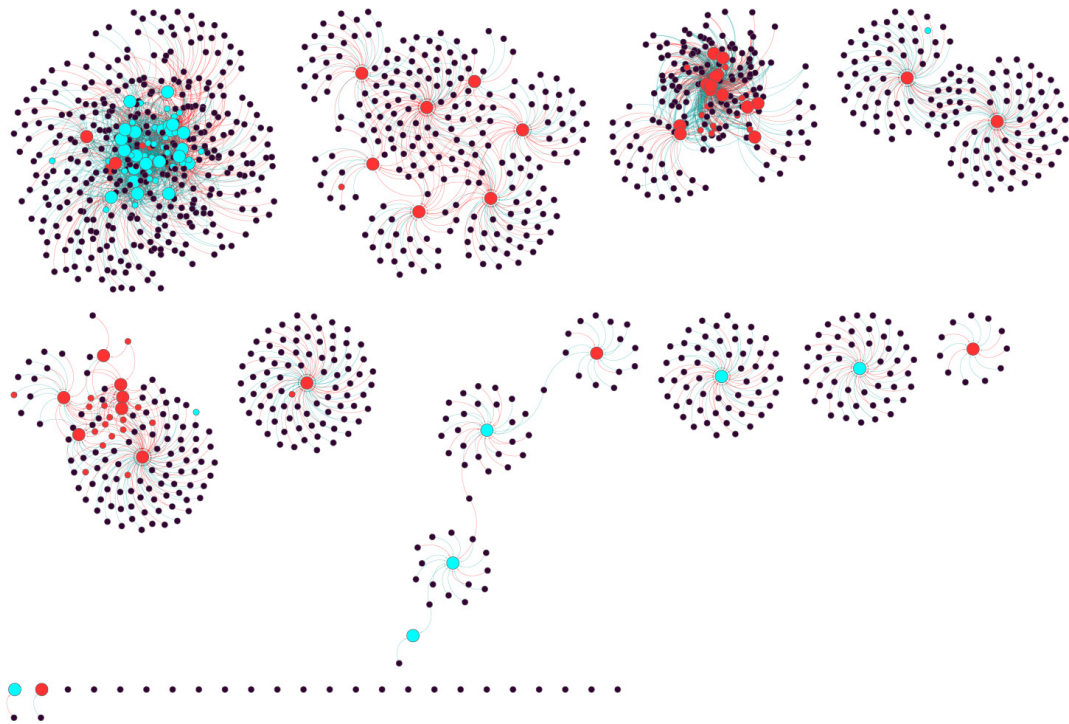


Figure B.18: SC community detection results for $k = 10$.

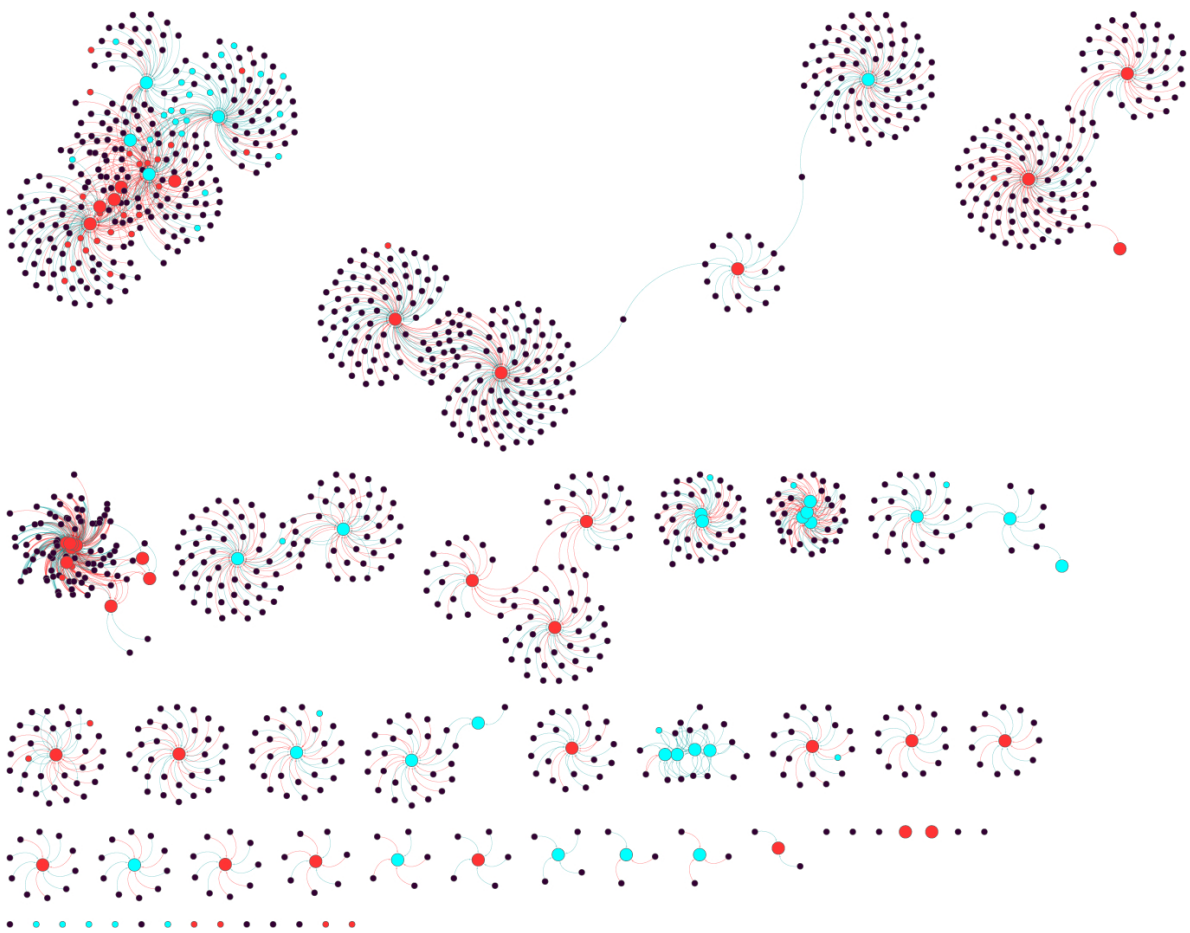


Figure B.19: RWIF community detection results for a granularity parameter of 2.0.

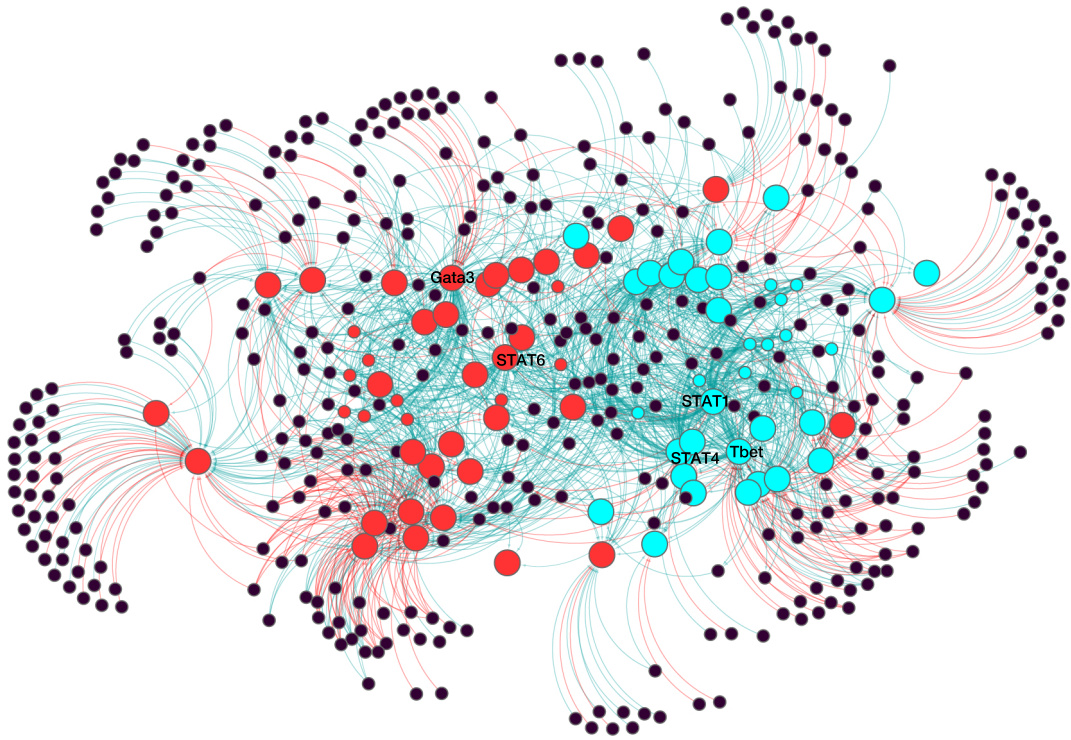


Figure B.20: Force-directed $Tbet^{+/+}Th1/2$ network.

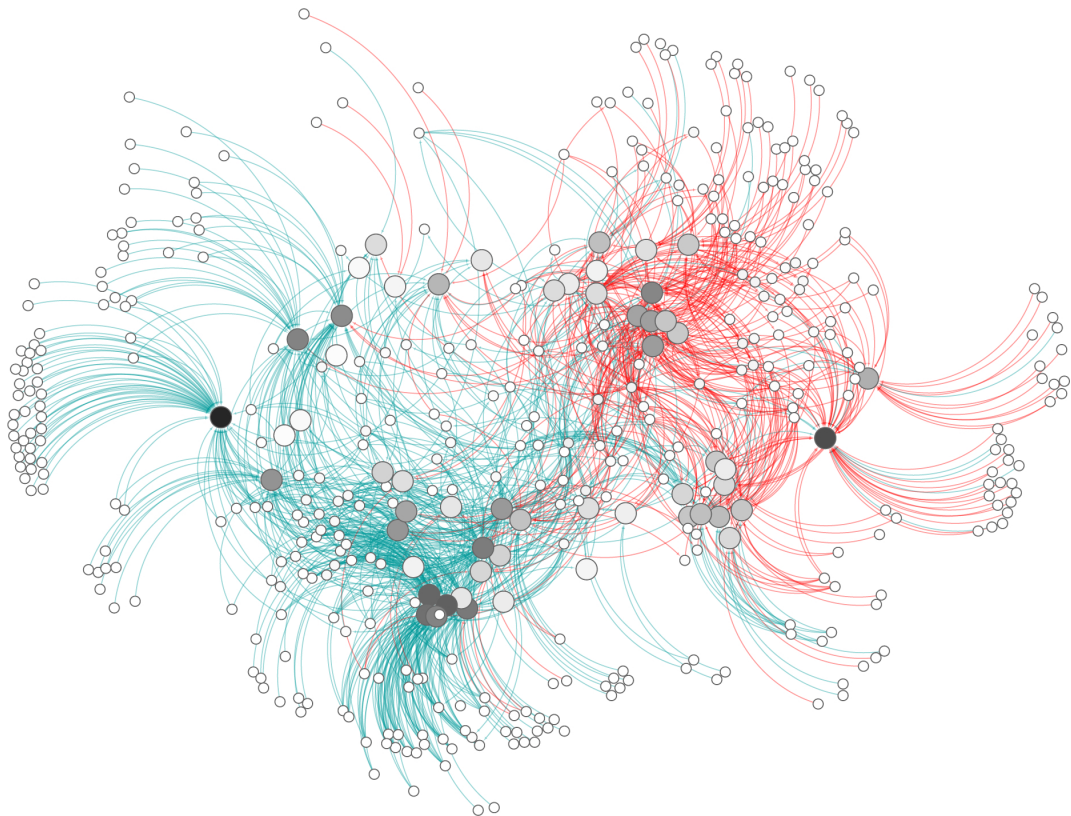


Figure B.21: Differential network of $Tbet^{+/+}Th1/2$ vs. $Tbet^{+/+}Th1$ (*Diff1*) considering edge addition and deletion. The node shadings represent different in-degree centralities while green edges denote upregulation and red edges denote downregulation.

APPENDIX C

Supplementary Tables

Transcript-ID	Transcript name	LFC	LFC SE	LFC/(LFC SE)	p-value	Padj
ENSMUST00000061673.7	Irga1-201	-9.88113721033285	0.371024095045687	-26.6320633680573	2.88799555914734e-156	8.87262962191918e-154
ENSMUST00000050385.5	Klri2-001	-9.86952790291867	0.445269047157455	-22.1653132323582	7.42516333616262e-109	1.36315132759259e-106
ENSMUST00000032270.8	Klrc1-002	-9.1483838625689	0.285819083718541	-32.0076033536576	8.54692134107669e-225	4.59519120959173e-222
ENSMUST00000145984.3	Klrc2-004	-8.23727753929038	0.523424994208697	-15.7372644226578	8.39972419051269e-56	6.07930038288356e-54
ENSMUST00000030709.8	Smpd13b-001	-7.82608422945899	0.506544850580456	-15.4499334471389	7.55213917517703e-54	5.21513317170252e-52
ENSMUST00000032207.8	Klrg1-001	-7.63050612295365	0.516133769049	-14.7839699328591	1.85879260913785e-49	1.14681409581808e-47
ENSMUST00000035938.2	Ccl5-001	-7.5890917907932	0.254136762542819	-29.8622352581301	6.08986954839057e-196	2.69637929945505e-193
ENSMUST00000032374.7	Kcnj8-201	-7.47206774937639	0.489813196173165	-15.2549335292608	1.5265039279163e-52	1.00789430398474e-50
ENSMUST00000025778.7	Gldc-201	-7.44930505897877	0.544848327767074	-13.6722546059467	1.48714938843434e-42	7.80053898727894e-41
ENSMUST00000056614.6	Cxcr3-001	-7.32756660832851	0.181339538772451	-40.4079918694582	0	0
ENSMUST00000018485.3	Il12rb2-001	-7.08432416363057	0.245973473625911	-28.8053872359701	1.83632201577739e-182	6.9109979063782e-180
ENSMUST00000100542.5	Ly6c2-001	-6.8490948131921	0.564429145534591	-12.1345519935989	6.92881754854607e-34	2.9548560729692e-32
ENSMUST00000065289.4	Clec12a-001	-6.69510904466732	0.539544117804052	-12.4088259397887	2.34059724674825e-35	1.03027341966515e-33
ENSMUST00000071920.6	Klrc2-201	-6.58391896929435	0.610707676799648	-10.7808026972851	4.24168013562854e-27	1.41583709006102e-25
ENSMUST00000121805.3	Dpysl3-001	-6.53764780072107	0.518572618607103	-12.6070053954667	1.93195228824626e-36	8.65583623430332e-35
ENSMUST00000028378.3	Galnt3-001	-6.488621444395227	0.302138464671565	-21.4756550477796	2.62986282549841e-102	4.44830954775877e-100
ENSMUST00000053708.6	Klre1-201	-6.36583156133598	0.582306870261048	-10.9320907693948	8.09612165811331e-28	2.7699776236645e-26
ENSMUST00000051014.1	Exph5-001	-6.26476948278762	0.466979943123212	-13.4149830315949	4.94046172585871e-41	4.5457791488703e-39
ENSMUST00000167691.4	Klrl1c-003	-6.13770330835277	0.628979948521249	-9.7581859688575	1.701717664050791e-22	4.73524172176114e-21
ENSMUST00000033545.5	Rab39b-001	-6.13689640818982	0.375989202117812	-16.3220017320255	6.88407590301832e-60	5.54186516813037e-58
ENSMUST00000050020.6	Amica1-001	-6.11534147297737	0.564208965516968	-10.8387881914887	2.25437338791147e-27	7.60926838152898e-26
ENSMUST00000031003.7	Ppp2r2c-001	-6.05206004458122	0.632619528688636	-9.56666648771751	1.1040817828119e-21	2.97331791743297e-20
ENSMUST00000019069.3	Heat9-001	-6.01819435945206	0.565733277070517	-10.6378652332695	1.98627964655086e-26	6.41662098694777e-25
ENSMUST00000025294.7	Ttc39c-001	-5.85501070514663	0.387066236870416	-15.1266376331005	1.08078160596474e-51	7.04332740094944e-50
ENSMUST00000032252.5	Klrl1-001	-5.80524874210694	0.642480603843651	-9.03567937674218	1.62986829174325e-19	4.04218076835302e-18
ENSMUST00000020668.1	Havcr2-001	-5.71783309056336	0.512996245764929	-11.1459550391787	7.49359969247585e-29	2.7117463887147e-27
ENSMUST00000011400.7	Adam19-001	-5.66893064364744	0.138119286673414	-41.0437295194823	0	0
ENSMUST00000009340.9	Mnda-001	-5.66771987827901	0.348488209484236	-16.2637349672956	1.78516046000028e-59	1.39242515880022e-57
ENSMUST00000001484.2	Tbx21-001	-5.64112913430323	0.66249671773026	-8.51495408706319	1.66656320185445e-17	1.66128337310841e-16
ENSMUST00000024936.5	Prss30-001	-5.57702193822686	0.43800881268682	-12.7326686209493	3.8930194547051e-37	1.7813226404599e-35
ENSMUST00000035419.5	Prfl1-201	-5.57560264920329	0.343708202670179	-16.2219074374364	3.53057720977696e-59	2.73965511938053e-57
ENSMUST00000064460.5	Stegalnac3-201	-5.46385479851887	0.589441708717446	-9.26954220869026	1.869465640523e-20	4.78621356333898e-19
ENSMUST00000139848.3	Rasd2-001	-5.43932040804226	0.659270930684648	-8.25050848578074	1.57721976950316e-16	3.25699127710571e-15
ENSMUST00000070166.5	Gramd3-201	-5.43287423422247	0.111838135450305	-48.5780115373664	0	0
ENSMUST00000027146.4	Ikzf2-001	-5.41485575138897	0.205455446732907	-26.3553769807248	4.4534235706571e-153	1.26494034903848e-150
ENSMUST00000108044.3	Il1lr1-202	-5.40328621354871	0.116727378873289	-46.2897930691489	0	0
ENSMUST00000105884.1	Zfp683-001	-5.39725973901454	0.625805793360999	-8.62449628346777	6.43753709561559e-18	1.43147242891281e-16
ENSMUST00000117441.3	Il12rb2-002	-5.39301422683712	0.473590131569076	-11.3875139436906	4.82561789977306e-30	1.81612129657959e-28
ENSMUST00000058748.1	Fam124b-001	-5.34959170314317	0.386341721583675	-13.8467874533829	1.33029902739441e-43	7.25591360811428e-42
ENSMUST00000015620.6	Prrtl-001	-5.3451112893857	0.664469000358217	-8.04418458423815	6.88215727905401e-16	1.71975257472209e-14
ENSMUST00000040227.1	Cldnd2-201	-5.23706875147392	0.669757452186229	-7.81935122092188	5.30962749946346e-15	9.97891789974068e-14
ENSMUST00000072729.5	Ms4a4c-001	-5.21430103797449	0.199405789936567	-26.1491957662473	1.00649842506135e-150	2.75487768924973e-148
ENSMUST00000034944.4	Dapk2-001	-5.17609952292115	0.434684560591168	-11.9077142189768	1.0789452468989e-32	4.41370699641741e-31
ENSMUST00000100526.2	Gml10874-201	-5.17123939343241	0.672903731548327	-7.68496168320181	1.5304279153677e-14	2.77913894305733e-13
ENSMUST00000111214.2	Ifi204-001	-5.1636886199818	0.22480578859196	-22.9695444184358	9.3991895976484e-117	1.88660533603999e-114
ENSMUST00000038144.8	Esm1-001	-5.12995515919527	0.278836631379001	-18.397708844153	1.37037745346253e-75	1.44263371919055e-73
ENSMUST00000114468.4	Osbpl3-003	-5.12326467569933	0.294506515591369	-17.3960995919286	8.83155303571716e-68	8.15645395009099e-66
ENSMUST00000046614.9	Gipc2-001	-5.09736882097307	0.639839645138736	-7.96663485875091	1.63053533508507e-15	3.1836678254696e-14
ENSMUST00000085077.3	Arnt2-201	-5.08386977836995	0.427860017978872	-11.8820865814599	1.4666006092735e-32	5.9509880646989e-31
ENSMUST00000148568.1	Hopxos-001	-5.08117644297651	0.457131328278888	-11.1153537914526	1.05631104265623e-28	3.78612058003498e-27

Table C.1: Top 50 up-regulated transcripts for the differential Th1/Th2 comparison.

Transcript-ID	Transcript name	LFC	LFC SE	LFC/(LFC SE)	p-value	P _{adj}
ENSMUST00000173287.3	Adamts13-001	10.7694050466739	0.35352585423062	30.4628499381224	8.09601382244724e-204	3.80866850259753e-201
ENSMUST00000192047.1	Sell-001	9.4035280071118	0.471002334514364	19.9649286596583	1.11186437533856e-88	1.49446484878096e-86
ENSMUST00000068581.7	Cja1-201	8.98624882163545	0.534721045744783	16.8054893166193	2.22469178251036e-63	1.93586763548618e-61
ENSMUST00000034874.9	Cyp11a1-001	8.79541378922041	0.407818075805384	21.5670033061941	3.66692861896108e-103	6.41883063137676e-101
ENSMUST00000170941.1	Tremf2-201	8.58898610662184	0.261959201513449	32.7874953695066	8.87633747386486e-236	4.94905127153932e-233
ENSMUST00000015540.2	Cd83-201	8.52343091219566	0.465091704273083	18.3263447485424	5.10050442021541e-75	5.22333289400835e-73
ENSMUST000000165205.1	Lrrc32-201	8.44845135611195	0.560638833834971	15.0693295687734	2.57714927423783e-51	1.6369791212817e-49
ENSMUST00000003509.9	St8sia6-001	8.43668652250956	0.32353926557469	26.076236859609	6.7829933269984e-150	1.79142423762516e-147
ENSMUST00000032386.6	Bhlhe41-001	8.43160087046705	0.516744352422674	16.316774108777	7.49953100868216e-60	6.00520956407985e-58
ENSMUST00000148750.3	Slc4a4-001	8.18021056604396	0.555724002494123	14.7199158743021	4.80257398484353e-49	2.9035320790295e-47
ENSMUST0000017637.8	Igfbp4-001	8.03284459230112	0.526879900343273	15.2460638317529	1.74861512914928e-52	1.14950446808268e-50
ENSMUST00000076840.7	Efn5-001	7.99638552677892	0.455094586200638	17.5708210320339	4.12164083757555e-69	3.90233843829323e-67
ENSMUST00000125209.3	Mctp1-001	7.86892445683219	0.502736471376478	15.6521853990169	3.21003995215982e-55	2.30114006856257e-53
ENSMUST00000172167.3	Inpp4b-008	7.68482923806959	0.397535654836175	19.3311697820829	2.93683265151189e-83	3.59439664519187e-81
ENSMUST0000009143.7	Bmp7-001	7.68068379341673	0.375018008614069	20.4808398983338	3.19108527065944e-93	4.57510453947688e-91
ENSMUST0000019999.5	D10Bwg1379e-201	7.556939821532	0.388642896654844	19.4444305725813	3.24861253460892e-84	4.07538442466689e-82
ENSMUST00000181289.1	Gm17322-001	7.51536209227018	0.511404533570999	14.6955327904359	6.8854779865893e-49	4.12964086096077e-47
ENSMUST00000161395.1	Gm15947-001	7.34454148444242	0.545518883205493	13.463404678653	2.56804961851213e-41	1.28436607830836e-39
ENSMUST00000063062.8	Chil3-001	7.3124199085865	0.470681570481086	15.5358109753743	1.98544583592114e-54	1.38374544509059e-52
ENSMUST00000023687.7	Ifngr2-001	7.28393878308143	0.325276091179976	22.393096143827	4.59532482774603e-111	8.64725249461109e-109
ENSMUST00000161306.1	Imprss11e-001	7.24207898806631	0.584246929516247	12.3955790303643	2.76143821304192e-35	1.21197349443536e-33
ENSMUST00000109691.2	Rapgef5-201	7.13912514123052	0.473977687017337	15.062154478528	2.87270489864312e-51	1.81704619933502e-49
ENSMUST00000034554.7	Pou2af1-201	7.09561617682322	0.522926919278916	13.5690397935674	6.11174268427461e-42	3.10831670165777e-40
ENSMUST00000092566.6	Slc16a10-201	7.05191263107527	0.590792787874057	11.9363553107195	7.65025704577001e-33	3.15525944019238e-31
ENSMUST00000028233.3	Hc-001	7.03674899842117	0.591381384669333	11.8988250668127	1.20025077094753e-32	4.89663282001197e-31
ENSMUST00000030683.3	Hgf-201	7.02219636915112	0.591673892427131	11.8683559626825	1.72830601057513e-32	6.99406416215e-31
ENSMUST00000150568.3	Il4-003	7.02168218042834	0.5916548290316	11.8678692979168	1.73838753558199e-32	7.01600159802983e-31
ENSMUST00000040489.7	Trpm6-201	6.81526603516788	0.348478169836682	19.5572251724174	3.58038711950085e-85	4.56772438109879e-83
ENSMUST00000069035.4	A630091E08Rik-201	6.74630854541237	0.601840827683696	11.209456446112	3.66436716734378e-29	1.34217477706066e-27
ENSMUST00000111704.3	Rassf8-001	6.66953948516259	0.604297202613922	11.0368531515836	2.53765446918389e-28	8.90486022822711e-27
ENSMUST00000023616.4	Slc15a2-001	6.56582183671873	0.608350034192183	10.7928354856384	3.72127677361079e-27	1.24766370935271e-25
ENSMUST00000103134.3	Ccr7-001	6.53905115302865	0.480491260851883	13.60909486977	3.53595818209031e-42	1.81673428236135e-40
ENSMUST00000001812.4	Smo-001	6.43815104323358	0.310708138068247	20.7208960900131	2.2446885189857e-95	3.31289617302066e-93
ENSMUST00000067230.5	Sox4-001	6.41995325057402	0.577152422099415	11.1234970256578	9.64171259818103e-29	3.46411316116986e-27
ENSMUST00000037941.9	Cd81-001	6.41921662220042	0.283810310868739	22.6179824212563	2.88368556582262e-113	5.49506368060103e-111
ENSMUST00000062613.6	Tdrp-001	6.36722568951879	0.270740135642595	23.5178492261827	2.67892147661925e-122	5.60117832069808e-120
ENSMUST00000030361.9	Bcam-001	6.29681015796802	0.549019529290517	11.4691915715734	1.88411188450479e-30	7.18061273654055e-29
ENSMUST000000033730.2	Grpr-001	6.17409855599955	0.624698598999464	9.88332383950944	4.9174200889073e-23	1.39148199282727e-21
ENSMUST00000183482.1	Htr1b-002	6.16909335062262	0.624365258198057	9.88058395245577	5.05375389230921e-23	1.42470432761841e-21
ENSMUST00000100960.6	Gbp11-001	6.15268871438197	0.297892600279535	20.6540501798583	8.97691282825557e-95	1.31202374482096e-92
ENSMUST00000169159.2	Ms4a1-001	6.11366538456125	0.557504538898086	10.9661266554798	5.56036519005238e-28	1.91546310231232e-26
ENSMUST00000029559.6	Il6ra-201	6.07018689755997	0.208667509492221	29.0902350458459	4.77105596153457e-186	1.89009148539319e-183
ENSMUST00000034026.8	Hpgd-201	6.03629207428551	0.631024648551419	9.56585782844209	1.11274801927862e-21	2.99130512182507e-20
ENSMUST00000099112.2	Itga7-201	5.942059974926	0.444774503128715	13.3597135922299	1.03963385468557e-40	5.11459086550213e-39
ENSMUST000000058714.8	Cd24a-201	5.93180274564668	0.457166299533453	12.9751531372726	1.69257724106248e-38	7.86421536634403e-37
ENSMUST00000110109.3	Plcb4-002	5.8786340147021	0.53915707084122	10.9033792425832	1.11054658639208e-27	3.79096787109894e-26
ENSMUST00000028045.3	Mrc1-001	5.77848936144666	0.640363753414873	9.02376084004078	1.81740494647978e-19	4.47777644260337e-18
ENSMUST00000073957.6	Sema3e-001	5.73362134812645	0.544594588309898	10.5282378326972	6.4021727590657e-26	2.02901702557842e-24
ENSMUST00000069408.5	Folr4-001	5.72085793546611	0.644436496763027	8.87730282844268	6.85011182140162e-19	1.62140854338648e-17
ENSMUST00000023629.8	Pros1-001	5.69611413032183	0.109279233844666	52.124396648118	0	0

Table C.2: Top 50 up-regulated transcripts for the differential Th2/Th1 comparison.

Th1 transcript	Th2 transcript
Ifng-201	A430108G06Rik-002
Tbx21-001	Gm12214-001
Runx3-001	Gm17334-201
Runx3-002	Gm22275-201
Cxcr3-001	Il4-001
Il2-001	Il4-003
Il12rb2-001	Il5-001
Il12rb2-002	Il13-001
Eomes-001	Irf1-002
Eomes-003	Irf1-008
Ccl5-001	Kif3a-001
Klri2-001	Kif3a-005
Itga1-201	Rad50-001
Klrc1-001	Sept8-002
Klrc1-002	Sept8-004
Klrc1-201	Gata3-001
Klrc1-202	Il10-001
Klrc2-201	Ccr4-201
Klrc2-001	Ccr1-201
Klrc2-002	Areg-201
Klrc2-004	Pparg-202
Stat1-001	Pparg-201
Stat1-004	Il9r-004
Stat1-006	Il9r-003
Stat1-007	Il9r-001
Stat1-008	Asb2-002
Stat1-009	Asb2-001
Stat4-001	Lrrc32-201
Stat4-002	Adamtsl3-001
Il18r1-202	Adamtsl3-005
Il18rap-001	Gja1-201
Ccr2-001	Trem2-201
Ccr5-001	Sell-001
Fasl-001	Sell-002
Fasl-002	Sell-003
Smpd13b-001	Stat6-001
Klrg1-001	Il1rl1-001
Kcnj8-201	Il1rl1-002
Gldc-201	Il1rl1-003
Ly6c2-001	Ifngr2-001
Clec12a-001	Chil3-001
Dpysl3-001	Inpp4b-008
Galnt3-001	Mctp1-001
Klre1-201	Efna5-001
Exph5-001	Igfbp4-001
Klrb1c-003	Slc4a4-001
	Bhlhe41-001
	St8sia6-001
	Cd83-201
	Cyp11a1-001

Table C.3: List of relevant 46 Th1 and 50 Th2 transcripts as determined in [334] and from our RNA-Seq data sets.

HMM state	Naïve	Th2	Tbet ^{+/-} -Th1	Tbet ^{+/-} -Th1/2	Tbet ^{-/-} -Th1	Tbet ^{-/-} -Th1/2	Tbet ^{+/+} -Th1	Tbet ^{+/+} -Th1/2
1	1.53646	2.37870	2.22236	1.62183	1.90720	1.68000	2.32633	1.63504
2	0.02016	0.08171	0.01143	0.02341	0.01418	0.04355	0.01990	0.02447
3	0.41723	0.29655	0.17644	0.12325	0.24216	0.23621	0.21328	0.29822
4	0.19722	0.36155	0.37282	0.36797	0.42040	0.42252	0.36011	0.35185
5	0.01239	0.08296	0.01780	0.06941	0.01309	0.07045	0.02127	0.02814
6	0.01700	0.06748	0.02310	0.03640	0.03574	0.05195	0.03149	0.03574
7	0.17757	0.09676	0.07846	0.03358	0.07442	0.06643	0.06099	0.06248
8	0.15488	0.21591	0.19913	0.09997	0.14805	0.10007	0.27153	0.14853
9	0.15384	0.03300	0.05054	0.01233	0.03610	0.01592	0.03932	0.05706
10	3.00115	1.17962	1.88272	1.21598	2.09157	1.29406	1.58059	1.85226
11	2.42636	0.28199	0.55016	0.41549	0.50666	0.41698	0.29073	0.65364
12	0.39200	0.05975	0.12564	0.04315	0.07738	0.04784	0.07420	0.16404
13	6.39794	6.60396	3.56356	3.32295	4.92826	3.60929	7.03653	4.62131
14	83.67664	87.58462	89.18898	92.08515	88.54472	91.40598	86.61357	89.73587
15	0.09743	0.17026	0.04971	0.03357	0.03728	0.07087	0.02196	0.02079
16	1.32173	0.50519	1.48714	0.49558	0.92278	0.46790	1.03818	0.31056

Table C.4: Percentage of genome occupancy of each HMM state in every experimental condition.

k	BIC
5	$3.252 \cdot 10^7$
6	$2.754 \cdot 10^7$
7	$2.448 \cdot 10^7$
8	$2.288 \cdot 10^7$
9	$2.142 \cdot 10^7$
10	$2.108 \cdot 10^7$
11	$2.042 \cdot 10^7$
12	$2.012 \cdot 10^7$
13	$1.992 \cdot 10^7$
14	$1.948 \cdot 10^7$
15	$1.912 \cdot 10^7$
16	$1.894 \cdot 10^7$
17	$1.881 \cdot 10^7$
18	$1.875 \cdot 10^7$
19	$1.864 \cdot 10^7$
20	$1.852 \cdot 10^7$
21	$1.844 \cdot 10^7$
22	$1.841 \cdot 10^7$
23	$1.839 \cdot 10^7$
24	$1.837 \cdot 10^7$
25	$1.836 \cdot 10^7$

Table C.5: Corresponding BIC scores for different models with parameter number k . We discarded the AIC since the BIC penalizes the number of parameters even stronger yet the respective values of Δ BIC are still exceedingly large.

Transcript	Enhancer
<i>Ifnγ-201</i>	CNS-54
<i>Ifnγ-201</i>	CNS-6
<i>Ifnγ-201</i>	Intron
<i>Ifnγ-201</i>	CNS+18-20
<i>Ifnγ-201</i>	CNS+29
<i>Ifnγ-201</i>	CNS+40
<i>Ifnγ-201</i>	CNS+46
<i>Ifnγ-201</i>	CNS+54
<i>Tbx21-001</i>	-11.9 kb
<i>Tbx21-001</i>	-13.8 kb
<i>Il4-003</i>	HS1
<i>Il4-003</i>	CNS2
<i>Il10-001</i>	-9 kb
<i>Il10-001</i>	+6.45 kb

Table C.6: List of independently validated Th1 and Th2 enhancers used for parametrical learning of the histone modification correlation measure.

Transcript	Partial correlation	<i>p</i> -value
<i>Il4-001</i>	-0.50885	0.13309
<i>Il4-003</i>	0.96622	$5.4697 \cdot 10^{-6}$
<i>Il5-001</i>	0.68426	0.029074
<i>Il13-001</i>	-0.56023	0.092109
<i>Rad50-001</i>	0.30169	0.39692
<i>Sept8-002</i>	0.047332	0.89669
<i>Sept8-004</i>	0.26257	0.46362

Table C.7: Partial correlation values and corresponding *p*-values of the enhancer segment *chr11:53623600-53628000* with co-regulated gene-transcripts.

Gini ESC ranking	Gini value	IG ESC ranking	IG value
"0-0-1-1-0-0-0-0"	3.802861764549086976e-02	"0-0-1-1-0-0-0-0"	3.566208694309333516e-02
"0-0-1-1-1-0-0-0"	2.793203035456571712e-02	"0-1-0-0-0-0-0-0"	2.744356014809607239e-02
"0-0-1-0-1-0-0-0"	2.507822630430756733e-02	"0-0-1-1-1-0-0-0"	2.681337743478382635e-02
"0-0-1-0-0-0-0-0"	2.351054320844190096e-02	"0-0-1-0-1-0-0-0"	2.303623118933336572e-02
"0-0-1-1-0-1-0-0"	2.244676573823319726e-02	"0-0-1-0-0-0-0-0"	2.139103184843979197e-02
"0-1-0-0-0-0-0-0"	2.237943723536544097e-02	"0-0-1-1-1-0-1-0"	2.018495645575485095e-02
"0-0-1-1-1-0-1-0"	2.130107183159843376e-02	"0-0-1-1-0-1-0-0"	1.985132533399914731e-02
"0-0-1-1-1-1-0-0"	1.910379787802115453e-02	"0-0-1-1-1-1-0-0"	1.776003852997779311e-02
"0-2-1-0-0-0-0-0"	1.513752024583004015e-02	"0-2-1-0-0-0-0-0"	1.491147214453106100e-02
"0-0-1-0-0-0-1-0"	1.326601481680461138e-02	"0-0-0-0-0-0-0-1"	1.472737856771999185e-02
"0-0-1-1-1-1-1-0"	1.310524163210281465e-02	"0-0-2-0-2-0-0-0"	1.392336688587722353e-02
"0-0-1-0-1-1-1-0"	1.257446606686534255e-02	"0-1-0-0-0-0-0-1"	1.269001530276536900e-02
"0-0-2-0-2-0-0-0"	1.167350358833286854e-02	"0-0-1-0-0-0-1-0"	1.243364691542413539e-02
"0-0-0-0-0-0-0-1"	1.166953784298974638e-02	"0-0-1-1-1-1-1-0"	1.185559839828293806e-02
"1-1-0-1-1-1-0-0"	1.163510533584564619e-02	"0-0-1-0-1-1-1-0"	1.124031811266250347e-02
"0-2-0-0-0-0-0-0"	1.162945592692493472e-02	"0-1-0-1-1-1-1-1"	1.096307680407765210e-02
"0-2-1-0-1-0-0-0"	1.037571590517448981e-02	"0-2-0-0-0-0-0-0"	1.069473244330305071e-02
"0-0-1-0-1-1-0-1"	1.035265793859341601e-02	"1-1-0-1-1-1-0-0"	1.028752507535200290e-02
"0-1-0-0-0-0-0-1"	1.012320421940228823e-02	"0-2-1-0-1-0-0-0"	9.802646567220329663e-03
"2-0-1-1-0-0-0-0"	1.005482635414427055e-02	"0-0-2-0-0-0-0-0"	9.396647845897089513e-03
"1-0-1-0-1-1-1-1"	9.985329147152352189e-03	"0-0-1-0-1-0-1-0"	9.281930127025800337e-03
"0-0-1-0-1-0-1-0"	9.839018095844235048e-03	"0-0-1-0-1-1-0-1"	8.990944204389098374e-03
"1-0-1-1-1-1-1-0"	9.690233177113420637e-03	"1-0-1-1-1-1-1-0"	8.872460765170455191e-03
"0-1-0-1-1-1-1-1"	8.544037886329047468e-03	"2-0-1-1-0-0-0-0"	8.660247110103184640e-03
"0-0-1-1-0-1-1-0"	8.320978172828669486e-03	"0-1-0-1-0-1-0-1"	8.648889635130115050e-03

Table C.8: Inter-class Gini impurity and information gain ranking for the Top 25 ESCs. ESCs are denoted with a ternary number code where 1 means active enhancer state (before: green state), 2 means repressive state (before: red state) and 0 means none of both. The order from left to-right of the conditions is the same as in the earlier Gini impurity rankings from top to bottom. Hence we have: Naïve, Th2, Tbet^{+/+}Th1, Tbet^{+/+}Th1/2, Tbet^{+/-}Th1, Tbet^{+/-}Th1/2, Tbet^{-/-}Th1, Tbet^{-/-}Th1/2.

APPENDIX C. SUPPLEMENTARY TABLES

ESC	Gini impurity		
"0-0-1-1-0-0-0"	3.802861764549086976e-02	"0-0-0-0-0-2-0-0"	3.255416127928973096e-03
"0-0-1-1-1-0-0-0"	2.793203035456571712e-02	"1-1-1-0-1-0-0-0"	3.229597616419088366e-03
"0-0-1-0-1-0-0-0"	2.507822630430756733e-02	"0-1-1-1-1-0-1-0"	3.208521402435159185e-03
"0-0-1-0-0-0-0-0"	2.351054320844190096e-02	"0-0-2-0-0-0-2-0"	3.193462629850407283e-03
"0-0-1-1-0-1-0-0"	2.244676573823319726e-02	"0-2-0-0-0-2-0-0"	3.171124836861863219e-03
"0-1-0-0-0-0-0-0"	2.237943723536544097e-02	"0-0-1-1-1-1-2-0"	3.145664460309346901e-03
"0-0-1-1-1-0-1-0"	2.130107183159843376e-02	"0-0-1-1-0-0-2-0"	3.128661290455589297e-03
"0-0-1-1-1-1-0-0"	1.910379787802115453e-02	"0-1-0-0-0-1-1-1"	3.107186699978411234e-03
"0-2-1-0-0-0-0-0"	1.513752024583004015e-02	"1-0-1-0-0-0-0-0"	3.026667746154354261e-03
"0-0-1-0-0-0-1-0"	1.326601481680461138e-02	"1-1-1-0-1-0-0-0"	3.006722020252121028e-03
"0-0-1-1-1-1-1-0"	1.310524163210281465e-02	"0-2-1-0-1-0-1-0"	2.999886095745582255e-03
"0-0-1-0-1-1-1-0"	1.257446606686534255e-02	"0-2-0-1-1-0-0-0"	2.929438924188270971e-03
"0-0-2-0-2-0-0-0"	1.167350358833286854e-02	"0-0-0-0-0-1-1-1"	2.912918211145241933e-03
"0-0-0-0-0-0-0-1"	1.166953784298974638e-02	"0-0-0-1-1-1-1-1"	2.876898565459942454e-03
"1-1-0-1-1-1-0-0"	1.163510533584564619e-02	"0-1-1-1-1-1-1-1"	2.780402420287593750e-03
"0-2-0-0-0-0-0-0"	1.162945592692493472e-02	"2-0-1-1-1-1-1-1"	2.721710809729180423e-03
"0-2-1-0-1-0-0-0"	1.037571590517448981e-02	"1-0-1-1-0-0-0-0"	2.696793052312602013e-03
"0-0-1-0-1-1-0-1"	1.035265793859341601e-02	"0-1-0-0-2-0-0-0"	2.612211435118915424e-03
"0-1-0-0-0-0-0-1"	1.012320421940228823e-02	"1-1-1-1-1-1-0-1"	2.610391861932187182e-03
"2-0-1-1-0-0-0-0"	1.005482635414427055e-02	"0-1-0-0-0-1-0-1"	2.573663758756408796e-03
"1-0-1-0-1-1-1-1"	9.985329147152352189e-03	"0-1-1-1-0-1-1-1"	2.557255110648452238e-03
"0-0-1-0-1-0-1-0"	9.839018095844235048e-03	"0-0-0-1-0-1-0-0"	2.53520521148662801e-03
"1-0-1-1-1-1-1-0"	9.690233177113420637e-03	"2-2-0-0-0-0-0-0"	2.476466078105282308e-03
"0-1-0-1-1-1-1-1"	8.544037886329047468e-03	"0-0-0-1-0-1-0-1"	2.437442336624923416e-03
"0-0-1-1-0-1-1-0"	8.32097817282869486e-03	"1-0-0-0-0-1-0-0"	2.431547755658995495e-03
"1-0-1-1-1-1-1-1"	8.175893979845054102e-03	"0-0-0-1-0-0-2-0"	2.375709057936025922e-03
"2-0-1-1-1-0-0-0"	7.972899195252369162e-03	"2-2-0-0-0-0-2-2"	2.355778103350880150e-03
"0-0-0-2-0-0-0-0"	7.869750158875352120e-03	"1-0-0-0-1-0-0-0"	2.327256700030807396e-03
"0-0-2-0-0-0-0-0"	7.667362706324573098e-03	"1-1-0-1-0-1-1-1"	2.307102411900818686e-03
"1-1-1-1-1-1-1-1"	7.415873501928987865e-03	"1-0-0-0-0-1-1-0"	2.256989851077387862e-03
"1-0-1-1-1-0-0-0"	7.152220647236272322e-03	"1-0-0-1-1-0-0-0"	2.224073046180088421e-03
"0-1-1-1-1-1-1-0"	7.107545329162461012e-03	"2-0-1-0-0-0-0-0"	2.210704649673765274e-03
"1-0-0-0-1-0-0-1"	6.903346000839438203e-03	"0-0-0-2-0-2-0-0"	2.180415000282979587e-03
"0-1-0-1-0-1-0-1"	6.885053935952274370e-03	"0-0-2-0-2-0-1-0"	2.179431104186891656e-03
"1-1-0-0-1-0-0-0"	6.567006907369595491e-03	"0-0-2-2-2-0-0-0"	2.148172062173052572e-03
"0-0-0-0-1-1-0-0"	6.443002625796823840e-03	"1-1-0-0-0-0-0-0"	2.110653168652559027e-03
"0-0-0-0-0-0-0-0"	6.316536983042748965e-03	"1-1-0-1-1-1-1-1"	2.108568273023932790e-03
"0-2-1-1-1-0-0-0"	6.135427340668114966e-03	"0-1-0-0-0-0-1-0"	2.078602225794011667e-03
"1-0-1-0-1-1-0-0"	6.113789152760756851e-03	"1-1-1-1-1-0-0-0"	2.062182738864153965e-03
"0-0-1-1-0-1-0-0"	5.961357171737862240e-03	"1-0-1-0-1-0-0-0"	2.041644910991007034e-03
"1-0-0-0-1-1-0-0"	5.931662770308820075e-03	"0-1-0-0-1-0-1-1"	2.037178869687213904e-03
"1-1-1-0-1-0-1-0"	5.910211780915261559e-03	"0-0-2-1-2-1-1-1"	1.92171562154879577e-03
"0-2-0-0-1-0-0-0"	5.898262929439492853e-03	"1-1-0-0-0-0-1-0"	1.865913087050020697e-03
"0-0-1-0-1-1-1-1"	5.885487857535035179e-03	"0-1-1-1-0-0-0-1"	1.863912793679301004e-03
"1-1-0-1-0-1-0-0"	5.730765271233393028e-03	"0-2-1-2-0-2-0-0"	1.826790932509838239e-03
"0-1-0-1-0-0-0-1"	5.626815741809439685e-03	"1-0-0-1-0-0-2-0"	1.819109593297561431e-03
"1-1-1-1-1-0-1-0"	5.590715706963020684e-03	"1-0-0-1-0-0-1-0"	1.801328694614692631e-03
"0-0-1-1-1-1-0-1"	5.380900380446183330e-03	"2-0-1-0-1-0-0-0"	1.785664439541832143e-03
"0-0-0-0-0-1-1-1"	5.348123452301503308e-03	"0-0-0-0-1-1-1-0"	1.752003999614429883e-03
"0-0-0-0-1-0-0-0"	5.263616228911117967e-03	"0-0-0-1-0-0-0-2"	1.740440216484795801e-03
"0-0-0-0-0-0-1-0"	5.251078321563670825e-03	"2-1-1-1-0-0-0-0"	1.730340066259539655e-03
"1-0-0-0-1-1-1-0"	5.064947230304499629e-03	"0-1-2-0-0-1-0-1"	1.727118203806616474e-03
"0-1-2-0-0-0-0-0"	4.982754541579766973e-03	"1-0-0-1-1-0-1-0"	1.719682445946222162e-03
"0-2-1-1-0-0-0-0"	4.947647765166825196e-03	"0-2-0-0-2-0-0-0"	1.713854806008292125e-03
"0-1-0-1-0-0-1-1"	4.891479607255552059e-03	"0-1-0-0-0-1-0-0"	1.702757129545797539e-03
"1-0-1-0-1-1-1-0"	4.813213859566753926e-03	"2-2-1-0-0-0-0-0"	1.688464256814478301e-03
"0-0-0-0-0-0-1-0"	4.803037239197656925e-03	"2-0-0-0-0-0-0-0"	1.683427230825546321e-03
"0-0-0-0-1-1-1-1"	4.793226725633079922e-03	"1-0-1-1-0-1-1-0"	1.666621867188302361e-03
"1-0-1-0-1-0-0-0"	4.679817325987039624e-03	"0-0-0-0-1-0-1-0"	1.656760437207322606e-03
"0-1-0-1-0-0-0-0"	4.612284818232827328e-03	"2-2-1-0-1-0-0-0"	1.645927234160789220e-03
"1-0-0-0-1-0-1-0"	4.535734259191358947e-03	"0-0-0-1-1-0-0-0"	1.633594032155686899e-03
"0-0-1-0-1-0-2-0"	4.533622999400435845e-03	"1-0-0-2-0-0-1-0"	1.605943732771373178e-03
"0-0-1-1-1-1-1-1"	4.505724298638865387e-03	"2-0-1-1-1-0-1-1"	1.605420102097889368e-03
"1-0-1-1-1-0-1-0"	4.450668102995452083e-03	"0-1-0-1-1-0-0-1"	1.581444574092137536e-03
"0-0-1-0-0-0-2-0"	4.422841330575289637e-03	"0-0-0-1-1-1-0-0"	1.580928153517508275e-03
"2-0-1-1-1-1-1-0"	4.297232770209263343e-03	"0-0-1-0-1-1-1-1"	1.572806895466880928e-03
"0-1-0-0-0-0-1-1"	4.252737981577586930e-03	"1-2-1-0-0-2-0-0"	1.563523568984651318e-03
"0-0-0-2-0-0-0-2"	3.976719121433802337e-03	"0-0-0-0-0-1-0-0"	1.535746409418876565e-03
"1-1-0-0-1-0-0-1"	3.943641515971364059e-03	"1-2-1-2-0-0-0-0"	1.535414167084742229e-03
"0-0-1-0-1-1-0-0"	3.938789740323417853e-03	"1-1-1-1-1-1-1-0"	1.534637219525917130e-03
"2-2-1-0-1-0-1-0"	3.903377075765371816e-03	"1-2-1-2-0-2-0-0"	1.519478602296308635e-03
"1-0-1-1-1-1-0-0"	3.887336088040332319e-03	"0-2-0-2-0-2-0-0"	1.514288228564632200e-03
"0-1-2-0-2-0-0-0"	3.829035801621729664e-03	"0-1-0-1-1-0-1-1"	1.505445541446416785e-03
"0-1-0-1-0-0-1-0"	3.814173193957432709e-03	"2-0-0-0-0-1-0-0"	1.477078999137116083e-03
"0-0-0-1-0-0-0-0"	3.733385297964809165e-03	"1-1-0-0-1-2-1-0"	1.474173819981073122e-03
"0-0-0-0-0-0-0-2"	3.706697759585449777e-03	"0-1-0-1-0-1-1-1"	1.429317772058547129e-03
"0-0-0-1-0-0-0-1"	3.577502152938325398e-03	"1-2-1-0-0-0-0-0"	1.410440352924401510e-03
"1-1-1-1-0-0-0-1"	3.534546127709830947e-03	"0-0-0-2-0-1-0-0"	1.400399018302346044e-03
"1-0-0-1-1-0-1-1"	3.448106985676647335e-03	"0-0-1-1-0-0-1-1"	1.343630294553439970e-03
"0-1-0-0-1-1-1-1"	3.440562444558588600e-03	"1-1-1-1-1-1-0-0"	1.332307881952155427e-03
"2-0-0-0-1-0-0-0"	3.338099605735622873e-03	"1-1-0-0-0-0-0-1"	1.332287885719712992e-03
"2-2-0-0-0-2-0-0"	3.318714701184941659e-03	"1-0-1-1-0-1-1-1"	1.280434025914260971e-03
"1-1-0-0-0-1-0-0"	3.303090904167034058e-03	"0-1-0-0-0-1-1-0"	1.241975234740232440e-03

"0-1-1-1-0-1-1-1"	1.238813148157800548e-03	"2-2-1-2-1-0-0-2"	6.098292735485797214e-04
"1-0-0-1-0-0-0-0"	1.198663099475052096e-03	"2-0-0-0-1-0-1-0"	6.056692146193652258e-04
"0-1-1-1-1-0-0-0"	1.175829164995518313e-03	"0-0-0-2-0-0-2-0"	5.886800531072417571e-04
"0-1-0-0-1-0-0-1"	1.174684418646486676e-03	"0-0-2-0-0-1-1-0"	5.866800890463031142e-04
"0-1-2-0-0-0-0-1"	1.171236706801835354e-03	"0-0-1-0-0-0-0-2"	5.715032170260514336e-04
"1-0-1-0-0-0-1-0"	1.154795126179178127e-03	"1-0-0-0-1-1-1-1"	5.679831644803019775e-04
"0-2-0-0-2-0-2-0"	1.145567369043649919e-03	"0-1-2-1-0-1-0-1"	5.605976191717010251e-04
"2-1-2-1-0-1-0-1"	1.124789820505849849e-03	"1-0-1-0-1-2-0-0"	5.333726980139855351e-04
"1-1-0-1-0-1-0-1"	1.124518586948588606e-03	"1-0-1-1-1-0-0-1"	5.286100512830027966e-04
"0-0-0-0-1-0-1-1"	1.098878648878804602e-03	"1-0-0-0-0-0-1-1"	5.218944237136508230e-04
"0-2-0-1-0-0-0-0"	1.088316343437201617e-03	"0-0-2-2-2-2-0-0"	5.145984721901509205e-04
"1-0-0-1-1-1-0-0"	1.060517175095384125e-03	"1-0-0-0-0-0-0-1"	5.130009828013096284e-04
"2-2-1-1-0-1-0-0"	1.054130855038260374e-03	"0-0-2-2-2-2-0-0"	5.120780559419803356e-04
"1-1-1-0-1-0-1-1"	1.052129762440301170e-03	"0-0-0-0-1-1-0-1"	5.088730996007346337e-04
"2-0-1-1-0-1-0-0"	1.050458814033474299e-03	"1-1-0-2-0-0-0-0"	5.029836373433783014e-04
"1-0-1-1-0-0-1-0"	1.050453103322121495e-03	"0-1-0-1-1-0-1-0"	4.924263375350729352e-04
"1-0-0-1-1-1-1-1"	1.040609304773767654e-03	"0-0-2-0-0-1-0-0"	4.918487321059653939e-04
"1-1-0-1-1-0-1-0"	1.039804149033626444e-03	"0-1-1-0-1-0-0-0"	4.874202798312369232e-04
"1-2-0-0-1-0-0-0"	1.032405257246198646e-03	"0-1-1-1-0-0-0-0"	4.805556401231196812e-04
"0-0-2-0-2-0-2-0"	1.031354131428687112e-03	"1-0-2-2-2-2-0-0"	4.787832852770260552e-04
"1-1-0-1-0-0-0-0"	1.002144902194226830e-03	"1-1-0-2-2-0-0-0"	4.727526786600742131e-04
"0-2-0-2-2-0-2-0"	9.975358189901863473e-04	"0-0-0-1-1-0-1-0"	4.673811938411460455e-04
"1-2-0-0-2-0-2-0"	9.963438996749508694e-04	"2-2-1-1-0-0-0-2"	4.646909681024174836e-04
"1-0-1-1-1-0-1-1"	9.938994324966471307e-04	"0-2-1-0-0-0-1-0"	4.645309499389131745e-04
"0-1-1-1-1-1-0-1"	9.898821418482827740e-04	"2-1-0-1-0-1-1-1"	4.627057012049213798e-04
"2-1-0-0-0-0-0-0"	9.816850559315702923e-04	"1-1-0-0-2-0-2-0"	4.576206194057584001e-04
"1-0-0-0-1-0-0-0"	9.745466251831127070e-04	"0-1-0-0-1-1-0-0"	4.568618383677017333e-04
"0-1-1-0-0-0-0-0"	9.736797681035254176e-04	"0-0-0-0-1-0-0-1"	4.500930484643724775e-04
"0-0-0-0-0-0-2-0"	9.518139511882349772e-04	"0-2-0-2-0-2-0-2"	4.483416671430280999e-04
"1-0-2-0-0-0-0-0"	9.423557743360281040e-04	"1-0-0-0-0-0-0-2"	4.371365619700917963e-04
"1-0-1-0-0-1-1-1"	9.009785045062505566e-04	"0-2-0-0-0-0-0-2"	4.329122336303392821e-04
"0-0-2-1-0-0-0-0"	8.983562689163552238e-04	"1-0-1-0-0-1-1-0"	4.325080101901058799e-04
"0-0-0-1-0-0-1-1"	8.862739414220073293e-04	"0-1-0-0-1-0-0-0"	4.298933571858166080e-04
"2-1-2-1-0-0-0-1"	8.786911136706738058e-04	"0-2-0-1-0-0-0-2"	4.249152827446902843e-04
"0-1-0-1-0-1-0-0"	8.765645272259956411e-04	"1-1-0-0-2-0-0-0"	4.232477608632473744e-04
"0-0-2-0-2-1-1-1"	8.697513273654336853e-04	"2-1-0-1-0-1-0-1"	4.181171178863156944e-04
"0-1-1-1-0-1-0-1"	8.676785144295577450e-04	"1-1-1-0-0-0-0-0"	4.16039663832493434e-04
"2-0-1-0-1-0-1-0"	8.603640898388814151e-04	"0-0-1-1-0-0-0-1"	4.159046349770162063e-04
"0-1-2-1-0-0-0-1"	8.562554142421866616e-04	"1-1-1-1-0-0-0-1"	4.155746803142742508e-04
"1-0-0-0-0-2-0-0"	8.555645742327527749e-04	"0-0-0-0-0-1-1-0"	4.151035850001529200e-04
"0-1-2-1-0-1-0-0"	8.484492236050886899e-04	"1-1-0-1-1-0-0-0"	4.147645369960962190e-04
"1-0-0-0-0-1-1-1"	8.462576993768257654e-04	"0-0-1-1-1-0-0-1"	4.129314915608391871e-04
"0-1-1-1-0-1-0-0"	8.460097416956374131e-04	"2-0-1-1-1-0-1-1"	4.085625524798387428e-04
"0-0-2-0-2-0-1-1"	8.403067570929579613e-04	"1-1-0-1-0-0-1-1"	4.060720814573056878e-04
"0-0-2-0-0-0-1-0"	8.377974593507037842e-04	"2-0-1-0-0-2-2-0"	4.058252099589620661e-04
"1-0-2-0-2-0-0-0"	8.350616363566310246e-04	"0-0-2-2-2-0-2-0"	4.022746458254268737e-04
"2-0-0-0-0-0-1-0"	8.293749495121350603e-04	"1-1-0-0-0-1-1-1"	3.961180622751633157e-04
"1-0-1-0-1-0-1-0"	8.290784017686373677e-04	"2-2-1-1-1-0-1-0"	3.942277088418248454e-04
"1-1-1-1-0-0-0-0"	8.170087820754428536e-04	"1-0-1-0-0-1-0-1"	3.941888124038696144e-04
"0-0-2-1-0-1-0-0"	8.128825787960872445e-04	"0-0-0-0-2-2-0-0"	3.933762540900088219e-04
"0-2-0-0-0-0-2-2"	8.107099932497982249e-04	"2-0-1-0-0-1-0-0"	3.896255546034654176e-04
"1-0-0-1-1-1-1-0"	8.03186332957864818e-04	"2-0-1-1-1-1-2-0"	3.874837135831673174e-04
"2-0-0-0-1-0-0-2"	7.95659557278034976e-04	"1-0-1-1-0-0-1-1"	3.86757962176224967e-04
"1-1-2-0-2-1-0-0"	7.927005112305681020e-04	"1-1-1-1-0-1-1-1"	3.861282137861579526e-04
"0-0-1-0-1-0-0-1"	7.727796986857827545e-04	"0-1-2-2-2-0-0-0"	3.808363545498816521e-04
"1-1-1-0-1-1-0-0"	7.646435996563743240e-04	"0-0-1-0-0-2-2-0"	3.762192507190657994e-04
"0-1-0-0-1-1-1-0"	7.645931894254468343e-04	"1-1-0-0-0-1-0-1"	3.761766585435694327e-04
"0-1-2-0-0-0-1-1"	7.626952841262842151e-04	"0-0-1-0-1-0-0-2"	3.755649888621947086e-04
"0-0-2-1-0-1-0-1"	7.435281831319815035e-04	"0-2-0-2-0-2-2-0"	3.698162403103677671e-04
"1-2-1-1-1-1-1-1"	7.423213259200802001e-04	"0-1-2-2-0-0-0-0"	3.647541131570058580e-04
"0-0-0-2-0-2-0-0"	7.360374419400354980e-04	"2-0-0-1-1-1-2-0"	3.647415653580643696e-04
"1-0-1-0-1-0-1-1"	7.295993502918246727e-04	"2-1-1-1-1-0-1-0"	3.620108070922493275e-04
"2-2-0-0-0-2-2-0"	7.295115149446949193e-04	"2-0-1-1-0-1-0-0"	3.592955436067518248e-04
"1-1-1-1-1-0-1-1"	7.255940972092138558e-04	"1-1-1-0-1-1-1-0"	3.580198037971481467e-04
"0-0-1-0-0-1-0-0"	7.205999368723647793e-04	"0-2-0-0-2-0-2-0"	3.560764214683422523e-04
"2-2-1-0-0-0-2-0"	7.130274146262993067e-04	"2-0-1-0-0-1-2-2"	3.528102357200577901e-04
"0-2-1-0-1-1-2-0"	7.082533764755194299e-04	"0-0-2-2-0-2-0-0"	3.515918114245249583e-04
"0-0-0-0-2-0-0-0"	7.056316932035956203e-04	"0-0-2-1-0-0-0-1"	3.451473661803044156e-04
"1-2-0-0-0-0-0-0"	6.793400494503035094e-04	"2-0-0-1-1-0-0-0"	3.36798902292393090e-04
"1-0-1-1-1-0-1-1"	6.734131865782161863e-04	"1-1-0-0-1-0-1-1"	3.360856444762706547e-04
"2-2-0-2-0-0-0-2"	6.732287729403011725e-04	"2-0-1-0-1-1-2-2"	3.347242098493244974e-04
"2-2-0-1-0-0-0-0"	6.719972979075171469e-04	"0-0-1-0-1-1-0-2"	3.344935870258891044e-04
"1-0-1-0-1-1-0-1"	6.669675186722715541e-04	"2-1-0-2-0-0-0-0"	3.341864664766503725e-04
"0-0-0-1-1-1-1-0"	6.586043266425181805e-04	"1-1-2-2-0-0-2-0"	3.279300788823727333e-04
"1-0-0-1-0-0-1-1"	6.532260063186922294e-04	"2-0-1-1-0-0-2-2"	3.277301023912869990e-04
"2-0-0-2-0-0-0-2"	6.382432914871250538e-04	"1-1-2-0-0-1-0-1"	3.241286389562740371e-04
"2-2-1-0-0-2-2-0"	6.357425037614070594e-04	"2-0-1-1-1-0-0-0"	3.174096294136009334e-04
"1-1-0-1-1-1-0-1"	6.286127493656162900e-04	"0-0-0-0-1-1-1-2"	3.173370568294103192e-04
"1-1-1-0-1-1-1-1"	6.278791340226273299e-04	"0-0-0-1-0-1-1-1"	3.155038090684859453e-04
"0-0-0-2-2-0-0-0"	6.250628336885808572e-04	"1-1-2-0-2-0-0-1"	3.132122537280270284e-04
"2-0-1-0-0-0-2-0"	6.198745822797746031e-04	"0-0-0-1-0-1-1-0"	3.062599362314028656e-04
"0-0-2-1-2-1-1-0"	6.162531809110030666e-04	"1-0-0-2-2-0-0-0"	2.971946287412095150e-04
"0-0-2-2-0-0-2-0"	6.135303785644268112e-04	"2-2-1-1-1-1-1-0"	2.965688325925133631e-04
"0-1-1-1-0-0-1-0"	6.113518068843842813e-04	"1-0-1-0-1-1-0-2"	2.950485534567023004e-04

APPENDIX C. SUPPLEMENTARY TABLES

"1-2-1-1-1-1-1-0"	2.923880109964559497e-04	"0-1-1-0-1-0-1-1"	1.112219233099856757e-04
"1-1-1-1-0-1-0-1"	2.899449133593628471e-04	"0-1-2-1-0-1-1-1"	1.111739342124111121e-04
"1-2-1-1-1-0-0-0"	2.889617035749171992e-04	"2-0-2-0-2-0-1-0"	1.105913008120799288e-04
"1-1-2-0-0-0-0-1"	2.818387725775144083e-04	"0-0-2-0-0-0-2-2"	1.105755061473580255e-04
"2-1-0-2-0-0-1-1"	2.797223655455626672e-04	"0-2-0-0-1-1-0-0"	1.098588437606900782e-04
"2-0-2-1-0-0-0-2"	2.791386286137151275e-04	"1-1-0-1-1-0-1-0"	1.092990152272424389e-04
"2-2-0-0-1-0-0-0"	2.698193488124791444e-04	"1-1-0-0-1-0-1-0"	1.091066423215419651e-04
"1-0-0-0-1-0-1-1"	2.682090840540364562e-04	"0-0-1-0-1-2-0-0"	1.084008638838882314e-04
"2-1-2-0-2-0-0-1"	2.652876974050219256e-04	"2-1-0-1-0-1-0-0"	1.060382308876450119e-04
"2-0-0-0-1-1-1-0"	2.632614553696687159e-04	"2-0-2-2-0-0-2-0"	1.055070162271064282e-04
"0-0-2-0-2-1-0-0"	2.622804816479547990e-04	"1-1-0-2-2-2-0-0"	1.054251887887687445e-04
"1-0-1-0-1-1-1-2"	2.615680360301458267e-04	"2-0-1-0-0-0-1-0"	1.026091397317242988e-04
"0-0-1-0-0-0-1-1"	2.583150547759415503e-04	"1-1-1-0-0-2-1-1"	1.022373312485556072e-04
"1-0-0-2-2-0-0-0"	2.57765937253613447e-04	"2-2-1-2-1-0-0-0"	1.021253551116844586e-04
"2-1-2-0-0-0-0-0"	2.527658452398541923e-04	"0-2-1-0-1-0-1-0"	1.019707878046558083e-04
"1-1-2-2-2-0-0-0"	2.518645295742151275e-04	"0-2-1-0-1-1-0-0"	9.951754705823618403e-05
"0-0-2-2-0-0-0-0"	2.514805281424835749e-04	"1-2-1-0-1-0-1-0"	9.75085716917669851e-05
"0-0-2-1-2-0-0-0"	2.355685577548477245e-04	"1-0-0-0-2-0-0-0"	9.740251808281662422e-05
"1-1-1-1-0-1-1-0"	2.309594866761659562e-04	"1-1-1-0-0-0-1-1"	9.617954638360599092e-05
"2-1-2-0-2-0-0-0"	2.285913263023863092e-04	"1-1-0-0-0-0-2-0"	9.506571206096810309e-05
"0-2-0-0-0-0-1-0"	2.276744440236288212e-04	"1-2-1-0-1-1-1-0"	9.47322452071614105e-05
"0-2-1-1-1-1-0-0"	2.264814132076733749e-04	"2-2-1-0-0-2-0-2"	9.430285896941086455e-05
"2-0-2-1-0-0-0-0"	2.243881130269444191e-04	"1-1-1-0-0-0-1-0"	9.414636883994090591e-05
"0-0-1-0-0-1-0-1"	2.210655381740926368e-04	"0-1-1-0-1-1-1-1"	9.201469462661022276e-05
"0-0-1-1-1-0-1-1"	2.204847882086239514e-04	"1-2-0-0-1-0-0-1"	9.027862395485786585e-05
"0-1-0-0-2-0-1-0"	2.204344324455154700e-04	"1-1-0-2-2-2-0-0"	9.006136086497142654e-05
"1-1-2-2-2-0-0-2"	2.149487653012603140e-04	"1-1-0-0-0-0-0-2"	8.853513139730956924e-05
"0-1-1-0-1-1-0-1"	2.147845616534816262e-04	"0-0-0-1-2-0-0-1"	8.838774958443216916e-05
"2-1-0-0-0-0-0-1"	2.133769973549537089e-04	"1-2-1-0-0-0-0-0"	8.762774094649592204e-05
"1-0-2-0-2-0-2-0"	2.124688255162513579e-04	"2-0-2-2-2-0-2-0"	8.737953940471526195e-05
"1-1-1-0-1-1-0-1"	2.124213305514641755e-04	"1-0-2-0-0-0-1-0"	8.737690333771448411e-05
"2-2-2-1-2-0-0-2"	2.120880331155758803e-04	"1-0-2-0-0-2-0-2"	8.696764539761677468e-05
"2-1-1-1-1-1-1-1"	2.115188057287490232e-04	"2-2-1-1-1-0-2-2"	8.690512589421269343e-05
"2-2-0-1-1-0-1-0"	2.07751278407274728e-04	"0-1-2-0-0-0-2-1"	8.484738847927961105e-05
"0-1-2-0-2-0-2-1"	2.041481574314464576e-04	"0-2-1-0-1-2-0-0"	8.471875149039330812e-05
"2-2-0-0-0-0-2-0"	2.010322615170620974e-04	"1-0-1-1-1-0-2-0"	8.409944534505637377e-05
"0-1-1-1-1-0-0-0"	1.989006551154897252e-04	"0-2-1-0-0-0-2-2"	8.328552079995243907e-05
"2-0-1-1-0-1-1-0"	1.963428406379258179e-04	"2-2-1-1-1-0-0-2"	8.236954703648423652e-05
"2-2-1-1-0-0-0-0"	1.946599494858719461e-04	"0-0-1-1-1-2-2-0"	8.214925198144793037e-05
"0-0-1-0-1-0-1-1"	1.941488626469296092e-04	"0-1-2-0-0-0-0-2"	8.190916168947381276e-05
"0-0-0-0-1-0-1-1"	1.917975779816406910e-04	"0-2-1-0-1-2-0-2"	8.179868202279428420e-05
"2-2-0-1-1-0-0-0"	1.893221529645688146e-04	"1-0-1-1-1-2-2-0"	8.165311169431293669e-05
"1-0-0-1-0-1-0-0"	1.881931676587739817e-04	"0-2-1-1-1-0-2-0"	8.125910430286086570e-05
"2-0-0-1-0-0-0-0"	1.848190972943753493e-04	"1-1-0-1-0-0-1-0"	8.095487865090211549e-05
"0-1-1-0-1-1-1-0"	1.839028556681824428e-04	"1-0-2-2-2-0-2-0"	8.093268280480370896e-05
"0-1-0-0-1-0-1-0"	1.835743005711974762e-04	"2-2-1-0-2-0-2-2"	8.032710032942149314e-05
"2-2-1-0-0-0-0-2"	1.810943445493525612e-04	"2-2-0-2-0-0-0-0"	7.970087297670881048e-05
"1-0-0-2-0-0-0-0"	1.785693029706623156e-04	"1-1-2-2-0-0-0-0"	7.93308199730530096e-05
"0-0-2-1-0-1-0-0"	1.75377771622086451e-04	"0-0-2-0-2-2-0-0"	7.85430270902117603e-05
"1-0-0-1-0-0-0-1"	1.735643849555763884e-04	"2-2-1-0-1-2-1-0"	7.829665540418167494e-05
"1-0-0-2-2-2-2-0"	1.641587190810523483e-04	"0-1-1-0-0-1-1-1"	7.815208884891926310e-05
"2-0-1-0-0-0-0-2"	1.641317579518421633e-04	"2-2-0-2-0-2-0-0"	7.768430995279944005e-05
"2-0-2-0-0-0-0-0"	1.638054373209146917e-04	"2-2-1-0-1-0-1-2"	7.701997195398496351e-05
"0-0-2-0-2-0-2-2"	1.589609586223750389e-04	"0-2-1-1-1-1-1-0"	7.683692573535867124e-05
"0-1-0-1-1-1-1-0"	1.588614065468758581e-04	"0-2-0-0-1-0-1-0"	7.65842386628763087e-05
"0-1-2-0-2-1-0-1"	1.576685842082757995e-04	"0-1-1-0-0-0-1-1"	7.551886276015367468e-05
"0-0-1-0-0-0-0-1"	1.569386906401566575e-04	"0-0-0-2-2-2-0-0"	7.503447510424817975e-05
"1-0-0-2-0-2-0-0"	1.557885286808258901e-04	"0-1-0-0-2-2-0-0"	7.425238312394681324e-05
"1-0-0-0-0-1-0-1"	1.556138501453966283e-04	"1-1-2-2-2-0-2-2"	7.404368290357767654e-05
"2-1-2-0-2-0-2-0"	1.539844166740612973e-04	"0-2-1-1-1-0-1-0"	7.363779492306930449e-05
"0-0-2-1-0-1-1-0"	1.504766316492242724e-04	"0-1-0-1-2-1-0-1"	7.356355676132535725e-05
"2-1-2-0-2-1-0-1"	1.428491121037825157e-04	"1-2-0-2-2-0-0-0"	7.313772018968100488e-05
"0-1-0-2-0-1-0-0"	1.423988510657214648e-04	"2-0-0-2-0-0-2-0"	7.291748056403892756e-05
"1-0-0-0-2-0-2-0"	1.404036486107080478e-04	"0-2-1-0-1-0-0-2"	7.237759125240347032e-05
"2-2-1-0-1-1-1-0"	1.390002372178961415e-04	"1-0-0-0-1-0-2-0"	7.235242683006603008e-05
"0-1-0-0-2-0-2-0"	1.350927795380420953e-04	"0-1-2-2-0-0-2-1"	7.180619546306866020e-05
"1-1-0-0-2-0-1-0"	1.349175349850627835e-04	"0-2-0-0-1-2-0-2"	7.132282382180084930e-05
"0-1-2-0-2-0-2-0"	1.349087137084117774e-04	"0-2-1-1-0-0-2-2"	7.018737305752769718e-05
"0-0-1-0-2-0-0-0"	1.330728020752360106e-04	"0-0-2-2-0-0-1-0"	7.017553848276667546e-05
"0-0-1-1-0-0-1-2"	1.290470580454682796e-04	"2-2-0-1-1-1-0-0"	7.014890324736316855e-05
"1-0-0-0-0-2-2-2"	1.289399004857325634e-04	"0-2-1-0-0-2-0-2"	7.011607525434761180e-05
"0-1-2-0-2-0-1-0"	1.286980936673062586e-04	"2-2-1-0-0-0-1-0"	6.759872098702092555e-05
"1-1-0-1-0-0-0-1"	1.275730265332529641e-04	"1-0-2-2-0-0-0-0"	6.756140955379292385e-05
"0-2-0-2-0-0-2-2"	1.245718623575348789e-04	"1-1-0-0-0-0-1-1"	6.713022884757892405e-05
"0-0-1-2-0-0-0-0"	1.232971345372777167e-04	"2-2-1-1-0-0-2-2"	6.688161164981324856e-05
"0-1-2-0-0-0-2-0"	1.227314961026832135e-04	"2-1-2-0-2-0-2-1"	6.683273137112255894e-05
"1-0-1-0-1-2-0-1"	1.185500881987491887e-04	"0-2-0-0-1-2-1-2"	6.609648279821289272e-05
"2-2-1-1-1-1-0-0"	1.148360939638184660e-04	"0-2-0-0-1-0-0-2"	6.553659764674976099e-05
"0-0-0-0-2-0-0-2"	1.141646031244035692e-04	"0-2-0-0-0-0-1-2"	6.544654801228142436e-05
"1-0-0-0-0-0-2-0"	1.135976633855073241e-04	"2-0-2-2-2-0-0-0"	6.531234927897961394e-05
"1-0-2-2-0-2-0-0"	1.132310280537858846e-04	"1-0-2-0-1-0-1-0"	6.478253420038747497e-05
"1-0-2-2-2-0-2-2"	1.129920146190785087e-04	"0-2-1-0-0-2-0-0"	6.432112677263016368e-05
"0-0-0-2-0-0-0-1"	1.120309894968079460e-04	"0-1-2-0-0-0-1-0"	6.419852684941583032e-05
"0-0-2-0-2-2-2-0"	1.112958638415902709e-04	"1-2-1-0-0-0-0-2"	6.379206998605402258e-05

"1-1-0-2-0-2-0-0"	6.294310992721710617e-05	"2-1-0-1-1-1-1-1"	2.351779414008386800e-05
"1-1-2-2-2-0-2-0"	6.235329647649696201e-05	"0-1-1-1-0-1-1-0"	2.349977986383005522e-05
"1-1-0-0-2-0-2-2"	6.114874604240231216e-05	"2-1-0-0-2-1-2-1"	2.342028915942585449e-05
"2-0-0-1-0-1-0-1"	6.104029911596123260e-05	"2-0-0-0-0-0-2-0"	2.282025049412910313e-05
"0-1-2-2-0-0-0-1"	5.946588326625740337e-05	"2-0-2-1-0-0-2-0"	2.231249439748070744e-05
"1-1-0-2-0-2-2-2"	5.945944775892515321e-05	"0-0-0-1-0-0-2-1"	2.220816233470131637e-05
"0-1-2-2-0-2-0-0"	5.872115167267263344e-05	"2-0-2-0-0-0-0-1"	2.179424334591297834e-05
"0-2-1-2-1-2-0-0"	5.853699497883076125e-05	"2-1-0-1-1-1-1-0"	2.041984831611825095e-05
"1-1-0-0-0-0-2-2"	5.831915728885013207e-05	"1-0-2-2-2-0-0-0"	2.035097293602215301e-05
"1-1-0-0-0-2-1-0"	5.618344335913046941e-05	"2-0-2-1-0-1-1-1"	1.995272495635394531e-05
"1-2-1-1-1-0-1-1"	5.608682086625653733e-05	"0-1-0-0-1-1-0-0"	1.973996328824174714e-05
"1-1-2-2-0-0-1-1"	5.587642802436226927e-05	"2-1-2-0-2-1-2-1"	1.911456586485541409e-05
"1-1-1-1-0-0-1-0"	5.563284001950707030e-05	"2-1-0-0-2-0-0-0"	1.881114039322979038e-05
"0-1-2-0-0-2-0-0"	5.543293523917976114e-05	"2-1-0-0-2-0-2-1"	1.880954815237732548e-05
"0-2-1-1-0-0-0-2"	5.495057634608807693e-05	"0-0-0-1-2-1-0-0"	1.876709470064077376e-05
"0-2-0-1-1-1-0-0"	5.462864781113100444e-05	"2-1-2-1-2-0-0-1"	1.861321304260054850e-05
"2-2-1-0-1-2-0-0"	5.426491059068312526e-05	"2-0-0-0-0-0-0-0"	1.836545528933941678e-05
"0-2-0-0-0-0-2-0"	5.410842657373018721e-05	"1-0-2-0-0-0-2-0"	1.765034770174715034e-05
"2-0-2-0-0-0-1-0"	5.386162674796424199e-05	"0-1-0-1-2-0-0-1"	1.718721395341780796e-05
"0-2-1-0-0-0-0-2"	5.373856805533378921e-05	"0-1-2-2-2-0-2-0"	1.699497286374649548e-05
"2-1-2-1-0-1-2-1"	5.373825603060334758e-05	"2-0-0-1-0-0-2-0"	1.696073934170004178e-05
"1-1-0-2-0-0-1-1"	5.364313370084309554e-05	"2-1-2-0-0-1-2-1"	1.692750764932039300e-05
"0-2-1-1-1-1-1-1"	5.348875165370289597e-05	"1-1-0-0-2-0-1-1"	1.673264845132362100e-05
"0-0-0-1-1-1-0-1"	5.288298538031525967e-05	"2-1-0-0-0-0-2-0"	1.587048396645381187e-05
"0-2-1-1-1-0-2-2"	5.287957990854752607e-05	"2-1-2-0-2-1-1-1"	1.585676397748900360e-05
"0-2-1-0-0-0-2-0"	5.268324277562110707e-05	"0-0-0-1-1-0-0-1"	1.566629481065551780e-05
"0-0-0-0-0-2-0-2"	5.251147905394761939e-05	"1-0-1-0-1-0-1-2"	1.520301631856531240e-05
"1-1-2-2-2-2-2-0"	5.199763251594785899e-05	"2-0-0-0-2-0-2-0"	1.490531770673959264e-05
"2-0-0-1-1-1-1-1"	5.116933505264430442e-05	"2-0-2-0-2-0-2-0"	1.417061370585438473e-05
"1-2-1-1-1-1-0-0"	5.115486227015382303e-05	"0-0-2-0-1-1-0-0"	1.271552533462085932e-05
"2-0-0-0-2-0-0-0"	5.062747894966663776e-05	"0-1-2-2-0-0-2-0"	9.372720430394723977e-06
"0-2-1-1-1-0-0-1"	5.060736419485445517e-05	"2-1-2-1-2-0-1-1"	8.757456667944396892e-06
"2-2-2-0-2-2-2-0"	5.051739826113965426e-05	"1-1-0-0-0-0-2-1"	8.613785281739383406e-06
"2-0-0-2-0-0-0-0"	5.036346677839791515e-05	"2-1-0-1-2-0-2-0"	7.951705785872844318e-06
"0-0-2-0-0-2-0-0"	4.936451691156410680e-05	"2-1-2-2-2-0-2-1"	7.659711342752065699e-06
"1-0-0-1-0-1-0-1"	4.866572676929956315e-05	"2-0-0-1-2-1-2-1"	7.382246638744156779e-06
"1-1-2-2-0-0-0-1"	4.756400820490459764e-05	"0-1-0-1-0-1-2-1"	6.650112440126120128e-06
"0-0-0-1-0-1-2-1"	4.721636220122908212e-05	"0-1-0-1-2-0-2-1"	6.429319370453981039e-06
"1-0-1-0-1-2-1-0"	4.647269444058063338e-05	"0-1-2-0-0-2-2-1"	6.422558551658294761e-06
"0-0-0-0-2-0-0-1"	4.599465388272564175e-05	"0-1-2-1-2-1-0-0"	6.155474716500039074e-06
"1-1-0-1-0-1-1-0"	4.529512133848585878e-05	"0-1-0-1-2-1-2-1"	5.879947373560735837e-06
"2-0-0-2-0-0-1-0"	4.457802393379634335e-05	"2-1-1-1-0-1-0-1"	5.485194633930210600e-06
"1-0-0-1-0-1-1-0"	4.399456910673449540e-05	"2-0-2-2-0-0-0-0"	5.738065172401693979e-06
"2-1-0-1-2-0-0-1"	4.344066886601349002e-05	"2-1-0-0-2-2-2-1"	5.415352001467553414e-06
"2-1-0-0-0-1-0-0"	4.276764047116802956e-05	"0-1-2-2-0-2-2-1"	5.345518312776561754e-06
"2-1-0-1-1-1-0-1"	4.202359223503719594e-05	"2-1-2-1-0-0-2-0"	5.209299854235530997e-06
"2-0-0-1-0-1-0-0"	4.160528410638864851e-05	"0-1-0-1-2-0-2-0"	5.182413660054110719e-06
"0-1-0-1-1-1-0-1"	4.021577420383718208e-05	"2-1-2-0-0-2-2-1"	5.019193001104952115e-06
"1-0-2-0-2-0-1-0"	4.012325702533402184e-05	"2-1-0-0-2-0-2-0"	5.000831243755575818e-06
"2-0-0-2-0-1-0-1"	3.977896849016686597e-05	"1-1-0-1-2-0-0-1"	4.815302627083318164e-06
"0-1-2-1-0-0-2-0"	3.795615251025500855e-05	"0-1-0-0-2-0-2-1"	4.605099916549000463e-06
"1-0-0-1-0-1-1-1"	3.690586497932450114e-05	"0-1-0-0-0-0-2-0"	4.570372034308779391e-06
"0-1-2-2-2-2-2-0"	3.674669994947733586e-05	"2-1-0-1-2-1-1-1"	4.529545107530535530e-06
"2-1-2-1-0-0-2-1"	3.606603029922169886e-05	"2-1-0-1-2-0-1-1"	4.494917319344843823e-06
"2-1-0-0-0-1-0-1"	3.599314254821144054e-05	"1-1-0-0-1-1-0-1"	4.226151258663228795e-06
"1-0-0-1-1-0-0-1"	3.598305501455477856e-05	"1-1-0-1-2-0-0-0"	3.938645944590388401e-06
"2-0-2-0-2-0-2-2"	3.543650890856465333e-05	"2-0-0-0-2-0-2-2"	3.821259324539321537e-06
"0-0-2-0-1-0-0-0"	3.543383385964586049e-05	"2-0-0-2-2-0-2-0"	3.664180382469819730e-06
"1-1-0-1-1-1-1-0"	3.535946640632003056e-05	"2-1-1-1-0-1-1-1"	3.651074986641756389e-06
"2-0-2-0-2-0-0-0"	3.531088076713995903e-05	"2-1-2-1-0-2-2-1"	3.5730164115868671112e-06
"1-0-0-1-1-1-0-1"	3.397287414968562633e-05	"1-1-0-0-1-0-2-1"	3.545128687170371502e-06
"2-0-2-0-0-0-2-0"	3.300678026883050614e-05	"0-0-0-1-2-0-2-1"	3.43577568225556066e-06
"1-1-0-0-1-1-0-0"	3.281172176441869367e-05	"2-1-2-2-2-2-2-1"	3.422458924676431112e-06
"2-0-2-0-2-2-2-0"	3.269511843905248156e-05	"0-0-0-1-2-1-0-1"	3.345160633953969154e-06
"2-1-2-0-0-1-0-0"	3.230023155317006297e-05	"2-1-2-1-2-1-2-1"	3.205557934797090127e-06
"0-1-0-1-1-1-0-0"	3.159523480222413693e-05	"1-1-0-1-1-0-0-1"	3.131721612893762549e-06
"1-0-2-2-2-0-0-1"	3.094929921874222935e-05	"1-0-0-2-2-0-2-0"	3.074001478851614746e-06
"2-0-0-1-0-0-0-1"	2.917058563867898040e-05	"2-1-2-0-2-1-0-0"	3.023301588776610380e-06
"2-1-0-0-2-1-0-1"	2.799222823608714676e-05	"2-0-0-1-2-1-0-1"	2.783642481991367107e-06
"0-1-0-1-0-1-1-0"	2.779274892440575833e-05	"1-0-2-1-0-0-0-0"	2.081860851432425035e-06
"2-0-2-1-0-1-0-1"	2.773301476010828345e-05	"0-1-2-2-2-2-2-1"	1.603981430863353783e-06
"0-1-1-1-0-0-1-1"	2.741622760203238476e-05	"2-1-0-1-0-1-2-0"	1.590095214020901419e-06
"1-1-2-0-0-0-1-1"	2.740405442141078521e-05	"2-1-0-1-2-1-0-1"	1.414370919146359065e-06
"2-0-2-0-0-1-0-0"	2.721241418982674336e-05	"0-1-0-0-0-0-2-1"	1.219882699132390600e-06
"0-0-2-1-1-1-1-0"	2.640834152361471894e-05	"2-1-2-0-0-2-2-0"	1.110745393474050917e-06
"2-1-0-1-0-0-0-1"	2.634403253518109952e-05	"0-1-0-2-0-0-2-0"	1.013600586847484977e-06
"1-1-0-0-1-1-1-0"	2.612044317898729691e-05	"2-1-0-1-0-1-2-1"	8.405117285916036192e-07
"0-0-0-1-2-0-0-0"	2.609107242900466947e-05	"0-1-0-1-2-1-1-1"	5.695144484093072948e-07
"2-1-0-0-2-0-0-1"	2.571613920994301743e-05	"2-1-2-1-2-1-0-1"	4.869616895907733993e-07
"2-0-2-2-2-2-2-0"	2.559440158382865279e-05	"0-1-0-1-0-0-2-1"	3.905363462974645062e-07
"2-1-2-0-2-0-1-1"	2.470367243212088373e-05	"2-1-2-1-2-0-2-0"	2.936083901977393945e-07
"0-0-0-0-2-2-0-0"	2.423656201810324255e-05	"2-0-0-0-2-0-0-0"	1.906339264336069467e-07
"0-1-0-0-2-0-1-1"	2.408696221952562482e-05	"0-1-0-1-0-2-2-1"	5.528955768353710774e-08
"2-1-2-1-0-1-0-0"	2.375725900466275870e-05		

Table C.9: Full inter-class Gini impurity ranking for the set of all ternary ESCs. ESCs are denoted with a ternary number code where 1 means active enhancer state (before: green state), 2 means repressive state (before: red state) and 0 means none of both. The order from left to right of the conditions is the same as in the earlier Gini impurity rankings from top to bottom. Hence we have: Naïve, Th₂, Tbet^{+/+}Th₁, Tbet^{+/+}Th₁/2, Tbet^{+/+}Th₁, Tbet^{+/+}Th₁/2, Tbet^{+/+}Th₁, Tbet^{+/+}Th₁/2.

APPENDIX C. SUPPLEMENTARY TABLES

ESC	T_{Gini}^*		
"0-0-1-1-0-0-0"	3.802861764549086282e-02	"0-0-0-1-1-1-0"	1.752003999614431401e-03
"0-0-1-1-1-0-0-0"	2.793203035456573793e-02	"0-0-0-1-0-0-0-2"	1.740440216484793632e-03
"0-0-1-0-1-0-0-0"	2.507822630430760549e-02	"2-1-1-1-0-0-0-0"	1.730340066259537487e-03
"0-0-1-0-0-0-0-0"	2.351054320844192178e-02	"1-0-0-1-1-0-1-0"	1.719682445946174240e-03
"0-0-1-1-0-1-0-0"	2.244676573823316257e-02	"0-2-0-0-2-0-0-0"	1.713854806008292559e-03
"0-0-1-1-1-0-1-0"	2.130107183159842682e-02	"2-2-1-0-0-0-0-0"	1.688464256814477217e-03
"0-0-1-1-1-1-0-0"	1.910379787802115800e-02	"0-0-0-1-0-1-0-1"	1.656760437207324124e-03
"0-2-1-0-0-0-0-0"	1.513752024583002453e-02	"2-2-1-0-1-0-0-0"	1.645927234160785967e-03
"0-0-1-0-0-1-0-0"	1.326601481680460791e-02	"0-0-0-1-1-0-0-0"	1.633594032155681695e-03
"0-0-1-1-1-1-1-0"	1.310524163210283199e-02	"1-0-0-2-0-0-1-0"	1.605943732771371443e-03
"0-0-1-0-1-1-1-0"	1.257446606686531999e-02	"2-0-1-1-0-1-1-1"	1.605420102097887417e-03
"1-1-0-1-1-1-0-0"	1.163510533584564098e-02	"0-0-0-1-1-1-0-0"	1.580928153517508058e-03
"0-2-0-0-0-0-0-0"	1.162945592692495901e-02	"0-0-1-1-0-1-1-1"	1.572806895466883096e-03
"0-2-1-0-1-0-0-0"	1.037571590517450890e-02	"1-2-1-0-0-2-0-0"	1.563523568984649583e-03
"0-0-1-0-1-1-0-1"	1.035265793859342469e-02	"1-2-1-2-0-0-0-0"	1.535414167084740494e-03
"2-0-1-1-0-0-0-0"	1.005482635414426881e-02	"1-1-1-1-1-1-1-0"	1.534637219525919732e-03
"1-0-1-0-1-1-1-1"	9.985329147152357393e-03	"1-2-1-2-0-2-0-0"	1.519478602296306683e-03
"0-0-1-0-1-0-1-0"	9.839018095844238518e-03	"2-0-0-0-0-1-0-0"	1.477078999137116950e-03
"1-0-1-1-1-1-1-0"	9.690233177113429311e-03	"1-1-0-0-1-2-1-0"	1.474173819981073122e-03
"0-0-1-1-0-1-1-1-0"	8.320978172828667752e-03	"1-2-1-0-0-0-0-0"	1.410440352924401510e-03
"1-0-1-1-1-1-1-1"	8.175893979845059306e-03	"1-1-1-1-1-1-0-0"	1.332307881952096446e-03
"2-0-1-1-1-0-0-0"	7.972899195225376101e-03	"1-0-1-1-0-1-1-1"	1.2804340205914261188e-03
"1-1-1-1-1-1-1-1"	7.415873501928991335e-03	"0-1-0-0-0-1-1-0"	1.241975234704229405e-03
"1-0-1-1-1-0-0-0"	7.152220647236270587e-03	"1-0-1-0-0-0-1-0"	1.154795126179179862e-03
"0-1-1-1-1-1-1-0"	7.107545329162439328e-03	"1-15567369043649919e-03"	1.145567369043649919e-03
"1-0-0-0-1-0-0-1"	6.903346000839430396e-03	"0-2-0-0-0-0-0-0"	1.088316343437202701e-03
"1-1-0-0-1-0-0-0"	6.567006907369595491e-03	"1-0-0-1-1-1-0-0"	1.060517175095384125e-03
"0-0-0-0-1-0-0-0"	6.443002625796819503e-03	"2-2-1-1-1-0-0-0"	1.054130855038260374e-03
"0-2-1-1-0-0-0-0"	6.135427340668108027e-03	"1-1-1-0-1-0-1-1"	1.052129762440301170e-03
"1-0-1-0-1-1-0-0"	6.113789152760755116e-03	"2-0-1-1-1-0-1-0"	1.050458814033475166e-03
"0-0-1-1-0-0-1-0"	5.961357171737872648e-03	"1-0-1-1-0-0-1-0"	1.050458814033475166e-03
"1-0-0-0-1-1-0-0"	5.931662770308814003e-03	"1-2-0-0-1-0-0-0"	1.032405257246198646e-03
"1-1-1-0-1-0-1-0"	5.910211780915260692e-03	"0-2-0-2-2-0-2-0"	9.975358189901863473e-04
"0-2-0-0-1-0-0-0"	5.898262929439490251e-03	"1-2-0-0-2-0-2-0"	9.963438996749508694e-04
"0-0-1-0-1-1-1-1"	5.885487857535036046e-03	"1-0-1-1-1-1-0-1"	9.938994324966486486e-04
"1-1-1-1-1-0-1-0"	5.590715706963024154e-03	"2-0-1-0-1-0-1-0"	8.60364089838806562e-04
"0-0-1-1-1-1-0-1"	5.380900380446195473e-03	"0-1-1-1-0-1-0-0"	8.460097416956360037e-04
"0-0-0-0-1-0-0-0"	5.263616228911121436e-03	"2-0-0-0-0-0-1-0"	8.293749495121341929e-04
"1-0-0-0-1-1-1-0"	5.064947230304499629e-03	"1-0-1-0-1-0-1-0"	8.290784017686374761e-04
"0-2-1-1-0-0-0-0"	4.947647765166832134e-03	"0-2-0-0-0-0-2-2"	8.107099932497978997e-04
"1-0-1-0-1-1-1-0"	4.813213859566708665e-03	"2-0-0-0-1-0-0-2"	7.956595572708039313e-04
"1-0-1-0-1-0-0-0"	4.679817325987042226e-03	"0-0-1-0-1-0-0-1"	7.72779698685782745e-04
"0-0-1-0-1-0-2-0"	4.533622999400433243e-03	"1-1-1-0-1-1-0-0"	7.64643599663748661e-04
"0-0-1-1-1-1-1-1"	4.505724298638870591e-03	"1-2-1-1-1-1-1-1"	7.423213259200809591e-04
"1-0-1-1-1-0-1-0"	4.450668102995451215e-03	"0-0-0-2-0-2-0-0"	7.360374419400339801e-04
"0-0-1-0-0-2-0-0"	4.422841330575287902e-03	"1-0-1-0-1-0-1-1"	7.295993502918256485e-04
"2-0-1-1-1-1-1-0"	4.297232770209264210e-03	"2-2-0-0-0-2-2-0"	7.295115149446940520e-04
"0-0-0-2-0-0-0-2"	3.976719121433797133e-03	"1-1-1-1-0-1-0-1"	7.255940972092143979e-04
"0-0-1-0-1-1-0-0"	3.938789740323422189e-03	"0-0-1-0-0-1-0-0"	7.205999368723666225e-04
"2-2-1-0-1-0-1-0"	3.903377075765382658e-03	"2-2-1-0-0-0-2-0"	7.130274146262984393e-04
"1-0-1-1-1-1-0-0"	3.88733608804036222e-03	"0-2-1-0-1-1-2-0"	7.082533764755185625e-04
"0-0-0-1-0-0-0-0"	3.733385297964810900e-03	"1-2-0-0-0-0-0-0"	6.793400494503052441e-04
"0-0-0-0-0-0-0-2"	3.706697759585451512e-03	"1-0-1-1-1-0-1-1"	6.734131865782166200e-04
"1-1-1-1-0-0-0-1"	3.534546127709833115e-03	"2-2-0-2-0-0-0-2"	6.732287729403004135e-04
"2-0-0-0-1-0-0-0"	3.338099605735624607e-03	"2-2-0-1-0-0-0-2"	6.719972979075174721e-04
"2-2-0-0-0-2-0-0"	3.318714701184945562e-03	"2-0-0-2-0-0-0-2"	6.382432914871242948e-04
"1-1-0-0-1-0-0-0"	3.303090904167039262e-03	"1-1-1-0-1-1-1-1"	6.357425037614064089e-04
"0-0-0-0-2-0-0-0"	3.255416127928977433e-03	"2-0-1-0-0-0-2-0"	6.278791340226292815e-04
"1-1-1-0-1-0-0-0"	3.229597616419093136e-03	"0-1-1-0-0-0-1-0"	6.198745822797738442e-04
"0-1-1-1-1-0-1-0"	3.208521402435155715e-03	"2-2-1-2-1-0-0-2"	6.113518068843843897e-04
"0-2-0-0-0-2-0-0"	3.171124836861864953e-03	"2-0-0-0-1-0-1-0"	6.098292735485789625e-04
"0-0-1-1-1-2-0-0"	3.145664460309343432e-03	"0-0-0-0-0-0-0-2"	6.056692146193639248e-04
"0-0-1-1-0-0-2-0"	3.128661290455586261e-03	"0-0-1-0-0-0-0-2"	5.715032170260504578e-04
"1-0-1-0-0-0-0-0"	3.027871920305361553e-03	"1-0-1-0-1-2-0-0"	5.333726980139849930e-04
"1-1-1-1-0-1-0-0"	3.026667746154352526e-03	"1-0-1-1-1-0-0-1"	5.286100512830043145e-04
"0-2-1-1-0-1-1-0"	2.999886095745587459e-03	"1-1-0-2-0-0-0-0"	5.029836373433781930e-04
"0-2-0-1-1-0-0-0"	2.929438924188272706e-03	"0-0-0-1-1-0-1-0"	4.673811938411465876e-04
"2-0-1-1-1-1-1-1"	2.721710809729178688e-03	"2-2-1-1-0-0-0-2"	4.646909681024187846e-04
"1-0-1-1-0-0-0-0"	2.696793052312601146e-03	"0-2-1-0-0-0-1-0"	4.645309499389131745e-04
"1-1-1-1-1-1-0-1"	2.610391861932193254e-03	"0-0-1-0-0-1-1-0"	4.568618383677024381e-04
"0-1-1-1-1-0-1-1"	2.557255110648411472e-03	"0-2-0-2-0-2-0-2"	4.483416671430275578e-04
"2-2-0-0-0-0-0-0"	2.476466078105286645e-03	"1-0-0-0-0-0-0-2"	4.371365619700920131e-04
"2-2-0-0-0-0-2-2"	2.355778103350877981e-03	"0-2-0-0-0-0-0-2"	4.329122336303386316e-04
"1-0-0-0-1-0-0-0"	2.327256700030807830e-03	"0-1-0-0-1-0-0-0"	4.298933571858187764e-04
"1-0-0-1-1-0-0-0"	2.224073046180291818e-03	"0-2-0-1-0-0-0-2"	4.249152827446898506e-04
"2-0-1-0-0-0-0-0"	2.210704649673763539e-03	"1-1-1-0-0-0-0-0"	4.160396963832504818e-04
"1-1-1-1-1-0-0-0"	2.062182738864156133e-03	"0-0-0-0-0-1-1-0"	4.151035850001484747e-04
"1-0-1-1-0-1-0-0"	2.041644910991012671e-03	"2-0-1-1-1-0-1-0"	4.085625524798384175e-04
"0-1-1-1-0-0-0-1"	1.863912793679312497e-03	"2-0-1-0-0-2-2-0"	4.058252099589620119e-04
"0-2-1-2-0-2-0-0"	1.826790932509837154e-03	"2-2-1-1-1-0-1-0"	3.942727088418253333e-04
"1-0-0-1-0-0-2-0"	1.819109593297562515e-03	"0-0-0-0-0-2-2-0"	3.933762540900087134e-04
"2-0-1-0-1-0-0-0"	1.785664439541836913e-03	"2-0-1-0-0-1-0-0"	3.896255546034649297e-04
		"2-0-1-1-1-1-2-0"	3.874837135831671548e-04

"1-0-1-1-0-0-1-1"	3.867757962176232014e-04	"1-0-0-0-1-0-2-0"	7.235242683006604363e-05
"0-0-1-0-0-2-2-0"	3.762192507190657452e-04	"0-2-0-0-1-2-0-2"	7.132282382180076799e-05
"0-0-1-0-1-0-0-2"	3.755649888621946544e-04	"0-2-1-1-0-0-2-2"	7.018737305752768363e-05
"0-2-0-2-0-2-2-2"	3.698162403103673335e-04	"2-2-0-1-1-1-0-0"	7.014890324736310079e-05
"2-0-0-1-1-1-2-0"	3.647415653580643154e-04	"0-2-1-0-0-2-0-2"	7.011607525434759825e-05
"2-1-1-1-1-1-0-1"	3.620108070922488938e-04	"2-2-1-0-0-0-1-0"	6.759872098702084423e-05
"2-0-1-1-0-1-0-0"	3.592955436067517706e-04	"2-2-1-1-0-0-2-2"	6.688161164981323500e-05
"1-1-1-0-1-1-1-0"	3.580198037971480925e-04	"0-2-0-0-1-2-1-2"	6.609648279821281141e-05
"0-2-0-0-0-2-0-2"	3.560764214683418728e-04	"0-2-0-0-1-0-0-2"	6.553659764674969323e-05
"2-0-1-0-0-1-2-2"	3.528102357200574106e-04	"0-2-0-0-0-1-2"	6.544654801228135660e-05
"2-0-0-1-1-0-0-0"	3.367989022923289296e-04	"0-2-1-0-0-2-0-0"	6.432112677263015013e-05
"1-1-0-0-1-0-1-1"	3.360856444762697873e-04	"0-2-1-2-1-2-0-0"	5.853699497883074770e-05
"2-0-1-0-1-1-2-2"	3.347242098493241180e-04	"1-2-1-1-1-0-1-1"	5.608682086625652377e-05
"0-0-1-0-1-1-0-2"	3.344935870258887792e-04	"1-1-1-1-0-0-1-0"	5.563284001950715839e-05
"2-0-1-1-0-0-2-2"	3.2773101023912866196e-04	"0-2-1-1-0-0-0-2"	5.495057634608807693e-05
"2-0-1-1-1-0-0-0"	3.174096294136008250e-04	"0-2-0-1-1-1-0-0"	5.462864781113101800e-05
"0-0-0-0-1-1-1-2"	3.173370568294098313e-04	"2-2-1-0-1-2-0-0"	5.426491059068307105e-05
"2-2-1-1-1-1-1-0"	2.965688325925132547e-04	"0-2-0-0-0-2-0"	5.410842657373014656e-05
"1-0-1-0-1-1-0-2"	2.950485534567018667e-04	"0-2-1-0-0-0-2"	5.3783856805533378244e-05
"1-2-1-1-1-1-1-0"	2.923880109964558955e-04	"0-2-1-1-1-1-1-1"	5.348875165370288242e-05
"1-1-1-1-0-1-0-1"	2.899449133593628471e-04	"0-2-1-1-1-0-2-2"	5.287957990854751930e-05
"1-2-1-1-0-1-0-0"	2.889617035749177955e-04	"0-2-1-0-0-2-0"	5.268332477562110030e-05
"2-2-0-0-1-0-0-0"	2.698193488124769218e-04	"1-2-1-1-1-1-0-0"	5.115486227015381626e-05
"1-0-1-0-1-1-1-2"	2.615680360301454472e-04	"0-2-1-1-0-0-1-1"	5.060736419485444839e-05
"0-0-1-0-0-0-1-1"	2.583150547759412250e-04	"2-0-0-2-0-0-0"	5.036346677839808456e-05
"0-2-0-0-0-0-1-0"	2.276744440236292278e-04	"1-0-1-0-1-2-1-0"	4.647269444058057917e-05
"0-2-1-1-1-0-0-0"	2.264814132076735647e-04	"0-1-1-0-0-1-1"	2.741622760203236782e-05
"0-0-1-1-1-0-1-1"	2.204847882086243308e-04	"1-1-0-0-1-1-1-0"	2.612044317898727996e-05
"0-1-1-0-1-0-1-1"	2.147845616534813009e-04	"1-0-1-0-1-0-1-2"	1.520301631856531410e-05
"2-1-1-1-1-1-1-1"	2.115188057287495382e-04		
"2-2-0-1-1-0-1-0"	2.077512784702724186e-04		
"2-2-0-0-0-0-2-0"	2.010322615170620974e-04		
"0-1-1-1-1-0-0-0"	1.989006551154897252e-04		
"2-0-1-1-0-1-1-0"	1.963428406379257366e-04		
"2-2-1-1-0-0-0-0"	1.946599494858719190e-04		
"0-0-1-0-1-0-1-1"	1.941488626469296092e-04		
"2-2-0-1-1-0-0-0"	1.893221529645685436e-04		
"0-1-1-0-1-1-1-0"	1.839028556681824157e-04		
"0-1-0-0-1-0-1-0"	1.835743005711978286e-04		
"2-2-1-0-0-0-2-2"	1.810943445493525612e-04		
"2-0-1-0-0-0-2-2"	1.641317579518421633e-04		
"1-0-0-0-0-1-0-1"	1.556138501453964657e-04		
"2-2-1-0-1-1-1-0"	1.390002372178968073e-04		
"0-0-1-0-2-0-0-0"	1.330728020752358479e-04		
"0-0-1-1-0-0-1-2"	1.290470580454681441e-04		
"1-0-0-0-0-2-2-2"	1.289399004857334308e-04		
"0-2-0-2-0-0-2-2"	1.245718623575349331e-04		
"0-0-1-2-0-0-0-0"	1.23297134537277439e-04		
"1-0-1-0-1-2-0-1"	1.185500881987490668e-04		
"1-0-0-0-0-0-2-0"	1.135976633855071614e-04		
"0-1-1-0-1-0-1-1"	1.112219233099856892e-04		
"0-2-0-0-1-1-0-0"	1.098588437606899562e-04		
"1-1-0-1-1-0-1-0"	1.092990152272424660e-04		
"0-0-1-0-1-2-0-0"	1.084008638838882449e-04		
"2-0-1-0-0-0-1-0"	1.026091397317243259e-04		
"1-1-1-0-0-2-1-1"	1.022373312485556343e-04		
"2-2-1-2-1-0-0-0"	1.021253551116843366e-04		
"0-2-1-0-1-0-1-0"	1.019707878046558354e-04		
"0-2-1-0-1-1-0-0"	9.951754705823621114e-05		
"1-2-1-0-1-0-1-0"	9.750857716917659009e-05		
"1-1-1-0-0-0-1-1"	9.617954638360600448e-05		
"2-2-1-0-0-2-0-2"	9.430285896941076969e-05		
"0-1-1-0-1-1-1-1"	9.201469462661023631e-05		
"1-2-0-0-1-0-0-1"	9.027862395485775743e-05		
"1-2-1-1-0-0-0-0"	8.762774094694582717e-05		
"2-2-1-1-1-0-2-2"	8.690512589421267987e-05		
"0-2-1-0-1-2-0-0"	8.471875149039337588e-05		
"1-0-1-1-1-0-2-0"	8.409944534505640087e-05		
"0-2-1-0-0-0-2-2"	8.328552079995250683e-05		
"2-2-1-1-1-0-0-2"	8.236954703648430428e-05		
"0-0-1-1-1-2-2-0"	8.214925198144794392e-05		
"0-2-1-0-1-2-0-2"	8.179868202279418933e-05		
"1-0-1-1-1-2-2-0"	8.165311169431295024e-05		
"0-2-1-1-1-0-2-0"	8.125910430286090635e-05		
"2-2-1-2-0-2-0-2"	8.032710032942139828e-05		
"2-2-0-2-0-0-0-0"	7.970087297670871561e-05		
"2-2-1-0-1-2-1-0"	7.829665540418158007e-05		
"2-2-0-2-0-2-0-0"	7.768450995279935873e-05		
"2-2-1-0-1-0-1-2"	7.701997195398494996e-05		
"0-2-1-1-1-1-1-0"	7.683692573535865768e-05		
"0-2-0-0-1-0-1-0"	7.658423866628761732e-05		
"0-2-1-1-1-0-1-0"	7.363779492306925028e-05		
"1-2-0-2-2-0-0-0"	7.313772018968092356e-05		
"0-2-1-0-1-0-0-2"	7.237759125240338901e-05		

Table C.10: Th1 intra-class Gini impurity ranking for the set of all ternary ESCs. ESCs are denoted with a ternary number code where 1 means active enhancer state (before: green state), 2 means repressive state (before: red state) and 0 means none of both. The order from left to-right of the conditions is the same as in the earlier Gini impurity rankings from top to bottom. Hence we have: Naïve, Th2, Tbet^{+/+}Th1, Tbet^{+/+}Th1/2, Tbet^{+/-}Th1, Tbet^{+/-}Th1/2, Tbet^{-/-}Th1, Tbet^{-/-}Th1/2.

APPENDIX C. SUPPLEMENTARY TABLES

ESC	T_{Gini}^*		
"0-1-0-0-0-0-0"	2.237943723536544791e-02	"0-0-2-0-2-1-1-1"	8.697513273654352032e-04
"0-0-2-0-2-0-0-0"	1.167350358833286160e-02	"0-1-1-1-0-1-0-1"	8.676785144295611060e-04
"0-0-0-0-0-0-0-1"	1.166953784298974117e-02	"0-1-2-1-0-0-0-1"	8.562554142421881795e-04
"0-1-0-0-0-0-0-1"	1.012320421940227783e-02	"1-0-0-0-0-2-0-0"	8.555645742327702306e-04
"0-1-0-1-1-1-1-1"	8.544037886329043999e-03	"0-1-2-1-0-1-0-0"	8.484492236050902077e-04
"0-0-0-2-0-0-0-0"	7.869750158875353854e-03	"1-0-0-0-0-1-1-1"	8.462576993768261991e-04
"0-0-2-0-0-0-0-0"	7.667362706324570495e-03	"0-0-2-0-2-0-1-1"	8.403067570929594792e-04
"0-1-0-1-0-1-0-1"	6.885053935952276105e-03	"0-0-2-0-0-1-0-0"	8.377974593507048684e-04
"1-1-0-1-0-1-0-0"	5.730765271233401702e-03	"1-0-2-0-2-0-0-0"	8.350616363566298319e-04
"0-1-0-1-0-0-0-1"	5.626815741809446624e-03	"1-1-1-1-0-0-0-0"	8.170087820754522862e-04
"0-0-0-0-1-1-1-1"	5.348123452301503308e-03	"0-0-2-1-0-1-0-0"	8.128825787960887624e-04
"0-0-0-1-0-0-1-0"	5.251078321563673428e-03	"1-0-0-1-1-1-1-0"	8.03186332957884334e-04
"0-1-2-0-0-0-0-0"	4.982754541579764371e-03	"0-1-2-0-2-1-0-0"	7.927005112305686441e-04
"0-1-0-1-0-0-1-1"	4.89147960725555898e-03	"0-1-0-0-1-1-1-0"	7.645931894254478101e-04
"0-0-0-0-0-0-1-0"	4.803037239197662997e-03	"0-1-2-0-0-0-1-1"	7.626952841262851909e-04
"0-0-0-0-1-1-1-1"	4.793226725633079055e-03	"0-0-2-1-0-1-0-1"	7.435281831319817203e-04
"0-1-0-1-0-0-0-0"	4.612284818232827328e-03	"0-0-0-0-2-0-0-0"	7.056316932035967045e-04
"1-0-0-0-1-0-1-0"	4.535734259191358947e-03	"1-0-1-0-1-1-0-1"	6.669675186722698193e-04
"0-1-0-0-0-0-1-1"	4.252737981577591267e-03	"0-0-0-1-1-1-1-0"	6.586043266425186142e-04
"1-1-0-0-1-0-0-1"	3.943641515971355385e-03	"1-0-0-1-0-0-1-1"	6.532260063186908199e-04
"0-1-2-0-2-0-0-0"	3.829035801621727495e-03	"1-1-0-1-1-1-0-1"	6.286127493656166153e-04
"0-1-0-1-0-0-1-0"	3.814173193957430975e-03	"0-0-0-2-2-0-0-0"	6.250628336885811825e-04
"0-0-0-1-0-0-0-1"	3.577502152938329301e-03	"0-0-2-1-2-1-1-0"	6.16253180910028498e-04
"1-0-0-1-1-0-1-1"	3.448106985676653406e-03	"0-0-2-0-0-2-0-0"	6.135303785644264860e-04
"0-1-0-0-1-1-1-1"	3.440562444558592503e-03	"0-0-0-2-0-0-2-0"	5.86800531072444677e-04
"0-0-2-0-0-0-2-0"	3.193462629850414222e-03	"0-0-2-0-0-1-1-0"	5.866800890463030058e-04
"0-1-0-0-0-1-1-1"	3.107186699978409933e-03	"1-0-0-0-1-1-1-1"	5.679831644803032785e-04
"1-0-0-0-0-0-0-0"	3.006722020252140977e-03	"0-1-2-1-0-1-0-1"	5.605976191717009166e-04
"0-0-0-0-0-0-1-1"	2.912918211145238030e-03	"1-0-0-0-0-0-1-1"	5.218944237136509314e-04
"0-0-0-1-1-1-1-1"	2.876898565459943755e-03	"0-0-2-2-2-2-2-0"	5.145984721901515711e-04
"0-1-1-1-1-1-1-1"	2.780402420287595485e-03	"1-0-0-0-0-0-0-1"	5.130009828013103873e-04
"0-1-0-2-0-0-0-0"	2.612211435118914557e-03	"0-0-2-2-2-0-0-0"	5.120780559419808777e-04
"0-1-0-0-0-1-0-1"	2.573653758756412266e-03	"0-0-0-0-1-1-0-1"	5.088730996007357179e-04
"0-0-0-1-0-1-0-0"	2.535205211148632010e-03	"0-1-0-1-1-0-1-0"	4.924263375350736941e-04
"0-0-0-1-0-1-0-1"	2.437442336624923850e-03	"0-0-2-0-0-1-0-0"	4.918487321059657191e-04
"1-0-0-0-0-0-1-0"	2.431547755658992893e-03	"0-1-1-0-1-0-0-1"	4.874202798312379532e-04
"0-0-0-1-0-0-2-0"	2.375709057935973881e-03	"0-1-1-0-0-0-0-0"	4.805556401231200606e-04
"1-1-0-1-0-1-1-1"	2.307102411900819987e-03	"1-0-2-2-2-2-2-0"	4.787832852770258925e-04
"1-0-0-0-0-1-1-0"	2.256989851077391331e-03	"1-1-0-2-2-0-0-0"	4.727526786600745384e-04
"0-0-0-0-2-0-2-0"	2.180415000282977852e-03	"2-1-0-1-0-1-1-1"	4.627057012049205667e-04
"0-0-2-0-2-0-1-0"	2.179431104186890789e-03	"1-1-0-0-2-0-2-0"	4.576206194057587254e-04
"0-0-2-2-0-0-0-0"	2.148172062173054307e-03	"0-0-0-0-1-0-0-1"	4.500930484643728569e-04
"1-1-0-0-0-0-0-0"	2.110653168652555124e-03	"1-0-1-0-0-1-1-0"	4.325080101901058256e-04
"1-1-0-1-1-1-1-1"	2.108568273023929754e-03	"1-1-0-0-2-0-0-0"	4.232477608632476997e-04
"0-1-0-0-0-0-1-0"	2.078602225794012968e-03	"2-1-0-1-0-1-0-1"	4.18117178863162907e-04
"0-1-0-0-1-0-1-1"	2.037178869687212603e-03	"0-0-1-1-0-0-0-1"	4.159046349770165315e-04
"0-0-2-1-2-1-1-1"	1.921715621548796210e-03	"1-1-1-1-1-0-0-1"	4.155746803142751182e-04
"1-1-0-0-0-0-1-0"	1.865913087050017228e-03	"1-1-0-1-1-0-0-0"	4.147645369960965443e-04
"1-0-0-1-0-0-1-0"	1.801328694614693065e-03	"0-0-1-1-0-0-0-1"	4.129314915608412470e-04
"0-1-2-0-0-1-0-1"	1.727118203806617342e-03	"1-1-0-1-0-0-1-1"	4.060720814573060130e-04
"0-1-0-0-0-1-0-0"	1.702757129545799056e-03	"0-0-2-2-2-0-2-0"	4.022746458254258979e-04
"2-0-0-0-0-0-0-0"	1.683427230825554344e-03	"1-1-0-0-0-1-1-1"	3.961180622751635867e-04
"1-0-1-0-1-0-1-0"	1.666621867188301494e-03	"1-0-1-0-0-1-0-1"	3.941888124038699396e-04
"0-1-0-1-1-0-0-1"	1.581444574092139271e-03	"1-1-1-1-0-1-1-1"	3.861282137861596873e-04
"0-0-0-0-0-1-0-0"	1.535746409418877433e-03	"0-1-2-2-2-0-0-0"	3.808363545498822484e-04
"0-1-0-1-1-0-1-1"	1.505445541446415051e-03	"1-1-0-0-0-1-0-1"	3.761766585435702458e-04
"0-1-0-1-0-1-1-1"	1.429317772058547780e-03	"0-1-2-2-0-0-0-0"	3.647541131570063459e-04
"0-0-0-2-0-1-0-0"	1.400399018302435610e-03	"0-0-2-0-2-0-0-0"	3.515918114245251752e-04
"0-0-1-1-0-0-1-1"	1.343630294553426960e-03	"0-0-2-1-0-0-0-1"	3.451473661803039278e-04
"1-1-0-0-0-0-0-1"	1.332287885719709740e-03	"2-1-0-2-0-0-0-0"	3.341864664766510773e-04
"0-1-1-1-0-1-1-1"	1.238813148157804018e-03	"1-1-2-2-0-0-2-0"	3.279300788823725707e-04
"1-0-0-1-0-0-0-0"	1.198663099475050578e-03	"1-1-2-0-0-1-0-1"	3.241286389562738745e-04
"0-1-1-1-1-1-0-0"	1.175829164995516795e-03	"0-0-0-1-0-1-1-1"	3.155038090684854574e-04
"0-1-0-0-1-0-0-1"	1.174684418646488194e-03	"0-1-2-0-2-0-0-1"	3.132122537280271368e-04
"0-1-2-0-0-0-0-1"	1.171236706801835788e-03	"0-0-0-1-0-1-1-0"	3.062599362314029740e-04
"2-1-2-1-0-1-0-1"	1.124789820505852017e-03	"1-0-0-2-2-0-0-0"	2.971946287412100571e-04
"1-1-0-1-0-1-0-1"	1.124518586948588606e-03	"1-1-2-0-0-0-0-1"	2.818387725775146793e-04
"0-0-0-0-0-1-0-1"	1.098878648878803951e-03	"2-1-0-2-0-0-1-1"	2.797223655455632636e-04
"1-0-0-1-1-1-1-1"	1.040690934773768304e-03	"2-0-2-1-0-0-0-2"	2.791386286137155070e-04
"1-1-0-1-1-0-1-1"	1.039804149033629263e-03	"1-0-0-0-1-0-1-1"	2.682090840540366189e-04
"0-0-2-0-2-0-2-0"	1.031354131428686462e-03	"2-1-2-0-2-0-0-1"	2.652876974050222508e-04
"1-1-0-1-0-0-0-0"	1.002144902194225962e-03	"2-0-0-0-1-1-1-0"	2.632614553696690954e-04
"0-1-1-1-1-1-0-1"	9.898821418482812561e-04	"0-0-2-0-2-1-0-0"	2.622804816479547448e-04
"2-1-0-0-0-0-0-0"	9.816850559315692080e-04	"1-0-0-2-2-0-0-0"	2.577659372536135018e-04
"1-0-0-0-0-1-0-0"	9.745466251831160680e-04	"2-1-2-0-0-0-0-0"	2.527658452398545176e-04
"0-1-1-0-0-0-0-0"	9.736797681035267186e-04	"1-1-2-2-0-0-0-0"	2.518645295742151817e-04
"0-0-0-0-0-0-2-0"	9.518139511882349772e-04	"0-0-2-2-0-0-0-0"	2.514805281424834123e-04
"1-0-2-0-0-0-0-0"	9.423557743360289713e-04	"0-0-2-1-2-0-0-0"	2.355685577548480227e-04
"1-0-1-0-0-1-1-1"	9.009785045062500145e-04	"1-1-1-0-1-1-1-0"	2.309594866761662815e-04
"0-0-2-1-0-0-0-0"	8.983562689163548986e-04	"2-1-2-0-2-0-0-0"	2.285913263023870952e-04
"0-0-0-1-0-0-1-1"	8.862739414220078714e-04	"2-0-2-1-0-0-0-0"	2.243881130269447172e-04
"2-1-2-1-0-0-0-1"	8.786911136706753236e-04	"0-1-0-0-1-0-1-1"	2.210655381740928808e-04
"0-1-0-1-0-1-0-0"	8.765645272259966168e-04	"0-1-0-0-2-0-1-0"	2.204344324455158224e-04
		"1-1-2-2-0-0-2"	2.149487653012605579e-04

"2-1-0-0-0-0-1"	2.133769973549533565e-04	"0-1-2-0-0-2-0-0"	5.543293523917986956e-05
"1-0-2-0-2-0-2-0"	2.124688255162517645e-04	"2-0-2-0-0-0-1-0"	5.386162674796424876e-05
"1-1-1-0-1-1-0-1-1"	2.124213305514646634e-04	"2-1-2-1-0-1-2-1"	5.373825603060346278e-05
"2-2-2-1-2-0-2-0"	2.120880331155761784e-04	"1-1-0-2-0-0-1-1"	5.364313370084312942e-05
"0-1-2-0-2-0-2-1"	2.041481574314463491e-04	"0-0-0-1-1-1-0-1"	5.288298538031532066e-05
"0-0-0-0-1-0-1-1"	1.917975779816406910e-04	"0-0-0-0-2-0-2-0"	5.251147905394773458e-05
"1-0-0-1-0-1-0-1-0"	1.881931676587730059e-04	"1-1-2-2-2-2-2-0"	5.199763251594789287e-05
"2-0-0-1-0-0-0-0"	1.848190972943757559e-04	"2-0-0-1-1-1-1-1"	5.116933505264431120e-05
"1-0-0-2-0-0-0-0"	1.785693029706626950e-04	"2-0-0-0-2-0-0-0"	5.062747894966662421e-05
"0-0-2-1-0-0-1-0"	1.753777716222085096e-04	"2-2-2-0-2-2-2-0"	5.051739826113976268e-05
"1-0-0-1-0-0-0-1"	1.735643849555767949e-04	"0-0-2-0-0-2-0-0"	4.936451691156416101e-05
"1-0-0-2-2-2-2-0"	1.641587190810525653e-04	"1-0-0-1-0-1-0-1"	4.866572676929962414e-05
"2-0-2-0-0-0-0-0"	1.638054373209144206e-04	"1-1-2-2-0-0-0-1"	4.756400820490469251e-05
"0-0-2-0-2-0-2-2"	1.589609586223752829e-04	"0-0-0-1-0-1-2-1"	4.721636220122918376e-05
"0-1-0-1-1-1-1-0"	1.588614065468758852e-04	"0-0-0-0-2-0-0-1"	4.599465388272571629e-05
"0-1-2-0-2-1-0-1"	1.576685842082756097e-04	"1-1-0-1-0-1-1-0"	4.529512133848843681e-05
"0-0-1-0-0-0-0-1"	1.569386906401569557e-04	"2-0-0-2-0-0-1-0"	4.457802393379634335e-05
"1-0-0-2-0-2-0-0"	1.557885286808262967e-04	"1-0-0-1-0-1-1-0"	4.399456910673458349e-05
"2-1-2-0-2-0-2-0"	1.539844166740613515e-04	"2-1-0-1-2-0-0-1"	4.344066886601345614e-05
"0-0-2-1-0-1-1-1-0"	1.504766316492241911e-04	"2-1-0-0-0-1-0-0"	4.276764047116815153e-05
"2-1-2-0-2-1-0-1"	1.428491121037828409e-04	"2-1-0-1-1-1-0-1"	4.202359223503725693e-05
"0-1-0-2-0-1-0-0"	1.423988510657215733e-04	"2-0-0-1-0-1-0-0"	4.160528410638863496e-05
"1-0-0-0-2-0-2-0"	1.404036486107083459e-04	"0-1-0-1-1-1-0-1"	4.021577420383722952e-05
"0-1-0-0-2-0-2-0"	1.350927795380423393e-04	"1-0-2-0-2-0-1-0"	4.012325702533410993e-05
"1-1-0-0-2-0-1-0"	1.349175349850630817e-04	"2-0-0-2-0-1-0-1"	3.977896849016687275e-05
"0-1-2-0-2-0-2-0"	1.349087137084119129e-04	"1-1-2-1-0-0-2-0"	3.79561525102508986e-05
"0-1-2-0-2-0-1-0"	1.286980936673063941e-04	"1-0-0-1-0-1-1-1"	3.6905864997932457567e-05
"1-1-0-1-0-0-0-1"	1.275730265332531809e-04	"0-1-2-2-2-2-2-0"	3.674669994947742395e-05
"0-1-2-0-0-0-2-0"	1.227314961026832135e-04	"2-1-2-1-0-0-2-1"	3.606603029922178017e-05
"2-2-1-1-1-1-0-0"	1.148360939638187235e-04	"2-1-0-0-0-1-0-1"	3.599314254821148120e-05
"0-0-0-0-2-0-0-2"	1.141646031244036098e-04	"1-0-0-1-1-0-0-1"	3.598305501455475823e-05
"1-0-2-2-0-2-0-0"	1.132310280537860336e-04	"2-0-2-0-2-0-2-2"	3.543650890856469399e-05
"1-0-2-2-2-0-2-2"	1.129920146190787662e-04	"0-0-2-0-1-0-0-0"	3.543383385964593503e-05
"0-0-0-2-0-0-0-1"	1.120309894968078782e-04	"1-1-0-1-1-1-1-0"	3.535946640632006444e-05
"0-0-2-0-2-2-2-0"	1.112958638415904606e-04	"2-0-2-0-2-0-0-0"	3.531088076714004035e-05
"0-1-2-1-0-1-1-1"	1.111739342124113289e-04	"1-0-0-1-1-1-0-1"	3.397287414968570087e-05
"2-0-2-0-2-0-1-0"	1.105913008120801728e-04	"2-0-2-0-0-2-0-0"	3.300678026883053324e-05
"0-0-2-0-0-0-2-2"	1.105755061473584591e-04	"1-1-0-0-1-1-0-0"	3.281172176441876144e-05
"1-1-0-0-1-0-1-0"	1.091066423215418025e-04	"2-0-2-0-2-2-2-0"	3.269511843905255609e-05
"2-1-0-1-0-1-0-0"	1.060382308876453643e-04	"2-1-2-0-0-1-0-0"	3.230023155317013074e-05
"2-0-2-2-0-0-2-0"	1.055070162271064553e-04	"0-1-0-1-1-1-0-0"	3.159523480222417081e-05
"1-1-0-2-2-2-2-0"	1.054251887887687716e-04	"1-0-2-2-2-0-0-1"	3.094929921874229711e-05
"1-0-0-0-2-0-0-0"	9.740251808281682750e-05	"2-0-0-1-0-0-0-1"	2.917058563867902106e-05
"1-1-0-0-0-0-2-0"	9.506571206096831993e-05	"2-1-0-0-2-1-0-1"	2.799222823608720775e-05
"1-2-1-0-1-1-1-0"	9.473224520711616815e-05	"0-1-0-1-0-1-1-0"	2.779274892440581932e-05
"1-1-1-0-0-0-1-0"	9.414636883994110920e-05	"2-0-2-1-0-1-0-1"	2.773301476010826990e-05
"1-1-0-2-2-2-0-0"	9.006136086497161628e-05	"1-1-2-0-0-0-1-1"	2.740405442141082926e-05
"1-1-0-0-0-0-2-2"	8.853513139730975897e-05	"2-0-2-0-0-1-0-0"	2.721241418982680435e-05
"0-0-0-1-2-0-0-1"	8.838774958443210139e-05	"0-0-2-1-1-1-0-1"	2.640834152361475621e-05
"2-0-2-2-2-0-2-0"	8.737953940471528905e-05	"2-1-0-1-0-0-0-1"	2.634403253518103854e-05
"1-0-2-0-0-0-1-0"	8.737690333771468739e-05	"0-0-0-1-2-0-0-0"	2.609107242900472707e-05
"1-0-2-0-0-2-0-2"	8.696764539761696442e-05	"2-1-0-0-2-0-0-1"	2.571613920994305131e-05
"0-1-2-0-0-0-2-1"	8.484738847927976012e-05	"2-0-2-2-2-2-2-0"	2.559440158382870700e-05
"0-1-2-0-0-0-2-0"	8.190916168947377210e-05	"2-1-2-0-2-0-1-1"	2.470367243212093794e-05
"1-1-0-1-0-0-1-0"	8.095487865090215615e-05	"0-0-0-0-2-2-0-0"	2.423656201810329676e-05
"1-0-2-2-2-0-2-0"	8.093268280480379027e-05	"0-1-0-0-2-0-1-1"	2.408696221952568242e-05
"1-1-2-2-0-0-0-0"	7.933081997305355517e-05	"2-1-2-1-0-1-0-0"	2.375725900466280953e-05
"0-0-2-0-2-2-0-0"	7.854330270902135222e-05	"2-1-0-1-1-1-1-1"	2.351779414008388155e-05
"0-1-1-0-0-1-1-1"	7.815208884891942573e-05	"0-1-1-1-0-1-1-0"	2.349977986383010604e-05
"0-1-1-0-0-0-1-1"	7.551886276015383731e-05	"2-1-0-0-2-1-2-1"	2.342028915942584094e-05
"0-0-0-2-2-2-0-0"	7.503447510424812554e-05	"2-0-0-0-0-0-2-0"	2.282025049412915395e-05
"0-1-0-0-2-2-0-0"	7.425238312394686745e-05	"2-0-2-1-0-0-2-0"	2.231249439748069727e-05
"1-1-2-2-2-0-2-2"	7.404368290357771719e-05	"0-0-0-1-0-0-2-1"	2.220816233470136381e-05
"0-1-0-1-2-1-0-1"	7.356355676132553343e-05	"2-0-2-0-0-0-1-1"	2.179424334591302577e-05
"2-0-0-2-0-0-2-0"	7.291748056403900887e-05	"2-1-0-1-1-1-1-0"	2.041984831611827128e-05
"0-1-2-2-0-0-2-1"	7.180619546306860599e-05	"1-0-2-2-2-0-0-0"	2.035097293602215640e-05
"0-0-2-2-0-0-1-0"	7.017553848276682454e-05	"2-0-2-1-0-1-1-1"	1.995272495635396902e-05
"1-0-2-2-0-0-0-0"	6.756140955379307293e-05	"0-1-0-0-1-1-0-0"	1.973996328824179119e-05
"1-1-0-0-0-0-1-1"	6.713022884757888339e-05	"2-1-2-0-2-1-2-1"	1.911456586485545475e-05
"2-1-2-0-2-0-2-1"	6.683273137112268091e-05	"2-1-0-0-2-0-0-0"	1.881114039322983103e-05
"2-0-2-2-2-0-0-0"	6.531234927897984434e-05	"2-1-0-0-2-0-2-1"	1.880954815237736614e-05
"1-0-2-0-1-0-1-0"	6.478253420038761049e-05	"0-0-0-1-2-1-0-0"	1.876709470064081442e-05
"0-1-2-0-0-0-1-0"	6.419852684941581677e-05	"2-1-2-1-2-0-0-1"	1.861321304260058916e-05
"1-2-1-1-0-0-0-2"	6.379206998605410389e-05	"2-0-0-0-0-0-1-1"	1.836545528933941000e-05
"1-1-0-2-0-2-0-0"	6.294310992721707907e-05	"1-0-2-0-0-2-0-0"	1.765034770174714017e-05
"1-1-2-2-2-0-2-0"	6.235329647649707043e-05	"0-1-0-1-2-0-0-1"	1.718721395341781474e-05
"1-1-0-0-2-0-2-2"	6.114874604240235281e-05	"0-1-2-2-2-0-2-0"	1.699497286374653275e-05
"2-0-0-1-0-1-0-1"	6.104029911596131391e-05	"2-0-0-1-0-0-2-0"	1.696073934170007566e-05
"0-1-2-2-0-0-0-1"	5.946588326625751856e-05	"2-1-2-0-0-1-2-1"	1.692750764932043027e-05
"1-1-0-2-2-0-2-2"	5.945944775892519386e-05	"1-1-0-0-2-0-1-1"	1.673264845132359729e-05
"0-1-2-2-0-2-0-0"	5.872115167267276219e-05	"2-1-0-0-0-0-2-0"	1.587048396645381187e-05
"1-1-0-0-0-0-2-2"	5.831915728885017950e-05	"2-1-2-0-2-1-1-1"	1.5856676397748903748e-05
"1-1-0-0-0-2-1-0"	5.618344335913057783e-05	"0-0-0-1-1-0-0-1"	1.566629481065551780e-05
"1-1-2-2-0-0-1-1"	5.587642802436222183e-05	"2-0-0-0-2-0-2-0"	1.490531770673961805e-05

"2-0-2-0-2-0-2-0"	1.417061370585439998e-05
"0-0-2-0-1-1-0-0"	1.271552533462088643e-05
"0-1-2-2-0-0-2-0"	9.372720430394734141e-06
"2-1-2-1-2-0-1-1"	8.757456667944400281e-06
"1-1-0-0-0-0-2-1"	8.613785281739398653e-06
"2-1-0-1-2-0-2-0"	7.951705785872857871e-06
"2-1-2-2-0-2-1"	7.659711342752069087e-06
"2-0-0-1-2-1-2-1"	7.382246638744159320e-06
"0-1-0-1-0-1-2-1"	6.650112440126127751e-06
"0-1-0-1-2-0-2-1"	6.429319370453982733e-06
"0-1-2-0-0-2-2-1"	6.422558551658301537e-06
"0-1-2-1-2-1-0-0"	6.155474716500050085e-06
"0-1-0-1-2-1-2-1"	5.879947373560748543e-06
"2-1-1-1-0-1-0-1"	5.845194633930223305e-06
"2-0-2-2-0-0-0-0"	5.738065172401704144e-06
"2-1-0-0-0-2-2-1"	5.415352001467559343e-06
"0-1-2-2-0-2-2-1"	5.345518312776568531e-06
"2-1-2-1-0-0-2-0"	5.209299854235542856e-06
"0-1-0-1-2-0-2-0"	5.182413660054116648e-06
"2-1-2-0-0-2-2-1"	5.019193001104957197e-06
"2-1-0-0-2-0-2-0"	5.000831243755585136e-06
"1-1-0-1-2-0-0-1"	4.815302627083329175e-06
"0-1-0-0-2-0-2-1"	4.605099916549005546e-06
"0-1-0-0-0-0-2-0"	4.570372034308787862e-06
"2-1-0-1-2-1-1-1"	4.529545107530544847e-06
"2-1-0-1-2-0-1-1"	4.494917319344853987e-06
"1-1-0-0-1-1-0-1"	4.22615125866323030e-06
"1-1-0-1-2-0-0-0"	3.938645944590396872e-06
"2-0-0-0-2-0-2-2"	3.821259324539325772e-06
"2-0-0-2-2-0-2-0"	3.664180382469823965e-06
"2-1-1-1-0-1-1-1"	3.651074986641757236e-06
"2-1-2-1-0-2-2-1"	3.573016415868675348e-06
"1-1-0-0-1-0-2-1"	3.545128687170375314e-06
"0-0-0-1-2-0-2-1"	3.435775682255564112e-06
"2-1-2-2-2-2-2-1"	3.422458924676438311e-06
"0-0-0-1-2-1-0-1"	3.345160633953970001e-06
"2-1-2-1-2-1-2-1"	3.205557934797096056e-06
"1-1-0-1-1-0-0-1"	3.131721612893766361e-06
"1-0-0-2-2-0-2-0"	3.074001478851619404e-06
"2-1-2-0-2-1-0-0"	3.023301588776613345e-06
"2-0-0-1-2-1-0-1"	2.783642481991373460e-06
"1-0-2-1-0-0-0-0"	2.081860851432429694e-06
"0-1-2-2-2-2-2-1"	1.603981430863355689e-06
"2-1-0-1-0-1-2-0"	1.590095214020904807e-06
"2-1-0-1-2-1-0-1"	1.414370919146361394e-06
"0-1-0-0-0-0-2-1"	1.219882699132392717e-06
"2-1-2-0-0-2-2-0"	1.110745393474052822e-06
"0-1-0-2-0-0-2-0"	1.013600586847486036e-06
"2-1-0-1-0-1-2-1"	8.405117285916054192e-07
"0-1-0-1-2-1-1-1"	5.695144484093086712e-07
"2-1-2-1-2-1-0-1"	4.869616895907744581e-07
"0-1-0-1-0-0-2-1"	3.905363462974649297e-07
"2-1-2-1-2-0-2-0"	2.936083901977400298e-07
"2-0-0-0-2-0-0-0"	1.906339264336073437e-07
"0-1-0-1-0-2-2-1"	5.528955768353716730e-08

Table C.11: Th2 intra-class Gini impurity ranking for the set of all ternary ESCs. ESCs are denoted with a ternary number code where 1 means active enhancer state (before: green state), 2 means repressive state (before: red state) and 0 means none of both. The order from left to-right of the conditions is the same as in the earlier Gini impurity rankings from top to bottom. Hence we have: Naïve, Th2, Tbet^{+/+}Th1, Tbet^{+/+}Th1/2, Tbet^{+/-}Th1, Tbet^{+/-}Th1/2, Tbet^{-/-}Th1, Tbet^{-/-}Th1/2.

ESC A	ESC B	Conditional probability
0-0-1-0-0-0-0-0	0-2-1-0-1-0-0-0	0.249397
0-0-1-0-0-0-0-0	0-0-1-1-0-1-1-0	0.233207
0-0-1-0-0-0-0-0	0-2-1-0-0-0-0-0	0.230581
0-0-1-0-0-0-0-0	0-0-1-0-1-0-1-0	0.20843
0-0-1-0-0-0-0-0	0-0-1-0-1-0-0-0	0.20657
0-0-1-0-0-0-0-0	0-0-1-1-1-0-1-0	0.197327
0-0-1-0-0-0-0-0	0-2-0-0-0-0-0-0	0.187355
0-1-0-1-0-0-0-1	0-1-0-1-0-1-0-1	0.16431
0-0-1-0-0-0-0-0	0-0-1-1-1-0-0-0	0.163348
0-1-0-0-0-0-0-0	1-1-0-1-0-1-0-0	0.149446
0-0-1-0-0-0-0-0	0-0-1-0-0-0-1-0	0.148143
0-0-1-0-0-0-0-0	1-0-1-1-1-1-1-0	0.145142
0-0-1-0-0-0-0-0	0-0-1-0-1-1-1-0	0.143419
0-0-1-0-1-0-0-0	0-0-1-0-1-0-1-0	0.129595
0-1-0-0-0-0-0-0	0-1-0-0-0-0-0-1	0.129084
0-0-1-0-0-0-0-0	2-0-1-1-0-0-0-0	0.128096
0-1-0-1-0-1-0-1	1-1-0-0-1-0-0-1	0.12052
0-0-0-0-1-1-1-1	0-1-0-1-1-1-1-1	0.119556
0-1-0-0-0-0-0-0	0-1-2-0-2-0-0-0	0.117949
0-1-0-0-0-0-0-0	0-1-0-0-0-0-1-1	0.117554
0-0-1-0-0-0-0-0	0-0-1-1-0-0-0-0	0.115298
0-0-1-0-1-0-0-0	0-0-1-1-0-1-1-0	0.113892
0-1-0-1-1-1-1-1	0-1-2-0-2-0-0-0	0.11345
0-0-1-0-1-0-0-0	0-0-1-1-1-0-1-0	0.113003
0-1-0-0-0-0-0-0	0-1-0-1-0-0-0-0	0.112988
0-0-1-0-1-0-0-0	0-0-1-0-1-1-1-0	0.112084
0-1-0-0-0-0-0-0	0-1-0-1-0-0-1-1	0.110879
0-0-0-0-0-0-0-1	0-1-0-0-0-0-0-0	0.110075
0-0-1-0-1-0-0-0	0-2-1-0-1-0-0-0	0.109915
0-1-0-0-0-0-0-0	0-1-0-1-1-1-1-1	0.109111
0-1-0-0-0-0-0-0	0-1-0-1-0-1-0-1	0.108468
0-0-1-0-0-0-0-0	0-0-1-0-1-1-0-1	0.108468
0-0-2-0-0-0-0-0	0-1-0-0-0-0-0-0	0.106164
0-0-1-1-0-1-1-0	0-0-1-1-1-0-0-0	0.106058
0-0-0-2-0-0-0-0	0-1-0-0-0-0-0-0	0.106058
0-0-0-0-0-1-1-1	0-1-0-1-1-1-1-1	0.104451
0-0-1-0-0-0-0-0	0-0-1-0-1-0-0-0	0.103744
0-0-1-1-1-0-0-0	0-2-1-0-1-0-0-0	0.103166
0-0-0-0-0-1-1-1	0-1-0-0-0-0-0-0	0.101237
0-0-1-0-1-0-0-0	0-2-1-0-0-0-0-0	0.0995999
0-0-0-2-0-0-0-0	0-1-0-1-1-1-1-1	0.0982242
0-0-1-0-0-0-0-0	0-0-1-0-0-0-0-0	0.0982161
0-0-1-0-0-0-0-0	0-0-1-0-0-0-0-0	0.0982161
0-0-1-0-1-0-0-0	0-0-1-0-1-1-0-1	0.0976215
0-1-0-0-0-0-0-0	0-1-0-1-0-0-0-1	0.0968182
0-0-2-0-0-0-0-0	0-1-2-0-2-0-0-0	0.0906313
0-0-1-0-0-0-0-0	1-0-1-0-1-1-1-1	0.0899882
0-0-1-1-1-0-0-0	0-2-1-0-0-0-0-0	0.085956
0-0-0-0-1-1-1-1	0-1-0-0-0-0-0-0	0.0858106
0-0-2-0-0-0-0-0	0-1-0-1-1-1-1-1	0.0856496
0-0-2-0-0-0-0-0	0-1-0-1-1-1-1-1	0.0847087
0-1-0-0-0-0-0-0	0-1-2-0-0-0-0-0	0.0843644
0-0-1-1-1-0-0-0	0-0-1-1-1-0-1-0	0.0840784
0-0-1-0-0-0-0-0	0-0-1-1-0-1-0-0	0.083509
0-0-1-1-1-0-0-0	0-0-1-1-1-1-0-0	0.0824667
0-0-1-0-0-0-1-0	0-0-1-0-1-0-0-0	0.0820842
0-0-1-0-1-0-0-0	0-2-0-0-0-0-0-0	0.0818443
0-2-1-0-0-0-0-0	0-2-1-0-1-0-0-0	0.0813112
0-0-1-0-0-0-0-0	0-0-1-1-1-0-0-0	0.0813112
0-0-1-0-0-0-0-0	0-0-1-1-1-1-1-0	0.080778
0-0-0-2-0-0-0-0	0-0-2-0-0-0-0-0	0.0807488
⋮	⋮	⋮

Table C.12: Top-ranked conditional probabilities of an ESC A given an ESC B. ESCs are denoted with a ternary number code where 1 means active enhancer state (before: green state), 2 means repressive state (before: red state) and 0 means none of both. The order from left to-right of the conditions is the same as in the earlier Gini impurity rankings from top to bottom. Hence we have: Naïve, Th2, Tbet^{+/+}Th1, Tbet^{+/+}Th1/2, Tbet^{+/-}Th1, Tbet^{+/-}Th1/2, Tbet^{-/-}Th1, Tbet^{-/-}Th1/2.

ESC	Gata3 (Th1)	Gata3 (Th2)	STAT1 (Th1)	STAT4 (Th1)	STAT6 (Th2)	Tbet (Th1)
"0-1-0-0-0-0-1"	0.0	0.44716478751007438	0.0	0.0	0.13320186337986661	0.0
"0-1-0-0-0-0-0"	0.0	0.27947799219379649	0.0	0.0	0.33300465844966659	0.0
"1-1-1-1-1-1-1"	0.0	0.01470936801019981	0.42105263157894735	0.17426866366848235	0.0	0.35832579114299568
"0-1-0-1-1-1-1"	0.0	0.48273471378928473	0.0	0.0	0.090819452304454526	0.0
"0-1-1-1-1-1-1"	0.0	0.37263732292506185	0.0	0.078835819516694383	0.16650232922483327	0.0
"0-0-1-0-0-0-0"	0.012058352535724313	0.0	0.5862068965517242	0.19573031052420678	0.0	0.13415152771855993
"0-0-1-0-0-0-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0
"1-2-1-1-1-1-1"	0.0	0.0	0.0	0.0	0.0	0.0
"1-0-1-1-1-1-0"	0.0	0.0	0.5	0.28380895026009978	0.0	0.97259857595955956
"0-2-0-0-0-0-0"	0.029141018628000431	0.0	0.4999999999999994	0.15767163903338877	0.0	0.19451971519191188
"0-0-0-0-0-0-0"	0.0059269868395933076	0.13263362341400506	0.57627118644067787	0.032068807938994325	0.033864880520305077	0.16209976265992657
"0-2-1-0-0-2-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.065938886505732833
"0-0-1-1-0-0-0"	0.0	0.0	0.5454545454545451	0.2580081366009075	0.0	0.17683610471991992
"0-0-1-1-0-1-0"	0.0	0.0	0.42857142857142849	0.20272067875721414	0.0	0.34735663427127123
"0-0-1-1-1-0-0"	0.0	0.0	0.61904761904761907	0.22524519861912684	0.0	0.1389426537085085
"0-0-1-1-1-1-0"	0.0	0.0	0.5	0.29108610283087155	0.0	0.1870381876845307
"1-0-1-1-1-1-0"	0.0	0.0	0.33333333333333326	0.31534327806677753	0.0	0.32419952531985319
"1-0-1-1-0-0-1"	0.0	0.0	0.2499999999999997	0.23650745855008315	0.16650232922483327	0.24314964398989899
"1-0-1-1-0-1-1"	0.0	0.0	0.5454545454545451	0.17200542440006047	0.0	0.26525415707987987
"0-0-1-1-1-0-1"	0.0	0.0	0.33333333333333326	0.31534327806677753	0.0	0.32419952531985319
"0-2-1-1-1-0-0"	0.0	0.0	0.1999999999999998	0.75682386736026608	0.0	0.0
"2-2-1-1-0-0-0"	0.0	0.0	0.6666666666666652	0.31534327806677753	0.0	0.0
"2-2-1-1-0-0-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0
"2-2-1-1-1-0-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0
"0-0-1-1-0-0-0"	0.0	0.093159330731265491	0.1666666666666666	0.15767163903338877	0.11100155281655552	0.32419952531985324
"0-0-1-1-0-0-1"	0.0	0.0	0.5999999999999998	0.18920596684006652	0.0	0.19451971519191188
"2-0-1-1-0-0-0"	0.0	0.0	0.8333333333333337	0.15767163903338877	0.0	0.0
"2-0-1-1-1-0-0"	0.0	0.0	0.5999999999999998	0.18920596684006652	0.0	0.0
"0-1-0-0-0-1-0"	0.0	0.0	0.0	0.0	0.66600931689933307	0.0
"1-1-0-0-0-1-0"	0.0	0.13973899609689822	0.5000000000000011	0.0	0.1665023292248333	0.0
"1-1-0-0-0-1-1"	0.0	0.4192169882906947	0.0	0.0	0.16650232922483327	0.0
"0-1-0-0-0-1-1"	0.0	0.0	0.0	0.0	0.66600931689933307	0.0
"1-1-0-1-1-0-1"	0.0	0.0	0.0	0.0	0.66600931689933307	0.0
"1-1-0-1-1-1-1"	0.0	0.0	0.33333333333333331	0.0	0.22200310563311104	0.0
"1-0-0-1-0-0-0"	0.0	0.18631866146253098	0.0	0.0	0.33300465844966654	0.0
"0-1-2-1-0-1-0"	0.0	0.27947799219379643	0.0	0.0	0.66600931689933307	0.0
"1-1-0-1-0-1-0"	0.0	0.0	0.0	0.0	0.0	0.0
"1-1-0-0-0-1-0"	0.0	0.55895598438759286	0.0	0.0	0.0	0.0
"1-1-0-0-0-1-0-1"	0.0	0.55895598438759286	0.0	0.0	0.0	0.0
"1-0-1-1-0-0-1"	0.0	0.0	0.0	0.9460298342003326	0.0	0.0
"1-1-1-1-0-1-1"	0.0	0.055895598438759284	0.0	0.18920596684006652	0.066600931689933321	0.29177957278786781
"1-0-0-0-0-1-0"	0.0	0.0	0.2999999999999999	0.0	0.0	0.0
"1-0-0-0-0-1-1"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0
"1-0-0-0-1-0-1"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0
"0-0-0-0-1-0-1-0"	0.0	0.0	0.7999999999999993	0.0	0.0	0.19451971519191188
"0-0-0-0-1-1-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0

"0-0-0-1-1-1-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.33300465844966654	0.0	0.0	0.48629928797977978
"1-1-0-0-1-1-1-0"	0.0	0.0	0.4999999999999994	0.0	0.0	0.66600931689933307	0.0	0.0	0.32419952531985324
"1-0-0-0-0-0-0"	0.0	0.046579665365632746	0.4999999999999994	0.078835819516694383	0.0	0.33300465844966654	0.0	0.0	0.48629928797977978
"0-0-1-1-1-2-0"	0.0	0.0	0.4999999999999994	0.0	0.0	0.26640372675973323	0.0	0.0	0.0
"1-1-0-0-1-0-0-1"	0.0	0.27947799219379643	0.0	0.0	0.0	0.66600931689933307	0.0	0.0	0.0
"1-1-0-0-0-0-1"	0.0	0.0	0.0	0.0	0.0	0.33300465844966654	0.0	0.0	0.0
"0-0-1-0-0-0-1"	0.0	0.27947799219379643	0.0	0.0	0.0	0.26640372675973323	0.0	0.0	0.0
"2-0-0-1-0-0-0-1"	0.0	0.55895598438759286	0.0	0.0	0.0	0.66600931689933307	0.0	0.0	0.0
"2-1-0-1-0-1-0-1"	0.0	0.33537359063255573	0.0	0.0	0.0	0.33300465844966654	0.0	0.0	0.0
"2-1-0-0-2-0-0-0"	0.0	0.0	0.0	0.0	0.0	0.26640372675973323	0.0	0.0	0.0
"0-0-0-2-0-0-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.66600931689933307	0.0	0.0	0.0
"0-0-1-0-0-0-1-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"0-0-1-0-1-0-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"1-0-1-0-1-2-0-1"	0.0	0.0	0.25	0.1182537292750416	0.0	0.0	0.0	0.0	0.60787410997472469
"1-0-1-1-1-0-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"1-0-1-1-1-1-1-1"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"1-0-1-1-1-0-1-0"	0.0	0.0	1.0	0.4730149171001663	0.0	0.0	0.0	0.0	0.0
"1-1-1-0-1-1-1-1"	0.0	0.0	0.4999999999999994	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-1-0-1-1-2-0"	0.0	0.0	0.4999999999999994	0.0	0.0	0.0	0.0	0.0	0.0
"0-1-0-0-1-0-0-0"	0.0	0.27947799219379643	0.0	0.0	0.0	0.33300465844966654	0.0	0.0	0.48629928797977978
"0-1-0-0-0-1-0-1"	0.0	0.079850854912513286	0.0	0.0	0.0	0.57086512877085682	0.0	0.0	0.0
"0-0-2-0-0-0-0-0"	0.0	0.27947799219379643	0.0	0.0	0.0	0.33300465844966654	0.0	0.0	0.0
"0-0-0-0-0-0-1-1"	0.0	0.55895598438759286	0.0	0.0	0.0	0.0	0.0	0.0	0.0
"0-0-0-0-0-1-1-1"	0.0	0.55895598438759286	0.0	0.0	0.0	0.0	0.0	0.0	0.0
"0-0-0-1-0-1-1-1"	0.0	0.18631866146253098	0.0	0.0	0.0	0.44400621126622203	0.0	0.0	0.0
"0-1-0-1-0-0-1-1"	0.0	0.37263732292506196	0.0	0.0	0.0	0.22200310563311101	0.0	0.0	0.0
"0-0-1-1-1-1-1-1"	0.0	0.31053110243755155	0.11111111111111112	0.0	0.0	0.22200310563311104	0.0	0.0	0.0
"0-0-0-1-1-1-1-0"	0.0	0.069869498048449108	0.2499999999999997	0.23650745855008315	0.0	0.24975349383724996	0.0	0.0	0.0
"1-1-0-1-0-1-0-0"	0.0	0.37263732292506196	0.0	0.0	0.0	0.22200310563311101	0.0	0.0	0.0
"1-1-0-1-0-0-1-1"	0.0	0.55895598438759286	0.0	0.0	0.0	0.0	0.0	0.0	0.0
"0-0-2-1-0-0-0-1"	0.0	0.55895598438759286	0.0	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-0-1-1-1-0-0"	0.0	0.0	0.9999999999999989	0.4730149171001663	0.0	0.0	0.0	0.0	0.48629928797977978
"1-2-1-1-1-0-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"1-0-1-1-0-1-1-1"	0.0	0.0	0.4999999999999994	0.23650745855008315	0.0	0.0	0.0	0.0	0.24314964398988989
"1-0-1-1-0-0-0-0"	0.0	0.0	0.3333333333333326	0.31534327806677753	0.0	0.0	0.0	0.0	0.32419952531985319
"1-0-1-1-1-0-0-0"	0.0	0.0	0.6666666666666652	0.10511442602225915	0.0	0.0	0.0	0.0	0.21613301687990211
"0-2-1-0-1-0-0-0"	0.0	0.0	0.3333333333333326	0.52557213011129589	0.0	0.0	0.0	0.0	0.10806650843995103
"0-2-0-0-1-0-0-0"	0.0	0.0	0.6666666666666652	0.31534327806677753	0.0	0.0	0.0	0.0	0.0
"0-2-0-0-0-1-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-0-0-0-1-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-0-0-1-0-0-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-0-0-1-0-1-0"	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-0-0-1-2-0-2"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-1-0-0-0-0-0"	0.0	0.0	0.9999999999999989	0.0	0.0	0.0	0.0	0.0	0.0
"0-2-1-0-0-2-0-2"	0.0	0.0	0.4999999999999994	0.0	0.0	0.0	0.0	0.0	0.48629928797977978
"0-2-1-0-1-0-0-2"	0.0	0.0	0.4999999999999994	0.0	0.0	0.0	0.0	0.0	0.48629928797977978

"2-2-0-1-1-0-0-0"	0.0	0.0	0.999999999999999889	0.0	0.0	0.0	0.0	0.0
"2-2-0-1-1-0-1-0"	0.0	0.0	0.999999999999999889	0.0	0.0	0.0	0.0	0.0
"2-0-1-1-0-1-1-0"	0.0	0.0	0.0	0.4730149171001663	0.0	0.0	0.4862992879777978	0.0
"2-0-1-1-1-1-0"	0.0	0.0	0.0	0.0	0.0	0.0	0.97259857595955956	0.0
"2-2-1-1-1-1-0"	0.0	0.0	0.49999999999999994	0.0	0.0	0.0	0.4862992879777978	0.0

Table C.13: ESC-specific TF binding weights obtained from the intra-class Gini impurity measure.

Statistic	Value
Connected components	1
Network diameter	4
Network radius	1
Shortest paths	97 249 (5%)
Characteristic path length	3.05
Average number of neighbours	7.44
Number of nodes	1340

Table C.14: Additional notable network statistics of the full CSC-gene network.

in-degree ranking	weighted in-degree	out-degree ranking	weighted out-degree
e0-0-0-0-0-0-0	430.4781	STAT1	1859.163797
r0-0-2-0-0-0-0-0	281.0	Gata3	461.600283
Cyp11a1	232.4478	Tbet	403.532628
Lrrc32	230.1966	STAT4	376.4427676
Ifngr2	224.9328	e0-0-0-0-0-0-0-0	364.3924
Il10	215.244	STAT6	326.581588
Pparg	209.4506	r0-0-2-0-0-0-0-0	211.9819
Runx3	202.2571	e0-0-1-0-0-0-0-0	129.3448
Il4	198.957	r0-2-0-0-0-0-0-0	95.6674
Tbet	189.8796	e1-0-0-0-0-0-0-0	93.1462
Cd83	181.1618	e0-1-0-0-0-0-0-0	79.0954
Gata3	179.969	e0-0-1-0-1-0-0-0	71.6271
Il5	173.3125	e0-0-1-1-1-0-0-0	62.6262
Il18r1	169.2575	r2-2-2-2-2-2-2-2	61.6774
Il13	159.7898	e0-0-2-0-0-0-0-0	50.7003
Klrc2	156.2886	e1-1-1-1-1-1-1-1	50.1544
Ccl5	153.1107	r2-0-2-0-0-0-0-0	49.8278
Sept8	150.8791	e0-1-0-1-1-1-1-1	44.1795
Smpd13b	149.8061	e0-1-0-1-0-1-0-1	41.2277
Inpp4b	147.7018	r0-0-2-2-0-0-0-0	38.7814
Klri2	144.8959	r0-0-2-0-2-0-0-0	38.0104
Il18rap	142.2748	e1-0-1-1-1-1-1-1	37.7825
Ifng	140.6242	r0-2-2-0-0-0-0-0	37.7482
Klrc1	140.3436	e0-0-1-1-0-0-0-0	36.7985
e1-0-0-0-0-0-0-0	130.09721	e0-0-1-1-1-0-1-0	35.4658
e0-0-1-0-0-0-0-0	123.44356	e1-0-1-1-1-1-1-0	34.1413
Rad50	123.4072	r0-0-2-0-0-0-2-0	32.4089
Klre1	121.0694	e0-0-1-0-1-0-1-0	32.225
r2-2-2-2-2-2-2-2	104.0	e0-0-0-0-1-0-0-0	31.8951
Clec12a	103.7875	e0-2-1-0-0-0-0-0	31.6602
Fasl	102.2885	e0-1-0-0-0-0-0-1	31.6189
Igfbp4	95.7316	r0-0-0-2-0-0-0-0	31.0159
STAT6	94.3875	r0-0-0-0-0-0-0-0	30.8315
Asb2	85.7799	e0-0-0-1-0-0-0-0	30.284
Klrb1c	83.1478	e0-2-1-0-1-0-0-0	28.558
Il9r	82.4624	r2-2-0-0-0-0-0-0	27.8229
e1-1-1-1-1-1-1-1	74.56352	r0-2-1-0-1-0-0-0	27.6822
e0-0-1-1-1-0-0-0	71.8119	r0-0-2-2-2-0-0-0	27.4607
e0-0-1-0-1-0-0-0	71.25732	e0-1-1-1-1-1-1-1	27.4528
St8sia6	67.1279	e1-0-1-1-1-0-0-0	27.2851
Itga1	64.9287	r2-2-2-2-2-0-2-2	27.1723
Il1r1	63.8236	r0-2-1-0-0-0-0-0	27.1283
Il12rb2	61.8969	e0-0-0-0-0-0-1-0	26.8396
Kcnj8	59.5323	e0-0-1-1-0-0-1-0	26.1544
r0-0-2-2-0-0-0-0	59.0	e0-1-2-0-2-0-0-0	25.4834
e0-1-0-0-0-0-0-0	58.7983	r0-0-2-0-2-0-2-0	24.7955
Sell	57.3402	e0-2-0-0-0-0-0-0	23.9652
Bhlhe41	56.9197	e0-0-0-0-0-0-0-1	23.6233
Ccr4	53.6894	r0-1-2-0-2-0-0-0	23.4256
e1-0-1-1-1-1-1-1	49.13567	e1-0-1-1-1-0-1-0	23.3435

STAT1	48.3981	e0-2-1-1-1-0-0-0	22.5645
Klrg1	47.6526	r0-0-2-2-2-0-2-0	22.2425
e0-0-1-1-1-0-1-0	43.57883	r2-2-2-2-0-2-2-2	21.1673
e0-0-1-1-0-0-0-0	42.15275	e0-0-1-1-1-1-0-0	20.7887
Chil3	41.9597	r2-2-2-2-2-2-2-0	20.5895
e0-0-0-0-1-0-0-0	41.44587	e0-1-0-1-0-0-0-1	20.5389
r0-0-2-0-2-0-0-0	41.29173	r2-2-2-2-2-0-2-0	20.3377
e0-0-0-1-0-0-0-0	40.92956	e0-1-0-1-0-1-1-1	20.0761
e1-0-1-1-1-1-1-0	40.11151	e0-1-0-0-0-0-1-1	19.7563
r0-0-0-0-0-0-0-0	39.59418	r0-2-1-1-1-0-0-0	19.5051
Galnt3	38.0271	r0-2-2-2-2-2-2-2	19.4723
e0-0-0-0-0-0-1-0	35.9048	e0-0-1-0-0-0-1-0	19.059
STAT4	35.8411	r0-0-2-2-0-0-2-0	18.6936
e0-0-1-0-1-0-1-0	35.52864	e0-0-1-1-1-1-1-0	18.4777
Gldc	35.2127	e1-0-0-0-0-0-1-0	18.069
Cxcr3	35.1503	e0-1-0-0-0-1-0-1	17.9377
e1-0-1-1-1-0-0-0	34.57696	e0-1-0-0-0-0-1-0	17.5601
Irf1	34.5176	e0-0-2-0-2-0-0-0	17.5301
r0-0-2-0-0-0-2-0	33.8171	e1-1-0-0-0-0-0-0	16.7017
e0-0-2-0-0-0-0-0	33.6866	e0-0-1-1-0-1-0-0	16.1873
Exph5	33.418	e0-0-1-0-1-1-1-0	15.9475
e0-1-0-1-1-1-1-1	33.26613	r2-0-2-2-2-0-2-0	15.7406
Eomes	32.851	e2-1-0-1-0-1-0-1	15.5692
e1-0-1-1-1-0-1-0	32.1095	e1-0-1-1-1-1-0-0	15.556
e0-0-1-1-0-0-1-0	31.9611	e1-0-1-0-0-0-0-0	15.2486
e0-1-0-1-0-1-0-1	30.3267	e0-0-0-0-0-1-0-0	15.2396
Kif3a	29.9579	r0-2-2-0-2-0-2-0	14.7075
e1-0-0-0-0-0-1-0	29.0	e0-0-0-1-0-0-1-0	14.6841
e0-0-1-0-0-0-1-0	29.0	e0-0-1-1-0-1-1-0	14.588
e0-2-0-0-0-0-0-0	28.863025	r2-2-2-0-2-0-2-0	14.5377
e0-2-1-0-0-0-0-0	28.0	r0-0-2-2-2-2-2-2	14.4194
e0-0-1-1-1-1-1-0	27.38748	r0-0-2-0-0-1-1-0	13.7954
⋮	⋮	⋮	⋮

Table C.15: Leading ranked nodes for weighted in- and out-degree respectively. The prefix *e* denotes ESCs and *r* denotes RSCs.

betweenness ranking	betweenness centrality	closeness ranking	closeness centrality	eigenvector ranking	eigenvector centrality	Katz ranking	Katz centrality
STAT6	0.0262963124211	e0-0-0-0-0-0-0-0	0.123364744779	e0-0-0-0-0-0-0-0	0.518623712455	e0-0-0-0-0-0-0-0	0.3172524427
Gata3	0.0210517855169	STAT1	0.115169184811	Gata3	0.316329073272	Gata3	0.184660992781
Tbet	0.0204595714849	e1-0-0-0-0-0-0-0	0.114767898452	Pparg	0.192332167107	Pparg	0.13683589225
STAT1	0.0149616372569	e1-1-1-1-1-1-1-1	0.110903659447	Il18r1	0.184410444256	r0-0-2-0-0-0-0-0	0.131169452585
e0-0-1-1-0-0-1	0.014633435701	r0-0-0-0-0-0-0-0	0.102611797059	e0-1-0-0-0-0-0-0	0.16794591865	Il18r1	0.128674954877
STAT4	0.009028333618	r0-0-0-0-0-0-0-0	0.102347430826	Cd83	0.159876275186	Cd83	0.123401773449
e1-1-0-0-0-1-1	0.00771831822378	e0-0-0-0-0-1-0-0	0.101348882633	Il18rap	0.153552266581	Cyp11a1	0.117897105946
e1-1-0-1-0-1-0-0	0.0067214339059	r0-0-2-2-0-0-0-0	0.100116677373	Clecl2a	0.144549439547	Il4	0.114276201716
e1-1-1-1-1-1-1-1	0.00280590003695	e1-0-1-1-1-1-1-1	0.09980309132615	Il4	0.143412342264	Ilfng2	0.113667385326
e0-0-1-0-0-0-0-0	0.00277185191635	Gata3	0.0977400203438	Inpp4b	0.137847928267	Il18rap	0.111013738851
e1-0-1-0-1-0-0-0	0.00212661212269	r0-0-0-2-0-0-2-0	0.0963969988016	Sept8	0.132763687124	Inpp4b	0.105601168591
e1-1-1-1-0-1-1	0.00201776976996	e0-1-0-0-0-0-0-0	0.095751124581	e0-1-0-1-1-1-1-1	0.132568809356	Il5	0.104018600404
r0-0-1-1-1-1-1-0	0.00164603127292	STAT4	0.095751124581	Cyp11a1	0.132388795342	Smpdl3b	0.101317149342
e0-0-0-0-1-0-1-0	0.00157458603625	e0-1-0-0-0-0-0-0	0.0951976498724	Il5	0.130129050366	Clecl2a	0.101305836354
e1-0-1-1-0-0-1	0.000891390960615	e0-0-0-0-1-0-1-0	0.0949233050601	Ilfng2	0.129956037999	Klrc2	0.100998963192
e0-0-0-1-1-1-0	0.000752407648659	r2-2-2-2-2-0-2-0	0.0936739197394	Smpdl3b	0.127769241808	Klrc2	0.100209481252
e0-0-1-1-1-1-1	0.000571003727432	STAT6	0.0935749626587	r0-0-2-0-0-0-0-0	0.125455221828	Sept8	0.0995688241716
e0-0-0-0-1-0-1-1	0.000328201555943	Tbet	0.0930462905533	Klrc2	0.119863973857	Lrrc32	0.0989863861657
e1-1-0-0-1-0-0-1	0.000276850292088	e0-0-0-1-1-0-0-0	0.0918575279439	Il13	0.119542426643	Il13	0.0975276365335
e0-0-2-1-0-0-0-1	0.000274059462531	e0-0-0-1-0-0-0-0	0.0917503812141	Klrc2	0.1149567459	Klrc2	0.0954395411338
e0-1-1-1-0-1-1-1	0.000270152301151	e1-0-1-0-0-0-0-0	0.0914955190441	Klrc2	0.11391532657	-0-0-0-0-0-0-0	0.093451049252
e0-2-0-0-0-0-0-0	0.000245593001046	e1-1-0-0-0-0-0-0	0.0911004604059	Rad50	0.111346713037	Klrc1	0.0924246198706
e1-1-0-0-0-0-1-0	0.000217126539561	e0-0-0-0-0-0-1-0	0.0909900189388	Klrb1c	0.11041022257	Tbet	0.08834286689
e0-1-0-0-0-1-0-1	0.00021377544092	r2-2-2-2-2-0-2-0	0.0896231556426	STAT6	0.108052112389	Rad50	0.0873872189606
e2-0-1-1-0-0-0	0.000185311082607	e1-0-1-1-0-1-0	0.089022667178	Klrc1	0.1052681625	Klrc1	0.0871545634982
r0-2-0-0-0-0-0-0	0.000175264096201	e1-0-0-0-1-0-1-0	0.0889023259725	Klrc1	0.101755778439	Klrb1c	0.0833076450493
e2-0-0-0-0-0-0-0	0.000174147764378	e1-1-0-1-0-1-0-0	0.0887827138972	Lrrc32	0.101611606223	STAT6	0.0811653351859
e1-0-0-0-1-0-0-0	0.000174147764378	e1-1-0-0-0-1-0-0	0.0887827138972	Ilfng	0.0962191374398	Ilfng	0.0796446170865
e0-0-0-0-0-1-0-0	0.000173031432555	e1-0-0-0-1-0-0-0	0.0887827138972	e0-1-0-0-0-0-0-1	0.0898652004666	e0-0-1-0-0-0-0-0	0.0768455660587
e1-0-0-0-0-0-0-0	0.000172473266644	r0-0-0-0-1-0-0-0	0.0886646192186	e0-0-0-0-0-0-0-1	0.0895232047597	Runx3	0.0740110379671
e1-0-0-0-0-0-0-1	0.000166891607529	r2-0-2-2-0-0-0-0	0.088428180234	Ilfng	0.0881684884597	Ccl5	0.0739373316416
e2-0-1-1-0-1-1	0.000160193616591	e0-0-1-0-1-1-1-0	0.0880705530905	Tbet	0.0846277381304	Ccl5	0.0702170022457
e1-2-1-1-1-1-0	0.00014065780969	e1-0-0-1-1-0-0-0	0.0877263692798	r0-2-0-0-0-0-0-0	0.083614258051	e0-0-1-0-1-0-0-0	0.0674160524884
e2-0-1-1-1-1-0-1	0.000139541477867	e0-1-1-1-1-1-1-0	0.0853326084349	e0-1-1-1-1-1-1-1	0.0831464001465	e0-1-0-1-0-1-0-1	0.0622430304897
e0-1-0-1-2-0-2-1	0.000137308814221	e0-0-1-1-1-1-1-1	0.0838126892007	Il12rb2	0.07554524443295	Igfbbp4	0.0620878233608
r0-1-0-0-0-0-0-0	0.000130052657372	r0-2-2-0-0-0-2-2	0.0831091919493	e0-0-1-0-0-0-0-0	0.069316655372	Il12rb2	0.062065853747
r1-1-2-0-2-2-2	0.000126145495992	r0-2-0-0-0-0-0-0	0.0831091919493	Il10	0.0678237987157	e1-1-1-1-1-1-1-1	0.0619986455703
e0-2-1-0-1-1-2-0	0.000122796500523	r0-2-0-0-0-0-2-0	0.0826946822637	Kcmj8	0.0634572870829	e0-0-2-0-0-0-0-0	0.0612099003549
e1-1-0-0-1-1-1-0	0.000120005670966	e0-0-1-1-1-1-1-0	0.0825523480097	Igfbbp4	0.0592103652629	Kcmj8	0.0590375559972
e0-0-1-1-1-1-2-0	0.00011386584594	r0-2-2-2-2-2-2-0	0.082080612841	Ccr4	0.0577554284504	Il9r	0.056886257278
e0-0-0-0-1-0-1-0-1	0.000110270252771	e1-0-1-1-1-0-0-0	0.0819362857111	e1-1-1-1-1-1-1-1	0.0562290698695	Fasl	0.0562474738482
e0-2-0-0-1-0-0-0	0.00010270252771	e1-0-1-1-1-1-1-0-0	0.0817329698657	Ccl5	0.055019666189	e0-0-0-0-0-0-0-1	0.0551176541476

r1-0-1-1-1-0-1-0	9.4882049496e-05	r0-0-2-2-0-2-0-2	0.0812759009504	Il9r	0.053262184376	e0-1-1-1-1-1-1-1	0.0544991287686
e0-1-2-1-0-1-0-1	9.43300390381e-05	r0-0-0-1-0-0-0-0	0.0810771823661	e0-0-0-1-0-0-0-0	0.0522704441636	Asb2	0.05434222438042
e0-1-0-0-2-0-0-0	9.09810435693e-05	r0-2-2-2-2-2-2-2	0.0804868145334	S8sia6	0.0474169629234	r0-2-0-0-0-0-0-0	0.0542788174757
r1-1-2-2-0-1-1-1	8.87483799234e-05	r0-0-2-2-0-2-0-2	0.0804868145334	r0-0-2-2-2-0-2-0	0.047282367687	S8sia6	0.0535384048944
e2-1-0-1-1-1-0-1	8.87483799234e-05	e0-0-1-0-0-0-1-0	0.0803375289167	Rumx3	0.0463656902931	e0-0-1-1-1-0-0-0	0.0527935474178
r0-1-0-0-0-0-0-1	8.7073882189e-05	r0-2-2-0-0-0-2-0	0.0802919312052	e0-0-0-0-0-0-1-0	0.0459592646082	Ccr4	0.0502964615902
r1-1-0-1-0-1-0-1	8.59575503661e-05	e1-0-0-0-1-1-0-0	0.0802919312052	e0-1-0-1-0-1-1-1	0.0451099178647	Il1rl1	0.05011125308
r2-1-0-0-0-2-2	7.8701393517e-05	e0-0-0-0-1-1-0-0	0.0801420604766	Fasl	0.0448767545959	e0-0-0-1-0-0-0-0	0.0500562526058
e2-1-1-1-1-1-1	6.86544071106e-05	e0-2-2-0-0-2-2	0.0799049821392	e0-0-0-0-0-1-1-1	0.0447616023798	r0-0-2-0-0-0-2-0	0.0478344261752
e0-2-1-0-1-0-0-0	5.69329229698e-05	r0-2-2-2-2-0-2-2	0.0797129028552	Asb2	0.0445292512646	Irga1	0.0468602062483
e1-0-1-0-1-1-1	5.58165911468e-05	r0-0-2-2-2-0-2-0	0.0795613209079	e0-1-0-1-0-0-0-1	0.0436062581351	e1-0-1-1-1-1-1	0.04355549789357
e2-2-1-0-1-0-1-2	4.91186002092e-05	r2-2-2-2-2-2-2-2	0.0793315014061	Il1rl1	0.0423302490148	e0-0-0-0-0-1-0	0.0430870280049
r0-2-1-0-1-0-1-2	4.85604342977e-05	r2-2-2-2-2-2-2-2	0.0793315014061	e0-0-0-0-0-1-0-0	0.0419864726147	r0-0-2-2-0-0-0-0	0.0417295063401
r1-2-1-1-1-1-1-0	4.80022683863e-05	r2-0-2-2-2-0-2-2	0.0791421660806	e2-1-0-1-0-1-0-1	0.0417318227149	Sell	0.0416542060846
r0-2-1-0-0-0-1-0	4.74441024748e-05	e1-0-1-0-1-1-1-1	0.0789889373042	e0-0-1-1-1-1-1-1	0.0406757655758	e0-0-0-0-1-0-0-0	0.0415008932398
e2-1-2-0-2-1-0-1	4.68859365633e-05	r0-2-2-0-2-0-2-0	0.0787661937951	e0-0-1-1-1-0-0-0	0.0404453838129	Gldc	0.0405787625945
r0-1-2-2-0-0-0-0	4.29787751831e-05	e0-0-0-0-0-0-2-0	0.0785795440468	r0-0-2-0-0-0-2-0	0.0402499948511	STAT1	0.0402882640745
e0-0-0-1-0-0-0-1	4.24206092716e-05	r0-2-0-2-0-2-0-2	0.0783937768032	Gldc	0.0397127310399	e0-1-0-1-0-1-1-1	0.0390925525017
r0-2-1-1-1-1-0-0	4.18624433601e-05	r2-2-2-2-0-2-0-0	0.0782088858202	Irga1	0.038738241687	e0-1-0-1-0-0-0-1	0.0383881180517
e1-0-2-2-2-0-0-0	4.13042774486e-05	r2-0-2-2-2-0-0-0	0.0782088858202	A430108G06Rik	0.0385708060374	Chil3	0.0383409833257
e0-1-1-1-1-1-1-0	3.90716138028e-05	r0-0-0-0-0-0-2-0	0.0780248649124	e0-0-0-1-0-0-0-0	0.0377143264496	e0-0-1-1-0-0-0-0	0.0378560226824
r0-1-0-2-2-0-0-2	3.85134478913e-05	r0-0-2-2-2-2-2-2	0.0778417079525	e0-0-0-1-1-1-1-1	0.0373013649761	e0-0-1-1-1-0-1-0	0.0376092854285
e0-2-1-1-1-0-1-0	3.51644524225e-05	r2-2-2-2-2-2-2-2	0.0776594088706	r0-0-2-2-2-0-0-0	0.0349053495406	r0-0-2-2-2-0-2-0	0.037584831527
e2-1-0-1-1-1-1-0	3.40481205996e-05	e1-0-1-0-1-0-0-0	0.0776594088706	e1-1-1-1-1-1-1-0	0.0332392933986	Irf1	0.0375532266743
e0-1-2-0-2-1-0-1	3.29317887766e-05	r0-2-2-2-2-2-0-2	0.0774779616536	Irf1	0.0327370695933	Kif3a	0.037378653831
e2-0-1-1-0-0-0-0	3.23736228652e-05	r0-2-2-2-2-0-0-0	0.0774779616536	Kif3a	0.0321729200129	Klrg1	0.0372322429567
e1-1-0-0-0-2-1	3.18154569537e-05	r0-0-2-2-0-2-2-2	0.0774779616536	Adams13	0.0320092796785	e1-0-1-1-1-1-1-0	0.0371759495905
e1-1-0-0-0-0-2	3.18154569537e-05	e1-0-0-1-0-0-0-0	0.0774779616536	e0-1-0-1-0-0-1-1	0.0319233795494	e0-0-0-0-0-1-0-0	0.0368323340986
e0-1-1-1-1-1-1-1	2.90246273963e-05	r0-0-2-2-2-0-2-2	0.0771175990413	e0-1-0-0-0-1-0-1	0.0317922569139	A430108G06Rik	0.0367667309544
r1-1-2-2-0-0-1-1	2.79082955734e-05	r0-0-0-2-0-2-0-2	0.0771175990413	STAT1	0.0317442951828	r0-0-2-2-2-0-0-0	0.0365813832665
r0-1-0-1-0-0-0-0	2.79082955734e-05	e1-0-1-0-1-1-1-0	0.0769588477941	e0-1-0-0-0-0-1-1	0.0313756136849	e0-0-0-0-0-1-1-1	0.0365385423419
r0-2-1-0-1-0-1-0	2.73501296619e-05	r0-0-2-2-0-0-2-2	0.0769386718973	e1-0-1-1-1-1-1-1	0.0304458124001	e2-1-0-1-0-1-0-1	0.0363411277346
e1-1-0-1-0-1-1	2.67919637505e-05	e1-1-0-1-0-1-0-0	0.0766008996648	e0-0-0-1-0-0-1-0	0.0303600265599	e0-0-1-1-1-1-1-1	0.0363222744865
e0-1-0-1-0-0-1-1	2.56756319275e-05	e0-1-1-1-1-1-0-1-1	0.0766008996648	e1-1-0-1-1-1-1-1	0.029841083083	e1-0-1-1-1-0-0-0	0.0352796398905
r0-1-0-1-2-1-1-1	2.51174660161e-05	e1-0-0-0-1-1-1-0	0.0765832969694	e0-1-0-0-0-0-1-0	0.029351409957	Exp5	0.0348007040587
e1-0-1-0-1-2-0-1	2.06521387243e-05	e0-0-0-0-1-1-1-0	0.0765832969694	Chil3	0.0292699966717	e0-1-0-0-0-0-1-1	0.0347170691346
r1-1-2-2-0-0-2-0	1.89776409899e-05	e1-1-1-0-1-1-1-1	0.0764231713593	Sell	0.0287258348141	r0-0-2-0-2-0-0-0	0.0341532161286
r2-0-2-1-0-2-0-0	1.7861309167e-05	r2-2-2-2-2-2-0-2	0.0764068377598	e1-1-1-1-1-0-1-0	0.0284008499213	e0-1-0-0-0-1-0-1	0.0338401792145
r1-0-2-0-0-0-1-0	1.73031432555e-05	r2-0-2-2-2-0-0-2	0.0764068377598	e0-0-0-0-0-0-0	0.0279909915073	e0-1-0-0-0-0-1-0	0.03363594687
r0-1-2-0-0-0-1-0	1.61868114326e-05	r2-0-2-2-0-0-2-2	0.0764068377598	e0-0-0-1-0-1-0-1	0.0275352594336	e0-1-0-0-0-0-1-0	0.033617567776
e1-0-1-1-1-0-0-0	1.56286455211e-05	e1-1-0-1-1-1-0-0	0.0764068377598	e2-1-2-1-0-1-0-1	0.0273543521115	STAT4	0.0333517269122
r2-0-2-2-0-0-2	1.50704796096e-05	e1-1-1-1-1-1-1-0	0.0762462658701	e0-1-0-1-1-0-0-1	0.0273543521115	Adams13	0.03334420360183
r1-1-2-1-0-2-1	1.45123136982e-05	e0-1-0-1-0-1-1-1	0.0762311898569	r0-0-2-2-0-0-0-0	0.0263411319852	e0-2-0-0-0-0-0-0	0.0333091512642
e1-0-1-0-1-0-1-2	1.45123136982e-05	r0-0-2-0-0-0-0-0	0.0760701774962	Exp5	0.0252024784893	Bhlhe41	0.0331725212659
e0-2-1-0-0-0-1-0	1.11633182294e-05	e1-1-1-1-1-1-1-0	0.0757204295537	e0-0-2-1-0-1-0-1	0.0248675766507	e1-1-1-1-1-1-1-0	0.0329004807877
r1-1-0-0-1-0-2-1	7.81432276055e-06	r2-2-2-2-0-2-2-0-2	0.0753649263358	e0-1-0-1-0-0-0-0	0.0244921141825	e0-0-0-0-1-0-0-1-0	0.0328116936562

r1-1-0-0-1-0-0-1	7.81432276055e-06	e2-2-0-1-0-0-0-0	0.0746859630355	e0-1-0-0-0-1-0-0	0.0244921141825	e0-1-0-1-0-0-1-1	0.0318523215363
r1-1-0-0-0-0-2-1	7.81432276055e-06	e1-1-1-0-0-0-0-0	0.0746859630355	e0-0-0-1-0-0-0-1	0.0244921141825	Trem12	0.0317024014799
r2-2-2-2-2-2-2	0.0	r0-2-0-0-0-0-0-0	0.0740188468671	e0-0-0-0-0-0-1-1	0.0244921141825	e1-1-1-1-1-0-1	0.0312272020052
r2-2-2-2-2-2-0	0.0	r0-2-2-0-0-0-0-0	0.0732021359553	Gja1	0.0237756849694	Galnt3	0.0310241695874
r2-2-2-2-2-0-2	0.0	e0-0-1-1-1-0-0-0	0.0728725372917	Klrg1	0.0227910961704	e0-0-1-1-1-1-0	0.0309663348915
r2-2-2-2-2-0-0	0.0	e1-1-0-1-0-1-1-1	0.0727205429556	e0-0-1-1-0-0-0-0	0.02272630359	e0-0-1-0-1-0-1-0	0.0305528101939
r2-2-2-2-0-2-2	0.0	e0-1-0-1-0-0-1-0	0.0727116707635	e0-0-1-1-1-0-1-0	0.0224288678275	e2-0-0-0-0-0-0-0	0.0305413696943
r2-2-2-2-0-2-0	0.0	e0-0-0-1-0-0-0-1	0.0723920590239	e1-1-0-1-0-1-0-1	0.0223808011899	Gja1	0.0303557435445
r2-2-2-2-0-0-2	0.0	e1-0-0-0-0-1-0-0	0.0722452452892	e1-0-0-0-1-0-0-0	0.0223568992304	e1-1-0-1-1-1-1-1	0.0303455112287
r2-2-2-2-0-0-0	0.0	e0-1-0-1-0-1-0-0	0.0720881904082	Gm17334	0.0223235493961	e0-0-1-1-1-1-0-0	0.0300519630638

Table C.16: Respective top 100 node rankings w.r.t. betweenness, closeness, eigenvector and Katz centralities. The prefix e denotes ESCs and r denotes RSCs.

module 1	module 2	module 3	module 4	module 5	module 6
Ccr1	A430108G06Rik	Ccl5	Asb2	Cxcr3	Ccr2
Gata3	Adamts13	Ccr5	Cd83	Eomes	e0-1-0-0-0-1-1-0
Gm12214	Bhlhe41	Clec12a	Sell	Galnt3	e0-1-0-0-1-0-1-1
Il13	Ccr4	Exph5	St8sia6	Il2	e2-0-0-0-0-1-0-0
Il4	Chil3	Fasl	e0-0-0-0-0-2-0-2	Kcnj8	
Il5	Cyp11a1	Glhc	e0-0-0-1-0-0-1-1	Runx3	
Rad50	Dpysl3	Gm17334	e0-0-0-1-0-1-0-0	e0-0-0-0-0-0-0-2	
Sept8	Efna5	Ifng	e0-0-2-0-0-0-2-2	e0-0-0-0-0-2-2-0	
e0-1-0-0-0-0-1-0	Gja1	Il12rb2	e0-0-2-0-2-0-2-2	e0-0-0-1-0-0-0-2	
e0-1-0-0-0-0-1-1	Ifngr2	Il18r1	e0-0-2-0-2-2-0-0	e0-0-0-1-0-0-2-0	
e0-1-0-0-0-1-0-0	Igfbp4	Il18rap	e0-0-2-1-2-0-0-0	e0-0-0-2-0-0-0-2	
e0-1-0-0-0-1-0-1	Il10	Itga1	e0-0-2-2-0-0-1-0	e0-0-0-2-0-0-2-0	
e0-1-0-0-1-0-0-1	Il1rl1	Klrb1c	e0-0-2-2-2-2-2-0	e0-0-0-2-0-2-0-0	
e0-1-0-0-1-1-1-0	Il9r	Klrc1	e0-1-0-1-0-1-0-0	e0-0-1-0-0-0-2-0	
e0-1-0-0-1-1-1-1	Inpp4b	Klrc2	e0-1-1-0-0-0-0-0	e0-0-1-0-0-2-2-0	
e0-1-0-1-0-0-0-0	Irf1	Klre1	e0-1-1-0-0-0-1-1	e0-0-1-0-1-0-2-0	
e0-1-0-1-0-0-1-0	Kif3a	Klrg1	e0-1-1-0-0-1-1-1	e0-0-1-1-0-0-2-0	
e0-1-0-1-0-0-1-1	Lrrc32	Klri2	e0-1-1-1-0-0-0-0	e0-0-1-1-0-1-1-1	
e0-1-0-1-0-1-1-0	Mctp1	STAT1	e0-1-2-0-0-0-0-2	e0-1-1-1-0-0-0-1	
e0-1-0-1-1-0-0-1	Pparg	STAT4	e0-1-2-2-0-2-0-0	e0-2-0-0-0-2-0-2	
e0-1-0-1-1-0-1-0	Slc4a4	STAT6	e1-0-0-0-0-0-1-1	e0-2-0-0-2-0-0-0	
e0-1-0-1-1-0-1-1	Trem12	Smpd13b	e1-0-0-0-0-0-2-0	e0-2-0-0-2-0-2-0	
e0-1-0-1-1-1-1-0	e0-1-0-0-0-0-0-0	Tbet	e1-0-0-0-0-1-1-0	e0-2-0-1-0-0-0-2	
e0-1-0-1-2-0-2-1	e0-1-0-0-0-0-0-1	e0-0-1-0-0-0-0-0	e1-0-0-0-1-0-1-1	e0-2-0-2-0-2-0-0	
e0-1-0-2-0-1-0-0	e0-1-0-0-0-0-2-0	e0-0-1-0-0-0-0-2	e1-0-0-0-2-0-0-0	e0-2-0-2-0-2-0-2	
e0-1-1-1-0-1-1-0	e0-1-0-0-0-0-2-1	e0-0-1-0-0-0-1-0	e1-0-0-1-0-0-0-1	e0-2-0-2-0-2-2-2	
e0-1-1-1-0-1-1-1	e0-1-0-0-0-1-1-1	e0-0-1-0-0-0-1-1	e1-0-0-1-0-1-0-0	e0-2-0-2-2-0-2-0	
e0-1-1-1-1-0-1-1	e0-1-0-0-1-0-0-0	e0-0-1-0-0-1-0-0	e1-0-0-1-1-0-1-1	e0-2-1-2-0-2-0-0	
e0-1-2-0-0-0-0-1	e0-1-0-0-1-1-0-0	e0-0-1-0-1-0-0-0	e1-0-1-0-0-1-1-0	e1-0-0-0-0-1-0-0	
e0-1-2-0-0-0-1-1	e0-1-0-0-2-0-0-0	e0-0-1-0-1-0-0-1	e1-0-2-0-0-0-1-0	e1-0-0-0-0-2-0-0	
e0-1-2-0-0-1-0-1	e0-1-0-0-2-0-1-0	e0-0-1-0-1-0-0-2	e1-0-2-0-0-2-0-2	e1-0-0-1-0-0-0-0	
e0-1-2-0-2-1-0-0	e0-1-0-0-2-0-1-1	e0-0-1-0-1-0-1-0	e1-0-2-0-1-0-1-0	e1-0-0-1-0-0-2-0	
e0-1-2-1-0-0-0-1	e0-1-0-0-2-0-2-0	e0-0-1-0-1-0-1-1	e1-0-2-2-0-0-0-0	e1-0-0-1-1-0-0-0	
e0-1-2-1-0-1-0-0	e0-1-0-0-2-0-2-1	e0-0-1-0-1-1-0-0	e1-0-2-2-0-2-0-0	e1-0-0-2-0-0-1-0	
e1-0-0-0-1-0-0-0	e0-1-0-0-2-2-0-0	e0-0-1-0-1-1-0-1	e1-1-0-0-0-0-0-2	e1-1-0-0-0-1-0-0	
e1-0-0-1-0-0-1-0	e0-1-0-1-0-0-0-1	e0-0-1-0-1-1-0-2	e1-1-0-0-0-0-2-0	e1-1-0-0-1-0-1-1	
e1-0-0-1-0-1-1-1	e0-1-0-1-0-0-2-1	e0-0-1-0-1-1-1-0	e1-1-0-0-0-0-2-2	e1-1-0-0-1-2-1-0	
e1-0-2-2-2-0-0-0	e0-1-0-1-0-1-0-1	e0-0-1-0-1-1-1-1	e1-1-0-0-2-0-0-0	e1-1-0-1-0-1-0-0	
e1-1-0-0-0-1-0-1	e0-1-0-1-0-1-1-1	e0-0-1-0-1-2-0-0	e1-1-0-0-2-0-2-0	e1-1-0-1-0-1-0-1	
e1-1-0-0-1-1-0-0	e0-1-0-1-0-1-2-1	e0-0-1-0-2-0-0-0	e1-1-0-0-2-0-2-2	e1-1-0-1-1-1-0-0	
e1-1-0-1-0-1-1-1	e0-1-0-1-0-2-2-1	e0-0-1-1-0-0-0-0	e1-1-0-1-0-0-0-0	e1-1-0-2-0-0-0-0	
e1-1-0-1-1-1-1-1	e0-1-0-1-1-0-0-0	e0-0-1-1-0-0-0-1	e1-1-0-1-1-0-1-1	e1-1-1-0-1-0-1-1	
e1-1-1-1-1-1-0-0	e0-1-0-1-1-1-0-1	e0-0-1-1-0-0-1-0	e1-1-0-2-0-2-0-0	e1-1-1-1-0-0-0-0	
e1-1-2-0-0-0-1-1	e0-1-0-1-1-1-1-1	e0-0-1-1-0-0-1-1	e1-1-0-2-2-0-0-0	e1-1-1-1-0-0-0-1	
e2-0-0-1-1-1-1-1	e0-1-0-1-2-0-0-1	e0-0-1-1-0-0-1-2	e1-1-0-2-2-0-2-2	e1-2-0-0-1-0-0-0	
e2-0-0-2-0-0-1-0	e0-1-0-1-2-0-2-0	e0-0-1-1-0-1-0-0	e1-1-1-0-1-1-0-0	e1-2-0-0-2-0-2-0	
e2-0-0-2-0-1-0-1	e0-1-0-1-2-1-0-1	e0-0-1-1-0-1-1-0	e1-1-1-1-0-1-1-0	e1-2-1-0-0-0-0-0	
e2-0-2-0-0-0-1-0	e0-1-0-1-2-1-1-1	e0-0-1-1-1-0-0-0	e1-1-1-1-1-0-0-1	e1-2-1-0-0-2-0-0	
e2-1-0-1-0-1-0-1	e0-1-0-1-2-1-2-1	e0-0-1-1-1-0-0-1	e1-1-2-2-0-0-0-0	e1-2-1-2-0-0-0-0	
e2-1-0-1-0-1-1-1	e0-1-0-2-0-0-2-0	e0-0-1-1-1-0-1-0	e1-1-2-2-2-0-0-0	e1-2-1-2-0-2-0-0	
e2-1-2-0-2-1-0-1	e0-1-1-0-1-0-0-1	e0-0-1-1-1-0-1-1	e1-1-2-2-2-0-0-2		
e2-1-2-1-0-0-0-1	e0-1-1-0-1-1-0-1	e0-0-1-1-1-1-0-0	e1-1-2-2-2-0-2-0		
e2-1-2-1-0-1-0-1	e0-1-1-1-0-1-0-1	e0-0-1-1-1-1-0-1	e1-1-2-2-2-0-2-2		
	e0-1-1-1-1-0-1-1	e0-0-1-1-1-1-1-0	e1-1-2-2-2-2-2-0		
		e0-0-1-1-1-1-1-1			
		e0-0-1-1-1-1-2-0			
		e0-0-1-1-1-2-2-0			
		e0-0-1-2-0-0-0-0			
		e0-2-1-0-0-0-0-0			
		e0-2-1-0-0-0-0-2			
		e0-2-1-0-0-0-1-0			
		e0-2-1-0-0-0-2-0			
		e0-2-1-0-0-0-2-2			
		e0-2-1-0-0-2-0-0			
		e0-2-1-0-0-2-0-2			
		e0-2-1-0-1-0-0-0			
		e0-2-1-0-1-0-0-2			
		e0-2-1-0-1-0-1-0			

e0-2-1-0-1-1-0-0
e0-2-1-0-1-1-2-0
e0-2-1-0-1-2-0-0
e0-2-1-0-1-2-0-2

⋮ ⋮ ⋮ ⋮ ⋮ ⋮

Table C.17: GLay community clustering results with six modules or clusters for all genes and selected ESCs.

module 1	module 2	module 3	module 4	module 5
Ccr5	Adamtsl3	A430108G06Rik	Ccl5	Asb2
Clec12a	Cyp11a1	Ccr1	Ccr2	Bhlhe41
Cxcr3	Il10	Ccr4	Il2	Chil3
Eomes	Lrrc32	Cd83	e0-0-0-2-0-2-0-0	Igfbp4
Exph5	e0-0-0-1-0-0-2-0	Dpysl3	e0-1-0-0-1-0-1-1	St8sia6
Fasl	e0-0-0-1-2-0-2-1	Efnra5	e0-1-1-1-1-1-0-1	e0-0-0-0-2-0-0-2
Galnt3	e0-0-0-1-2-1-0-1	Gata3	e0-2-0-0-0-0-2-0	e0-0-2-0-0-2-0-0
Gldc	e0-0-0-2-0-0-2-0	Gja1	e0-2-0-0-0-0-2-2	e0-0-2-1-2-0-0-0
Ifng	e0-0-2-0-0-0-2-0	Gm12214	e0-2-0-2-0-2-0-0	e0-1-0-0-2-2-0-0
Il12rb2	e0-0-2-1-0-0-0-0	Gm17334	e0-2-1-0-0-0-0-2	e0-1-1-0-1-0-0-1
Il18r1	e0-0-2-2-0-0-2-0	Ifngr2	e0-2-1-0-0-0-2-0	e0-1-1-0-1-1-0-1
Il18rap	e0-0-2-2-2-0-2-0	Il13	e0-2-1-0-0-0-2-2	e0-1-1-1-0-0-0-0
Itga1	e0-1-0-0-0-0-2-0	Il1rl1	e0-2-1-0-0-2-0-0	e0-1-2-0-0-0-1-0
Kcnj8	e0-1-0-0-0-0-2-1	Il4	e0-2-1-0-1-2-0-0	e0-1-2-0-0-2-0-0
Klrb1c	e0-1-0-0-1-0-0-0	Il5	e0-2-1-1-0-0-0-2	e0-1-2-2-0-0-0-1
Klrc1	e0-1-0-0-2-0-2-0	Il9r	e0-2-1-1-0-0-2-2	e0-1-2-2-0-0-2-1
Klrc2	e0-1-0-0-2-0-2-1	Inpp4b	e0-2-1-1-1-0-0-1	e1-0-0-0-0-0-0-2
Klre1	e0-1-0-1-0-0-2-1	Irf1	e0-2-1-1-1-0-2-0	e1-0-0-0-0-0-2-2
Klrg1	e0-1-0-1-0-1-2-1	Kif3a	e0-2-1-1-1-0-2-2	e1-0-0-0-2-0-2-0
Klri2	e0-1-0-1-0-2-2-1	Mctp1	e0-2-1-1-1-1-0-0	e1-0-0-1-1-1-1-1
Runx3	e0-1-0-1-2-0-0-1	Pparg	e0-2-1-1-1-1-1-0	e1-0-0-2-2-2-0-0
STAT1	e0-1-0-1-2-0-2-0	Rad50	e0-2-1-1-1-1-1-1	e1-0-0-2-2-2-2-0
STAT4	e0-1-0-1-2-1-0-1	STAT6	e0-2-1-2-0-2-0-0	e1-0-1-1-0-1-1-0
Smpdl3b	e0-1-0-1-2-1-1-1	Sell	e0-2-1-2-1-2-0-0	e1-0-2-2-2-0-2-0
Tbet	e0-1-0-1-2-1-2-1	Sept8	e1-0-0-0-0-2-0-0	e1-0-2-2-2-0-2-2
e0-0-1-0-0-0-0-0	e0-1-0-2-0-0-2-0	Sic4a4	e1-0-0-2-0-0-1-0	e1-1-0-0-0-2-1-0
e0-0-1-0-0-0-0-2	e0-1-1-1-0-1-0-1	Trem12	e1-0-1-0-1-1-0-0	e1-1-0-0-1-0-1-1
e0-0-1-0-0-0-1-0	e0-1-2-0-0-0-2-0	e0-1-0-0-0-0-0-0	e1-1-0-0-1-2-1-0	e1-1-0-0-2-0-0-0
e0-0-1-0-0-0-1-1	e0-1-2-0-0-2-2-1	e0-1-0-0-0-0-0-1	e1-2-1-0-0-0-0-0	e1-1-0-0-2-0-2-0
e0-0-1-0-0-0-2-0	e0-1-2-0-2-0-2-0	e0-1-0-0-0-0-1-0	e1-2-1-0-0-2-0-0	e1-1-0-2-0-0-1-1
e0-0-1-0-0-1-0-0	e0-1-2-1-0-0-2-0	e0-1-0-0-0-0-1-1	e1-2-1-1-1-0-1-1	e1-1-0-2-2-0-0-0
e0-0-1-0-0-1-1-0	e0-1-2-1-2-1-0-0	e0-1-0-0-0-1-0-0	e1-2-1-2-0-0-0-0	e1-1-0-2-2-2-0-0
e0-0-1-0-0-2-2-0	e0-1-2-2-0-0-2-0	e0-1-0-0-0-1-0-1	e1-2-1-2-0-2-0-0	e1-1-0-2-2-2-2-0
e0-0-1-0-1-0-0-0	e0-1-2-2-0-2-2-1	e0-1-0-0-0-1-1-0	e2-0-0-0-0-1-0-0	e1-1-1-0-0-0-1-0
e0-0-1-0-1-0-0-1	e0-1-2-2-2-0-2-0	e0-1-0-0-0-1-1-1	e2-0-1-1-1-0-0-0	e1-1-1-0-1-1-0-1
e0-0-1-0-1-0-0-2	e0-1-2-2-2-2-2-0	e0-1-0-0-1-0-0-1	e2-0-1-1-1-0-1-1	e1-1-1-1-0-1-1-0
e0-0-1-0-1-0-1-0	e0-1-2-2-2-2-2-1	e0-1-0-0-1-0-1-0	e2-2-1-1-0-0-0-2	e1-1-2-0-0-0-0-1
e0-0-1-0-1-0-1-1	e2-1-0-0-0-0-0-0	e0-1-0-0-1-1-0-0	e2-2-1-1-0-0-2-2	e1-1-2-0-0-1-0-1
e0-0-1-0-1-0-2-0	e2-1-0-0-0-0-0-1	e0-1-0-0-1-1-1-0	e2-2-1-1-1-0-0-2	e1-1-2-2-0-0-0-1
e0-0-1-0-1-0-2-1	e2-1-0-0-0-0-0-2	e0-1-0-0-1-1-1-1	e2-2-1-1-1-0-2-2	e1-1-2-2-0-0-1-1
e0-0-1-0-1-1-0-0	e2-1-0-0-0-1-0-0	e0-1-0-0-2-0-0-0	r0-0-0-0-0-0-0-2	e1-1-2-2-0-0-2-0
e0-0-1-0-1-1-0-1	e2-1-0-0-0-1-0-1	e0-1-0-0-2-0-1-0	r0-0-1-0-0-0-0-2	e1-2-1-0-1-1-1-0
e0-0-1-0-1-1-0-2	e2-1-0-0-0-2-2-1	e0-1-0-0-2-0-1-1	r0-0-1-1-0-0-0-0	e2-0-0-0-1-0-1-0
e0-0-1-0-1-1-1-0	e2-1-0-0-2-0-0-0	e0-1-0-1-0-0-0-0	r0-2-0-0-2-2-2-0	e2-0-0-0-1-1-1-0
e0-0-1-0-1-1-1-1	e2-1-0-0-2-0-0-1	e0-1-0-1-0-0-0-1	r0-2-0-0-2-2-2-2	e2-0-2-1-0-0-0-0
e0-0-1-0-1-2-0-0	e2-1-0-0-2-0-2-0	e0-1-0-1-0-0-1-0	r0-2-0-1-0-0-2-2	e2-0-2-1-0-0-0-2
e0-0-1-0-2-0-0-0	e2-1-0-0-2-0-2-1	e0-1-0-1-0-0-1-1	r0-2-0-2-0-0-0-0	e2-1-0-2-0-0-0-0
e0-0-1-1-0-0-0-0	e2-1-0-0-2-1-0-1	e0-1-0-1-0-1-0-0	r0-2-0-2-0-2-0-0	e2-1-0-2-0-0-1-1
e0-0-1-1-0-0-0-1	e2-1-0-0-2-1-2-1	e0-1-0-1-0-1-0-1	r0-2-0-2-0-2-0-2	e2-2-1-1-1-1-0-0
e0-0-1-1-0-0-0-1-1	e2-1-0-1-0-0-0-1	e0-1-0-1-0-1-1-0	r0-2-0-2-0-2-2-2	e2-2-2-1-2-0-0-2
e0-0-1-1-0-0-0-1-2	e2-1-0-1-0-1-0-0	e0-1-0-1-0-1-1-1	r0-2-1-0-0-0-0-2	r0-0-0-0-0-0-1-1
e0-0-1-1-0-0-0-2-0	e2-1-0-1-0-1-0-0	e0-1-0-1-1-0-0-1	r0-2-1-0-0-0-2-0	r0-0-0-2-0-0-1-0
e0-0-1-1-0-1-0-0	e2-1-0-1-0-1-2-0	e0-1-0-1-1-0-0-1	r0-2-1-0-0-0-2-2	r0-0-2-1-2-0-0-0
e0-0-1-1-0-1-0-1	e2-1-0-1-1-1-1-1	e0-1-0-1-1-0-1-1	r0-2-1-0-0-2-0-0	r0-0-2-2-0-0-2-1
e0-0-1-1-0-1-0-0-1	e2-1-0-1-2-0-0-1	e0-1-0-1-1-1-0-0	r0-2-1-0-1-2-0-0	r0-0-2-2-2-2-0-2
e0-0-1-1-1-0-1-0	e2-1-0-1-2-0-1-1	e0-1-0-1-1-1-0-1	r0-2-1-1-0-0-0-2	r0-1-2-0-0-0-1-0
e0-0-1-1-1-0-1-1	e2-1-0-1-2-0-2-0	e0-1-0-1-1-1-1-0	r0-2-1-1-0-0-2-2	r0-1-2-0-0-2-0-0
e0-0-1-1-1-1-0-0	e2-1-0-1-2-1-0-1	e0-1-0-1-1-1-1-1	r0-2-1-1-1-0-0-1	r0-1-2-2-0-0-0-1

e0-0-1-1-1-1-0-1	e2-1-0-1-2-1-1-1	e0-1-0-1-2-0-2-1	r0-2-1-1-1-0-2-0	r0-1-2-2-0-0-2-1
e0-0-1-1-1-1-1-0	e2-1-1-1-0-1-0-1	e0-1-0-2-0-1-0-0	r0-2-1-1-1-0-2-2	r1-0-0-0-2-0-2-0
e0-0-1-1-1-1-1-1	e2-1-1-1-0-1-1-1	e0-1-1-0-0-0-0-0	r0-2-1-1-1-1-1-0	r1-0-0-2-0-0-0-2
e0-0-1-1-1-1-2-0	e2-1-2-0-0-0-0-0	e0-1-1-0-0-0-1-1	r0-2-1-1-1-1-1-1	r1-0-0-2-0-0-2-2
e0-0-1-1-1-2-2-0	e2-1-2-0-0-1-2-1	e0-1-1-0-0-1-1-1	r0-2-1-2-0-2-0-0	r1-0-0-2-2-0-0-2
e0-0-1-2-0-0-0-0	e2-1-2-0-0-2-2-0	e0-1-1-1-0-1-1-0	r1-0-0-0-0-0-1-0	r1-0-0-2-2-2-2-0
e0-2-1-0-0-0-0-0	e2-1-2-0-0-2-2-1	e0-1-1-1-0-1-1-1	r1-2-1-0-0-0-0-0	r1-0-2-0-0-0-2-2
e0-2-1-0-0-0-1-0	e2-1-2-0-2-0-0-0	e0-1-1-1-1-0-1-1	r1-2-1-1-1-0-1-1	r1-0-2-0-0-2-2-0
e0-2-1-0-0-2-0-2	e2-1-2-0-2-0-0-1	e0-1-1-1-1-1-0-0	r2-2-0-0-0-2-2-0	r1-0-2-0-2-2-2-0
e0-2-1-0-1-0-0-0	e2-1-2-0-2-0-2-0	e0-1-1-1-1-1-1-0	r2-2-1-1-0-0-0-2	r1-0-2-2-2-2-0-2
e0-2-1-0-1-0-0-2	e2-1-2-0-2-0-2-1	e0-1-2-0-0-0-0-0	r2-2-1-1-0-0-2-2	r1-1-0-0-0-0-2-0
e0-2-1-0-1-0-1-0	e2-1-2-0-2-1-0-0	e0-1-2-0-0-0-0-1	r2-2-1-1-1-0-0-2	r1-1-0-0-0-1-0-1
e0-2-1-0-1-1-0-0	e2-1-2-0-2-1-2-1	e0-1-2-0-0-0-0-2		r1-1-0-0-2-0-0-0
e0-2-1-0-1-1-2-0	e2-1-2-1-0-0-2-0	e0-1-2-0-0-0-1-1		r1-1-0-2-2-0-0-0
e0-2-1-0-1-2-0-2	e2-1-2-1-0-0-2-1	e0-1-2-0-0-0-2-1		r1-1-0-2-2-2-0-0
e0-2-1-1-0-0-0-0	e2-1-2-1-0-1-0-0	e0-1-2-0-0-1-0-1		r1-1-0-2-2-2-2-0
e0-2-1-1-0-1-1-0	e2-1-2-1-0-1-2-1	e0-1-2-0-2-0-0-0		r1-1-2-0-0-0-0-1
e0-2-1-1-0-0-0-0	e2-1-2-1-0-2-2-1	e0-1-2-0-2-0-0-1		r1-1-2-0-0-1-0-1
e0-2-1-1-1-0-0-0	e2-1-2-1-2-0-2-0	e0-1-2-0-2-0-1-0		r1-1-2-0-2-0-0-0
e0-2-1-1-1-0-1-0	e2-1-2-1-2-1-0-1	e0-1-2-0-2-0-2-1		r1-1-2-2-0-0-0-1
	e2-1-2-1-2-1-2-1	e0-1-2-0-2-1-0-0		r1-1-2-2-0-0-1-1
	e2-1-2-2-2-0-2-1	e0-1-2-0-2-1-0-1		r1-1-2-2-0-0-2-0
	e2-1-2-2-2-2-2-1	e0-1-2-1-0-0-0-1		r1-2-2-0-2-2-2-2
	e2-2-2-0-2-2-2-0	e0-1-2-1-0-1-0-0		r2-0-0-2-0-2-0-0
		e0-1-2-1-0-1-0-1		r2-0-0-2-1-0-1-0
		e0-1-2-1-0-1-1-1		r2-0-2-0-0-2-0-2
		e0-1-2-2-0-0-0-0		r2-0-2-1-0-0-0-0
		e0-1-2-2-0-2-0-0		r2-1-2-2-0-0-0-0
		e0-1-2-2-0-0-0		r2-2-2-2-0-0-1-0
⋮	⋮	⋮	⋮	⋮

Table C.18: Spectral community clustering ($k = 5$) results with six modules or clusters for all genes and selected CSCs.

module 1	module 2	module 3	module 4	module 5	module 6	module 7	module 8	module 9	module 10
Adams13	A430108G06Rik	Il2	Ccr2	Kcnj8	Klrc1	Inpp4b	Asb2	St8sia6	Il9r
Bhlhe41	Ccr1	e0-0-0-2-0-2-0-0	Exph5	e0-0-0-0-0-2-2-0	Klrc2	e0-0-2-0-1-0-0-0	e0-0-2-1-2-0-0-0	e0-1-1-1-0-0-0-0	e0-1-0-0-2-0-1-0
Cd5	Gm12214	e0-1-0-0-1-0-1-1	Pparg	e0-0-1-0-0-2-2-0	Klre1	e0-0-2-0-1-1-0-0	e0-1-1-0-0-0-0	e1-0-0-0-0-0-1-1	e0-1-2-0-2-0-0-1
Ccr4	Gm17334	e0-1-0-1-1-0-1-1	e0-0-0-2-2-0-0	e0-2-0-0-2-0-2	Klri2	e0-0-2-1-1-1-1-0	e1-0-0-0-0-1-1-0	e1-1-0-0-2-0-0-0	e0-1-2-0-2-0-1-0
Ccr5	Il13	e0-1-1-1-0-1-1	e0-0-2-0-2-2-0	e0-2-0-1-0-0-2		e0-1-0-0-1-1-0-0	e1-0-0-1-1-1-1-1	e1-1-0-0-2-0-2-0	e1-0-0-0-0-1-1-1
Cd83	Il4	e0-1-1-1-1-0-1	e0-1-0-0-0-1-1-0	e0-2-0-2-0-2-0-2		e0-1-0-1-1-1-0-0	e1-0-1-1-0-1-1-1	e1-1-0-1-1-0-1-1	e1-0-1-0-0-1-1-1
Cxcr3	Il5	e0-2-0-2-0-2-0-0	e0-1-1-1-1-1-1-0	e0-2-0-2-0-2-2-2		e1-0-0-1-1-1-0-1	e1-1-1-1-0-1-1-0	e1-1-0-2-2-0-0-0	e1-1-0-0-2-0-1-0
Cyp11a1	Irf1	e0-2-1-2-0-2-0-0	e0-1-2-0-0-0-2-1	e1-0-0-0-0-0-2		e1-1-0-1-0-1-1-0	e2-0-0-0-1-1-1-0	r0-0-2-2-2-0-0-2	e2-0-2-0-2-0-1-0
Dpysl3	Kif3a	e1-0-0-0-2-0-0	e0-1-2-1-0-1-1-1	e1-0-0-0-1-0-0		e1-1-0-1-1-1-1-0	e2-0-2-1-0-0-0-2	r0-0-2-2-2-0-2	r0-0-2-0-2-0-0-1
EfnA5	Rad50	e1-0-0-2-0-0-1-0	e1-2-1-1-0-0-2	e1-0-0-1-1-0-0-0		e1-1-1-1-0-1-0-1	e2-0-2-1-0-0-0-2	r1-0-0-2-2-0-0-2	r0-1-2-0-2-0-0-1
Fas1	Sept8	e1-0-1-0-1-0-0	e2-0-0-0-0-1-0	e1-1-0-0-1-0-0-0		e2-0-0-0-0-0-2	e2-2-2-1-2-0-0-2	r1-0-2-0-0-0-2-2	r0-1-2-0-2-0-1-0
Galnt3		e1-1-0-0-1-2-1-0	e2-0-0-0-0-1-0-0	e1-1-0-0-1-0-1-1		e2-0-0-0-0-0-0-0	r0-0-2-1-2-0-0-0	r1-0-2-0-0-2-2-0	r2-0-2-0-2-0-1-0
Gata3		e1-2-1-0-0-0-0-0	e2-0-0-0-1-0-0-0	e1-1-0-2-0-0-0-0		e2-0-2-0-2-0-0-0	r2-0-0-2-1-0-1-0	r1-0-2-2-2-2-0-2	r2-1-2-0-2-1-0-1
Gja1		e1-2-1-0-0-2-0-0	e2-0-0-0-1-0-0-2	e1-1-1-0-1-0-0-0		e2-0-2-0-2-0-2-2	r2-0-0-2-1-1-0-0	r1-1-0-2-2-0-0-0	
Gldc		e1-2-1-2-0-0-0-0	e2-0-0-0-1-0-1-0	e1-1-1-0-1-0-1-1		r0-0-2-0-1-0-1-0	r2-0-2-0-2-0-0-2	r1-2-2-0-2-2-2-2	
Ifng		e1-2-1-2-0-2-0-0	e2-0-0-2-0-0-2-0	e1-1-1-0-1-1-0-0		r0-2-0-1-0-1-0-0	r2-0-2-0-2-2-2-0		
Ifngr2		e2-0-1-1-1-0-0-0	e2-0-1-0-1-0-0-0	e2-0-1-0-0-2-2-0		r0-2-1-1-0-1-1-0	r2-0-2-1-0-0-0-0		
Igfbp4		e2-0-1-1-1-0-1-1	e2-0-1-0-1-0-1-0	r0-0-0-0-2-2-0		r1-0-0-1-1-0-0-0	r2-1-2-2-0-0-0-0		
Il10		r0-2-0-0-0-0-2	e2-0-2-2-0-0-0-0	r0-0-0-0-2-2-2		r1-0-2-1-1-0-0-0	r2-0-2-0-2-2-2-2		
Il12rb2		r0-2-0-2-0-0-0-0	r0-0-0-2-0-2-0-0	r0-0-2-0-2-0-2		r1-0-2-1-1-0-0-0	r2-0-2-0-2-2-2-2		
Il18r1		r0-2-0-2-0-2-0-0	r0-0-0-2-0-2-2-2	r0-0-0-2-0-2-0-2		r2-0-0-2-2-2-2-2	r2-0-2-0-2-2-2-2		
Il18rap		r0-2-0-2-0-2-2-2	r0-1-2-0-0-0-2-1	r0-2-0-0-0-2-0-2		r2-2-0-2-0-2-2-2	r2-2-0-2-0-2-2-2		
Il1r1		r0-2-1-2-0-2-0-0	r0-1-2-1-0-1-1-1	r0-2-0-1-0-0-0-2		r2-0-0-2-2-2-2-2	r2-1-2-2-0-0-0-0		
Itga1		r1-0-0-0-0-0-1-0	r2-0-1-0-1-0-0-0	r2-0-1-0-0-0-0-0		r2-0-2-1-1-1-1-0	r2-0-2-1-0-0-0-0		
Klrg1		r1-0-0-0-0-0-1-0	r2-0-1-0-1-0-0-0	r0-2-0-2-2-2-2		r2-0-2-1-1-1-1-0	r2-0-2-0-2-2-2-0		
Lrrc32		r1-2-1-0-0-0-0-0	r2-0-2-2-0-0-2-2	r2-0-1-0-0-2-2-0		r2-0-2-0-2-2-0-2	r2-0-2-0-2-2-2-0		
Mctp1									
Runx3									
STAT4									
Sell									
Slc4a4									
Smpd13b									
Tbet									
Trem12									
e0-0-1-0-0-0-0-0									
e0-0-1-0-0-0-0-2									
e0-0-1-0-0-0-1-0									
e0-0-1-0-0-0-1-1									
e0-0-1-0-0-0-1-1									
e0-0-1-0-0-0-2-0									
e0-0-1-0-1-0-0-0									
e0-0-1-0-1-0-0-1									
e0-0-1-0-1-0-0-2									
e0-0-1-0-1-0-1-0									
e0-0-1-0-1-0-1-1									
e0-0-1-0-1-0-1-1									

e0-1-0-1-0-2-0	e0-1-2-0-0-0-1-1
e0-0-1-0-1-1-0-0	e0-1-2-0-0-1-0-1
e0-0-1-0-1-1-0-1	e0-1-2-0-2-0-0-0
e0-0-1-0-1-1-0-2	e0-1-2-0-2-1-0-0
e0-0-1-0-1-1-1-0	e0-1-2-1-0-0-0-1
e0-0-1-0-1-2-0-0	e0-1-2-1-0-1-0-0
e0-0-1-0-2-0-0-0	e1-0-0-0-1-1-1-1
e0-0-1-1-0-0-0-0	e1-1-0-1-1-1-1-1
e0-0-1-1-0-0-1-0	e1-1-1-1-1-1-0-0
e0-0-1-1-0-0-1-1	e2-0-0-1-1-1-1-1
e0-0-1-1-0-0-1-2	e2-0-0-2-0-0-1-0
e0-0-1-1-0-0-2-0	e2-0-0-2-0-1-0-1
e0-0-1-1-0-1-0-0	e2-0-2-0-0-0-1-0
e0-0-1-1-0-1-1-0	e2-1-0-1-0-1-0-1
e0-0-1-1-0-1-1-1	e2-1-0-1-0-1-1-1
e0-0-1-1-0-0-0-0	e2-1-2-1-0-0-0-1
e0-0-1-1-1-0-0-1	e2-1-2-1-0-1-0-1
e0-0-1-1-1-0-1-0	r0-0-0-0-1-0-0-1
e0-0-1-1-1-0-1-1	r0-0-0-0-1-0-1-1
e0-0-1-1-1-1-0-0	r0-0-0-0-2-0-2-0
e0-0-1-1-1-1-0-1	r0-0-0-1-1-1-1-0
e0-0-1-1-1-1-1-0	r0-0-0-1-1-1-1-1
e0-0-1-1-1-1-1-1	r0-0-0-2-0-0-0-1
e0-0-1-1-1-1-2-0	r0-0-0-2-0-1-0-0
e0-0-1-1-1-2-2-0	r0-0-0-2-0-1-1-1
e0-0-1-2-0-0-0-0	r0-0-2-0-0-0-1-0
e0-1-0-0-0-0-0-0	r0-0-2-0-0-1-0-0
e0-1-0-0-0-0-0-1	r0-0-2-0-0-1-1-0
e0-1-0-0-0-0-1-0	r0-0-2-0-2-0-1-0
e0-1-0-0-0-0-1-1	r0-0-2-0-2-0-1-1
e0-1-0-0-0-2-1	r0-0-2-0-2-1-0-0
e0-1-0-0-1-0-0-0	r0-0-2-0-2-1-1-1
e0-1-0-0-1-0-1-0	r0-0-2-1-0-0-0-1
e0-1-0-0-1-0-1-1	r0-0-2-1-0-0-1-0
e0-1-0-0-2-0-1-1	r0-0-2-1-0-1-0-0
e0-1-0-0-2-0-2-0	r0-0-2-1-0-1-0-1
e0-1-0-0-2-0-2-1	r0-0-2-1-0-1-1-0
⋮	⋮
⋮	⋮

Table C.20: RWIF clustering (Granularity 1.8) results with fourteen modules or clusters for all genes and selected CSCs.

module 11	module 12	module 13	module 14
Clec12a	STAT6	Chil3	Eomes
Klrb1c	e0-0-1-0-0-1-0-0	e1-1-2-0-0-0-0-1	e0-0-0-1-0-0-0-2
STAT1	e0-0-1-1-0-0-0-1	e1-1-2-0-0-1-0-1	e1-0-0-1-0-0-2-0
e0-0-0-0-1-1-0-0	e1-0-1-0-0-1-0-1	e1-1-2-2-0-0-2-0	e1-1-1-0-1-0-0
e0-0-1-0-0-1-1-0	e1-0-1-0-1-1-0-1	r1-1-0-0-0-1-0-1	e2-1-1-1-0-0-0-0
e0-0-1-0-1-1-1-1	e1-1-0-0-0-1-1-1	r1-1-2-0-0-0-0-1	
e1-0-0-0-1-1-0-0	e1-1-0-1-1-0-0-0	r1-1-2-0-0-1-0-1	
e1-0-0-0-1-1-1-0	r2-0-0-0-1-1-0-0	r1-1-2-2-0-0-2-0	
e1-0-1-0-1-0-1-2	r2-0-0-0-1-1-1-0		
e1-1-1-0-1-0-1-0	r2-0-0-2-0-0-0-2		
e1-1-1-0-1-1-1-1	r2-0-2-0-2-2-0-2		
r0-0-0-2-0-0-2-2			
⋮	⋮	⋮	⋮

Contd.: RWIF clustering (Granularity 1.8) results with fourteen modules or clusters for all genes and selected CSCs.

r1-1-0-0-0-2-2	r2-0-0-2-0-1-0-1	r2-2-1-1-0-0-0
r1-1-0-0-0-2-0-0	r2-0-2-0-0-1-0	r2-2-1-1-0-0-2
r1-1-0-0-0-2-1-0	r2-0-2-0-2-2-2-0	r2-2-1-1-1-1-0
r1-1-0-2-0-2-1-0	r2-1-0-1-0-1-0-1	r2-2-2-0-0-0-2
r1-1-2-2-0-0-0-0	r2-1-2-1-0-0-0-1	
r1-1-2-2-2-0-0-0	r2-1-2-1-0-1-0-1	
r1-1-2-2-2-0-0-2		
r1-1-2-2-2-0-2-0		
r1-1-2-2-2-0-2-2		
r1-1-2-2-2-2-0		
r1-2-2-0-0-0-0		
r2-0-2-0-0-1-0-0		
r2-0-2-0-2-1-1-1		
r2-0-2-1-0-1-1-1		
r2-0-2-2-0-0-2		
r2-0-2-2-0-2-0-0		
r2-0-2-2-0-2-0-2		
r2-0-2-2-0-2-2-2		
r2-0-2-2-2-0-0-2		
r2-0-2-2-2-0-0		
r2-0-2-2-2-0-2		
r2-1-2-0-0-1-0-0		
r2-1-2-0-2-0-1-1		
r2-1-2-0-2-1-1-1		
r2-1-2-1-2-0-0-1		
r2-1-2-1-2-0-1-1		
r2-2-2-0-0-2-0		
r2-2-2-0-0-2-0-0		
r2-2-2-0-2-0-0-2		
r2-2-2-0-2-2-0-2		
r2-2-2-2-0-0-0-0		
r2-2-2-2-0-0-0-2		
r2-2-2-2-0-0-2-0		
r2-2-2-2-2-0-2-0		

Table C.22: RWIF clustering (Granularity 2.0) results with 25 modules or clusters for all genes and selected CSCs.

module 21	module 22	module 23	module 24	module 25	module 26	module 27	module 28
Trem2	Chil3	Il1rl1	Cxcr3	Eomes	Klrg1	Gldc	Ccr4
e0-0-0-2-0-0-0	e1-1-2-0-0-0-1	e0-1-1-0-1-0-0-1	e0-0-1-1-0-0-2-0	e0-0-0-1-0-0-0-2	e2-0-1-1-1-0-1-1	r2-0-2-2-0-2-1-0	e0-1-0-0-2-0-1-1
e0-0-0-2-0-2-0	e1-1-2-0-0-1-0-1	e0-1-1-0-1-1-0-1	r0-0-0-0-0-2-2	e1-0-0-1-0-0-2-0	r2-0-0-2-2-2-2-2	r2-2-0-1-1-1-1-0	e1-1-0-0-2-0-1-1
e1-0-0-0-2-0-2-0	e1-1-2-2-0-0-2-0	e1-1-1-0-1-1-0-1	r2-2-0-0-0-0-1	e1-1-1-0-1-0-0	r2-2-0-2-0-2-0-2	r2-2-0-1-1-1-1-1	
e1-0-0-1-1-0-0	r1-1-0-0-0-1-0-1	e2-1-0-2-0-0-0-0	r2-2-0-0-0-0-1-0	e2-1-1-1-0-0-0-0			
e1-0-0-2-2-2-2-0	r1-1-2-0-0-0-1	e2-1-0-2-0-0-1-1	r2-2-0-2-2-0-0-0				
r0-0-0-0-0-2-0	r1-1-2-0-0-1-0-1						
r0-0-0-2-0-2-0	r1-1-2-2-0-0-2-0						
r1-0-0-0-2-0-2-0							

Contd.: RWIF clustering (Granularity 2.0) results with 25 modules or clusters for all genes and selected CSCs.

up-regulated ranking	betweenness centr. up	down-regulated ranking	betweenness centr. down
Gata3	0.0223281806795	Tbet	0.0573501687312
STAT1	0.0100694704161	STAT1	0.0259134423804
STAT6	0.0100201643869	STAT6	0.0205985300014
e0-1-0-0-0-0-0	0.00592561446682	STAT4	0.0175417701792
e0-1-0-1-1-0-1-1	0.00474342452205	e0-0-1-1-1-1-1-1	0.0160204898002
e1-1-0-1-0-1-0-0	0.00305692786914	e0-0-1-0-0-0-0-0	0.00806062877019
e1-1-0-0-0-1-0-0	0.00252291202605	e1-0-1-1-1-1-1-1	0.00711247535524
e0-1-0-1-0-0-1-0	0.00136851459094	e0-0-1-1-1-1-1-0	0.00510261240649
e1-1-0-0-0-0-1-0	0.00122216373623	Gata3	0.00499429899854
e0-1-0-1-1-1-1-1	0.00091008945064	e0-0-1-0-1-1-1-0	0.00470479033097
e1-1-0-0-1-0-0-1	0.000723468671049	e1-0-1-1-1-1-1-0	0.00435318955728
e0-1-0-0-0-0-1-0	0.000629631364228	r0-0-1-1-1-1-1-0	0.00433639970797
e0-1-0-1-0-1-0-1	0.000566881304531	e0-0-1-1-0-0-0-0	0.00252092586948
e0-1-0-1-0-0-0-1	0.000489045153238	e1-0-1-1-1-0-0-0	0.00200131240216
e0-1-0-0-0-0-0-1	0.000438923553793	e0-0-1-1-1-0-0-0	0.00152419339242
r0-1-2-0-0-0-0-0	0.000389931429669	e0-0-1-0-1-0-0-0	0.00142084747716
e0-1-0-1-0-0-1-1	0.000384970335038	e0-0-1-1-0-1-0-0	0.00137720623808
e0-1-0-1-0-0-0-0	0.00031004130216	e1-0-1-1-1-0-1-1	0.00125145498171
e1-1-0-0-1-1-1-0	0.000309296571227	e1-0-1-0-1-0-1-0	0.0011752536181
e0-1-0-0-0-0-1-1	0.000309089419098	e1-0-1-1-1-0-1-0	0.00111516135779
r0-1-0-1-0-1-0-0	0.000289034415662	e0-0-1-1-1-1-0-0	0.00100807209339
r0-1-2-2-2-0-2-0	0.000249646518204	e1-0-1-0-0-0-0-0	0.000898559755184
e0-1-0-0-0-1-0-1	0.000236150947166	e0-0-1-1-0-1-1-0	0.000850707433112
e1-1-0-0-1-0-1-0	0.000206252916319	e2-0-1-1-1-0-0-0	0.000839936834645
e0-1-0-0-0-1-0-0	0.000166097530923	e0-0-1-1-1-0-1-0	0.000776602869661
e0-1-0-1-0-1-1-1	0.000126168059784	e2-0-1-1-1-1-1-0	0.000586860047459
e2-1-0-1-0-1-0-1	0.000122717160622	e0-2-1-0-1-0-0-0	0.000492859172066
e2-1-2-1-0-1-0-1	0.000113048957276	e0-0-1-1-1-0-0-1	0.000436425805651
e0-1-2-1-0-1-0-1	9.82836495394e-05	e1-0-1-1-0-0-0-0	0.000418496313817
e1-1-0-0-0-0-0-1	9.09562172376e-05	e1-0-1-1-0-1-1-1	0.000333138503513
e0-1-0-1-1-0-0-1	8.66697517489e-05	e2-0-1-1-1-0-1-1	0.000305252374604
e1-1-0-1-1-1-1-1	8.35282758034e-05	e2-2-1-1-1-1-1-0	0.000223916924816
e1-1-0-0-0-0-1-1	6.49423166049e-05	e1-0-1-1-0-0-1-1	0.000208443382213
e1-1-0-1-0-1-0-1	6.40201358047e-05	e2-0-1-1-0-0-0-0	0.000188107357315
e0-1-2-0-2-1-0-1	6.26243375284e-05	e0-0-1-0-1-0-1-0	0.000186556659641
e1-1-0-0-0-1-0-1	5.51087938416e-05	e0-0-1-1-0-0-1-0	0.000165119070143
e2-1-2-0-2-1-0-1	5.47533355063e-05	e0-0-1-0-1-1-0-0	0.00016499261135
e0-1-0-0-0-1-1-1	4.87094695387e-05	e0-0-1-1-1-1-0-1	0.000161477831446
e0-1-2-0-2-0-2-1	4.70818186962e-05	e1-0-1-1-1-1-0-0	0.000144040170529
e2-1-0-1-1-1-0-1	4.37548463167e-05	r0-2-1-0-1-0-1-0	0.000125834945744
r0-1-0-0-0-0-0-0	3.80350421512e-05	e1-0-1-1-1-0-0-1	0.000124720492524
r0-1-2-0-2-0-0-0	3.76801775264e-05	e2-0-1-0-0-0-0-0	0.00012268625172
r2-1-0-1-0-1-0-1	3.5593653038e-05	e2-2-1-1-1-0-1-0	0.000119596030125
r0-1-2-2-2-0-0-0	3.44175383942e-05	e0-0-1-1-1-1-2-0	0.000109384394248
e2-1-2-0-2-0-2-1	3.42502075493e-05	e2-2-1-1-1-0-0-0	0.000107095664151
r1-1-0-0-0-0-0-0	3.31389183457e-05	e0-0-1-0-0-0-1-0	9.78196681383e-05
e1-1-0-1-1-1-1-0-1	3.21511798624e-05	e0-2-1-0-0-0-0-0	9.42793180525e-05
e1-1-0-1-0-0-0-1	3.01998953071e-05	e1-2-1-1-1-1-1-1	9.16463141233e-05
e2-1-0-1-0-1-1-1	2.78023111951e-05	e1-0-1-0-1-1-1-1	8.32764727225e-05
e0-1-0-1-2-0-2-1	1.90747792885e-05	e0-2-1-0-1-1-2-0	7.77383607361e-05
r1-1-2-2-0-2-2-2	1.76740897844e-05	e0-2-1-1-1-0-0-0	6.7499729586e-05
r0-1-0-1-0-0-0-0	1.57527878233e-05	e0-2-1-0-0-0-1-0	6.48371924277e-05
e1-1-0-1-1-0-1-1	1.39816881598e-05	e1-0-1-0-1-1-1-0	6.20361487937e-05
r1-1-2-2-2-0-0-0	1.28860083791e-05	e1-2-1-1-1-1-1-0	4.89345148182e-05
r1-1-2-2-2-0-0-2	1.13956337445e-05	r1-0-1-1-1-0-1-0	3.47764228179e-05
r0-1-0-0-0-0-0-1	1.00644122383e-05	e2-2-1-0-1-0-1-0	3.23611634988e-05
e1-1-2-2-2-0-2-0	9.8189387691e-06	e2-2-1-0-1-0-1-2	2.16623688764e-05
e1-1-2-2-0-0-0-0	9.8189387691e-06	e1-2-1-1-1-1-0-0	2.16623688764e-05
e1-1-0-0-0-0-0-2	9.8189387691e-06	e0-2-1-1-1-0-1-0	2.16623688764e-05
e0-1-0-1-1-1-0-1	9.4156609268e-06	e0-2-1-0-1-0-1-0	2.16623688764e-05
e2-1-0-1-1-1-1-0	9.3128079938e-06	e0-2-1-0-1-2-0-2	1.77336616528e-05
r2-1-2-1-0-2-2-1	8.52697314158e-06	e0-2-1-0-1-0-0-2	1.77336616528e-05
r1-1-0-0-0-0-2-2	6.85437456381e-06	e0-2-1-0-0-2-0-2	1.77336616528e-05
e2-1-0-0-2-0-0-0	6.5459591794e-06	e2-2-1-1-0-0-0-0	1.59676117209e-05
r1-1-2-2-0-0-1-1	5.78616034608e-06	r0-2-1-0-0-0-1-0	1.3464247245e-05
r1-1-2-2-2-0-2-0	5.01332354461e-06	r1-2-1-1-1-1-1-0	1.31817805196e-05
r0-1-2-0-0-0-2-0	3.92757550764e-06	r0-2-1-1-1-1-0-0	1.31817805196e-05
r0-1-2-0-0-0-1-0	3.68210203841e-06	r0-2-1-0-1-0-1-2	7.81491273661e-06
r0-1-0-1-2-1-1-1	3.68210203841e-06	e2-0-1-1-1-1-0-1	6.59036208135e-06

APPENDIX C. SUPPLEMENTARY TABLES

r1-1-0-0-0-0-2	3.1722725254e-06	e0-0-1-0-0-0-2	6.54349997897e-06
e1-1-0-1-1-0-0-1	2.66514052304e-06	e1-2-1-1-1-0-0-0	5.80807387951e-06
e1-1-0-0-1-0-2-1	2.66514052304e-06	e1-0-1-0-1-2-0-1	3.51677187146e-06
e1-1-0-0-0-0-2-1	2.66514052304e-06	e1-0-1-0-1-1-1-2	3.47871785475e-06
r1-1-0-0-1-0-2-1	1.47284081536e-06	e1-0-1-0-1-1-0-2	3.47871785475e-06
r1-1-0-0-1-0-0-1	1.47284081536e-06	e2-2-1-1-1-0-2-2	3.02672810751e-06
r1-1-0-0-0-0-2-1	1.47284081536e-06	e2-2-1-1-1-0-0-2	3.02672810751e-06

Table C.25: Betweenness centrality ranking with respective values for the Th2 vs. Tbet^{+/+}Th1 differential network with respective up- and down-regulated parts.

in-degree ranking up	in-degree up	out-degree ranking up	out-degree up	in-degree ranking down	in-degree down	out-degree ranking down	out-degree down
Il10	68.0	Gata3	29.0	Ifng	46.0	STAT1	33.0
Il5	50.0	e0-0-0-1-0-0-0-0	27.0	Smpdl3b	32.0	e0-0-1-0-0-0-0-0	26.0
Il4	48.0	STAT6	24.0	Il18r1	25.0	e0-0-1-0-1-0-0-0	20.0
Il13	45.0	e0-1-0-1-1-1-1-1	17.0	Il18rap	23.0	e0-0-1-0-0-0-1-0	15.0
Sept8	40.0	e0-1-0-1-0-1-0-1	16.0	Ccl5	21.0	Tbet	14.0
Gata3	40.0	e0-0-0-1-0-0-1-0	15.0	Tbet	20.0	e1-0-1-0-1-0-0-0	12.0
Cyp11a1	38.0	e0-1-0-1-0-0-0-1	14.0	Klrb1c	17.0	e1-0-1-0-0-0-0-0	12.0
Rad50	37.0	e0-1-0-1-0-0-1-1	12.0	Galnt3	17.0	e0-2-1-0-0-0-0-0	12.0
Lrrc32	34.0	e0-0-0-1-1-0-0-0	12.0	Clec12a	16.0	e0-0-1-0-1-0-1-0	12.0
Ifngr2	33.0	STAT1	12.0	Fasf	14.0	r0-2-1-0-0-0-0-0	11.0
Inpp4b	31.0	e1-0-0-1-0-0-0-0	11.0	STAT6	13.0	r0-2-1-0-1-0-0-0	9.0
Pparg	26.0	e0-0-0-1-1-1-1-1	11.0	Kcnj8	13.0	e1-0-1-0-1-0-1-0	9.0
STAT6	19.0	e0-0-0-1-1-1-1-0	11.0	Klri2	12.0	e0-2-1-0-1-0-0-0	9.0
Cd83	14.0	e0-0-0-1-0-1-0-0	11.0	Klre1	12.0	STAT4	9.0
Ccr4	14.0	e0-1-0-1-0-0-0-0	10.0	Klrc2	12.0	e1-0-1-0-1-1-1-1	8.0
Klf3a	12.0	r2-1-2-1-0-1-0-1	9.0	Exph5	12.0	e0-0-1-0-1-1-1-0	8.0
Asb2	11.0	r0-2-0-1-0-0-0-0	9.0	Klrc1	11.0	e1-1-0-1-0-1-0-0	7.0
Sell	10.0	e0-1-0-1-0-0-1-0	9.0	Il2	9.0	e1-0-1-0-1-1-1-0	7.0
Irf1	10.0	r0-1-2-1-0-1-0-1	8.0	Cd83	9.0	r0-0-1-0-1-0-0-0	6.0
Ifng	10.0	e2-1-2-1-0-1-0-1	8.0	Asb2	8.0	e1-1-0-1-1-1-1-1	6.0
Eomes	10.0	e1-1-0-1-0-1-1-1	8.0	STAT1	7.0	e1-0-1-0-0-0-1-0	6.0
Tbet	9.0	e0-0-0-1-0-0-0-1	8.0	Rumx3	7.0	e0-0-1-0-1-1-0-0	6.0
Klri2	9.0	r0-1-2-1-0-1-0-0	7.0	Itga1	6.0	r0-0-1-0-1-0-1-0	5.0
Klrc2	8.0	r0-1-0-1-0-1-0-1	7.0	Il12rb2	6.0	e1-0-1-0-1-1-0-0	5.0
A430108C06Rik	8.0	e2-1-0-1-0-1-1-1	7.0	Klrg1	5.0	e1-1-1-0-1-1-1-0	4.0
Klrc1	7.0	e2-1-0-1-0-1-0-1	7.0	e0-0-1-0-0-0-0-0	4.0	e1-1-0-1-1-0-0	4.0
Kcnj8	7.0	e1-1-0-1-1-1-1-1	7.0	St8sia6	4.0	e1-1-0-1-0-1-1-1	4.0
Il9r	7.0	e1-0-0-1-0-0-1-0	7.0	e2-2-1-0-1-0-1-2	3.0	e1-1-0-1-0-0-0	4.0
Il18rap	7.0	e0-2-0-1-0-0-0-0	7.0	e2-0-1-0-0-0-0-0	3.0	e1-0-1-0-1-0-0-0	4.0
Il12rb2	7.0	e0-1-0-1-1-0-1-1	7.0	e0-2-1-0-1-0-1-0	3.0	r2-2-1-0-1-0-1-0	3.0
Smpdl3b	6.0	e0-1-0-1-1-0-0-1	7.0	e0-2-1-0-1-0-0-0	3.0	r2-2-1-0-1-0-0-0	3.0
Klre1	6.0	e0-0-0-1-0-1-1-0	7.0	e0-0-1-0-1-1-1-0	3.0	r2-2-1-0-0-0-0-0	3.0
Itga1	6.0	STAT4	7.0	e0-0-1-0-1-0-0-0	3.0	e2-2-1-0-1-0-1-0	3.0
Il18r1	6.0	r0-1-0-1-1-1-1-1	6.0	STAT4	3.0	e2-2-1-0-1-0-0-0	3.0
Ccr1	6.0	r0-0-2-1-2-1-1-1	6.0	Irf1	3.0	e2-0-1-0-0-0-0-0	3.0
Il1r1	5.0	r0-0-2-1-0-0-1-0	6.0	Il9r	3.0	e1-1-1-0-0-0-0-0	3.0
St8sia6	4.0	e0-1-2-1-0-0-0-1	6.0	Il1r1	3.0	e1-0-1-0-1-0-1-1	3.0
STAT1	4.0	e0-1-0-1-0-1-1-1	6.0	Ifngr2	3.0	e1-0-1-0-0-1-1-1	3.0
Rumx3	4.0	e0-0-0-1-1-1-0-0	6.0	e1-0-1-0-1-0-1-0	2.0	e0-2-1-0-0-0-1-0	3.0
Gja1	4.0	r2-1-2-1-0-0-0-1	5.0	e0-2-1-0-1-1-2-0-2	2.0	e0-0-1-0-1-1-0-1	3.0
Ccl5	4.0	r2-1-0-1-0-1-1-1	5.0	e0-2-1-0-1-1-2-0	2.0	e0-0-1-0-1-1-0-1	3.0
STAT4	3.0	r2-1-0-1-0-1-0-1	5.0	e0-2-1-0-1-0-0-2	2.0	e0-0-1-0-0-1-0-0	3.0
				e0-2-1-0-0-2-0-2	2.0	Gata3	2.0
				e0-2-1-0-0-0-1-0	2.0	r2-2-1-0-0-0-0-2	2.0
					5.0	r1-1-0-1-1-1-1-0	2.0

Gm17334	r2-0-0-1-1-1-1-1	5.0	e0-0-1-0-1-1-1-0-0	2.0	r1-0-1-0-1-1-1-2	2.0
Gldc	r1-1-0-1-0-1-1-1	5.0	e0-0-1-0-1-0-1-0	2.0	r1-0-1-0-1-1-1-1	2.0
Fasl	r0-1-2-1-0-0-0-1	5.0	Sept8	2.0	r1-0-1-0-1-1-1-0	2.0
r2-1-0-1-0-1-0-1	r0-1-0-1-0-0-0-1	5.0	Rad50	2.0	r1-0-1-0-1-1-0-2	2.0
e2-1-0-1-1-1-0-1	r0-0-2-1-2-1-1-0	5.0	Pparg	2.0	r0-2-1-2-0-2-0-0	2.0
e2-1-0-1-0-1-0-1	r0-0-2-1-0-1-1-0	5.0	Klf3a	2.0	r0-2-1-0-0-0-1-0	2.0
e0-1-2-1-0-1-0-1	r0-0-2-1-0-1-0-1	5.0	Inpp4b	2.0	r0-0-1-0-0-0-0-0	2.0
e0-1-0-1-1-1-1	r0-0-2-1-0-0-0-1	5.0	Igfbp4	2.0	e2-2-1-0-0-0-0-0	2.0
e0-1-0-1-0-1-1-1	e2-1-2-1-0-0-0-1	5.0	Gldc	2.0	e2-0-1-0-1-0-0-0	2.0
e0-1-0-1-0-1-0-1	e1-1-0-1-0-0-0-0	5.0	Cxcr3	2.0	e1-0-1-0-1-2-0-0	2.0
e0-1-0-1-0-1-1-1	e1-0-0-1-1-0-1-1	5.0	r0-2-1-0-1-0-1-2	1.0	e1-0-1-0-1-1-2	2.0
e0-1-0-1-0-0-1-1	e1-0-0-1-1-0-0-0	5.0	r0-2-1-0-0-0-1-0	1.0	e1-0-1-0-1-1-0-2	2.0
e0-1-0-1-0-0-0-1	e1-0-0-1-1-0-0-0-1	5.0	r0-2-1-0-0-0-1-0	1.0	e1-0-1-0-0-1-1-0	2.0
e0-1-0-1-0-0-0-0	e0-1-2-1-0-0-0-1	5.0	e2-2-1-0-1-2-1-0	1.0	e0-2-1-2-0-2-0-0	2.0
e0-0-1-0-1-1-1	e0-1-0-1-1-1-1-0	5.0	e2-2-1-0-1-0-1-0	1.0	e0-1-0-1-1-0-1	2.0
e0-0-1-0-1-1-1	e0-0-2-1-2-1-1-1	5.0	e1-2-1-0-1-0-1-0	1.0	e0-1-0-0-0-0-0	2.0
e0-0-1-0-1-1-1	e0-0-2-1-2-1-1-0	5.0	e1-1-1-0-1-1-1-1	1.0	e0-0-1-0-1-0-2-0	2.0
e0-0-1-0-1-0-1-0	e0-0-2-1-0-1-0-1	5.0	e1-1-1-0-1-1-1-0	1.0	e0-0-1-0-1-0-2-0	2.0
e0-0-1-0-0-1-0	e0-0-2-1-0-0-0-0	5.0	e1-1-1-0-1-0-1-0	1.0	e0-0-1-0-1-0-0-1	2.0
e0-0-1-0-0-0-1	e0-0-2-1-0-0-0-0	5.0	e1-0-1-0-1-2-1-0	1.0	e0-0-1-0-0-0-2-0	2.0
Trem2	e0-0-0-1-1-0-1-0	2.0	e1-0-1-0-1-2-0-1	1.0	e0-0-1-0-0-0-1-1	2.0
Cxcr3	e0-0-0-1-0-1-1-1	2.0	e1-0-1-0-1-1-1-1	1.0	e0-0-1-0-0-1-1-0	2.0
Clec12a	r2-1-0-1-1-1-1	2.0	e1-0-1-0-1-1-1-2	1.0	e0-0-1-0-0-0-2-0	2.0
Bhlhe41	r0-2-0-1-0-1-1-0	2.0	e1-1-1-0-1-0-1-0	1.0	e0-0-1-0-1-0-0-2	2.0
r2-1-2-1-0-2-2-1	r0-1-0-1-0-1-1-1	1.0	e1-0-1-0-1-1-1-0	1.0	e0-0-1-0-0-0-1-1	2.0
r2-0-2-1-0-0-2-0	r0-0-0-1-0-0-0-0	1.0	e1-0-1-0-1-1-1-1	1.0	e0-0-1-0-0-0-0-2	2.0
r0-1-0-1-2-1-1-1	r0-0-0-1-1-1-0-1	1.0	e1-0-1-0-1-0-1-2	1.0	e0-0-1-0-0-0-0-1	2.0
r0-1-0-1-2-0-2-1	e1-1-0-1-1-1-0-1	1.0	e1-0-1-0-0-0-0-0	1.0	r2-2-1-0-1-2-1-0	1.0
r0-1-0-1-0-1-0-0	e1-1-0-1-0-1-0-0	1.0	e0-2-1-0-0-0-2-2	1.0	r2-2-1-0-1-2-0-0	1.0
r0-1-0-1-0-0-0-0	e1-1-0-1-0-0-1-1	1.0	e0-1-1-0-1-1-1-1	1.0	r2-2-1-0-1-1-1-0	1.0
e2-2-0-1-1-0-0-0	e1-0-0-1-1-1-0-0	1.0	e0-0-1-0-0-0-1-0	1.0	r2-2-1-0-0-2-0-2	1.0
e2-1-0-1-0-1-0-1	e1-0-0-1-0-0-1-0	1.0	e0-0-1-0-0-0-0-2	1.0	r2-0-1-0-1-0-0-0	1.0
e2-1-0-1-1-1-1-0	e1-0-0-1-0-0-1-0	1.0	Sell	1.0	r2-0-1-0-0-2-2-0	1.0
e2-1-0-1-1-1-1-0	e0-2-0-1-1-0-0-0	1.0	Gata3	1.0	r1-2-1-0-1-0-1-0	1.0
e2-1-0-1-0-1-1-1	e0-1-2-1-0-1-0-1	1.0	Cyp11a1	1.0	r1-2-1-0-1-0-0-1	1.0
e2-0-0-1-0-0-0-1	e0-1-0-1-1-0-1-0	1.0	r2-2-1-0-1-2-1-0	0	r1-2-1-0-1-0-0-0	1.0
e1-1-0-1-1-1-1	e0-1-0-1-0-1-0-0	1.0	r2-2-1-0-1-2-0-0	0	r1-2-1-0-0-0-0-0	1.0
e1-1-0-1-1-1-1	e0-0-2-1-0-0-1-0	1.0	r2-2-1-0-1-1-1-0	0	r1-0-1-0-1-2-0-0	1.0
e1-1-0-1-1-0-1-1	e0-0-2-1-0-0-0-1	1.0	r2-2-1-0-1-0-1-2	0	r1-0-1-0-1-0-0-0	1.0
e1-1-0-1-0-1-0-1	e0-0-0-1-1-1-0-1	1.0	r2-2-1-0-1-0-1-0	0	r0-2-1-0-1-2-0-2	1.0
e1-1-0-1-0-1-0-1	r0-2-0-1-1-0-0-0	3.0	r2-2-1-0-1-0-0-0	0	r0-2-1-0-1-2-0-0	1.0
e1-1-0-1-0-0-0-1	r0-2-0-1-0-0-1-0	3.0	r2-2-1-0-0-2-0-2	0	r0-2-1-0-1-1-2-0	1.0
e1-0-0-1-1-1-0-0	r0-1-0-1-0-0-0-0	3.0	r2-2-1-0-0-0-0-2	0	r0-2-1-0-1-1-0-0	1.0
e1-0-0-1-0-1-0-0	r0-0-2-1-0-0-0-0	3.0	r2-2-1-0-0-0-0-0	0	r0-2-1-0-1-0-1-2	1.0
e0-2-0-1-1-0-0-0	e2-2-0-1-0-0-0-0	3.0	r2-0-1-0-1-0-0-0	0	r0-2-1-0-1-0-1-0	1.0
e0-1-0-1-1-1-0-1	e2-2-0-1-0-0-0-0	3.0	r2-0-1-0-1-0-0-2	0	r0-2-1-0-1-0-1-0	1.0
e0-1-0-1-1-1-0-1	e2-0-0-0-1-1-1-1	1.0	r2-0-1-0-0-2-2-0	0	r0-2-1-0-1-0-0-2	1.0

e0-1-0-1-1-0-1-1	1.0	e2-0-0-1-0-0-0-0	r1-2-1-0-1-0-1-0	0	r0-2-1-0-0-2-0-2	1.0
e0-1-0-1-1-0-0-1	1.0	e1-1-0-1-1-0-1-1	r1-2-1-0-1-0-0-1	0	r0-2-1-0-0-2-0-0	1.0
e0-0-2-1-0-1-0-1	1.0	e1-1-0-1-0-0-0-1	r1-2-1-0-1-0-0-0	0	r0-2-1-0-0-2-2	1.0
e0-0-2-1-0-1-0-0	1.0	e1-0-0-1-0-1-0-0	r1-2-1-0-0-0-0-0	0	r0-2-1-0-0-2-0	1.0
e0-0-2-1-0-0-0-1	1.0	e0-0-2-1-0-1-1-0	r1-1-1-0-1-1-1-0	0	r0-2-1-0-0-0-2	1.0
e0-0-0-1-2-0-0-1	1.0	e0-0-1-2-0-0-1	r1-0-1-0-1-2-0-0	0	r0-0-1-2-0-0-0-0	1.0
e0-0-0-1-1-1-1-1	1.0	e0-0-0-1-0-0-2-0	r1-0-1-0-1-1-1-2	0	r0-0-1-0-2-0-0-0	1.0
e0-0-0-1-1-1-0-0	1.0	e0-0-0-1-0-0-1-1	r1-0-1-0-1-1-1-1	0	r0-0-1-0-1-2-0-0	1.0
e0-0-0-1-1-0-1-0	1.0	Tbet	r1-0-1-0-1-1-1-0	0	r0-0-1-0-0-2-0	1.0
e0-0-0-1-0-0-1-1	1.0	r2-2-0-1-0-0-0-0	r1-0-1-0-1-1-0-2	0	r0-0-1-0-0-0-2	1.0
Klrg1	1.0	r0-1-0-1-0-1-0-0	r1-0-1-0-1-0-0-0	0	e2-2-1-2-1-0-0-2	1.0
Il2	1.0	r0-0-0-1-0-1-0-0	r0-2-1-2-0-2-0-0	0	e2-2-1-2-1-0-0-0	1.0
Igfbp4	1.0	e2-1-0-1-2-0-0-1	r0-2-1-0-1-2-0-2	0	e2-2-1-2-0-2-0-2	1.0
Chil3	1.0	e2-1-0-1-1-1-1-1	r0-2-1-0-1-2-0-0	0	e2-2-1-0-1-2-1-0	1.0
Adamts3	1.0	e2-1-0-1-1-1-1-0	r0-2-1-0-1-1-2-0	0	e2-2-1-0-1-2-0-0	1.0

Table C.26: In- and out-degree ranking with respective values for the *Diff1* network with respective up- and down-regulated parts.

Full network ranking	PageRank full	Th1 network ranking	PageRank Th1	Th1 / 2 network ranking	PageRank Th1 / 2
Il10	0.0259491509026	Ccl5	0.0567528565681	Il10	0.0538996151769
Lrrc2	0.0241101038277	Ilfng	0.0510594680293	Ccl5	0.0319097268816
Ilfng	0.023979040484	Smpdl3b	0.0351152252845	Cyp11a1	0.0240529900241
Ilfng	0.0231336756512	Tbet	0.0317014382147	Gata3	0.0226487105771
Cyp11a1	0.021452768389	Fasl	0.026317571311	Tbet	0.0208941366979
Ccl5	0.0180633191816	Il18r1	0.021712261874	Lrrc32	0.019843957905
Runx3	0.0179848074786	Il18rap	0.0173557104185	Ilfng	0.0172304160158
Tbet	0.0169248746008	Itgal1	0.0165511375026	Inpp4b	0.0168129370795
Cd83	0.0164415884512	Galnt3	0.0128761346931	Fasl	0.0154375237248
Smpdl3b	0.0143144779241	STAT6	0.0112222605202	Ilfng2	0.0146777918425
Gata3	0.0133295856853	Il2	0.0107961121234	Itga1	0.0131566281411
Pparg	0.0130490351385	Clec12a	0.0105840902615	Smpdl3b	0.0127309265958
Inpp4b	0.0124054848706	Klrc2	0.0102511747316	Asb2	0.011713883701
Igfbp4	0.0112339945144	Klrc2	0.00998548183466	Il5	0.0114143616462
Fasl	0.0105834128509	Klrc1	0.00975669187164	Il4	0.0113921592802
Sell	0.00966100179939	Il12rb2	0.00941120080158	Sept8	0.0110619367399
Il18r1	0.00895394067884	Klrb1c	0.00906933492828	Pparg	0.0106323831357
Il4	0.00858984724626	Exph5	0.00874940571603	STAT6	0.0100268205308
Bhlhe41	0.00843308266836	Asb2	0.00801830962325	Il13	0.0094029796984
Sept8	0.00808599537834	Runx3	0.00742701852956	Il18r1	0.00873159470159
STAT6	0.00802226467661	Cd83	0.0073689848346	Il12rb2	0.008247373583378
Asb2	0.00798029167401	Klrg1	0.00735618469555	Klrc2	0.00749311005036
Il5	0.00791366426097	Klrc1	0.00727806068384	Klrc2	0.00744828602576
Il9r	0.00761532639503	Kcnj8	0.00710595132933	Rad50	0.0068986676101
Il13	0.00719287086078	Ilfng	0.00703279469179	Eomes	0.00670516203025
St8sia6	0.00692641570587	Igfbp4	0.00559945673911	Il18rap	0.00647428467093
Kcnj8	0.00692579862137	e0-0-1-0-1-0-0-0	0.00520507925795	Klrc1	0.00647361678882
Il18rap	0.00683780321894	e1-1-1-1-1-1-1-1	0.00514939990038	Cd83	0.00587125244032
Galnt3	0.00654311927257	e0-1-1-1-1-1-1-1	0.00492234468149	Clec12a	0.0054230701592
Itga1	0.00615804126879	Il10	0.00477346872919	Klrc1	0.0049988148606
Rad50	0.00603337540803	Il1rl1	0.0046546138402	e1-1-1-1-1-1-1-1	0.00460404102243
Klrc2	0.00527044159665	Il9r	0.00448548567158	Runx3	0.00436148235051
Il2	0.00474345037679	Sept8	0.00444346202421	Gldc	0.00426546114759
Klrc2	0.00460288651239	Cxcr3	0.00439598551719	Kcnj8	0.00422085919247
Ilfng	0.00435941886449	Inpp4b	0.00437026455407	e0-1-0-1-1-1-1-1	0.00421713627827
Klrc1	0.0043288877684	STAT4	0.00415367647912	Sell	0.00412757998393
Clec12a	0.00427370727988	e1-1-1-1-1-1-1-0	0.00411128828553	Klrg1	0.0040797596706
Chil3	0.00424641606969	Eomes	0.00382167969562	Il9r	0.00404618787258
Il1rl1	0.00397186721097	St8sia6	0.00381684128517	Klrb1c	0.00401141369674
Exph5	0.0039545776648	Pparg	0.00381664968201	Cxcr3	0.0038632066816
Klrc1	0.00392491384978	e0-0-1-1-1-1-1-1	0.00361483841772	Ccr4	0.00384818293015
Ccr4	0.00391082357797	STAT1	0.003575848260175	e0-0-0-1-0-0-0-0	0.00376188977326
Cxcr3	0.00384564198757	e0-0-1-1-1-1-0-0-0	0.0033552526168	e0-1-0-1-0-1-0-1	0.00372775879596
Klrg1	0.00356465997564	e1-0-1-1-1-1-1-1	0.00331552526168	St8sia6	0.00372128763273

e0-0-0-0-0-1-0	0.000677248028194	e2-2-1-1-1-1-1-0	0.00171041841095	e1-2-1-1-1-1-1-1	0.0016823306519
e0-2-0-0-0-0-0	0.000675225980951	e2-2-1-1-0-0-0-0	0.00163519154366	Exp5	0.00166570028632
e0-0-1-1-1-1-1	0.000670718747847	Gml2214	0.00162631677942	e1-1-0-1-0-1-0-1	0.00164403734241
e0-0-1-1-1-0-1-0	0.000658274856448	e0-1-1-1-1-0-1-0	0.00161095767323	e2-1-0-1-0-1-1-1	0.00162182805225
e0-0-1-1-1-1-0	0.000650226609626	e2-2-1-1-1-0-1-0	0.00160608164351	e0-0-0-1-0-1-1-1	0.00157253953219
e1-1-1-1-1-1-0	0.000646923269415	e0-0-1-0-0-0-1-0	0.00160145159912	e1-1-0-1-1-1-0-1	0.00153678950172
r0-0-2-2-2-0-2-0	0.000645277441872	e2-0-1-0-0-0-0-0	0.00160141309004	e1-0-0-1-1-1-0-1-1	0.00152738665758
Efnra5	0.000644433278588	e0-2-1-0-0-0-0-0	0.00159678304565	e0-0-1-1-1-1-0-1	0.00150924950766
e0-0-0-0-1-1-1	0.000631348342918	e1-2-1-1-1-1-1-0	0.00159674453657	e0-0-0-1-1-1-1-0	0.00150751236032
Ccr1	0.000622529891055	Ccr1	0.00158991330476	r2-1-0-1-0-1-0-1	0.00149513532006
e2-0-1-1-1-1-0	0.00060606624479	r0-2-1-0-1-0-1-0	0.00158881204145	e0-1-0-1-1-1-0-1	0.00149427022646
e2-0-0-0-0-0-0	0.000605410186744	e0-2-1-0-0-2-0-2	0.00158824085891	e0-1-0-1-1-0-0-1	0.00148073109845
e0-1-0-1-0-0-1	0.000604895619213	e1-0-1-1-1-1-0-0	0.0015734402783	e2-0-0-1-0-0-0-1	0.00147920670482
e0-0-0-1-0-0-0	0.000603193053831	e1-0-1-1-0-0-1-1	0.00157144139755	Trem2	0.00147920670482
e1-1-1-1-0-1-0	0.000602643706386	e1-0-1-0-0-0-0-0	0.00156410317137	e0-0-1-1-0-0-1-0	0.00146811386739
e0-0-1-0-0-1-0	0.000590540046837	e1-0-1-1-1-0-0-1	0.00155629654213	e1-0-1-1-1-0-1-0	0.00144683279722
e0-0-1-1-1-1-1	0.000590124995103	e0-0-1-1-1-0-0-1	0.00155629654213	e1-0-1-1-0-1-1-1	0.00144199989897
e0-0-1-0-1-1-1-0	0.000587750076272	e1-0-1-1-1-0-0-1	0.00155086690885	e0-2-1-1-1-1-0-0	0.00143390989864
e0-0-1-1-1-0-0	0.00058725343794	e0-2-1-0-1-0-1-0	0.00153607253561	e1-0-0-1-1-1-0-0	0.00142629503454
e1-1-1-1-0-1-1	0.00058397671137	e1-2-1-1-1-1-0-0	0.00153607253561	e0-2-1-0-1-0-0	0.00142629503454
e0-1-0-0-1-0-0	0.000577734913364	r0-2-1-1-1-1-0-0	0.00153471640414	e1-1-0-1-0-1-0-0	0.00140643490683
e0-1-0-1-0-0-0	0.000577734913364	Chil3	0.00153256675312	e0-1-1-1-1-1-1-1	0.00139650371761
e0-0-0-0-0-1-1	0.000577734913364	e1-1-0-1-0-1-0-1	0.00153142329708	e2-1-1-1-1-1-1-1	0.00139650371761
e0-0-1-0-0-0-1	0.000577734913364	e2-0-1-1-0-1-1-0	0.00153140398214	e2-0-1-1-1-0-0-0	0.0013942013984
e0-0-1-1-0-1-0	0.00057700692213	r1-0-1-1-1-0-1-0	0.00152743798245	e2-2-1-1-1-1-1-0	0.00137539492215
e0-1-0-1-0-0-1-0	0.000573206768717	e0-0-1-1-1-1-2-0	0.00152715214331	e0-0-2-1-0-0-0-1	0.00137185897064
e0-0-0-1-0-1-1	0.000572933831219	e2-0-1-1-1-0-1-1	0.00152715214331	e1-1-0-1-0-0-0-1	0.00137185897064
e1-2-1-1-1-1-1	0.000569056352684	e0-2-1-0-1-2-0-2	0.00152715214331	e0-1-2-1-0-1-0-1	0.00134802636723
e0-0-0-1-0-1-0-1	0.000563269874368	e0-2-1-0-1-0-0-2	0.00152715214331	e0-0-0-1-2-0-0-1	0.00132419376383
r0-0-2-2-0-0-0	0.000561987671048	e0-2-1-0-1-1-2-0	0.00152715214331	e0-0-0-1-0-0-1-1	0.00131742388061
Slc4a4	0.000553788934456	Sell	0.00152535183655	Galnt3	0.00131161558735
:	:	:	:	e2-2-1-1-0-0-0-0	0.00130830344065
:	:	:	:	:	:
:	:	:	:	:	:

Table C.27: PageRank centrality ranking for the full network as well as for the Tbet⁺/⁺Th1 and Tbet⁺/⁺Th1/2 networks.

Diff1 up ranking	PageRank Diff1 up	Diff1 down ranking	PageRank Diff1 down	Diff2 up ranking	PageRank Diff2 up	Diff2 down ranking	PageRank Diff2 down
II10	0.0796565048134	Infng	0.093881693684	Ccl5	0.0486549795608	Lrrc32	0.0688490251087
Cyp11a1	0.0354453928649	Smpdl3b	0.0431152480282	Tbet	0.0297233387744	Sell	0.0387036594763
Gata3	0.0328741542683	Ccl5	0.0368892528982	Infng	0.0280641169349	Cyp11a1	0.0359158765661
Lrrc32	0.0279333779893	Galnt3	0.0269485461611	II10	0.0252098238197	Bhlhe41	0.0312399963915
Infng2	0.023720699683	Tbet	0.0251533078256	Gata3	0.025782072135	Cd83	0.0230517238209
Inpp4b	0.0231486970034	II18r1	0.0235198437952	Fasl	0.020646358035	Infng2	0.0209779628262
II5	0.0176236743949	III8rap	0.0220657492392	Cyp11a1	0.0170583237651	II9r	0.0209051995209
II4	0.0171648130297	II2	0.0178564337555	Igfa1	0.0169859927764	Pparg	0.0203788020273
II13	0.0163530946577	Fasl	0.0168046026238	Asb2	0.0145622512266	Igfbp4	0.0178164724463
Sept8	0.016270270565	Expn5	0.0141972343526	Inpp4b	0.0143973598764	Chil3	0.017049355347
Asb2	0.0135047825534	Kcnj8	0.0139119146653	Smpdl3b	0.0139847857778	II10	0.0163997053088
Pparg	0.0134206539407	r0-2-1-0-1-0-1-0	0.0126152064093	Infng2	0.0124532757126	Ccr4	0.0132178854995
Rad50	0.0133316050216	Igfa1	0.011771401507	Sept8	0.010915545408	Smpdl3b	0.0128576873149
Infng	0.0117040477774	Cd83	0.0116686487023	STAT6	0.0106567118369	St8sia6	0.0108669678435
Tbet	0.0104985414835	STAT6	0.011528564784	II18r1	0.0102511643341	Gata3	0.010829129533
STAT6	0.00942129528254	Klrb1c	0.010475270389	Klrc2	0.00959642565137	II1r1	0.0105475318875
Cd83	0.0078513643193	Clecl2a	0.00967796484074	Klrc2	0.00938537599291	Rumx3	0.00886457628924
Eomes	0.00756349366573	Rumx3	0.00901440927813	III8rap	0.00918743530476	II13	0.00800317201946
Sell	0.00736572279393	II1r1	0.00879428243428	Klrc1	0.00839923729288	II4	0.008000317201946
Igfa1	0.00733688169674	Asb2	0.00765732461368	II5	0.0081360345156	Sept8	0.00727768747576
e0-0-0-1-0-0-0-0	0.00664657667126	Igfbp4	0.00759832411186	III3	0.00792378766889	II5	0.00723103080908
Ccr4	0.0065095856026	STAT1	0.00697936325912	Pparg	0.00778767569546	Kcnj8	0.00721105680266
Kcnj8	0.00598410218288	Klrg1	0.00640225046227	II4	0.00768504400453	Rad50	0.00720619291925
Gldc	0.00540969183685	Klre1	0.00633106310376	III2rb2	0.00763486844411	STAT6	0.00616319414956
r2-1-0-1-0-1-0-1	0.00508654406896	Klrc2	0.00633106310376	Cd83	0.00755439709472	Asb2	0.00548157477488
e0-1-0-1-0-0-1-0	0.00508654406896	Klrc2	0.00606473951021	Klre1	0.00720937542535	II18r1	0.00539694433878
Smpdl3b	0.00469183425558	Klrc1	0.00585267977479	Rad50	0.00710012802796	Clecl2a	0.00537830478282
Klrc2	0.00467676581325	e0-0-1-0-0-0-0-0	0.00584752383469	Cxcr3	0.00693577370492	Inpp4b	0.00439900002713
Kif3a	0.00447584402278	III2rb2	0.00520640746704	Gldc	0.00650264073503	Tbet	0.00430604493503
Klrc2	0.0044509044807	Sept8	0.00520640746704	Klrg1	0.00611159944869	II2	0.00429394034496
Irf1	0.00413050473523	Kif3a	0.00520640746704	Kcnj8	0.0060470682867	III8rap	0.00411004232149
Klre1	0.00403039509475	e0-0-1-0-1-0-0-0	0.00488074930749	Clecl2a	0.00572857510829	Adamts13	0.00404764322782
Ccl5	0.00398096297391	e0-2-1-0-1-0-1-0	0.00488074930749	Eomes	0.00555652347844	Ccr2	0.00343600566677
III2rb2	0.00394682510669	e2-2-1-0-1-0-1-2	0.00488074930749	Sell	0.005325353517318	A430108G06Rik	0.00311411343959
III8rap	0.00391961452149	e2-0-1-0-0-0-0-0	0.00488074930749	Igfbp4	0.004972553689403	e1-1-0-0-0-0-1-0	0.00307153185322
St8sia6	0.0038037181092	e0-0-1-0-1-1-1-0	0.00488074930749	Rumx3	0.00492360341319	Klrb1c	0.00301743366285
II9r	0.00376146946557	e0-2-1-0-1-0-0-0	0.00488074930749	e0-0-0-1-0-0-0-0	0.0048114481913	r0-1-0-0-0-0-0-1	0.00297704190521
III8r1	0.00371042538296	II9r	0.00460842830583	II9r	0.00480176236753	Dpysl3	0.00294417062365
STAT1	0.00370503820135	e1-0-1-0-1-0-1-0	0.00452174379581	Klrb1c	0.00470692242147	II2	0.00292273250963
e0-0-0-1-1-1-1-0	0.00367137251956	e0-2-1-0-1-2-0-2	0.00452174379581	II2	0.00450638381463	e0-1-0-0-0-0-1-1	0.00292273250963
Klrc1	0.00364228077576	e0-0-1-0-1-1-0-0	0.00452174379581	Lrrc32	0.00401965913011	e0-1-0-0-0-1-0-0	0.00292273250963
Rumx3	0.003509287058	e0-2-1-0-0-0-1-0	0.00452174379581	e0-0-0-1-1-1-1-0	0.00390887438667	e0-1-0-0-0-1-0-1	0.00292273250963
Gjal	0.0035048282717	e0-2-1-0-0-2-0-2	0.00452174379581	STAT4	0.00387678926748	e0-1-2-0-2-0-2-1	0.00292273250963
A430108G06Rik	0.00343173820902	e0-2-1-0-1-0-0-2	0.00452174379581	e0-0-1-1-1-1-1-1	0.00380588614539	e0-1-0-0-0-0-0-0	0.00292273250963

e0-2-0-1-1-1-0-0	0.00208841400291	e2-0-1-0-0-2-2-0	Irf1	0.00257882798144	e1-1-1-0-1-1-1-1	0.0021670184701
r2-1-2-1-0-2-2-1	0.00208841400291	e0-0-1-0-1-1-0-2	STAT1	0.00256570076868	r0-1-2-0-0-0-0	0.0021670184701
e0-0-0-1-1-1-0-0	0.00208841400291	e0-0-1-0-0-1-0-0	r0-0-1-1-1-1-0	0.00256428243934	e1-1-1-0-1-0-1-0	0.0021670184701
e0-1-0-1-1-0-1-1	0.00208841400291	e0-0-1-0-0-1-0-1	e0-0-0-1-2-0-0-1	0.00256428243934	Klri2	0.00209837786168
r2-0-2-1-0-0-2-0	0.00208841400291	r1-0-1-0-1-1-1-2	e1-0-0-1-0-1-0-0	0.00256428243934	Klrc1	0.00209837786168
e2-2-0-1-1-0-0-0	0.00208841400291	r2-2-1-0-1-1-1-0	r1-0-1-1-1-0-1-0	0.00256428243934	Gja1	0.00209837786168
r0-1-0-1-0-0-0-0	0.00208841400291	r0-2-1-0-1-1-2-0	Ccr2	0.00254554513684	Klrc2	0.00209837786168
e2-2-0-1-1-0-1-0	0.00208841400291	e0-0-1-0-1-0-0-1	Trem12	0.00249759642787	Efna5	0.00209837786168
e0-0-0-1-1-0-1-0	0.00208841400291	e0-0-1-0-1-0-0-2	Ccr1	0.00244078045999	Slc4a4	0.00209837786168
r0-1-0-1-1-0-1-0-0	0.00208841400291	e1-0-1-0-1-2-0-0	Ccr4	0.00240341521827	e2-1-2-0-0-1-0-0	0.00201821912651
Il2	0.0020794990036	e0-1-1-0-1-0-1-1	A430108C06Rik	0.00228225026043	e0-1-2-0-0-0-1-0	0.00201821912651
Igfbp4	0.00204763210622	e2-2-1-0-0-2-0-2	Galnt3	0.00225012759943	e0-1-2-0-0-0-1-1	0.00201821912651
Klrg1	0.00203514587957	e2-0-1-0-0-0-2-0	Exp5	0.00209727218373	r2-1-2-0-2-0-1-1	0.00201821912651
Adamts13	0.00199184712635	e0-2-1-0-0-0-2-0	Ccr5	0.00206999495467	r1-1-0-0-1-0-1-1	0.00201821912651
Chil3	0.00198208886839	r1-1-1-0-1-1-1-0	Chil3	0.00198714422211	e0-1-0-0-1-0-1-1	0.00201821912651
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:

Table C.28: PageRank centrality ranking for the up- and down-regulated parts of *Diff1* and *Diff2*.

λ_1 ranking	PageRank	λ_2 ranking	λ_2 values
Gata3	0.741782	Tbet	0.586051361808
Tbet	0.626507	e0-0-1-0-1-0-0-0	0.123258669886
e0-0-0-0-0-0-0-0	0.148212	e1-0-0-0-0-0-0-0	0.105058289443
e1-0-0-0-0-0-0-0	0.0747871	r0-0-2-0-0-0-2-0	0.0942825135056
e0-0-1-0-1-0-0-0	0.0659869	e1-1-1-1-1-1-1-1	0.0739448658361
r0-0-2-0-0-0-2-0	0.0510942	e0-0-1-0-0-0-0-0	0.0458365610444
e1-1-1-1-1-1-1-1	0.0438635	e0-0-1-1-1-0-0-0	0.0404420885126
e0-1-0-1-1-1-1-1	0.0417517	e1-0-1-1-1-1-1-1	0.0378698944021
e0-1-0-0-0-0-0-0	0.0401505	e0-0-0-1-0-0-0-0	0.031347890041
e0-0-0-1-0-0-0-0	0.0314694	e1-0-1-1-1-1-1-0	0.023358377738
e0-0-0-0-0-0-0-1	0.0309608	e0-0-1-1-0-0-0-0	0.0223391609167
e0-0-1-0-0-0-0-0	0.0307572	e1-0-1-1-1-0-0-0	0.0222313274697
e0-1-0-1-0-1-0-1	0.0294292	e0-0-1-1-0-1-0-0	0.0205364842889
e0-1-0-0-0-0-0-1	0.0285103	e2-0-1-1-1-1-1-0	0.0174858049872
e0-1-1-1-1-1-1-1	0.0253448	e0-0-1-1-1-1-1-0	0.0158431357395
r0-2-0-0-0-0-0-0	0.0249659	e0-0-1-1-1-0-1-0	0.0152875286378
e0-0-2-0-0-0-0-0	0.0244514	e1-2-1-1-1-1-1-1	0.0148164537369
e0-0-1-1-1-0-0-0	0.023422	e1-1-1-1-1-0-1-0	0.0148164537369
e1-0-1-1-1-1-1-1	0.0221	e0-2-0-0-0-0-0-0	0.0137529227493
r0-0-2-2-2-0-2-0	0.0179525	e0-0-1-0-1-1-1-0	0.0136724421374
e0-0-0-0-0-1-1-1	0.0171762	e0-1-0-1-0-0-1-0	0.0134817781117
e0-0-0-1-1-1-1-1	0.0148788	e0-0-1-1-0-1-1-0	0.0128144402991
e1-0-1-1-1-1-1-0	0.0146416	r1-1-0-0-0-0-0-0	0.0121789118439
e0-0-0-0-0-0-1-0	0.0143682	e1-2-0-0-0-0-0-0	0.0121471024865
e0-0-1-1-0-0-0-0	0.0141178	e1-0-1-0-1-0-1-0	0.0121471024865
e1-0-1-1-1-0-0-0	0.0140623	e0-0-1-1-1-1-0-0	0.0117657744352
e2-1-0-1-0-1-0-1	0.0135003	e0-0-1-0-1-1-0-0	0.0112573187364
e0-0-1-1-0-1-0-0	0.0131912	e1-1-1-1-1-0-1-1	0.00979505629973
e0-0-1-1-1-1-1-1	0.013177	r2-1-0-1-0-1-0-1	0.00950160139285
e1-1-0-1-1-1-1-1	0.0125813	e0-0-1-1-1-1-0-1	0.00947777868171
e0-0-0-0-0-1-0-0	0.0125813	e0-0-0-0-1-0-1-0	0.0078761679315
e0-2-0-0-0-0-0-0	0.0125234	e1-0-1-1-0-1-1-1	0.00769821118148
e2-1-2-1-0-1-0-1	0.0118155	e1-0-1-1-1-0-1-1	0.00747576524395
e0-1-0-1-1-0-0-1	0.0118155	e0-2-1-0-1-0-0-0	0.0074016074496
e0-0-0-1-0-0-1-0	0.0118155	e2-2-1-1-1-1-1-0	0.00680842743137
e2-0-1-1-1-1-1-0	0.0116233	e0-2-1-0-0-0-1-0	0.00680842743137
e0-1-0-1-0-0-1-1	0.0115602	e2-0-1-1-1-0-0-0	0.0062745571813
r0-0-2-2-2-0-0-0	0.0115286	e2-2-1-1-0-0-0-0	0.00502885993114
e1-1-1-1-1-1-0-1	0.0112029	e2-2-1-1-1-0-1-0	0.00413907343647
e0-1-0-1-0-1-1-1	0.0110497	e2-0-1-0-0-0-0-0	0.00413907343647
e0-0-2-1-0-1-0-1	0.0110497	e1-2-1-1-1-1-1-0	0.00413907343647
e0-0-0-1-0-1-0-1	0.0108582	e0-2-1-0-0-2-0-2	0.00413907343647
e0-1-0-1-0-0-0-1	0.0107944	r0-1-0-0-0-0-0-0	0.00384954241503
e0-0-1-1-1-1-1-0	0.010779	e1-0-1-1-1-0-0-1	0.00324928968635
e1-1-1-1-1-1-1-0	0.0106668	e0-0-1-1-1-0-0-1	0.00324928968635
e0-0-1-1-1-0-1-0	0.0104935	e2-0-1-1-1-0-1-1	0.0028043978113
e0-0-0-0-1-0-0-0	0.0104371	e2-0-1-1-0-1-1-0	0.0028043978113
e1-1-0-1-0-1-0-1	0.0102839	e1-2-1-1-1-1-0-0	0.0028043978113
e1-2-1-1-1-1-1-1	0.0102513	e1-1-0-0-1-1-1-0	0.0028043978113
e1-1-1-1-1-0-1-0	0.0102513	e0-2-1-0-1-2-0-2	0.0028043978113
e1-1-1-1-1-0-1-1	0.00973775	e0-0-0-1-0-0-1-1	-0.00334207424189
e0-0-1-0-1-1-1-0	0.00966336	e1-1-0-0-0-0-1-0	-0.00374305989753
e0-1-0-1-0-0-1-0	0.00956537	e1-1-0-1-0-1-0-0	-0.00387672178275
e2-0-0-0-0-0-0-0	0.00951808	e0-1-0-0-0-1-0-1	-0.0044877270472
e0-0-1-1-0-1-1-0	0.00922238	e1-1-0-0-0-0-1-1	-0.00454500251369
r1-1-0-0-0-0-0-0	0.00889574	e2-1-2-0-2-0-2-1	-0.00494598816933
e1-2-0-0-0-0-0-0	0.00887939	e1-1-0-1-0-0-0-1	-0.00494598816933
e1-0-1-0-1-0-1-0	0.00887939	e1-1-0-0-1-0-1-0	-0.00494598816933
e0-0-0-0-1-0-1-1	0.00883578	e0-0-2-1-0-0-0-1	-0.00494598816933
e0-1-0-1-0-0-0-0	0.00875226	e0-0-2-0-0-1-1-0	-0.00494598816933
e0-1-0-0-0-1-0-0	0.00875226	e1-1-0-0-0-1-0-0	-0.00654993079189
e0-0-0-1-0-0-0-1	0.00875226	e1-0-0-1-1-1-0-0	-0.00654993079189
e0-0-0-0-0-0-1-1	0.00875226	e0-0-2-1-0-1-0-0	-0.00654993079189
e0-0-1-1-1-1-0-0	0.0086834	e1-0-0-0-1-0-0-0	-0.0076192258736
e0-0-1-0-1-1-0-0	0.00842207	e2-0-0-1-0-0-0-1	-0.00815387341446
e2-0-0-1-0-0-0-1	0.00798646	e0-1-0-1-0-0-0-0	-0.0097577873419
e1-0-0-0-1-0-0-0	0.00773118	e0-1-0-0-0-1-0-0	-0.0097577873419
r2-1-0-1-0-1-0-1	0.00751969	e0-0-0-1-0-0-0-1	-0.0097577873419
e0-0-1-1-1-1-0-1	0.00750745	e0-0-0-0-0-0-1-1	-0.0097577873419

e1-1-0-0-0-1-0-0	0.00722064	e2-0-0-0-0-0-0-0	-0.0113617299645
e1-0-0-1-1-1-0-0	0.00722064	e1-1-0-1-0-1-0-1	-0.0129656438919
e0-0-2-1-0-1-0-0	0.00722064	e0-0-0-0-1-0-0-0	-0.0132864553725
e0-0-0-0-1-0-1-0	0.00668428	e1-1-1-1-1-1-1-0	-0.0137676152032
e1-0-1-1-0-1-1-1	0.00659281	e0-1-0-1-0-0-0-1	-0.0140349389736
e1-0-1-1-1-0-1-1	0.00647848	e0-0-0-1-0-1-0-1	-0.0141686008588
e2-1-2-0-2-0-2-1	0.00645482	e0-1-0-1-0-1-1-1	-0.0145695865145
e1-1-0-1-0-0-0-1	0.00645482	e0-0-2-1-0-1-0-1	-0.0145695865145
e1-1-0-0-1-0-1-0	0.00645482	e1-1-1-1-1-1-0-1	-0.0148903693
e0-0-2-1-0-0-0-1	0.00645482	r0-0-2-2-2-0-0-0	-0.0155725671658
e0-0-2-0-0-1-1-0	0.00645482	e0-1-0-1-0-0-1-1	-0.0156388815962
e0-2-1-0-1-0-0-0	0.00644037	e2-1-2-1-0-1-0-1	-0.016173529137
e1-1-0-0-0-0-1-1	0.00626337	e0-1-0-1-1-0-0-1	-0.016173529137
e0-1-0-0-0-1-0-1	0.00623602	e0-0-0-1-0-0-1-0	-0.016173529137
e2-2-1-1-1-1-1-0	0.00613549	e1-1-0-1-1-1-1-1	-0.0177774430645
e0-2-1-0-0-0-1-0	0.00613549	e0-0-0-0-0-1-0-0	-0.0177774430645
e1-1-0-1-0-1-0-0	0.00594429	e0-0-1-1-1-1-1-1	-0.0190249635582
e1-1-0-0-0-0-1-0	0.00588047	e2-1-0-1-0-1-0-1	-0.0197021684725
e2-0-1-1-1-0-0-0	0.0058611	e0-0-0-0-0-0-1-0	-0.0215199471553
e0-0-0-1-0-1-1-1	0.00568902	e0-0-0-1-1-1-1-1	-0.022589242237
e0-0-0-1-0-0-1-1	0.00568902	e0-0-0-0-0-1-1-1	-0.0274010127145
e2-1-1-1-1-1-1-1	0.00530611	r0-0-2-2-2-0-2-0	-0.0290268784114
e1-1-0-0-1-0-0-1	0.00530611	e0-0-2-0-0-0-0-0	-0.0426384102386
e0-1-1-1-0-1-1-1	0.00530611	r0-2-0-0-0-0-0-0	-0.0437159121256
e2-2-1-1-0-0-0-0	0.00522086	e0-1-1-1-1-1-1-1	-0.0445096192413
e1-0-1-1-0-0-0-0	0.0049886	e0-1-0-0-0-0-0-1	-0.0511393405723
r0-2-1-1-1-1-0-0	0.00497502	e0-1-0-1-0-1-0-1	-0.0530639225048
e0-1-2-1-0-1-0-1	0.0049232	e0-0-0-0-0-0-0-1	-0.0562717503596
e0-1-2-0-2-1-0-1	0.0049232	e0-1-0-0-0-0-0-0	-0.0755190044402
e0-1-2-0-2-0-2-1	0.0049232	e0-1-0-1-1-1-1-1	-0.0788726035224
e2-2-1-1-1-0-1-0	0.00476354	e0-0-0-0-0-0-0-0	-0.1094441511187
e2-0-1-0-0-0-0-0	0.00476354	Gata3	-0.746936958252

Table C.29: Eigenvectors λ_1 and λ_2 for the Tbet/Gata3 subnetwork with leading ranked CSCs for λ_1 and CSCs with largest positive and negative λ_2 component.

λ_1 ranking	PageRank	λ_2 ranking	λ_2 values	λ_3 ranking	λ_3 values
1110011001	0.59243947212	1110111001	0.418879192305	1110011001	0.750503915711
1111011001	0.268173910002	1111111001	0.196030422849	1110111001	0.216237858952
1110011011	0.268173910002	1110111011	0.196030422849	1110011101	0.180611319771
1110011000	0.268173910002	1110111000	0.196030422849	1110001001	0.180611319771
1110010001	0.268173910002	1110110001	0.196030422849	0110011001	0.180611319771
1010011001	0.268173910002	1010111001	0.196030422849	1110010001	0.153861896076
1110111001	0.136999195027	1111111011	0.0886012490474	1010011001	0.153861881124
1111011011	0.116301236571	1111111000	0.0886012490474	1110011000	0.153861874981
1111011000	0.116301236571	1111110001	0.0886012490474	1110011011	0.153861871375
1111010001	0.116301236571	1110111010	0.0886012490474	1111011001	0.153861868383
1110011010	0.116301236571	1110110011	0.0886012490474	1110111101	0.0496183650167
1110010011	0.116301236571	1110110000	0.0886012490474	1110101001	0.0496183650167
1110010000	0.116301236571	1011111001	0.0886012490474	0110111001	0.0496183650167
1011011001	0.116301236571	1010111011	0.0886012490474	1110110001	0.0432530518509
1010011011	0.116301236571	1010111000	0.0886012490474	1010111001	0.0432530476478
1010011000	0.116301236571	1010110001	0.0886012490474	1110111000	0.0432530459208
1010010001	0.116301236571	1110111101	0.0806985474236	1110111011	0.0432530449072
1110011101	0.108193225209	1110101001	0.0806985474236	1111111001	0.0432530440659
1110001001	0.108193225209	0110111001	0.0806985474236	1110001101	0.036800171604
0110011001	0.108193225209	1111111010	0.0390699989398	0110011101	0.036800171604
1111111001	0.0605069371764	1111110011	0.0390699989398	0110001001	0.036800171604
1110111011	0.0605069371764	1111110000	0.0390699989398	1110010101	0.0340582325458
1110111000	0.0605069371764	1110110010	0.0390699989398	1110000001	0.0340582325458
1110110001	0.0605069371764	1011111011	0.0390699989398	0110010001	0.0340582325458
1010111001	0.0605069371764	1011111000	0.0390699989398	1010011101	0.0340582292362
1111011010	0.048980394856	1011110001	0.0390699989398	1010001001	0.0340582292362
1111010011	0.048980394856	1010111010	0.0390699989398	0010011001	0.0340582292362
1111010000	0.048980394856	1010110011	0.0390699989398	1110011100	0.0340582278763
1110010010	0.048980394856	1010110000	0.0390699989398	1110001000	0.0340582278763
1011011011	0.048980394856	1111111101	0.0352379189091	0110011000	0.0340582278763
1011011000	0.048980394856	1111101001	0.0352379189091	1110011111	0.0340582270782

APPENDIX C. SUPPLEMENTARY TABLES

1011010001	0.048980394856	1110111111	0.0352379189091	1110001011	0.0340582270782
1010011010	0.048980394856	1110111100	0.0352379189091	0110011011	0.0340582270782
1010010011	0.048980394856	1110110101	0.0352379189091	1111011101	0.0340582264157
1010010000	0.048980394856	1110101011	0.0352379189091	1111001001	0.0340582264157
1111011101	0.0450479151303	1110101000	0.0352379189091	0111011001	0.0340582264157
1111001001	0.0450479151303	1110100001	0.0352379189091	1110101101	0.00989432326996
1110011111	0.0450479151303	1010111101	0.0352379189091	0110111101	0.00989432326996
1110011100	0.0450479151303	1010101001	0.0352379189091	0110101001	0.00989432326996
1110010101	0.0450479151303	0111111001	0.0352379189091	1110101011	0.00926057061036
1110001011	0.0450479151303	0110111011	0.0352379189091	1110100001	0.00926057061036
1110001000	0.0450479151303	0110111000	0.0352379189091	0110110001	0.00926057061036
1110000001	0.0450479151303	0110110001	0.0352379189091	1010111101	0.00926056971048
1010011101	0.0450479151303	0010111001	0.0352379189091	0010111001	0.00926056971048
1010001001	0.0450479151303	1111110010	0.0169143957798	1010101001	0.00926056971047
0111011001	0.0450479151303	1011111010	0.0169143957798	1110111100	0.00926056934071
0110011011	0.0450479151303	1011110011	0.0169143957798	0110111000	0.00926056934071
0110011000	0.0450479151303	1011110000	0.0169143957798	1110101000	0.0092605693407
0110010001	0.0450479151303	1010110010	0.0169143957798	1110111111	0.00926056912371
0010011001	0.0450479151303	1111111111	0.0151408138642	1110101011	0.00926056912371
1111111011	0.0258095086286	1111111100	0.0151408138642	0110111011	0.00926056912371
1111111000	0.0258095086286	1111110101	0.0151408138642	1111111101	0.00926056894358
1111110001	0.0258095086286	1111101011	0.0151408138642	0111111001	0.00926056894358
1110111010	0.0258095086286	1111101000	0.0151408138642	1111101001	0.00926056894357
1110110011	0.0258095086286	1111100001	0.0151408138642	0110001101	0.00690444485428
1110110000	0.0258095086286	1110111110	0.0151408138642	1110000101	0.00667498075898
1011111001	0.0258095086286	1110110111	0.0151408138642	0110010101	0.00667498075898
1010111011	0.0258095086286	1110110100	0.0151408138642	0110000001	0.00667498075898
1010110000	0.0258095086286	1110101010	0.0151408138642	0010011101	0.00667498011036
1010110001	0.0258095086286	1110100011	0.0151408138642	1010001101	0.00667498011035
1110111101	0.0238566421145	1110100000	0.0151408138642	0010001001	0.00667498011035
1110101001	0.0238566421145	1011111101	0.0151408138642	1110001100	0.00667497984383
0110111001	0.0238566421145	1011101001	0.0151408138642	0110011100	0.00667497984383
1111010010	0.0201929535747	1010111111	0.0151408138642	0110001000	0.00667497984383
1011011010	0.0201929535747	1010111100	0.0151408138642	1110001111	0.00667497968742
1011010011	0.0201929535747	1010110101	0.0151408138642	0110011111	0.00667497968742
1011010000	0.0201929535747	1010101011	0.0151408138642	0110001011	0.00667497968742
1010010010	0.0201929535747	1010101000	0.0151408138642	1111001101	0.00667497955758
1111011111	0.0184124229933	1010100001	0.0151408138642	0111011101	0.00667497955758
1111011100	0.0184124229933	0111111011	0.0151408138642	0111001001	0.00667497955758
1111010101	0.0184124229933	0111111000	0.0151408138642	0110101101	0.00183369510722
1111010101	0.0184124229933	0111110001	0.0151408138642	1110100101	0.0017845658203
1111001000	0.0184124229933	0110110101	0.0151408138642	0110110101	0.0017845658203
1111000001	0.0184124229933	0110110011	0.0151408138642	0110100001	0.0017845658203
1110011110	0.0184124229933	0110110000	0.0151408138642	1010101101	0.00178456564689
1110010111	0.0184124229933	0011111001	0.0151408138642	0010111101	0.00178456564689
1110010100	0.0184124229933	0010111011	0.0151408138642	0010101001	0.00178456564689
1110010101	0.0184124229933	0010111000	0.0151408138642	1110101100	0.00178456557563
1110000011	0.0184124229933	0010110001	0.0151408138642	0110111100	0.00178456557563
1110000000	0.0184124229933	1110101101	0.013510631096	0110101000	0.00178456557563
1011011101	0.0184124229933	0110111101	0.013510631096	1110101111	0.00178456553382
1011001001	0.0184124229933	0110101001	0.013510631096	0110111111	0.00178456553382
1010011111	0.0184124229933	1011110010	0.00721904913387	0110101011	0.00178456553382
1010011100	0.0184124229933	1111111110	0.00642358208512	1111101101	0.0017845654991
1010010101	0.0184124229933	1111110111	0.00642358208512	0111111101	0.0017845654991
1010001011	0.0184124229933	1111110100	0.00642358208512	0111101001	0.0017845654991
1010001000	0.0184124229933	1111101010	0.00642358208512	0110000101	0.00122452916781
1010000001	0.0184124229933	1111100011	0.00642358208512	0010001101	0.00122452904882
0111011011	0.0184124229933	1111100000	0.00642358208512	0110001100	0.00122452899992
0111011000	0.0184124229933	1110110110	0.00642358208512	0110001111	0.00122452897123
0111010001	0.0184124229933	1110100010	0.00642358208512	0111001101	0.00122452894741
0110011010	0.0184124229933	1011111111	0.00642358208512	0110100101	0.00032405626072
0110010011	0.0184124229933	1011111100	0.00642358208512	0010101101	0.00032405622923
0110010000	0.0184124229933	1011110101	0.00642358208512	0110101100	0.000324056216291
0011011001	0.0184124229933	1011101011	0.00642358208512	0110101111	0.000324056208698
0010011011	0.0184124229933	1011101000	0.00642358208512	0111101101	0.000324056202394
0010011000	0.0184124229933	1011100001	0.00642358208512	1001001001	2.12340664662e-16
0010010001	0.0184124229933	1010111110	0.00642358208512	0101001000	8.40682998014e-17
1110001101	0.0167294097115	1010110111	0.00642358208512	0101001001	7.87854604753e-17
0110011101	0.0167294097115	1010110100	0.00642358208512	0101010011	5.71809832666e-17
0110001001	0.0167294097115	1010111100	-0.00660751072818	1111011111	-0.00920228026839
1111111010	0.0107411701405	1011001011	-0.00660751072818	0111011011	-0.00920228026839
1111110011	0.0107411701405	1011001000	-0.00660751072818	1111001011	-0.0092022802684

1111110000	0.0107411701405	1011000001	-0.00660751072818	1010110001	-0.0121963115581
1110110010	0.0107411701405	1010011110	-0.00660751072818	1110110000	-0.0121963124149
1011111011	0.0107411701405	1010001010	-0.00660751072818	1110110011	-0.0121963129178
1011111000	0.0107411701405	1010000011	-0.00660751072818	1111110001	-0.0121963133352
1011110001	0.0107411701405	0111011010	-0.00660751072818	1010111000	-0.0121963145003
1010111010	0.0107411701405	0111010011	-0.00660751072818	1010111011	-0.0121963150032
1010110011	0.0107411701405	0111010000	-0.00660751072818	1011111001	-0.0121963154206
1010110000	0.0107411701405	0110010010	-0.00660751072818	1110111010	-0.0121963158601
1111111101	0.00983153857517	0011011011	-0.00660751072818	1111111000	-0.0121963162775
1111101001	0.00983153857517	0011011000	-0.00660751072818	1111111011	-0.0121963167804
1110111111	0.00983153857517	0011010001	-0.00660751072818	1011110010	-0.0121970856193
1110111100	0.00983153857517	0010011010	-0.00660751072818	1010010110	-0.0130268890986
1110110101	0.00983153857517	0010010011	-0.00660751072818	1010000010	-0.0130268890986
1110101011	0.00983153857517	0010010000	-0.00660751072818	0010010010	-0.0130268890986
1110101000	0.00983153857517	1011010010	-0.00740376231762	1011010100	-0.013026889167
1110100001	0.00983153857517	1110001101	-0.0138562285143	1011000000	-0.013026889167
1010111101	0.00983153857517	0110011101	-0.0138562285143	0011010000	-0.013026889167
1010101001	0.00983153857517	0110001001	-0.0138562285143	1011010111	-0.0130268892494
0111111001	0.00983153857517	1111011111	-0.0154768015649	1011000011	-0.0130268892494
0110111011	0.00983153857517	1111011100	-0.0154768015649	0011010011	-0.0130268892494
0110111000	0.00983153857517	1111010101	-0.0154768015649	1111010110	-0.0130268893898
0110110001	0.00983153857517	1111001011	-0.0154768015649	1111000010	-0.0130268893898
0010111001	0.00983153857517	1111001000	-0.0154768015649	0111010010	-0.0130268893898
1011010010	0.00819106381112	1111000001	-0.0154768015649	1011011110	-0.0130268897314
1111011110	0.00741876485419	1110011110	-0.0154768015649	1011001010	-0.0130268897314
1111010111	0.00741876485419	1110010111	-0.0154768015649	0011011010	-0.0130268897314
1111010100	0.00741876485419	1110010100	-0.0154768015649	1010010100	-0.0162141034315
1111001010	0.00741876485419	1110001010	-0.0154768015649	1010000000	-0.0162141034315
1111000011	0.00741876485419	1110000011	-0.0154768015649	0010010000	-0.0162141034315
1111000000	0.00741876485419	1110000000	-0.0154768015649	1010010111	-0.0162141036093
1110010110	0.00741876485419	1011011101	-0.0154768015649	1010000011	-0.0162141036093
1110000010	0.00741876485419	1011001001	-0.0154768015649	0010010011	-0.0162141036093
1011011111	0.00741876485419	1010011111	-0.0154768015649	1011010101	-0.0162141037569
1011011100	0.00741876485419	1010011100	-0.0154768015649	1011000001	-0.0162141037569
1011010101	0.00741876485419	1010010101	-0.0154768015649	0011010001	-0.0162141037569
1011001011	0.00741876485419	1010001011	-0.0154768015649	1110010110	-0.0162141039123
1011001000	0.00741876485419	1010001000	-0.0154768015649	1110000010	-0.0162141039123
1011000001	0.00741876485419	1010000001	-0.0154768015649	0110010010	-0.0162141039123
1010011110	0.00741876485419	0111011011	-0.0154768015649	1111010100	-0.0162141040599
1010010111	0.00741876485419	0111011000	-0.0154768015649	1111000000	-0.0162141040599
1010010100	0.00741876485419	0111010001	-0.0154768015649	0111010000	-0.0162141040599
1010000011	0.00741876485419	0110011010	-0.0154768015649	1111010111	-0.0162141042377
1010000000	0.00741876485419	0110010101	-0.0154768015649	1111000011	-0.0162141042377
0111011010	0.00741876485419	0011011001	-0.0154768015649	0111010011	-0.0162141042377
0111010011	0.00741876485419	0011011011	-0.0154768015649	1010011110	-0.0162141046497
0111010000	0.00741876485419	0010011000	-0.0154768015649	1010001010	-0.0162141046497
0110010010	0.00741876485419	0010011001	-0.0154768015649	0010011010	-0.0162141046497
0011011011	0.00741876485419	1111010010	-0.0172256615275	1011011100	-0.0162141047973
0011011000	0.00741876485419	1011011010	-0.0172256615275	1011001000	-0.0162141047973
0011010001	0.00741876485419	1011010011	-0.0172256615275	0011011000	-0.0162141047973
0010011010	0.00741876485419	1011010000	-0.0172256615275	1011011111	-0.0162141049751
0010010011	0.00741876485419	1010010010	-0.0172256615275	1011001011	-0.0162141049751
0010010000	0.00741876485419	1111011101	-0.0357354568504	0011011011	-0.0162141049751
1111001101	0.0067002776986	1111001001	-0.0357354568504	1111011110	-0.0162141052781
1110001111	0.0067002776986	1110011111	-0.0357354568504	1111001010	-0.0162141052781
1110001100	0.0067002776986	1110011100	-0.0357354568504	0111011010	-0.0162141052781
1110000101	0.0067002776986	1110010101	-0.0357354568504	1010110010	-0.0182908991352
1010001101	0.0067002776986	1110001011	-0.0357354568504	1011110000	-0.0182908992312
0111011101	0.0067002776986	1110001000	-0.0357354568504	1011110011	-0.0182908993469
0111010010	0.0067002776986	1110000001	-0.0357354568504	1111110010	-0.018290899544
0110011111	0.0067002776986	1010011101	-0.0357354568504	1011111010	-0.0182909000238
0110011100	0.0067002776986	1010001001	-0.0357354568504	1010110000	-0.0221957426608
0110010101	0.0067002776986	0111011001	-0.0357354568504	1010110011	-0.0221957429042
0110001011	0.0067002776986	0110011011	-0.0357354568504	1011110001	-0.0221957431063
0110001000	0.0067002776986	0110011000	-0.0357354568504	1110110010	-0.022195743319
0110000001	0.0067002776986	0110010001	-0.0357354568504	1111110000	-0.0221957435211
0010011101	0.0067002776986	0010011001	-0.0357354568504	1111110011	-0.0221957437645
0010001001	0.0067002776986	1111010011	-0.0394323709138	1010111010	-0.0221957443285
1111110010	0.0043888612532	1111010000	-0.0394323709138	1011111000	-0.0221957445305
1011111010	0.0043888612532	1110010010	-0.0394323709138	1011111011	-0.0221957447739
1011110011	0.0043888612532	1011011011	-0.0394323709138	1111111010	-0.0221957451887
1011100011	0.0043888612532	1011011010	-0.0394323709138	1010010001	-0.0441111312141

1011110000	0.0043888612532	1011011000	-0.0394323709138	1110010000	-0.0441111343133
1010110010	0.0043888612532	1011010001	-0.0394323709138	1110010011	-0.044111136132
1111111111	0.00398701318423	1010011010	-0.0394323709138	1111010001	-0.0441111376418
1111111100	0.00398701318423	1010010011	-0.0394323709138	1010011000	-0.0441111418555
1111110101	0.00398701318423	1010010000	-0.0394323709138	1010011011	-0.0441111436743
1111101011	0.00398701318423	1111011010	-0.0394323709139	1011011001	-0.044111145184
1111101000	0.00398701318423	1110011101	-0.0809979189956	1110011010	-0.0441111467734
1111100001	0.00398701318423	1110001001	-0.0809979189956	1111011000	-0.0441111482832
1110111110	0.00398701318423	0110011001	-0.0809979189956	1111011011	-0.0441111501019
1110110111	0.00398701318423	1111011011	-0.0883617782457	1011010010	-0.0453652843072
1110110100	0.00398701318423	1111011000	-0.0883617782457	1010010010	-0.0675537453698
1110101010	0.00398701318423	1111010001	-0.0883617782457	1011010000	-0.0675537457244
1110100011	0.00398701318423	1110011010	-0.0883617782457	1011010011	-0.0675537461517
1110100000	0.00398701318423	1110010011	-0.0883617782457	1111010010	-0.0675537468797
1011111101	0.00398701318423	1110010000	-0.0883617782457	1011011010	-0.0675537486515
1011101001	0.00398701318423	1011011001	-0.0883617782457	1010010000	-0.0812408242887
1010111111	0.00398701318423	1010011011	-0.0883617782457	1010010011	-0.0812408251797
1010111100	0.00398701318423	1010011000	-0.0883617782457	1011010001	-0.0812408259193
1010110101	0.00398701318423	1010010001	-0.0883617782457	1110010010	-0.0812408266979
1010101011	0.00398701318423	1111011001	-0.192284080152	1111010000	-0.0812408274375
1010101000	0.00398701318423	1110011011	-0.192284080152	1111010011	-0.0812408283284
1010100001	0.00398701318423	1110011000	-0.192284080152	1010011010	-0.0812408303927
0111111011	0.00398701318423	1110010001	-0.192284080152	1011011000	-0.0812408311323
0111111000	0.00398701318423	1010011001	-0.192284080152	1011011011	-0.0812408320232
0111110001	0.00398701318423	1110011001	-0.400881271349	1111011010	-0.0812408335415

Table C.30: Eigenvectors λ_1 , λ_2 and λ_3 for the Tbet/Gata3 MISA motif with leading ranked microstates for λ_1 and microstates with largest positive and negative λ_2 component. The binary microstate code denotes activity of the aforementioned ESCs.

APPENDIX D

Code documentation

D.1 Algorithmic commands

```
1 ## ChIP-Seq
2 # Bowtie
3 bowtie -t -m 1 -S -q -p 8 <Genome Location> <FASTQ File> <SAM File Output>
4
5 # SICER (modified script)
6 SICER_modified.sh <Input file folder> <Input file name> <Control file name> <Output
  folder> t mm10 1 200 150 0.8
7
8 ## RNA-Seq
9 # STAR
10 STAR --genomeDir <Directory> --runThreadN 8 --readFilesIn <Input file rep1 lane1>,<
  Input file rep1 lane2> <Input file rep2 lane1>,<Input file rep2 lane2> --
  outFilePrefix <Prefix> --outSAMtype BAM SortedByCoordinate
11
12 # HTSeq
13 python -m HTSeq.scripts.count -f bam <BAM Input file> <GTF file> > <Count file output>
14
15 ## HMM
16 java -mx2000M -jar ChromHMM.jar BinarizeBed -peaks <Genome file> <Input file folder> <
  Cellmark file> <Output folder>
17
18 java -mx8000M -jar ChromHMM.jar LearnModel -p 8 -r 400 -color 0,0,255 -printposterior
  -printstatebyline <Binarized file folder> <Output folder> 16 mm10
```

Listing D.1: Algorithmic commands with respective options from the data pre-processing workflow including the HMM.

D.2 Correlation algorithm

```
1 histone_correlation.sh Ifng-201 p1_1_p2_1.4094_p3_2.7373.bed -r 1 -m --resolution 600
```

Listing D.2: Simple example for the execution of the correlation algorithm.

```
1 ##### fragmentmerging: Merge adjacent similar fragments by mean and sd of the difference
  across conditions
2 fragmentmerging <- function(frac, meanlimarg, sylimarg) {
3 ##### fragmentmerging: Merge adjacent similar fragments by mean and sd of the difference
  across conditions
4 ## Requires modification data as produced by enhancerfractionation and the thresholds
  for similarity in mean and sd
5 ## ARGS:
```

```

6 ## - frac = dataframe with two columns for the fragment position and replicates*
  conditions columns with the corresponding modification data
7 ## - meanlimarg = vector of two values giving the negative and positive boundary for
  similar mean values
8 ## - sylimarg = numeric value giving the maximum sd for similar fragments
9 ## VALUE:
10 ## - frac = dataframe similar to the input frac, but with fewer lines where fragments
  were combined
11 ## - file names meansddiff.png showing the differences between neighbouring fragments
12
13 # Combine neighbouring enhancers with similar density distribution
14 cols <- dim(frac)[2]
15 remove <- c(1)
16 cat("Merging similar neighbouring fragments.\n")
17 flush.console()
18 while(length(remove) != 0) {
19
20   # Calculate mean and sd difference between neighbouring regions
21   meandiff <- c()
22   for(i in 2:dim(frac)[1]) meandiff <- c(meandiff, mean(as.numeric(frac[i,3:cols]-
     frac[i-1,3:cols])))
23   sddiff <- c()
24   for(i in 2:dim(frac)[1]) sddiff <- c(sddiff, sd(as.numeric(frac[i,3:cols]-frac[i
     -1,3:cols])))
25
26   # The first time these were calculated generate a plot including the automatically
     calculated quantiles and the set limits
27   if (remove[1] == 1) {
28     # 1. automatic criterion would be to have an average difference between
     conditions that is within the inner 50% quantile across all neighbouring
     fragments
29     meanlim <- summary(meandiff)[c("1st Qu.", "3rd Qu.")]
30     # 2. automatic criterion would be to have a sd difference between conditions that
     is within the left 25% quantile across all neighbouring fragments
31     sylim <- quantile(sddiff, 0.5)
32
33     # Plot distribution of differences
34     png(filename="meansddiff.png")
35     plot(meandiff, sddiff, type="p", xlab="mean difference", ylab="sd difference")
36     # Plot calculated lines in blue and argument lines in red
37     lines(c(meanlim[1], meanlim[1], meanlim[2], meanlim[2]), c(0, sylim, sylim, 0),
     col="blue")
38     lines(c(sylimarg[1], sylimarg[1], sylimarg[2], sylimarg[2]), c(0,
     sylim, sylim, 0), col="red")
39     # Add title and close the png device
40     title(sub = paste0("mean = ", meanlim[1], " - ", meanlim[2], ", sd = ", sylim),
     main = "Distribution of differences")
41     dev.off()
42
43     # Set both criteria to input value
44     meanlim <- meanlimarg
45     cat(paste0("Matching limit for mean difference is ", meanlim[1], "-", meanlim[2],
     ".\n"))
46     sylim <- sylimarg
47     cat(paste0("Matching limit for sd difference is ", sylim, ".\n"))
48
49     # How many neighbours are within the limit?
50     meanmatch <- length(which(meandiff >= meanlim[1] & meandiff <= meanlim[2] &
     sddiff <= sylim))
51     # If no pairs are within the limit return frac as it was
52     if (meanmatch == 0) {
53       cat("There no fragments similar enough to be fused. Please consider setting
     different thresholds for mean and sd.\n")
54       return(frac)
55     }
56     # Create Progress bar
57     pb <- txtProgressBar(min = -meanmatch, max = 0, style=3)
58   }
59
60   # The following is done in all cases including the first
61   # How many neighbours are within the limit?
62   meanmatch <- length(which(meandiff >= meanlim[1] & meandiff <= meanlim[2] & sddiff

```



```

    <= sdlim))
63 # Set Progress bar to value according to possible matches
64 setTxtProgressBar(pb, -meanmatch)
65
66 # Combine matching rows
67 remove <- c()
68 # Go through all lines starting at number two and compare with the neighbouring
    fragments to decide whether it has to be fused
69 for(i in 2:dim(frac)[1]) {
70
71     # only left sided comparison for last element
72     if (i == dim(frac)[1]) {
73
74         # Check whether left neighbour is adjacent and similar
75         if (frac$starts[i] == frac$stops[i-1]) {
76             # logical statement to test whether its similar:
77             left <- meandiff[i-1] >= meanlim[1] & meandiff[i-1] <= meanlim[2] & sddiff[i
                -1] <= sdlim
78         } else left <- FALSE
79
80         # Combine to left neighbour if it matches
81         if (left == TRUE) { # Left neighbour fits and wasn't matched to its left
                neighbour (otherwise they wouldn't combine)
82             # calculate sizes of fragments to fuse
83             sizel <- frac$stops[i-1]-frac$starts[i-1]
84             sizer <- frac$stops[i]-frac$starts[i]
85             # convert densities to modcounts in order to calculate the average weighted
                by the fragment size
86             frac[(i-1),3:cols] <- frac[(i-1),3:cols]*sizel
87             frac[i,3:cols] <- frac[i,3:cols]*sizer
88             # overwrite left neighbour with average modcount and expand its range
89             frac[(i-1),] <- cbind(frac[(i-1),1], frac[i,2], t(colMeans(frac[(i-1):i, 3:
                cols])))
90             # convert average modcount back to density by dividing through average length
                of the fused fragments
91             frac[(i-1),3:cols] <- frac[(i-1),3:cols]*2/(sizel+sizer)
92             # add current line to the list of elements to be removed
93             remove <- c(remove, i)
94             # change current line's stop to make sure that next line can't combine with
                it
95             frac$stops[i] <- 0
96         }
97
98     } else if (frac$starts[i] != frac$stops[i-1] & frac$stops[i] != frac$starts[i+1])
        { # Skip isolated rows
99         next
100     } else { # all rows that are neither the last nor isolated
101
102         # Check whether left neighbour is similar
103         if (frac$starts[i] == frac$stops[i-1]) {
104             left <- meandiff[i-1] >= meanlim[1] & meandiff[i-1] <= meanlim[2] & sddiff[i
                -1] <= sdlim
105         } else left <- FALSE
106         # Check whether right neighbour is similar
107         if (frac$stops[i] == frac$starts[i+1]) {
108             right <- meandiff[i] >= meanlim[1] & meandiff[i] <= meanlim[2] & sddiff[i] <=
                sdlim
109         } else right <- FALSE
110
111         # Combine best fitting row
112         if (right == FALSE & left == FALSE) { # Both neighbours are different
113             next
114         } else if (right == FALSE & left == TRUE) { # Only left neighbour fits and wasn
                't matched yet
115             # calculate sizes of fragments to fuse
116             sizel <- frac$stops[i-1]-frac$starts[i-1]
117             sizer <- frac$stops[i]-frac$starts[i]
118             # convert densities to modcounts in order to calculate the average weighted
                by the fragment size
119             frac[(i-1),3:cols] <- frac[(i-1),3:cols]*sizel
120             frac[i,3:cols] <- frac[i,3:cols]*sizer
121             # overwrite left neighbour with average modcount and expand its range

```

```

122     frac[(i-1),] <- cbind(frac[(i-1),1], frac[i,2], t(colMeans(frac[(i-1):i, 3:
123         cols])))
123     # convert average modcount back to density by dividing through average length
124         of the fused fragments
124     frac[(i-1),3:cols] <- frac[(i-1),3:cols]*2/(sizel+sizer)
125     # add current line to the list of elements to be removed
126     remove <- c(remove, i)
127     # change current line's stop to make sure that next line can't combine with
128         it
128     frac$stops[i] <- 0
129   } else if (right == TRUE & left == FALSE) { # Only right neighbour fits but
130         wasn't matched yet
130     # Right neighbour will either match the current line or fit better to its
131         next line
131     next
132   } else if (right == TRUE & left == TRUE) { # Both neighbours fit
133     # Check whether left or right neighbour is the better match
134     if (abs(meandiff[i-1]) < abs(meandiff[i])) { # left match is better and didn'
135         t match before
135     # calculate sizes of fragments to fuse
136     sizel <- frac$stops[i-1]-frac$starts[i-1]
137     sizer <- frac$stops[i]-frac$starts[i]
138     # convert densities to modcounts in order to calculate the average weighted
139         by the fragment size
139     frac[(i-1),3:cols] <- frac[(i-1),3:cols]*sizel
140     frac[i,3:cols] <- frac[i,3:cols]*sizer
141     # overwrite left neighbour with average modcount and expand its range
142     frac[(i-1),] <- cbind(frac[(i-1),1], frac[i,2], t(colMeans(frac[(i-1):i, 3:
143         cols])))
143     # convert average modcount back to density by dividing through average
144         length of the fused fragments
144     frac[(i-1),3:cols] <- frac[(i-1),3:cols]*2/(sizel+sizer)
145     # add current line to the list of elements to be removed
146     remove <- c(remove, i)
147     # change current line's stop to make sure that next line can't combine with
148         it
148     frac$stops[i] <- 0
149   } else {
150     # Right neighbour might still find a better match to its right side so go
151         next
151     # This might cause a whole chain of next cases, but it will end latest at
152         the end of the current neighbourhood. In the next repeat of the while
153         loop the case might change. So this next increases the runtime but is
154         the only correct option.
152     next
153   }
154 }
155 }
156
157 # Continue with next line of frac
158 }
159
160 # After going through alllines of frac and matching leftsided, similar neighbours
161     remove extra lines and start over
161     if (length(remove) != 0) frac <- frac[-remove,]
162 }
163 # Finish Progress bar
164 setTxtProgressBar(pb, 0)
165 close(pb)
166
167 return(frac)
168 }

```

Listing D.3: R-function for enhancer fragment merging as implemented in the correlation algorithm.

D.3 Class-specificity computation

```

1 import numpy as np
2 from numpy import *
3 from sklearn.preprocessing import scale
4 from sklearn.ensemble import ExtraTreesClassifier
5
6 X = ... #read feature matrix input
7 Y = ... #read binary target vector input
8 y=Y.astype(int)
9 M, N = X.shape
10 X2=scale(X)
11
12 # train ERT classifier
13 forest = ExtraTreesClassifier(n_estimators=10000,random_state=0)
14 forest.fit(X, y)
15
16 # extract Gini impurities
17 importances = forest.feature_importances_
18
19 # calculate intra-class Gini impurity for Th1 as an example
20 gini_intraclass_th1=sorted(zip(np.mean(X2[y=='Th1', :], axis=0)*importances*(X2[y=='Th2',
    :, :].sum(axis=1).sum()/(shape(X2[y=='Th2', :, :].sum(axis=1))[0]))/(abs(np.mean(X2[y=='
    'Th2', :, :], axis=0))*X2[y=='Th1', :, :].sum(axis=1).sum()/(shape(X2[y=='Th1', :, :].sum(
    axis=1))[0])), range(N))
21
22
23 # do LOOCV predictions
24 predictall=[]
25 for i in range(M):
26     Xtest=np.concatenate((X2[0:i,:], X2[i+1:M,:]), axis=0)
27     Ytest=np.concatenate((Y[0:i], Y[i+1:M]), axis=0)
28     forest.fit(Xtest, Ytest)
29     predictall.append(forest.predict(X2[i:i+1,:]))

```

Listing D.4: Intra-class Gini impurity computation in Python.

D.4 Computational dependencies & packages

General:

Bowtie v1.2.1.1

STAR v2.4.0j

SICER v1.1

HT-Seq v0.6.1

ChromHMM v1.10

HOMER v4.7

Cytoscape v3.4.0

cytoscape.js v3.2.14

Python:

python v2.7.13

```
networkx v1.11  
scikit-learn v0.19.1  
scipy v0.19.0  
pandas v0.19.2  
seaborn v0.8.1  
matplotlib v2.0.2  
numpy v1.13.3
```

R:

```
R v3.3.2  
vsn v3.48.1  
DESeq2 v1.20
```

Acronyms

- AIC** Akaike Information Criterion.
- BIC** Bayesian Information Criterion.
- CNS** Conserved Non-Coding Sequence.
- CSC** Chromatin State Class.
- ERT** Extremely Randomized Trees.
- ESC** Enhancer State Class.
- FDR** False Discovery Rate.
- FFL** Feed-Forward Loop.
- GRN** Gene Regulatory Network.
- HAT** Histone Acetyltransferase.
- HER** Hierarchical Edge Removal.
- HMM** Hidden Markov Model.
- LFC** Logarithmic Fold Change.
- LOOCV** Leave-One-Out Cross-Validation.
- MAP** Maximum Likelihood A Posteriori Estimate.
- MCL** Markov Cluster Algorithm.
- MISA** Mutual-Inhibition and Self-Activation.
- PCA** Principal Component Analysis.
- RSC** Repressive State Class.
- RWIF** Random Walk Information Flow.
- SC** Spectral Clustering.

TAD Topologically Associating Domain.

TES Transcription End Site.

TF Transcription Factor.

TSS Transcription Start Site.

VST Variance Stabilizing Transformation.

Bibliography

- [1] ABBAS, A. K., MURPHY, K. M., AND SHER, A. Functional diversity of helper T lymphocytes., 1996.
- [2] AHSENDORF, T., WONG, F., EILS, R., AND GUNAWARDENA, J. A framework for modelling gene regulation which accommodates non-equilibrium mechanisms. *BMC Biology* 12, 1 (dec 2014), 102.
- [3] AKHABIR, L., AND SANDFORD, A. Genetics of interleukin 1 receptor-like 1 in immune and inflammatory diseases. *Current genomics* 11, 8 (dec 2010), 591–606.
- [4] ALBERT, R., AND BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 1 (jan 2002), 47–97.
- [5] ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K., AND WALTER, P. *Molecular biology of the cell*. Garland Science, 2002.
- [6] ALON, U. Network motifs: theory and experimental approaches. *Nature Reviews Genetics* 8, 6 (jun 2007), 450–461.
- [7] AMIT, Y., AND GEMAN, D. Shape Quantization and Recognition with Randomized Trees. *Neural Computation* 9, 7 (oct 1997), 1545–1588.
- [8] ANDERS, S., AND HUBER, W. Differential expression analysis for sequence count data. *Genome Biology* 11, 10 (oct 2010), R106.
- [9] ANDERS, S., PYL, P. T., AND HUBER, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* 31, 2 (jan 2015), 166–9.
- [10] ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I., HOOF, I., BORNHOLDT, J., BOYD, M., CHEN, Y., ZHAO, X., SCHMIDL, C., SUZUKI, T., NTINI, E., ARNER, E., VALEN, E., LI, K., SCHWARZFISCHER, L., GLATZ, D., RAITHEL, J., LILJE, B., RAPIN, N., BAGGER, F. O., JØRGENSEN, M., ANDERSEN, P. R., BERTIN, N., RACKHAM, O., BURROUGHS, A. M., BAILLIE, J. K., ISHIZU, Y., SHIMIZU, Y., FURUHATA, E., MAEDA, S., NEGISHI, Y., MUNGALL, C. J., MEEHAN, T. F., LASSMANN, T., ITOH, M., KAWAJI, H., KONDO, N., KAWAI, J., LENNARTSSON, A., DAUB, C. O., HEUTINK, P., HUME, D. A., JENSEN, T. H., SUZUKI, H., HAYASHIZAKI, Y., MÜLLER, F., FORREST, A. R. R., CARNINCI, P., REHLI, M., AND SANDELIN, A. An atlas of active enhancers across human cell types and tissues. *Nature* 507, 7493 (2014), 455–61.

- [11] ANDRECU, M., HALLEY, J. D., WINKLER, D. A., AND HUANG, S. A general model for binary cell fate decision gene circuits with degeneracy: Indeterminacy and switch behavior in the absence of cooperativity. *PLoS ONE* 6, 5 (2011).
- [12] ANGELINI, C., AND COSTA, V. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Frontiers in Cell and Developmental Biology* 2 (sep 2014), 51.
- [13] ANSEL, K. M., DJURETIC, I., TANASA, B., AND RAO, A. Regulation of Th2 differentiation and Il4 locus accessibility. *Annual review of immunology* 24 (2006), 607–656.
- [14] ANTEBI, Y. E., REICH-ZELIGER, S., HART, Y., MAYO, A., EIZENBERG, I., RIMER, J., PUTHETI, P., PE'ER, D., AND FRIEDMAN, N. Mapping differentiation under mixed culture conditions reveals a tunable continuum of T cell fates. *PLoS biology* 11, 7 (2013), e1001616.
- [15] AOYAGI, S., NARLIKAR, G., ZHENG, C., SIF, S., KINGSTON, R. E., AND HAYES, J. J. Nucleosome remodeling by the human SWI/SNF complex requires transient global disruption of histone-DNA interactions. *Molecular and cellular biology* 22, 11 (jun 2002), 3653–62.
- [16] ARNONE, M. I., AND DAVIDSON, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124, 10 (may 1997), 1851–64.
- [17] ARNOSTI, D. N., AND KULKARNI, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry* 94, 5 (2005), 890–898.
- [18] BAILEY, N. T. J. *Statistical Methods in Biology*. Cambridge University Press, Cambridge, 1995.
- [19] BALASUBRAMANI, A., MUKASA, R., HATTON, R. D., AND WEAVER, C. T. Regulation of the Ifng locus in the context of T-lineage specification and plasticity. *Immunological reviews* 238, 1 (nov 2010), 216–32.
- [20] BALASUBRAMANI, A., WINSTEAD, C. J., TURNER, H., JANOWSKI, K. M., HARBOUR, S. N., SHIBATA, Y., CRAWFORD, G. E., HATTON, R. D., AND WEAVER, C. T. Deletion of a Conserved cis-Element in the Ifng Locus Highlights the Role of Acute Histone Acetylation in Modulating Inducible Gene Transcription. *PLoS Genetics* 10, 1 (2014).
- [21] BARABÁSI, A. L., AND OLTVAI, Z. N. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5, 2 (2004), 101–113.
- [22] BARABÁSI, A.-L., AND PÓSFAL, M. *Network science*. Cambridge University Press, Cambridge, 2016.
- [23] BARBER, D. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

- [24] BARBER, R. D., HARMER, D. W., COLEMAN, R. A., AND CLARK, B. J. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiological Genomics* 21, 3 (may 2005), 389–395.
- [25] BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T.-Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I., AND ZHAO, K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129, 4 (may 2007), 823–837.
- [26] BASSO, K., MARGOLIN, A. A., STOLOVITZKY, G., KLEIN, U., DALLA-FAVERA, R., AND CALIFANO, A. Reverse engineering of regulatory networks in human B cells. *Nature Genetics* 37, 4 (apr 2005), 382–390.
- [27] BATTISTON, F., NICOSIA, V., AND LATORA, V. Structural measures for multiplex networks. *Physical Review E* 89, 3 (mar 2014), 032804.
- [28] BATTISTON, F., NICOSIA, V., AND LATORA, V. The new challenges of multiplex networks: Measures and models. *The European Physical Journal Special Topics* 226, 3 (feb 2017), 401–416.
- [29] BEISEL, C., AND PARO, R. Silencing chromatin: Comparing modes and mechanisms. *Nature Reviews Genetics* 12, 2 (2011), 123–135.
- [30] BERNARD, A., AND HARTEMINK, A. J. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing* (2005), 459–70.
- [31] BERNHART, S. H., KRETZMER, H., HOLDT, L. M., JÜHLING, F., AMMERPOHL, O., BERGMANN, A. K., NORTHOFF, B. H., DOOSE, G., SIEBERT, R., STADLER, P. F., AND HOFFMANN, S. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Scientific Reports* 6 (2016), 1–18.
- [32] BERNSTEIN, B. E., MEISSNER, A., AND LANDER, E. S. The Mammalian Epigenome. *Cell* 128, 4 (feb 2007), 669–681.
- [33] BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L., ECKER, J. R., FARNHAM, P. J., HIRST, M., LANDER, E. S., MIKKELSEN, T. S., AND THOMSON, J. A. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* 28, 10 (oct 2010), 1045–8.
- [34] BIANCONI, G. Statistical mechanics of multiplex networks : Entropy and overlap. 1–15.
- [35] BIESINGER, J., WANG, Y., AND XIE, X. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics* 2013 14:5 14, 5 (apr 2013).
- [36] BISHOP, C. M. *Pattern recognition and machine learning*. Springer, 2006.
- [37] BOCCALETTI, S., BIANCONI, G., CRIADO, R., DEL GENIO, C., GÓMEZ-GARDEÑES, J., ROMANCE, M., SENDIÑA-NADAL, I., WANG, Z., AND ZANIN, M. The structure and dynamics of multilayer networks. *Physics Reports* 544, 1 (nov 2014), 1–122.

- [38] BOCCALETTI, S., LATORA, V., MORENO, Y., CHAVEZ, M., AND HWANG, D. Complex networks: Structure and dynamics. *Physics Reports* 424, 4-5 (feb 2006), 175–308.
- [39] BOCK, C. N., BABU, S., BRELOER, M., RAJAMANICKAM, A., BOOTHRA, Y., BRUNN, M.-L., KÜHL, A. A., MERLE, R., LÖHNING, M., HARTMANN, S., AND RAUSCH, S. Th2/1 Hybrid Cells Occurring in Murine and Human Strongyloidiasis Share Effector Functions of Th1 Cells. *Frontiers in cellular and infection microbiology* 7 (2017), 261.
- [40] BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N., ALONSO-BETANZOS, A., BENÍTEZ, J. M., AND HERRERA, F. A review of microarray datasets and applied feature selection methods. 111–135.
- [41] BONDY, J. A. J. A., AND MURTY, U. S. R. *Graph theory*. Springer, 2008.
- [42] BONELLI, M., SHIH, H.-Y., HIRAHARA, K., SINGELTON, K., LAURENCE, A., POHOLEK, A., HAND, T., MIKAMI, Y., VAHEDI, G., KANNO, Y., AND O'SHEA, J. J. Helper T Cell Plasticity: Impact of Extrinsic and Intrinsic Signals on Transcriptomes and Epigenomes. In *Current topics in microbiology and immunology*, vol. 381. 2014, pp. 279–326.
- [43] BONN, S., ZINZEN, R. P., GIRARDOT, C., GUSTAFSON, E. H., PEREZ-GONZALEZ, A., DELHOMME, N., GHAVI-HELM, Y., WILCZYŃSKI, B., RIDDELL, A., AND FURLONG, E. E. M. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics* 44, 2 (feb 2012), 148–156.
- [44] BONNEAU, R., REISS, D. J., SHANNON, P., FACCIOTTI, M., HOOD, L., BALIGA, N. S., AND THORSSON, V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology* 7, 5 (2006), R36.
- [45] BORK, P., JENSEN, L. J., VON MERING, C., RAMANI, A. K., LEE, I., AND MARCOTTE, E. M. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology* 14, 3 (jun 2004), 292–299.
- [46] BORNHOLDT, S. Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society, Interface* 5 Suppl 1, Suppl 1 (aug 2008), S85–94.
- [47] BOWMAN, G. R., PANDE, V., AND NOÉ, F. *An introduction to Markov State Models and their application to long timescale molecular simulation*. Springer, 2014.
- [48] BOYLE, A. P., DAVIS, S., SHULHA, H. P., MELTZER, P., MARGULIES, E. H., WENG, Z., FUREY, T. S., AND CRAWFORD, G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 2 (jan 2008), 311–22.
- [49] BRADLEY, L. M., DALTON, D. K., AND CROFT, M. A direct role for IFN-gamma in regulation of Th1 cell development. *Journal of immunology (Baltimore, Md. : 1950)* 157, 4 (aug 1996), 1350–8.

- [50] BREIMAN, L. Bagging Predictors. *Machine Learning* 24, 2 (1996), 123–140.
- [51] BREIMAN, L. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
- [52] BREINDL, C., AND ALLGÖWER, F. Verification of multistability in gene regulation networks: A combinatorial approach. *Proceedings of the IEEE Conference on Decision and Control* (2009), 5637–5642.
- [53] BRIN, S., AND PAGE, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30, 1-7 (apr 1998), 107–117.
- [54] BROIDO, A. D., AND CLAUSET, A. Scale-free networks are rare. *arXiv:1801.03400* (jan 2018).
- [55] BUDDEN, D. M., AND CRAMPIN, E. J. Information theoretic approaches for inference of biological networks from continuous-valued data. *BMC systems biology* 10, 1 (2016), 89.
- [56] BULGER, M., AND GROUDINE, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* 144, 3 (feb 2011), 327–339.
- [57] BURNHAM, K. P., ANDERSON, D. R., AND BURNHAM, K. P. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, 2002.
- [58] BURROWS, M., AND WHEELER, D. J. A block-sorting lossless data compression algorithm. *SRC Research Report* (1994).
- [59] CALDARELLI, G., CAPOCCI, A., DE LOS RIOS, P., AND MUÑOZ, M. A. Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Physical Review Letters* 89, 25 (dec 2002), 258702.
- [60] CALO, E., AND WYSOCKA, J. Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell* 49, 5 (2013), 825–837.
- [61] CANTINI, L., MEDICO, E., FORTUNATO, S., AND CASELLE, M. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific Reports* 5 (2015), 1–10.
- [62] CAO, Q., ANYANSI, C., HU, X., XU, L., XIONG, L., TANG, W., MOK, M. T., CHENG, C., FAN, X., GERSTEIN, M., CHENG, A. S., AND YIP, K. Y. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nature Genetics* 49, 10 (2017), 1428–1436.
- [63] CARDILLO, A., GÓMEZ-GARDEÑES, J., ZANIN, M., ROMANCE, M., PAPO, D., DEL POZO, F., AND BOCCALETTI, S. Emergence of network features from multiplexity. *Scientific Reports* 3, 1 (dec 2013), 1344.
- [64] CARLETON, J. B., BERRETT, K. C., AND GERTZ, J. Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor α -Bound Enhancers. *Cell Systems* 5, 4 (2017), 333–344.e5.

- [65] CHAI, L. E., LOH, S. K., LOW, S. T., MOHAMAD, M. S., DERIS, S., AND ZAKARIA, Z. A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine* 48, 1 (2014), 55–65.
- [66] CHAUDHRY, A., SAMSTEIN, R., TREUTING, P., LIANG, Y., PILS, M., HEINRICH, J.-M., JACK, R., WUNDERLICH, F., BRÜNING, J., MÜLLER, W., AND RUDENSKY, A. Interleukin-10 Signaling in Regulatory T Cells Is Required for Suppression of Th17 Cell-Mediated Inflammation. *Immunity* 34, 4 (apr 2011), 566–578.
- [67] CHEN, T., AND DENT, S. Y. Chromatin modifiers and remodellers: Regulators of cellular differentiation. *Nature Reviews Genetics* 15, 2 (2014), 93–106.
- [68] CHTANOVA, T., TANGYE, S. G., NEWTON, R., FRANK, N., HODGE, M. R., ROLPH, M. S., AND MACKAY, C. R. T follicular helper cells express a distinctive transcriptional profile, reflecting their role as non-Th1/Th2 effector cells that provide help for B cells. *Journal of immunology (Baltimore, Md. : 1950)* 173, 1 (jul 2004), 68–78.
- [69] CHU, B. K., TSE, M. J., SATO, R. R., AND READ, E. L. Markov State Models of gene regulatory networks. *BMC Systems Biology* 11, 1 (2017), 1–17.
- [70] CHUNG, K. L. *Markov chains with stationary transition probabilities*. Springer, 1967.
- [71] CLAUSET, A., SHALIZI, C. R., AND NEWMAN, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Review* 51, 4 (nov 2009), 661–703.
- [72] CLOOS, P. A., CHRISTENSEN, J., AGGER, K., AND HELIN, K. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes & Development* 22, 9 (may 2008), 1115–1140.
- [73] COHEN, R., AND HAVLIN, S. *Complex networks : structure, robustness, and function*. Cambridge University Press, 2010.
- [74] COLLINS, P. L., HENDERSON, M. A., AND AUNE, T. M. Diverse functions of distal regulatory elements at the IFNG locus. *J Immunol* 188, 4 (2012), 1726–1733.
- [75] CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L., ZHANG, X., AND MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. *Genome biology* 17 (jan 2016), 13.
- [76] COSTA-SILVA, J., DOMINGUES, D., AND LOPES, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS one* 12, 12 (2017), e0190152.
- [77] COX, D. R. D. R., AND HINKLEY, D. V. *Theoretical statistics*. Chapman and Hall, 1974.
- [78] CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M., SHARP, P. A., BOYER, L. A., YOUNG, R. A., AND JAENISCH, R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* 107, 50 (2010), 21931–21936.

- [79] CROTTY, S. Follicular Helper CD4 T Cells (Tfh). *Annual Review of Immunology* 29, 1 (2011), 621–663.
- [80] CURRADI, M., IZZO, A., BADARACCO, G., AND LANDSBERGER, N. Molecular mechanisms of gene silencing mediated by DNA methylation. *Molecular and cellular biology* 22, 9 (may 2002), 3157–73.
- [81] DAY, N., HEMMAPLARDH, A., THURMAN, R. E., STAMATOYANNOPOULOS, J. A., AND NOBLE, W. S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* 23, 11 (jun 2007), 1424–1426.
- [82] DE DOMENICO, M., GRANELL, C., PORTER, M. A., AND ARENAS, A. Author Correction: The physics of spreading processes in multilayer networks. *Nature Physics* (2018), 1.
- [83] DE WIT, E., AND DE LAAT, W. A decade of 3C technologies: insights into nuclear organization. *Genes & development* 26, 1 (jan 2012), 11–24.
- [84] DEHMER, M., Ed. *Structural Analysis of Complex Networks*. Birkhäuser Boston, Boston, 2011.
- [85] DEHMER, M., AND EMMERT-STREIB, F. *Analysis of complex networks : from biology to linguistics*. Wiley-VCH, 2009.
- [86] DENG, W., LEE, J., WANG, H., MILLER, J., REIK, A., GREGORY, P., DEAN, A., AND BLOBEL, G. Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell* 149, 6 (jun 2012), 1233–1244.
- [87] DEUFLHARD, P., AND WEBER, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications* 398 (mar 2005), 161–184.
- [88] DIESTEL, R. *Graph Theory*, vol. 173 of *Graduate Texts in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.
- [89] DIXON, J. R., SELVARAJ, S., YUE, F., KIM, A., LI, Y., SHEN, Y., HU, M., LIU, J. S., AND REN, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 7398 (2012), 376–380.
- [90] DJURETIC, I. M., LEVANON, D., NEGREANU, V., GRONER, Y., RAO, A., AND ANSEL, K. M. Transcription factors T-bet and Runx3 cooperate to activate Ifng and silence Il4 in T helper type 1 cells. *Nature Immunology* 8, 2 (feb 2007), 145–153.
- [91] DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M., AND GINGERAS, T. R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 1 (2013), 15–21.
- [92] DOYLE, J. C., ALDERSON, D. L., LI, L., LOW, S., ROUGHAN, M., SHALUNOV, S., TANAKA, R., AND WILLINGER, W. The robust yet fragile nature of the Internet. *Proceedings of the National Academy of Sciences of the United States of America* 102, 41 (oct 2005), 14497–502.

- [93] DURBIN, B. P., HARDIN, J. S., HAWKINS, D. M., AND ROCKE, D. M. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics (Oxford, England) 18 Suppl 1* (2002), S105–10.
- [94] DURBIN, R., EDDY, S. R., KROGH, A., AND MITCHISON, G. *Biological sequence analysis*. Cambridge University Press, Cambridge, 1998.
- [95] EBBO, M., CRINIER, A., VÉLY, F., AND VIVIER, E. Innate lymphoid cells: major players in inflammatory diseases. *Nature Reviews Immunology 17*, 11 (aug 2017), 665–678.
- [96] EFRON, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics 7*, 1 (jan 1979), 1–26.
- [97] EIZENBERG-MAGAR, I., RIMER, J., ZARETSKY, I., LARA-ASTIASO, D., REICH-ZELIGER, S., AND FRIEDMAN, N. Diverse continuum of CD4+ T-cell states is determined by hierarchical additive integration of cytokine signals. *Proceedings of the National Academy of Sciences* (2017), 201615590.
- [98] EMMERT-STREIB, F., DEHMER, M., AND HAIBE-KAINS, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology 2*, August (2014), 1–7.
- [99] ENDO, Y., HIRAHARA, K., YAGI, R., TUMES, D. J., AND NAKAYAMA, T. Pathogenic memory type Th2 cells in allergic inflammation. *Trends in immunology 35*, 2 (feb 2014), 69–78.
- [100] ENDO, Y., IWAMURA, C., KUWAHARA, M., SUZUKI, A., SUGAYA, K., TUMES, D., TOKOYODA, K., HOSOKAWA, H., YAMASHITA, M., AND NAKAYAMA, T. Eomesodermin Controls Interleukin-5 Production in Memory T Helper 2 Cells through Inhibition of Activity of the Transcription Factor GATA3. *Immunity 35*, 5 (nov 2011), 733–745.
- [101] ENGSTRÖM, P. G., STEIJGER, T., SIPOS, B., GRANT, G. R., KAHLES, A., RÄTSCH, G., GOLDMAN, N., HUBBARD, T. J., HARROW, J., GUIGÓ, R., BERTONE, P., DAVIS, C. A., DOBIN, A., ENGSTRÖM, P. G., GINGERAS, T. R., GOLDMAN, N., GRANT, G. R., GUIGÓ, R., HARROW, J., HUBBARD, T. J., JEAN, G., KAHLES, A., KOSAREV, P., LI, S., LIU, J., MASON, C. E., MOLODTSOV, V., NING, Z., PONSTINGL, H., PRINS, J. F., RÄTSCH, G., RIBECA, P., SELEDTSOV, I., SIPOS, B., SOLOVYEV, V., STEIJGER, T., VALLE, G., VITULO, N., WANG, K., WU, T. D., ZELLER, G., RÄTSCH, G., GOLDMAN, N., HUBBARD, T. J., HARROW, J., GUIGÓ, R., AND BERTONE, P. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods 10*, 12 (dec 2013), 1185–1191.
- [102] ENRIGHT, A. J., VAN DONGEN, S., AND OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research 30*, 7 (apr 2002), 1575–84.
- [103] ERNST, J., AND KELLIS, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology 28*, 8 (2010), 817–825.

- [104] ERNST, J., AND KELLIS, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 9, 3 (2012), 215–216.
- [105] ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M., KU, M., DURHAM, T., KELLIS, M., AND BERNSTEIN, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 7345 (2011), 43–49.
- [106] ERNST, J., VAINAS, O., HARBISON, C. T., SIMON, I., AND BAR-JOSEPH, Z. Reconstructing dynamic regulatory maps. *Molecular systems biology* 3, 74 (2007), 74.
- [107] ESTRADA, E. *The structure of complex networks : theory and applications*. Oxford University Press, 2011.
- [108] ESTRADA, E., AND KNIGHT, P. A. *A first course in network theory*. Oxford University Press, 2015.
- [109] ESTRADA, J., WONG, F., DEPACE, A., AND GUNAWARDENA, J. Information Integration and Energy Expenditure in Gene Regulation. *Cell* 166, 1 (2016), 234–244.
- [110] EZKURDIA, I., JUAN, D., RODRIGUEZ, J. M., FRANKISH, A., DIEKHANS, M., HARROW, J., VAZQUEZ, J., VALENCIA, A., AND TRESS, M. L. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics* 23, 22 (nov 2014), 5866–5878.
- [111] FANG, M., XIE, H., DOUGAN, S. K., PLOEGH, H., AND VAN OUDENAARDEN, A. Stochastic Cytokine Expression Induces Mixed T Helper Cell States. *PLoS Biology* 11, 7 (2013).
- [112] FENG, S., SÁEZ, M., WIUF, C., FELIU, E., AND SOYER, O. S. Core signalling motif displaying multistability through multi-state enzymes. *Journal of The Royal Society Interface* 13, 123 (oct 2016), 20160524.
- [113] FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S., AND AMORIM, D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15 (2014), 3133–3181.
- [114] FIELDS, P. E., LEE, G. R., KIM, S. T., BARTSEVICH, V. V., AND FLAVELL, R. A. Th2-Specific Chromatin Remodeling and Enhancer Activity in the Th2 Cytokine Locus Control Region. *Immunity* 21, 6 (dec 2004), 865–876.
- [115] FILION, G. J., VAN BEMMEL, J. G., BRAUNSCHWEIG, U., TALHOUT, W., KIND, J., WARD, L. D., BRUGMAN, W., DE CASTRO, I. J., KERKHOVEN, R. M., BUSSEMAKER, H. J., AND VAN STEENSEL, B. Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* 143, 2 (2010), 212–224.
- [116] FORTUNATO, S. Community detection in graphs. *Physics Reports* 486, 3-5 (2010), 75–174.
- [117] FRIEDMAN, J. H., AND H., J. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (feb 2002), 367–378.

- [118] FUREY, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* 13, 12 (dec 2012), 840–852.
- [119] GAO, Z., CHEN, S., QIN, S., AND TANG, C. Network Motifs Capable of Decoding Transcription Factor Dynamics. *Scientific Reports* 8, 1 (2018), 1–10.
- [120] GARDINER, C. W. C. W., AND GARDINER, C. W. C. W. *Stochastic methods : a handbook for the natural and social sciences*. Springer, 2009.
- [121] GEEVEN, G., VAN KESTEREN, R. E., SMIT, A. B., AND DE GUNST, M. C. M. Identification of context-specific gene regulatory networks with GEMULA - gene expression modeling using LASSO. *Bioinformatics* 28, 2 (jan 2012), 214–221.
- [122] GEIER, F., TIMMER, J., AND FLECK, C. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC systems biology* 1 (feb 2007), 11.
- [123] GEIGER, D., AND HECKERMAN, D. Learning Gaussian Networks.
- [124] GEURTS, P., ERNST, D., AND WEHENKEL, L. Extremely randomized trees. *Machine Learning* 63, 1 (apr 2006), 3–42.
- [125] GHAHRAMANI, Z. An Introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15, 01 (feb 2001), 9–42.
- [126] GILL, R., DATTA, S., AND DATTA, S. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics* 11 (2010).
- [127] GIRVAN, M., AND NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, 12 (jun 2002), 7821–6.
- [128] GOODMAN, R. H., AND SMOLIK, S. CBP/p300 in cell growth, transformation, and development. *Genes & development* 14, 13 (jul 2000), 1553–77.
- [129] GRIESENAUER, B., AND PACZESNY, S. The ST2/IL-33 Axis in Immune Cells during Inflammatory Diseases. *Frontiers in immunology* 8 (2017), 475.
- [130] GUANTES, R., AND POYATOS, J. F. Multistable decision switches for flexible control of epigenetic differentiation. *PLoS Computational Biology* 4, 11 (2008).
- [131] GUÉRY, L., AND HUGUES, S. Th17 Cell Plasticity and Functions in Cancer Immunity. *BioMed research international* 2015 (2015), 314620.
- [132] GUO, L., JUNTILA, I. S., AND PAUL, W. E. Cytokine-induced cytokine production by conventional and innate lymphoid cells. *Trends in immunology* 33, 12 (dec 2012), 598–606.
- [133] HAIRER, M. Ergodic Properties of Markov Processes. *Lecture given at The University of Warwick (Spring 2006)*.

- [134] HAMADA, M., ONO, Y., FUJIMAKI, R., AND ASAI, K. Learning chromatin states with factorized information criteria. *Bioinformatics* 31, 15 (aug 2015), 2426–2433.
- [135] HAMALAINEN, H., ZHOU, H., CHOU, W., HASHIZUME, H., HELLER, R., AND LAHESMAA, R. Distinct gene expression profiles of human type 1 and type 2 T helper cells. *Genome biology* 2, 7 (2001).
- [136] HARENBERG, S., BELLO, G., GJELTEMA, L., RANSHOUS, S., HARLALKA, J., SEAY, R., PADMANABHAN, K., AND SAMATOVA, N. Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics* 6, 6 (nov 2014), 426–439.
- [137] HARRINGTON, L. E., HATTON, R. D., MANGAN, P. R., TURNER, H., MURPHY, T. L., MURPHY, K. M., AND WEAVER, C. T. Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nature Immunology* 6, 11 (2005), 1123–1132.
- [138] HARTEMINK, A. J., GIFFORD, D. K., JAAKKOLA, T. S., AND YOUNG, R. A. Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2002), 437–49.
- [139] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.
- [140] HAYASHI, Y., SENDA, T., SANO, N., AND HORIKOSHI, M. Theoretical framework for the histone modification network: modifications in the unstructured histone tails form a robust scale-free network. *Genes to Cells* 14, 7 (jul 2009), 789–806.
- [141] HECKER, M., LAMBECK, S., TOEPFER, S., VAN SOMEREN, E., AND GUTHKE, R. Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems* 96, 1 (2009), 86–103.
- [142] HEGAZY, A. N., PEINE, M., HELMSTETTER, C., PANSE, I., FRÖHLICH, A., BERGTHALER, A., FLATZ, L., PINSCHOWER, D. D., RADBRUCH, A., AND LÖHNING, M. Interferons Direct Th2 Cell Reprogramming to Generate a Stable GATA-3+T-bet+ Cell Subset with Combined Th2 and Th1 Cell Functions. *Immunity* 32, 1 (2010), 116–128.
- [143] HEINIG, M., COLOMÉ-TATCHÉ, M., TAUDI, A., RINTISCH, C., SCHAFFER, S., PRAVENEK, M., HUBNER, N., VINGRON, M., AND JOHANNES, F. histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC Bioinformatics* 16, 1 (dec 2015), 60.
- [144] HEINTZMAN, N. D., HON, G. C., HAWKINS, R. D., KHERADPOUR, P., STARK, A., HARP, L. F., YE, Z., LEE, L. K., STUART, R. K., CHING, C. W., CHING, K. A., ANTOSIEWICZ-BOURGET, J. E., LIU, H., ZHANG, X., GREEN, R. D., LOBANENKOV, V. V., STEWART, R., THOMSON, J. A., CRAWFORD, G. E., KELLIS, M., AND REN, B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 7243 (2009), 108–112.

- [145] HEINTZMAN, N. D., STUART, R. K., HON, G., FU, Y., CHING, C. W., HAWKINS, R. D., BARRERA, L. O., VAN CALCAR, S., QU, C., CHING, K. A., WANG, W., WENG, Z., GREEN, R. D., CRAWFORD, G. E., AND REN, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* 39, 3 (mar 2007), 311–318.
- [146] HEINZ, L., BAUMANN, C., KÖBERLIN, M., SNIJDER, B., GAWISH, R., SHUI, G., SHARIF, O., ASPALTER, I., MÜLLER, A., KANDASAMY, R., BREITWIESER, F., PICHLMAIR, A., BRUCKNER, M., REBSAMEN, M., BLÜML, S., KARONITSCH, T., FAUSTER, A., COLINGE, J., BENNETT, K., KNAPP, S., WENK, M., AND SUPERTI-FURGA, G. The Lipid-Modifying Enzyme SMPDL3B Negatively Regulates Innate Immunity. *Cell Reports* 11, 12 (jun 2015), 1919–1928.
- [147] HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H., AND GLASS, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38, 4 (may 2010), 576–589.
- [148] HEINZ, S., ROMANOSKI, C. E., BENNER, C., AND GLASS, C. K. The selection and function of cell type-specific enhancers. *Nature reviews. Molecular cell biology* 16, 3 (mar 2015), 144–54.
- [149] HENS, C., DANA, S. K., AND FEUDEL, U. Extreme multistability: Attractor manipulation and robustness. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25, 5 (may 2015), 053112.
- [150] HENS, C. R., BANERJEE, R., FEUDEL, U., AND DANA, S. K. How to obtain extreme multistability in coupled dynamical systems. *Physical Review E* 85, 3 (mar 2012), 035202.
- [151] HERTWECK, A., EVANS, C. M., ESKANDARPOUR, M., LAU, J. C., OLEINIK, K., JACKSON, I., KELLY, A., AMBROSE, J., ADAMSON, P., COUSINS, D. J., LAVENDER, P., CALDER, V. L., LORD, G. M., AND JENNER, R. G. T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Reports* 15, 12 (2016), 2756–2770.
- [152] HIRAHARA, K., POHOLEK, A., VAHEDI, G., LAURENCE, A., KANNO, Y., MILNER, J. D., AND O’SHEA, J. J. Mechanisms underlying helper T-cell plasticity: Implications for immune-mediated disease. *Journal of Allergy and Clinical Immunology* 131, 5 (2013), 1276–1287.
- [153] HIRAHARA, K., VAHEDI, G., GHORESCHI, K., YANG, X.-P., NAKAYAMADA, S., KANNO, Y., O’SHEA, J. J., AND LAURENCE, A. Helper T-cell differentiation and plasticity: insights from epigenetics. *Immunology* 134, 3 (nov 2011), 235–45.
- [154] HNISZ, D., ABRAHAM, B. J., LEE, T. I., LAU, A., SAINT-ANDRÉ, V., SIGOVA, A. A., HOKE, H. A., AND YOUNG, R. A. Super-enhancers in the control of cell identity and disease. *Cell* 155, 4 (2013), 934–47.
- [155] HNISZ, D., SCHUIJERS, J., LIN, C. Y., WEINTRAUB, A. S., ABRAHAM, B. J., LEE, T. I., BRADNER, J. E., AND YOUNG, R. A. Convergence of developmental and oncogenic

- signaling pathways at transcriptional super-enhancers. *Molecular cell* 58, 2 (apr 2015), 362–70.
- [156] HO, T. K. A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. *Pattern Analysis & Applications* 5, 2 (jun 2002), 102–112.
- [157] HOFFMAN, M. M., BUSKE, O. J., WANG, J., WENG, Z., BILMES, J. A., AND NOBLE, W. S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* 9, 5 (2012), 473–476.
- [158] HOFFMAN, M. M., ERNST, J., WILDER, S. P., KUNDAJE, A., HARRIS, R. S., LIBBRECHT, M., GIARDINE, B., ELLENBOGEN, P. M., BILMES, J. A., BIRNEY, E., HARDISON, R. C., DUNHAM, I., KELLIS, M., AND NOBLE, W. S. Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research* 41, 2 (jan 2013), 827–41.
- [159] HORI, S., NOMURA, T., AND SAKAGUCHI, S. Control of Regulatory T Cell Development by the Transcription Factor Foxp3. *Science* 299, 5609 (feb 2003), 1057–1061.
- [160] HUANG, S. Hybrid T-helper cells: stabilizing the moderate center in a polarized system. *PLoS biology* 11, 8 (2013), e1001632.
- [161] HUSIC, B. E., AND PANDE, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* 140, 7 (feb 2018), 2386–2396.
- [162] HWANG, E. S., SZABO, S. J., SCHWARTZBERG, P. L., AND GLIMCHER, L. H. T Helper Cell Fate Specified by Kinase-Mediated Interaction of T-bet with GATA-3. *Science* 307, 5708 (jan 2005), 430–433.
- [163] ING-SIMMONS, E., SEITAN, V. C., FAURE, A. J., FLICEK, P., CARROLL, T., DEKKER, J., FISHER, A. G., LENHARD, B., AND MERKENSCHLAGER, M. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome research* 25, 4 (apr 2015), 504–13.
- [164] IVANOV, I. I., MCKENZIE, B. S., ZHOU, L., TADOKORO, C. E., LEPALLEY, A., LAFAILLE, J. J., CUA, D. J., AND LITTMAN, D. R. The Orphan Nuclear Receptor ROR γ t Directs the Differentiation Program of Proinflammatory IL-17+ T Helper Cells. *Cell* 126, 6 (sep 2006), 1121–1133.
- [165] JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. *An Introduction to Statistical Learning*, vol. 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013.
- [166] JAVED, M. A., YOUNIS, M. S., LATIF, S., QADIR, J., AND BAIG, A. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications* 108, September 2017 (2018), 87–111.
- [167] JENNER, R. G., TOWNSEND, M. J., JACKSON, I., SUN, K., BOUWMAN, R. D., YOUNG, R. A., GLIMCHER, L. H., AND LORD, G. M. The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. *Proceedings of the National Academy of Sciences of the United States of America* 106, 42 (2009), 17876–17881.

- [168] JENSEN, S. T., CHEN, G., AND STOECKERT, JR., C. J. Bayesian variable selection and data integration for biological regulatory networks. *The Annals of Applied Statistics* 1, 2 (dec 2007), 612–633.
- [169] JIN, Q., YU, L.-R., WANG, L., ZHANG, Z., KASPER, L. H., LEE, J.-E., WANG, C., BRINDLE, P. K., DENT, S. Y. R., AND GE, K. Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *The EMBO Journal* 30, 2 (jan 2011), 249–262.
- [170] JOHANSSON, Å., LØSET, M., MUNDAL, S. B., JOHNSON, M. P., FREED, K. A., FENSTAD, M. H., MOSES, E. K., AUSTGULEN, R., AND BLANGERO, J. Partial correlation network analyses to detect altered gene interactions in human disease: using preeclampsia as a model. *Human Genetics* 129, 1 (jan 2011), 25–34.
- [171] JONES, E. A., AND FLAVELL, R. A. Distal enhancer elements transcribe intergenic RNA in the IL-10 family gene cluster. *Journal of immunology (Baltimore, Md. : 1950)* 175, 11 (dec 2005), 7437–46.
- [172] JORDAN, I. K., MARIÑO-RAMÍREZ, L., WOLF, Y. I., AND KOONIN, E. V. Conservation and Coevolution in the Scale-Free Human Gene Coexpression Network. *Molecular Biology and Evolution* 21, 11 (nov 2004), 2058–2070.
- [173] JOSEFOWICZ, S. Z., AND RUDENSKY, A. Control of Regulatory T Cell Lineage Commitment and Maintenance. *Immunity* 30, 5 (2009), 616–625.
- [174] JOST, D., CARRIVAIN, P., AND CAVALLI, G. Modeling epigenome folding : formation and dynamics of topologically associated chromatin domains.
- [175] JOST, D., VAILLANT, C., AND MEISTER, P. Coupling 1D modifications and 3D nuclear organization: data, models and function. *Current Opinion in Cell Biology* 44 (feb 2017), 20–27.
- [176] KADERALI, L., AND RADDE, N. Inferring Gene Regulatory Networks from Expression Data. *Computational Intelligence in Bioinformatics* 74, 2008 (2008), 33–74.
- [177] KAGEY, M. H., NEWMAN, J. J., BILODEAU, S., ZHAN, Y., ORLANDO, D. A., VAN BERKUM, N. L., EBMEIER, C. C., GOOSSENS, J., RAHL, P. B., LEVINE, S. S., TAATJES, D. J., DEKKER, J., AND YOUNG, R. A. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 7314 (sep 2010), 430–435.
- [178] KALKHOVEN, E. CBP and p300: HATs for different occasions. *Biochemical Pharmacology* 68, 6 (sep 2004), 1145–1155.
- [179] KANHERE, A., HERTWECK, A., BHATIA, U., GÖKMEN, M. R., PERUCHA, E., JACKSON, I., LORD, G. M., AND JENNER, R. G. T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nature communications* 3 (2012), 1268.
- [180] KANNO, Y., VAHEDI, G., HIRAHARA, K., SINGLETON, K., AND O’SHEA, J. J. Transcriptional and Epigenetic Control of T Helper Cell Specification: Molecular Mechanisms Underlying Commitment and Plasticity. *Annual Review of Immunology* 30, 1 (apr 2012), 707–731.

- [181] KAPP, J. A. Special regulatory T-cell review: Suppressors regulated but unexpressed. *Immunology* 123, 1 (jan 2008), 28–32.
- [182] KARLEBACH, G., AND SHAMIR, R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* 9, 10 (2008), 770–780.
- [183] KELSO, J. A. S. Multistability and metastability: understanding dynamic coordination in the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1591 (2012), 906–918.
- [184] KHAN, A., AND ZHANG, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Research* 44, D1 (jan 2016), D164–D171.
- [185] KHANIN, R., AND WIT, E. How Scale-Free Are Biological Networks. *Journal of Computational Biology* 13, 3 (apr 2006), 810–818.
- [186] KIVELA, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y., AND PORTER, M. A. Multilayer networks. *Journal of Complex Networks* 2, 3 (sep 2014), 203–271.
- [187] KLEFTOGIANNIS, D., KALNIS, P., AND BAJIC, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research* 43, 1 (jan 2015), e6–e6.
- [188] KLUS, S., NÜSKE, F., KOLTAI, P., WU, H., KEVREKIDIS, I., SCHÜTTE, C., AND NOÉ, F. Data-Driven Model Reduction and Transfer Operator Approximation. *Journal of Nonlinear Science* 28, 3 (jun 2018), 985–1010.
- [189] KOHU, K., OHMORI, H., WONG, W. F., ONDA, D., WAKOH, T., KON, S., YAMASHITA, M., NAKAYAMA, T., KUBO, M., AND SATAKE, M. The Runx3 Transcription Factor Augments Th1 and Down-Modulates Th2 Phenotypes by Interacting with and Attenuating GATA3. *The Journal of Immunology* 183, 12 (dec 2009), 7817–7824.
- [190] KOLLER, D., AND FRIEDMAN, N. *Probabilistic graphical models : principles and techniques*. MIT Press, 2009.
- [191] KOMMER, C., TUGENDHAT, T., AND WAHL, N. *Tutorium Physik fürs Nebenfach Übersetzt aus dem Unverständlichen*. Springer Spektrum, 2015.
- [192] KOTHAMACHU, V. B., FELIU, E., CARDELLI, L., AND SOYER, O. S. Unlimited multistability and Boolean logic in microbial signalling. *Journal of The Royal Society Interface* 12, 108 (jul 2015), 20150234.
- [193] KULAEVA, O. I., NIZOVITSEVA, E. V., POLIKANOV, Y. S., ULIANOV, S. V., AND STUDITSKY, V. M. Distant activation of transcription: mechanisms of enhancer action. *Molecular and cellular biology* 32, 24 (dec 2012), 4892–7.
- [194] KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M. J., AMIN, V., WHITAKER, J. W., SCHULTZ, M. D., WARD, L. D., SARKAR, A., QUON, G., SANDSTROM, R. S., EATON, M. L., WU, Y.-C., PFENNING, A. R., WANG, X., CLAUSNITZER, M., LIU, Y., COARFA, C., HARRIS, R. A., SHORESH, N., EPSTEIN, C. B., GJONESKA, E., LEUNG, D., XIE, W., HAWKINS, R. D., LISTER, R., HONG, C., GASCARD, P., MUNGALL, A. J.,

- MOORE, R., CHUAH, E., TAM, A., CANFIELD, T. K., HANSEN, R. S., KAUL, R., SABO, P. J., BANSAL, M. S., CARLES, A., DIXON, J. R., FARH, K.-H., FEIZI, S., KARLIC, R., KIM, A.-R., KULKARNI, A., LI, D., LOWDON, R., ELLIOTT, G., MERCER, T. R., NEPH, S. J., ONUCHIC, V., POLAK, P., RAJAGOPAL, N., RAY, P., SALLARI, R. C., SIEBENTHALL, K. T., SINNOTT-ARMSTRONG, N. A., STEVENS, M., THURMAN, R. E., WU, J., ZHANG, B., ZHOU, X., BEAUDET, A. E., BOYER, L. A., DE JAGER, P. L., FARNHAM, P. J., FISHER, S. J., HAUSSLER, D., JONES, S. J. M., LI, W., MARRA, M. A., MCMANUS, M. T., SUNYAEV, S., THOMSON, J. A., TLSTY, T. D., TSAI, L.-H., WANG, W., WATERLAND, R. A., ZHANG, M. Q., CHADWICK, L. H., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A., MILOSAVLJEVIC, A., REN, B., STAMATOYANNOPOULOS, J. A., WANG, T., KELLIS, M., AND KELLIS, M. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 7539 (feb 2015), 317–330.
- [195] LÄHDESMÄKI, H., SHMULEVICH, I., AND YLI-HARJA, O. On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning* 52, 1/2 (2003), 147–167.
- [196] LAM, K. Y., WESTRICK, Z. M., MÜLLER, C. L., CHRISTIAEN, L., AND BONNEAU, R. Fused Regression for Multi-source Gene Regulatory Network Inference. *PLOS Computational Biology* 12, 12 (dec 2016), e1005157.
- [197] LANG, A. H., LI, H., COLLINS, J. J., AND MEHTA, P. Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. *PLoS Computational Biology* 10, 8 (2014).
- [198] LANGFELDER, P., AND HORVATH, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 1 (dec 2008), 559.
- [199] LANGMEAD, B., TRAPNELL, C., POP, M., AND SALZBERG, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10, 3 (mar 2009), R25.
- [200] LANGVILLE, A. N., AND MEYER, C. D. A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Review* 47, 1 (jan 2005), 135–161.
- [201] LANGVILLE, A. N., AND MEYER, C. D. C. D. *Google's PageRank and beyond : the science of search engine rankings*. Princeton University Press, 2006.
- [202] LASSERRE, J., CHUNG, H.-R., AND VINGRON, M. Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks. *PLoS Computational Biology* 9, 9 (sep 2013), e1003168.
- [203] LAWYER, G. Understanding the influence of all nodes in a network. *Scientific Reports* 5, 1 (aug 2015), 8665.
- [204] LAZAREVIC, V., GLIMCHER, L. H., AND LORD, G. M. T-bet: a bridge between innate and adaptive immunity. *Nature reviews. Immunology* 13, 11 (2013), 777–89.
- [205] LE PHILLIP, P., BAHL, A., AND UNGAR, L. H. Using prior knowledge to improve genetic network reconstruction from microarray data. *In silico biology* 4, 3 (2004), 335–53.

- [206] LEE, G. R., FIELDS, P. E., AND FLAVELL, R. A. Regulation of IL-4 Gene Expression by Distal Regulatory Elements and GATA-3 at the Chromatin Level. *Immunity* 14, 4 (apr 2001), 447–459.
- [207] LEE, G. R., FIELDS, P. E., GRIFFIN, T. J., AND FLAVELL, R. A. Regulation of the Th2 Cytokine Locus by a Locus Control Region. *Immunity* 19, 1 (jul 2003), 145–153.
- [208] LELLI, K. M., SLATTERY, M., AND MANN, R. S. Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Annual Review of Genetics* 46, 1 (2012), 43–68.
- [209] LI, C., CESBRON, F., OEHLER, M., BRUNNER, M., AND HÖFER, T. Frequency Modulation of Transcriptional Bursting Enables Sensitive and Rapid Gene Regulation. *Cell Systems* 6, 4 (apr 2018), 409–423.e11.
- [210] LI, G., BRAUNSTEIN, L. A., BULDYREV, S. V., HAVLIN, S., AND STANLEY, H. E. Transport and percolation theory in weighted networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 75, 4 (2007), 1–4.
- [211] LI, Y., AND LI, J. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC Genomics* 13, Suppl 7 (2012), S27.
- [212] LICHTBLAU, Y., ZIMMERMANN, K., HALDEMANN, B., LENZE, D., HUMMEL, M., AND LESER, U. Comparative assessment of differential network analysis methods. *Briefings in Bioinformatics* 18, 5 (jul 2016), bbw061.
- [213] LINO, C. N. R., BARROS-MARTINS, J., OBERDÖRFER, L., WALZER, T., AND PRINZ, I. Eomes expression reports the progressive differentiation of IFN- γ -producing Th1-like gamma-delta T cells. *European Journal of Immunology* 47, 6 (jun 2017), 970–981.
- [214] LOVE, M. I., HUBER, W., AND ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15, 12 (dec 2014), 550.
- [215] LV, J., QIAO, H., LIU, H., WU, X., ZHU, J., SU, J., WANG, F., CUI, Y., AND ZHANG, Y. Discovering cooperative relationships of chromatin modifications in human T cells based on a proposed closeness measure. *PLoS ONE* 5, 12 (2010), 1–15.
- [216] MACKAY, D. J. C. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.
- [217] MALIN, J., ANIBA, M. R., AND HANNENHALLI, S. Enhancer networks revealed by correlated DNase hypersensitivity states of enhancers. *Nucleic Acids Research* 41, 14 (2013), 6828–6838.
- [218] MAMMANA, A., AND CHUNG, H. R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology* 16, 1 (2015), 1–12.

- [219] MARBACH, D., COSTELLO, J. C., KÜFFNER, R., VEGA, N. M., PRILL, R. J., CAMACHO, D. M., ALLISON, K. R., ADERHOLD, A., ALLISON, K. R., BONNEAU, R., CAMACHO, D. M., CHEN, Y., COLLINS, J. J., CORDERO, F., COSTELLO, J. C., CRANE, M., DONDELINGER, F., DRTON, M., ESPOSITO, R., FOYCEL, R., DE LA FUENTE, A., GERTHEISS, J., GEURTS, P., GREENFIELD, A., GRZEGORCZYK, M., HAURY, A.-C., HOLMES, B., HOTHORN, T., HUSMEIER, D., HUYNH-THU, V. A., IRRTHUM, A., KELLIS, M., KARLEBACH, G., KÜFFNER, R., LÈBRE, S., DE LEO, V., MADAR, A., MANI, S., MARBACH, D., MORDELET, F., OSTRER, H., OUYANG, Z., PANDYA, R., PETRI, T., PINNA, A., POULTNEY, C. S., PRILL, R. J., REZNY, S., RUSKIN, H. J., SAEYS, Y., SHAMIR, R., SÎRBU, A., SONG, M., SORANZO, N., STATNIKOV, A., STOLOVITZKY, G., VEGA, N., VERA-LICONA, P., VERT, J.-P., VISCONTI, A., WANG, H., WEHENKEL, L., WINDHAGER, L., ZHANG, Y., ZIMMER, R., KELLIS, M., COLLINS, J. J., AND STOLOVITZKY, G. Wisdom of crowds for robust gene network inference. *Nature Methods* 9, 8 (2012), 796–804.
- [220] MARCO, E., MEULEMAN, W., HUANG, J., GLASS, K., PINELLO, L., WANG, J., KELLIS, M., AND YUAN, G.-C. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. *Nature Communications* 8 (apr 2017), 15011.
- [221] MARIANI, L., LÖHNING, M., RADBRUCH, A., AND HÖFER, T. Transcriptional control networks of cell differentiation: Insights from helper T lymphocytes. *Progress in Biophysics and Molecular Biology* 86, 1 (2004), 45–76.
- [222] MASUDA, N., PORTER, M. A., AND LAMBIOTTE, R. Random walks and diffusion on networks. *Physics Reports* 716-717 (2017), 1–58.
- [223] MEYER, C. D. C. D. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, 2000.
- [224] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D., AND ALON, U. Network motifs: simple building blocks of complex networks. *Science (New York, N.Y.)* 298, 5594 (oct 2002), 824–7.
- [225] MODENCODE CONSORTIUM, ROY, S., ERNST, J., KHARCHENKO, P. V., KHERADPOUR, P., NEGRE, N., EATON, M. L., LANDOLIN, J. M., BRISTOW, C. A., MA, L., LIN, M. F., WASHIETL, S., ARSHINOFF, B. I., AY, F., MEYER, P. E., ROBINE, N., WASHINGTON, N. L., DI STEFANO, L., BEREZIKOV, E., BROWN, C. D., CANDEIAS, R., CARLSON, J. W., CARR, A., JUNGREIS, I., MARBACH, D., SEALFON, R., TOLSTORUKOV, M. Y., WILL, S., ALEKSEYENKO, A. A., ARTIERI, C., BOOTH, B. W., BROOKS, A. N., DAI, Q., DAVIS, C. A., DUFF, M. O., FENG, X., GORCHAKOV, A. A., GU, T., HENIKOFF, J. G., KAPRANOV, P., LI, R., MACALPINE, H. K., MALONE, J., MINODA, A., NORDMAN, J., OKAMURA, K., PERRY, M., POWELL, S. K., RIDDLE, N. C., SAKAI, A., SAMSONOVA, A., SANDLER, J. E., SCHWARTZ, Y. B., SHER, N., SPOKONY, R., STURGILL, D., VAN BAREN, M., WAN, K. H., YANG, L., YU, C., FEINGOLD, E., GOOD, P., GUYER, M., LOWDON, R., AHMAD, K., ANDREWS, J., BERGER, B., BRENNER, S. E., BRENT, M. R., CHERBAS, L., ELGIN, S. C. R., GINGERAS, T. R., GROSSMAN, R., HOSKINS, R. A., KAUFMAN, T. C., KENT, W., KURODA, M. I., ORR-WEAVER, T., PERRIMON, N., PIRROTTA, V., POSAKONY, J. W., REN, B., RUSSELL, S., CHERBAS, P., GRAVELEY, B. R., LEWIS, S., MICKLEM, G., OLIVER, B., PARK, P. J., CELNIKER, S. E., HENIKOFF, S., KARPEN, G. H., LAI, E. C., MACALPINE, D. M., STEIN, L. D., WHITE, K. P., KELLIS, M., ACEVEDO, D., AUBURN, R., BARBER,

- G., BELLEN, H. J., BISHOP, E. P., BRYSON, T. D., CHATEIGNER, A., CHEN, J., CLAWSON, H., COMSTOCK, C. L. G., CONTRINO, S., DENAPOLI, L. C., DING, Q., DOBIN, A., DOMANUS, M. H., DRENKOW, J., DUDOIT, S., DUMAIS, J., ENG, T., FAGEGALTIER, D., GADEL, S. E., GHOSH, S., GUILLIER, F., HANLEY, D., HANNON, G. J., HANSEN, K. D., HEINZ, E., HINRICHS, A. S., HIRST, M., JHA, S., JIANG, L., JUNG, Y. L., KASHEVSKY, H., KENNEDY, C. D., KEPHART, E. T., LANGTON, L., LEE, O.-K., LI, S., LI, Z., LIN, W., LINDER-BASSO, D., LLOYD, P., LYNE, R., MARCHETTI, S. E., MARRA, M., MATTIUZZO, N. R., MCKAY, S., MEYER, F., MILLER, D., MILLER, S. W., MOORE, R. A., MORRISON, C. A., PRINZ, J. A., ROOKS, M., MOORE, R., RUTHERFORD, K. M., RUZANOV, P., SCHEFTNER, D. A., SENDEROWICZ, L., SHAH, P. K., SHANOWER, G., SMITH, R., STINSON, E. O., SUCHY, S., TENNEY, A. E., TIAN, F., VENKEN, K. J. T., WANG, H., WHITE, R., WILKENING, J., WILLINGHAM, A. T., ZALESKI, C., ZHA, Z., ZHANG, D., ZHAO, Y., AND ZIEBA, J. Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE. *Science* 330, 6012 (dec 2010), 1787–1797.
- [226] MOORTHY, S. D., DAVIDSON, S., SHCHUKA, V. M., SINGH, G., MALEK-GILANI, N., LANGROUDI, L., MARTCHENKO, A., SO, V., MACPHERSON, N. N., AND MITCHELL, J. A. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Research* 27, 2 (feb 2017), 246–258.
- [227] MOSMANN, T. R., CHERWINSKI, H., BOND, M. W., GIEDLIN, M. A., AND COFFMAN, R. L. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *Journal of immunology (Baltimore, Md. : 1950)* 136, 7 (1986), 2348–57.
- [228] MUCHA, P. J., RICHARDSON, T., MACON, K., PORTER, M. A., AND ONNELA, J.-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science (New York, N.Y.)* 328, 5980 (may 2010), 876–8.
- [229] MURPHY, K. *Janeway's Immunobiology*. Garland Science, 2017.
- [230] MURPHY, K. M., AND REINER, S. L. The lineage decisions of helper T cells. *Nature Reviews Immunology* 2, 12 (2002), 933–944.
- [231] MURPHY, K. P. *Machine learning : a probabilistic perspective*. MIT Press, 2012.
- [232] NAKATO, R., AND SHIRAHIGE, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in bioinformatics* 18, 2 (2017), 279–290.
- [233] NAKAYAMADA, S., KANNO, Y., TAKAHASHI, H., JANKOVIC, D., LU, K., JOHNSON, T., SUN, H.-W., VAHEDI, G., HAKIM, O., HANDON, R., SCHWARTZBERG, P., HAGER, G., AND O'SHEA, J. Early Th1 Cell Differentiation Is Marked by a Tfh Cell-like Transition. *Immunity* 35, 6 (dec 2011), 919–931.
- [234] NAKAYAMADA, S., TAKAHASHI, H., KANNO, Y., AND O'SHEA, J. J. Helper T cell diversity and plasticity. *Current Opinion in Immunology* 24, 3 (2012), 297–302.
- [235] NEWMAN, M. E. J. Analysis of weighted networks. *Physical Review E* 70, 5 (nov 2004), 056131.

- [236] NEWMAN, M. E. J. *Networks : An Introduction*. Oxford University Press, 2010.
- [237] NGUYEN, M. L. T., JONES, S. A., PRIER, J. E., AND RUSS, B. E. Transcriptional Enhancers in the Regulation of T Cell Differentiation. *Frontiers in Immunology* 6, September (2015).
- [238] NI, Y., STINGO, F. C., AND BALADANDAYUTHAPANI, V. Bayesian nonlinear model selection for gene regulatory networks. *Biometrics* 71, 3 (sep 2015), 585–95.
- [239] NIEBUR, E., SCHUSTER, H. G., AND KAMMEN, D. M. Collective frequencies and metastability in networks of limit-cycle oscillators with time delay. *Physical Review Letters* 67, 20 (nov 1991), 2753–2756.
- [240] NORA, E. P., LAJOIE, B. R., SCHULZ, E. G., GIORGETTI, L., OKAMOTO, I., SERVANT, N., PIOLOT, T., VAN BERKUM, N. L., MEISIG, J., SEDAT, J., GRIBNAU, J., BARILLOT, E., BLÜTHGEN, N., DEKKER, J., AND HEARD, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 7398 (apr 2012), 381–385.
- [241] NORRIS, J. R. J. R. *Markov chains*. Cambridge University Press, 1998.
- [242] OESTREICH, K. J., AND WEINMANN, A. S. Master regulators or lineage-specifying? Changing views on CD4+ T cell transcription factors. *Nature reviews. Immunology* 12, 11 (2012), 799–804.
- [243] OKA, H., EMORI, Y., HAYASHI, Y., AND NOMOTO, K. Breakdown of Th Cell Immune Responses and Steroidogenic CYP11A1 Expression in CD4+ T Cells in a Murine Model Implanted with B16 Melanoma. *Cellular Immunology* 206, 1 (nov 2000), 7–15.
- [244] OLTVAI, Z. N., BARABÁSI, A.-L., JEONG, H., TOMBOR, B., AND ALBERT, R. The large-scale organization of metabolic networks. *Nature* 407, 6804 (oct 2000), 651–654.
- [245] OMRANIAN, N., ELOUNDOU-MBEBI, J. M. O., MUELLER-ROEBER, B., AND NIKOLOSKI, Z. Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific Reports* 6, 1 (apr 2016), 20533.
- [246] ONG, C.-T., AND CORCES, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics* 12, 4 (2011), 283–293.
- [247] O’SHEA, J. J., LAHESMAA, R., VAHEDI, G., LAURENCE, A., AND KANNO, Y. Genomic views of STAT function in CD4+ T helper cell differentiation. *Nature Reviews Immunology* 11, 4 (apr 2011), 239–250.
- [248] O’SHEA, J. J., AND PAUL, W. E. Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. *Science (New York, N.Y.)* 327, 5969 (feb 2010), 1098–102.
- [249] PADI, M., AND QUACKENBUSH, J. Detecting phenotype-driven transitions in regulatory network structure. *npj Systems Biology and Applications* 4, 1 (2018), 16.

- [250] PAN, W., WANG, Z., GAO, H., LI, Y., AND DU, M. On multistability of delayed genetic regulatory networks with multivariable regulation functions. *Mathematical Biosciences* 228, 1 (2010), 100–109.
- [251] PAN, W., WANG, Z., GAO, H., AND LIU, X. Monostability and multistability of genetic regulatory networks with different types of regulation functions. *Non-linear Analysis: Real World Applications* 11, 4 (2010), 3170–3185.
- [252] PARK, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics* 10, 10 (2009), 669–680.
- [253] PASINI, D., MALATESTA, M., JUNG, H. R., WALFRIDSSON, J., WILLER, A., OLSSON, L., SKOTTE, J., WUTZ, A., PORSE, B., JENSEN, O. N., AND HELIN, K. Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. *Nucleic Acids Research* 38, 15 (aug 2010), 4958–4969.
- [254] PECCOUD, J., AND YCART, B. Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology* 48, 2 (oct 1995), 222–234.
- [255] PEINE, M., RAUSCH, S., HELMSTETTER, C., FRÖHLICH, A., HEGAZY, A. N., KÜHL, A. A., GREVELDING, C. G., HÖFER, T., HARTMANN, S., AND LÖHNING, M. Stable Tbet(+)/GATA-3(+)/Th1/Th2 hybrid cells arise in vivo, can develop directly from naive precursors, and limit immunopathologic inflammation. *PLoS biology* 11, 8 (2013), e1001633.
- [256] PELLET, E., PEINE, M., HELMSTETTER, C., FLOSSDORF, M., LÖHNING, M., AND HÖFER, T. Systematic inference of regulatory networks that drive cytokine stimulus integration by T cells. *In preparation*.
- [257] PENNACCHIO, L. A., BICKMORE, W., DEAN, A., NOBREGA, M. A., AND BEJERANO, G. Enhancers: five essential questions. *Nature Reviews Genetics* 14, 4 (apr 2013), 288–295.
- [258] PERNER, J., LASSERRE, J., KINKLEY, S., VINGRON, M., AND CHUNG, H.-R. Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Nucleic Acids Research* 42, 22 (2014), 13689–13695.
- [259] PERRIN, B.-E., RALAIVOLA, L., MAZURIE, A., BOTTANI, S., MALLET, J., AND D’ALCHÉ-BUC, F. Gene networks inference using dynamic Bayesian networks. *Bioinformatics (Oxford, England)* 19 Suppl 2 (oct 2003), ii138–48.
- [260] PETERS, A., LEE, Y., AND KUCHROO, V. K. The many faces of Th17 cells. *Current Opinion in Immunology* 23, 6 (dec 2011), 702–706.
- [261] PLACEK, K., COFFRE, M., MAIELLA, S., BIANCHI, E., AND ROGGE, L. Genetic and epigenetic networks controlling T helper 1 cell differentiation. *Immunology* 127, 2 (jun 2009), 155–62.

- [262] PLANK, M. W., KAIKO, G. E., MALTBY, S., WEAVER, J., TAY, H. L., SHEN, W., WILSON, M. S., DURUM, S. K., AND FOSTER, P. S. Th22 Cells Form a Distinct Th Lineage from Th17 Cells In Vitro with Unique Transcriptional Properties and Tbet-Dependent Th1 Plasticity. *Journal of immunology (Baltimore, Md. : 1950)* 198, 5 (2017), 2182–2190.
- [263] POMBO, A., AND DILLON, N. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology* 16, 4 (apr 2015), 245–257.
- [264] POTT, S., AND LIEB, J. D. What are super-enhancers ? *Nature Publishing Group* 47, 1 (2015), 8–12.
- [265] PRINZ, J.-H., WU, H., SARICH, M., KELLER, B., SENNE, M., HELD, M., CHODERA, J. D., SCHÜTTE, C., AND NOÉ, F. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics* 134, 17 (may 2011), 174105.
- [266] PROBST-KEPPER, M., AND BUER, J. FOXP3 and GARP (LRRC32): the master and its minion. *Biology direct* 5 (feb 2010), 8.
- [267] RADA-IGLESIAS, A., BAJPAI, R., SWIGUT, T., BRUGMANN, S. A., FLYNN, R. A., AND WYSOCKA, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470, 7333 (feb 2011), 279–283.
- [268] RADICCHI, F. Predicting percolation thresholds in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 91, 1 (2015), 1–5.
- [269] RAILEANU, L. E., AND STOFFEL, K. Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence* 41, 1 (may 2004), 77–93.
- [270] RAPHAEL, I., NALAWADE, S., EAGAR, T. N., AND FORSTHUBER, T. G. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. *Cytokine* 74, 1 (2015), 5–17.
- [271] RAZA, K., AND ALAM, M. Recurrent neural network based hybrid model for reconstructing gene regulatory network. *Computational Biology and Chemistry* 64 (oct 2016), 322–334.
- [272] REVERTER, A., AND CHAN, E. K. F. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24, 21 (nov 2008), 2491–2497.
- [273] REYNOLDS, N., O'SHAUGHNESSY, A., AND HENDRICH, B. Transcriptional repressors: multifaceted regulators of gene expression. *Development* 140, 3 (feb 2013), 505–512.
- [274] RHIE, S. K., GUO, Y., TAK, Y. G., YAO, L., SHEN, H., COETZEE, G. A., LAIRD, P. W., AND FARNHAM, P. J. Identification of activated enhancers and linked transcription factors in breast, prostate, and kidney tumors by tracing enhancer networks using epigenetic traits. *Epigenetics and Chromatin* 9, 1 (2016), 1–17.

- [275] RON, G., GLOBERSON, Y., MORAN, D., AND KAPLAN, T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nature Communications* 8, 1 (dec 2017), 2237.
- [276] ROY, A., MACKIN, P. D., AND MUKHOPADHYAY, S. Methods for pattern selection , class-specific feature selection and classification for automated learning. *Neural Networks* 41 (2013), 113–129.
- [277] RUDAN, M. V., BARRINGTON, C., TANAY, A., RUDAN, M. V., BARRINGTON, C., HENDERSON, S., ERNST, C., ODOM, D. T., AND TANAY, A. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture Article Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *CellReports* 10, 8 (2015), 1297–1309.
- [278] SALLEH, F. H. M., ZAINUDIN, S., AND ARIF, S. M. Multiple Linear Regression for Reconstruction of Gene Regulatory Networks in Solving Cascade Error Problems. *Advances in Bioinformatics 2017* (2017), 1–14.
- [279] SALTER-TOWNSHEND, M., WHITE, A., GOLLINI, I., AND MURPHY, T. B. Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining* 5, 4 (aug 2012), 243–264.
- [280] SARAIVA, M., AND O’GARRA, A. The regulation of IL-10 production by immune cells. *Nature Reviews Immunology* 10, 3 (mar 2010), 170–181.
- [281] SAWICKI, J., OMELCHENKO, I., ZAKHAROVA, A., AND SCHÖLL, E. Chimera states in complex networks: interplay of fractal topology and delay. *European Physical Journal: Special Topics* 226, 9 (2017), 1883–1892.
- [282] SCHMIDL, C., RENNER, K., PETER, K., EDER, R., LASSMANN, T., BALWIERZ, P. J., ITOH, M., NAGAO-SATO, S., KAWAJI, H., CARNINCI, P., SUZUKI, H., HAYASHIZAKI, Y., ANDREESSEN, R., HUME, D. A., HOFFMANN, P., FORREST, A. R. R., KREUTZ, M. P., EDINGER, M., REHLI, M., AND CONSORTIUM, F. Transcription and enhancer profiling in human monocyte subsets. *Blood* 123, 17 (apr 2014), e90–e99.
- [283] SCHMIDT, D., WILSON, M. D., BALLESTER, B., SCHWALIE, P. C., BROWN, G. D., MARSHALL, A., KUTTER, C., WATT, S., MARTINEZ-JIMENEZ, C. P., MACKAY, S., TALIANIDIS, I., FLICEK, P., AND ODOM, D. T. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)* 328, 5981 (may 2010), 1036–40.
- [284] SCHMITT, E., KLEIN, M., AND BOPP, T. Th9 cells, new players in adaptive immunity. *Trends in Immunology* 35, 2 (feb 2014), 61–68.
- [285] SEGAL, E., SHAPIRA, M., REGEV, A., PE’ER, D., BOTSTEIN, D., KOLLER, D., AND FRIEDMAN, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* 34, 2 (jun 2003), 166–176.
- [286] SEYED-ALLAEI, H., BIANCONI, G., AND MARSILI, M. Scale-free networks with an exponent less than two. *Physical Review E* 73, 4 (apr 2006), 046113.

- [287] SHANAHAN, M. Metastable chimera states in community-structured oscillator networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20, 1 (mar 2010), 013108.
- [288] SHIH, H. Y., SCIUMÈ, G., POHOLEK, A. C., VAHEDI, G., HIRAHARA, K., VILLARINO, A. V., BONELLI, M., BOSSELUT, R., KANNO, Y., MULJO, S. A., AND O'SHEA, J. J. Transcriptional and epigenetic networks of helper T and innate lymphoid cells. *Immunological Reviews* 261, 1 (2014), 23–49.
- [289] SHLYUEVA, D., STAMPFEL, G., AND STARK, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* 15, 4 (apr 2014), 272–286.
- [290] ŠIKIĆ, M., LANČIĆ, A., ANTULOV-FANTULIN, N., AND ŠTEFANČIĆ, H. Epidemic centrality - is there an underestimated epidemic impact of network peripheral nodes? *The European Physical Journal B* 86, 10 (oct 2013), 440.
- [291] SNEPPEN, K., AND MITARAI, N. Multistability with a metastable mixed state. *Physical Review Letters* 109, 10 (2012), 1–5.
- [292] SPILIANAKIS, C. G., LALIOTI, M. D., TOWN, T., LEE, G. R., AND FLAVELL, R. A. Interchromosomal associations between alternatively expressed loci. *Nature* 435, 7042 (jun 2005), 637–645.
- [293] SPITZ, F., AND FURLONG, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* 13, 9 (sep 2012), 613–626.
- [294] STEINHAUSER, S., KURZAWA, N., EILS, R., AND HERRMANN, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in bioinformatics* 17, 6 (2016), 953–966.
- [295] STUART, J. M., SEGAL, E., KOLLER, D., AND KIM, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)* 302, 5643 (oct 2003), 249–55.
- [296] SU, G., KUCHINSKY, A., MORRIS, J. H., STATES, D. J., AND MENG, F. GLayer: community structure analysis of biological networks. *Bioinformatics* 26, 24 (dec 2010), 3135–3137.
- [297] SUTO, A., KASHIWAKUMA, D., KAGAMI, S.-I., HIROSE, K., WATANABE, N., YOKOTE, K., SAITO, Y., NAKAYAMA, T., GRUSBY, M. J., IWAMOTO, I., AND NAKAJIMA, H. Development and characterization of IL-21-producing CD4⁺ T cells. *The Journal of Experimental Medicine* 205, 6 (jun 2008), 1369–1379.
- [298] SZABO, S. J., KIM, S. T., COSTA, G. L., ZHANG, X., FATHMAN, C. G., AND GLIMCHER, L. H. A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* 100, 6 (2000), 655–669.
- [299] SZABO, S. J., SULLIVAN, B. M., PENG, S. L., AND GLIMCHER, L. H. Molecular Mechanisms Regulating Th1 Immune Responses. *Annual Review of Immunology* 21, 1 (apr 2003), 713–758.

- [300] SZABO, S. J., SULLIVAN, B. M., STEMMANN, C., SATOSKAR, A. R., SLECKMAN, B. P., AND GLIMCHER, L. H. Distinct effects of T-bet in TH1 lineage commitment and IFN-gamma production in CD4 and CD8 T cells. *Science (New York, N.Y.)* 295, 5553 (jan 2002), 338–42.
- [301] TAMADA, Y., KIM, S., BANNAI, H., IMOTO, S., TASHIRO, K., KUHARA, S., AND MIYANO, S. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics (Oxford, England)* 19 Suppl 2 (oct 2003), ii227–36.
- [302] TANAKA, S., TSUKADA, J., SUZUKI, W., HAYASHI, K., TANIGAKI, K., TSUJI, M., INOUE, H., HONJO, T., AND KUBO, M. The Interleukin-4 Enhancer CNS-2 Is Regulated by Notch Signals and Controls Initial Expression in NKT Cells and Memory-Type CD4 T Cells. *Immunity* 24, 6 (jun 2006), 689–701.
- [303] THE ENCODE PROJECT CONSORTIUM. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 7414 (sep 2012), 57–74.
- [304] THOMAS, J. M. Sturm’s Theorem for Multiple Roots. *National Mathematics Magazine* 15, 8 (may 1941), 391.
- [305] THOMAS, R., THOMAS, S., HOLLOWAY, A. K., AND POLLARD, K. S. Features that define the best ChIP-seq peak calling algorithms. *Briefings in bioinformatics* 18, 3 (2017), 441–450.
- [306] THOMPSON, D., REGEV, A., AND ROY, S. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. *Annual Review of Cell and Developmental Biology* 31, 1 (2015), 399–428.
- [307] THOMSON, M., AND GUNAWARDENA, J. Unlimited multistability in multisite phosphorylation systems. *Nature* 460, 7252 (2009), 274–277.
- [308] TIMÁR, G., DOROGOVITSEV, S. N., AND MENDES, J. F. F. Scale-free networks with exponent one. *Physical Review E* 94, 2 (aug 2016), 022302.
- [309] TRIPATHI, S. K., AND LAHESMAA, R. Transcriptional and epigenetic regulation of T-helper lineage specification. *Immunological reviews* 261, 1 (sep 2014), 62–83.
- [310] ULLNER, E., KOSKESKA, A., KURTHS, J., VOLKOV, E., KANTZ, H., AND GARCÍA-OJALVO, J. Multistability of synthetic genetic networks with repressive cell-to-cell communication. *Physical Review E* 78, 3 (sep 2008), 031904.
- [311] UMETSU, D. T., AND DEKRUYFF, R. H. The regulation of allergy and asthma. *Immunological Reviews* 212, 1 (aug 2006), 238–255.
- [312] USUI, T., NISHIKOMORI, R., KITANI, A., AND STROBER, W. GATA-3 suppresses Th1 development by downregulation of Stat4 and not through effects on IL-12Rbeta2 chain or T-bet. *Immunity* 18, 3 (mar 2003), 415–28.
- [313] VAHEDI, G., KANNO, Y., FURUMOTO, Y., JIANG, K., PARKER, S. C. J., ERDOS, M. R., DAVIS, S. R., ROYCHOUDHURI, R., RESTIFO, N. P., GADINA, M., TANG, Z., RUAN, Y., COLLINS, F. S., SARTORELLI, V., AND O’SHEA, J. J. Super-enhancers delineate

- disease-associated regulatory nodes in T cells. *Nature* 520, 7548 (apr 2015), 558–562.
- [314] VAHEDI, G., TAKAHASHI, H., NAKAYAMADA, S., SUN, H.-W., SARTORELLI, V., KANNO, Y., AND O'SHEA, J. J. STATs shape the active enhancer landscape of T cell populations. *Cell* 151, 5 (nov 2012), 981–93.
- [315] VAN ARENSBERGEN, J., VAN STEENSEL, B., AND BUSSEMAKER, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends in cell biology* 24, 11 (nov 2014), 695–702.
- [316] VAN BERLO, R. J. P., VAN SOMEREN, E. P., AND REINDERS, M. J. T. Studying the Conditions for Learning Dynamic Bayesian Networks to Discover Genetic Regulatory Networks. *Simulation* 79, 12 (dec 2003), 689–702.
- [317] VAN SOMEREN, E. P., WESSELS, L. F., AND REINDERS, M. J. Linear modeling of genetic networks from experimental data. *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 8 (2000), 355–66.
- [318] VASTENHOEW, N. L., AND SCHIER, A. F. Bivalent histone modifications in early embryogenesis. *Current Opinion in Cell Biology* 24, 3 (jun 2012), 374–386.
- [319] VILLA-VIALANEIX, N., LIAUBET, L., LAURENT, T., CHEREL, P., GAMOT, A., AND SANCRISTOBAL, M. The structure of a gene co-expression network reveals biological functions underlying eQTLs. *PloS one* 8, 4 (2013), e60045.
- [320] VISEL, A., BLOW, M. J., LI, Z., ZHANG, T., AKIYAMA, J. A., HOLT, A., PLAJZER-FRICK, I., SHOUKRY, M., WRIGHT, C., CHEN, F., AFZAL, V., REN, B., RUBIN, E. M., AND PENNACCHIO, L. A. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 7231 (feb 2009), 854–8.
- [321] VON LUXBURG, U. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (dec 2007), 395–416.
- [322] WALKER, J. A., AND MCKENZIE, A. N. J. TH2 cell development and function. *Nature Reviews Immunology* 18, 2 (oct 2017), 121–133.
- [323] WANG, J., CHEN, B., WANG, Y., WANG, N., GARBEY, M., TRAN-SON-TAY, R., BERCELLI, S. A., AND WU, R. Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information. *Nucleic acids research* 41, 8 (apr 2013), e97.
- [324] WANG, J., CHEUNG, L. W.-K., AND DELABIE, J. New probabilistic graphical models for genetic regulatory networks studies. *Journal of Biomedical Informatics* 38, 6 (dec 2005), 443–455.
- [325] WANG, P., SONG, C., ZHANG, H., WU, Z., TIAN, X.-J., AND XING, J. Epigenetic state network approach for describing cell phenotypic transitions. *Interface Focus* 4, 3 (apr 2014), 20130068–20130068.

- [326] WANG, Y., JOSHI, T., ZHANG, X.-S., XU, D., AND CHEN, L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* 22, 19 (oct 2006), 2413–2420.
- [327] WANG, Y.-M., ZHOU, P., WANG, L.-Y., LI, Z.-H., ZHANG, Y.-N., AND ZHANG, Y.-X. Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PloS one* 7, 8 (2012), e42414.
- [328] WANG, Z., GERSTEIN, M., AND SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10, 1 (jan 2009), 57–63.
- [329] WANG, Z., ZANG, C., ROSENFELD, J. A., SCHONES, D. E., BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T.-Y., PENG, W., ZHANG, M. Q., AND ZHAO, K. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics* 40, 7 (jul 2008), 897–903.
- [330] WASSERMAN, L. *All of statistics : a concise course in statistical inference*. Springer, 2004.
- [331] WASSERMAN, L. *All of nonparametric statistics*. Springer, 2006.
- [332] WATTS, D. J., AND STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. *Nature* 393, 6684 (jun 1998), 440–442.
- [333] WEEDON, M. N., CEBOLA, I., PATCH, A.-M., FLANAGAN, S. E., DE FRANCO, E., CASWELL, R., RODRÍGUEZ-SEGÚI, S. A., SHAW-SMITH, C., CHO, C. H.-H., ALLEN, H. L., HOUGHTON, J. A. L., ROTH, C. L., CHEN, R., HUSSAIN, K., MARSH, P., VAL- LIER, L., MURRAY, A., ELLARD, S., FERRER, J., HATTERSLEY, A. T., AND HATTERSLEY, A. T. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature Genetics* 46, 1 (jan 2014), 61–64.
- [334] WEI, G., ABRAHAM, B., YAGI, R., JOTHI, R., CUI, K., SHARMA, S., NARLIKAR, L., NORTHRUP, D., TANG, Q., PAUL, W., ZHU, J., AND ZHAO, K. Genome-wide Analyses of Transcription Factor GATA3-Mediated Gene Regulation in Distinct T Cell Types. *Immunity* 35, 2 (aug 2011), 299–311.
- [335] WEI, L., VAHEDI, G., SUN, H.-W., WATFORD, W. T., TAKATORI, H., RAMOS, H. L., TAKAHASHI, H., LIANG, J., GUTIERREZ-CRUZ, G., ZANG, C., PENG, W., O’SHEA, J. J., AND KANNO, Y. Discrete Roles of STAT4 and STAT6 Transcription Factors in Tuning Epigenetic Modifications and Transcription during T Helper Cell Differentiation. *Immunity* 32, 6 (jun 2010), 840–851.
- [336] WERHLI, A. V., GRZEGORCZYK, M., AND HUSMEIER, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics* 22, 20 (oct 2006), 2523–2531.
- [337] WERNICKE, S., AND RASCHE, F. FANMOD: a tool for fast network motif detection. *Bioinformatics* 22, 9 (may 2006), 1152–1153.

- [338] WHITAKER, J. W., NGUYEN, T. T., ZHU, Y., WILDBERG, A., AND WANG, W. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. *Methods* 72 (jan 2015), 86–94.
- [339] WHITE, J. S. Tables of Normal Percentile Points. *Journal of the American Statistical Association* 65, 330 (jun 1970), 635–638.
- [340] WHITLEY, D., AND WATSON, J. P. Complexity Theory and the No Free Lunch Theorem. In *Search Methodologies*. Springer US, Boston, MA, 2005, pp. 317–339.
- [341] WHYTE, W. A., ORLANDO, D. A., HNISZ, D., ABRAHAM, B. J., LIN, C. Y., KAGEY, M. H., RAHL, P. B., LEE, T. I., AND YOUNG, R. A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 2 (apr 2013), 307–19.
- [342] WIENCH, M., JOHN, S., BAEK, S., JOHNSON, T. A., SUNG, M.-H., ESCOBAR, T., SIMMONS, C. A., PEARCE, K. H., BIDDIE, S. C., SABO, P. J., THURMAN, R. E., STAMATOYANNOPOULOS, J. A., AND HAGER, G. L. DNA methylation status predicts cell type-specific enhancer activity. *The EMBO Journal* 30, 15 (jun 2011), 3028–3039.
- [343] WILKINSON, D. M., AND HUBERMAN, B. A. A method for finding communities of related genes. *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl, suppl 1 (apr 2004), 5241–8.
- [344] WILSON, C. B., ROWELL, E., AND SEKIMATA, M. Epigenetic control of T-helper-cell differentiation. *Nature reviews. Immunology* 9, 2 (2009), 91–105.
- [345] WITTE, S., O’SHEA, J. J., AND VAHEDI, G. Super-enhancers: Asset management in immune cell genomes. *Trends in immunology* 36, 9 (sep 2015), 519–26.
- [346] WITTKOPP, P. J., AND KALAY, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13, 1 (jan 2012), 59–69.
- [347] WOLPERT, D., AND MACREADY, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* 1, 1 (apr 1997), 67–82.
- [348] XU, D., AND TIAN, Y. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science* 2, 2 (jun 2015), 165–193.
- [349] XU, J., YANG, Y., QIU, G., LAL, G., YIN, N., WU, Z., BROMBERG, J. S., AND DING, Y. Stat4 is critical for the balance between Th17 cells and regulatory T cells in colitis. *Journal of immunology (Baltimore, Md. : 1950)* 186, 11 (jun 2011), 6597–606.
- [350] XU, S., GRULLON, S., GE, K., AND PENG, W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods in molecular biology (Clifton, N.J.)* 1150 (2014), 97–111.
- [351] YANG, Y., OCHANDO, J. C., BROMBERG, J. S., AND DING, Y. Identification of a distant T-bet enhancer responsive to IL-12/Stat4 and IFN γ /Stat1 signals. *Blood* 110, 7 (oct 2007), 2494–500.

- [352] YANG, Y., XU, J., NIU, Y., BROMBERG, J. S., AND DING, Y. T-bet and eomesodermin play critical roles in directing T cell differentiation to Th1 versus Th17. *Journal of immunology (Baltimore, Md. : 1950)* 181, 12 (dec 2008), 8700–10.
- [353] YANG, Y., AND YE, D. Inverses of bipartite graphs. *Combinatorica*, 11671347 (2017), 1–13.
- [354] YEUNG, C. H., AND SAAD, D. Networking - A statistical physics perspective. *Journal of Physics A: Mathematical and Theoretical* 46, 10 (2013).
- [355] YORDANOV, P., AND STELLING, J. Steady-State Differential Dose Response in Biological Systems. *Biophysical Journal* 114, 3 (feb 2018), 723–736.
- [356] YOUNG, M. D., WILLSON, T. A., WAKEFIELD, M. J., TROUNSON, E., HILTON, D. J., BLEWITT, M. E., OSHLACK, A., AND MAJEWSKI, I. J. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic acids research* 39, 17 (sep 2011), 7415–27.
- [357] YU, B., ZHANG, K., MILNER, J. J., TOMA, C., CHEN, R., SCOTT-BROWNE, J. P., PEREIRA, R. M., CROTTY, S., CHANG, J. T., PIPKIN, M. E., WANG, W., AND GOLDRATH, A. W. Epigenetic landscapes reveal transcription factors that regulate CD8 + T cell differentiation. *Nature Immunology* 18, 5 (2017), 573–582.
- [358] YU, D., LIM, J., WANG, X., LIANG, F., AND XIAO, G. Enhanced construction of gene regulatory networks using hub gene information. *BMC bioinformatics* 18, 1 (mar 2017), 186.
- [359] YU, H., ZHU, S., ZHOU, B., XUE, H., AND HAN, J.-D. J. Inferring causal relationships among different histone modifications and gene expression. *Genome Research* 18, 8 (aug 2008), 1314–1324.
- [360] YU, J., SMITH, V. A., WANG, P. P., HARTEMINK, A. J., AND JARVIS, E. D. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 18 (dec 2004), 3594–3603.
- [361] YUAN, Y., LI, C.-T., AND WINDRAM, O. Directed partial correlation: inferring large-scale gene regulatory network through induced topology disruptions. *PloS one* 6, 4 (apr 2011), e16835.
- [362] ZABIDI, M. A., ARNOLD, C. D., SCHERNHUBER, K., PAGANI, M., RATH, M., FRANK, O., AND STARK, A. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518, 7540 (feb 2015), 556–559.
- [363] ZANG, C., SCHONES, D. E., ZENG, C., CUI, K., ZHAO, K., AND PENG, W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25, 15 (aug 2009), 1952–1958.
- [364] ZHANG, P., BEHRE, G., PAN, J., IWAMA, A., WARA-ASWAPATI, N., RADOMSKA, H. S., AURON, P. E., TENEN, D. G., AND SUN, Z. Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proceedings of the National Academy of Sciences* 96, 15 (1999), 8705–8710.

- [365] ZHANG, S.-Q., CHING, W.-K., TSING, N.-K., LEUNG, H.-Y., AND GUO, D. A new multiple regression approach for the construction of genetic regulatory networks. *Artificial Intelligence in Medicine* 48, 2-3 (feb 2010), 153–160.
- [366] ZHANG, Z., SHAN, T., AND CHEN, G. Random walks on weighted networks. *Physical Review E* 87, 1 (jan 2013), 012112.
- [367] ZHENG, W.-P., AND FLAVELL, R. A. The Transcription Factor GATA-3 Is Necessary and Sufficient for Th2 Cytokine Gene Expression in CD4 T Cells. *Cell* 89, 4 (may 1997), 587–596.
- [368] ZHOU, W., AND DICKERSON, J. A. A novel class dependent feature selection method for cancer biomarker discovery. *Computers in Biology and Medicine* 47 (2014), 66–75.
- [369] ZHU, J., JANKOVIC, D., OLER, A., WEI, G., SHARMA, S., HU, G., GUO, L., YAGI, R., YAMANE, H., PUNKOSDY, G., FEIGENBAUM, L., ZHAO, K., AND PAUL, W. The Transcription Factor T-bet Is Induced by Multiple Pathways and Prevents an Endogenous Th2 Cell Program during Th1 Cell Responses. *Immunity* 37, 4 (oct 2012), 660–673.
- [370] ZHU, J., MIN, B., HU-LI, J., WATSON, C. J., GRINBERG, A., WANG, Q., KILLEEN, N., URBAN, J. F., GUO, L., AND PAUL, W. E. Conditional deletion of Gata3 shows its essential function in TH1-TH2 responses. *Nature Immunology* 5, 11 (nov 2004), 1157–1165.
- [371] ZHU, J., YAMANE, H., AND PAUL, W. Differentiation of effector CD4 T cell populations. *Annu Rev Immunol.* 28, 1 (2010), 445–489.
- [372] ZUO, Y., YU, G., TADESSE, M. G., AND RESSOM, H. W. Biological network inference using low order partial correlation. *Methods* 69, 3 (2014), 266–273.
- [373] ZUR BONSEN, A., OMELCHENKO, I., ZAKHAROVA, A., AND SCHÖLL, E. Chimera states in networks of logistic maps with hierarchical connectivities. *European Physical Journal B* 91, 4 (2018).

List of Figures

II.1	Decision tree splitting visualization.	19
II.2	HMM depiction of Markov chain.	21
II.3	HMM parameter scheme depiction.	22
III.1	Mouse model used for underlying Th1 and Th2 cell conditions. . . .	25
III.2	Epigenetic histone modification peak landscape results around <i>Ifnγ</i>	30
III.3	LFC transcript expression plot of differential regulation between Th2 and Th1 wild-type cells.	33
III.4	Dispersion estimate plot for the differential expression between Th2 and Th1 cells.	33
III.5	Standard deviations for absolute gene transcript expression values determined via a VST and via a classical library size normalization.	35
III.6	Hierarchical clustering of the Euclidean distance VST count matrix for all RNA-Seq samples.	35
III.7	Principal component analysis of all sample replicates for absolute values obtained by naïve normalization in comparison to VST normalized values.	36
III.8	Pre-processing pipeline for histone modification ChIP-Seq data and RNA-Seq data samples.	37
IV.1	Emission probabilities for models with five up to 25 hidden states for the underlying five histone modification marks.	42
IV.2	Best emission parameter correlation for 5-25 state HMMs.	43
IV.3	16-state HMM parameters with emission probabilities and transition probabilities.	44
IV.4	Functional and positional fold enrichments for the 16-state HMM.	45
IV.5	Chromatin state landscape of <i>Gata3</i> and <i>Tbx21</i> according to the 16-state HMM with state color-coding as defined in Fig.IV.3.	46
IV.6	Annotated <i>Ifnγ</i> locus with colour-coded HMM state segmentation.	48
IV.7	Venn diagram depiction of overlap of p300 binding sites in Th1 and Th2 cells with HMM enhancer states in Th1, Th1/2 and Th2 cells respectively.	50
IV.8	The <i>Il1rl1</i> locus with the inferred HMM segmentation.	51
V.1	Parameter distributions for the correlation measure obtained from an $n = 1000$ bootstrap sampling procedure.	57

V.2	Dependence of <i>Ifnγ</i> gene expression on H3K4me1, H3K27ac, H3K27me3 and the combined parametrized measure respectively at the <i>Ifnγ</i> enhancer segment at CNS-34.	58
V.3	Pearson correlation values of all enhancer elements from the training sample rank ordered decreasingly by their respective correlation via the parametrized correlation measure.	59
V.4	Correlations of enhancer state elements at the <i>Ifnγ</i> locus for different upper bound resolutions.	63
V.5	Algorithmic flowchart of the correlation analysis.	66
V.6	Significant correlations of enhancer state elements at the <i>Ifnγ</i> locus for combined 600 bp and 2000 bp resolutions.	68
V.7	Close-up of an upstream region of <i>Ifnγ</i> showing the validation of the CNS-54 enhancer element.	69
V.8	Schematic depiction of significantly correlating connected enhancer segments at the <i>Ifnγ</i> locus including its regulatory enhancer activity logic.	70
V.9	Annotated <i>Tbx21</i> locus according to our algorithmic correlation procedure.	72
V.10	Schematic depiction of significantly correlating enhancer elements for the <i>Tbx21</i> locus.	73
V.11	Significant correlations of enhancer segments with several notable Th2-specific transcripts within the Th2 cytokine locus.	75
V.12	Hierarchical clustering of significant enhancer segment correlations with gene transcripts at the Th2 cytokine locus.	76
V.13	Clustering of the correlation matrix of VST-normalized expression values of Th2-specific transcripts in the Th2 cytokine cluster.	77
V.14	Significant remaining partial correlations at the Th2 cytokine locus. We colour-coded the respective most significant attribution of each enhancer segment to a certain transcript.	79
V.15	Phase-space of partial correlations and <i>p</i> -values for all enhancer segments from the negative co-regulated correlation cluster in the Th2 cytokine cluster.	80
V.16	Linear fitting result of predicted vs. measured <i>Ifnγ</i> expression where only one replicate was respectively used for fitting the other.	83
VI.1	Class hierarchy for chromatin states and gene transcripts.	88
VI.2	Feature ranking w.r.t. Gini impurity of all ESC features.	90
VI.3	Feature ranking w.r.t. Gini impurity of all RSC features.	91
VI.4	Top 20 ranked ESC features for Th1 as well as Th2 transcripts.	94
VII.1	Full force-directed CSC-gene network.	104
VII.2	In- and out-degree distributions.	106
VII.3	In-degree centrality for the full weighted multi-digraph.	107
VII.4	Betweenness centrality for the full weighted multi-digraph.	109
VII.5	Closeness centrality for the full weighted multi-digraph.	110
VII.6	Eigenvector centrality for the full weighted multi-digraph.	110
VII.7	Katz centrality for the full weighted multi-digraph.	111
VII.8	Core CSC-TF network.	115

VII.9	TF activation and inhibition network after removing the CSC layer.	116
VII.10	TF activation network with ESC layer and low-ranked ESCs removed.	117
VII.11	Community structures of the full network according to different community algorithms.	119
VII.12	Depiction of all considered multiplex network dimensions with fixed node positions w.r.t. the full network.	123
VII.13	Resulting community structures according to the HER method for two exemplary multiplex layers, i.e. Tbet ^{+/+} Th1/2 and Th2 control.	125
VII.14	Differential network of Th2 control w.r.t. Tbet ^{+/+} Th1.	126
VII.15	Differential network of Tbet ^{-/-} Th1 vs. Tbet ^{+/+} Th1.	127
VII.16	Differential network of Tbet ^{+/+} Th1/2 vs. Tbet ^{+/+} Th1 (<i>Diff1</i>).	127
VII.17	Differential network of Tbet ^{+/+} Th1/2 vs. Th2 control (<i>Diff2</i>).	128
VII.18	Stationary distribution of nodes in the full CSC-gene network with fixed node positions compared to earlier depictions.	133
VII.19	Eigenvalue spectrum of the full CSC-gene network.	135
VII.20	Visualization of all eigenvector components of the second and third eigenvector of the full network.	136
VII.21	Heatmap of the leading eigenvector components of the full network of notable Th1 and Th2 genes.	137
VII.22	Eigenvalue spectrum of the Tbet ^{+/+} Th1/2 multiplex dimension.	138
VII.23	Heatmap of the leading eigenvector components of the Tbet ^{+/+} Th1/2 multiplex dimension of notable Th1 and Th2 genes.	138
VII.24	Leading logarithmic real eigenvalue parts of the upregulated section of the <i>Diff1</i> network as well as of the downregulated section.	139
VII.25	Heatmaps of the leading eigenvector components of the <i>Diff1</i> network of notable Th1 and Th2 genes for upregulated connections as well as for downregulated connections.	140
VII.26	Close-up of the leading logarithmic eigenvalues of the reduced epigenetic Tbet-Gata3 network.	141
VII.27	Close-up of the leading logarithmic eigenvalues of the epigenetic Tbet-Gata3 MISA motif.	143
B.1	Per base sequence quality and sequence length distribution for one ChIP-Seq sample of H3K4me1 under Tbet ^{+/+} Th1 conditions.	161
B.2	Per base sequence quality for one RNA-Seq sample for Tbet ^{+/+} Th1 conditions.	161
B.3	Heatmap of pairwise correlations between different ChIP-Seq samples.	162
B.4	Hierarchical clustering of a subsample of VST normalized genes for all RNA-Seq samples.	163
B.5	PCA analysis components PC3 and PC4 for the VST case.	164
B.6	Pearson and Spearman correlations for the parametrized histone measure in comparison to the individual histone modification Pearson correlations.	164
B.7	Results from the correlation algorithm for a resolution of 600 bp for all HMM states as well as for enhancer states only.	164

B.8	Schematic depiction of significantly correlating activating elements at the <i>Gata3</i> locus unveiling the enhancer activity regulation.	165
B.9	Schematic depiction of significantly correlating activating elements at the <i>STAT1</i> locus unveiling the enhancer activity regulation.	165
B.10	Schematic depiction of significantly correlating activating elements at the <i>STAT4</i> locus unveiling the enhancer activity regulation.	166
B.11	Schematic depiction of significantly correlating activating elements at the <i>STAT6</i> locus unveiling the enhancer activity regulation.	166
B.12	Schematic depiction of significantly correlating activating elements at the <i>GAPDH</i> locus unveiling the enhancer activity regulation.	167
B.13	Top 20 ranked RSC features for Th1 as well as Th2 transcripts.	167
B.14	Out-degree centrality for the full weighted multi-digraph.	168
B.15	Top-ranked directed 4-node subgraphs w.r.t. <i>Z</i> -score.	169
B.16	Rearranged full network after the deletion of the two Th1 and Th2 master TFs Tbet and Gata3.	170
B.17	Rearranged full network after the deletion of STAT1 corresponding to a STAT1 knockout of the system.	170
B.18	SC community detection results for $k = 10$	171
B.19	RWIF community detection results for a granularity parameter of 2.0.	171
B.20	Force-directed Tbet ^{+/+} Th1/2 network.	172
B.21	Differential network of Tbet ^{+/+} Th1/2 vs. Tbet ^{+/+} Th1 (<i>Diff1</i>) considering edge addition and deletion.	172

List of Tables

II.1	Master TFs and signature cytokines for the most notable T-helper and CD4+ subtypes.	6
III.1	Experimental treatment conditions for the considered data sets. . .	25
III.2	List of histone modifications used in the experiments.	26
V.1	Intrinsic logic for the correlation algorithm of a Th1 or Th2 chromatin feature.	67
VI.1	Total amount of top ranked Th1 and Th2 ESCs and RSCs respectively in order to reproduce a certain percentage of total Gini impurity.	96
C.1	Top 50 up-regulated transcripts for the differential Th1/Th2 comparison.	173
C.2	Top 50 up-regulated transcripts for the differential Th2/Th1 comparison.	174
C.3	List of relevant 46 Th1 and 50 Th2 transcripts as determined in [334] and from our RNA-Seq data sets.	175
C.4	Percentage of genome occupancy of each HMM state in every experimental condition.	176
C.5	Corresponding BIC scores for different models with parameter number k	176
C.6	List of independently validated Th1 and Th2 enhancers used for parametrical learning of the histone modification correlation measure.	176
C.7	Partial correlation values and corresponding p -values of the enhancer segment <i>chr11:53623600-53628000</i> with co-regulated gene-transcripts.	177
C.8	Inter-class Gini impurity and information gain ranking for the Top 25 ESCs.	177
C.9	Full inter-class Gini impurity ranking for the set of all ternary ESCs.	181
C.10	Th1 intra-class Gini impurity ranking for the set of all ternary ESCs.	183
C.11	Th2 intra-class Gini impurity ranking for the set of all ternary ESCs.	186
C.12	Top-ranked conditional probabilities of an ESC A given an ESC B. .	187
C.13	ESC-specific TF binding weights obtained from the intra-class Gini impurity measure.	192
C.14	Additional notable network statistics of the full CSC-gene network. .	193
C.15	Leading ranked nodes for weighted in- and out-degree respectively.	194
C.16	Respective top 100 node rankings w.r.t. betweenness, closeness, eigenvector and Katz centralities.	197

C.17	GLayer community clustering results with six modules or clusters for all genes and selected ESCs.	199
C.18	Spectral community clustering ($k = 5$) results with six modules or clusters for all genes and selected CSCs.	200
C.19	Spectral community clustering ($k = 10$) results with ten modules or clusters for all genes and selected CSCs.	202
C.20	RWIF clustering (Granularity 1.8) results with fourteen modules or clusters for all genes and selected CSCs.	204
C.22	RWIF clustering (Granularity 2.0) results with 25 modules or clusters for all genes and selected CSCs.	208
C.25	Betweenness centrality ranking with respective values for the Th2 vs. Tbet ^{+/+} Th1 differential network with respective up- and down-regulated parts.	212
C.26	In- and out-degree ranking with respective values for the <i>Diff1</i> network with respective up- and down-regulated parts.	215
C.27	PageRank centrality ranking for the full network as well as for the Tbet ^{+/+} Th1 and Tbet ^{+/+} Th1/2 networks.	218
C.28	PageRank centrality ranking for the up- and down-regulated parts of <i>Diff1</i> and <i>Diff2</i>	221
C.29	Eigenvectors λ_1 and λ_2 for the Tbet/Gata3 subnetwork with leading ranked CSCs for λ_1 and CSCs with largest positive and negative λ_2 component.	223
C.30	Eigenvectors λ_1 , λ_2 and λ_3 for the Tbet/Gata3 MISA motif with leading ranked microstates for λ_1 and microstates with largest positive and negative λ_2 component.	226

Acknowledgements

It naturally goes without saying that such a work could have never been achieved without the help, advice, support and patience of many other people.

First of all I would like to express my sincere gratitude to my doctoral supervisor Prof. Thomas Höfer. Thomas, I profited so much during all these years from your advice and vast knowledge in the field, which shaped not only the results in the thesis as it is presented here, but you also always challenged me in a positive way to find new creative paths and also to find a new way to think about science in general. For this (and of course your great wit and humour) I am very grateful.

Also I would like to thank Prof. Ursula Kummer, my second referee, who was always enthusiastic about the project and contributed on several instances with extremely helpful comments and her general openness.

At this point I want to thank my other two TAC members: Michael Floßdorf, whose advice was always invaluable, and Congxin Li, who was also my roommate during all these years, to whom I owe deep gratitude for always having an open ear and taking his valuable time for intense discussions on the project.

Furthermore I would like to thank the whole Höfer group from DKFZ where it is especially hard to single out only some people since the discussions with all of you were extremely helpful and the time with you simply entertaining. This held true not only in the office, in seminars, at conferences or during lunch time, but also at our unforgettable group retreats. So to the current group with Jens, Verena, Nils, Lisa, Congxin, Qin, Matthias, Carsten, Erika, Ines, Lena, Melania, Soheil, Adrien, Xi, Alessandro, Nick, Diana, Conny and to all the others who already left the group: thank you! Also thank you to Julia Schnessner who as a student intern contributed valuably to the development of the correlation algorithm.

Of course a big thank you goes to our collaborators Prof. Max Löhning and his whole group at DRFZ Berlin and Prof. Ahmed Hegazy for numerous helpful discussions, the friendly hospitality in Berlin and of course also thank you for the contribution to the thesis with your perfectly produced experimental data. In this regard I also owe a great deal to Qin Zhang from the Höfer group who was responsible for the invaluable high-quality ChIP- and RNA-Seq data set production. Furthermore I would like to thank Naveed Ishaque for providing a large deal of advice especially at the beginning of my work in the group.

I furthermore want to thank the two graduate programs I was given the ability to take part in: The Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp) as well as The Helmholtz International Graduate School for Cancer Research from the DKFZ in Heidelberg who not only provided a stimulating research atmosphere and program but also helped me as a physicist in shaping a strong interdisciplinary research background. In this regard

also thank you to the European SysmedIBD project for a range of rewarding meetings in always a nice atmosphere.

But not only the support from work has to be considered. Obviously the largest part of my support I owe to my wife, Maren, who always is my focal point concerning all questions that can possibly arise. Thank you for all your understanding, love and support during this time (and naturally long before), which always kept me going and growing and will even more so in the future.

A large thank you also goes to my invaluable friends over all the years: Niklas, thank you especially not only for the incountable discussions concerning academic questions and innumerable personal stuff¹, but also for the great time since the first day of our physics studies. This praise naturally also has to go to Martin, Tim and Sebastian who contributed so much to the great time and to all the great university and non-university related experiences over all these years. In this regard two of the most important persons not only during the time of my PhD but dating back before the time at university are Patrick and Nico whom I cannot thank enough for everything. All of you had a large impact on my personal and academic development.

Speaking of which: Thank you so much to my parents, Jürgen and Susanne Kommer, and to my sister, Dorothee, for your continuous love and support over all of these years, for making me who I am intellectually and personally and giving me all the opportunities I had. Without the knowledge, humour and creativity you provided me with this work could have never been accomplished.

¹“Phrasing” – Sterling Archer