UNIVERSITY OF CALIFORNIA SAN DIEGO

Novel network-based integrated analyses of multi-omics data reveal new insights into $CD8^+$ T cell differentiation and mouse embryogenesis

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Kai Zhang

Committee in charge:

Professor Wei Wang, Chair
Professor Pavel Arkadjevich Pevzner, Co-Chair
Professor Vineet Bafna
Professor Cornelis Murre
Professor Bing Ren

2018

The dissertation of Kai Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2018

EPIGRAPH

The only true wisdom is in knowing you know nothing.

—Socrates

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Wei Wang for his guidance throughout my graduate education and insight during the many problem-solving processes required to complete this dissertation.

I would like to thank Dr. Pavel Pevzner, Dr. Vineet Bafna, Dr. Bing Ren and Dr. Cornelis Murre for taking time to serve on my dissertation committee and for their valuable advice. I also acknowledge Dr. Ananda Goldrath and her lab for their collaborative efforts and experimental work as detailed in Chapter 3.

Furthermore, this dissertation would not have been possible without the help of my fellow lab members Dr. Yun Zhu, Dr. Nan Li, Dr. Richard Ainsworth, Dr. Ying Zhao and Mengchi Wang. I thank them for their guidance, assistance and contributions to the work submitted for publication.

Chapter 2, in full, is a reprint of the material as it appears in Systematic Identification of Protein Combinations Mediating Chromatin Looping. Zhang, Kai; Nan Li, Richard I. Ainsworth, Wei Wang. Nature Communications 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Epigenetic Landscapes Reveal Transcription Factors That Regulate CD8+ T Cell Differentiation. Yu B, Zhang K, Milner J, Toma C, Chen R, Scott-Browne J, Pereira R, Crotty S, Chang J, Pipkin M, Wang W, Goldrath A. Nat Immunol 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Systems-level identification of transcription factors critical for mouse embryonic development. Zhang K, Wang M, Zhao Y, Wang W. Biorxiv 2017. The dissertation author was the primary investigator and author of this paper.

VITA

2009	B. S. in Biological Sciences, Xiamen University, Xiamen, China
2012	M. S. in Molecular Biology and Biochemistry, Xiamen University, Xiamen, China
2013-2018	Research Assistant, University of California San Diego
2018	Ph. D. in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Zhang K, Wang M, Zhao Y, Wang W: Systems-level identification of transcription factors critical for mouse embryonic development. *Biorxiv* 2017, :16719710.1101/167197.

Yu B*, **Zhang K***, Milner J, Toma C, Chen R, Scott-Browne J, Pereira R, Crotty S, Chang J, Pipkin M, Wang W, Goldrath A: Epigenetic landscapes reveal transcription factors that regulate CD8+ T cell differentiation. *Nat Immunol* 2017, 18:573–58210.1038/ni.3706. (* equally contribution)

Zhang K, Li N, Ainsworth R, Wang W: Systematic identification of protein combinations mediating chromatin looping. *Nat Commun* 2016, 7:1224910.1038/ncomms12249.

Milner J, Toma C, Yu B, **Zhang K**, Omilusik K, Phan A, Wang D, Getzler A, Nguyen T, Crotty S, Wang W, Pipkin M, Goldrath A: Runx3 programs CD8+ T cell residency in non-lymphoid tissues and tumours. *Nature* 2017, 552:25310.1038/nature24993.

Zhu Y, Chen Z, **Zhang K**, Wang M, Medovoy D, Whitaker J, Ding B, Li N, Zheng L, Wang W: Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* 2016, 7:1081210.1038/ncomms10812.

Ainsworth R, Ai R, Ding B, Li N, **Zhang K**, Wang W: Bayesian Networks Predict Neuronal Transdifferentiation. *G3 Genes Genomes Genetics* 2018, 8:g3.200401.201810.1534/g3.118.200401.

Wei W, Ji Z, He Y, **Zhang K**, Ha Y, Li Q, Ohno-Machado L: Finding relevant biomedical datasets: the UC San Diego solution for the bioCADDIE Retrieval Challenge. *Database* 2018, 2018:bay017–10.1093/database/bay017.

ABSTRACT OF THE DISSERTATION

Novel network-based integrated analyses of multi-omics data reveal new insights into CD8$^+$ T cell differentiation and mouse embryogenesis

by

Kai Zhang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2018

Professor Wei Wang, Chair
Professor Pavel Arkadjevich Pevzner, Co-Chair

Advancements in next-generation sequencing technologies have fueled the development of high throughput profiling assays. The sheer volume of data generated by these experiments grants us an unprecedented opportunity to deepen our understanding of complex biological systems, but also raise many computational challenges. With multi-omics data becoming very common in modern biological research, one of the most urgent tasks is to develop novel algorithms to perform integrated data analyses of these data. In the thesis, we developed network-based frameworks to integrate a variety of omics data. We combined the strength of different high throughput assays

and analyzed multi-omics data to infer molecular interactions and build regulatory networks. Using network-based approaches, we addressed two important biological problems: identifying protein complexes mediating the formation of chromosome loops and identifying driver TFs in different biological processes. We further conducted computational and biological experiments to validate our findings. Our study provides new insights into the processes of $CD8^+$ T cell development and mouse embryogenesis. The identified driver TFs can be used as the blueprint for future mechanistic study.

Chapter 1

General introduction

Network science has drawn a lot of attention during the past two decades, leading to revolutionary changes in a variety of fields, including computer science, physics, chemistry, biology, and social science. Data in real life often exhibit structure or connection properties. Therefore, graphs or networks are natural formulations of many problems across different domains, leading to broad applications in almost every field of scientific research. The studies of network topology led to many intriguing discoveries, such as the small-world[1] and scale-free[2] properties. These findings pave the way to understand the mechanics of complex systems.

1.1 The applications of graph theory in bioinformatics

In the field of biology, networks have become a major mode of analysis. Generally, most biological events are not the result of a single molecule but depend on the coordinated effects of multiple molecules interacting with each other. The network is an intuitive representation of the interactions or relations between biological molecules or organisms. There is an increasing effort to organize our understanding of the biological system into different network representations, such as Protein-Protein interaction (PPI) networks, metabolic pathways, and genetic networks. As the related biotechnologies quickly advanced, the sheer volume of network data grow exponentially

and have become an essential part of scientific research. Particularly the emergence of high throughput screening technologies, such as the Perturb-Seq[3], enabled us to disrupt hundreds of genes in thousands of cells in parallel. This kind of experiments provides us with an unprecedented opportunity for constructing large-scale genetic networks. Besides, a lot of new data is also being generated by traditional techniques, such as mass spectrometry, chromatin immunoprecipitation and yeast two-hybrid assay. On the other hand, the data-mining algorithms constantly improve, allowing more data being extracted and curated from the literature.

To handle the huge amount of currently available network data, a number of databases were developed to organize and store this information and to provide convenient access to researchers. KEGG, initiated in 1995, is one of the most popular databases of molecular networks[4]. It stores molecular interaction, reaction and relation networks representing the systemic functions of the cell and the organism. Experimental knowledge of such systemic functions is captured from literature and organized into different forms. Founded in 2001, REACTOME database is another popular resource for studying biological networks. As an open-source, open access, manually curated and peer-reviewed pathway database, REACTOME aims at providing intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic and clinical research. While KEGG and REACTOME mostly focus on direct interactions of molecules and are more conservative in their curation process, the STRING database includes both direct (physical) interaction and indirect (functional) interactions, as long as both are specific and biologically meaningful. It uses computational approaches to derive putative associations from various sources, including systematic co-expression analysis, detection of shared selective signals across genomes, automated text-mining of the scientific literature and computational transfer of interaction knowledge between organisms based on gene orthology. These different types of databases complement with each other to form entire toolchain for modern network study.

Apart from being used to encode relationships between bio-molecules, graphs are also used to model complex biological processes and to process big data. Under these scenarios, the network

is perceived as a "concept" to provide a theoretical basis for devising bioinformatics algorithms to tackle different problems. Among different models, probabilistic graphical models become extremely popular. First, as probabilistic graphical models are very flexible, scalable and are able to combine heterogeneous sources of data, they offer an appealing framework for addressing various problems in systems biology. For modeling directional relationships, Bayesian networks provide a compact representation for expressing joint probability distributions (JPDs) and for inference. They are becoming increasingly important and have many applications, including inferring cellular networks[5], modeling protein signaling pathways[6], and data integration[7]. For systems that cannot be represented by a directed network, Markov random field models provide a means to encode undirected relationships and can represent certain dependencies that a Bayesian network cannot, such as cyclic dependencies. Markov random field models have broad applications in analyzing genomics data. For example, one of the most popular chromosome segmentation tools – ChromHMM[8] – was built upon the hidden Markov model.

Graph theory is a rich source of concepts and algorithms applicable to many different disciplines. Many biological problems can be converted to graph theory problems. As a result, a handful of existing algorithms is immediately applicable. For instances, the genome assembly problem is to aligning and merging fragments from a longer DNA sequence to reconstruct the original sequence. This problem can be formulated as the Eulerian cycle problem, which is known to have a linear time solution[9]. The application of graph theory to the genome assembly problem was a huge breakthrough, as at that time the computational resource was still scarce and the best algorithm took quadratic time. Cluster analysis is a process of dividing a set of objects into possibly overlapping subsets, where objects in each subset are considered more similar to each other than those not in the subset. Cluster analysis is one of the most important components in bioinformatics analysis and is often the first step of the data-mining process. Cluster analysis can be formulated as the graph cluster problem, which seeks to a method of dividing the nodes of a graph into clusters.

1.2 Leveraging graphs to conduct integrated analyses

The biological research is entering the era of "Big Data". More and more omics data are generated and it has become very common to have multi-omics data in a single study. For example, the ENCODE project[10] has conducted ChIP-Seq, RNA-Seq, whole genome bisulfite sequencing and ATAC-Seq experiments in parallel in mammalian cells. As a result, there is a great need to develop new methods to perform the integrated analysis.

As discussed in the previous section, graphs or networks have gradually become a major player in bioinformatics analysis. It has great potentials for solving the emerging problems in "Big Data" era. In this thesis, I sought to develop new bioinformatics tools to integrate a wide range of genomics or epigenomics data to address important biological questions, using graph theory approaches. In particular, I have presented three studies, illustrating the applications of novel graph algorithms to different problems in systems biology.

1. Chromatin looping plays a pivotal role in gene expression and other biological processes through bringing distal regulatory elements into spatial proximity. The formation of chromatin loops is mainly mediated by DNA-binding proteins (DBPs) that bind to the interacting sites and form complexes in three-dimensional (3D) space. Previously, identification of DBP cooperation has been limited to those binding to neighbouring regions in the proximal linear genome (1D cooperation). In Chapter 2, we developed a new algorithm based on Gaussian graphical model to integrates protein ChIP-seq and Hi-C data to systematically identify both the 1D- and 3D-cooperation between DNA-binding proteins. Our method allows identification of cooperation between multiple DBPs and reveals cell-type-specific and -independent regulations. Using this framework, we retrieve many known and previously unknown 3D-cooperations between DBPs in chromosomal loops that may be a key factor in influencing the 3D organization of chromatin.

2. Dynamic changes in the expression of transcription factors (TFs) can influence the spec-

ification of distinct $CD8^+$ T cell fates, but the observation of equivalent expression of TFs among differentially fated precursor cells suggests additional underlying mechanisms. In Chapter 3 we profiled the genome-wide histone modifications, open chromatin and gene expression of naive, terminal-effector, memory-precursor and memory CD8+ T cell populations induced during the in vivo response to bacterial infection. Integration of these data suggested that the expression and binding of TFs contributed to the establishment of subset-specific enhancers during differentiation. We developed a new bioinformatics method to construct TF regulatory network and use the PageRank algorithm to reveal key TFs that influence the generation of effector and memory populations. The TFs YY1 and Nr3c1, both constitutively expressed during $CD8^+$ T cell differentiation, regulated the formation of terminal-effector cell fates and memory-precursor cell fates, respectively. Our data define the epigenetic landscape of differentiation intermediates and facilitate the identification of TFs with previously unappreciated roles in $CD8^+$ T cell differentiation.

3. Dynamic changes in the transcriptional regulatory circuit can influence the specification of distinct cell types. Numerous transcription factors (TFs) have been shown essential for the rewiring of the genetic network during embryonic development but a systematic identification of these TFs is still lacking. In Chapter 4, we performed an integrated analysis of epigenomics and transcriptomics data to reveal key regulators from 2 cells to postnatal day 0 in mouse embryogenesis. We predicted 3D chromatin interactions including enhancer-promoter interactions in 12 tissues across 8 developmental stages, which facilitates linking TFs to their target genes for constructing transcriptional regulatory networks. To identify driver TFs particularly those not necessarily differentially expressed ones, we developed a new algorithm, dubbed as Taiji, to assess the global importance of TFs in development. Through comparative analysis across tissues and developmental stages, we systematically uncovered TFs that are critical for lineage-specific and stage-dependent tissue specification. Most interestingly, we have identified TF combinations that function in spatiotemporal

order to form transcriptional waves regulating developmental progress and differentiation, which suggests a distributed synchronization mechanism between tissues to orchestrate embryonic development. Not only does our analysis provide the first comprehensive map of transcriptional regulatory circuits during mouse embryonic development, the identified novel regulators and the predicted 3D chromatin interactions also provide a valuable resource to guide further mechanistic studies.

1.3 References

1. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440. ISSN: 0028-0836 (Apr. 1998).

2. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. en. *Science* **286,** 509–512. ISSN: 0036-8075, 1095-9203 (15 10 1999).

3. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N. & Regev, A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. en. *Cell* **167,** 1853–1866.e17. ISSN: 0092-8674, 1097-4172 (15 12 2016).

4. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. en. *Nucleic Acids Res.* **28,** 27–30. ISSN: 0305-1048 (Jan. 2000).

5. Friedman, N. Inferring cellular networks using probabilistic graphical models. en. *Science* **303,** 799–805. ISSN: 0036-8075, 1095-9203 (June 2004).

6. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. en. *Science* **308,** 523–529. ISSN: 0036-8075, 1095-9203 (22 4 2005).

7. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). en. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 8348–8353. ISSN: 0027-8424 (Aug. 2003).

8. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. en. *Nat. Methods* **9,** 215–216. ISSN: 1548-7091, 1548-7105 (28 2 2012).

9. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. en. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 9748–9753. ISSN: 0027-8424 (14 8 2001).

10. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

Chapter 2

Systematic identification of protein combinations mediating chromatin looping

2.1 Introduction

The human genome is tightly packaged into chromatin and forms complex structures of which the functional outputs, such as gene expression, depend on local chromatin states and chromatin three-dimensional (3D) organization[1–11]. Chromatin loops are formed to bring distal regulatory elements such as enhancers and their target promoters to spatial proximity. The formation of chromatin loops is mainly regulated by proteins that bind to the 3D interaction sites and form complexes[1]. Previous studies have shown that perturbation of the binding of these proteins could disrupt the loops, which suggests an important role for DNA-binding proteins (DBPs) in genome organization. Mediators of chromatin loops including CTCF, cohesin and several transcription factors (TFs) such as GATA1 and KLF1 have been identified[12–16]. Particularly, a recent study has uncovered about 10,000 chromatin loops using kilobase-resolution Hi-C data and discovered that CTCF and cohesin subunits RAD21 and SMC3 are present in the majority of the loops[17].

These studies have shown that the cooperation of multiple DBPs is critical to orchestrate loop formation. However, there still lacks a systematic method to investigate the role of combinatorial regulation between DBPs in chromatin loop formation. Previous studies have focused on identifying DBPs binding to proximal genomic regions[18–21], which is hereinafter referred to as 1D-cooperation. Despite the great insight provided by these studies in revealing the combinatorial regulation of DBPs, they could not detect the cooperation between DBPs binding to distal genomic loci that are localized spatially and form long-range interactions (referred to as 3D-cooperation in this study to be distinct from the 1D-cooperation of DBPs in neighbouring genomic loci). 3D-cooperation of DBPs is key to mediating chromatin looping, either enhancing the existing 3D contacts or creating new ones to bring functional elements, such as enhancers, to their target loci, such as promoters. Despite its importance, no study has thoroughly investigated the DBPs' 3D-cooperation and its relationship to 1D-cooperation.

The ENCODE project has generated hundreds of ChIP-seq data sets to map binding sites of DBPs in multiple cell lines[19, 22, 23]. Recently, kilobase-resolution Hi-C data were available in two of these cell lines, namely GM12878 and K562[17]. These data sets provide an unprecedented opportunity to systematically map both 1D- and 3D-cooperation between DBPs. However, it is a great challenge to analyse this large amount of data and extract cooperation among multiple rather than pairs of DBPs.

To tackle this challenge and comprehensively catalogue DBP cooperation, we present here a new model to construct networks that represent both 1D- and 3D-association between DBPs. Analysing these networks in GM12878 and K562 has revealed complex cooperative relationships among TFs, histone modifications, chromatin-remodelling enzymes and chromatin architectural proteins. Through the identification of communities and cliques in the DBP cooperation network, we have uncovered many DBP interactions in the chromatin loop regions. Intriguingly, many of these 3D-cooperative DBPs directly interact with one another, which suggests their binding may be important for loop formation or stabilization in 3D space. Furthermore, we performed

a comparative network analysis between GM12878 and K562, and revealed cell-type-specific cooperation between DBPs that are critical for regulating cell-type-specific functions.

2.2 Methods

2.2.1 Data sets.

BAM files of ChIP-seq experiments in K562 and GM12878 were downloaded from the ENCODE project website[24]. Chromatin loops were downloaded from the study by Rao *et al.*[17]. Because only these two cell lines had both DBP ChIP-seq and 5-kb resolution Hi-C data, we focused on these data sets in this study.

2.2.2 Data preprocessing.

We divided each chromosome into consecutive 1-kb regions. For each protein, we computed the Reads Per Kilobase per Million (RPKM) mapped reads on these regions. The fold enrichment was calculated using MACS's algorithm[25] with customized parameters. In particular, we set $\lambda_{local} = max(\lambda_{BG}, \lambda_{14k}, \lambda_{24k})$ where λ_{BG} is the average RPKM of the whole genome; λ_{14k} and λ_{24k} are average RPKM of 14 and 24 kb windows. We used a larger window size than MACS's default size and a loose P value (0.01) to call peaks to increase the sensitivity for detecting broad peaks. For each 1 kb region, if it was called as a peak, we used the fold enrichment as its ChIP-seq enrichment score; otherwise, a zero score is assigned to that region. After computing the enrichment scores for every protein, we removed regions with low variation of scores by requiring the *s.d.* of scores to be at least 1. This excludes some unwanted artifacts from our analysis. For instance, regions with low mappability or an abnormally high signal[24]. Next, for each DBP pair we calculated the Spearman's correlation of ChIP-seq enrichment scores in the remaining bins as the 1D-correlation score. To compute 3D correlation scores, we first

downloaded the 3D interaction loops identified in a 5-kb resolution Hi-C study[17]. Next, for each DBP we computed its enrichment scores on loop regions as follows: suppose we have n loops, denoted by $L^1, L^2, ..., L^n$ and each loop L^i consists of two interacting loci L^i_a and L^i_b. To compute the enrichment score of a given DBP on loop L^i, we first binned L^i_a into 1-kb regions, and then took the maximum of ChIP-seq enrichment scores of these bins as the enrichment score for L^i_a. Likewise, we can compute the enrichment score for L^i_b. Then, given a pair of DBPs denoted by A and B, for every loop, we first compared the enrichment scores of A on the two interacting loci. We considered the interacting locus with larger enrichment score of protein A as A's primary binding locus, and the other interacting locus as the primary binding locus of protein B. The enrichment scores of primary binding loci for each protein were then used to compute the correlation coefficient.

2.2.3 Network construction.

We adapted the GGM to construct the DBP cooperation networks. GGM assumes that the observations have a multivariate Gaussian distribution with mean m and covariance matrix Σ. Let Σ^{-1} be the inverse of covariance matrix. If the ijth component of Σ^{-1} is zero, then variables i and j are conditionally independent given other variables. Therefore, each non-zero component represents an edge in the network. To efficiently and accurately estimate the inverse of the covariance matrix using DBP ChIP-seq data, we employed the Graphical lasso algorithm[26] and the Copula method[27]. We used a lasso penalty equal to 0.3 in this study. We chose this value because less than 15% of DBP pairs have a correlation score at least 0.3. To estimate the false discovery rate, we generated a null model by random shuffling of DBP binding sites to represent uncooperative DBPs. When we applied our algorithm to this data set, the cutoff we chose identifies zero cooperation, suggesting our method has a very low false discovery rate. Because we aimed at identifying DBP interactions, edges with negative correlations were removed in the network analysis.

2.2.4 Network analysis.

We used Eppstein's algorithm[28] for maximal clique searching, which gives an exact solution in near optimal time. For community detection, we used Newman's algorithm[29].

2.2.5 Comparing the DBP cooperation network with PPI network.

Protein-protein interaction data was obtained from the BioGrid database[30] version 3.2.99. For each edge formed by node A and B in a DBP cooperation network, if also present in the PPI network, it was considered as a direct interaction. Otherwise, we checked whether there exists a third node in the PPI network that connects to both A and B; if so, this edge was considered as an indirect interaction. To determine the statistical significance of these overlaps, we first replaced the nodes in the DBP cooperation network with randomly selected genes from the PPI network. Next, we counted the direct and indirect interactions in the simulated network. This process was repeated for 10^9 times to generate the background distribution, which was then used to calculate the P values.

2.2.6 Simulated networks.

To generate an Erdős-Rényi random graph, we used the $G(n,p)$ model. This model specifies an n-node network, in which each edge is included with a probability P independent from every other edge. We used $P = 0.2$ in this paper, which gives rise to a sparse network. We follow the procedures given in [27] to generate a Gaussian distributed data set that was used for constructing the simulated network. We used Genetweaver 3.1 to extract random subnetworks with different sizes (50 and 100 nodes) from the yeast gene regulatory network provided by the software. For other parameters, we used the software's default setting. To draw the receiver operating characteristic (ROC) curve, we counted the number of true positives, false positives, true negatives and false negatives. If a predicted edge is present in the true network, it is a true positive,

otherwise it is a false positive. Edges that are present in the true network but not identified by the algorithm are defined as false negatives. True negatives are edges that are not present in either predicted or true networks.

2.3 Results

2.3.1 Gaussian graphical model

To systematically identify DBP cooperation, we analysed DBP ChIP-seq data using Gaussian graphical model (GGM)[31]. GGM is an undirected probabilistic graphical model with the assumption that the data follows a multivariate Gaussian distribution with mean μ and covariance matrix Σ. Let Σ^{-1} be the inverse of covariance matrix. If the ijth component of Σ^{-1} is zero, then variables i and j are conditionally independent given all other variables in the network[31]. This important property serves as the foundation for GGM to infer direct interactions from high-dimensional data. Unlike relevance networks or correlation networks, in which edges are determined based on marginal correlations, GGM provides a stronger criterion of dependency, and thus further reduces the false positive rate. However, a great limitation of classic learning methods for GGM is the lack of sparsity in the resulting graph. A dense graph not only complicates downstream analysis but also raises the issue of overfitting the data. To cope with this, Friedman *et al.*[26] proposed an efficient algorithm, named graphical Lasso, to introduce sparsity to the GGM. Recently, Liu *et al.*[27] developed a data transformation method called Copula that can be used with the graphical Lasso algorithm to relax the normality assumption of GGM. Based on these recent advances, we developed a new framework to systematically identify cooperation between hundreds of DBPs.

Before applying the GGM to the DBP ChIP-seq data, we assessed its performance using synthetic data. First, we generated an Erdős-Rényi random graph as our ground truth (see Methods). To generate samples according to the simulated graph, we constructed a covariance

matrix by assigning each ijth component a non-zero covariance if node i and j were connected in the simulated graph. All other components were then set to zero. We next drew samples from a multivariate Gaussian distribution parameterized by a zero mean vector and the constructed covariance matrix. These samples were used as input for network re-construction. As a comparison, we selected ARACNE[32], a popular algorithm for constructing gene regulatory networks that employs an information theory approach to infer interactions from gene expression data. We generated 10 networks with 50 nodes and another 10 networks with 100 nodes. When applying both methods to these data sets, we observed a superior performance of GGM with an average AUC of 0.923, which is significantly higher than ARACNE (AUC = 0.822) (Fig. 2.1a). This simulation showed that, when experimental data follows a Gaussian distribution, the GGM can precisely reconstruct the underlying graphical model. However, the real data can be quite noisy and may not be Gaussian distributed. To cope with this, we incorporated the Copula algorithm[27] and carried out a further benchmark to evaluate its performance on a more noisy data set. To produce synthetic gene expression data sets, we used GeneNetWeaver 3.1[33], an in silico simulator that employs a dynamic model to simulate gene regulatory networks. The ground truth were subnetworks taken from yeast gene regulatory network with size 50 and 100, respectively. For each size, we performed 10 different simulations. Again, GGM outperformed ARACNE (average AUC of 0.695 versus 0.615; Fig. 2.1b).

It is worth noting that GGM is much faster than ARACNE when the sample size is large. The time complexity for ARACNE is $O(N^3 + N^2M^2)$, where N is the number of variables or nodes in the network, M is the number of samples; as it scales with M^2, it is not suitable for our application where we have more than 10,000 samples (the number of ChIP-seq peaks). In contrast, GGM, with a time complexity $O(N^3 + N^2M)$, can easily handle a large number of samples. In practice, we observed that the GGM was 50-100 times faster than ARACNE on the synthetic data sets (Table 2.1).

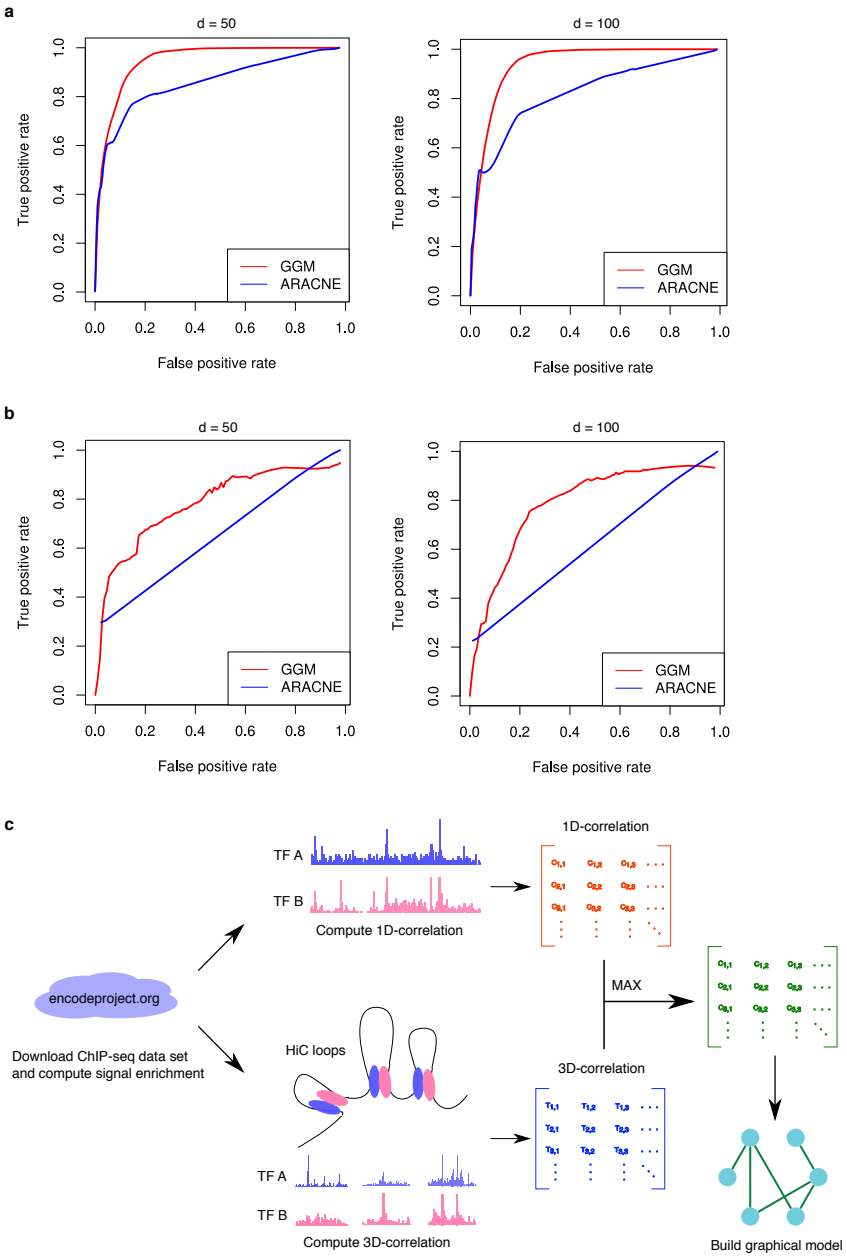

Figure 2.1: The performance of the GGM is consistently better than ARACNE. Each plot shows the average curve from 10 independent simulations. (**a**) ROC curve for samples generated from random networks. For each simulation 500 (left) or 1,000 (right) samples were generated from a network of 50 (left) or 100 (right) nodes. (**b**) ROC curve for samples generated from yeast sub-networks. For each simulation 500 (left) or 1,000 (right) samples were generated from a network of 50 (left) or 100 (right) nodes. (**c**) Workflow of the DBPnet pipeline.

Table 2.1: Speed comparison of GGM and ARACNE on data sets with different sizes.

Algorithm	Number of nodes	Number of samples	Average running time and standard deviation (second)
ARACNE	50	100	0.44(0.001)
		200	1.6(0.001)
		400	6.3(0.003)
	100	100	1.7(0.002)
		200	6.5(0.005)
		400	25(0.03)
	200	100	6.8(0.004)
		200	26(0.005)
		400	102(0.23)
GGM	50	100	0.03(0.002)
		200	0.07(0.001)
		400	0.14(0.001)
	100	100	0.09(0.002)
		200	0.17(0.001)
		400	0.33(0.002)
	200	100	0.34(0.008)
		200	0.49(0.002)
		400	0.89(0.002)

2.3.2 Constructing the DBP cooperation network.

We applied the GGM framework to DBP ChIP-seq and Hi-C data, aiming to systematically detect DBP cooperation (Fig. 2.1c). We considered both 1D (DBPs that bind to loci in the nearby linear genome) and 3D (DBPs that bind to loci that are spatially close but linearly distal in the genome) cooperation between DBPs. We first computed 1D and 3D correlation scores for each pair of DBPs separately using the 84 ChIP-seq data sets, including six histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3) as well as chromatin loops called by the 5-kb resolution Hi-C data in a lymphoblastoid cell line GM12878[17] (see details in Methods). We then merged 1D and 3D correlation matrices by keeping the larger correlation score at each entry. This matrix was used to construct the GGM, which represents the DBP cooperation network.

The DBP cooperation network (Fig. 2.2a) contains 484 associations between 84 DBPs. An edge between two proteins may indicate either a direct physical interaction or a co-occurrence of binding sites without direct interaction. To examine whether our model can recover protein-protein interactions (PPI), for each edge we searched for supporting evidence from the public PPI databases (Methods). Remarkably, 11% of edges (empirical P value is 1×10^{-9}) in the GGM network are also present in the PPI network (Fig. 2.2b). Another 80% of the associated DBPs are separated by one protein in the PPI network (the intermediate protein may not be analysed by the ChIP-seq experiments). This evidence strongly supports that the DBP cooperation recovered by our method is reliable and likely represents physical contacts. Furthermore, we found that 11.5% and 11.5% of 3D-dominant and 1D-3D cooperative edges, respectively, are coincident with protein-protein interactions, which is much higher than the 1D-dominant edges (1.8%); 94.6%, 63.5% and 79.4% of 1D-dominant, 3D-dominant and 1D-3D cooperative DBP pairs are separated by one protein in the protein-protein interaction network, respectively. This observation suggests that our analysis did identify physical interactions in the 3D space and many 1D-dominant ones may be formed through indirect interactions.

To characterize the topological properties of the DBP cooperation network, we plotted its node degree distribution. In agreement with other types of biological networks, we observed that the node degree distribution of the DBP cooperation network follows a power law, reflecting its scale-free property. A prominent feature of scale-free networks is the existence of hubs, which are the highly connected nodes that may be critical for network stability. To identify hubs, we ranked the nodes in our network by two popular centrality metrics – node degree centrality and eigenvector centrality. Node degree centrality for a given node is simply the number of nodes that link to the given node, while eigenvector centrality reflects both the node degree and its connection with other well-connected nodes. We ranked the nodes by both their node degree and eigenvector centrality. The results show that EP300, CREB1 and EBF1 are the top three DBPs that have the best average rank (Fig. 2.2a). EP300 is an important cofactor that cooperates with

many TFs[34, 35] to perform a variety of biological functions. CREB1 plays a central role in the immune system through binding to the c-AMP response element, a ubiquitous DNA motif, to regulate gene transcription[36, 37]. It was not surprising that these two general DBPs would be found as hubs. Previous studies showed that EBF1 is mainly expressed in B-lymphocytes (GM12878 is a lympoblastoid cell line) and is pivotal for maintenance of B-cell identity[38]. In the DBP cooperation network, EBF1 is linked to many important transcriptional regulators, including general activators such as EP300 and SP1, as well as B-lymphocyte-specific TFs such as PAX5, TCF12 and BCL11A ([39–42]). By analysing the topology of the DBP cooperation network, we uncovered TFs that are crucial for cell functions.

2.3.3 Identifying 1D and 3D cooperation between DBPs.

DBPs can cooperate through 1D or 3D interactions, which can be determined for each DBP pair using the constructed network. In this study, we define an edge as 1D or 3D cooperation if the 1D or 3D correlation score is larger than a pre-selected cutoff (0.3, see Methods). In GM12878, we found roughly the same numbers of 1D and 3D edges, 413 and 417 respectively. We noticed a great overlap between 1D and 3D edges (Fig. 2.2c). We thus labelled these DBP cooperations as 1D-dominant, 3D-dominant or 1D-3D cooperative.

A 1D-dominant association cooperation between two DBPs represents a frequent co-occurrence in linear space but not in the long-range interacting loci that form loops in the 3D space. In this category, we recovered some previously known interactions such as the RNA Pol II-TAF1 interaction[43] (Fig. 2.2d). Interestingly, we found 71 3D-dominant edges in GM12878. Most of these edges show weak 1D correlations but have significantly larger 3D correlations. For instance, the 1D and 3D correlation scores of EP300-MYC are 0.127 and 0.348 (z-score: 0.081 and 1.301), respectively. Indeed, a number of independent studies have shown that EP300 and MYC can cooperate to regulate gene transcription and the physical interaction between them has been previously reported[44, 45]. We also identified novel 3D DBP cooperation. For example,

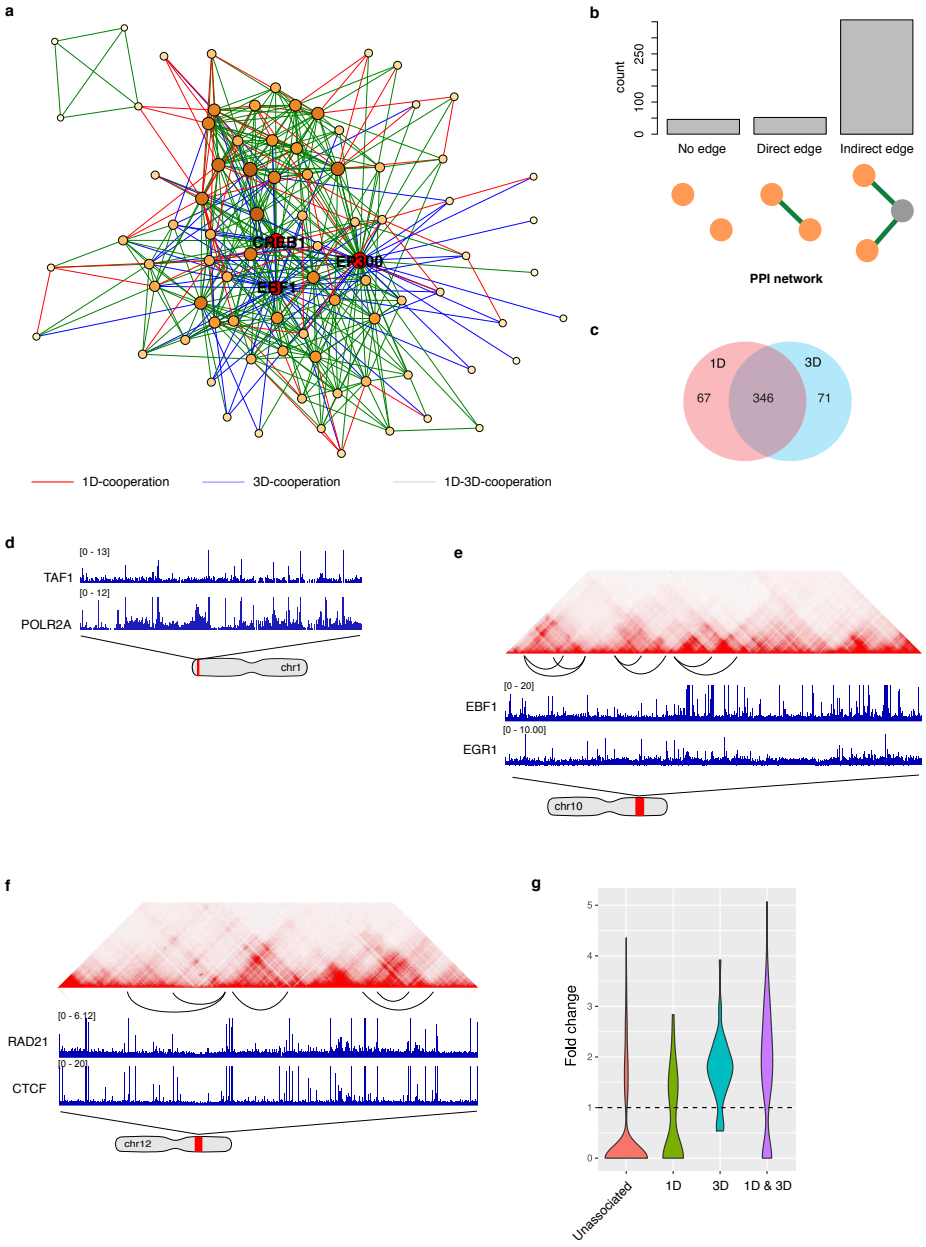

Figure 2.2: Constructing the DBP cooperation network in GM12878. (**a**) DBP cooperation network in GM12878, with network hubs (EP300, EBF1, CREB1) being highlighted. (**b**) A significant portion of DBP cooperation is supported by evidence of direct protein-protein interactions. (**c**) The majority of DBP cooperation is a mixture of 1D and 3D cooperation. (**d**) An example of 1D-cooperation. (**e**) An example of 3D-cooperation. (**f**) An example of mixed cooperation. (**g**) Disease-associated genotype variations are enriched in 1D-dominant ($n = 67$), 3D-dominant ($n = 71$) and 1D-3D cooperative ($n = 346$) sites.

EBF1 is an important TF in B lymphocytes, and Egr-1 is one of the key transcriptional regulators induced on B-cell antigen-receptor activation[46]. Both EBF1 and Egr-1 have crucial roles in B-cell development and differentiation. However, the interplay between these two proteins has not been reported. In our network, we found a 3D-dominant edge between EBF1 and Egr-1 (Fig. 2.2e), which suggests that they may form long-range loops to regulate cell-type-specific genes (4.4% of loop regions contain peaks of both EBF1 and Egr-1). Therefore, our framework provides a systematic way to uncover 3D cooperation between DBPs that are otherwise impossible to identify using previous approaches. The 1D-3D cooperation is formed between DBP pairs with both 1D and 3D associations. Most associations fall into this category. A well-known example is CTCF-RAD21 (Fig. 2.2f). While 1D-dominant cooperation can be identified by previous approaches[19], the last two categories of DBP cooperation can only be identified through the integration of DBP ChIP-seq and Hi-C data, which highlights the advantage of our method.

To further confirm the importance of DBP cooperation, we analyzed the enrichment of genotype variations in regions bound by cooperative DBPs. The disease-associated genotype variations were downloaded from the NHGRI-EBI GWAS database[47]. Given two DBPs A and B, if they are 1D-cooperative, we considered sites bound by both A and B as foreground regions. If A and B are 3D-cooperative, we first identified chromatin loops where one of the two anchors is bound by A and the other is bound by B. Within these loops, we identified binding sites of A or B as the foreground. If A and B are 1D- and 3D-cooperative, we considered only those loops for which each of the two anchors contain the binding sites of both A and B, and these sites are used as the foreground. In all cases, background are the binding sites of A or B that are not in the foreground. We then calculated the percentage of regions containing genotype variations for foreground and background, and took the ratio as the genotype variation enrichment. In Fig. 2.2g, we showed that the vast majority of DBP cooperation are more enriched with disease-associated genotype variations. We performed the Mann-Whitney U-test to compare the significance level of enrichments of 1D-dominant, 3D-dominant and 1D-3D cooperative DBPs

with that of uncooperative DBP pairs, the P values are 1.6×10^{-3}, 5.2×10^{-28} and 2.0×10^{-45}, respectively. These results suggest that DBP cooperation has important functional implications in a variety of diseases. Therefore, we anticipate that the binding sites of cooperative DBPs can be used to prioritize genotype variations to identify causal associations.

2.3.4 Identifying DBP communities.

Modularity is an important property of biological networks. Characterization of the modularity in DBP cooperation networks can illuminate how multiple DBPs cooperate to carry out complex regulation. Modularity can be studied at different levels. For instance, cliques highlight local modules in the network while community structure is a more global view of the modularity. Communities are groups of nodes in a network that are more densely connected internally than with the rest of the network[29]. In other words, community structure is a partition or clustering of the nodes in a network. We applied the community detection algorithm[29] to the DBP network in GM12878, and the nodes are separated into four communities (Fig. 2.3a). From the community structure, an immediate observation is the existence of a very small community (yellow) that is formed by only five proteins, namely CTCF, RAD21, ZNF143, SMC3 and YY1. Intriguingly, all these five proteins are important in mediating chromatin looping[17, 48, 49], suggesting that a major function of this community may relate to chromatin structure organization. This observation suggests that DBPs in the same community are functionally cooperative, and the communities in the DBP network may have different biological functions.

To reveal the functions of a particular community, for each protein in the community we analysed its ChIP-seq peaks. We then ranked genomic loci by the number of proteins that bind to them and selected the top 5,000 loci as input to GREAT analysis to search for enriched GO terms. We found that the green community is linked to mRNA metabolic processes and translation-related functions. The same analysis showed that the cyan community is also enriched with similar GO terms. Interestingly, these two communities share 3,248 out of the 5,000 loci used

in GREAT analysis despite being segregated in the network. A closer examination of these two communities revealed distinct protein composition. For instance, all the six histone modifications as well as RNA polymerase II belong to the green community, suggesting its pivotal role in gene transcription. In contrast, the cyan community contains numerous proteins, including ESRRA, BRCA1, NRF1, ETS1 and STAT3, that are involved in the oestrogen-signalling pathway.

Furthermore, GREAT analysis of binding sites of DBPs in the red community revealed that "immune response", "leukocyte activation" and "lymphocyte activation" are the most enriched GO terms. These terms are highly specific to the B cell, suggesting that this community is crucial in determining cell-type specificity. Indeed, many proteins in this community are known to be important for immune system development, such as STAT5A[50], BATF[51], BCL3[52] and RELA[53].

2.3.5 Identifying potential DBP complexes.

On a finer scale, the modularity of a network is revealed by cliques. A clique is a complete sub-graph in which every pair of nodes is connected. Intuitively, DBPs that form a clique in the network are more likely to function as a complex. Undoubtedly, the identification of such complexes is crucial for understanding the mechanisms of transcriptional regulation. Therefore, we searched for maximal cliques in the network and identified 220 cliques in GM12878. We ranked DBP cliques by their average correlation scores for each DBP pair. We observed that edges in most of the top cliques are associated with high 1D and 3D correlation scores, which suggests that they are likely to form complexes mediating chromatin loop formation. Figure 2.3b shows the top three highest ranked cliques. Next, we checked the percentage of shared peaks in the union of all DBP peaks and identified the loops that overlap with these shared peaks. As a comparison, for each k-component DBP clique, we randomly selected k DBPs and did the same analysis. We took 50,000 samples and used them as a null model for enrichment and empirical P value calculation. As a result, all the DBPs in the cliques share a significant amount of peaks that

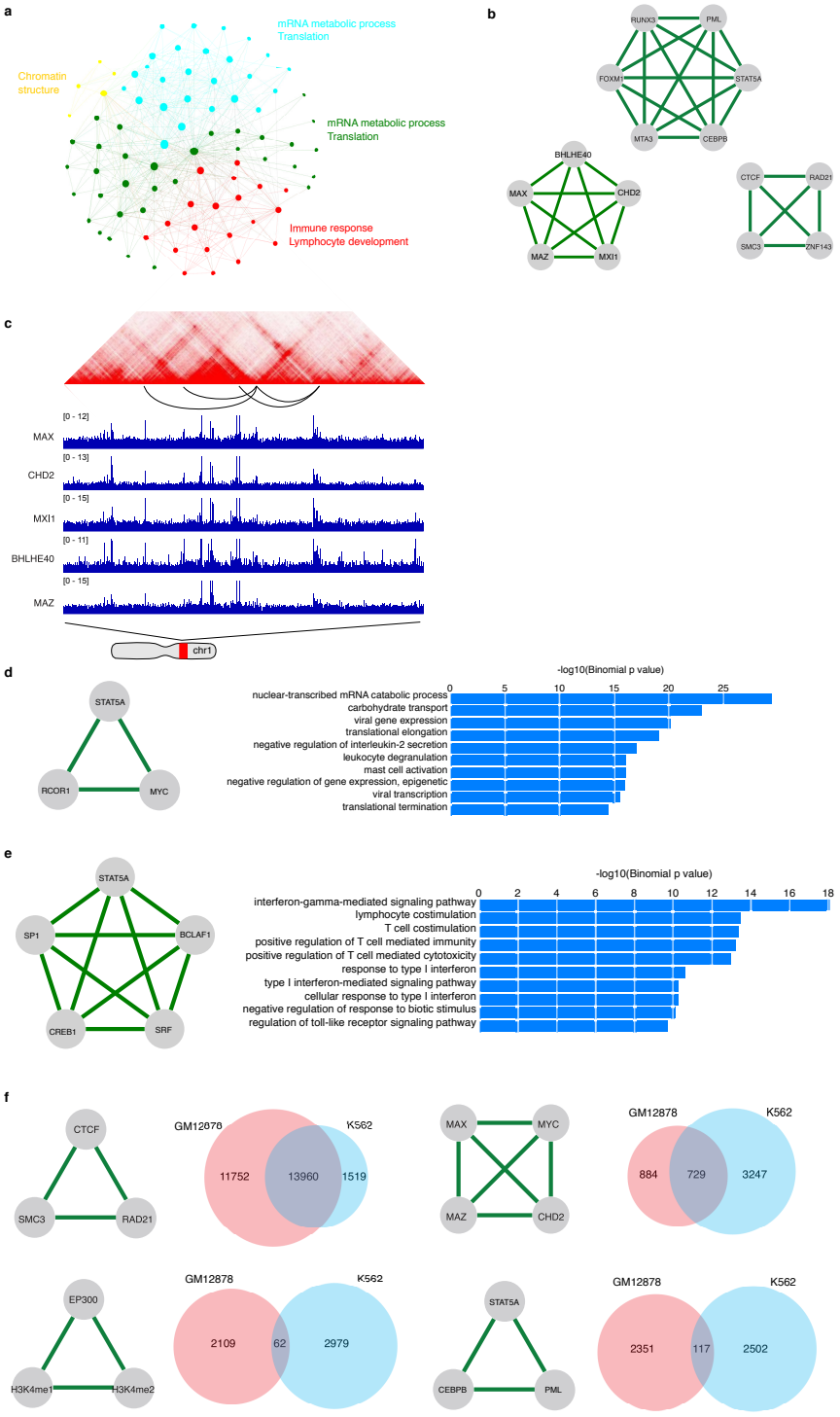

Figure 2.3: Network analysis reveals functions of DBP modules in GM12878 and K562. (**a**) Communities in the DBP cooperation network and their functions. (**b**) TopDBP cliques. (**c**) An example of DBP cliques. (**d**) An example of K562-specific DBP cliques and the enriched GO terms of their binding sites. (**e**) An example of GM12878-specific DBP cliques and enriched GO terms of their binding sites. (**f**) Top conserved DBP modules in K562 and GM12878.

occur in loops (Table 2.2), which confirmed the co-occurring bindings of the DBPs in a clique.

Table 2.2: The top 3 most frequently occuring DBP cliques. The central column gives the number of regions bound by all DBP members in a clique, the percentage of regions bound by all DBP members, their fold enrichment over background and the empirical P values. The right column gives the number of loops that are overlapped with the DBP-binding sites, the percentage, the fold enrichment over background and empirical P values.

DBP clique	No. of sites/pct./enrich-ment (P value)	No. of overlapped loops/pct./enrichment (P value)
CTCF, RAD21, SMC3, ZNF143	13,829/21.9%/8.8 (4.0e-5)	5,668/47.2%/30.5 (<2.0e-5)
PML, FOXM1, MTA3, STAT5A, CEBPB, RUNX3	2,195/3.1%/16.8 (1.8e-3)	388/3.2%/16.3 (5.2e-4)
MAX, MAZ, MXI1, CHD2, BHLHE40	2,900/8.8%/24 (8.0e-5)	501/4.2%/12.5 (1.1e-3)

The top-ranked clique is CTCF-RAD21-SMC3-ZNF143. RAD21 and SMC3 are components of the cohesin complex. Cohesin is a multi-subunit protein complex and plays an essential role in sister chromatid cohesion and chromosome segregation during cell division[54]. Cohesin is also crucial for regulating gene expression and mediating chromatin long-range interactions[55]. Cohesin-dependent chromatin interactions are usually mediated by the cooperation of cohesin and CTCF[49]. The involvement of ZNF143 in this complex has also been reported[48]. ZNF143 is believed to provide sequence specificity for chromatin interactions[56]. Overall, our analysis successfully recovered this important and well-characterized loop-forming complex.

The other two cliques, PML-FOXM1-MTA3-STAT5A-CEBPBRUNX3 and MAX-MAZ-MXI1-CHD2-BHLHE40 (Fig. 2.3c) have not been reported. STAT5A and RUNX3 are two of the major transcription factors that play essential roles in lymphocyte development. The physical interaction between STAT5 and RUNX3 has been reported[57]. Moreover, CEBPB binds to RUNX2 that has been shown to be associated with RUNX3[58, 59]. To investigate the function of this module, we extracted all loci bound by these six DBPs and performed GREAT analysis.

The most significant GO terms are “immune response”, “leukocyte activation” and “lymphocyte activation”. These results suggest that this module may play important roles in the development of lymphocytes.

The functions of the MAX-MAZ-MXI1-CHD2-BHLHE40 clique are more general. The enriched GO terms for their binding sites are “ribonucleoprotein complex biogenesis”, “nuclear-transcribed mRNA catabolic process ribosome biogenesis” and “translation”. The interaction between MAX and MXI1 is well studied[60] but interactions between other proteins have not been reported. However, the functions of these proteins are highly related. For example, BHLHE40 is a repressor that can interact with and recruit HDACs, which suggests a role for BHLHE40 in chromatin remodelling. CHD2 is also a chromatin remodeler. These observations suggest that DBPs in this clique may act together to alter chromatin states and regulate gene translation.

2.3.6 Comparative analysis of DBP cooperation networks.

DBPs have different cooperative modes in different cells. To perform a comparative analysis of the networks in different cell types, we focused on 68 proteins for which ChIP-seq data sets are available in both K562 and GM12878, and constructed TF cooperation networks in these two cell types.

To find cell-type-specific DBP cliques, we first identified cell-type-specific edges in the 68-node networks. We then searched for cliques in both GM12878- and K562-specific networks that consist of edges present in one but not the other cell line. We found 74 and 7 cell-type-specific cliques for GM12878 and K562, respectively. Cell-type-specific cliques shed light on how cells achieve transcriptional specificity through the combinatorial regulation of DBPs. For example, STAT5A is a member of STAT protein family. It is activated by a number of cytokines and plays a central role in the development of many different organs. However, how STAT5A cooperates with other DBPs to carry out cell-typespecific regulation is largely unknown. Our analysis showed that STAT5A, together with MYC and RCOR1, forms a clique in K562, which is absent in GM12878.

MYC is an oncogene and has been shown to play a critical role in leukaemia formation[61, 62]. STAT5A-MYC cooperation may be important to maintain the state of leukaemic cells. To further characterize the functions of the STAT5A-MYC-RCOR1 clique, we performed GREAT analysis on loci bound by all the three proteins in K562 and identified functions specific to leukocyte, such as “leukocyte degranulation”, “regulation of interleukin-2 secretion” and “mast cell activation” (Fig. 2.3d). These functions are drastically different from those enriched in GM12878 where STAT5A is associated with BCLAF1, SRF, CREB1 and SP1; GREAT analysis on the shared peaks suggests this clique is involved in lymphocyte specific functions (Fig. 2.3e).

Next, we sought to identify common DBP cliques in GM12878 and K562. First, we extracted a common network using edges shared by the two networks. We then searched for cliques in this network. We identified CTCF-RAD21-SMC3, MAX-MYC-MAZ-CHD2, EP300-H3K4me1-H3K4me2 and STAT5A-CEBPB-PML as top-ranked cliques (Fig. 2.3f). CTCF-RAD21-SMC3 interaction is known to be conserved across different cell types and it is not surprising that this clique is shared between the two cells. In the MAX-MYC-MAZ-CHD2 clique, MAX-MYC-MAZ is also a well-known complex that is found in multiple cell lines but their interaction with CHD2 has not been reported. The involvement of the chromatin-remodelling gene CHD2 in the MAX-MYC-MAZ complex suggests MAX-MYC-MAZ may utilize CHD2 to modify chromatin structure and alter gene expression. EP300-H3K4m1-H3K4me2 represents an enhancer’s signature, and has been found in many cell types. In the clique of STAT5A-CEBPB-PML, there is evidence for the STAT5A-CEBPB interaction: STAT5A was demonstrated to cooperate with CEBPB to regulate gene transcription[63]; STAT5A can induce deacetylation of CEBPB[64]. Their interaction with PML is less well-studied but STAT5 is shown to be activated by the PML/RARa fusion protein in acute myeloid leukaemia[65]. These common cliques in both GM12878 and K562 indicate their cell-type-independent cooperation.

We next investigated whether these common cliques bind to the same loci in the two cells. For each clique, we identified the sites bound by all the member DBPs and counted how many of

them are shared between the two cell types. We observed that the CTCF-RAD21-SMC3 clique shared 13,960 (51.3%) common binding sites in K562 and GM12878 (Fig. 2.3f), which is in agreement with the general roles of CTCF and the cohesin complex in stabilizing loops[17]. The MAX-MYC-MAZ-CHD2 clique shows moderate conservation with 729 (15%) common binding sites across the two cell types. In contrast, the binding sites of P300-H3K4me1-H3K4me2 and STAT5A-CEBPB-PML are highly cell-type-specific. Since P300-H3K4me1-H3K4me2 mark active enhancers and enhancers are highly cell-typespecific, it is understandable that there are only 62 (1.2%) P300-H3K4me1-H3K4me2 peaks shared across cell types. The fact that only a small percentage (2.4%, 117 sites) of STAT5A-CEBPB-PML sites are shared between GM12878 and K562 is surprising. To investigate the reason why STAT5A-CEBPB-PML has distinct binding profiles in the two cell types, we first analysed the enriched GO terms for the sites bound by all three DBPs in K562 and GM12878, respectively. Enriched GO terms in each cell type are highly specific: the top terms in GM12878 are “immune response” and “lymphocyte activation”; sites in K562 are enriched with GO terms such as “platelet activation” and “erythrocyte differentiation”, which are highly specific to K562.

The above analyses show that the same DBPs can bind to different loci to regulate cell-type-specific functions. There are several possible reasons for such cell-type-specific binding, such as differential accessibility of chromatin, DBPs recognizing celltype- specific motifs[66], and DBPs partnering with different cofactors. To understand the differential binding of the STAT5A-CEBPB-PML clique, we identified their cell-specific partners by examining the binding peaks of all the available ChIP-seq data in the regions bound by STAT5A-CEBPB-PML in GM12878 and K562 (Fig. 2.4a, b). It is obvious that the co-occurring DBPs are very different in the two cell types: RUNX3, BCL11A, BATF in GM12878 compared with TEAD4, TAL1, GATA2 in K562. To assess the contribution of cofactors to such cell-type-specific binding, for each potential cofactor we used its ChIP-seq peaks to discriminate binding regions of STAT5A-CEBPB-PML in GM12878 and K562. The top 12 TFs that have best discriminative accuracy are RUNX3, TEAD4, TAL1,

BCL11A, BATF, IRF4, PAX5, POU2F2, BCL3, GATA2, MYC and EBF1. Strikingly, either RUNX3 or TEAD4 alone can achieve an over 99% accuracy, which is consistent with their distinct binding patterns in Fig. 2.4. When the binding sites of these 12 TFs were used together to train a logistic regression model, we achieved a 100% accuracy rate for discriminating the STAT5A-CEBPB-PML binding regions in the two cell lines. Therefore, the cell-type-specific binding of this DBP clique can be explained by its partnership with different cofactors. Furthermore, we observed that all 12 TFs except MYC are differentially expressed in the two cell types, suggesting that the cell-type-specific binding of this DBP clique is largely due to cell-type-specific expression of cofactors. Because of the limited number of ChIP-seq experiments, possible partners might not be profiled. Therefore, we performed *de novo* motif analysis using MEME-ChIP[67] in STAT5A-CEBPB-PML sites and then matched the found motifs to the known ones. Clearly, the de novo motifs found in K562 and GM12878 were very different. Encouragingly, the motifs of several co-factors identified from ChIP-seq experiments were also retrieved from the *de novo* motif analysis. These results suggest that the STAT5A-CEBPB-PML complex indeed has different regulatory mechanisms in different cell types. To further interrogate the regulatory mechanisms of their cooperation, we used Spamo[68] to find spacing constraints between *de novo* motifs. As a result, in GM12878 we found two de novo motifs, corresponding to STAT5A and MEF2, showing a statistically significant spacing constraint with the MEF2 motif occurring 13 bp downstream of the STAT5 motif (Fig. 2.4c). This finding is new as there is no previous report about the partnership of STAT5A and MEF2. We also found, in K562 the TAL1::GATA1 motif frequently appears upstream of RUNX1 sites at a distance of 38 bp. GATA and RUNX usually cooperate with each other and form a cis-regulatory module[69, 70]. Therefore, our analysis has identified both new and known spacing constraints between TFs.

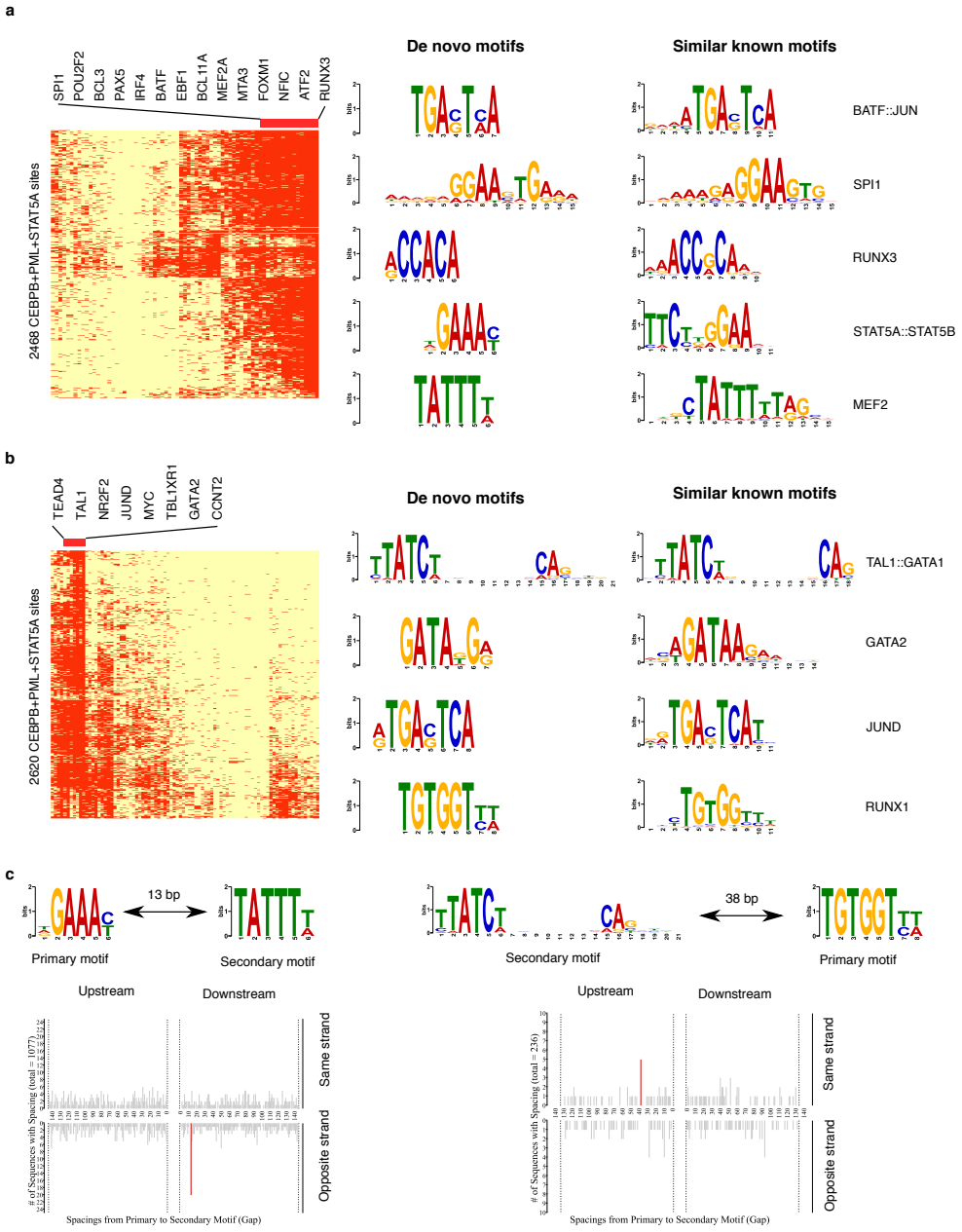

Figure 2.4: CEBPB-PML-STAT5A cooperates with different DBPs in GM12878 and K562. (**a**) DBP-binding profile (left) and enriched *de novo* motifs (right) in 2468 CEBPB-PML-STAT5A-binding sites in GM12878. (**b**) DBP-binding profile (left) and enriched *de novo* motifs (right) in 2620 CEBPB-PML-STAT5A-binding sites in K562. (**c**) Enriched spacing between *de novo* motifs found in CEBPB-PML-STAT5A sites in K562 and GM12878.

2.4 Discussion

We present here a first systematic search of DBP complexes mediating chromatin loop formation using a novel framework. Our method can identify both 1D and 3D cooperation between DBPs. Many of the identified cooperations are likely a result of physical interactions as most of the edges in the DBP cooperation network are supported by the PPI data. Our results showed that 3D-cooperation between TFs is ubiquitous, indicated by 86% of identified associations having strong 3D correlation scores, which can only be discovered by integrating DBP binding and chromatin structure data. The 3D-cooperation most often accompanies 1D-cooperation as the majority (71%) of DBP cooperation is a mixture of 3D and 1D events. Furthermore, we observed enrichment of disease-associated genotype variations in DBP cooperative binding sites, which suggests the functional importance of DBP cooperation.

Identification of cooperation between multiple DBPs has been a challenging problem. Combinatorial approaches are limited to consider cooperation between a small number of DBPs because of the exponential increase of the possible combinations. In contrast, our model can easily search combinatorial cooperation in thousands of DBPs. By identifying modules and cliques in the network, we have uncovered closely collaborated DBPs, particularly those associated through 3D interactions in chromatin loops that may be crucial for loop formation or stabilization.

Our comparative analyses between GM12878 and K562 reveals different mechanisms of achieving cell-type specificity: using different combinations of DBPs or using the same protein complex but collaborating with different partners. Interestingly, we also found spacing constraints between the binding sites of certain partners, which implies higher-order regulatory rules for not only 1D but also 3D DBP cooperation. One major limitation of this work is that it relies on high resolution Hi-C data that is only available in a limited number of cell types. Furthermore, the chromatin loops used in this project are taken directly from the study of Rao *et al.*[17] that were defined using very conservative criteria. Additional loops might be identified using less

conservative criteria or other technologies. However, as the sequencing technology rapidly evolves, these limitations will be overcome by the availability of more and more ChIP-seq and Hi-C data at even higher resolution. In conclusion, our model provides a powerful tool for integrative analysis of DBP binding and chromatin structure data in different cell types, which will facilitate the uncovering of the molecular mechanisms for transcriptional regulation and 3D chromosome organization.

Chapter 2, in full, is a reprint of the material as it appears in Systematic Identification of Protein Combinations Mediating Chromatin Looping. Zhang, Kai; Nan Li, Richard I. Ainsworth, Wei Wang. Nature Communications 2016. The dissertation author was the primary investigator and author of this paper.

2.5 References

1. Schleif, R. DNA looping. en. *Annu. Rev. Biochem.* **61,** 199–223. ISSN: 0066-4154 (1992).
2. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. en. *Nat. Rev. Genet.* **2,** 292–301. ISSN: 1471-0056 (Apr. 2001).
3. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. en. *Nature* **447,** 413–417. ISSN: 0028-0836, 1476-4687 (24 5 2007).
4. Dekker, J. Gene regulation in the third dimension. en. *Science* **319,** 1793–1794. ISSN: 0036-8075, 1095-9203 (28 3 2008).
5. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. en. *Nat. Biotechnol.* **29,** 1109–1113. ISSN: 1087-0156, 1546-1696 (20 11 2011).
6. Zhang, Y., McCord, R. P., Ho, Y.-J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C., Becker, M. S., Alt, F. W. & Dekker, J. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. en. *Cell* **148,** 908–921. ISSN: 0092-8674, 1097-4172 (Feb. 2012).
7. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. en. *Nat. Rev. Genet.* **14,** 390–403. ISSN: 1471-0056, 1471-0064 (June 2013).

8. Gorkin, D. U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. en. *Cell Stem Cell* **14,** 762–775. ISSN: 1934-5909, 1875-9777 (May 2014).

9. Mercer, T. R. & Mattick, J. S. Understanding the regulatory and transcriptional complexity of the genome through structure. en. *Genome Res.* **23,** 1081–1088. ISSN: 1088-9051, 1549-5469 (July 2013).

10. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. en. *Nature* **485,** 376–380. ISSN: 0028-0836, 1476-4687 (Nov. 2012).

11. Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. en. *Nature* **488,** 116–120. ISSN: 0028-0836, 1476-4687 (Feb. 2012).

12. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. en. *Cell* **137,** 1194–1211. ISSN: 0092-8674, 1097-4172 (26 6 2009).

13. Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N. & de Laat, W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. en. *Genes Dev.* **20,** 2349–2354. ISSN: 0890-9369 (Jan. 2006).

14. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. en. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 20398–20403. ISSN: 0027-8424, 1091-6490 (23 12 2008).

15. Vakoc, C. R., Letting, D. L., Gheldof, N., Sawado, T., Bender, M. A., Groudine, M., Weiss, M. J., Dekker, J. & Blobel, G. A. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. en. *Mol. Cell* **17,** 453–462. ISSN: 1097-2765 (Apr. 2005).

16. Drissen, R., Palstra, R.-J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S. & de Laat, W. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. en. *Genes Dev.* **18,** 2485–2490. ISSN: 0890-9369 (15 10 2004).

17. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. en. *Cell* **159,** 1665–1680. ISSN: 0092-8674, 1097-4172 (18 12 2014).

18. Zhou, Q. & Wong, W. H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. en. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 12114–12119. ISSN: 0027-8424 (17 8 2004).

19. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N.,

Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J. J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M. & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. en. *Nature* **489,** 91–100. ISSN: 0028-0836, 1476-4687 (June 2012).

20. Das, D., Banerjee, N. & Zhang, M. Q. Interacting models of cooperative gene regulation. en. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 16234–16239. ISSN: 0027-8424 (16 11 2004).

21. Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, B. & Robert, F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. en. *Genome Res.* **16,** 656–668. ISSN: 1088-9051 (May 2006).

22. Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., Cooper, G. M., Reddy, T. E., Crawford, G. E. & Myers, R. M. Distinct properties of cell-type-specific and shared transcription factor binding sites. en. *Mol. Cell* **52,** 25–36. ISSN: 1097-2765, 1097-4164 (Oct. 2013).

23. Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L. W., Janette, J., Jiang, L., Kasper, D., Kawli, T., Kheradpour, P., Kundaje, A., Li, J. J., Ma, L., Niu, W., Rehm, E. J., Rozowsky, J., Slattery, M., Spokony, R., Terrell, R., Vafeados, D., Wang, D., Weisdepp, P., Wu, Y.-C., Xie, D., Yan, K.-K., Feingold, E. A., Good, P. J., Pazin, M. J., Huang, H., Bickel, P. J., Brenner, S. E., Reinke, V., Waterston, R. H., Gerstein, M., White, K. P., Kellis, M. & Snyder, M. Comparative analysis of regulatory information and circuits across distant species. en. *Nature* **512,** 453–456. ISSN: 0028-0836, 1476-4687 (28 8 2014).

24. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

25. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). en. *Genome Biol.* **9,** R137. ISSN: 1465-6906 (17 9 2008).

26. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. en. *Biostatistics* **9,** 432–441. ISSN: 1465-4644, 1468-4357 (July 2008).

27. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., *et al.* High-dimensional semiparametric Gaussian copula graphical models. *Ann. Stat.* **40,** 2293–2326. ISSN: 0090-5364 (2012).

28. Eppstein, D., Löffler, M. & Strash, D. *Listing All Maximal Cliques in Sparse Graphs in Near-Optimal Time* in *Algorithms and Computation* **6506** (Springer Berlin Heidelberg, 2010), 403–414. doi:10.1007/978-3-642-17517-6_36. http://dx.doi.org/10.1007/978-3-642-17517-6_36.

29. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. en. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74,** 036104. ISSN: 1539-3755 (Sept. 2006).

30. Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M. S., Dolinski, K. & Tyers, M. The BioGRID interaction database: 2015 update. en. *Nucleic Acids Res.* **43,** D470–8. ISSN: 0305-1048, 1362-4962 (Jan. 2015).

31. Dempster, A. P. Covariance Selection. *Biometrics* **28,** 157–175. ISSN: 0006-341X, 1541-0420 (1972).

32. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. & Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. en. *BMC Bioinformatics* **7 Suppl 1,** S7. ISSN: 1471-2105 (20 3 2006).

33. Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. en. *Bioinformatics* **27,** 2263–2270. ISSN: 1367-4803, 1367-4811 (15 8 2011).

34. Chan, H. M. & La Thangue, N. B. p300/CBP proteins: HATs for transcriptional bridges and scaffolds. en. *J. Cell Sci.* **114,** 2363–2373. ISSN: 0021-9533 (July 2001).

35. Kalkhoven, E. CBP and p300: HATs for different occasions. en. *Biochem. Pharmacol.* **68,** 1145–1155. ISSN: 0006-2952 (15 9 2004).

36. Mayr, B. & Montminy, M. Transcriptional regulation by the phosphorylation-dependent factor CREB. en. *Nat. Rev. Mol. Cell Biol.* **2,** 599–609. ISSN: 1471-0072 (Aug. 2001).

37. Wen, A. Y., Sakamoto, K. M. & Miller, L. S. The role of the transcription factor CREB in immune function. en. *J. Immunol.* **185,** 6413–6419. ISSN: 0022-1767, 1550-6606 (Jan. 2010).

38. Nechanitzky, R., Akbas, D., Scherer, S., György, I., Hoyler, T., Ramamoorthy, S., Diefenbach, A. & Grosschedl, R. Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. en. *Nat. Immunol.* **14,** 867–875. ISSN: 1529-2908, 1529-2916 (Aug. 2013).

39. Cobaleda, C., Schebesta, A., Delogu, A. & Busslinger, M. Pax5: the guardian of B cell identity and function. en. *Nat. Immunol.* **8,** 463–470. ISSN: 1529-2908 (May 2007).

40. Liu, P., Keller, J. R., Ortiz, M., Tessarollo, L., Rachel, R. A., Nakamura, T., Jenkins, N. A. & Copeland, N. G. Bcl11a is essential for normal lymphoid development. en. *Nat. Immunol.* **4,** 525–532. ISSN: 1529-2908 (June 2003).

41. Zhuang, Y., Cheng, P. & Weintraub, H. B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB. en. *Mol. Cell. Biol.* **16,** 2898–2905. ISSN: 0270-7306 (June 1996).

42. Sun, H., Lu, B., Li, R. Q., Flavell, R. A. & Taneja, R. Defective T cell activation and autoimmune disorder in Stra13-deficient mice. en. *Nat. Immunol.* **2,** 1040–1047. ISSN: 1529-2908 (Nov. 2001).

43. Sims 3rd, R. J., Belotserkovskaya, R. & Reinberg, D. Elongation by RNA polymerase II: the short and long of it. *Genes Dev.* **18,** 2437–2468. ISSN: 0890-9369 (2004).

44. Faiola, F., Liu, X., Lo, S., Pan, S., Zhang, K., Lymar, E., Farina, A. & Martinez, E. Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription. en. *Mol. Cell. Biol.* **25,** 10220–10234. ISSN: 0270-7306 (Dec. 2005).

45. Zhang, K., Faiola, F. & Martinez, E. Six lysine residues on c-Myc are direct substrates for acetylation by p300. en. *Biochem. Biophys. Res. Commun.* **336,** 274–280. ISSN: 0006-291X (14 10 2005).

46. Dinkel, A., Warnatz, K., Ledermann, B., Rolink, A., Zipfel, P. F., Bürki, K. & Eibel, H. The transcription factor early growth response 1 (Egr-1) advances differentiation of pre-B and immature B cells. en. *J. Exp. Med.* **188,** 2215–2224. ISSN: 0022-1007 (21 12 1998).

47. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. en. *Nucleic Acids Res.* **42,** D1001–6. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

48. Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q. & Snyder, M. P. Genome-wide map of regulatory interactions in the human genome. en. *Genome Res.* **24,** 1905–1917. ISSN: 1088-9051, 1549-5469 (Dec. 2014).

49. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., Cobb, B. S., Yokomori, K., Dillon, N., Aragon, L., Fisher, A. G. & Merkenschlager, M. Cohesins functionally associate with CTCF on mammalian chromosome arms. en. *Cell* **132,** 422–433. ISSN: 0092-8674, 1097-4172 (Aug. 2008).

50. Lin, J.-X., Li, P., Liu, D., Jin, H. T., He, J., Ata Ur Rasheed, M., Rochman, Y., Wang, L., Cui, K., Liu, C., Kelsall, B. L., Ahmed, R. & Leonard, W. J. Critical Role of STAT5 transcription factor tetramerization for cytokine responses and normal immune function. en. *Immunity* **36,** 586–599. ISSN: 1074-7613, 1097-4180 (20 4 2012).

51. Betz, B. C., Jordan-Williams, K. L., Wang, C., Kang, S. G., Liao, J., Logan, M. R., Kim, C. H. & Taparowsky, E. J. Batf coordinates multiple aspects of B and T cell function required for normal antibody responses. en. *J. Exp. Med.* **207,** 933–942. ISSN: 0022-1007, 1540-9538 (Oct. 2010).

52. Ge, B., Li, O., Wilder, P., Rizzino, A. & McKeithan, T. W. NF-kappa B regulates BCL3 transcription in T lymphocytes through an intronic enhancer. en. *J. Immunol.* **171,** 4210–4218. ISSN: 0022-1767 (15 10 2003).

53. Gerondakis, S. & Siebenlist, U. Roles of the NF-kappaB pathway in lymphocyte development and function. en. *Cold Spring Harb. Perspect. Biol.* **2,** a000182. ISSN: 1943-0264 (May 2010).

54. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. en. *Cell* **91,** 35–45. ISSN: 0092-8674 (Mar. 1997).

55. 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. & Young, R. A. Mediator and cohesin connect gene expression and chromatin architecture. en. *Nature* **467,** 430–435. ISSN: 0028-0836, 1476-4687 (23 9 2010).

56. Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sal Lari, R., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-Kains, B. & Lupien, M. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. en. *Nat. Commun.* **2,** 6186. ISSN: 2041-1723 (Mar. 2015).

57. Ogawa, S., Satake, M. & Ikuta, K. Physical and functional interactions between STAT5 and Runx transcription factors. en. *J. Biochem.* **143,** 695–709. ISSN: 0021-924X (May 2008).

58. Hirata, M. C/EBPbeta and RUNX2 cooperate to degrade cartilage with MMP-13 as the target and HIF-2alpha as the inducer in chondrocytes. *Hum. Mol. Genet.* **21,** 1111–1123. ISSN: 0964-6906 (2012).

59. Hirata, M., Kugimiya, F., Fukai, A., Ohba, S., Kawamura, N., Ogasawara, T., Kawasaki, Y., Saito, T., Yano, F., Ikeda, T., Nakamura, K., Chung, U.-I. & Kawaguchi, H. C/EBPbeta Promotes transition from proliferation to hypertrophic differentiation of chondrocytes through transactivation of p57. en. *PLoS One* **4,** e4543. ISSN: 1932-6203 (20 2 2009).

60. Zervos, A. S., Gyuris, J. & Brent, R. Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. en. *Cell* **79,** following 388. ISSN: 0092-8674 (21 10 1994).

61. Luo, H., Li, Q., O'Neal, J., Kreisel, F., Le Beau, M. M. & Tomasson, M. H. c-Myc rapidly induces acute myeloid leukemia in mice without evidence of lymphoma-associated anti-apoptotic mutations. en. *Blood* **106,** 2452–2461. ISSN: 0006-4971 (Jan. 2005).

62. Salvatori, B., Iosue, I., Djodji Damas, N., Mangiavacchi, A., Chiaretti, S., Messina, M., Padula, F., Guarini, A., Bozzoni, I., Fazi, F. & Fatica, A. Critical Role of c-Myc in Acute Myeloid Leukemia Involving Direct Regulation of miR-26a and Histone Methyltransferase EZH2. en. *Genes Cancer* **2,** 585–592. ISSN: 1947-6019, 1947-6027 (May 2011).

63. Wyszomierski, S. L. & Rosen, J. M. Cooperative effects of STAT5 (signal transducer and activator of transcription 5) and C/EBPβ (CCAAT/enhancer-binding protein-beta) on β-casein gene transcription are mediated by the glucocorticoid receptor. *Mol. Endocrinol.* **15,** 228–240. ISSN: 0888-8809 (2001).

64. Xu, M., Nie, L., Kim, S.-H. & Sun, X.-H. STAT5-induced Id-1 transcription involves recruitment of HDAC1 and deacetylation of C/EBPbeta. en. *EMBO J.* **22,** 893–904. ISSN: 0261-4189 (17 2 2003).

65. Martens, J. H. A., Brinkman, A. B., Simmer, F., Francoijs, K.-J., Nebbioso, A., Ferrara, F., Altucci, L. & Stunnenberg, H. G. PML-RARalpha/RXR Alters the Epigenetic Landscape in Acute Promyelocytic Leukemia. en. *Cancer Cell* **17,** 173–185. ISSN: 1535-6108, 1878-3686 (17 2 2010).

66. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. en. *Genome Res.* **22,** 1723–1734. ISSN: 1088-9051, 1549-5469 (Sept. 2012).

67. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. en. *Bioinformatics* **27,** 1696–1697. ISSN: 1367-4803, 1367-4811 (15 6 2011).

68. Whitington, T., Frith, M. C., Johnson, J. & Bailey, T. L. Inferring transcription factor complexes from ChIP-seq data. en. *Nucleic Acids Res.* **39,** e98. ISSN: 0305-1048, 1362-4962 (Aug. 2011).

69. Ferjoux, G., Auge, B., Boyer, K. & Haenlin, M. Waltzer L. A GATA/RUNX cis-regulatory module couples Drosophila blood cell commitment and differentiation into crystal cells. *Dev. Biol.* **305,** 726–734. ISSN: 0012-1606 (2007).

70. Waltzer, L., Ferjoux, G., Bataillé, L. & Haenlin, M. Cooperation between the GATA and RUNX factors Serpent and Lozenge during Drosophila hematopoiesis. en. *EMBO J.* **22,** 6516–6525. ISSN: 0261-4189 (15 12 2003).

Chapter 3

Epigenetic landscapes reveal transcription factors that regulate CD8$^+$ T cell differentiation

3.1 Introduction

In response to infection, naive CD8$^+$ T cells differentiate into a heterogeneous population of pathogen-specific effector CD8$^+$ T cells. While the majority of these T cells undergo apoptosis after resolution of the infection, a small fraction persists as memory cells and provide lasting protection against re-infection[1]. Published studies have demonstrated that commitment to the effector or memory CD8$^+$ T cell fate occurs early after infection, and differential expression of the activation marker KLRG1 ("killer-cell lectin-like receptor G1") and cytokine receptor IL7R can be used to distinguish two effector subsets with distinct memory potential: terminally differentiated effector (TE) (KLRG1^{hi}IL7R^{lo}) CD8$^+$ T cells and memory-precursor effector (MP) (KLRG1^{lo}IL7R^{hi}) CD8$^+$ T cells[2, 3]. Numerous TFs have been identified as critical regulators of CD8$^+$ T cell fate, including T-bet, BLIMP1, ID2, IRF4 and ZEB2 for TE and effector populations;

and TCF1, EOMES, ID3, E proteins, BCL6 and FOXO1 for MP and memory populations[2–5]. Notably, not all these factors exhibit differential expression in the TE subset relative to their expression in the MP subset, which suggests that additional mechanisms contribute to their activity in promoting cell fates. Furthermore, how these TFs function within a coherent regulatory network is unknown, and additional TFs relevant to $CD8^+$ T cell differentiation remain unidentified.

We reasoned that integrated analysis of the expression and binding of TFs and expression of their target genes would provide additional insights for the identification of TFs with previously unappreciated involvement in $CD8^+$ T cell differentiation. The ATAC-seq approach (assay for transposase-accessible chromatin with high-throughput sequencing) has been used to globally probe open chromatin to map TF-binding regions with high genomic resolution with a requirement for minimal material[6, 7]. By scanning TF-binding motifs within accessible chromatin regions, it is possible to infer the binding of hundreds of TFs and identify potential gene targets of these TFs simultaneously, which has previously been technically impossible to achieve[8]. ATAC-seq has proven powerful for pinpointing TF-binding sites within regulatory elements characterized by active epigenetic marks such as promoters marked by trimethylation of histone H3 at Lys4 (H3K4me3) and enhancers associated with monomethylation of histone H3 at Lys4 (H3K4me1) and acetylation of histone H3 at Lys27 (H3K27ac)[9–11]. Additionally, trimethylation of histone H3 at Lys27 (H3K27me3) is associated with gene repression[10]. Published studies combining ATAC-seq and analysis of histone modifications have facilitated the prediction of TFs and enhancers that define tissue-specific macrophages and of lineage-determining TFs in hematopoiesis[12, 13]. In naive $CD8^+$ T cells, co-deposition of H3K4me3 and H3K27me3 at promoter regions is a signature of genes encoding products important for cellular differentiation, suggestive of an epigenetic mechanism underlying $CD8^+$ T cell differentiation[14, 15]. However, those studies focused exclusively on promoters. Accumulating evidence suggests that enhancers also have a key role in “fine-tuning” gene expression, providing better specificity than promoters[12, 16]. However, the enhancer landscapes important for the differentiation of effector and

memory $CD8^+$ T cells remain largely unknown.

Here we characterized the epigenetic landscapes of naive, TE, MP and memory $CD8^+$ T cells generated during bacterial infection to identify both enhancers and promoters important for $CD8^+$ T cell differentiation. Using ATAC-seq to identify accessible regulatory regions, we predicted TF candidates and further constructed a transcriptional-regulatory network for each subset. To facilitate the identification of key TFs, we developed a new bioinformatics method using the PageRank algorithm to rank the importance of TF in each regulatory network. We identified TFs known to be central to $CD8^+$ T cell differentiation and TFs not previously associated with specification to the $CD8^+$ T cell fate. Among those, we experimentally confirmed that the TFs YY1 ("yin and yang-1") and NR3C1 ("nuclear receptor subfamily 3 group C member 1"; a glucocorticoid receptor) promoted the TE cell phenotype and MP cell phenotype, respectively. Together our results yielded a comprehensive catalog of the regulatory elements of $CD8^+$ T cells and revealed unexpected regulators that control the fate of $CD8^+$ T cells. Furthermore, our computational framework can be applied generally to any cell or tissue type to delineate regulatory networks and identify biologically important TFs.

3.2 Methods

3.2.1 Mice, cell transfer, infection and drug treatment.

All mice were maintained in specific-pathogen-free conditions according to the instructions of Institutional Animal Care and Use Committee (IACUC) of the University of California, San Diego (UCSD). OT-I mice (specific for OVA amino acids 257–264)–MHC H2-Kb), *Tbx21*$^{-/-}$, $CD45.1^+$ congenic and C57BL/6J mice were either bred at UCSD or received from The Jackson Laboratory. We transferred 5×10^3 OT-I $CD8^+$ T cells into congenically distinct mice by intravenous (*i.v.*) injection and then infected mice intravenously with 5×10^3 colony-forming units of L. monocytogenes expressing OVA (Lm-OVA) 1 d later. For T-bet-deficient experiments, we

co-transferred 1×10^4 *Tbx21*$^{+/+}$ OT-I CD8$^+$ T cells and *Tbx21*$^{-/-}$ OT-I CD8$^+$ T cells into host mice and then infected the mice intravenously with 5×10^3 colony-forming units of Lm-OVA. For drug treatment, dexamethasone (Sigma-Aldrich) was dissolved in DMSO and diluted in PBS and then administrated to mice by intraperitoneal (*i.p.*) injection at 10 mg/kg daily after *i.v.* infection with 5×10^3 colony-forming units of Lm-OVA.

3.2.2 Antibodies and flow cytometry.

Antibodies to KLRG1 (2F1), CD127 (A7R34), CD8 (53-6.7), CD45.1 (A20-1.7), CD45.2 (104), CXCR3 (CXCR3-173), CD27 (LG-7F9), T-bet (4B10) and BCL6 (K112-91) were purchased from eBioscience. Antibodies to FOXO1 (C29H4), TCF1 (C63D9), IFN-γ (XMG1.2) and TNF (MP6-XT22) were from Cell Signaling Technology. All antibodies for flow cytometry were used at a dilution of 1:200, except BCL6, used at a dilution of 1:50. Antibodies for ChIP-seq, to H3K4me3 (Ab8580), H3K4me1 (Ab8895) and H3K27ac (Ab4729), were from Abcam. Antibody to H3K27me3 (07-449) was from Millipore. All antibodies for ChIP-seq were used at a concentration of 5 μg per 2×10^6 cells. For intracellular staining of cytokines, splenocytes were in vitro restimulated with 1 μg/ml OVA peptide (SIINFEKL) with Protein Transport Inhibitor (eBioscience) for 4 h and then fixed and permeabilized using BD cytofix/cytoperm kit (BD Biosciences). FOXP3-transcription factor staining buffer kit (eBioscience) were used for intracellular staining of transcription factors. For intracellular staining of shRNA-transduced cells containing Ametrine-reporter, cells were fixed using freshly made 2% formaldehyde for 45 min on ice and then permeabilized. All flow cytometry data were acquired by BD LSRFortessa X-20 and all cell sorting was performed on a BD FACSAria.

3.2.3 shRNA-mediated knockdown by retroviral transduction.

The detailed protocol was described previously[17]. PLAT-E cells were transfected with shRNAmir using TransIT-LT1 Reagent (Mirus). Retrovirus-containing supernatant was harvested after 48 h and mixed with 2-mercaptoethanol and polybrene (Millipore) for subsequent transductions. Purified naive OT-I $CD8^+$ T cells were in vitro activated by anti-CD3 (145-2C11) and anti-CD28 (37.51) (1 μg/ml for each; both from eBioscience) for at least 18 h and then "spinfected" at 805 g with retrovirus for 1 h at 37 °C. After 4 h of incubation, the retrovirus-containing medium was replaced by T cell medium. Transduction efficiency was measured by flow cytometry analyzing the ametrine reporter after 24 h, and 1×10^4 shRNA-transduced cells were transferred into host mice, followed by Lm-OVA infection. For Ncor1 shRNA knockdown, purified P14 $CD8^+$ T cells were in vitro activated and transduced by shRNA retrovirus similarly to OT-I $CD8^+$ T cells. Transduced P14 $CD8^+$ T cells (5×10^5) were transferred into host mice, followed by *i.p.* infection with 1.5×10^5 plaque-forming units of LCMV-C13, which results in acute infection[17]. The full hairpin sequence for shRNA was as follows: shYy1, `5'-TGCTGTTGACAGTGAGCGCCCTCCTGATTATTCTGAATAATAGTGAAGCCACAGATGTATTATTCAGAATAATCAGGAGGTTGCCTACTGCCTCGGA-3'`; and shNr3c1, `5'-TGCTGTTGACAGTGAGCGAATGCATGATGTGGTTGAAAAATAGTGAAGCCACAGATGTATTTTTCAACCACATCATGCATGTGCCTACTGCCTCGGA-3'`.

3.2.4 RT-PCR and qPCR.

For RT-PCR, RNA was extracted using Trizol (Life Technologies), followed by precipitation of isopropanol. The cDNA was synthesized using Superscript II kit (Life Technologies) following the manufacturer's instruction. For qPCR, the cDNA was quantitatively amplified using Stratagene Brilliant II Syber Green master mix (Agilent Technologies). The abundance of transcripts was normalized to that of the housekeeping gene *Hprt*. The following primers were used: *Zeb2* forward, `5'-CATGAACCCATTTAGTGCCA-3'`, and *Zeb2* reverse, `5'-AGCAAGTCTCCCT`

GAAATCC-3'; *Bcl2* forward, 5'-ACTTCGCAGAGATGTCCAGTCA-3', and *Bcl2* reverse: 5'-TGGCA AAGCGTCCCCTC-3'; *Gzma* forward, 5'-TGCTGCCCACTGTAACGTG-3', and *Gzma* reverse: 5'-G GTAGGTGAAGGATAGCCACAT-3'; *Klrb1c* forward, 5'-GACACAGCAAGTATCTACCT-3', and *Klrb1c* reverse: 5'-TACTAAGACTCGCACTAAGAC-3'; *Pou6f1* forward, 5'-GTCAGATCCTCACGAATGCTC-3 ', and *Pou6f1* reverse: 5'-GAGTCACGGCTTGGACCTG-3'; *Crtam* forward, 5'-CCTTTTCATCATCG TTCAGCTCT-3', and *Crtam* reverse: 5'-GGAGCCTGGCTGCTATTCTC-3'; *Yy1* forward, 5'-CATGT GGTCCTCGGATGAAA-3', and *Yy1* reverse: 5'-GGGAGTTTCTTGCCTGTCATA-3'; *Nr3c1* forward, 5 '-CCGGGTCCCCAGGTAAAGA-3', and *Nr3c1* reverse: 5'-TGTCCGGTAAAATAAGAGGCTTG-3'; *Hprt* forward, 5'-GGCCAGACTTTGTTGGATTT-3', and *Hprt* reverse: 5'-CAACTTGCGCTCATCTTAGG-3'.

3.2.5 Microarray analysis.

The protocol was described previously[18]. KLRG1^{hi}IL7R^{lo} TE CD8^{+} T cells and KLRG1^{lo}IL7R^{hi} MP CD8^{+} T cells (2×10^4) were sorted into TRIzol on day 8 of Lm-OVA infection. RNA was amplified and labeled with biotin, followed by hybridized to Affymetrix Mouse Gene ST 1.0 microarrays (Affymetrix). Microarray analysis was performed using GenePattern Multiplot Studio module. All data was generated in collaboration with the Immgen project (http://www.immgen.org) and passed ImmGen quality control pipeline. The gene expression data of naive and memory CD8^{+} T cells were used from a published study[18] and were normalized with the gene expression data of TE and MP subsets by RMA normalization. Given that the TE and MP subsets are highly similar "effector" populations on day 8 of infection and the finding that no genes showed a significant difference under the 1% false-discovery rate using the Student's t-test, we used a cutoff of a 1.5-fold change in expression to identify genes expressed differentially in TE and MP subsets.

3.2.6 Chromatin immunoprecipitation (ChIP), ChIP-seq library construction and sequence alignment.

Cells were fixed in 1% formaldehyde for 10 min and then quenched with 125 mM glycine for 5 min. Cells were lysed for 5 min on ice and sonicated to generate 200- to 500-bp fragments using Bioruptor sonicator (Diagenode). Sonicated DNA was used as input control. Magnetic-dynabeads (30 μl) were washed with blocking buffer twice and then mixed with 5 μg antibody in 500 μl blocking buffer and rotated at 4 °C. The sonicated lysates were first diluted to a final 0.1% SDS concentration. The diluted lysates were added to antibody-conjugated Dynabeads incubated at 4 °C. Beads were washed by Wash Buffer I, II and III for 5 min and then washed twice by TE buffer for 5 min. The beads were resuspended in 200 μl Elution Buffer and reverse-crosslinked at 65 °C overnight and then treated with RNAse for 30 min at 37 °C and Proteinase K at 55 °C for 1 h. DNA was purified by Zymo DNA Clean & Concentrator kit (Zymo Research). The purified DNA was end-repaired using End-it End-repair kit (Epicentre) and then added an "A" base to the 3' end of DNA fragments using Klenow (NEB). Then DNA was ligated with adaptors using quick DNA ligase (NEB) at 25 °C for 15 min followed by size selection of 200-400 bp using AMPure SPRI beads (Beckman Coulter). The adaptor ligated DNA was amplified using NEBNext High-Fidelity 2X PCR master mix (NEB). To prevent PCR overamplification, 1 μl DNA was first quantitatively amplified using Syber Green I master mix to determine the best amplification cycle. Then the amplified library was size-selected as 200–400 bp using SPRI beads and quantified by Qubit dsDNA HS assay kit (ThermoFisher). Finally, the library was sequenced using Hiseq 2500 for single-end 50-bp sequencing to obtain around 20 million reads for each sample. We used BWA to map raw reads to the Mus musculus genome (mm10) with following parameters: "-q 5 -l 32 -k 2". Reads with low quality (MAPQ < 30) were filtered out. If multiple reads were mapped to the same location, only one read was kept.

3.2.7 ATAC-seq and peak calling.

Cells were sorted (2.5×10^4) into 1 ml FACS buffer and spun down 500g for 20 min at 4 °C. The cell pellet was resuspended in 25 μl lysis buffer and then spun down 600g for 30 min at 4 °C. The nuclear pellet was resuspended into 25 μl transposition reaction mixture containing Tn5 transposase from Nextera DNA Sample Prep Kit (Illumina) and incubated at 37 °C for 30 min. Then the transposase-associated DNA was purified using Zymo DNA clean-up kit. To amplify the library, the DNA was first amplified for five cycles using indexing primer from Nextera kit and NEBNext High-Fidelity 2X PCR master mix. To reduce the PCR amplification bias, 5 μl of amplified DNA after the first five cycles was used for qPCR of 20 cycles to determine the number of cycles for the second round of PCR. Usually the maximum cycle of the second round of PCR is five cycles. Then, the total amplified DNA was size selected to fragments of less than 800 bp using SPRI beads. Quantification of the ATAC-seq library was based on KAPA library quantification kit (KAPAbiosystems). The size of the pooled library was examined by TapeStation. Finally, the library was sequenced using Hiseq 2500 for single-end 50-bp sequencing to obtain at least 10 million reads. To obtain confident peaks, we performed each ATAC-seq experiment at least twice and used the Irreproducibility Discovery Rate (IDR) framework to identify the reproducible peaks. In particular, we called peaks for each individual replicate as well as the pooled data from the two replicates using MACS2 with a relaxed threshold (P value, 0.01)[19]. These three sets of peaks were input for IDR analysis using a threshold of 0.05 to identify the confident set of peaks.

3.2.8 Predicting enhancers and putative TF-binding sites.

Enhancers were predicted by the RFECS algorithm using three histone marks (H3K4me1, H3K4me3 and H3K27ac). The RFECS model was trained on the active and distal P300 ChIP-seq peaks (at least 2 kb away from any transcription start site (TSS)), which were taken as

representative of enhancers (the positive set). For the non-enhancer class (the negative set), we chose promoters that overlapped DNase I hypersensitivity (DHS) peaks and random 100-bp bins that were distal (2 kb away) to any P300 site and TSS. The data sets for model training were downloaded from ENCODE with following accession codes: ENCSR000CCD (P300), ENCSR000CBF (H3K4me1), ENCSR000CBG (H3K4me3), ENCSR000CDE (H3K27ac) and ENCSR000CMW (DNase-seq). The trained model was used to scan the whole genome except the 2000 bp upstream of TSS and 500 bp downstream of TSS and classify each 100-bp bin as an enhancer or non-enhancer based on the histone modification pattern. To further reduce the false positives, we filtered the predicted enhancers using a false-discovery rate of 1%. To identify putative binding sites of TFs, we first collected 761 unique motifs from two TF-motif databases (JASPAR and UniPROBE) and one resource paper[20–22]. We then searched for TF-binding sites in 150-bp regions centered around the ATAC-seq peak summits, using the algorithm described previously[23] with a P-value cutoff of 1×10^{-5}.

3.2.9 Motif enrichment analysis at open chromatin regions.

To compute the enrichment of a TF motif over cell-type-specific open chromatin regions, we first identified the number of regions that contain at least one motif, denoted by m. If N is the number of all regions, then $\frac{m}{n}$ is considered as the enrichment score of the query motif. To construct the null model for P-value calculation, we randomly selected 10,000 regions from all open chromatin sites and computed the fraction of those regions, denoted by p, containing at least one occurrence of the motif. The P value for enrichment or depletion is then computed using the binomial test with p as the population proportion of null hypothesis.

3.2.10 Constructing TF regulatory networks.

We selected active promoters as the 5-kb regions around TSS (4 kb upstream and 1 kb downstream) that were marked by H3K4me3 peaks. Enhancers were predicted using the RFECS method based on enhancer-associated histone-modification signatures. Enhancers were linked to the nearest genes. We connected a TF to a gene if the TF had any predicted binding site in the gene's promoter or linked enhancers. We assembled all the regulatory interactions between TFs and genes into a genetic network.

3.2.11 Personalized PageRank.

The Personalized PageRank algorithm measures global influence of each node in a network, used by Google and many other companies to order search-engine results[24]. In an internet network, nodes are "web pages" and edges are "links between websites". The PageRank algorithm was designed to find out how likely a specific web page is visited if web surfers who start on a random page sampled from a given distribution have probability α of choosing a random link from the page they are currently visiting and $1-\alpha$ probability of jumping to a random page chosen from all web pages. PageRank is the stationary distribution of a random walk which, at each step, with a certain probability α jumps to a random node, and with probability $1-\alpha$ follows a randomly chosen outgoing edge from the current node. Personalized PageRank is an extension of PageRank in which all the jumps are made by a pre-defined probability distribution[25]. To give a formal definition, let $G=(V,E)$ denote a directed graph, where V is a set of nodes and E contains a directed edge (u,v) if and only if node u links to node v. We let A be the transition matrix. We defined

$$A_{ij}=\frac{1}{O(j)}$$

if node j links to node i, and $A_{ij}=0$ otherwise, where $O(j)$ is the out-degree of node j.

Given a seed vector s, the Personalized PageRank vector v is calculated by

$$v = (1-\alpha)Av + \alpha s$$

In a TF regulatory network, we set the weight of each gene to e^{z_i}, where z_i is the z-score of the expression of gene i under different conditions or in different cell states. The weights of genes are then normalized and used as the seed vector for computing personalized PageRank. For comparison between PageRank and the TF activity (TFA) metric, the TFA is a measurement of the activities of TFs[26], computed from the gene-regulatory network (GRN) and genes' expression levels. Mathematically, TFA is defined by the following equation:

$$X_i = \sum_{k \in \text{all TFs}} P_{i,k} A_k$$

where P is a matrix representing GRN, X is a vector containing the gene expression levels, and A is the TFA vector. The above equation can be written in matrix notation: $X = PA$, and TFA vector A can be solved by computing the pseudoinverse of matrix P: $A = P^{-1}X$. To compare the performance of PageRank and TFA on predicting driver TFs, we used the gene-expression profile and GRN as the input data to run both algorithm in each cell types. The "gold standard" is a set of 16 TFs that have known roles in TN, TE, MP and memory cells. 14 of those were identified from literature, and 2 were confirmed by experiments in this study. We found that PageRank successfully retrieved 12 of 16 (75%) TFs whose pattern were consistent with published reports, demonstrative of a clear advantage over the TFA metric.

3.3 Results

3.3.1 Differential gene expression by TE and MP $CD8^+$ T cells

The effector $CD8^+$ T cell population is characterized by extensive phenotypic and functional heterogeneity, including the TE and MP subsets[2]. Microarray analysis of the TE and MP subsets revealed genes expressed differentially by TE cells versus MP cells on day 8 of bacterial infection of mice, and comparison with gene-expression data for total effector and memory $CD8^+$ T cell populations indicated that many genes upregulated in the TE subset relative to their expression in the MP subset also had higher expression by total effector cells than by memory $CD8^+$ T cells[18]. This result indicated the unique transcriptional identities of effector and memory $CD8^+$ T cells could be captured by analysis of the TE and MP subsets. Notably, the differences in abundance of mRNA and protein for the majority of TFs known to control the differentiation of the TE subset versus that of the MP subset were subtle, which suggested that expression differences alone did not account for the differential dependence of distinct subsets on TFs. High-throughput RNA-based sequencing (RNA-seq) of TE and MP subsets was consistent with our microarray analyses: for genes upregulated in the TE subset relative to their expression in the MP subset, total effector $CD8^+$ T cells showed higher expression of these genes than did memory $CD8^+$ T cells, and many of the key TFs had similar expression by the TE subset and MP subset. Thus, beyond TF expression, additional regulatory mechanisms, such as the control of TF binding, might contribute to the differentiation of these two subsets and the subsequent formation of long-lived memory cells.

3.3.2 Distinct enhancer repertoires of $CD8^+$ T cell subsets

Spatial and temporal regulation of gene expression requires the specific binding of TFs at regulatory elements, which is affected by chromatin state and accessibility. We analyzed histone modifications (H3K4me1, H3K4me3, H3K27ac and H3K27me3) by chromatin immuno-

precipitation followed by deep sequencing (ChIP-seq) for characterization of potential enhancer and promoter elements, and combined that with ATAC-seq to integrate the chromatin state and accessibility of each $CD8^+$ T cell subset; this allowed us to predict the binding of TFs at specific regulatory elements. We transferred OT-I $CD8^+$ T cells (which have transgenic expression of a T cell antigen receptor that specifically recognizes a peptide fragment of ovalbumin (OVA) presented by major histocompatibility complex class I H-2Kb) into host mice, followed by infection of the host mice with Listeria monocytogenes engineered to express recombinant OVA (Lm-OVA)[18]. Naive, TE, MP and memory CD8+ T cell populations were sorted for ChIP-seq and ATAC-seq. Notably, OT-I and polyclonal $CD8^+$ T cells responding to infection showed highly correlated gene expression throughout the immune response[18], and OT-I and polyclonal effector and memory $CD8^+$ T cells displayed similar ATAC-seq profiles.

Published studies have shown that bivalent chromatin domains, comprising H3K4me3 and H3K27me3 modifications, exist in the promoters of genes encoding effector molecules in naive cells, and occupancy by H3K27me3 at these promoters is diminished when the cells differentiate into effector CD8+ T cells[14, 15]. We also observed this pattern in the change of bivalent modification of genes encoding effector molecules, including *Tbx21* (which encodes the TF T-bet) (Fig. 3.1a). Conversely, we found that genes with higher expression in naive T cells than in effector T cells, such as *Tcf7* (which encodes the TF TCF1), became repressed in effector $CD8^+$ T cells, concomitant with increased occupancy by H3K27me3 at promoters (Fig. 3.1a). Furthermore, the proportion of genes with occupancy by H3K27me3 at promoters was greater during differentiation into effector $CD8^+$ T cells than during differentiation into memory $CD8^+$ T cells (Fig. 3.1b), which suggested that epigenetic repression of genes for which naive $CD8^+$ T cells had higher expression might be essential for the terminal differentiation of effector $CD8^+$ T cells.

Focusing on distal regulatory regions of well-characterized genes in effector and memory $CD8^+$ T cells, we found both gains and losses of enhancer and repressive H3K27me3 marks.

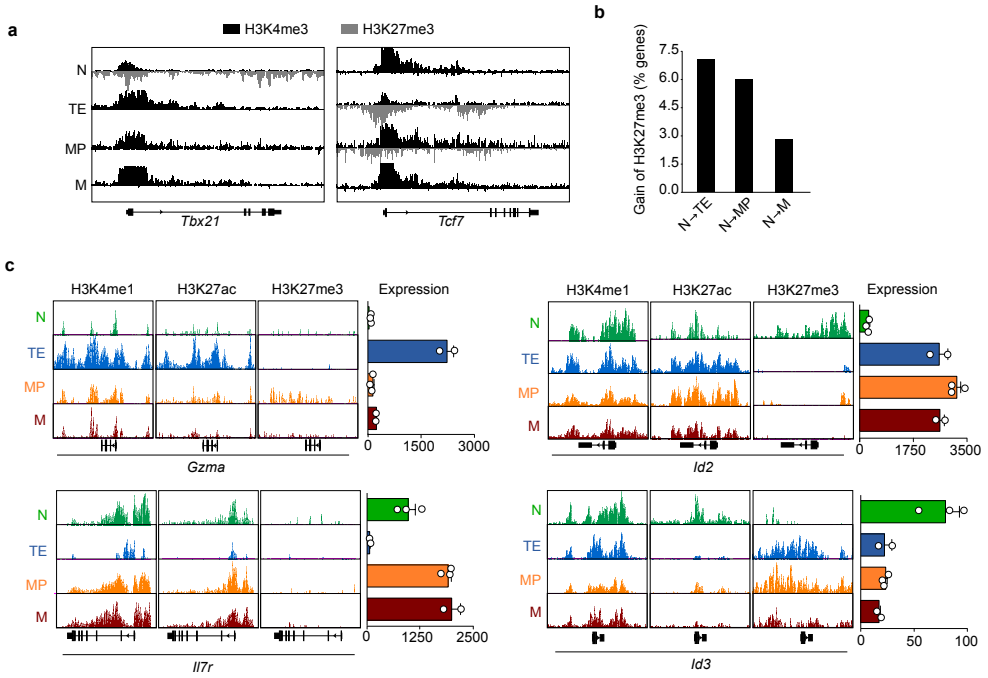

Figure 3.1: Epigenetic landscapes of CD8$^+$ T cells in response to bacterial infection. (**a**) ChIP-seq analysis of H3K4me3 and H3K27me3 (key) at *Tbx21* (left) and *Tcf7* (right) in naive CD8$^+$ T cells (N), TE CD8$^+$ T cells (TE), MP CD8$^+$ T cells (MP) and memory CD8$^+$ T cells (M) (left margin). (**b**) Frequency of genes with increased H3K27me3 at the promoter regions, among genes with a decrease in expression at various stages of differentiation (horizontal axis). (**c**) ChIP-seq analysis (left three plots) of H3K4me1, H3K27ac and H3K27me3 (above plots) at *Gzma* (top left), *Il7r* (bottom left), *Id2* (top right) and *Id3* (bottom right) in cells as in **a** (left margin), and microarray analysis of the expression of those genes (far right bar plot). Each symbol (far right) represents a biological replicate of spleens pooled from three mice. Data are pooled from two independent experiments (**a**,**b**,**c** (left three plots); $n = 10$ mice) or three independent experiments (**c** (far right plots); $n = 3$ mice; mean + s.e.m.).

For example, *Gzma*, a characteristic effector-molecule-encoding gene (that encodes granzyme A) with high expression in TE cells, was associated with increased deposition of H3K4me1 and H3K27ac after the differentiation of naive $CD8^+$ T cells into the TE subset (Fig. 3.1c). Conversely, *Il7r* (which encodes IL7R) exhibited greater deposition of H3K4me1 and H3K27ac in MP and memory $CD8^+$ T cells than in the TE subset (Fig. 3.1c), consistent with its role promoting the long-term survival of memory $CD8^+$ T cells[3, 27]. Alternatively, *Id2* and *Id3* (which encode established transcriptional regulators of $CD8^+$ T cell differentiation) exhibited substantial occupancy by H3K4me1 in all $CD8^+$ T cells but were associated with dynamic changes in the intensity of H3K27ac and H3K27me3 during differentiation[28, 29] (Fig. 3.1c). Thus, as expected, combinatorial epigenetic marks set the stage for gene expression.

To systematically identify putative enhancers, we applied the machine-learning algorithm RFECS (random forest-based enhancer identification from chromatin states)[30]. RFECS identified 27,236 enhancers, 26,561 enhancers, 23,302 enhancers and 21,883 enhancers in naive $CD8^+$ T cells, TE $CD8^+$ T cells, MP $CD8^+$ T cells and memory $CD8^+$ T cells, respectively; this constituted a non-redundant set of 52,331 putative enhancers. Upon the differentiation of naive $CD8^+$ T cells during infection with Lm-OVA, TE cells gained a greater number of newly formed enhancers than did MP cells or memory cells, while all populations lost a similar number of enhancers (Fig. 3.2a). To understand the dynamics of the usage of enhancers during differentiation, we performed k-means clustering analysis of 52,331 enhancers according to their H3K4me1 intensity. Enhancers were separated into five distinct clusters (I-V) (Fig. 3.2b). In cluster V, the intensity of H3K4me1 was maintained equivalently across the $CD8^+$ T cell subsets, and genes associated with this cluster (*Cd8a* and *Lck*) had high expression in all subsets (Fig. 3.2b). The intensity of H3K4me1 increased in clusters I and II during differentiation, and the TE subset showed for enrichment for H3K4me1 relative to its intensity in MP and memory $CD8^+$ T cells (Fig. 3.2b). Genes associated with clusters I and II (*Klrg1* and *Tbx21*) were associated with differentiation into the TE subset[2] (Fig. 3.2b). In cluster III, the intensity of H3K4me1 was higher in all

differentiated subsets than in naive $CD8^+$ T cells, and genes associated this enhancer cluster (such as *Prf1* (which encodes perforin 1)) encoded products involved in the activation of $CD8^+$ T cells (Fig. 3.2b). Conversely, for cluster IV, the intensity of H3K4me1 decreased during the differentiation of naive $CD8^+$ T cells into the TE subset and was higher in MP and memory $CD8^+$ T cells than in the TE subset (Fig. 3.2b). Enhancers of genes encoding canonical regulators of memory potential and homeostasis (*Il7r* and *Cxcr4* (which encodes the chemokine receptor CXCR4)) were in cluster IV[3, 31] (Fig. 3.2b).

To determine if differential establishment of enhancers regulates subset-specific gene expression, we assigned enhancers to the nearest genes and compared gene expression during $CD8^+$ T cell differentiation. Enhancers in clusters I, II and III were associated with genes upregulated in activated $CD8^+$ T cells, and enhancers in cluster IV were associated with genes with high expression in naive $CD8^+$ T cells. Notably, around 66% of genes with enhancers in clusters I and II were upregulated in the TE subset relative to their expression in the MP subset, while 65% of genes with enhancers in cluster IV were upregulated in the MP subset relative to their expression in the TE subset. We performed gene-ontology analysis using GREAT (genomic regions enrichment of annotations tool)[32] with the whole genome as the background set and found that clusters I and II showed enrichment for enhancers of genes encoding components of the IL12 signaling pathway (Fig. 3.2c), consistent with the role of IL12 in promoting differentiation into the TE subset[2]. In addition, cluster IV showed enrichment for enhancers of genes encoding components of the cytokine TGF-β and EGF signaling pathway (Fig. 3.2c), which suggested that these signaling pathways might favor the naive and/or memory T cell state, consistent with data showing TGF-β signaling is required for the differentiation of memory T cells[33]. We further observed that the association of genes with multiple enhancers correlated with higher expression than that of genes associated with a single enhancer (Fig. 3.2d).

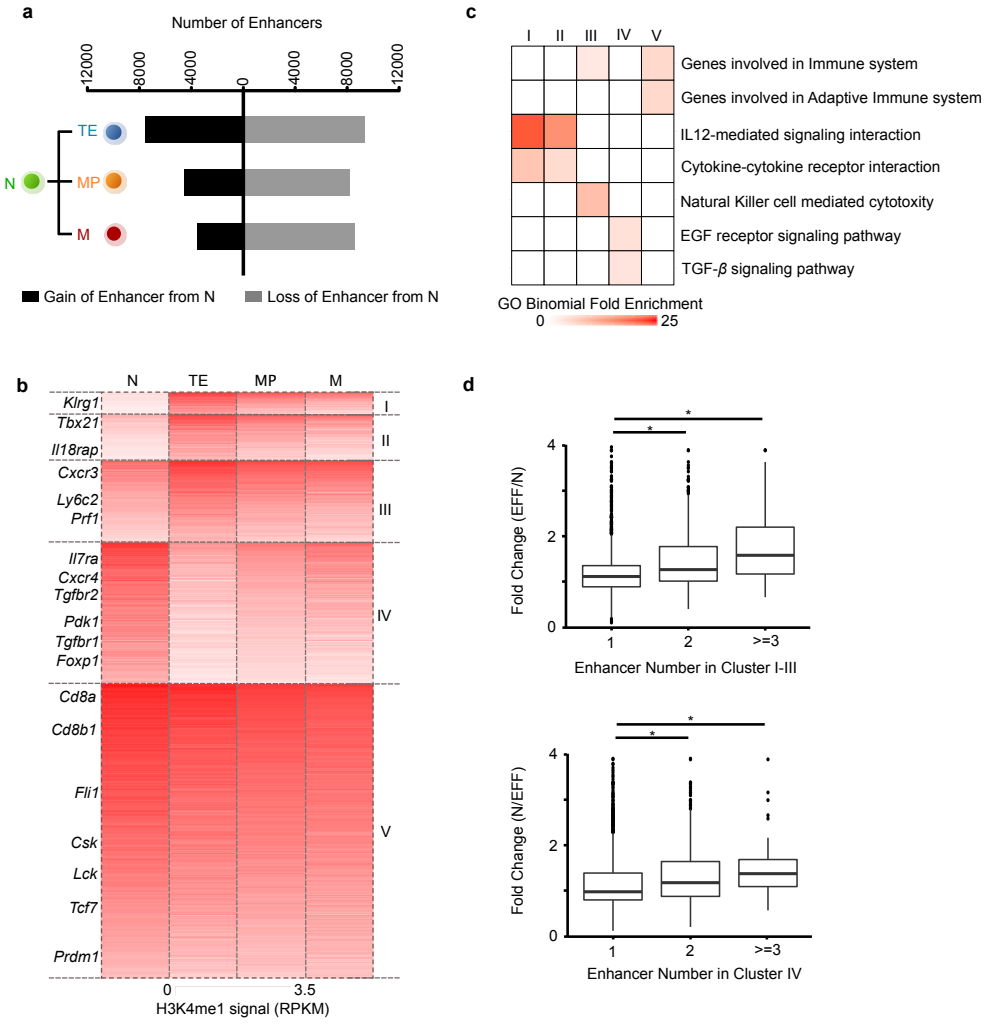

Figure 3.2: Dynamic use of enhancers is associated with differentially expressed genes during CD8$^+$ T cell differentiation. (**a**) Quantification of enhancers gained or lost (key) during the differentiation of naive CD8$^+$ T cells into TE, MP or memory CD8$^+$ T cells (left margin). (**b**) Clustering (k-means analysis; $k = 5$) of H3K4me1 signal intensity (key) in total enhancers (52,331) across CD8$^+$ T cell subsets (above plot) into clusters I–V (right margin); left margin, select genes associated with specific enhancers. RKPM, reads per kilobase per million mapped reads. (**c**) Gene-ontology (GO) analysis of the clusters in **b** (above plot), assessed with a binomial test, with the top two pathways for which the cluster showed enrichment presented (cut off binomial P value, <0.001). (**d**) Expression of mRNA from genes with one, two or three or more enhancers (horizontal axis) in clusters I–III (top) and cluster IV (bottom), presented as expression in effector CD8$^+$ T cells relative to that in naive CD8$^+$ T cells (EFF/N; top) and vice versa (N/EFF; bottom). $^*P < 0.0001$ (unpaired two-tailed Student's t-test). Data are pooled from two independent experiments (**a–c**; $n = 10$ mice) or three independent experiments (**d**; $n = 3$ mice; mean $\pm$ s.e.m. (symbols indicate outliers)).

3.3.3 TF-motif enrichment at subset-specific regulatory regions

We reasoned that accessible regulatory regions would show enrichment for TF-binding motifs relative to the abundance of such motifs in the whole genome (as background) and that we could use ATAC-seq to identify TFs important for $CD8^+$ T cell differentiation. Thus, we identified subset-specific open enhancers and promoters and then scanned 761 unique known TF-binding motifs at the center of the ATAC peaks of these regulatory regions. For example, the T-bet-binding motif appeared at a TE-specific accessible enhancer near *Zeb2* (encoding the TF ZEB2), which was expressed exclusively in the TE subset (Fig. 3.3a), in support of published findings showing that T-bet directly regulates *Zeb2* to promote differentiation into the TE subset[34, 35]. Our motif-enrichment analysis predicted the enrichment or depletion of putative binding motifs for known TFs at promoters and enhancers relative to their abundance at randomly selected open chromatin[4, 36–40]. Naive $CD8^+$ T cell subsets showed depletion of binding motifs for T-bet, BATF, SREBP2 and AP1, and differentiated $CD8^+$ T cell subsets showed enrichment for these motifs (Fig. 3.3b), consistent with the crucial role of these TFs in the activation and effector function of $CD8^+$ T cells[36–38, 41]. TE $CD8^+$ T cells showed depletion of binding motifs for TCF1, LEF1 and E2A, and naive, MP and memory $CD8^+$ T cells showed enrichment for these motifs (Fig. 3.3b), which corresponded with the well-characterized roles of these TFs in regulating the differentiation of memory populations[4, 39, 40]. Enrichment for binding motifs for some TFs (TCF1 and T-bet) was highly correlated with gene expression and function[2, 4]; in contrast, enrichment for binding motifs for other TFs (enrichment for the SREBP2-binding motif in effector T cells and for the E2A-binding motif in MP and memory cells) was consistent with their demonstrated roles (SREBP2 maintains the activation of effector T cells, and E2A promoting MP and memory cell differentiation), yet their expression remained unchanged during $CD8^+$ T cell differentiation[37, 40] (Fig. 3.3c). These data indicated that subset-specific enhancers and promoters might be established by key TFs and that putative binding of TFs, in addition to differential expression, must be considered in the identification of TF involvement.

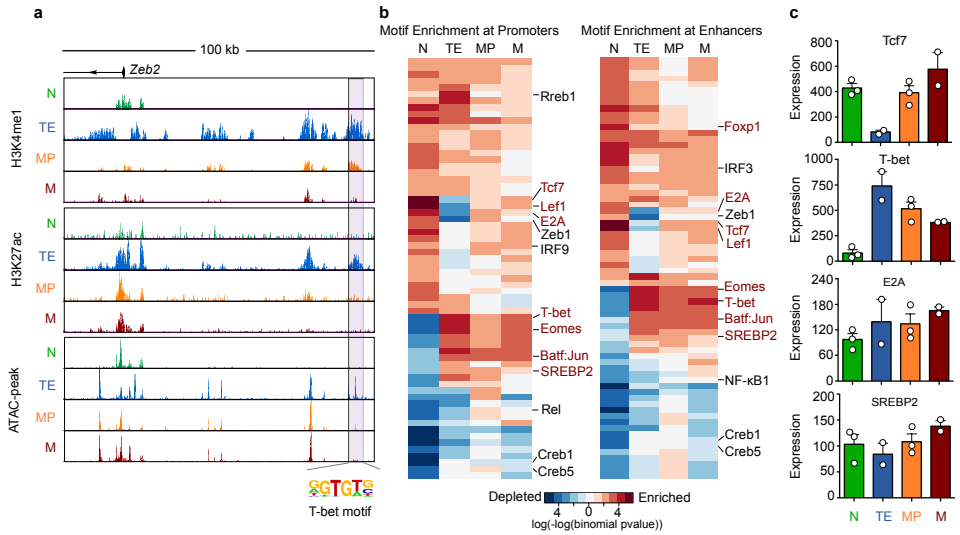

Figure 3.3: Accessible regulatory regions allow prediction of TF regulators. (**a**) ChIP-seq analysis of H3K4me1 and H3K27ac in the upstream region near *Zeb2* in naive, TE, MP and memory CD8$^+$ T cells (left margin) (top two plots), and ATAC-seq analysis of that region (bottom plot); shaded boxed area (right) indicates location of the T-bet motif (bottom right) in a TE-cell-specific enhancer. (**b**) Enrichment (red) or depletion (blue) of binding motifs for TFs (right margin; brownish red indicates TFs known to be key to the differentiation of effector or/and memory CD8$^+$ T cells) at subset-specific promoters (left) or enhancers (right) in naive, TE, MP and memory CD8$^+$ T cells (above plot), calculated by a binomial test (with randomly selected open chromatin regions as background) and presented as P values (key). (**c**) Microarray analysis of the expression of mRNA encoding the TFs TCF1 (*Tcf7*), T-bet (*Tbx21*), E2A (*Tcf3*) and SREBP2 (*Srebf2*) in cells as in **b** (horizontal axis). Each symbol represents a biological replicate of spleens pooled from three mice. Data are pooled from two independent experiments (**a**,**b**; $n = 10$ mice (ChIP-seq) or $n = 5$ mice (ATAC-seq)) or three independent experiments (**c**; $n = 3$ mice; mean + s.e.m.).

3.3.4 Construction of TF regulatory networks in CD8$^+$ T cell subsets

To elucidate TF-mediated regulatory mechanisms underlying CD8$^+$ T cell differentiation, we sought to construct a TF regulatory network in various CD8$^+$ T cell subsets. Published studies have applied correlation of gene co-expression to construct regulatory networks[42, 43]; however, this approach does not consider direct TF-binding. We combined information on TF-binding motifs, chromatin states and chromatin accessibility to predict and link TF-binding sites to their potential gene targets. We reconstructed TF regulatory networks and identified critical regulatory circuits responsible for CD8$^+$ T cell differentiation. For example, we identified a substantial number of putative targets regulated by T-bet in both the TE subset and MP subset (Fig. 3.4a). We compared the TE and MP subsets for genes predicted to be regulated by T-bet and found that 61.4% of the candidate genes were shared by these subsets; these included *Ifng* and *Cxcr3*, which are well-established targets regulated by T-bet that encode products important for effector function (interferon-γ (IFN-γ) and the chemokine receptor CXCR3, respectively)[44, 45] (Fig. 3.4b). Notably, on the basis of the subset-specific T-bet regulatory circuits, we predicted that T-bet uniquely controls the expression of *Zeb2*, *Gzma* and *Klrb1c* (which encodes the receptor NK1.1) in TE cells and *Bcl2* (which encodes the antiapoptotic protein BCL2), *Crtam* (which encodes the cytotoxic molecule CRTAM) and *Pou6f1* (which encodes the TF EMB) in MP cells. To confirm our analyses, we transferred *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ OT-I CD8$^+$ T cells together into host mice, followed by infection of the hosts with Lm-OVA. Given the loss of the TE subset in T-bet deficiency, we sort-purified total donor CD8$^+$ T cells or the MP subset from *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ mice and compared the expression of mRNA from candidate genes. There was a 200-fold decrease in *Zeb2* expression in total donor CD8$^+$ T cells in the absence of T-bet, while there was only a five-fold decrease in *Zeb2* expression in the MP subset in the absence of T-bet, all relative to its expression in their wild-type counterparts (Fig. 3.4c,d); this indicated regulation of *Zeb2* expression by T-bet in the TE subset rather than in the MP subset[35, 36]. To avoid the bias of a complete loss of TE cells among T-bet-deficient T cells, we compared mRNA abundance in

TE subsets derived from *Tbx21*$^{+/+}$ and *Tbx21*$^{+/-}$ populations and confirmed decreased expression of *Zeb2*, *Gzma* and *Klrb1c* in the TE subset with loss of T-bet relative to their expression in the wild-type TE subset (Fig. 3.4e,f). Notably, loss of T-bet affected the expression of *Bcl2*, *Crtam* and *Pou6f1* in the MP subset (Fig. 3.4d,f), which suggested that T-bet regulated these genes in an MP-cell-specific manner. The absence of T-bet resulted in a defect in the accumulation of MP cells over the course of infection (Fig. 3.4g), consistent with the finding that T-bet also regulates memory differentiation[36]. Thus, we demonstrated that T-bet positively regulated different genes in distinct CD8$^+$ T cell subsets, which highlighted the proposal that this approach allows the prediction of potential gene targets unique to different CD8$^+$ T cell subsets.

3.3.5 Identification of key TFs from PageRank-based TF ranking

Constitutively expressed TFs can exert cell-type-specific functions via regulation of the expression of distinct genes, but incorporating that knowledge for the identification of key TFs remains challenging because the TF targets are largely unknown. To overcome that limitation, we leveraged the TF regulatory network and developed a new bioinformatics method using the Personalized PageRank algorithm[24] to assess the importance of each TF in the regulatory network (Fig. 3.5a). The TF ranks determined by our method were influenced by the number of genes and the importance (determined from their expression) of genes regulated by the TF. Thus, TFs that regulate more important genes would receive higher ranks.

Using PageRank analysis, we predicted TFs important for CD8$^+$ T cell differentiation and compared our PageRank analysis with motif-enrichment analysis used by published studies[12, 13] to determine how many TFs reported previously as essential regulators of CD8$^+$ T cell differentiation could be recovered from predicted TF pools. We found that approximately half of the predicted TFs were shared by both analyses, and 25% of these shared TFs were identified in published studies. PageRank analysis revealed more known TFs than did motif-enrichment analysis: 22% of TFs among the entire pool of predicted TFs were previously reported to reg-

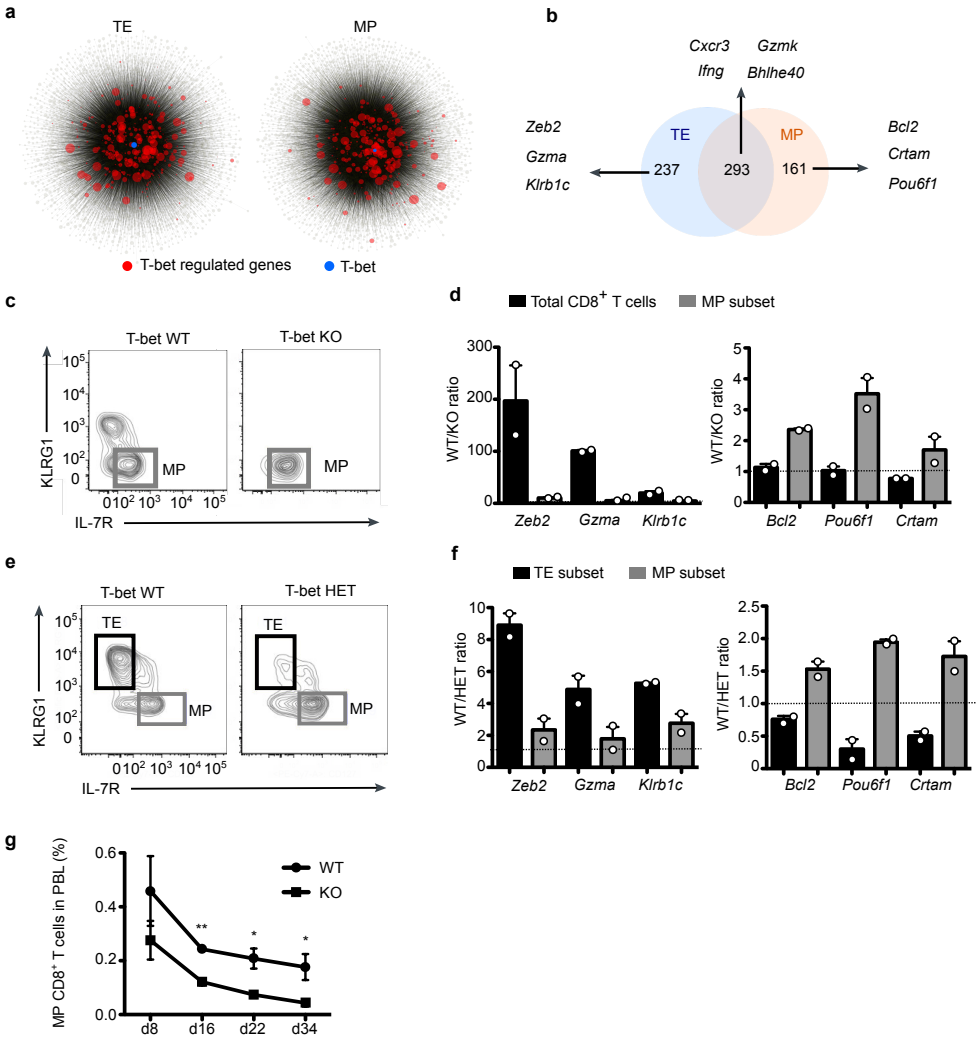

Figure 3.4: Network analysis reveals subset-specific T-bet regulatory circuits. (**a**) Global regulatory network in the TE and MP subsets: red indicates genes regulated by T-bet (blue), and size indicates their expression. (**b**) Comparison of genes regulated by T-bet in the TE and MP subsets; numbers in plot indicate total genes in set. (**c,d**) Flow cytometry of sorted *Tbx21*$^{+/+}$ (T-bet WT) and *Tbx21*$^{-/-}$ (T-bet KO) CD8^{+} T cell populations from recipient mice given co-transfer of *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ OT-I cells, followed by infection with Lm-OVA and analysis 9 d later, to identify KLRG1^{lo}IL7R^{hi} (MP) cells (outlined areas) (**c**), and RT-qPCR of mRNA encoding genes regulated by T-bet in total CD8^{+} T cells and the MP subset (key) from those mice (**d**); mRNA results are presented as expression in *Tbx21*$^{+/+}$ cells relative to that in *Tbx21*$^{-/-}$ cells (dashed lines, one-fold). (**e,f**) Flow cytometry of sorted *Tbx21*$^{+/+}$ (T-bet WT) and *Tbx21*$^{+/-}$ (T-bet HET) CD8^{+} T cell populations from recipient mice as in **c**, analyzed 8 d after infection, to identify KLRG1^{hi}IL7R^{lo} (TE) cells (top left gate) and KLRG1^{lo}IL7R^{hi} (MP) cells (bottom right gate) (**e**), and RT-qPCR of mRNA encoding genes regulated by T-bet in the TE and MP subsets (key) from those mice (**f**); mRNA results are presented as in **d**. (**g**) Frequency of MP cells among *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ peripheral blood lymphocytes (key) during Lm-OVA infection as in **c**. NS, not significant ($P > 0.05$); $^{*}P < 0.05$ and $^{**}P < 0.01$ (paired two-tailed Student's t-test). Data are pooled from (**a,b**) or representative of (**c–g**) two independent experiments ($n = 3$ mice (**c–f**) or $n = 4$ mice (**g**); mean $\pm$ s.e.m.).

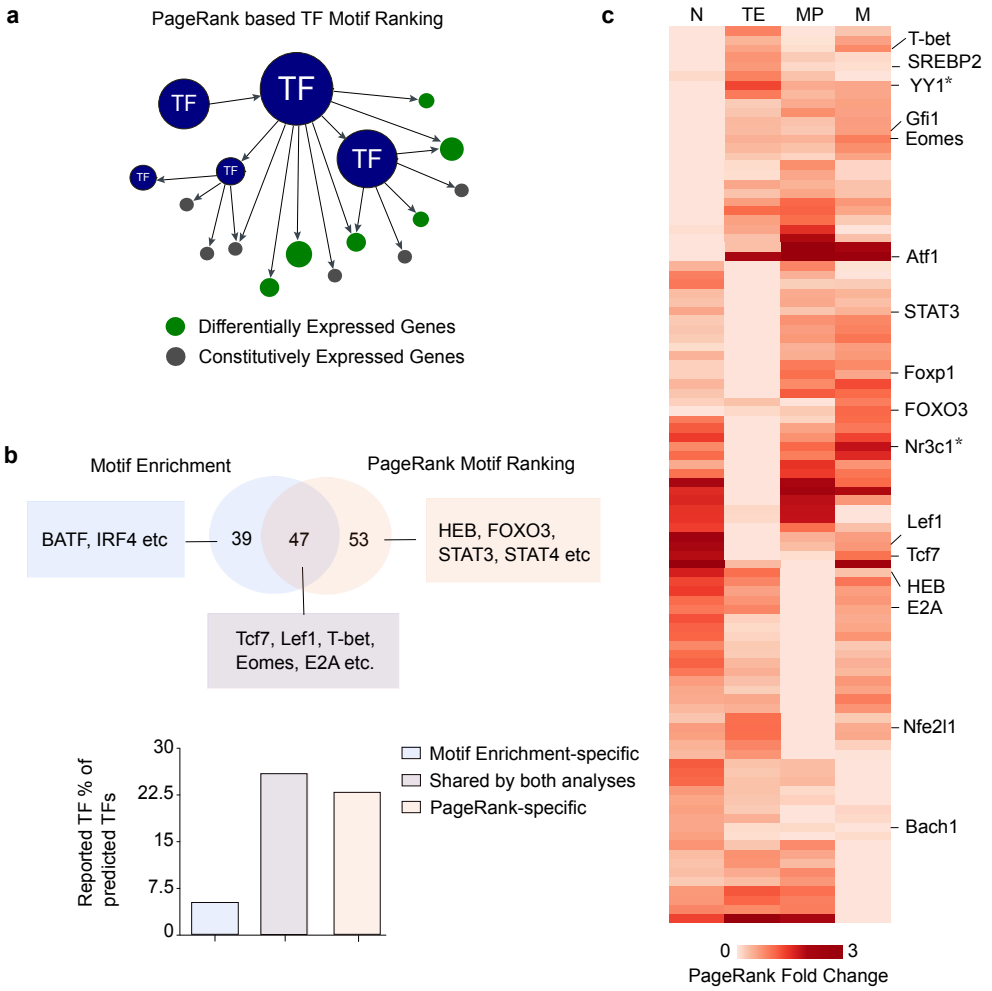

Figure 3.5: PageRank-based TF ranking highlights key TF candidates. (**a**) PageRank-based ranking of TF motifs: circle size indicates importance of gene targets (assessed by microarray analysis showing relative expression across various cell types) (green and gray) or importance of TFs (calculated from PageRank algorithm) (blue). (**b**) Comparison of PageRank analysis with motif-enrichment analysis (from Fig. 3.3) (top); numbers in plot indicate total motifs identified by PageRank (right) or motif-enrichment analysis (left) or both (middle). Below, frequency of known TFs reported previously recovered from predicted TF candidates by each analysis. (**c**) TFs with a PageRank score of at least 1.5-log-fold-change across naive, TE, MP and memory $CD8^+$ T cells (above plot), identified by PageRank analysis. Data are pooled from two independent experiments.

ulate $CD8^+$ T cell differentiation, and they were identified by PageRank analysis but not motif enrichment analysis, compared with 5% identified only by motif-enrichment analysis (Fig. 3.5b). For example, PageRank analysis assigned STAT3 a higher score in memory subsets than in the TE subset (Fig. 3.5c). That was consistent with the role of STAT3 in promoting the maturation and self-renewal of memory $CD8^+$ T cells[46]. Additionally, more TFs with known roles in $CD8^+$ T cell differentiation were identified by PageRank analysis than by another method, TF activity (TFA) analysis, which predicts the activity of TFs using the regulatory network constructed from gene-expression data[26]. These data highlighted the robustness of PageRank analysis and suggested that TFs predicted by PageRank analysis might be critical for $CD8^+$ T cell differentiation.

3.3.6 Validation of PageRank-predicted TFs

To highlight the power of PageRank analysis, we focused on YY1 and NR3C1, two regulators identified by PageRank analysis but not by the motif-enrichment analysis. Although the expression of YY1 and NR3C1 did not change during $CD8^+$ T cell differentiation, YY1 was ranked highly in the TE subset while NR3C1 was ranked highly in the MP subset (Fig. 3.5c). YY1 is a TF involved in transcriptional activation and repression and is important in immune-cell development, including the differentiation of B cells, the TH2 subset of helper T cells and regulatory T cells[47–49]. NR3C1 is a glucocorticoid receptor, which translocates into the nucleus to regulate gene expression after binding to glucocorticoids in the cytosol. NR3C1 has a critical role in development, metabolism and the immune response[50–52]. The role of YY1 and NR3C1 in the differentiation of effector or memory $CD8^+$ T cells in response to infection is unknown.

On the basis of the PageRank predictions, we hypothesized that abolishing the expression of YY1 or NR3C1 would affect formation of the TE subset or MP subset, respectively. To determine if YY1 is essential for differentiation of the TE subset, we transduced cogenically distinct OT-I $CD8^+$ T cells with retrovirus encoding short hairpin RNA (shRNA) targeting *Yy1* (shYy1) or shRNA

targeting the control gene *Cd19* (shCon) and co-transferred the cells into recipient mice, followed by infection of the recipients with Lm-OVA, then monitored the differentiation of effector T cells. Knockdown of *Yy1* resulted in a 54% reduction in its expression relative to that in cells transduced with shCon (Fig. 3.6a). Flow cytometry of $CD8^+$ T cell subsets on day 7 of infection showed a significantly lower frequency and number of the TE subset among shYy1-transduced cells than among shCon-transduced cells (Fig. 3.6b,c). In addition, the expression of MP-cell-associated molecules, including CD27, CXCR3 and TCF1, was significantly higher in shYy1-transduced cells than in shCon-transduced cells (Fig. 3.6d). Furthermore, analysis of cytokine production showed that the mean expression of IFN-γ and the number of IFN-γ-producing cells were lower for shYy1-transduced cells than for shCon-transduced cells (Fig. 3.6e). Together, these data confirmed that YY1 was important for differentiation of the TE subset.

We used a similar method (transfer of OT-I cells and infection with Lm-OVA) to determine how lowering *Nr3c1* expression, via shRNA targeting *Nr3c1* (shNr3c1), affected the MP subset differentiation. Knockdown of *Nr3c1* resulted in 86% reduction in its expression relative to that in shCon-transduced cells (Fig. 3.7a). Notably, both frequency and number of MP cells were significantly lower among shNr3c1-transduced cells than among shCon-transduced cells (Fig. 3.7b,c). Consistent with a loss of IL7R-expressing cells, expression of MP-cell-associated molecules, including CD27, CXCR3, and TCF1, was significantly lower in shNr3c1-transduced cells than in shCon-transduced cells (Fig. 3.7d), in support of the proposal of a role for NR3C1 in differentiation of the MP subset. We monitored the frequency of MP cells from day 8 to day 30 of infection and observed a decrease of the frequency of MP subset on day 30 after the loss of NR3C1 (Fig. 3.7e,f). NR3C1 has been shown to interact with co-factors such as NCOR1 ("nuclear receptor co-repressor 1") to modulate the expression of genes encoding products involved in the response to hormones[53]. Notably, knockdown of *Ncor1* via shRNA similarly affected differentiation of the MP subset; the frequency of MP cells was lower among cells in which *Ncor1* was knocked down than among cells transduced with shCon. To further confirm the role

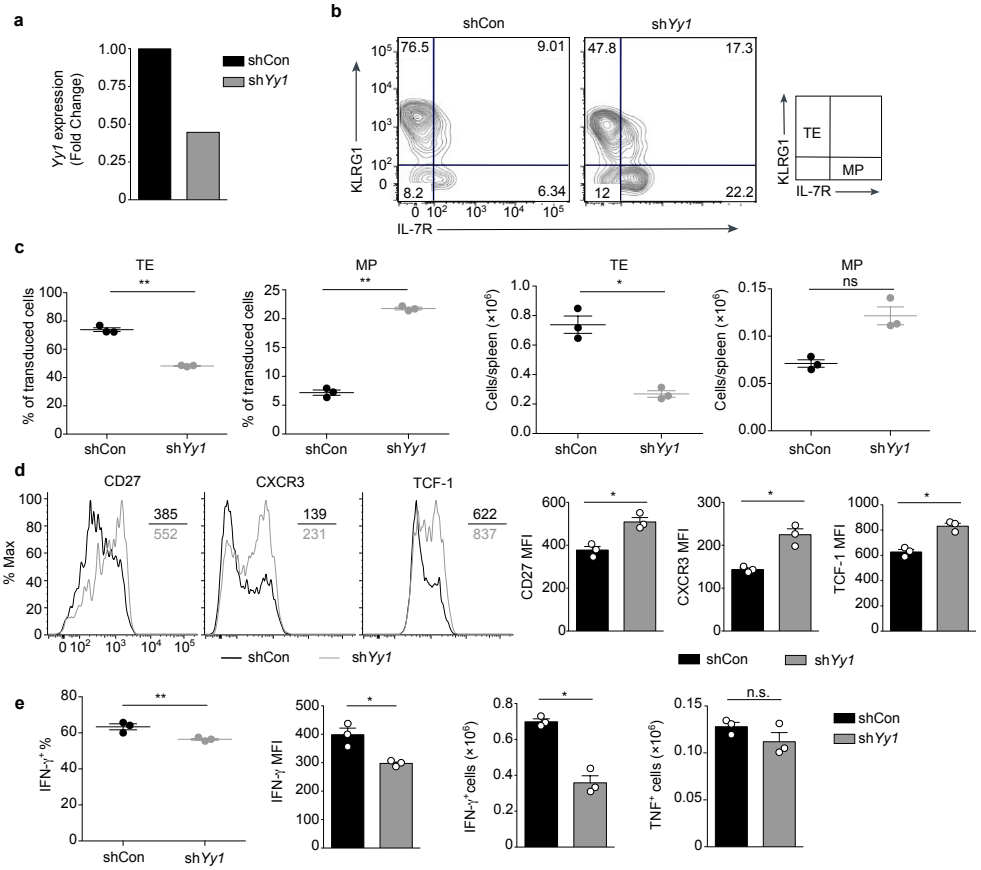

Figure 3.6: YY1 is a transcriptional regulator of the differentiation of TE CD8$^+$ T cells. (**a**) RT-qPCR analysis of *Yy1* mRNA in CD8$^+$ T cells activated *in vitro* for 72 h (with antibody to the invariant signaling protein CD3 and antibody to the co-receptor CD28) and transduced with shCon or shYy1 (key); results are presented relative to those of shCon-transduced cells. (**b**) Flow cytometry of shRNA-transduced cells from host mice given co-transfer of OT-I CD8$^+$ T cells (that had been activated in vitro (as in **a**) and transduced for 24 h with shCon or shYy1 (above plots)), followed by intravenous infection of the hosts with Lm-OVA and analysis, 7 d later, of the expression of KLRG1 and IL7R. Numbers in quadrants indicate percent cells in each throughout (far right, gating of the TE and MP subsets). (**c**) Frequency (left half) and quantification (right half) of TE and MP CD8$^+$ T cells as in **b**. (**d**) Expression of CD27, CXCR3 and TCF1 (left) in cells as in **b**; right, summary of results at left. Numbers above and below bracketed lines (left) indicate mean fluorescent intensity (MFI) of the factor assessed (colors match key). (**e**) Frequency (far left) and quantification (middle right) of IFN-γ producing cells, mean fluorescent intensity of IFN-γ (middle left), and quantification of TNF producing cells (far right) among cells as in **b** stimulated in vitro for 4 h with OVA peptide, assessed by intracellular cytokine staining followed by flow cytometry. Each symbol (**c–e**) represents an individual mouse; small horizontal lines (**c,e** (far left)) indicate the mean ($\pm$ s.e.m.). $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ (two-tailed paired Student's t-test). Data are representative of two (**a,d,e**) or three (**b,c**) independent experiments ($n = 3$ mice; mean + s.e.m. in **d,e** (middle and right)).

of NR3C1 in differentiation of the MP subset, we treated mice with synthetic glucocorticoid dexamethasone for 7 d and observed that the frequency of the MP subset increased significantly after treatment with dexamethasone. Collectively, these data demonstrated that NR3C1 promoted differentiation of the MP subset.

3.4 Discussion

The function and differentiation state of immune cells are controlled by TFs that relay environmental cues through regulation of gene expression. Efficient transcriptional regulation requires interaction between TFs and chromatin remodelers to control the binding of TFs with high fidelity. Key information is encoded in regulatory elements that contain TF-binding sequences and are associated with specific histone modifications that influence the accessibility, structure and location of those elements[16]. To identify the TF-mediated regulatory circuits critical for $CD8^+$ T cell differentiation, we systematically characterized the epigenome of $CD8^+$ T cell subsets during pathogen infection. Our global map of regulatory elements revealed a dynamic pattern of enhancer establishment that foreshadowed specific gene-expression programs. Our network analysis of T-bet regulatory circuits in distinct effector-cell subsets revealed T-bet-targets that overlapped in and were distinct in the TE subset and the MP subset. This analysis suggested a previously unknown function for T-bet in maintaining the accumulation of MP cells, potentially through regulation of the anti-apoptotic protein BCL2 and additional targets. Studies of distinct targets will further elucidate nuanced functions of T-bet in driving effector and memory fates.

Numerous crucial TFs that modulate $CD8^+$ T cell differentiation have been identified on the basis of differential gene expression and TF-gene co-expression correlation[18, 42, 43]. However, alterations in the binding of TFs without changes in expression also result in the differential expression of downstream gene targets, which makes it clear that the identification of relevant TFs exclusively on the basis of gene-expression analysis provides only partial understanding

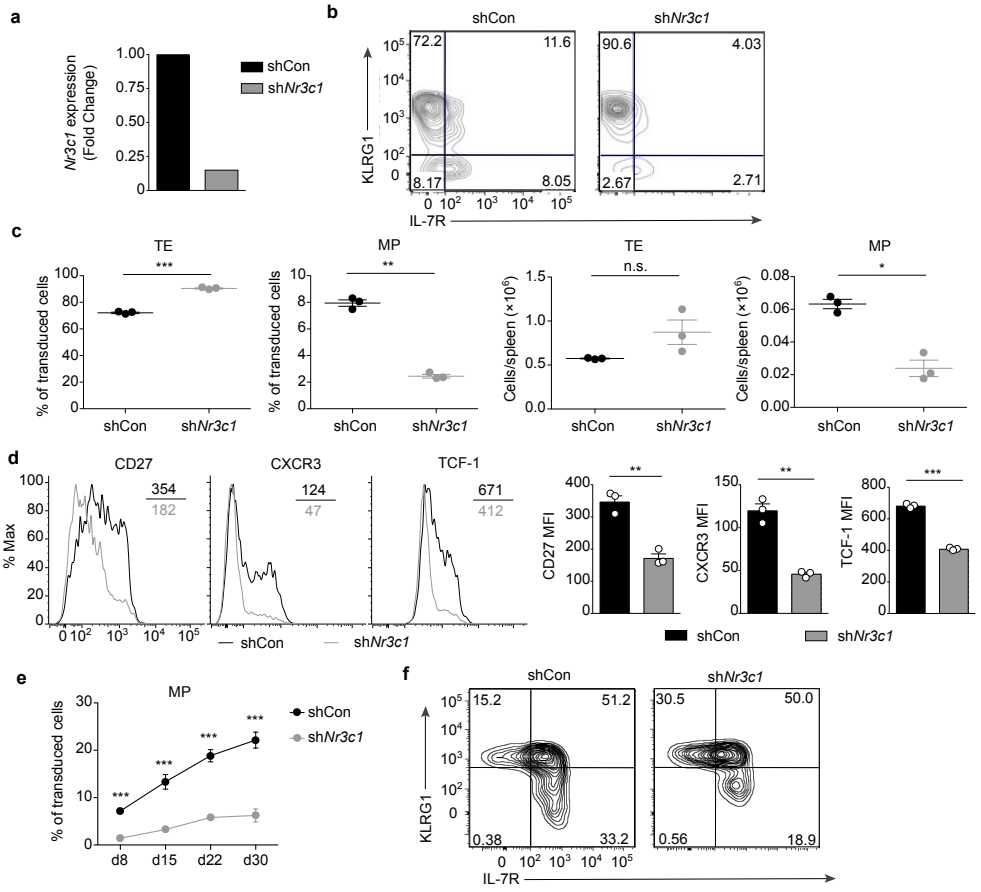

Figure 3.7: NR3C1 is essential for the formation of MP CD8$^+$ T cells. (**a**) RT-qPCR analysis of *Nr3c1* mRNA in CD8$^+$ T cells activated in vitro for 72 h (as in Fig. 3.6a) and transduced with shCon or shNr3c1 (key); results are presented relative to those of shCon-transduced cells. (**b**) Flow cytometry of shRNA-transduced cells from host mice given co-transfer of OT-I CD8$^+$ T cells (that had been activated *in vitro* (as in Fig. 3.6a) and transduced for 24 h with shCon or shNr3c1 (above plots)), followed by intravenous infection of the hosts with Lm-OVA and analysis, 7 d later, of the expression of KLRG1 and IL7R. (**c**) Frequency (left) and quantification (right) of TE and MP CD8$^+$ T cells as in **b**. (**d**) Expression of CD27, CXCR3 and TCF-1 (left) in cells as in **b**; right, summary of results at left. Numbers above and below bracketed lines (left) indicate mean fluorescent intensity (MFI) of the factor assessed (colors match key). (**e**) Frequency of MP cells among cells as in **b**, analyzed at various times (horizontal axis) after infection. (**f**) Flow cytometry of cells as in **b**, analyzed 30 d after infection (assessing expression of KLRG1 and IL7R). $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ (two-tailed paired Student's t-test). Data are representative of two independent experiments ($n = 3$ mice (**a–d**) or $n = 5$ mice (**e**); mean + s.e.m.).

of the TF networks involved. Indeed, our data demonstrated that gene expression alone could not fully explain the mechanisms behind cell-fate determination and supported the idea that the binding of TFs and gene expression should be considered together to facilitate the identification of important TFs. Differential TF binding can be achieved via numerous mechanisms, including variable chromatin state and accessibility, TF localization, the availability of co-factors, and post-translational modification of TFs. Our approach represents an advance in the efforts to achieve a comprehensive view of the regulatory networks that establish the effector and memory $CD8^+$ T cell fates by integrating data describing mRNA expression as well as chromatin states and accessibility.

For prioritization of those data, it is essential to develop new methods that rank the potential importance of TFs on the basis of the quantity and quality of the TF-regulated genes. Here, we applied the Personalized PageRank algorithm to rank the absolute importance of TFs in each subset and their relative importance across cell types by considering both binding of TFs and gene expression. Notably, our method ranked TFs by integrating two features: distinct weights for TF-regulated genes, as assessed by differential gene expression; and a hierarchy of TF-to-TF circuitry. This strategy allowed the identification of TFs that regulate relatively few but important genes, which are often overlooked by other analyses. Future modifications of gene weights by gene ontology could facilitate identification of TFs important in specific functions or pathways.

We also confirmed the functions of two TFs identified by PageRank (YY1 and NR3C1) and demonstrated their essential roles in differentiation of the TE subset and MP subset, respectively. YY1 has been shown to modulate long-range chromatin interactions of cytokine-encoding loci in TH2 cells[49]. How YY1 regulates differentiation of the TE subset and if YY1 controls chromatin interactions in the TE subset remain to be determined. The glucocorticoid receptor NR3C1 has been shown to regulate thymocyte apoptosis and inflammation responses[50–52]. Here we found that NR3C1 promoted differentiation of the MP subset, consistent with the role of glucocorticoids in inducing IL7R expression[51]. Treatment with dexamethasone increased the proportion of MP

subset during differentiation, which demonstrated a previously unknown role for glucocorticoid hormones in modulating $CD8^+$ T cell differentiation and a potential strategy for manipulating memory-cell differentiation. Thus, using our framework, we were able to both identify critical TFs and predict microenvironmental signals involved in regulating the differentiation of $CD8^+$ T cells.

Despite the successful confirmation of TFs predicted by our computational framework, additional factors could be integrated to refine our results. Global investigation of TF-binding motifs using new approaches, such as protein-binding microarrays, might be beneficial in broadening the database of known TF-binding motifs[54]. Moreover, TFs function with co-factors to regulate specific gene expression; co-binding analyses could be incorporated into these analyses to improve our network construction[55]. Furthermore, the assignment of enhancers to the nearest genes is a limited heuristic, and being able to better associate long-range enhancers with gene targets would enhance the power of our approach considerably. Published studies have shown that the interaction of enhancers and promoters is confined in topologically associated domains[56]; thus, exploration of the chromatin organization of enhancer marks as well as the use of new computational methods should facilitate the assignment of enhancers to their targets[57]. Here we have provided evidence for the involvement of many TFs in $CD8^+$ T cell immunity that were previously overlooked in this context; future studies should aim to refine and resolve the transcriptional networks by incorporating these additional approaches.

Chapter 3, in full, is a reprint of the material as it appears in Epigenetic Landscapes Reveal Transcription Factors That Regulate CD8+ T Cell Differentiation. Yu B, Zhang K, Milner J, Toma C, Chen R, Scott-Browne J, Pereira R, Crotty S, Chang J, Pipkin M, Wang W, Goldrath A. Nat Immunol 2017. The dissertation author was the primary investigator and author of this paper.

3.5 References

1. Ahmed, R. & Gray, D. Immunological memory and protective immunity: understanding their relation. en. *Science* **272,** 54–60. ISSN: 0036-8075 (May 1996).

2. Joshi, N. S., Cui, W., Chandele, A., Lee, H. K., Urso, D. R., Hagman, J., Gapin, L. & Kaech, S. M. Inflammation directs memory precursor and short-lived effector CD8(+) T cell fates via the graded expression of T-bet transcription factor. en. *Immunity* **27,** 281–295. ISSN: 1074-7613 (Aug. 2007).

3. Kaech, S. M., Tan, J. T., Wherry, E. J., Konieczny, B. T., Surh, C. D. & Ahmed, R. Selective expression of the interleukin 7 receptor identifies effector CD8 T cells that give rise to long-lived memory cells. en. *Nat. Immunol.* **4,** 1191–1198. ISSN: 1529-2908 (Dec. 2003).

4. Zhou, X., Yu, S., Zhao, D.-M., Harty, J. T., Badovinac, V. P. & Xue, H.-H. Differentiation and persistence of memory CD8(+) T cells depend on T cell factor 1. en. *Immunity* **33,** 229–240. ISSN: 1074-7613, 1097-4180 (27 8 2010).

5. Chang, J. T., Wherry, E. J. & Goldrath, A. W. Molecular regulation of effector and memory T cell differentiation. en. *Nat. Immunol.* **15,** 1104–1115. ISSN: 1529-2908, 1529-2916 (Dec. 2014).

6. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. en. *Nat. Methods* **10,** 1213–1218. ISSN: 1548-7091, 1548-7105 (Dec. 2013).

7. Winter, D. R., Jung, S. & Amit, I. Making the case for chromatin profiling: a new tool to investigate the immune-regulatory landscape. en. *Nat. Rev. Immunol.* **15,** 585–594. ISSN: 1474-1733, 1474-1741 (15 9 2015).

8. Neph, S. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150,** 1274–1286. ISSN: 0092-8674 (2010).

9. Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. en. *Nature* **488,** 116–120. ISSN: 0028-0836, 1476-4687 (Feb. 2012).

10. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. en. *Nat. Rev. Genet.* **13,** 613–626. ISSN: 1471-0056, 1471-0064 (Sept. 2012).

11. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49,** 825–837. ISSN: 1097-2765 (2013).

12. Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N. & Amit, I. Immunogenetics. Chromatin state dynamics during blood formation. en. *Science* **345,** 943–949. ISSN: 0036-8075, 1095-9203 (22 8 2014).

13. Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S. & Amit, I. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. en. *Cell* **159,** 1312–1326. ISSN: 0092-8674, 1097-4172 (Apr. 2014).

14. Araki, Y., Wang, Z., Zang, C., Wood 3rd, W. H., Schones, D., Cui, K., Roh, T.-Y., Lhotsky, B., Wersto, R. P., Peng, W., Becker, K. G., Zhao, K. & Weng, N.-P. Genome-wide analysis of histone methylation reveals chromatin state-based regulation of gene transcription and function of memory CD8+ T cells. en. *Immunity* **30,** 912–925. ISSN: 1074-7613, 1097-4180 (19 6 2009).

15. Russ, B. E., Olshanksy, M., Smallwood, H. S., Li, J., Denton, A. E., Prier, J. E., Stock, A. T., Croom, H. A., Cullen, J. G., Nguyen, M. L. T., Rowe, S., Olson, M. R., Finkelstein, D. B., Kelso, A., Thomas, P. G., Speed, T. P., Rao, S. & Turner, S. J. Distinct epigenetic signatures delineate transcriptional programs during virus-specific CD8(+) T cell differentiation. en. *Immunity* **41,** 853–865. ISSN: 1074-7613, 1097-4180 (20 11 2014).

16. 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. & Ren, B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. en. *Nature* **459,** 108–112. ISSN: 0028-0836, 1476-4687 (July 2009).

17. Chen, R., Bélanger, S., Frederick, M. A., Li, B., Johnston, R. J., Xiao, N., Liu, Y.-C., Sharma, S., Peters, B., Rao, A., Crotty, S. & Pipkin, M. E. In vivo RNA interference screens identify regulators of antiviral CD4(+) and CD8(+) T cell differentiation. en. *Immunity* **41,** 325–338. ISSN: 1074-7613, 1097-4180 (21 8 2014).

18. Best, J. A., Blair, D. A., Knell, J., Yang, E., Mayya, V., Doedens, A., Dustin, M. L., Goldrath, A. W. & Immunological Genome Project Consortium. Transcriptional insights into the CD8(+) T cell response to infection and memory T cell formation. en. *Nat. Immunol.* **14,** 404–412. ISSN: 1529-2908, 1529-2916 (Apr. 2013).

19. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). en. *Genome Biol.* **9,** R137. ISSN: 1465-6906 (17 9 2008).

20. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. & Taipale, J. DNA-binding specificities

of human transcription factors. en. *Cell* **152,** 327–339. ISSN: 0092-8674, 1097-4172 (17 1 2013).

21. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. en. *Nucleic Acids Res.* **37,** D77–82. ISSN: 0305-1048, 1362-4962 (Jan. 2009).

22. Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A. & Wasserman, W. W. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. en. *Nucleic Acids Res.* **44,** D110–5. ISSN: 0305-1048, 1362-4962 (Apr. 2016).

23. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. en. *Bioinformatics* **27,** 1017–1018. ISSN: 1367-4803, 1367-4811 (Jan. 2011).

24. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web (Bussum)* **54,** 1–17 (Nov. 1999).

25. Jeh, G. & Widom, J. *Scaling Personalized Web Search* in *Proceedings of the 12th International Conference on World Wide Web* **12** (ACM, New York, NY, USA, 2003), 271–279. ISBN: 9781581136807. doi:10.1145/775152.775191. http://doi.acm.org/10.1145/775152.775191.

26. Arrieta-Ortiz, M. L., Hafemeister, C., Bate, A. R., Chu, T., Greenfield, A., Shuster, B., Barry, S. N., Gallitto, M., Liu, B., Kacmarczyk, T., Santoriello, F., Chen, J., Rodrigues, C. D. A., Sato, T., Rudner, D. Z., Driks, A., Bonneau, R. & Eichenberger, P. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. en. *Mol. Syst. Biol.* **11,** 839. ISSN: 1744-4292 (17 11 2015).

27. Rubinstein, M. P., Lind, N. A., Purton, J. F., Filippou, P., Best, J. A., McGhee, P. A., Surh, C. D. & Goldrath, A. W. IL-7 and IL-15 differentially regulate CD8+ T-cell subsets during contraction of the immune response. en. *Blood* **112,** 3704–3712. ISSN: 0006-4971, 1528-0020 (Jan. 2008).

28. Yang, C. Y., Best, J. A., Knell, J., Yang, E., Sheridan, A. D., Jesionek, A. K., Li, H. S., Rivera, R. R., Lind, K. C., D'Cruz, L. M., Watowich, S. S., Murre, C. & Goldrath, A. W. The transcriptional regulators Id2 and Id3 control the formation of distinct memory CD8+ T cell subsets. en. *Nat. Immunol.* **12,** 1221–1229. ISSN: 1529-2908, 1529-2916 (June 2011).

29. Miyazaki, M. The opposing roles of E2A and Id3 that orchestrate and enforce the naïve T cell fate. *Nat. Immunol.* **12,** 992–1001. ISSN: 1529-2908 (2012).

30. Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M. & Ren, B. RFECS: a random-forest based algorithm for enhancer identification

from chromatin state. en. *PLoS Comput. Biol.* **9,** e1002968. ISSN: 1553-734X, 1553-7358 (14 3 2013).

31. Chaix, J., Nish, S. A., Lin, W.-H. W., Rothman, N. J., Ding, L., Wherry, E. J. & Reiner, S. L. Cutting edge: CXCR4 is critical for CD8+ memory T cell homeostatic self-renewal but not rechallenge self-renewal. en. *J. Immunol.* **193,** 1013–1016. ISSN: 0022-1767, 1550-6606 (Jan. 2014).

32. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. en. *Nat. Biotechnol.* **28,** 495–501. ISSN: 1087-0156, 1546-1696 (May 2010).

33. Ma, C. & Zhang, N. Transforming growth factor-β signaling is constantly shaping memory T-cell population. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 11013–11017. ISSN: 0027-8424, 1091-6490 (Jan. 2015).

34. Omilusik, K. D., Best, J. A., Yu, B., Goossens, S., Weidemann, A., Nguyen, J. V., Seuntjens, E., Stryjewska, A., Zweier, C., Roychoudhuri, R., Gattinoni, L., Bird, L. M., Higashi, Y., Kondoh, H., Huylebroeck, D., Haigh, J. & Goldrath, A. W. Transcriptional repressor ZEB2 promotes terminal differentiation of CD8+ effector and memory T cell populations during infection. en. *J. Exp. Med.* **212,** 2027–2039. ISSN: 0022-1007, 1540-9538 (16 11 2015).

35. Dominguez, C. X., Amezquita, R. A., Guan, T., Marshall, H. D., Joshi, N. S., Kleinstein, S. H. & Kaech, S. M. The transcription factors ZEB2 and T-bet cooperate to program cytotoxic T cell terminal differentiation in response to LCMV viral infection. en. *J. Exp. Med.* **212,** 2041–2056. ISSN: 0022-1007, 1540-9538 (16 11 2015).

36. Intlekofer, A. M., Takemoto, N., Wherry, E. J., Longworth, S. A., Northrup, J. T., Palanivel, V. R., Mullen, A. C., Gasink, C. R., Kaech, S. M., Miller, J. D., Gapin, L., Ryan, K., Russ, A. P., Lindsten, T., Orange, J. S., Goldrath, A. W., Ahmed, R. & Reiner, S. L. Effector and memory CD8+ T cell fate coupled by T-bet and eomesodermin. en. *Nat. Immunol.* **6,** 1236–1244. ISSN: 1529-2908 (Dec. 2005).

37. Kidani, Y., Elsaesser, H., Hock, M. B., Vergnes, L., Williams, K. J., Argus, J. P., Marbois, B. N., Komisopoulou, E., Wilson, E. B., Osborne, T. F., Graeber, T. G., Reue, K., Brooks, D. G. & Bensinger, S. J. Sterol regulatory element-binding proteins are essential for the metabolic programming of effector T cells and adaptive immunity. en. *Nat. Immunol.* **14,** 489–499. ISSN: 1529-2908, 1529-2916 (May 2013).

38. Kurachi, M., Barnitz, R. A., Yosef, N., Odorizzi, P. M., DiIorio, M. A., Lemieux, M. E., Yates, K., Godec, J., Klatt, M. G., Regev, A., Wherry, E. J. & Haining, W. N. The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. en. *Nat. Immunol.* **15,** 373–383. ISSN: 1529-2908, 1529-2916 (Apr. 2014).

39. Zhou, X. & Xue, H. Generation of memory precursors and functional memory CD8+ T cells depends on TCF-1 and LEF-1. *J. Immunol.* **189,** 2722–2726. ISSN: 0022-1767 (2012).

40. D'Cruz, L. M., Lind, K. C., Wu, B. B., Fujimoto, J. K. & Goldrath, A. W. Loss of E protein transcription factors E2A and HEB delays memory-precursor formation during the CD8+ T-cell immune response. en. *Eur. J. Immunol.* **42,** 2031–2041. ISSN: 0014-2980, 1521-4141 (Aug. 2012).

41. Rincón, M. & Flavell, R. A. AP-1 transcriptional activity requires both T-cell receptor-mediated and co-stimulatory signals in primary T lymphocytes. en. *EMBO J.* **13,** 4370–4381. ISSN: 0261-4189 (15 9 1994).

42. Doering, T. A., Crawford, A., Angelosanto, J. M., Paley, M. A., Ziegler, C. G. & Wherry, E. J. Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. en. *Immunity* **37,** 1130–1144. ISSN: 1074-7613, 1097-4180 (14 12 2012).

43. Hu, G. & Chen, J. A genome-wide regulatory network identifies key transcription factors for memory CD8 T-cell development. en. *Nat. Commun.* **4,** 2830. ISSN: 2041-1723 (2013).

44. Szabo, S. J., Sullivan, B. M., Stemmann, C., Satoskar, A. R., Sleckman, B. P. & Glimcher, L. H. Distinct effects of T-bet in TH1 lineage commitment and IFN-gamma production in CD4 and CD8 T cells. en. *Science* **295,** 338–342. ISSN: 0036-8075, 1095-9203 (Nov. 2002).

45. Lord, G. M., Rao, R. M., Choe, H., Sullivan, B. M., Lichtman, A. H., Luscinskas, F. W. & Glimcher, L. H. T-bet is required for optimal proinflammatory CD4+ T-cell trafficking. en. *Blood* **106,** 3432–3439. ISSN: 0006-4971 (15 11 2005).

46. Cui, W., Liu, Y., Weinstein, J. S., Craft, J. & Kaech, S. M. An interleukin-21-interleukin-10-STAT3 pathway is critical for functional maturation of memory CD8+ T cells. en. *Immunity* **35,** 792–805. ISSN: 1074-7613, 1097-4180 (23 11 2011).

47. Hwang, S. S., Jang, S. W., Kim, M. K., Kim, L. K., Kim, B.-S., Kim, H. S., Kim, K., Lee, W., Flavell, R. A. & Lee, G. R. YY1 inhibits differentiation and function of regulatory T cells by blocking Foxp3 expression and activity. en. *Nat. Commun.* **7,** 10789. ISSN: 2041-1723 (19 2 2016).

48. Hwang, S. S., Kim, Y. U., Lee, S., Jang, S. W., Kim, M. K., Koh, B. H., Lee, W., Kim, J., Souabni, A., Busslinger, M. & Lee, G. R. Transcription factor YY1 is essential for regulation of the Th2 cytokine locus and for Th2 cell differentiation. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 276–281. ISSN: 0027-8424, 1091-6490 (Feb. 2013).

49. Liu, H., Schmidt-Supprian, M., Shi, Y., Hobeika, E., Barteneva, N., Jumaa, H., Pelanda, R., Reth, M., Skok, J., Rajewsky, K. & Shi, Y. Yin Yang 1 is a critical regulator of B-cell development. en. *Genes Dev.* **21,** 1179–1189. ISSN: 0890-9369 (15 5 2007).

50. Herold, M. J., McPherson, K. G. & Reichardt, H. M. Glucocorticoids in T cell apoptosis and function. en. *Cell. Mol. Life Sci.* **63,** 60–72. ISSN: 1420-682X (Jan. 2006).

51. Franchimont, D., Galon, J., Vacchio, M. S., Fan, S., Visconti, R., Frucht, D. M., Geenen, V., Chrousos, G. P., Ashwell, J. D. & O'Shea, J. J. Positive effects of glucocorticoids on T cell function by up-regulation of IL-7 receptor alpha. en. *J. Immunol.* **168,** 2212–2218. ISSN: 0022-1767 (Jan. 2002).

52. Smoak, K. A. & Cidlowski, J. A. Mechanisms of glucocorticoid receptor signaling during inflammation. en. *Mech. Ageing Dev.* **125,** 697–706. ISSN: 0047-6374 (Oct. 2004).

53. Wang, Q., Blackford Jr, J. A., Song, L.-N., Huang, Y., Cho, S. & Simons Jr, S. S. Equilibrium interactions of corepressors and coactivators with agonist and antagonist complexes of glucocorticoid receptors. en. *Mol. Endocrinol.* **18,** 1376–1395. ISSN: 0888-8809 (June 2004).

54. Tsankov, A. M., Gu, H., Akopian, V., Ziller, M. J., Donaghey, J., Amit, I., Gnirke, A. & Meissner, A. Transcription factor binding dynamics during human ES cell differentiation. en. *Nature* **518,** 344–349. ISSN: 0028-0836, 1476-4687 (19 2 2015).

55. Zhang, K., Li, N., Ainsworth, R. I. & Wang, W. Systematic identification of protein combinations mediating chromatin looping. en. *Nat. Commun.* **7,** 12249. ISSN: 2041-1723 (27 7 2016).

56. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. en. *Nature* **518,** 331–336. ISSN: 0028-0836, 1476-4687 (19 2 2015).

57. Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., Ding, B., Li, N., Zheng, L. & Wang, W. Constructing 3D interaction maps from 1D epigenomes. en. *Nat. Commun.* **7,** 10812. ISSN: 2041-1723 (Oct. 2016).

Chapter 4

Systems-level identification of transcription factors critical for mouse embryonic development

4.1 Introduction

Transcription factors (TFs) are essential regulators of cell fate and play pivotal roles in development[1]. Identification of critical TFs that drive tissue differentiation can provide key mechanistic insights into the developmental process. Much effort has been made to identify driver TFs at different stages of development. Yet, a complete catalog of TFs underlying each developmental branch still lacks. Understanding the complex transcriptional regulatory logic during development requires dissecting the components of transcriptional networks and characterizing the genome-wide influence of key TFs.

The activity of a TF is usually defined according to its regulatory effect on target genes, which is affected by multiple factors, including synthesis of mRNA, cooperative regulation and post-translational modifications, *i.e.*, methylation, ubiquitination or phosphorylation[2]. As a

result, the expression level of a TF is not always correlated with its activity[3, 4]. In light of this, many methods have been proposed to infer the activity of TFs using statistical or machine learning approaches. For instance, Schacht et al. developed a statistical model to estimate the regulatory activity of TFs using their cumulative effects on their target genes[4]. Arrieta-Ortiz *et al.* used the linear model to infer the TF activity (TFA) by predicting target genes' expression levels[5]. Although these methods were able to predict the local activity of a TF, *i.e.*, the expression level of their direct target genes, measuring the system-wide influence of a given TF is not their focus. As genes rarely function alone but usually crosstalk with each other to form complex regulatory logic, the global activity or influence of a TF can in principle better predict the outcome of cells upon perturbation than its local activity[6, 7].

Previously, we have used the personalized PageRank to infer the global impact of TFs and achieved great success. Compared to traditional motif enrichment analysis and the TFA approach developed by Arrieta-Ortiz *et al.*, our method delivered the best performance[8]. In addition, several predicted TFs were later validated experimentally and demonstrated previously unappreciated roles[8, 9]. Building upon our previous studies, we have made further improvements by incorporating chromosome long-range interaction information into network construction, which helps assign distal regulatory elements to target genes. Furthermore, we have developed a software package, dubbed as Taiji, to help biologist to perform integrated analysis and identify driver TFs using different genomic information, including chromatin state (from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) or assay for transposase-accessible chromatin with high throughput sequencing (ATAC-seq)), gene expression profile (from RNA sequencing (RNA-seq) or microarray), and chromatin long-range interactions (from Hi-C or computational prediction).

For the first time, the ENCODE project has systematically mapped the epigenomic dynamics during mouse embryonic development in twelve tissues and eight developmental stages. While this dataset provides an unprecedented opportunity to dissect the complex transcriptional

regulatory logic during development, it also poses significant challenges for integrated computational analyses. Using our framework, we have identified TFs that are crucial in defining tissue differentiation. Furthermore, we applied Taiji to the existing data in earlier development, from 2-cell stage to embryonic stem cell, which complements the data generated by the mouse ENCODE project. Our analyses thus provide the first comprehensive catalog of key regulators for a variety of tissues during mouse embryogenesis. While many of the identified regulators are well supported by the literature, the newly discovered key TFs in development provide a valuable resource for follow-up biological study. Most interestingly, we uncovered TF combinations that activate in a spatiotemporal manner, which behave like transcriptional waves to direct the developmental progress and tissue specification.

4.2 Methods

4.2.1 Constructing TF regulatory networks

We used TF motifs, represented by position weight matrix (PWM), to predict TF binding sites at open chromatin regions within genes' promoters or enhancers, and then connected these TFs to downstream genes. We chose this strategy because it is one of the most scalable approaches for studying TF-gene regulations in a reasonably accurate way. Although TF binding sites obtained directly from experiments, such as ChIP-seq, are generally considered more accurate than computational predictions, such approaches have some significant limitations as discussed in [10]:

1. the availability of suitable affinity reagents.
2. the difficulty of interrogating the activities of multiple TFs within the same cellular environment.
3. the sizable number of TFs and cellular states that need to be studied.

Other popular methods for de novo network construction are based on gene expression correlations. These methods partially overcome the limitation of studying one TF at a time but lack directness and typically require several hundred independent gene expression perturbation studies to build a network for one cell type[11, 12].

Our network construction method contains 3 major steps:

Identifying active promoters and enhancers

Hereinafter, without further notifications, we refer to a gene's promoter as the 6 kb interval that covers the upstream 5 kb and the downstream 1 kb of the gene's transcription start site (TSS). As active promoters are usually indicated by open chromatin or active histone marks, we used ATAC-seq or H3K27ac ChIP-seq peaks to determine the activity of a given promoter. In particular, we called a promoter active if it is overlapped with at least one peak. Any gene of which the promoter is inactive was excluded from the network. To increase the sensitivity, *i.e.*, preserving more active genes in the analysis, in this step we used a relaxed cutoff (q value equals to 0.1) in MACS's[13] peak calling procedure. Distal peaks that are not overlapped with promoters were considered as enhancers.

Identifying genes' regulatory domains

The definition of gene regulatory domain was borrowed from the GREAT software[14]. Specifically, each gene was assigned a basal regulatory domain which is the gene's promoter. The gene regulatory domain was then extended in both directions to the nearest gene's basal domain but no more than 1000 kb in one direction.

Scanning TF binding sites and linking TFs to target genes

We first called ATAC-seq or H3K27ac peaks using MACS[13] with a q-value cutoff of 0.01. Next, for each peak, we identified its summit and scanned TF binding sites using FIMO's

algorithm[15] with the p-value cutoff 1×10^{-5} in the 100 bp interval centered around the summit. In this analysis, we used the mouse TF motif database curated by the CIS-BP database[16] that contains the motifs of 639 TFs. To link enhancers to their target genes, we first used long-range chromosome interactions to identify the interacting promoters for each enhancer. TF binding sites in these enhancers were then linked to corresponding promoters/genes. For the rest of the predicted TF binding sites of which the assignments cannot be made using the 3D chromosome information, we assigned them to genes according to the regulatory domains defined in the last step. Lastly, we formed a directed edge from a TF to a gene in the network if the TF has at least one binding site that is linked to the given gene.

4.2.2 Personalized PageRank

To perform personalized PageRank, we first assigned weights to the edges and nodes in the TF regulatory network.

Determine node weights

The weight of a node was calculated from the relative expression level of its representing gene. A gene's relative expression levels among different cell types were computed by applying the z-score transformation to its absolute expression levels. Suppose a gene's relative expression level in cell type i is z_i, the node weight for this gene in cell type i was then given by e^{z_i}.

Determine edge weights

As gene expression levels are heavily skewed in linear scale, we applied the log transformation to make the distribution more symmetric. The weight of a edge was given by the logarithm of the gene expression level of the source (TF).

Applying personalized PageRank

Let s be the vector containing node weights, and A be the edge weight matrix. The personalized PageRank score vector v was calculated by $v = (1-\alpha)Av + \alpha s$, where α is the damping factor (default to 0.85). The code implementing this calculation was adapted from the igraph library[17].

Statistical significance of PageRank scores

To determine the statistical significance of PageRank scores, we randomly rewired edges in the network and compute the PageRank scores. We did this multiple times and used these scores to calculate the null distribution. The p-values of PageRank scores were then inferred as to the null distribution.

4.2.3 Software implementation

Taiji has a user-friendly command line interface which is implemented in Haskell. Taiji can be easily set up and run in desktop computers. Furthermore, Taiji is designed to handle the big data through tightly integrated with workload managers that support the distributed resource management application API (DRMAA), such as the open grid scheduler and the slurm workload manager. This feature allows Taiji to analyze numerous datasets in parallel on high-performance computing clusters. Besides, Taiji has a built-in workflow manager, featuring automatic checkpointing and data recovery, allowing continuation of the analysis after interruption by error crash or power outage.

4.2.4 Computational validation of the Taiji framework

Validation of the network construction method

Considering the computational feasibility, we only selected 24 out of 72 samples to perform validation. They are: embryonic facial prominence at E10.5 and E15.5, forebrain at E10.5 and P0, midbrain at E10.5 and P0, hindbrain at E10.5 and P0, heart at E10.5 and P0, limb at E10.5 and E15.5, stomach at E14.5 and P0, liver at E11.5 and P0, neural tube at E11.5 and E15.5, kidney at E14.5 and P0, intestine at E14.5 and P0, lung at E14.5 and P0.

For every sample pair, we selected the top 500 most expressed genes from each sample. The log-fold-changes of these genes are the response variables of our prediction model. According to the constructed network, we identified the upstream regulators of those genes and computed their log-fold-changes. These were the predictors for the model. These procedures were repeated for all 276 sample pairs, and the results were combined. In total, we got 158,600 records for training a random forest model with 10-fold cross-validation. The control networks were generated by randomly assigning regulators to genes while keeping the number of regulators for each gene unchanged.

Benchmarking different ranking algorithms

The gene expression profiles of wild-type and knock-out experiments are simulated using the GeneNetWeaver[18]. The yeast and E. coli networks were provided by GeneNetWeaver. And additional 20 sub-networks are sampled from the yeast network. Each contains at least 50 regulators out of 1000 genes.

4.2.5 Prediction of chromatin interactions using EpiTensor

TADs were taken from the previous study[19]. The input histone modification data were downloaded from the ENCODE data portal, including a common set of 8 histone marks at 7

time points in 5 tissues: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac and H3K9me3, at developmental stages of E11.5, E12.5, E13.5, E14.5, E15.5, E16.5 and P0, in heart, liver, forebrain, midbrain and hindbrain. EpiTensor performed tensor analysis in the 4 dimensional space composed of tissue, time point, histone mark, and locus dimensions. The peaks in the eigenlocus vectors represent co-variation of histone modification signals that indicate 3D interactions of the corresponding loci. We considered the first 40 eigenlocus vectors, capturing on average 96.98% of total variance. The peaks in each eigenlocus vector were called by comparing to randomly shuffled background. FDR and p-value were calculated for each peak. In this study we used a stringent FDR=0.005 as cutoff.

4.2.6 Lineage tree construction

Lineage trees were constructed by running the FastME algorithm[20] on normalized ranking matrix, which was obtained from row-wise z-score transformation of the original matrix.

4.2.7 Identification of driver TFs for tissues

The output of Taiji pipeline is a matrix, consisting of TFs' ranking scores from different experiments. The rows and columns of the matrix represent TFs and experiments respectively. To identify driver TFs, we first removed the rows (TFs) with *CV*s less than 1. We then grouped the columns (experiments) from various stages together if they are from the same tissue, and averaged the ranking scores in each group. Assuming the average ranking scores are normally distributed, we calculated the deviation from the center for each score and computed the p-value. A 0.01 p-value cutoff was used for calling driver TFs.

4.3 Results

4.3.1 An overview of the Taiji framework

To identify key regulators, Taiji integrates various genomic information to build transcriptional regulatory networks by predicting regulations between TFs and genes. The TFs in the network are then ranked by the personalized PageRank algorithm[21] (Fig. 4.1a). Briefly, Taiji starts by identifying active regulatory elements, including active promoters and active enhancers, defined by ATAC-seq or H3K27ac ChIP-seq peaks. Enhancers were then assigned to their interacting promoters using chromosome loops predicted by EpiTensor[22]. To construct transcriptional regulatory networks, Taiji scans each regulatory element to identify putative TF binding sites using motifs from the CIS-BP database[23]. TFs with putative binding sites in promoters or enhancers are then linked to their target genes in the networks. Finally, the PageRank algorithm was used to assess the genome-wide influences of TFs. Furthermore, we used the node weights and edge weights to personalize the ranking algorithm. The node weights were determined by the z-scores of gene expression levels, which allows the ranking algorithm to give higher ranks to TFs that regulate more differentially expressed genes (DEGs). The edge weights were set to be proportional to TFs' expression levels, which helps filter out TFs that are not expressed or with low expression levels.

4.3.2 Predicting long-range chromosome interactions in mouse embryonic development

Due to the technical difficulties, the HiC experiments were not performed in the mouse ENCODE project. Therefore, we resorted to EpiTensor[22], an unsupervised learning method, to predict the enhancer-promoter interactions from histone modification data. As we have demonstrated in our previous study, EpiTensor not only shows high concordance with the experimen-

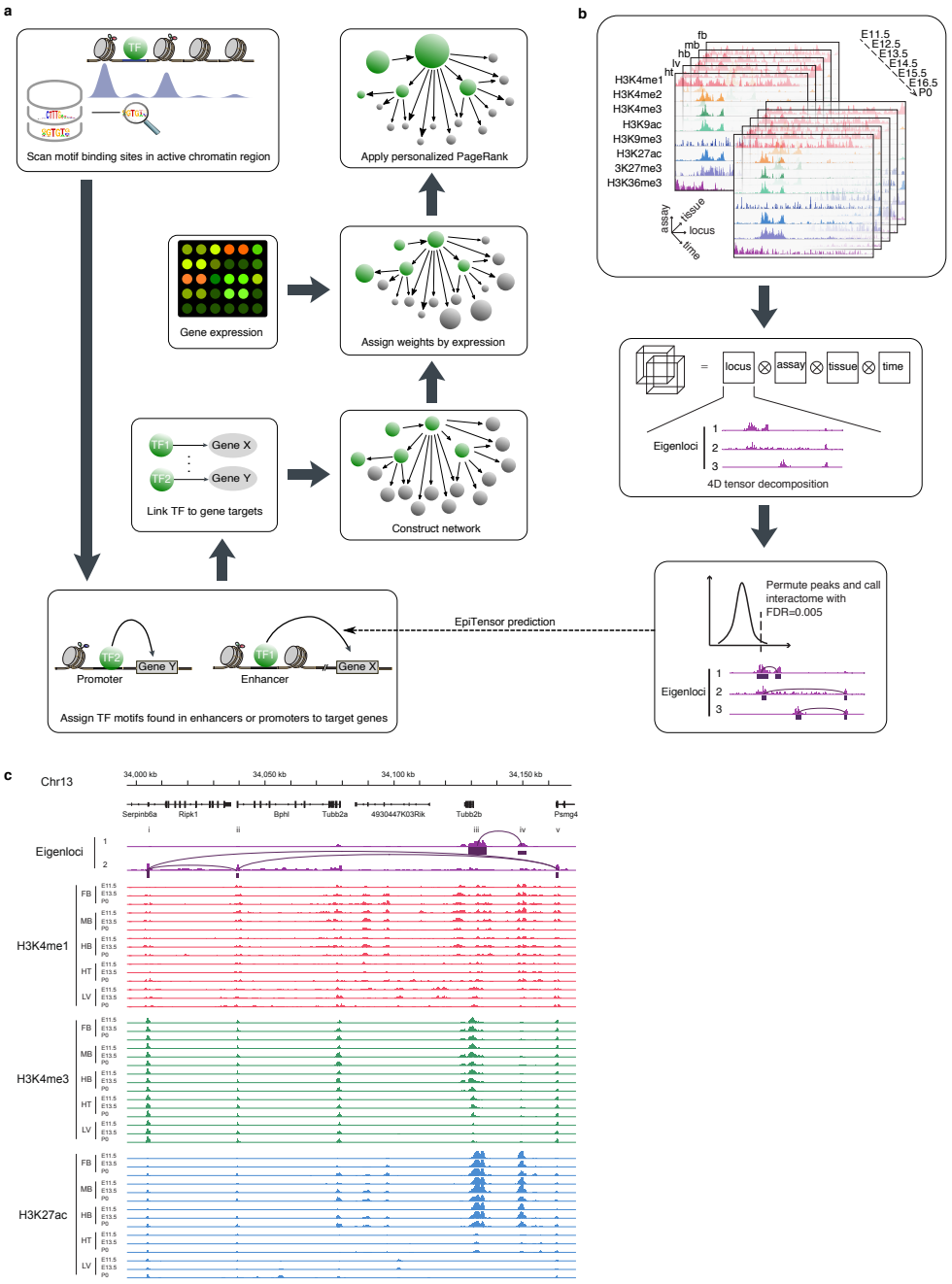

Figure 4.1: The design of Taiji and EpiTensor algorithms. (**a**) The workflow of Taiji pipeline. (**b**) The workflow of EpiTensor algorithm. (**c**) An example of chromosome interactions predicted by EpiTensor. Predicted interacting loci are linked by lines. Locus i, ii and v are three active promoters marked by H3K4me3 and H3K27ac in the tissue-stages shown in the figure. Locus iv is a validated enhancer documented in the Vista database. Locus iii is the promoter of *Tubb2b*, preferentially active in the brain. Eigenlocus 1 contains a strong peak representing the interaction between locus iii and iv; the co-variation of histone modifications at the two loci can be exemplified by the coherent decrease of H3K4me1/H3K4me3 and H3K27ac signal levels across tissues/stages. Only H3K4me1, H3K4me3, and H3K27ac in E11.5, E13.5 and P0 are shown due to the limited space.

tal data including Hi-C, ChIA-PET and eQTL results, it also significantly outperforms other correlation-based methods and nearest-gene assignment[22]. Besides, a unique advantage of EpiTensor is that it can predict chromosome interactions at a high resolution of 200 bp within topologically associating domains (TADs) (Fig. 4.1b); this is much higher than the highest resolution (1k bp) of the Hi-C data available in the public domain. Using Epitensor, we predicted the chromosome interactions in twelve tissues at eight stages. All predictions are available at http://wanglab.ucsd.edu/star/MouseENCODE/driverTF/index.html. The predicted interactions cover 25,358 transcription start sites (TSSs), >38% of all annotated transcripts (from Gencode vM14[24]), and 334 experimentally validated enhancers in the Vista database[25], account for 55% of confirmed active enhancers during mouse embryonic development. In Figure 4.1c, we show one example of the predicted interactions. Specifically, locus iii is the promoter of *Tubb2b*, a critical gene for the cortical formation and brain morphogenesis[26]. Locus iv is an experimentally validated active enhancer in embryonic mouse midbrain, hindbrain and facial mesenchyme from the Vista database (peak ID: mm1605). These two loci show correlated histone modification profile across tissues/stages and were identified as interacting loci by Epitensor. No loops were found between locus iv and v despite their closer distance. This result, together with the large distance (160 kbp) between locus i and v, indicates the power of EpiTensor for identifying long-range chromosome interactions.

4.3.3 Computational validation of the Taiji framework

Network construction and PageRank procedure are the two critical components deciding the overall accuracy of Taiji's results. To validate the network construction method, we trained a random forest model to predict the expression changes of target genes from the expression changes of their upstream regulators in the network (Fig. 4.2a). Specifically, we selected 24 samples from twelve tissues at their early and late developmental stages (see Methods for details). For each pair of the samples, we selected 500 most expressed genes from each sample. The

log-fold-changes of these genes were used as the prediction targets (response variable in the model). We next identified the upstream regulators of those genes in the constructed network, whose log-fold-changes were used as the predictors in the model. This procedure was repeated for all 276 sample pairs and in total we obtained 158,600 data points, which were used for training the random forest model with 10-fold cross-validation. The average Spearman's correlation between the experimental and predicted values is 0.756. As a comparison, we generated randomized networks, in which regulators were randomly assigned to genes while the number of regulators for each gene was kept unchanged. We repeated such randomization ten times and the average Spearman's correlation from the ten randomizations is 0.192 ± 0.001. This observation confirmed that the constructed network is biologically relevant and reliable.

To investigate whether the PageRank scores accurately reflect TFs' importance, we used the GeneNetWeaver[27] to generate benchmark networks and associate them with dynamic models. According to the dynamic models, we run simulations to generate the expression profiles and performed in silico knock-out experiments for each regulator. The Euclidean distance between wide-type and knock-out expression profiles was used to represent the magnitude of the perturbation. Regulators were ranked by their perturbation magnitudes, representing their importance to the cells. We used this rank list as the ground truth to assess the performance of different algorithms. Among the five methods we have benchmarked, PageRank performed the best, showing a strong Spearman's correlation ($\rho = 0.914$) with the ground truth in the benchmark dataset of yeast (Fig. 4.2b). Gene expression of a TF is not a strong predictor of its importance, as indicated by the weak correlation ($\rho = 0.393$) with the perturbation magnitude (Fig. 4.2e). The regression-based TFA model[5] does not correlate with the ground truth ($\rho = -0.134$) (Fig. 4.2f). Another method, developed by Schacht and his colleagues, defines the TF activity as the weighted average of its target genes' expression levels (WATG)[4]. Although it was reported as a good predictor for gene expression levels, it shows only a weak correlation with TFs' overall importance ($\rho = 0.225$, Fig. 4.2d). We hypothesized that the averaging effect in WATG might contribute negatively to its

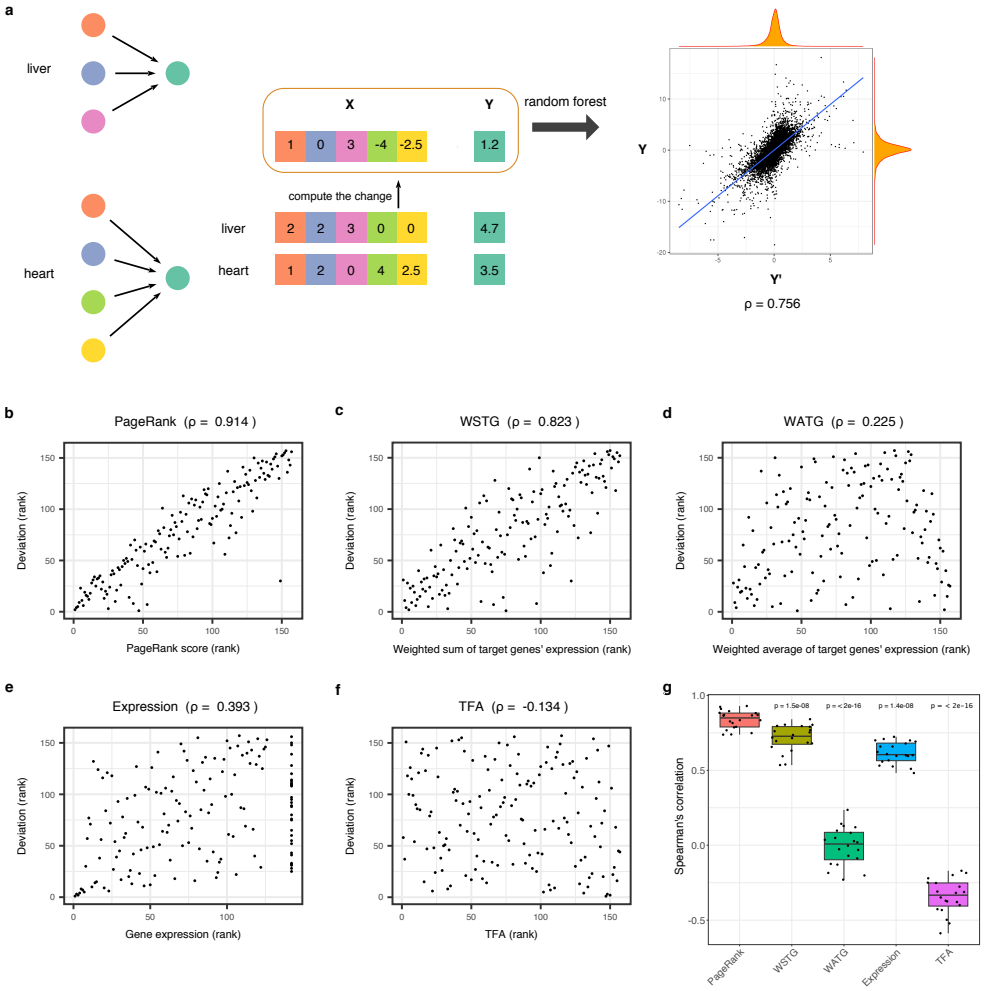

Figure 4.2: Computational validation of the Taiji framework. (**a**) An illustration of the network validation process. The log-fold expression changes of genes and their regulators between two tissues were computed and used to train a random forest model. The performance was assessed by the Spearman's correlation between predictions and "ground truth" with 10-fold cross-validation. The average correlation is 0.756. (**b-f**) The performance of various metrics, PageRank (**b**), weighted sum of target genes' expression (WSTG) (**c**), weighted average of target genes' expression (WATG) (**d**), gene expression (**e**), and TFA (**f**), on predicting TFs' importances, assessed by the Spearman's correlation with TFs' perturbation magnitudes resulted from simulated knockouts in yeast ($n = 157$). (**g**) The performances of various metrics on predicting TFs' importances in 20 sub-networks extracted from the yeast gene regulatory network.

performance on predicting the global influence. Therefore, we developed a similar metric based on WATG which calculates the weighted sum of the target genes' expression levels (WSTG), instead of averaging them. Note that the formulation of WSTG is conceptually equivalent to the network degree centrality. As expected, it performs much better than WATG ($\rho = 0.823$, Fig. 4.2c), because it takes into account the accumulated effects of all regulatees. In summary, algorithms that focus on measuring the "local" TF activities, *i.e.*, TFA and the method developed by Schacht *et al.* performed much worse than the network centrality-based methods (PageRank and WSTG), and PageRank showed the best performance.

A major advantage of PageRank algorithm, compared to the other simpler centrality metrics like WSTG, is that it is defined recursively and hence depends not only on the number of nodes linked to it but also on those nodes' PageRank values. Therefore, PageRank performed better than WSTG for the yeast network. To further confirm PageRank's superior performance, we conducted additional benchmarks using the E. coli network, and another 20 random subnetworks extracted from the yeast network (Fig. 4.2g,4.3). In all the tested cases, PageRank consistently outperformed WSTG and other methods.

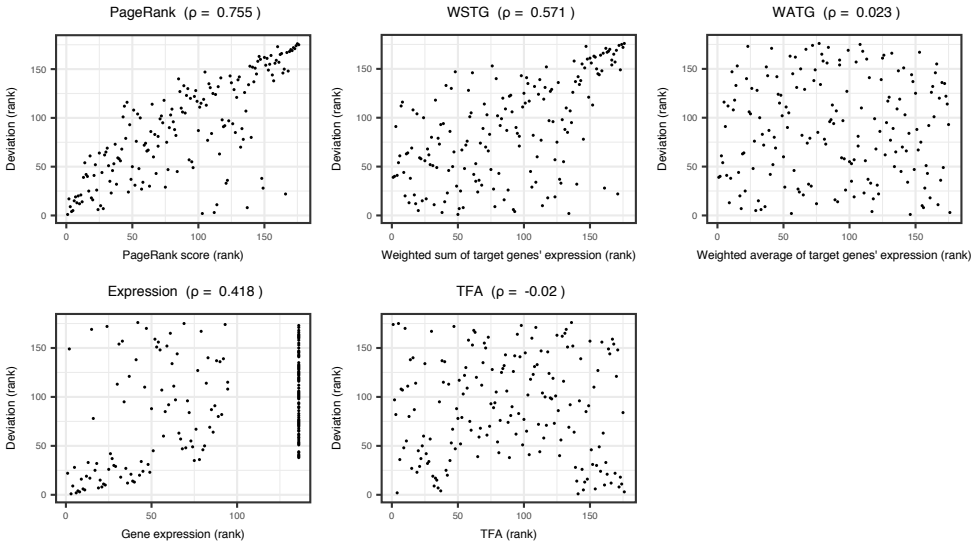

Figure 4.3: Comparing different ranking methods in *E. coli* network.

4.3.4 Taiji reveals driver TFs during embryogenesis

We constructed transcriptional regulatory networks in twelve tissues at eight developmental stages. The average number of nodes and edges in these networks are 16,187 and 320,227 respectively. 3.95% (639) of the nodes are TFs. On average, each TF regulates 501 genes, and each gene is regulated by 19 TFs (Fig. 4.4). To characterize the dynamics of TFs' global influences across different tissues and stages, we removed TFs that do not present significant variations (coefficients of variance less than 1) in their PageRank scores, which gave us a list of 245 most variable TFs. We plotted the PageRank score matrix in Figure 4.5a, ordering the samples (columns) by tissue types. Remarkably, we found that different tissues have quite distinct TF regulatory patterns (Fig. 4.5a left). In contrast, when ordering the same data by developmental stages, no discernible patterns can be observed (Fig. 4.5a right), suggesting that transcriptional regulation largely takes place in a tissue-dependent fashion. To investigate the relationship between PageRank scores and expression levels, we calculated the Spearman's correlation between PageRank scores and expression levels for the 245 TFs (Fig. 4.5b). While 60% of TFs' expression levels strongly correlate with their PageRank scores ($\rho > 0.75$), a significant portion (9.3%) of them show weak or no correlations ($\rho > 0.25$). The rest of the TFs' expression levels have moderate correlations with PageRank scores ($0.25 < \rho < 0.75$). The fact that over a half of the TFs' expression levels are highly correlated with their PageRank scores may explain why gene expression achieves the third-best in our in silico validation (Fig. 4.2g).

Using the FastME algorithm[20], we constructed a lineage tree based on the PageRank scores of 245 TFs (Fig. 4.5c). The result shows that the samples are mostly grouped by their tissue types. A mixture is only observed for four closely related samples (forebrain at E10.5, midbrain at E11.5, midbrain at E10.5, and craniofacial prominence at E10.5), which may be due to the difficulty of dissecting these tissues at their early stages. As a comparison, we found many unexpected clusters in the lineage tree constructed using the gene expression profile of the same 245 TFs (Fig. 4.5d). For instances, intestine at P0 is grouped with liver samples; Craniofacial

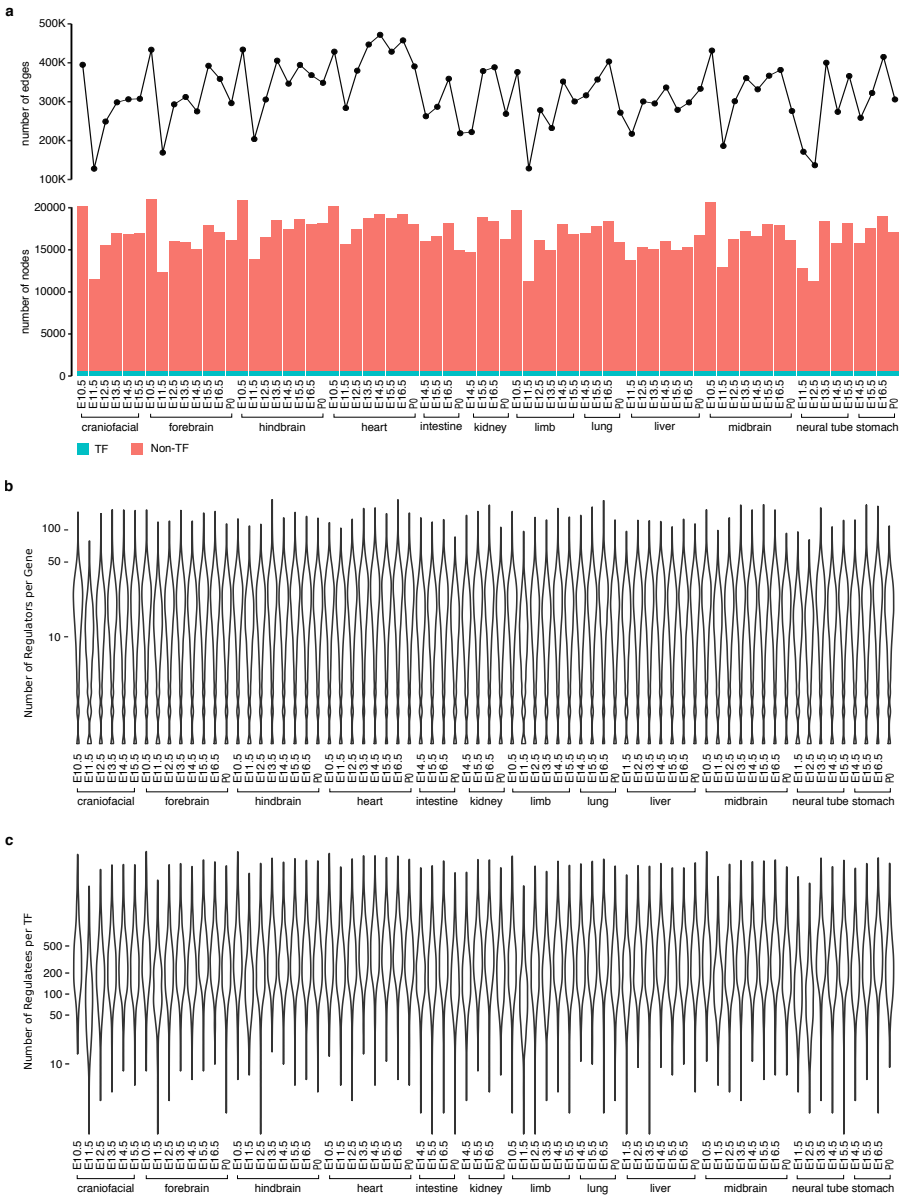

Figure 4.4: The topological properties of genetic networks. (**a**) The number of edges (top) and nodes (bottom) in each genetic network; (**b**) The distribution of the number of regulators per gene for each genetic network; (**c**) The distribution of the number of regulatees per TF for each genetic network.

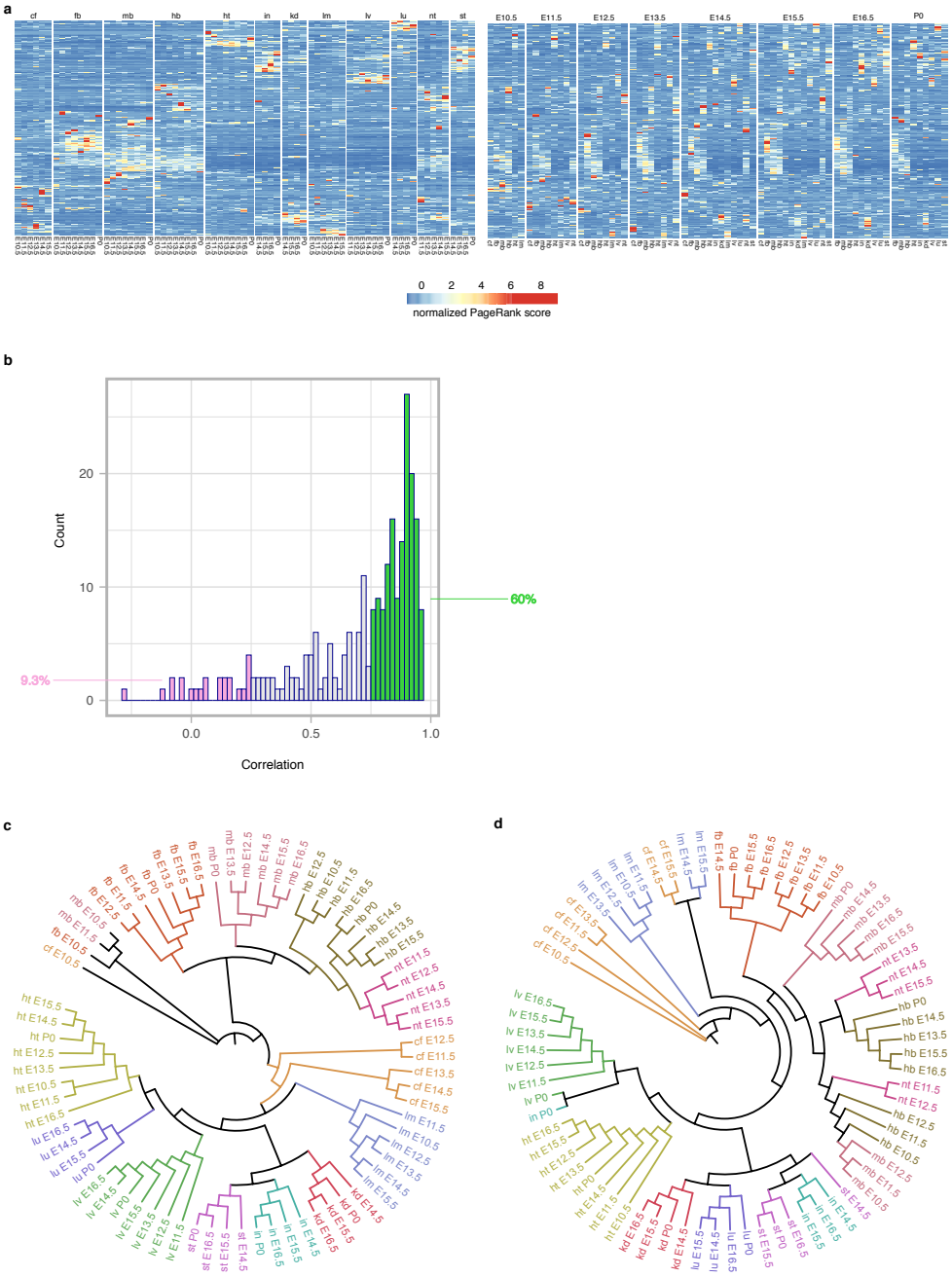

Figure 4.5: TF activity determined from Taiji accurately predicts tissue specification. (**a**) Two different views, arranged by tissue types (Left) and arranged by stages (Right), of the 245 most variable TFs' ranking scores during embryogenesis. (**b**) Histogram of the Spearman's correlations between PageRank scores and expression levels of 245 TFs across tissues and stages. 60% of the TFs show strong correlations (> 0.75). 9.3% of the TFs show weak or no correlation (< 0.25). (**c**) A lineage tree constructed from the ranking scores of the 245 TFs. (**d**) A lineage tree constructed from the gene expression levels, determined by RNA-seq, of the same 245 TFs in **c**. cf, craniofacial prominence; fb, forebrain; hb, hindbrain; ht, heart; in, intestine; kd, kidney; lm, limb; lv, liver; lu, lung; mb, midbrain; nt, neural tube; st, stomach.

prominence at E14.5 and E15.5 are clustered with limb samples. These results further demonstrate that PageRank is superior to gene expression as an indicator of TFs' activities.

We then asked whether there exist constitutively active TFs that exhibit high ranking scores across all tissues and stages. For this purpose, we first selected TFs with average ranking scores larger than 3×10^{-3}, corresponding to the top 10% of all TFs. Next, we retained TFs of which the coefficients of variance (*CV*s) are less than 0.5, giving us 35 constitutively active TFs in total (Fig. 4.6a). Similar to their transcriptional activities, the expression levels of these TFs remains relatively high across tissues and stages. Functional classification analysis revealed that these TFs are enriched in metabolic process, cellular process and developmental process, suggesting their general roles in basic cellular functions and embryonic development.

The earliest developmental stage that mouse ENCODE project has surveyed starts from E10.5. To obtain a complete view of mouse embryogenesis, we incorporated another published dataset which profiled the transcriptome and genome-wide chromatin accessibility in earlier mouse embryos[28], spanning stages from E1 (2-cell) to E5 (blastocyst stage). Together, we have identified potential driver TFs in twelve tissues from eight stages, as well as those in 2-cell, 4-cell, 8-cell embryos, inner cell masses (ICMs) and mouse embryonic stem cells (mESCs). For the first time, the complex transcriptional regulation during mouse embryonic development is systematically mapped (Fig. C.1 and C.2). The identified driver TFs include many well-known key regulators. For example, we successfully identified *Sox2*, *Nanog* and *Pou5f1* (also known as *Oct4*) as crucial TFs in mESCs. To systematically validate our predictions, we performed literature search in 5 tissues that have been extensively studied, including heart, lung, liver, kidney and limb. Remarkably, 41 out of 56 (73.2%) TFs identified by our method are shown to be associated with either development or disease of these tissues (see Appendix A). For instances, 9 out of 13 identified driver TFs in lung have been previously reported to play a pivotal role during the bronchiole tree and terminal alveolar region formation of mouse lung (Fig. 4.6b). And another two TFs relate with lung cancer[29, 30]. Apart from the spatial specificity (across

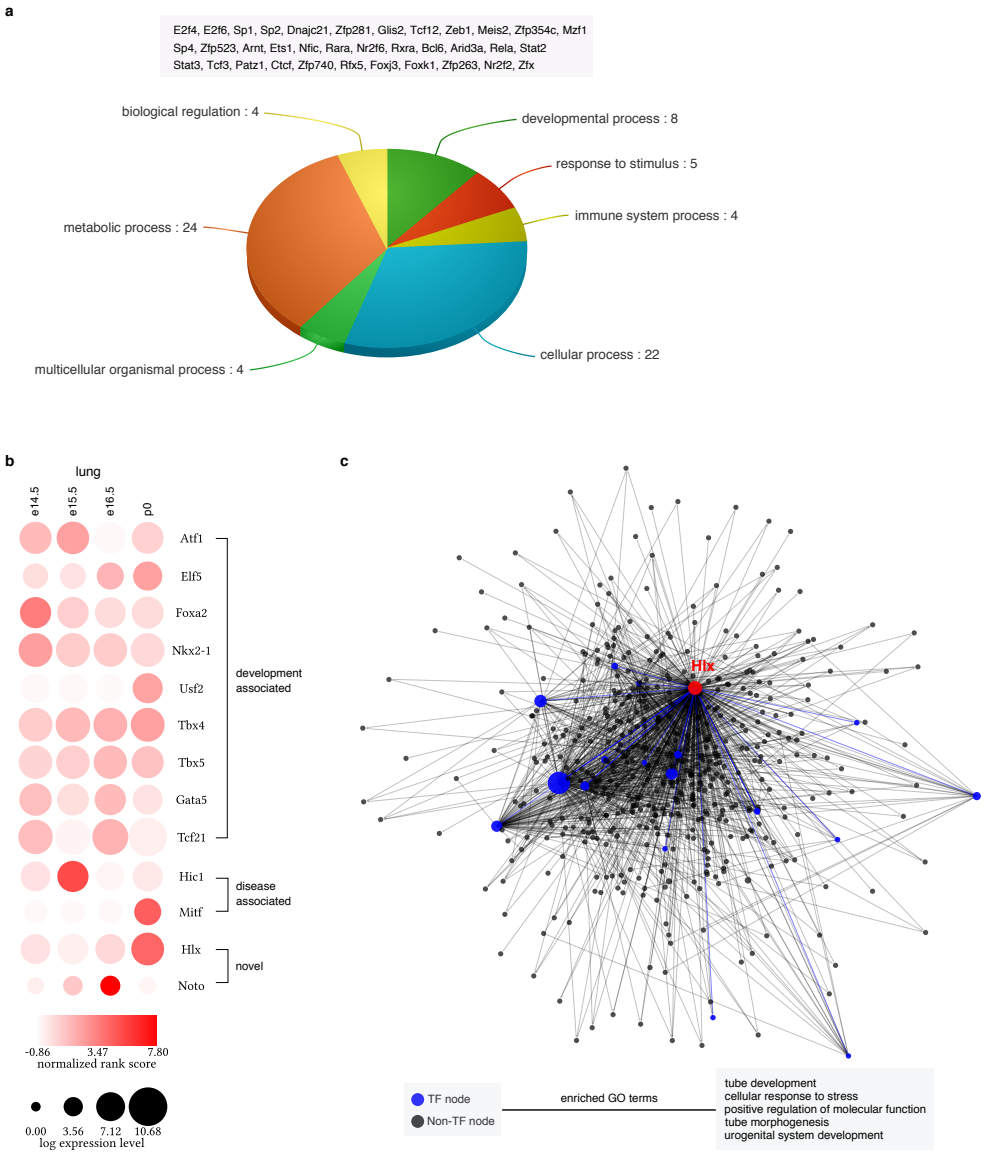

Figure 4.6: Identification of driver TFs during mouse embryogenesis. (**a**) Functional classification of 35 constitutively active TFs reveals their functions in essential biological processes. (**b**) Identified driver TFs in lung, including 9 TFs related with lung development, 2 TFs related with lung diseases and 2 novel TFs. (**c**) The network for *Hlx*, a novel driver TF for lung, and its regulatees in lung at p0. Larger nodes represent TFs with higher rank. The bottom shows the top 5 enriched GO terms for *Hlx*'s regulatees.

tissues), the temporal pattern of TFs' activity (across stages) can also be accurately captured by our approach. For instances, *Nkx2-1* is initiated at the early stage of lung development and *Nkx2-1* mutant embryos are arrested at early pseudoglandular (E11-E15) stage[31], which is consistent with our analysis. Similarly, another key regulator *Foxa2* is present in the epithelial cells from the beginning of lung bud formation, and transgenic mice with *Foxa2* ectopically expressed in the lung epithelial cells exhibited defects in branching morphogenesis[32].

We have also found many novel TFs. *Hlx*, identified as a key regulator for lung development, has not yet been studied in the lung. We showed that *Hlx* is regulating a number of high-rank TFs at P0, including several aforementioned constitutively active TFs, *i.e.*, *Sp1*, *Meis2* and *Zfx* (Fig. 4.6c). The functional enrichment analysis of all *Hlx*'s regulatees shows that they likely participate in epithelial tubes development in the lung. Besides, some of the predicted novel TFs have been studied in closely related tissues. For instances, *Vsx1*, predicted as an important regulator in the hindbrain, was related to retina development[33]. Taken together, these insights revealed by our analysis can guide future functional studies of the developmental mechanisms.

In development, ESCs differentiate into three germ layers (ectoderm, endoderm and mesoderm) and we have identified TFs that are specific to each layer (Fig. 4.7). For this purpose, tissues originated from the same layer were grouped and compared against tissues from other layers. The student t-tests with a p-value cutoff of 0.001 were used to identify TFs whose ranks change significantly in one germ layer compared with other layers. The functional relevance of the found layer-specific TFs is supported by the literature. For example, we have identified *Zic1*, *Zic4* and *Zic5* as specific regulators for ectoderm, which is in agreement with the role of *Zic* family in neural development[34]. In addition, many previously known layer-specific markers were also found from our analysis, e.g., *Pax6* and *Otx2* for ectoderm[35, 36], *Foxa2* for endoderm[37]. To analyze the tissue specificity of the driver TFs in Fig. C.1, we compared each tissue with other tissues originated from the same layer and used the student t-tests with a p-value cutoff of 0.001 to identify tissue-specific driver TFs, as shown in Fig. 4.7. Together these results provide a

comprehensive map for the future mechanistic study of embryonic development.

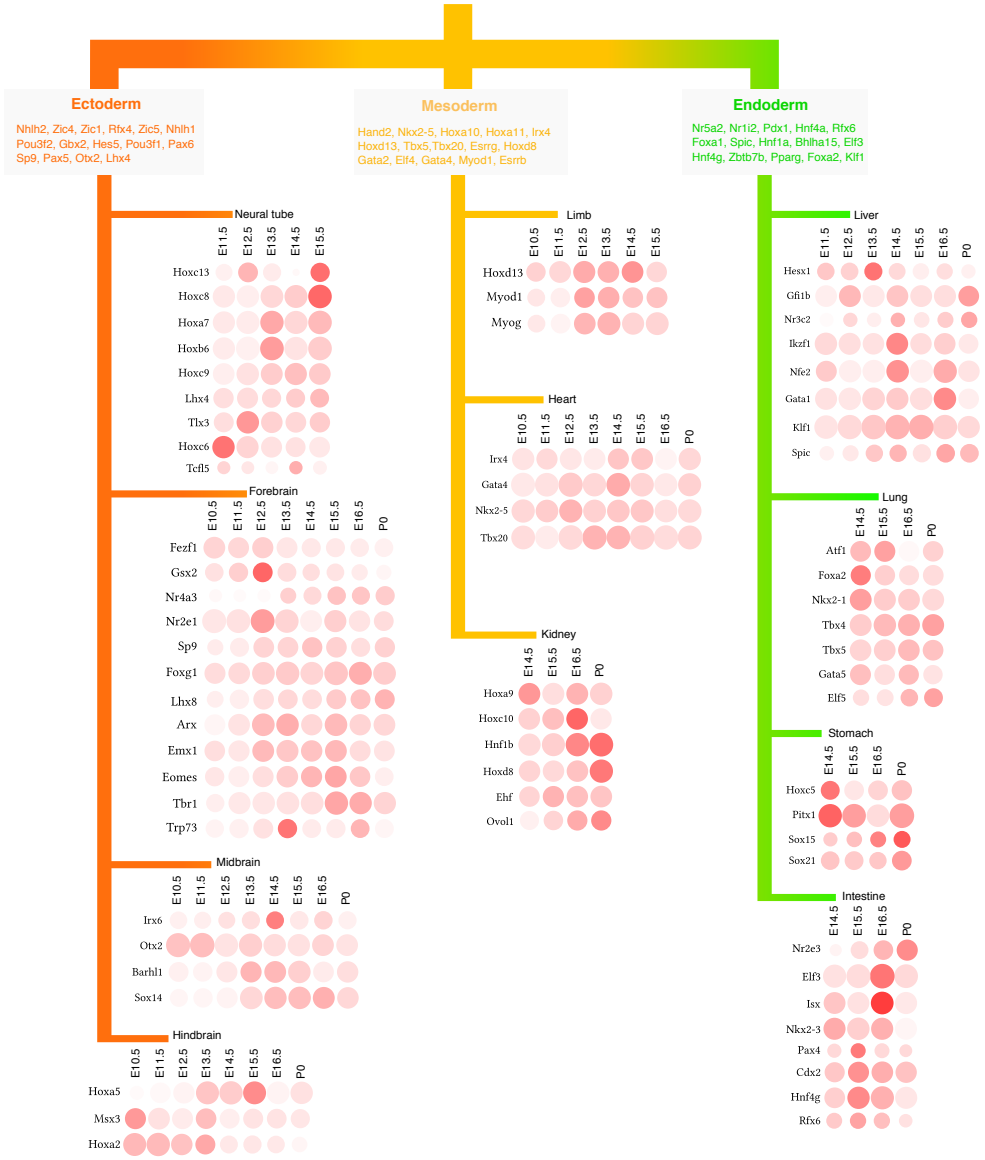

Figure 4.7: Germ-layer-specific and tissue-specific driver TFs in mouse embryonic development.

4.3.5 Transcriptional waves during embryogenesis

In addition to the tissue specificity, we have analyzed the temporal activity of TFs during the mouse embryogenesis. To identify clusters of TFs that show similar spatial-temporal patterns, we first performed the principal component analysis (PCA) to reduce the dimension of the TF

ranking score matrix. We reserved the first 20 components for clustering analysis as adding more components did not gain much explained variance (Fig. 4.8a). We used the silhouette analysis[38] to select the clustering method and the number of clusters. Among the five algorithms we tested, the k-means algorithm performed best, identifying 25 distinct dynamic patterns during embryogenesis (Fig. 4.8b).

The 25 clusters represent transcriptional waves that orchestrate the tissues differentiation (Fig. D.1). The first wave starts from as early as the 2-cell stage, represented by the cluster C7 (Fig. 4.9a). In C7, TFs show the highest activity at 2-cell stage (or possibly earlier as we do not have data in zygote). Example TFs from C7 include germ cell specific factors like *Obox1* and *Nr6a1*, both of which are essential for embryogenesis[39, 40], highlighting the roles of parental control in early development. As C7 wanes, C21 is turned on during 4-cell and 8-cell stages. And C13 and C16 emerge in ICM and ESC respectively. In C13 and C16, we found many well-known pluripotency regulators, such as *Pou5f1*, *Nanog* and *Sox2*. These results provide valuable insights into the transcriptional program during early embryogenesis.

We also found clusters that are responsible for the differentiation of tissues from specific layers. For example, C5 is highly active in all four brain tissues from E10.5 to P0, suggesting its critical role in neural differentiation. Indeed, many TFs that are crucial for neural development were recovered in C5, such as *Zic5*[34], *Pax6*[36] and *Gbx2*[41]. In contrast to C5, C1 is more active in mesoderm and endoderm-derived tissues. TFs in this cluster are associated with functions specific to the development of these two layers. For example, *Hnf4a* was reported to play pivotal roles in liver, colon and kidney development[42–44].

Besides germ-layer-specific transcriptional waves, there are also tissue-specific transcriptional programs that drive the differentiation of individual tissue. In Fig. 4.9c, we highlighted 4 clusters, C10, C12, C6 and C18, which are responsible for stomach/intestine, heart, forebrain and liver development, respectively. Note that we have found such a program for nearly every tissue, including craniofacial/limb (C9), lung (C3), kidney (C4), midbrain (C24) and neural

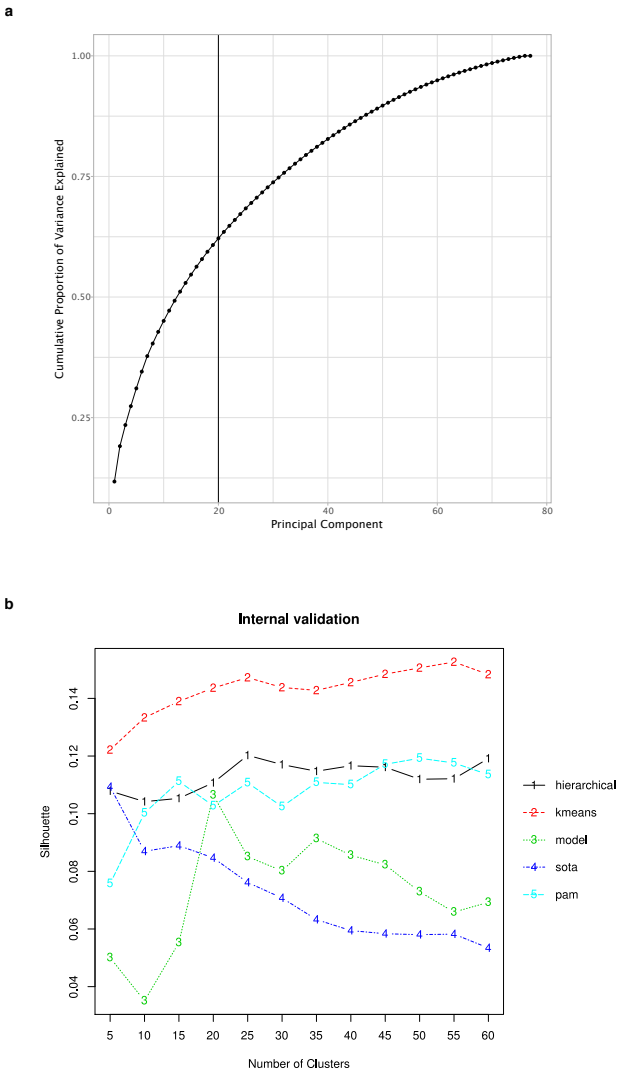

Figure 4.8: Selecting algorithms and parameters for clustering analysis. (**a**) Plotting the cumulative proportion of variance explained against the number of principal components (PCs). The first 20 PCs were kept according to the "elbow" method. (**b**) Selecting the best clustering algorithm and number of clusters according to the Silhouette metric. The k-means algorithm ($k = 25$) was picked using the "elbow" method. hierarchical, hierarchical clustering algorithm; kmeans, k-means algorithm; model, model-based algorithm; sota, self-organizing tree algorithm; pam, partitioning around medoids algorithm.

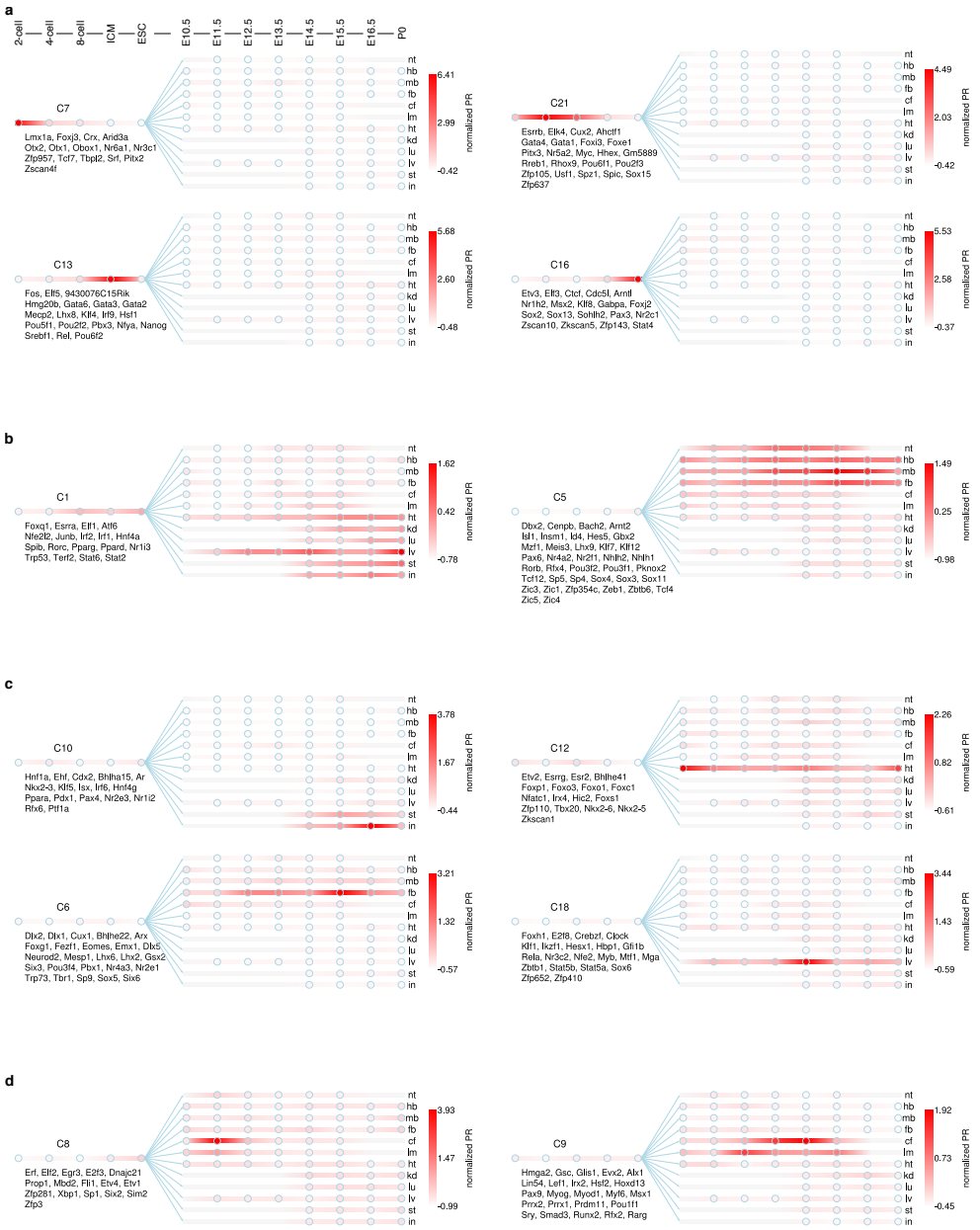

Figure 4.9: Temporal transcriptional waves direct tissue differentiation during embryogenesis. (**a**) Four waves in early embryonic development. (**b**) Layer specific transcriptional waves. (**c**) Four examples of tissue specific transcriptional waves. (**d**) Stage-specific waves in craniofacial development. Circles in each panel represent developmental stages. From left to right, they are 2-cell, 4-cell, 8-cell, ICM, ESC, E10.5, E11.5, E12.5, E13.5, E14.5, E15.5, E16.5 and P0 respectively. TF members are shown below the names of clusters.

tube/hindbrain (C19). See Fig. D.1 for details.

Most of the tissue-specific TFs are long-lived, playing roles in almost every stage of development. However, a few TFs exhibit transient transcriptional spikes, probably related to their stage-specific functions. In Fig. 4.9d we show two such examples which activate at different stages in craniofacial development (C8 for E11.5, C9 for E13.5 and E14.5). As the mechanism of craniofacial development remains largely unknown, our discovery, therefore, provides an invaluable resource to guide mechanistic studies.

4.4 Discussion

Here we present a novel and general framework for identifying driver TFs in any biological process. Our method Taiji is capable of flexibly integrating diverse genomic and epigenomic data, including ChIP-seq, RNA-seq, ATAC-seq and Hi-C. Considering the current limitation of Hi-C experiments, we provide a computational alternative when Hi-C experiment is infeasible or unavailable. In this work, we predicted 3D chromatin interactions in twelve tissues and eight developmental stages of the mouse embryo, which provides the first 3D chromatin organization information. By leveraging the strength of various experiments, we have successfully mapped lineage-, tissue- and stage-specific driver TFs throughout the mouse embryonic development, from as early as the 2-cell stage to postnatal day 0. In addition to retrieving known key regulators, we have also identified new TFs responsible for tissue differentiation and development progress, which can guide the future experimental investigations to understand the regulatory mechanisms of development.

Particularly interesting is the observation of transcriptional waves represented by TF combinations that activate in a spatiotemporal fashion. We did not find stage-specific tissue-independent TFs, *i.e.*, TFs that are active in all tissues at a specific stage (Fig. 4.5b). Therefore, we speculate that synchronization between tissues during development is not achieved by global

regulators, *i.e.*, there lacks a "central coordination", but by sequential activations of regulators in individual tissues through "distributed coordination" and use tissue-to-tissue communication to ensure synchronization between tissues, *i.e.*, tissues crosstalk and inform each other what is the current developmental stage. The distributed coordination alleviates the burden of developing a central authority during evolution and also provides a more robust timing strategy.

Chapter 4, in full, is a reprint of the material as it appears in Systems-level identification of transcription factors critical for mouse embryonic development. Zhang K, Wang M, Zhao Y, Wang W. Biorxiv 2017. The dissertation author was the primary investigator and author of this paper.

4.5 References

1. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. en. *Nat. Rev. Genet.* **13,** 613–626. ISSN: 1471-0056, 1471-0064 (Sept. 2012).

2. Filtz, T. M., Vogel, W. K. & Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. en. *Trends Pharmacol. Sci.* **35,** 76–85. ISSN: 0165-6147, 1873-3735 (Feb. 2014).

3. Bussemaker, H. J., Li, H. & Siggia, E. D. Regulatory element detection using correlation with expression. en. *Nat. Genet.* **27,** 167–171. ISSN: 1061-4036 (Feb. 2001).

4. Schacht, T., Oswald, M., Eils, R., Eichmüller, S. B. & König, R. Estimating the activity of transcription factors by the effect on their target genes. en. *Bioinformatics* **30,** i401–7. ISSN: 1367-4803, 1367-4811 (Jan. 2014).

5. ArrietaOrtiz, M. L., Hafemeister, C., Bate, A. R., Chu, T., Greenfield, A., Shuster, B., Barry, S. N., Gallitto, M., Liu, B., Kacmarczyk, T., Santoriello, F., Chen, J., DA Rodrigues, C., Sato, T., Rudner, D. Z., Driks, A., Bonneau, R. & Eichenberger, P. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. en. *Mol. Syst. Biol.* **11,** 839. ISSN: 1744-4292, 1744-4292 (Jan. 2015).

6. Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. en. *PLoS Comput. Biol.* **4,** e1000140. ISSN: 1553-734X, 1553-7358 (Jan. 2008).

7. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. en. *Nat. Rev. Genet.* **8,** 437–449. ISSN: 1471-0056 (June 2007).

8. 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. & Goldrath, A. W. Epigenetic landscapes reveal transcription factors that regulate CD8+ T cell differentiation. en. *Nat. Immunol.* **18,** 573–582. ISSN: 1529-2908, 1529-2916 (May 2017).

9. Milner, J. J., Toma, C., Yu, B., Zhang, K., Omilusik, K., Phan, A. T., Wang, D., Getzler, A. J., Nguyen, T., Crotty, S., Wang, W., Pipkin, M. E. & Goldrath, A. W. Runx3 programs CD8+ T cell residency in non-lymphoid tissues and tumours. *Nature* **552,** 253. ISSN: 0028-0836 (June 2017).

10. Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E. & Stamatoyannopoulos, J. A. Circuitry and dynamics of human transcription factor regulatory networks. en. *Cell* **150,** 1274–1286 (Sept. 2012).

11. Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., Anne, S. L., Doetsch, F., Colman, H., Lasorella, A., Aldape, K., Califano, A. & Iavarone, A. The transcriptional network for mesenchymal transformation of brain tumours. en. *Nature* **463,** 318–325 (Jan. 2010).

12. Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. & Califano, A. Reverse engineering of regulatory networks in human B cells. en. *Nat. Genet.* **37,** 382–390 (Apr. 2005).

13. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). en. *Genome Biol.* **9,** R137 (Sept. 2008).

14. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. en. *Nat. Biotechnol.* **28,** 495–501 (May 2010).

15. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. en. *Bioinformatics* **27,** 1017–1018 (Apr. 2011).

16. Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R. & Hughes, T. R. Determination and inference of eukaryotic transcription factor sequence specificity. en. *Cell* **158,** 1431–1443 (Sept. 2014).

17. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695,** 1–9 (2006).

18. Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. en. *Bioinformatics* **27,** 2263–2270 (Aug. 2011).

19. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. en. *Nature* **485,** 376–380. ISSN: 0028-0836, 1476-4687 (Nov. 2012).

20. Desper, R. & Gascuel, O. Getting a tree fast: Neighbor Joining, FastME, and distance-based methods. en. *Curr. Protoc. Bioinformatics* **Chapter 6,** Unit 6.3. ISSN: 1934-3396, 1934-340X (Oct. 2006).

21. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web (Bussum)* **54,** 1–17 (Nov. 1999).

22. Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., Ding, B., Li, N., Zheng, L. & Wang, W. Constructing 3D interaction maps from 1D epigenomes. en. *Nat. Commun.* **7,** 10812. ISSN: 2041-1723 (Oct. 2016).

23. Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R. & Hughes, T. R. Determination and inference of eukaryotic transcription factor sequence specificity. en. *Cell* **158,** 1431–1443. ISSN: 0092-8674, 1097-4172 (Nov. 2014).

24. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. en. *Mamm. Genome* **26,** 366–378. ISSN: 0938-8990, 1432-1777 (Oct. 2015).

25. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88–D92. ISSN: 0305-1048 (Jan. 2007).

26. Cushion, T. D., Dobyns, W. B., Mullins, J. G. L., Stoodley, N., Chung, S.-K., Fry, A. E., Hehr, U., Gunny, R., Aylsworth, A. S., Prabhakar, P., Uyanik, G., Rankin, J., Rees, M. I. & Pilz, D. T. Overlapping cortical malformations and mutations in TUBB2B and TUBA1A. en. *Brain* **136,** 536–548. ISSN: 0006-8950, 1460-2156 (Feb. 2013).

27. Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. en. *Bioinformatics* **27,** 2263–2270. ISSN: 1367-4803, 1367-4811 (15 8 2011).

28. Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., Xu, F.,

Chang, Z., Na, J., Yang, X. & Xie, W. The landscape of accessible chromatin in mammalian preimplantation embryos. en. *Nature* **534,** 652–657. ISSN: 0028-0836, 1476-4687 (30 6 2016).

29. Yu, S.-L., Chen, J. J. W., Chiu, S.-C., Chen, H.-Y., Chen, H.-W. & Yang, P.-C. Analysis of MITF expression and its downstream genes in lung cancer. en. *Cancer Res.* **64,** 396–397. ISSN: 0008-5472, 1538-7445 (Jan. 2004).

30. Wang, X., Wang, Y., Xiao, G., Wang, J., Zu, L., Hao, M., Sun, X., Fu, Y., Hu, G. & Wang, J. Hypermethylated in cancer 1(HIC1) suppresses non-small cell lung cancer progression by targeting interleukin-6/Stat3 pathway. en. *Oncotarget* **7,** 30350–30364. ISSN: 1949-2553 (24 5 2016).

31. Minoo, P., Su, G., Drum, H., Bringas, P. & Kimura, S. Defects in tracheoesophageal and lung morphogenesis in Nkx2.1(-/-) mouse embryos. *Dev. Biol.* **209,** 60–71. ISSN: 0012-1606 (1999).

32. Wan, H., Dingle, S., Xu, Y., Besnard, V., Kaestner, K. H., Ang, S. L., Wert, S., Stahlman, M. T. & Whitsett, J. A. Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *J. Biol. Chem.* **280,** 13809–13816. ISSN: 0021-9258 (2005).

33. Shi, Z., Jervis, D., Nickerson, P. E. B. & Chow, R. L. Requirement for the paired-like homeodomain transcription factor VSX1 in type 3a mouse retinal bipolar cell terminal differentiation. en. *J. Comp. Neurol.* **520,** 117–129. ISSN: 0021-9967, 1096-9861 (Jan. 2012).

34. Merzdorf, C. S. Emerging roles for zic genes in early development. en. *Dev. Dyn.* **236,** 922–940. ISSN: 1058-8388 (Apr. 2007).

35. Gammill, L. S. & Sive, H. otx2 expression in the ectoderm activates anterior neural determination and is required for Xenopus cement gland formation. en. *Dev. Biol.* **240,** 223–236. ISSN: 0012-1606 (Jan. 2001).

36. Zhang, X., Huang, C. T., Chen, J., Pankratz, M. T., Xi, J., Li, J., Yang, Y., Lavaute, T. M., Li, X.-J., Ayala, M., Bondarenko, G. I., Du, Z.-W., Jin, Y., Golos, T. G. & Zhang, S.-C. Pax6 is a human neuroectoderm cell fate determinant. en. *Cell Stem Cell* **7,** 90–100. ISSN: 1934-5909, 1875-9777 (Feb. 2010).

37. Burtscher, I. & Lickert, H. Foxa2 regulates polarity and epithelialization in the endoderm germ layer of the mouse embryo. en. *Development* **136,** 1029–1038. ISSN: 0950-1991 (Mar. 2009).

38. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65. ISSN: 0377-0427 (Jan. 1987).

39. Lan, Z.-J., Chung, A. C.-K., Xu, X., DeMayo, F. J. & Cooney, A. J. The embryonic function of germ cell nuclear factor is dependent on the DNA binding domain. en. *J. Biol. Chem.* **277,** 50660–50667. ISSN: 0021-9258 (27 12 2002).

40. Stein, P., Medvedev, S. & Schultz, R. M. OBOX Proteins Are Recruited During Oocyte Maturation and Are Essential for Early Development in Mouse. *Biol. Reprod.* **87,** 210–210. ISSN: 0006-3363 (2012).

41. Szabó, N.-E., Zhao, T., Zhou, X. & Alvarez-Bolado, G. The role of Sonic hedgehog of neural origin in thalamic differentiation in the mouse. en. *J. Neurosci.* **29,** 2453–2466. ISSN: 0270-6474, 1529-2401 (25 2 2009).

42. Lucas, B., Grigo, K., Erdmann, S., Lausen, J., Klein-Hitpass, L. & Ryffel, G. U. HNF4α reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma. *Oncogene* **24,** 6418. ISSN: 0950-9232 (13 6 2005).

43. Garrison, W. D., Battle, M. A., Yang, C., Kaestner, K. H., Sladek, F. M. & Duncan, S. A. Hepatocyte nuclear factor 4alpha is essential for embryonic development of the mouse colon. en. *Gastroenterology* **130,** 1207–1220. ISSN: 0016-5085 (Apr. 2006).

44. DeLaForest, A., Nagaoka, M., Si-Tayeb, K., Noto, F. K., Konopka, G., Battle, M. A. & Duncan, S. A. HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. en. *Development* **138,** 4143–4153. ISSN: 0950-1991, 1477-9129 (Oct. 2011).

Chapter 5

Concluding remarks

How to perform the integrated analysis of data generated from diverse assays has been a great challenge in the field of genomics study. In this thesis, we have demonstrated that the network model is an excellent approach for integrating different data sets, extracting information, and generating hypothesis. Particularly in Chapter 3 and 4, we developed a method – Taiji – to perform integrated analysis of multiple data, including ATAC-Seq, ChIP-Seq, RNA-Seq and HiC. Using this method, we constructed the TF regulatory network and identified driver TFs in different biological processes, including $CD8^+$ T cells development and mouse embryonic development. Many of these TFs cannot be found by traditional methods. We further validated several identified driver TFs by knock-down experiments. Together these demonstrate the advantages of our integrated analysis framework.

Taiji was originally developed to identify driver TFs, as the constructed network contains only TF-gene relationships. One of the future plans is to expand the network to include other types of connections, *e.g.*, protein-protein interactions. The ultimate goal of Taiji framework is to construct a "knowledge graph" representing our current understanding of biological systems, and to develop a "search engine" to summarize, extract, or search for information in the graph. In the following sections, I will discuss ideas for accomplishing this ambitious goal.

5.1 Constructing "knowledge graph"

To construct a "knowledge graph" containing much richer information, more data, including knowledge in the literature, need to be considered and incorporated.

5.1.1 Identifying gene-gene associations based on gene co-expression

The mRNA co-expression can be used to infer gene-gene associations. Many network construction methods are based on the analysis of gene co-expression[1, 2]. The gene-gene associations inferred from co-expression data can be used to fill in missing edges in the TF regulatory network. This will greatly expand our networks.

5.1.2 Augmenting the network using existing knowledge

A lot of knowledge of the molecular interactions is present in the literature. The renaissance of neural networks in machine learning research has fueled the development of new algorithms in natural language processing (NLP), which paves the way to better retrieve information from the literature. Using the NLP algorithms, we can identify molecular interactions from the literature and use them to expand the networks. Another possibility is to obtain such information directly from public databases, such as REACTOME. Most of the data in these databases is manually curated. So it can be treated as a more reliable source of information.

5.2 Developing novel gene ranking algorithms

As discussed before, the "knowledge graph" will contain multiple lines of evidence. Therefore, a sensible method for combining these pieces of information together is needed.

5.3 References

1. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. & Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. en. *BMC Bioinformatics* **7 Suppl 1,** S7. ISSN: 1471-2105 (20 3 2006).

2. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. en. *BMC Bioinformatics* **9,** 559. ISSN: 1471-2105 (29 12 2008).

Appendix A

Literature evidence supports identified driver TFs

Here we list the supporting evidence for the identified driver TFs in five tissues, including heart, limb, liver, lung and kidney. The rest of the tissues (neural tube, forebrain, hindbrain, midbrain, craniofacial, stomach and intestine) are either less studied or closely related to each other (functionally and spatially), which makes it very difficult to assess their tissue specificity.

A.1 Heart

Among the 11 predicted driver TFs in heart, 8 of them (72.7%) are supported by literature (Table A.1).

A.2 Limb

Among the 10 predicted driver TFs in limb, 8 of them (80%) are supported by literature (Table A.2).

Table A.1: Driver TFs in heart.

Predicted driver TF	Evidence
Etv2	Affect heart development[1]
Esx1	Unknown
Nkx2-6	Relate to congenital heart disease[2]
Bhlhe41	Unknown
Gata2	Relate to coronary artery disease[3]
Gata4	Essential for heart development[4]
Nkx2-5	Relate to heart development[5]
Tbx20	Essential for heart development[6]
Esrrg	Affect heart development[7]
Spz1	Unknown
Irx4	Affect heart development[8]

Table A.2: Driver TFs in limb.

Predicted driver TF	Evidence
Hoxa11	Relate to limb development[9]
Hoxa13	Relate to limb development [9]
Evx2	Important for limb development[10]
Msx1	Important for limb development[11]
Myod1	Important for limb development[12]
Myog	Important for limb development[13]
Sim2	Affect limb development[14]
Fos	Relate to limb mesenchymal chondrogenesis[15]
Rfx2	Unknown
Cebpg	Unknown

A.3 Liver

Among the 11 predicted driver TFs in liver, 7 of them (63.6%) are supported by literature (Table A.3).

A.4 Lung

Among the 13 predicted driver TFs in Lung, 11 of them (84.6%) are supported by literature (Table A.4).

Table A.3: Driver TFs in liver.

Predicted driver TF	Evidence
Gfi1b	Relate to liver development[16]
Nr3c2	Unknown
Ikzf1	Relate to liver cancer[17]
Nfe2	Important for liver function[18]
Gata1	Important for liver development[19]
Klf1	Relate to liver development[20]
Spic	Relate to liver development[21]
E2f2	Important for liver function[22]
Hesx1	Unknown
Zfp652	Unknown
Foxh1	Unknown

Table A.4: Driver TFs in lung.

Predicted driver TF	Evidence
Tcf21	Important for heart development[23]
Usf2	Relate to heart development[24]
Gata5	Important for heart development[25]
Foxa2	Important for heart development[26]
Nkx2-1	Essential for heart development[27, 28]
Mitf	Relate to heart cancer[29]
Elf5	Relate to heart development[30]
Tbx4	Important for heart development[31]
Hic1	Relate to heart cancer[32]
Atf1	Important for heart development[33]
Tbx5	Important for heart development[31]
Noto	Unknown
Hlx	Unknown

A.5 Kidney

Among the 10 predicted driver TFs in kidney, 7 of them (70%) are supported by literature (Table A.5).

Table A.5: Driver TFs in kidney.

Predicted driver TF	Evidence
Hoxd8	Relate to kidney development[34]
Osr2	Relate to early pectoral fin specification and pronephric development[35]
Ovol1	Relate to kidney development[36]
Hnf1b	Relate to kidney disease[37]
Bbx	Relate to kidney development[38]
Hoxa10	Important for kidney development[39]
Hoxc10	Important for kidney development [39]
Cic	Unknown
Hoxa9	Unknown
Ehf	Unknown

A.6 References

1. Oh, S.-Y., Kim, J. Y. & Park, C. The ETS Factor, ETV2: a Master Regulator for Vascular Endothelial Cell Development. en. *Mol. Cells* **38,** 1029–1036 (Dec. 2015).

2. Li, T., Liu, C., Xu, Y., Guo, Q., Chen, S., Sun, K. & Xu, R. Identification of candidate genes for congenital heart defects on proximal chromosome 8p. en. *Sci. Rep.* **6,** 36133 (Nov. 2016).

3. Connelly, J. J., Wang, T., Cox, J. E., Haynes, C., Wang, L., Shah, S. H., Crosslin, D. R., Hale, A. B., Nelson, S., Crossman, D. C., Granger, C. B., Haines, J. L., Jones, C. J. H., Vance, J. M., Goldschmidt-Clermont, P. J., Kraus, W. E., Hauser, E. R. & Gregory, S. G. GATA2 is associated with familial early-onset coronary artery disease. en. *PLoS Genet.* **2,** e139 (Aug. 2006).

4. Kuo, C. T., Morrisey, E. E., Anandappa, R., Sigrist, K., Lu, M. M., Parmacek, M. S., Soudais, C. & Leiden, J. M. GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. en. *Genes Dev.* **11,** 1048–1060 (Apr. 1997).

5. Schwartz, R. J. & Olson, E. N. Building the heart piece by piece: modularity of cis-elements regulating Nkx2-5 transcription. en. *Development* **126,** 4187–4192 (Oct. 1999).

6. Cai, X., Zhang, W., Hu, J., Zhang, L., Sultana, N., Wu, B., Cai, W., Zhou, B. & Cai, C.-L. Tbx20 acts upstream of Wnt signaling to regulate endocardial cushion formation and valve remodeling during mouse cardiogenesis. en. *Development* **140,** 3176–3187 (Aug. 2013).

7. Kwon, D.-H., Eom, G. H., Kee, H. J., Nam, Y. S., Cho, Y. K., Kim, D.-K., Koo, J. Y., Kim, H.-S., Nam, K.-I., Kim, K. K., Lee, I.-K., Park, S. B., Choi, H.-S. & Kook, H. Estrogen-related receptor gamma induces cardiac hypertrophy by activating GATA4. en. *J. Mol. Cell. Cardiol.* **65,** 88–97 (Dec. 2013).

8. Bruneau, B. G., Bao, Z. Z., Fatkin, D., Xavier-Neto, J., Georgakopoulos, D., Maguire, C. T., Berul, C. I., Kass, D. A., Kuroski-de Bold, M. L., de Bold, A. J., Conner, D. A., Rosenthal, N., Cepko, C. L., Seidman, C. E. & Seidman, J. G. Cardiomyopathy in Irx4-deficient mice is preceded by abnormal ventricular gene expression. en. *Mol. Cell. Biol.* **21,** 1730–1736 (Mar. 2001).

9. Ohgo, S., Itoh, A., Suzuki, M., Satoh, A., Yokoyama, H. & Tamura, K. Analysis of hoxa11 and hoxa13 expression during patternless limb regeneration in Xenopus. en. *Dev. Biol.* **338,** 148–157 (Feb. 2010).

10. Hérault, Y., Hraba-Renevey, S., van der Hoeven, F. & Duboule, D. Function of the Evx-2 gene in the morphogenesis of vertebrate limbs. en. *EMBO J.* **15,** 6727–6738 (Dec. 1996).

11. Bensoussan-Trigano, V., Lallemand, Y., Saint Cloment, C. & Robert, B. Msx1 and Msx2 in limb mesenchyme modulate digit number and identity. *Dev. Dyn.* **240,** 1190–1202 (2011).

12. Koishi, K., Zhang, M., McLennan, I. S. & Harris, A. J. MyoD protein accumulates in satellite cells and is neurally regulated in regenerating myotubes and skeletal muscle fibers. en. *Dev. Dyn.* **202,** 244–254 (Mar. 1995).

13. Meadows, E., Cho, J.-H., Flynn, J. M. & Klein, W. H. Myogenin regulates a distinct genetic program in adult muscle stem cells. en. *Dev. Biol.* **322,** 406–414 (Oct. 2008).

14. Havis, E., Coumailleau, P., Bonnet, A., Bismuth, K., Bonnin, M.-A., Johnson, R., Fan, C.-M., Relaix, F., Shi, D.-L. & Duprez, D. Sim2 prevents entry into the myogenic program by repressing MyoD transcription during limb embryonic myogenesis. en. *Development* **139,** 1910–1920 (June 2012).

15. Tufan, A. C., Daumer, K. M., DeLise, A. M. & Tuan, R. S. AP-1 transcription factor complex is a target of signals from both WnT-7a and N-cadherin-dependent cell-cell adhesion complex during the regulation of limb mesenchymal chondrogenesis. en. *Exp. Cell Res.* **273,** 197–203 (Feb. 2002).

16. Vassen, L., Beauchemin, H., Lemsaddek, W., Krongold, J., Trudel, M. & Möröy, T. Growth factor independence 1b (gfi1b) is important for the maturation of erythroid cells and the regulation of embryonic globin expression. en. *PLoS One* **9,** e96636 (May 2014).

17. Huo, Q., Ge, C., Tian, H., Sun, J., Cui, M., Li, H., Zhao, F., Chen, T., Xie, H., Cui, Y., Yao, M. & Li, J. Dysfunction of IKZF1/MYC/MDIG axis contributes to liver cancer progression through regulating H3K9me3/p21 activity. en. *Cell Death Dis.* **8,** e2766 (May 2017).

18. Hirotsu, Y., Hataya, N., Katsuoka, F. & Yamamoto, M. NF-E2-related factor 1 (Nrf1) serves as a novel regulator of hepatic lipid metabolism through regulation of the Lipin1 and PGC-1β genes. en. *Mol. Cell. Biol.* **32,** 2760–2770 (July 2012).

19. Pevny, L., Lin, C. S., D'Agati, V., Simon, M. C., Orkin, S. H. & Costantini, F. Development of hematopoietic cells lacking transcription factor GATA-1. en. *Development* **121,** 163–172 (Jan. 1995).

20. Yien, Y. Y. & Bieker, J. J. EKLF/KLF1, a tissue-restricted integrator of transcriptional control, chromatin remodeling, and lineage determination. en. *Mol. Cell. Biol.* **33,** 4–13 (Jan. 2013).

21. Kohyama, M., Ise, W., Edelson, B. T., Wilker, P. R., Hildner, K., Mejia, C., Frazier, W. A., Murphy, T. L. & Murphy, K. M. Role for Spi-C in the development of red pulp macrophages and splenic iron homeostasis. en. *Nature* **457,** 318–321 (Jan. 2009).

22. Delgado, I., Fresnedo, O., Iglesias, A., Rueda, Y., Syn, W.-K., Zubiaga, A. M. & Ochoa, B. A role for transcription factor E2F2 in hepatocyte proliferation and timely liver regeneration. en. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G20–31 (July 2011).

23. Xu, Y., Wang, Y., Besnard, V., Ikegami, M., Wert, S. E., Heffner, C., Murray, S. A., Donahue, L. R. & Whitsett, J. A. Transcriptional programs controlling perinatal lung maturation. en. *PLoS One* **7,** e37046 (Aug. 2012).

24. Gao, E., Wang, Y., Alcorn, J. L. & Mendelson, C. R. Transcription factor USF2 is developmentally regulated in fetal lung and acts together with USF1 to induce SP-A gene expression. en. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **284,** L1027–36 (June 2003).

25. Chen, B., Moore, T. V., Li, Z., Sperling, A. I., Zhang, C., Andrade, J., Rodriguez, A., Bahroos, N., Huang, Y., Morrisey, E. E., Gruber, P. J. & Solway, J. Gata5 deficiency causes airway constrictor hyperresponsiveness in mice. en. *Am. J. Respir. Cell Mol. Biol.* **50,** 787–795 (Apr. 2014).

26. Chung, C., Kim, T., Kim, M., Kim, M., Song, H., Kim, T.-S., Seo, E., Lee, S.-H., Kim, H., Kim, S. K., Yoo, G., Lee, D.-H., Hwang, D.-S., Kinashi, T., Kim, J.-M. & Lim, D.-S. Hippo-Foxa2 signaling pathway plays a role in peripheral lung maturation and surfactant homeostasis. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 7732–7737 (May 2013).

27. Yin, Z., Gonzales, L., Kolla, V., Rath, N., Zhang, Y., Lu, M. M., Kimura, S., Ballard, P. L., Beers, M. F., Epstein, J. A. & Morrisey, E. E. Hop functions downstream of Nkx2.1 and GATA6 to mediate HDAC-dependent negative regulation of pulmonary gene expression. en. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **291,** L191–9 (Aug. 2006).

28. Herriges, M. & Morrisey, E. E. Lung development: orchestrating the generation and regeneration of a complex organ. en. *Development* **141,** 502–513 (Feb. 2014).

29. Yu, S.-L., Chen, J. J. W., Chiu, S.-C., Chen, H.-Y., Chen, H.-W. & Yang, P.-C. Analysis of MITF expression and its downstream genes in lung cancer. en. *Cancer Res.* **64,** 396–397 (Apr. 2004).

30. Metzger, D. E., Stahlman, M. T. & Shannon, J. M. Misexpression of ELF5 disrupts lung branching and inhibits epithelial differentiation. en. *Dev. Biol.* **320,** 149–160 (Aug. 2008).

31. Arora, R., Metzger, R. J. & Papaioannou, V. E. Multiple roles and interactions of Tbx4 and Tbx5 in development of the respiratory system. en. *PLoS Genet.* **8,** e1002866 (Aug. 2012).

32. Wang, X., Wang, Y., Xiao, G., Wang, J., Zu, L., Hao, M., Sun, X., Fu, Y., Hu, G. & Wang, J. Hypermethylated in cancer 1(HIC1) suppresses non-small cell lung cancer progression by targeting interleukin-6/Stat3 pathway. en. *Oncotarget* **7,** 30350–30364 (May 2016).

33. Bird, A. D., Flecknoe, S. J., Tan, K. H., Olsson, P. F., Antony, N., Mantamadiotis, T., Mollard, R., Hooper, S. B. & Cole, T. J. cAMP response element binding protein is required for differentiation of respiratory epithelium during murine development. en. *PLoS One* **6,** e17843 (Mar. 2011).

34. Di-Poï, N., Zákány, J. & Duboule, D. Distinct roles and regulations for HoxD genes in metanephric kidney development. en. *PLoS Genet.* **3,** e232 (Dec. 2007).

35. Lam, P.-Y., Kamei, C. N., Mangos, S., Mudumana, S., Liu, Y. & Drummond, I. A. odd-skipped related 2 is required for fin chondrogenesis in zebrafish. en. *Dev. Dyn.* **242,** 1284–1292 (Nov. 2013).

36. Teng, A., Nair, M., Wells, J., Segre, J. A. & Dai, X. Strain-dependent perinatal lethality of Ovol1-deficient mice and identification of Ovol2 as a downstream target of Ovol1 in skin epidermis. en. *Biochim. Biophys. Acta* **1772,** 89–95 (Jan. 2007).

37. Adalat, S., Woolf, A. S., Johnstone, K. A., Wirsing, A., Harries, L. W., Long, D. A., Hennekam, R. C., Ledermann, S. E., Rees, L., van't Hoff, W., Marks, S. D., Trompeter, R. S., Tullus, K., Winyard, P. J., Cansick, J., Mushtaq, I., Dhillon, H. K., Bingham, C., Edghill, E. L., Shroff, R., Stanescu, H., Ryffel, G. U., Ellard, S. & Bockenhauer, D. HNF1B mutations associate with hypomagnesemia and renal magnesium wasting. en. *J. Am. Soc. Nephrol.* **20,** 1123–1131 (May 2009).

38. Brunskill, E. W., Lai, H. L., Jamison, D. C., Potter, S. S. & Patterson, L. T. Microarrays and RNA-Seq identify molecular mechanisms driving the end of nephron production. en. *BMC Dev. Biol.* **11,** 15 (Mar. 2011).

39. Wellik, D. M. Hox genes and kidney development. en. *Pediatr. Nephrol.* **26,** 1559–1565 (Sept. 2011).

Appendix B

The roles of node weights and edge weights in Taiji's ranking algorithm

B.1 Node weights

The node weights are computed from genes' expression levels (see "Determine node weights" for details). The addition of node weights to the network allows genes with higher expression levels passing more "information" to their upstream regulators (Fig. B.1).

B.2 Edges weights

The edge weights are calculated from TFs' expression levels (see "Determine edge weights" for details). A gene is usually regulated by multiple TFs. We set the "information" passed from a gene to its regulators proportional to regulators' expression levels (Fig. B.2). This feature is particularly useful when there exists TFs that have similar motifs, e.g., TFs from the same family. In this case, they will be differentiated by their expression levels. Furthermore, the edge weights can help remove spurious driver TFs with low or no expression from our candidate list.

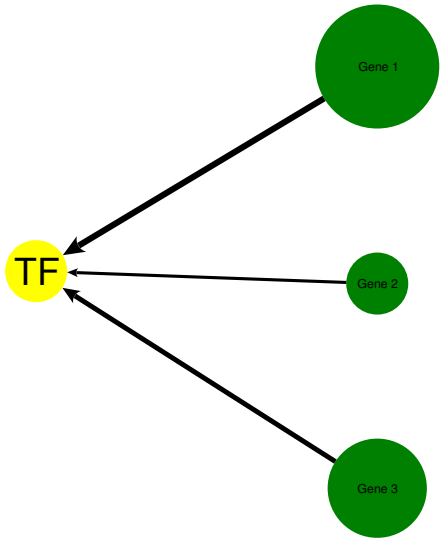

Figure B.1: Genes with higher expression, represented by circles with larger sizes, pass more "information", denoted by thicker edges, to their upstream regulators.

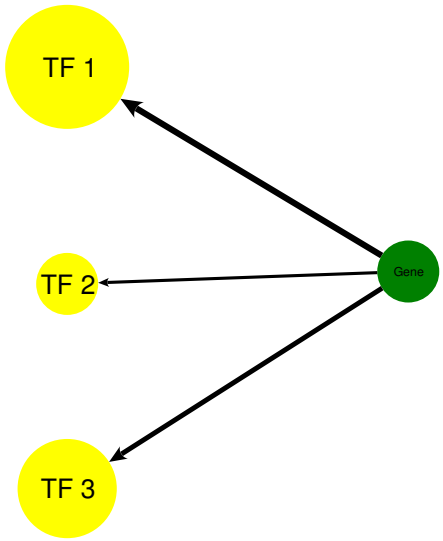

Figure B.2: TFs with higher expression receive more "information" from their target genes.

B.3 Advantages of Taiji's ranking algorithm over simple motif enrichment analyses.

The motif enrichment analysis only counts the number of binding sites for a given TF, regardless of the properties of its regulatees. In Fig. B.3, we showed that this simple assumption

is problematic in biological networks. Assuming that in a real network, TF1 regulates TF2 and TF3; TF2 regulates Gene1 and Gene2; TF3 regulates Gene3, Gene4 and Gene5. Fig. B.3a shows that Taiji takes into account the hierarchical structure of the network, giving TF1 the highest rank. For the same network, the motif enrichment analysis will decide that TF3 is most enriched because it has the most regulatees (Fig. B.3b). From the biological perspective, we think that Taiji's ranking algorithm is more sensible, because:

1. The regulatees of TF1, i.e., TF2 and TF3, are more important to the network than those of TF3, i.e., Gene3, Gene4 and Gene5.

2. Disrupting TF1 would affect all genes in the network, including TF3, while disrupting TF3 only affects three genes.

Another advantage of Taiji's ranking algorithm is that it considers the expression levels of TFs and their regulatees, which is used to filter spurious TF-gene links. On the contrary, the motif enrichment analysis normally does not consider the functional relevance of predicted binding sites.

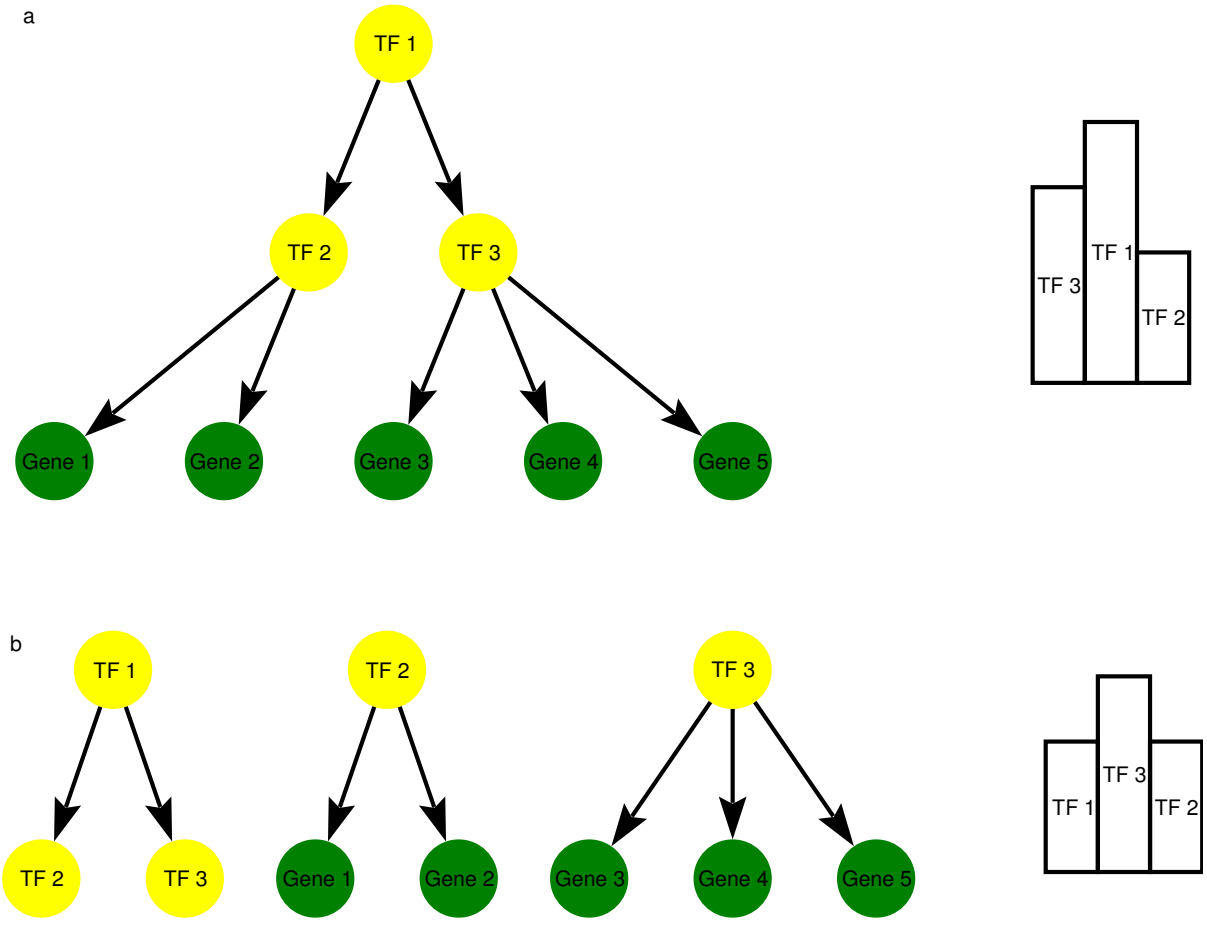

Figure B.3: Difference between Taiji and motif enrichment analysis for ranking TFs. **a**, Taiji's view of transcriptional regulatory network. The ranking: TF1 > TF3 > TF2; **b**, A flat view of transcriptional regulatory network in the motif enrichment analysis. The ranking: TF3 > TF1 = TF2.

Appendix C

A list of driver TFs during mouse embryogenesis

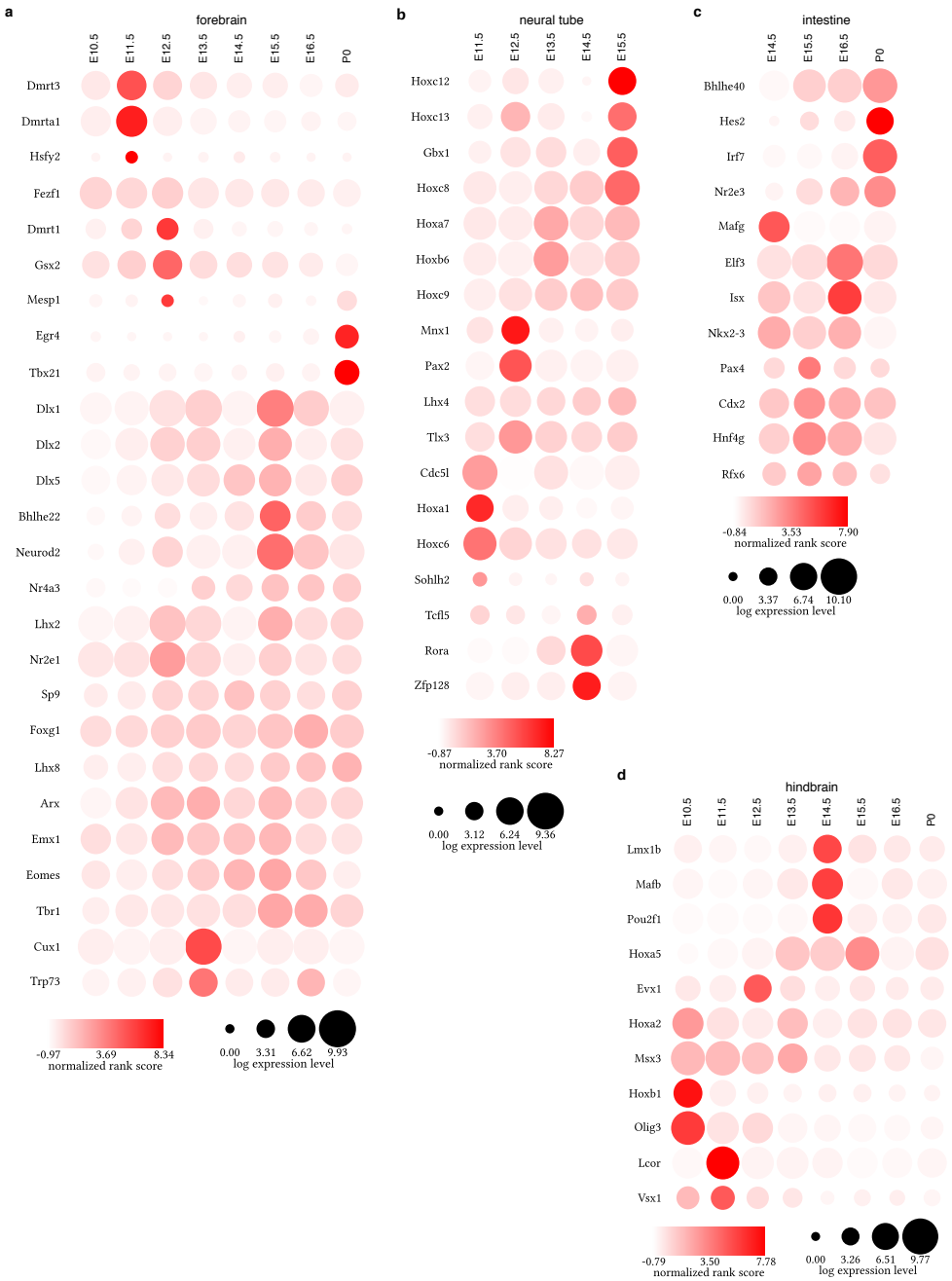

Figure C.1: Identification of driver TFs in twelve tissues. *(cont.)* (**a**) forebrain. (**b**) neural tube. (**c**) intestine. (**d**) hindbrain. (**e**) midbrain. (**f**) stomach. (**g**) lung. (**h**) limb. (**i**) kidney. (**j**) heart. (**k**) liver. (**l**) craniofacial prominence.

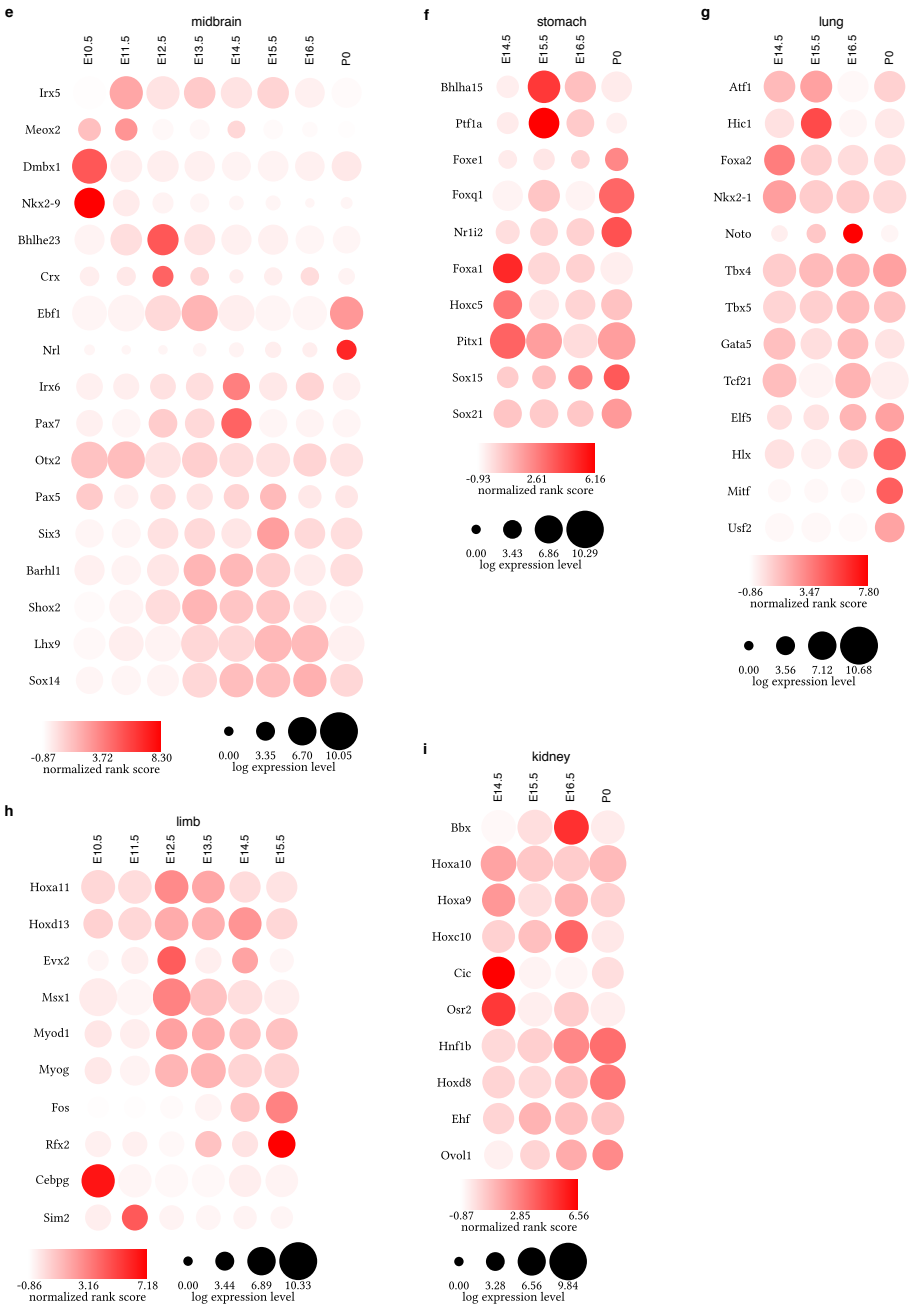

Figure C.1: Identification of driver TFs in twelve tissues. *(cont.)* (**a**) forebrain. (**b**) neural tube. (**c**) intestine. (**d**) hindbrain. (**e**) midbrain. (**f**) stomach. (**g**) lung. (**h**) limb. (**i**) kidney. (**j**) heart. (**k**) liver. (**l**) craniofacial prominence.

Figure C.1: Identification of driver TFs in twelve tissues. (**a**) forebrain. (**b**) neural tube. (**c**) intestine. (**d**) hindbrain. (**e**) midbrain. (**f**) stomach. (**g**) lung. (**h**) limb. (**i**) kidney. (**j**) heart. (**k**) liver. (**l**) craniofacial prominence.

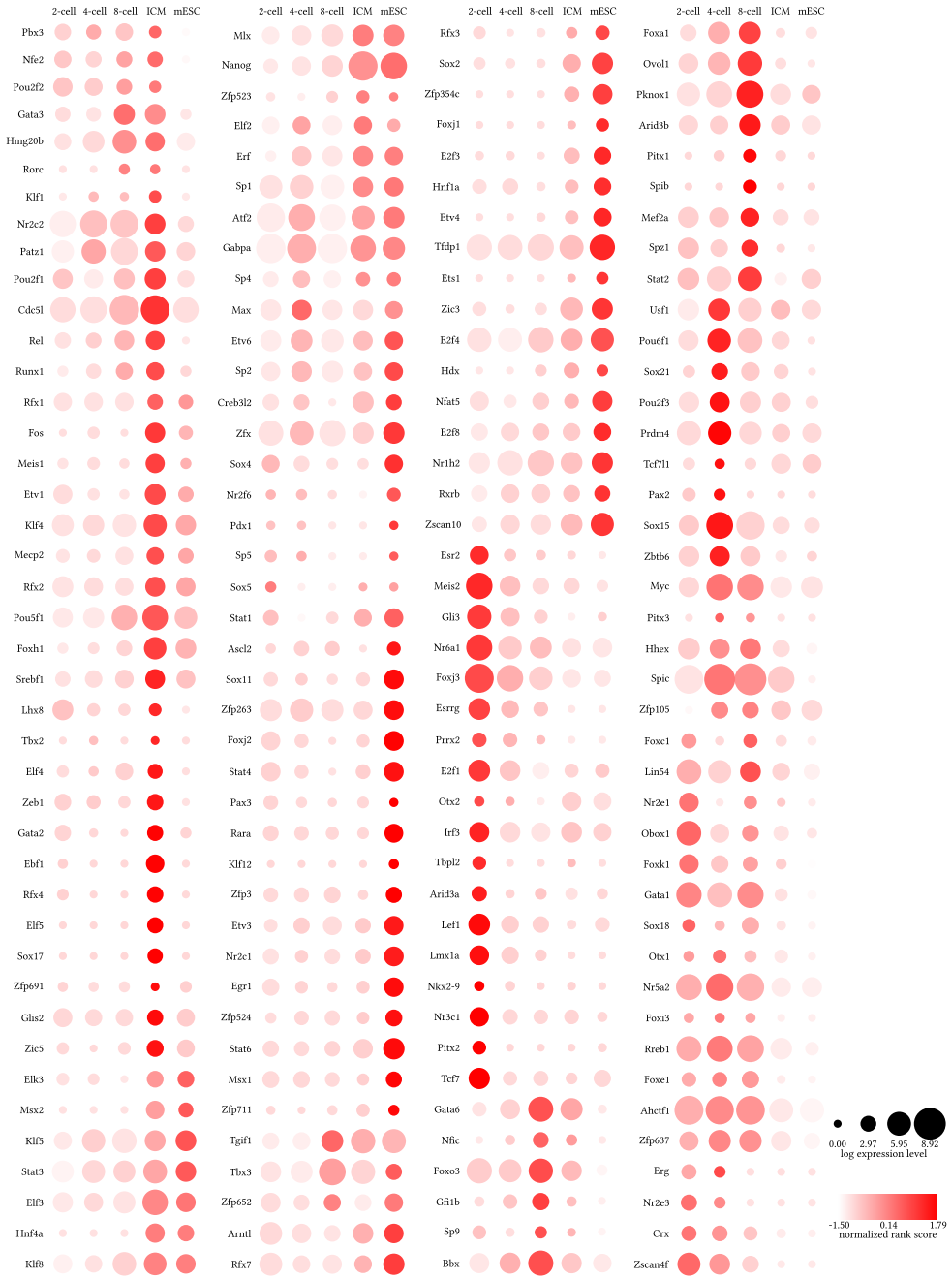

Figure C.2: Identification of driver TFs in early mouse embryonic development, including 2-cell, 4-cell, 8-cell stages, ICM and ESC.

Appendix D

A total of 25 transcriptional waves during mouse embryogenesis

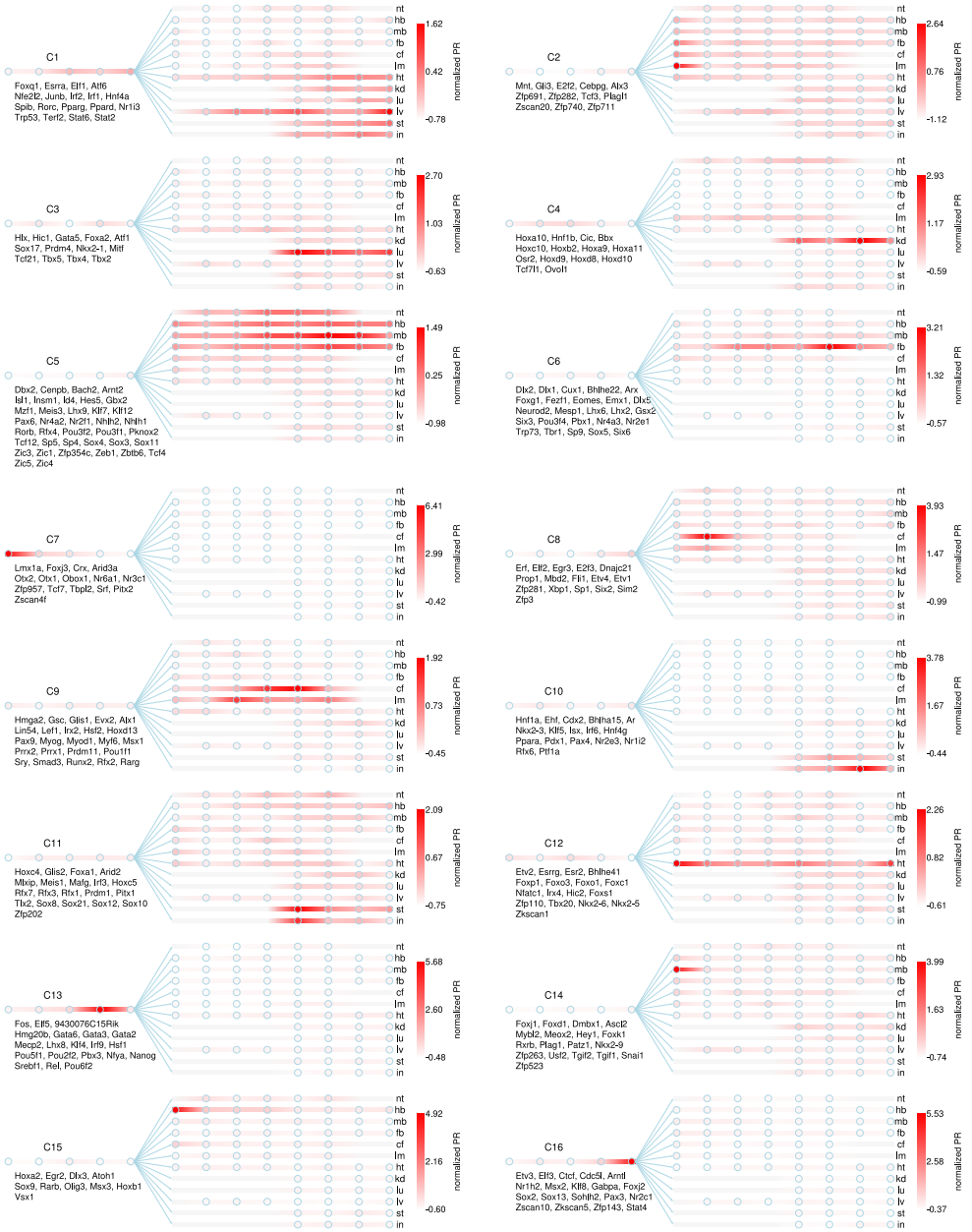

Figure D.1: Transcriptional waves direct tissue differentiation during mouse embryogenesis. *(cont.)* A total of 25 dynamic patterns of ranking scores were identified using the k-means clustering algorithm. Circles in each panel represent developmental stages. From left to right, they are 2-cell, 4-cell, 8-cell, ICM, ESC, E10.5, E11.5, E12.5, E13.5, E14.5, E15.5, E16.5 and P0 respectively. TF members are shown below the names of clusters.

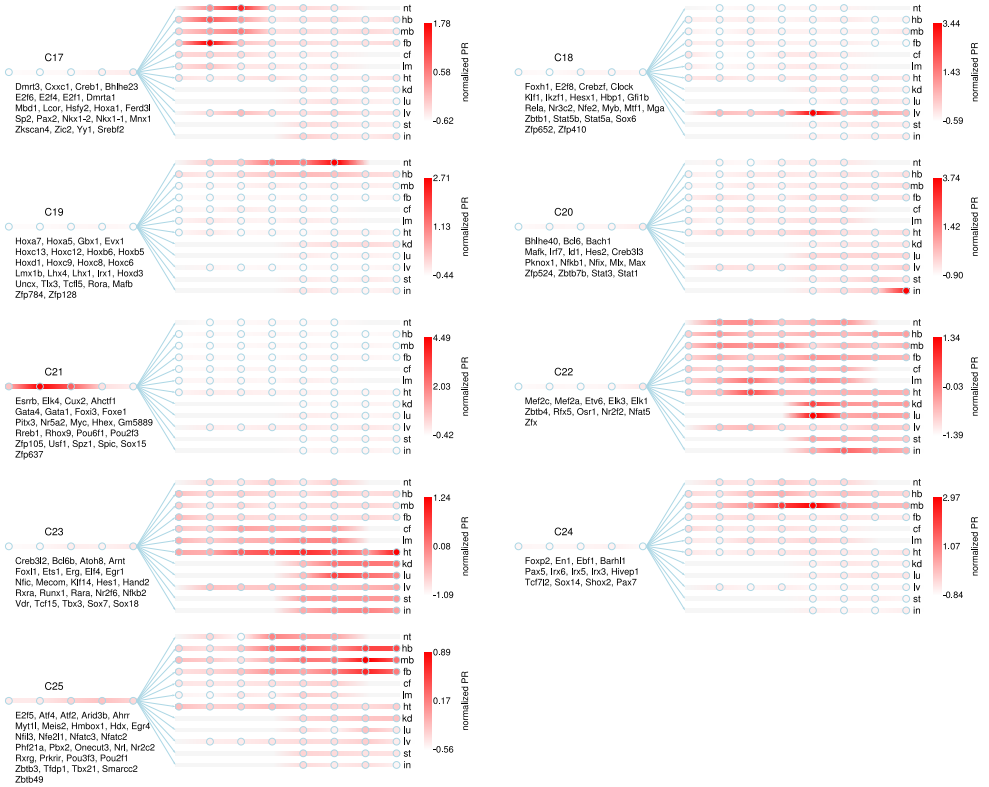

Figure D.1: Transcriptional waves direct tissue differentiation during mouse embryogenesis. A total of 25 dynamic patterns of ranking scores were identified using the k-means clustering algorithm. Circles in each panel represent developmental stages. From left to right, they are 2-cell, 4-cell, 8-cell, ICM, ESC, E10.5, E11.5, E12.5, E13.5, E14.5, E15.5, E16.5 and P0 respectively. TF members are shown below the names of clusters.